IMPROVEMENTS IN FINE-SCALE ESTIMATION AND EVALUATION OF GEOGRAPHIC VARIABLES USING CLIMATE DATA IN EAST AFRICA

By

Sarah L. Hession

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Geography

2011

ABSTRACT

IMPROVEMENTS IN FINE-SCALE ESTIMATION AND EVALUATION OF GEOGRAPHIC VARIABLES USING CLIMATE DATA IN EAST AFRICA

By

Sarah L. Hession

Global environmental change has surfaced as a critical issue to both the scientific community and the general public. One aspect of particular concern involves climate change, which will exert impacts on ecosystems and economies, presenting considerable challenge to human adaptation. In Africa, a continent that is vulnerable due to multiple stressors and low adaptive capacity, climate change is expected to significantly affect both people and ecosystems. Adaptation strategies are being developed using information from studies that evaluate the impacts of climate variability and climate change in Africa. Recommendations are made for local development of adaptation strategies due to the heterogeneity of climate change and its effects on East Africa's climate. However, global climate change models are coarse in scale and mask much of the local variation in regional climate, indicating the need for higher resolution climate data. This dissertation addresses this need by comparing spatially explicit statistical methods of interpolation and prediction, both theoretically and empirically; expanding upon the method of universal kriging by incorporating complex feedback relationships that may produce simultaneity between precipitation and its covariates; and evaluating precipitation patterns over space in East Africa through a case study. Mechanisms of precipitation have been considered in detail, expanding upon many other spatially explicit applications of prediction methods to date. Further, spatially

explicit inferential regression models have been developed to better understand spatial patterns and variability in East African precipitation. Predicted maps of precipitation, generated at a resolution of 1 kilometer, accurately reflect the mesoscale influences of topography and the presence of large water bodies (i.e., Lake Victoria) as well as the seasonal influences of the passing of the intertropical convergence zone (ITCZ). In terms of prediction, the spatially explicit methods considered herein clearly outperformed a global data set (i.e., the CRU TS 3.1) in terms of error and ability to reflect local variability. The method of local ordinary kriging generally outperformed the multivariate kriging techniques, indicating that precipitation patterns in areas of high topographic variability, such as East Africa, may be modeled as well or better using local search neighborhoods in the kriging process rather than using complex multivariate regression models. However, additional work to improve the multivariate regression models and overall levels of correlation are expected to yield improved prediction results. Furthermore, the case study successfully demonstrated that the newly developed method of universal kriging with instrumental variables performs similarly to other standard methods of estimation, and perhaps better in the presence of significant measurable simultaneity.

DEDICATION

This dissertation is dedicated to my children, Maura Elizabeth Hession and Patrick Joseph Hession III, who have endured this process with me and have made it all worthwhile.

ACKNOWLEDGEMENTS

I would like to acknowledge my committee for their direction and insight into my dissertation: Dr. Robert T. Walker (chair), Dr. Ashton Shortridge, Dr. Jennifer Olson, and Dr. Jeff Andresen. I would also like to acknowledge the following for their invaluable input and direction: Dr. David Campbell, Dr. Nathan Moore, Dr. Gopal Alagarswamy, and Dr. John Kern. The work presented herein was supported by The Graduate School at Michigan State University, the Climate Land Interaction Program (CLIP; *An Integrated Analysis of Regional Land-Climate Interactions*, National Science Foundation Award Number BCS0308420) and East Africa Climate Land Interaction Program – Savanna Ecosystems (EACLIPSE; *Dynamic Interactions Among People, Livestock, and Savanna Ecosystems under Climate Change*, National Science Foundation Award Number BCS0709671).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Modeling a changing climate

Global environmental change has surfaced as a critical issue to both the scientific community and the general public. One aspect of particular concern involves climate change, which will exert impacts on ecosystems and economies, presenting considerable challenge to human adaptation. Widespread impacts due to climate change are expected in many regions of the world (Giorgi 2001; Hulme 1998; Lobell et al. 2008).

In Africa, climate change is expected to result in warmer temperatures and changes in precipitation patterns, significantly affecting both people and ecosystems (Moore et al. 2005, 2006). According to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC; Boko et al. 2007), “Africa is one of the most vulnerable continents to climate change and climate variability, a situation aggravated by the interaction of ‘multiple stresses’, occurring at various levels, and low adaptive capacity.” Some of these stresses and vulnerabilities are described below, with a focus on eastern Africa.

In East Africa, studies have projected warmer and wetter conditions (Hulme et al. 2001, Moore et al. in review) with possible decreases in precipitation during the northern hemisphere summer months of June, July, and August (Giorgi 2001; Hulme 1998). Warmer temperatures combined with higher potential evapotranspiration and variable

precipitation may already be causing decreased productivity of much of the East African savanna (Moore et al. 2005).

Total rainfall in most areas has not significantly changed in the last 50 years, although increases in interannual variability have been observed after the 1970s (Boko et al. 2007, Fauchereau et al. 2003, Richard et al. 2001). Increased variability in precipitation is due, in part, to changing climatic patterns. Two such climatic patterns are the El-Nino Southern Oscillation (ENSO) and the Indian Ocean Dipole. ENSO has experienced a marked decrease in frequency, concurrent with an increase in the intensity and duration of this phenomenon (Camberlin et al. 2001, Mukabana and Pielke 1996). ENSO has a large impact on the variability of precipitation patterns; the impact is different based on the phase of ENSO that is occurring. The two phases of ENSO, the negative or warm phase (El Nino) and the positive or cool phase (La Nina), generally influence the timing of seasonal precipitation in opposite ways (Majugu and Magezi 1985). In La Nina years, the secondary rainy season generally occurs early in the season (i.e., the rainy season begins in August or early September and ends in October or early November). The end of this early rainy season is usually followed by a marked decrease in rainfall, resulting in droughts throughout much of the region (Majugu and Magezi 1985). Conversely, a typical El Nino year is characterized by a late onset of the second rainy season (i.e., the rainy season begins in late September or early October and ends in late November to early to mid December), with intense rains and potential flooding in parts of the region (Majugu and Magezi 1985).

Distinct from ENSO, an intensifying dipole rainfall pattern has been identified that occurs in a decadal pattern. The dipole is marked by increasing rainfall over the

northern sector and declining amounts over the southern sector (Boko et al. 2007, Schreck and Semazzi 2004). This increased variability of precipitation over both time and space greatly impacts water availability, crops and vegetation, farmers and others who rely on these crops, livestock and wildlife that rely on vegetation and water availability, as described below.

The eastern African region, like many regions within the tropics, experiences a disproportionate share of climate extremes (Majugu and Magezi 1985). Changes in the frequency and magnitude of extreme events, such as droughts and floods, have major implications for numerous Africans. For many years, droughts caused "human migration, cultural separation, population dislocation and the collapse of prehistoric and early historic societies" (Pandey et al. 2003). Major impacts of drought include decreased food production and shortages of potable water, which will be discussed next.

The long-term relationship between climate and vegetation in the region has been examined by Mworia-Maitima (1991, 1999) and colleagues, who show a clear relationship between plant species composition and climate. Future climate change may dramatically affect agricultural production across space and time; for example, the length of the growing season in East Africa could increase in some areas and decrease in others primarily as a result of altered precipitation amounts and timing (Jones and Thornton 2003).

Scholes and Biggs (2004) refer to Sub-Saharan Africa as the "food crisis epicenter of the world," concluding that food security will worsen during the first half of the twenty-first century due to projected changes in climate. Further challenges in

ensuring food security around the world, particularly in Africa, are expected due to the impacts of climate change on agriculture (Thornton et al. 2009). In East Africa, much of the population is largely dependent on rainfed cropping and pasture (approximately 80% of the population are agriculturalists); consequently, changes in productivity are expected to have a profound effect (Thornton et al. 2009)]. Vulnerability mapping has been used to identify areas that are presently vulnerable, both environmentally and socially, that are expected to be severely affected by climate change (Thornton et al. 2009). Parts of eastern Africa, such as arid-semiarid rangelands and coastal regions, are included as "hotspots" in these vulnerability maps (Thornton et al. 2009).

Climate change and variability are also expected to impose additional pressures on water availability, water accessibility, and water demand in Africa (Boko et al. 2007). Currently, about 25% of the African population experiences high water stress. Furthermore, one-third of the people in Africa live in drought-prone areas and are vulnerable to the effects of droughts (World Water Forum 2000). Although access to freshwater was improved during the 1990s, access to improved water supplies was available to only about 62% of the African population in 2000 (Boko et al. 2007, WHO/UNICEF 2000). Women would be particularly affected by changes in the location, quality and quantity of available domestic water as well as fuelwood, declining crop yields, and changes in animal health and productivity (Wangui 2003, 2004).

For many Africans, adaptation to climate change is not an option but a necessity (e.g., Thornton et al. 2006). Increasing numbers of studies are being carried out to evaluate the impacts of climate variability and climate change in Africa, and to develop adaptation strategies (Boko et al. 2007). Adaptation has been shown to be successful

and sustainable when conducted together with effective government input, consideration of civil and political rights, and literacy (Brooks et al. 2005).

Thornton et al. (2009) suggest that adaptation strategies to climate change be developed locally rather than for large, spatially contiguous regions, due to the heterogeneity of climate change and its effects. Crop yield projections are highly variable due, in part, to large variations in soils, topography, and current average temperatures and corresponding variability in projected rainfall and temperatures (Thornton et al. 2009). Consequently, adaptation strategies would best be informed by higher-resolution climate data and developed through localized, community-based efforts (Thornton et al. 2009).

Development of a range of adaptation strategies may improve the ability of local populations to cope with climate change. Recent evidence from East Africa suggests that diversification of coping mechanisms at the household level is greatest among the poorest and wealthiest sections of society (Campbell 1999, Smucker and Wisner 2008). Wealthy households have better access to non-farm activities and remittances that reduce reliance on local resources during drought. The poorest increasingly rely on wage labor or extractive activities (charcoal or fuelwood sales), subsistence agriculture at the arable margins of the savanna, or food aid resulting in differential opportunities for adaptation and new patterns of resource use.

Even if adaptation strategies are put in place, recent research suggests that, in tropical regions, human responses to droughts and more gradual declines in vegetative and livestock productivity, may lead to ‘tipping points’ in terms of the adaptive capacity of households to manage with change. At some point, dependence on natural

resource-based livelihood options may not be able to satisfy households' objectives and needs (Moore et al. 2005, 2006).

Policy change or other higher level institutional responses could influence local adaptation by either enabling or constraining those adaptations. Furthermore, adaptation at the local scale may heighten or lessen the vulnerability of some societal groups to future perturbations or stressors (Kates 2000, Wisner et al. 2004).

Under the near-term and future climate changes projected by the Climate-Land Interaction Project (CLIP, funded by the NSF Biocomplexity Program, BE/CNH Award # 03088420), savanna ecosystems will be one of the most negatively impacted ecosystems with large extents reaching 'tipping points' of dramatic changes in physical conditions, functions, and services (Moore et al. 2005, 2006). Savanna areas are dominated by pastoralism and, in some areas, wildlife. They are also at the expanding edge of cropped agriculture and are the location of the most rapid in-migration and land use change in the region over the past twenty years (Olson et al. 2004).

The savanna ecosystem responds to the highly variable rainfall with disequilibrium, leading to heterogeneous vegetation that changes over space and through time (Oba et al. 2003, Sankaran et al. 2005). This response will alter the distribution of savanna grasses, thus affecting grazing patterns of wildlife and domestic animals alike. Further degradation of pasture may occur due to the progressive growth of bush that often results from increases in rainfall. Conflicts between communities and ethnic groups over resource use have resulted from reductions in grazing lands (Oba et al. 2000).

An increase in health stresses may also be a consequence of climate change. Many contemporary African communities are affected by health stresses. Incidences of malaria, including the recent resurgence in the highlands of East Africa, involve a range of causal factors. Links to climate and other causal 'drivers' of change have recently attracted a great deal of attention and debate (e.g., Hay et al. 2002, Pascual et al. 2006). Results from the "Mapping Malaria Risk in Africa" project show a possible expansion and contraction, depending on location, of climatically suitable areas for malaria by 2020, 2050, and 2080 (Thomas et al. 2004). However, new evidence regarding micro-climate change due to land-use changes, such as swamp reclamation for agricultural use and deforestation in the highlands of western Kenya, suggest that conditions for larvae are being created and therefore the risk of malaria is increasing (Boko et al. 2007, Munga et al. 2006).

Biodiversity in Africa is also under threat from climate variability and change and other stress. Africa's development is constrained by climate change, habitat loss, over-harvesting of selected species, the spread of exotic species, and activities such as hunting and deforestation, which threaten to undermine the integrity of the continent's rich but fragile ecosystems (Boko et al. 2007, UNEP/GRID-Arendal 2002).

Increased variation in rainfall related to effects of ENSO in East Africa also has wide ranging socioeconomic impacts, especially in the agriculture and water resources industries (Majugu and Magezi 1985). Economic losses from drought in the 1980's totaled several hundred million U.S. dollars (Boko et al. 2007, Tarhule and Lamb 2003). Although future climate change seems to be marginally important when compared to other development issues (Davidson et al. 2003), it is clear that climate change and

variability, and associated increased disaster risks, will seriously hamper future development (Boko et al. 2007).

In efforts to plan for and mitigate these risks, powerful tools are being developed that utilize traditional information on climate and an emerging ability to predict future climatic events. These tools can be used to assist planning and management across all socioeconomic activities and underpin sustainable development. Advance knowledge of the probable climate extremes also plays a big role in mitigating the consequences of climate hazards such as drought, floods, and tropical cyclones (Majugu and Magezi 1985). A central goal of global change science is to obtain more reliable assessments of likely future climatic conditions and to assess the impacts on society, such as poverty, food production, and the incidence of disease (McCarthy et al. 2001). Concern with issues such as rising sea levels, increased frequency and intensity of extreme climatic events, and variability in crop production is influencing policy discussions in vulnerable countries.

To date, most research on climate change and its impacts has been global in scale (e.g., Lobell et al. 2008, Lobell and Field 2006). The relatively coarse-scale data simulated by general circulation models (GCMs) are useful for evaluating global trends in climatic variables; however, the resolution of GCM simulated data is not sufficient for regional evaluations of climatic patterns, particularly in the presence of high landscape variability (Moore et al. in review, Thornton et al. 2009). In concurrence, the Fourth Report of the Intergovernmental Panel on Climate Change (IPCC 4) stated, “Climate scenarios developed from GCMs are very coarse and do not usually adequately capture important regional variations in Africa’s climate. The need exists to further develop

regional climate models and sub-regional models at a scale that would be meaningful to decision-makers."

East Africa's climate is highly variable on a local scale due, in part, to highly variable topography. Substantial differences in local terrain occur across the region, from Mount Kilimanjaro (5,895 m) and Mount Kenya (5,199 m), the two highest peaks in Africa, and the Kenya Highlands, to the Great Rift Valley, Lake Victoria, and the Indian Ocean coastline. East African climate is also influenced by multiple sources of seasonality, such as the northerly and southerly migrations of the Intertropical Convergence Zone (ITCZ) throughout the year which, give rise to bimodal precipitation patterns at locations near the equator and unimodal patterns further north and south of the equator (Stock 2004, Mutai and Ward 2000, Hastenrath et al. 1993), and the complex seasonality resulting from Indian Ocean influences (Black et al. 2003). This large variation in regional climate may be masked at the coarse scales of GCMs (Moore et al. in review) and is difficult to detect from the sparse networks of climate observations on the ground. Furthermore, climate data measured at East African meteorological stations are sporadic over both time and space, and do not fully represent landscape variability (Hession and Moore 2010), thus demonstrating the need to develop higher resolution climate data.

1.2 In search of higher resolution climate data

Higher resolution climate information for impact assessment can be derived using two basic approaches: (1) downscaling of output from GCMs, and (2) estimation of conditions at unsampled locations using available measurements on the ground. The

former approach is more commonly employed; simulated data sets are more complete over space and time, allowing for more straightforward evaluation of general climate trends. Measured data are generally incomplete, but these direct measurements of the variable of interest can also be used to model trends using state-of-the-art statistical methods, the subject of this dissertation. Downscaling of simulated climate data is described first. Statistical analysis techniques and their use in modeling climate trends based on measured data are summarized in the following section.

1.2.1 Downscaling of GCM results

Downscaling methodologies developed over the last 30 years have begun to bridge the gap between global climate modeling and regional applications. Downscaling methods generally fall into one of two categories: statistical (empirical) downscaling and dynamical downscaling. Statistical downscaling involves correlating GCM simulated data with data from observed variables measured at specific locations (e.g., temperature or precipitation) to downscale simulated GCM results (Rogers et al. 2003). A diversity of approaches are utilized in statistical downscaling (e.g, multiple stepwise regression, logistic regression, canonical correlation analysis, and artificial neural networks), deployed to predict localized conditions from large-scale atmospheric parameters and GCM-derived aggregate data (e.g., Kahn et al. 2006, Salathe et al. 2007, von Storch et al. 1993). Dynamical downscaling forms the basis for regional climate modeling. Regional climate modeling uses GCM simulated data to drive limited-area, high-resolution regional climate model (RCM) simulations (Mearns et al. 2003; Dickinson et al. 1989; Giorgi 1990, 2006) that are able to resolve smaller, regional scale

climatic forcings such as: complex topography, land-water interfaces, and vegetation or land cover patterns. High-resolution climate impact assessments avail of these higher resolution climate predictions to formulate regional adaptation strategies (e.g., Mearns et al. 1999, 2001a, 2001b). Some major theoretical limitations impact both forms of downscaling, however, including the propagation of errors from GCMs (Mearns et al. 2003). The research presented herein makes use of observed climate data, rather than simulated data from GCMs or RCMs, and spatially explicit statistical techniques to develop improved methods for prediction of data in a data-scarce environment, and to apply these newly developed methods in East Africa. Furthermore, the accuracy of results will be compared to a global interpolated data set generated at a spatial resolution of 0.5 degrees by the Climatic Research Unit at the University of East Anglia (Mitchell and Jones 2005).

1.2.2 Use of statistical techniques with measured data

Various statistical techniques have been used to generate higher resolution climate data from observations on the ground. This dissertation focuses on techniques from geostatistics, particularly kriging, and new approaches in regression analysis that treat spatial problems to estimate higher resolution precipitation data. The methods of greatest immediate relevance within these two broad categories are now briefly discussed in order to motivate the dissertation work. The preferred techniques in this dissertation are multivariate, and utilize data on variables expected to influence the spatial distribution of precipitation (i.e., covariate data). Consequently, a description of the scales and mechanisms of precipitation in East Africa is included in Chapter 2,

Literature Review. This knowledge informs the selection of appropriate covariates in multivariate statistical analyses.

Uses of Kriging in the Estimation of Precipitation. Many climate studies have evaluated the ability of interpolation methods to estimate climatic variables over space. Historically, these studies have compared basic interpolation techniques (e.g., inverse distance weighting (IDW), Theissen polygons, and nearest neighbor interpolation) to various kriging techniques (e.g., Diodato 2005; Goovaerts 1999a, 1999b, 2000; Pardo-Iguzquiza 1998). Basic interpolators provide estimates of a variable of interest at unsampled locations; however, they do not assess uncertainty. Kriging techniques have the added benefits of being spatially explicit (i.e., accounting for location and configuration of samples as well as spatial autocorrelation in the data) and providing estimates of uncertainty for the interpolated values. Multivariate kriging techniques (e.g, universal kriging) incorporate covariate data to improve predictions and reduce uncertainty (Diodato 2005; Goovaerts 1999a, 1999b, 2000; Kyriakidis et al. 2001; Pardo-Iguzquiza 1998). In the development of multivariate kriging models, many researchers have relied on establishing correlations between precipitation and elevation to improve estimates of precipitation at unsampled locations (Diodato 2005; Goovaerts 1999a, 1999b, 2000; Kyriakidis et al. 2001; Pardo-Iguzquiza 1998). Although it is certainly important, and linked closely to spatial distribution of precipitation in certain parts of the world, elevation does not encompass the full range of mechanisms impacting rainfall, particularly in East Africa.

Regression Analysis. Scientists studying rainfall mechanisms frequently use multivariate statistical techniques in the interest of hypothesis testing. Many of these

researchers investigate climate on a more theoretical level with statistical inference and, consequently, rely on regression models. In so doing, these researchers have elaborated on relationships between precipitation and a wide variety of variables, including elevation and its derivatives, vegetative cover, specific humidity, and geographic descriptors such as distance to coastline, a variable of relevance to regions such as East Africa which is heavily influenced by maritime conditions (Marquinez et al. 2003, Oettli and Camberlin 2005, Anders et al. 2006, Propastin et al. 2006, Ji and Peters 2004). Such modeling has identified variables that are correlated with precipitation and has thus improved the understanding of spatial patterns in rainfall, but it is not designed explicitly to produce optimal estimates of values for sites out of sample, the objective of many geostatistical exercises such as kriging. Moreover, given the statistical estimation procedures generally used, these studies tend to neglect issues of spatial autocorrelation and endogenous feedbacks between precipitation and its covariates.

Modeling of Feedback Mechanisms. One widely noted feedback is the relationship between precipitation and vegetation. Precipitation plays an obvious a role in vegetation dynamics given the requirements of photosynthesis. But feedbacks to the atmosphere also exist (Rodriguez-Iturbe and Porporato 2004, p. 2). A wide literature covers many aspects of the relationship between vegetation and rainfall and a variety of approaches have been used to characterize it (e.g., Lyon et al. 2008, Notaro et al. 2008, Wang et al. 2008). To date, however, regression-based studies have tended to overlook endogenous relations between precipitation and vegetation, despite advances

in spatial econometrics that treat both simultaneity and spatial autocorrelation (Kelejian and Prucha 2004, Rey and Boarnet 2004).

To inform the selection of independent variables or covariates for multivariate statistical analyses of spatial patterns in precipitation data, the scales and mechanisms of precipitation will be considered. This knowledge will also inform hypothesis testing that will be conducted as part of the multivariate statistical analyses. Scales and mechanisms of precipitation are described in detail in Chapter 2.

Problem Statement: In sum, the predictive models of spatially explicit kriging have not accounted adequately for the mechanisms of precipitation, and the inferential models of regression addressing these mechanisms have not addressed spatial relations affecting climate processes. Further, neither approach has considered complex feedback relationships that may produce simultaneity between precipitation and its covariates. This dissertation addresses these limitations in predictive modeling by developing a kriging approach that accommodates simultaneous relations through theoretical innovation. Each of these methods will be applied in a case study of climate in East Africa, a data poor environment affected by climate change.

1.3 Links and contributions to the Geographic traditions

This dissertation spans and contributes to several strands of Geographic thought. The context of the work presented herein links Regional Geography of Africa, Physical Geography, and quantitative methods of spatial data analysis in an effort to answer

research questions that have been informed by Human-Environment researchers and, in turn, to inform their research (Figure 1-1).

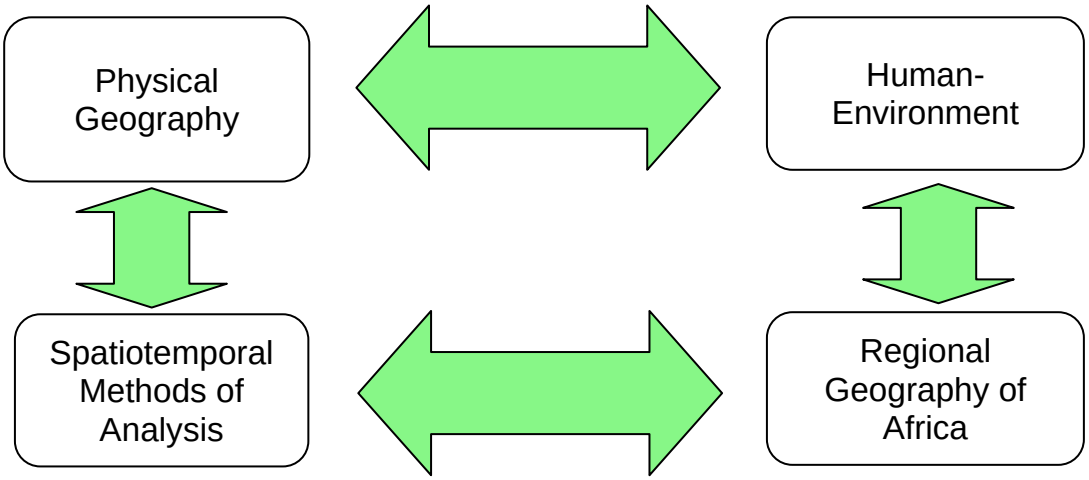

Figure 1-1. Relationships between the Geographic traditions represented in this research. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

The results and findings of this dissertation add to the knowledge and understanding of the spatial patterns of precipitation in East Africa. In addition, updated and improved maps of spatial patterns in precipitation within the study domain are generated, contributing to Physical Geography and Regional Geography of Africa. Results of this dissertation contribute to work conducted by Human-Environment researchers such as high-resolution climate impact assessments (Mearns et al.,1999, 2001a, 2001b) which may be used in the evaluation of coupled human natural systems and formulation of associated policy. Furthermore, this dissertation provides new and innovative techniques for spatiotemporal analysis of data which can be applied in any context when the goal is to understand spatial patterns in any continuous data, to predict data at unsampled locations, and/or to understand the relationship between a

variable of interest and other factors that may influence the spatial distribution of that variable.

Lastly, this research is designed to complement ongoing NSF activities at Michigan State University: the Climate Land Interaction Program (CLIP; *An Integrated Analysis of Regional Land-Climate Interactions*, National Science Foundation Award Number BCS0308420) and East Africa Climate Land Interaction Program – Savanna Ecosystems (EACLIPSE; *Dynamic Interactions Among People, Livestock, and Savanna Ecosystems under Climate Change*, National Science Foundation Award Number BCS0709671). Results of this dissertation can be incorporated in climate change studies such as these to inform adaptation strategies.

1.4 Research objectives

This dissertation research has four basic goals, which are to (1) contribute theoretically and methodologically to the prediction of variables in data-scarce environments utilizing improved kriging-based and spatial regression techniques; (2) apply these techniques in a case study based in East Africa; (3) evaluate case study results obtained using each technique and compare to those obtained from a gridded global climate data set interpolated to a 0.5 degree resolution; and (4) conduct hypothesis testing using selected spatial regression models to identify significant factors that influence the spatial distribution of precipitation.

This dissertation is organized as follows:

Chapter 2 provides a review of literature that supports and expands the introductory background discussion.

Chapter 3 presents the statistical modeling frameworks of kriging and spatial regression, demonstrates the theoretical links between these frameworks, and details the methodological innovations developed by this research.

Chapter 4 presents the case study. The case study area is described, including details of the physical precipitation mechanisms in East Africa, to provide a foundation for the selection of (1) independent variables for multivariate statistical analysis and (2) variables for which feedbacks will be modeled. Precipitation estimates at unsampled locations are generated using selected kriging method(s) and various formulations of the spatial regression models described in Chapter 3 of this dissertation. Error estimates from each analysis will be compared and evaluated for minimum error variance and the presence of spatial patterns not captured by the model. Additionally, error estimates from each analysis will be compared to error estimates generated by a global climate data set interpolated to a 0.5 degree resolution (Mitchell and Jones 2005) for comparison to the methods presented herein.

Chapter 5 focuses on understanding of spatial patterns in East African rainfall. Initial hypotheses related to variables that influence the spatial patterns of precipitation

and vegetation are described, followed by hypothesis testing using the spatial regression models developed in Chapter 4. This chapter follows the case study as an extension to the case study. The primary objective of the case study is to develop and compare prediction methods, some of which rely on multivariate regression models. Chapter 5 demonstrates the added benefit of spatial regression models: that of gaining an improved understanding of precipitation patterns in East Africa.

Chapter 6 presents findings, conclusions, contributions of the work presented herein, and discusses limitations of this work and future research needs.

Chapter 2

Literature Review

This chapter will link climatic concepts with the use of statistics to better understand the spatial patterns of precipitation and the scales over which they occur, and to develop finer-scale estimates for climatic variables with sparse data observations. The chapter begins with a discussion of the scales and mechanisms of precipitation in East Africa in order to motivate the need for higher resolution data in East Africa and the case study work in Chapter 4. Next, a review of statistical methods historically used to evaluate climatic data for various purposes is presented. The chapter ends with the identification of gaps in the literature and a discussion of how these gaps will be addressed by this dissertation.

2.1 Scales and mechanisms of precipitation in East Africa

Mechanisms of precipitation occur over various geographic scales around the world. Mesoscale processes take place over regions ranging from a few kilometers to approximately one hundred kilometers in diameter (Ahrens 2007), including land/sea breezes and orographic uplifting over mountainous terrain. Synoptic scale processes impact on the spatial distribution of precipitation over areas ranging from hundreds to thousands of square kilometers. These processes include high and low pressure areas and associated weather fronts. In East Africa, the presence of low pressure in January causes winds ranging from northeasterly (north of the Intertropical Convergence Zone (ITCZ)) to southeasterly (south of the ITCZ), coming in from the Indian Ocean (Mutai

and Ward 2000). In July, the low pressure over Asia and India results in the summer monsoon season there, and south to southwesterly winds in East Africa (Mutai and Ward 2000). In the tropics, the position of the global-scale ITCZ also plays a role in rainfall patterns. Its associated bands of rainfall move northward and southward over the year, dominating the spatial pattern of precipitation on a global scale. The passage of the ITCZ contributes to a generally bimodal seasonal pattern of rainfall near the equator, and a unimodal pattern at the northern and southern extents of the study area.

Although the broad features of this climate system can be described, specific questions remain unanswered, and constitute the empirical basis of this dissertation. As demonstrated herein, rainfall processes at varying scales can be modeled through the use of predictive or independent variables in a statistical correlation and regression analysis. Patterns in precipitation can be better understood through consideration of variables that are highly correlated with precipitation and for which data are more readily available, such as topographic variables including elevation and its derivatives (measures of mesoscale processes), geographic measures of location (proxies for synoptic and global scale processes), and season-specific analysis (allowing for seasonal variation in synoptic and global scale processes). Multivariate techniques that utilize covariate information can make use of relationships between these variables and precipitation to more accurately estimate precipitation patterns over space and time (Diodato 2005; Goovaerts 1999a, 1999b, 2000; Kyriakidis et al. 2001; Pardo-Iguzquiza 1998). Furthermore, this dissertation will demonstrate that multivariate techniques able to characterize endogenous relations between precipitation and vegetation provide a

new and exciting alternative to existing methodologies for producing maps of precipitation over space.

2.2 A review of statistical methods used in the analysis of precipitation data

Various statistical techniques have been used to generate higher resolution climate data using measured data, primarily at meteorological stations, rather than simulated data from GCMs or RCMs. This dissertation focuses on techniques from geostatistics, particularly kriging, and improved approaches in regression analysis that treat spatial problems to estimate higher resolution precipitation data. The methods of greatest immediate relevance within these two broad categories are summarized briefly in this chapter, motivating the theoretical work presented in Chapter 3. Given the vast literatures in question, the discussion in this chapter limits itself primarily to climate-related applications.

2.2.1 Kriging

Kriging refers to a family of techniques developed in France by Matheron (1963) based on the dissertation work of the South African mining engineer D. G. Krige (1951). Since its origination, many forms of kriging have been developed to predict attribute values at unsampled locations. Univariate kriging techniques consider only one variable, the variable of interest (e.g., simple kriging, ordinary kriging). Multivariate kriging techniques utilize secondary information in the form of independent variables or covariate data (e.g., kriging with an external drift/regression kriging, cokriging). All forms of kriging belong to a family of generalized least-squares regression algorithms,

four of which are described below. Each kriging method is described in detail in Chapter 3. A fifth form of kriging, developed in Chapter 3, is also briefly described below.

(1) **Simple kriging** is a univariate technique which is based on the assumption that the mean value of the variable of interest is constant and known.

(2) **Ordinary kriging** is also a univariate technique which differs from simple kriging in that the mean value, which is assumed constant, is unknown; to relax the assumption that the mean is constant through the entire study area, local neighborhoods can be established in which the mean is assumed constant.

(3) **Kriging with a trend surface model** is a multivariate technique in which a trend surface model is used to estimate a varying mean across the study area; trends surface modeling uses location coordinates and functions of these coordinates to estimate the mean at each location with the assumption that the data point to be predicted and the observed data are uncorrelated.

(4) **Universal kriging** (UK; Schabenberger and Gotway 2005), also known as kriging with an external drift (KED; Goovaerts 1997), is another multivariate technique in which the list of independent or explanatory variables is expanded to include other variables which are: (a) correlated with the variable of interest, and (b) measured at locations for which the variable of interest is to be estimated; estimators are developed using the *p* explanatory variables assuming a general linear model that holds for both the data and the unobservables, with the assumption that the data and the unobservables are spatially correlated.

(5) **Universal kriging with simultaneity** (developed herein) expands upon universal kriging by incorporating simultaneity between the variable of interest (i.e., the dependent or endogenous variable) and an explanatory variable (i.e., the independent or exogenous variable), such as that which occurs between precipitation and vegetation; this is accomplished by deriving and including an instrumental variable among the *p* explanatory variables in place of the variable that is simultaneously related to the dependent variable (e.g., vegetation).

As the field of geostatistics has evolved, some inconsistencies have arisen with regard to terminology (Hengl Heuvelink and Stein 2003) as well as an understanding of mathematical similarities and distinctions between various kriging techniques. This may be due in part to the fact that early developments in geostatistics generally occurred outside of mainstream statistics (Christensen 1991). As succinctly stated by Hengl, Hevelink, and Stein (2003),

> "The most probable cause (for this confusion) is that similar applications have been developed among different professions with different goals. The second important cause of this confusion is that some authors, more involved in the practice of kriging ('geostatisticians'), consider these techniques a special interpolation technique, while the other group ('statisticians') consider kriging to be only a case of regression analysis with spatially correlated data."

In addition, more than one school of thought has evolved, resulting in inconsistent naming conventions across kriging techniques. For example, Schabenberger and Gotway (2005) describe 'universal kriging' as above. However, the

technique termed ‘universal kriging’ by Journel and Huijbregts (1978, pg. 313) utilizes only polynomial functions of spatial coordinates, which is a method also known as ‘kriging with a trend model’ (Christensen 1991, Goovaerts 1997, Journel and Rossi 1989). Many authors agree that the term ‘universal kriging’ should be reserved for the case when the only covariates used are polynomial functions of the spatial coordinates (Hengl Heuvelink and Stein 2003). ‘Kriging with an external drift’ (e.g., Goovaerts 1997), is mathematically identical to ‘universal kriging’ as defined by Schabenberger and Gotway (2005) and ‘regression kriging’ (Hengl Heuvelink and Stein 2003). In mathematical terms it matters not which covariates are selected to model the mean function, $\boldsymbol{\mu}$; the mathematical derivation of the best linear unbiased predictor (BLUP) is the same (Christensen 1991). This dissertation develops a system of classification which recognizes this fact and seeks to clarify the links between forms of kriging as well as spatial regression (described below).

2.2.2 Spatial regression

A line of generalized least-squares regression models have arisen for analysis of spatial data: spatial regression models have been developed to explore causality using multiple independent (exogenous) variables (e.g., Anselin and Bera 1998, Anselin 2006, Caldas et al. 2007, Moore et al. 2007, Walker et al. 2000) and to account for simultaneity between multiple dependent (endogenous) variables (Kelejian and Prucha 2004, Rey and Boarnet 2004). Spatial regression models can also be used to generate estimates of dependent variables at unsampled locations taking into account information on independent variables and simultaneity between multiple dependent

variables. This section first provides a brief summary of spatial regression models, followed by a review of their current use and a discussion of potential future uses in climate research.

Three forms of spatial regression models provide the basis for the modeling of different forms of spatial autocorrelation:

(1) The **spatial lag model**, also known as a spatial autoregressive (SAR) model, accounts for the presence of spatial autocorrelation in the dependent variable by incorporating a spatial lag operator; spatial independence of the error terms is assumed.

(2) The **spatial error model** (SEM) allows for spatial dependence of the error terms, which may occur through the omission of a spatially varying covariate.

(3) The **general spatial model** (SAC) can be employed when both forms of spatial autocorrelation are present by incorporating both a spatial lag term and a spatially correlated error structure.

If neither form of spatial autocorrelation is present, an aspatial OLS regression model may be used. More detailed descriptions of these models are provided in Chapter 3 and by Anselin (2006), Anselin and Bera (1998), LeSage (1998), and LeSage and Pace (2009).

A decision process recommended by Anselin (2005) may be used to select between OLS, SAR, and SEM models for analysis of the data. In summary, the process involves completing an OLS analysis and calculating diagnostics for spatial dependence. The OLS results may be relied upon only if the assumptions underlying OLS are not violated. First, the lagrange multiplier (LM) statistics for both a spatial lag and spatial error model are tested for significance. If neither LM is significant, the OLS

results may be used. If one of LM statistics is significant, the corresponding model should be used to evaluate the data (e.g., if the LM statistic for only the spatial lag model is significant, the spatial lag model should be used to evaluate the data). If both of the LM statistics are significant, robust LM statistics should be tested for significance and the appropriate model selected.

In the case that both robust LM statistics are significant, generalized spatial modeling (SAC) may be used to account for this more complex form of spatial autocorrelation (LeSage 1998).

2.3 Historical use of methods to model climate patterns using measured data

Recently-used methods for modeling and predicting patterns in precipitation and vegetation based on measured data are summarized in
Table 2-1. Methods vary widely from simple OLS regression to spatially explicit geostatistical methods, Laplacian smoothing splines, and process-based physical models. The goals of many of these studies were to predict climatic variables over space and to compare the ability of various interpolation methods. Other studies were conducted with the goal of better understanding patterns in climatic variables and factors that influence these patterns using regression models and formal hypothesis testing. While some of these methods are univariate (i.e., they incorporate data from only one variable, the variable of interest), many others are multivariate and incorporate data from one or more independent or predictive variables (also known as covariates or exogenous variables). This section first describes the independent variables that are used in the studies below. Next, a review of past research is provided, with studies

grouped as follows: (1) studies to generate precipitation estimates at unsampled locations, (2) studies to improve understanding of precipitation patterns, and (3) studies to improve understanding of vegetation patterns.

Independent variables

In the development of multivariate models, many researchers have relied on establishing a correlation between precipitation and elevation as a covariate to improve estimates of precipitation at unsampled locations (Arora et al. 2006; Daly et al. 1994; Diodato 2005; Goovaerts 1999a, 1999b, 2000; Hutchinson 1998b; Hutchinson and Bischof 1983; KeifferWeisse and Bois 2001; Kyriakidis et al. 2004; Marquinez et al. 2003; Oettli and Camberlin 2005; Pardo-Iguzquiza 1998). Relatively high levels of correlation have been identified; the relationship is generally an increasing one (Arora et al. 2006; Spreen 1947): as elevation increases, precipitation increases. This is mainly due to the "orographic effect" of the mountain terrain. However, precipitation can have a very complex relationship with elevation (Arora et al. 2006), which can be complicated by station distribution and other factors (Hulme and New 1997).

Many other factors have also been found to play a role in the distribution of rainfall. Other researchers have expanded the list of potential predictive variables to include derivatives of elevation and other variables summarizing geographic location (Arora et al. 2006; Diodato 2005; Kieffer Weisse and Bois 2001; Kyriakidis et al. 2004; Hutchinson 1998b; Hutchinson and Bischof 1983; Marquinez et al. 2003; Oettli and Camberlin 2005; Spreen 1947). For example, Spreen (1947) found that distribution of rainfall also depends on variables such as slope, exposure, and orientation. Similarly,

Keiffer Weisse and Bois (2001) found that topographic variables were correlated with heavy rainfall events (e.g., 10- and 100-yr events), particularly when measured at short time steps (i.e., less than three hours). Regional topographic variables (e.g., distance to the Mediterranean, characterization of the general shape of the Alps, distance to corresponding features of the Alps) were found to influence heavy rains, whereas local measures of topography (e.g., altitude, slope, or azimuth) were less influential. Marquinez et al. (2003) evaluated the relationships between precipitation distribution and distance to coastline, distance to a location in the relative west, and elevation at two geographic scales (i.e., local and sub-basin). Hutchinson (1998b) derived the east and north components of the unit normal vector to represent slope and aspect for use as independent variables.

In the vegetation studies described herein, normalized difference vegetation index (NDVI; Rouse et al. 1974) was used to represent overall vegetation health. Theoretically, NDVI ranges between -1 and +1, although values typically range between 0 for bare ground to 0.7 for lush dense vegetation. Independent variables which have been considered to impact patterns in vegetation include precipitation (Propastin et al. 2006; Ji and Peters 2004), as well as potential evapotranspiration (PET), daily maximum and minimum air temperatures, soil temperature, and solar irradiation (Ji and Peters 2004). As shown by Ji and Peters (2004), precipitation and PET were the most significant predictors of NDVI.

Prediction and mapping of precipitation

Numerous studies predicting and mapping precipitation patterns have compared basic interpolation techniques (e.g., inverse distance weighting (IDW), proximal interpolation via Theissen polygons, and nearest neighbor interpolation) to various spatially-explicit kriging techniques that account for both location and configuration of samples as well as spatial autocorrelation (e.g., Diodato 2005; Goovaerts 1999a, 1999b, 2000; Pardo-Iguzquiza 1998). In general, spatially explicit kriging techniques outperformed basic interpolators. Further, multivariate kriging techniques incorporating covariate data were generally shown to improve predictions and reduce uncertainty (Diodato 2005; Goovaerts 1999a, 1999b, 2000; Kyriakidis et al. 2001; Pardo-Iguzquiza 1998). For example, Pardo-Iguzquiza (1998) found that kriging with an external drift (a multivariate technique also known as universal kriging) performed better than Thiessen polygons and ordinary kriging (univariate techniques). Goovaerts (1999a, 2000) found that kriging with an external drift and cokriging (multivariate methods) outperformed the basic interpolators (proximal interpolation with Thiessen polygons and IDW); however, simple kriging with varying local means (a univariate technique) outperformed both multivariate methods in this instance.

OLS regression has also been used to predict and map precipitation patterns (Daly, et al. 1994; KeifferWeisse and Bois 2001). KeifferWeisse and Boise (2001) used OLS regression to estimate heavy rainfall amounts at time steps ranging from 1 to 24 hours. Daly et al. (1994) stratified monthly and annual rainfall data over space into individual “topographic facets” to account for differences between facets in an effort to account for spatial autocorrelation. Independent regression analyses of precipitation

versus elevation for each facet were then conducted. It remains, however, that OLS regression methods do not explicitly account for spatial autocorrelation.

Another group of methods has been used to estimate the spatial distribution of rainfall: methods such as Laplacian smoothing splines and thin plate moving splines have been used to smooth and interpolate precipitation data over space using location coordinate information (Hutchinson 1998a; Hutchinson 1998b; Hutchinson and Bischof 1983). Rather than explicitly incorporating spatial autocorrelation into parameter estimation, however, these studies utilized a simplified approach in which one sample location was removed from close pairs of locations to avoid problems with short range correlation over space. Hutchinson (1998a, 1998b) evaluated spatial patterns in one day of rainfall data. Hutchinson and Bischof (1983) studied long-term mean seasonal (i.e., precipitation data were stratified into seasons) and annual precipitation. Process-based physical models have also been used to predict the spatial distribution of precipitation. For example, Barros and Lettenmaier (1993) modeled the advection of moisture over topographic barriers utilizing a 4D Lagrangian model. This model simulated orographically-induced precipitation at a scale sufficient to resolve dominant topographic features. Although methods utilizing splines or process-based physical models were used to interpolate rain data over space, they are not easily adapted for formal hypothesis testing of relationships between precipitation and possible explanatory variables.

Hypothesis testing

Hypothesis testing through regression analysis can be used to identify significant predictors of precipitation patterns over space. Studies with this goal include those by Arora et al. (2006), Marquinez et al. (2003) and Oettli and Camberlin (2005). These studies generally aimed to model and understand patterns in the mean precipitation function rather than generating estimates at unsampled locations, and they utilized aspatial OLS regression techniques to accomplish these goals. If present, spatially autocorrelated observations violate an underlying assumption of OLS regression, that of independent observations (Haining 1990; Neter, Wasserman, and Kutner, 1990; Bailey and Gatrell 1995; Schabenberger and Gotway 2005). Two major consequences of violating this assumption are (Ji and Peters 2004):

1) underestimation of the regression coefficients' standard errors, leading to the identification of variables as significant when they are not; and
2) underestimation of the error variance term, yielding inflated coefficients of determination (R^2). Spatial regression techniques can account for spatial autocorrelation, and thus are expected to provide better results in these instances.

Relatively few efforts have been made to apply spatially explicit regression techniques to climate data, probably because of the novelty of the methods. In particular, the subject of spatial regression only enters the academic literature in the 80's and 90's. Early efforts in the analysis of spatial data include Ord (1975) and

Anselin (1988). More recently, use of spatial regression expanded into ecological work, with applications to predicting the presence of species or species abundance (e.g., Augustin et al. 1996; Huffer and Wu 1998; Torbick, et al. 2010) and diseases (e.g., Gumpertz et al. 1997). An extensive recent review of spatial regression and geostatistics as applied to ecological modeling is given by Miller, et al. (2007).

Use of spatially explicit regression is relatively new to the modeling of climate variables and relationships with vegetation. Anders et al. (2006) utilize a form of regression modeling that reflected spatial lags to estimate average annual precipitation (as measured by the Tropical Rainfall Measuring Mission or TRMM) as a function of surface saturation vapor pressure (Vp), slope (S), relative elevation (E), the product of slope and Vp (Svp), and the product of slope and E (Se). This model incorporates two sets of OLS coefficients combined additively, with the second set incorporating a spatial lag. Significant predictors of precipitation are surface saturation vapor pressure (Vp) and the product of Vp with slope (Svp). Ji and Peters (2004) model NDVI as a function of precipitation, PET, maximum and minimum temperature, soil temperature, and solar irradiation. They implement OLS with ridge regression for their initial model, then integrate spatial variability using linear mixed models with a variogram component. Significant predictors of NDVI are precipitation and PET, which account for relatively low fractions of overall variability (46% for grassland areas, 24% for croplands). Propastin et al. (2006) compare OLS regression and geographically weighted regression (GWR) for modeling NDVI as a function of precipitation. GWR outperforms OLS, yielding coefficients of determination (R^2) of approximately 88%. Although the value of R^2 is no longer biased by the failure to account for spatial autocorrelation in the data, it is likely

that the reported values are overstated due to the handling of the precipitation data. Gridded maps of precipitation were created by interpolating data from nine climate stations using inverse distance weighting prior to performing the regression analysis. The resulting smoothed data are likely to underestimate true variability and overestimate the correlation between NDVI and precipitation. Further, GWR is a spatially explicit regression technique that allows for spatially varying covariates, but does not address the issue of spatially correlated error terms (Schabenberger and Gotway 2005).

In sum, work has begun on modeling climate using spatially explicit regression techniques; however, much remains to be done. Standard spatial models (e.g., SAR, SEM) have not been widely applied, if at all. Further, the studies cited (Ji and Peters 2004; Anders et al. 2006; Propastin et al. 2006) involve modeling of climatic variables for the purpose of hypothesis testing, which is somewhat limited given that spatial regression also provides a potentially useful tool for generating predictions and estimates of uncertainty at unsampled locations. Be this as it may, to date there are no direct applications of spatial modeling as proposed in this context. Related efforts include the mapping of Malaria incidence (Kazembe 2007) and species habitat distributions (Gottschalk et al. 2007), but these applications do not utilize spatial regression models in the form proposed here. Little used in the climate research domain, spatial regression is a method that may afford researchers a value-added approach to estimating climatic variables in data-sparse regions.

Finally, the ability to model multiple endogenous variables simultaneously in spatial regression yields particularly useful and exciting possibilities in climate research

due to interactions among multiple climate variables. In the climate research domain, an example of a feedback mechanism between climatic variables is that which occurs between precipitation and vegetation. Precipitation clearly plays a role in vegetation dynamics; however, as noted by Rodriguez-Iturbe and Porporato (2004), "vegetation exerts important control on the entire water balance and is responsible for many feedbacks to the atmosphere." A wide literature covers many aspects of the feedback between vegetation and precipitation (e.g., Brunsell 2006; Dekker et al. 2007; Kim and Wang 2007; Mendez-Barroso et al. 2009; Notaro and Liu 2008; Roy 2009; Wang et al. 2008; Zeng and Yoon 2009). A variety of approaches have been used to detect and model feedbacks, ranging from fully-coupled physical climate models to ensemble simulations using climate models, simple correlation coefficient analysis, and aspatial statistical feedback approaches developed initially for studies of ocean-atmosphere interactions. However, no one has yet incorporated new statistical approaches to simultaneity (e.g., Kelejian and Prucha 2004) to prediction, which is an objective of the proposed dissertation.

2.4 Gaps in the literature and contributions to address them

A gap exists that is not completely filled by the variety of statistical techniques most commonly used to evaluate climate data. The focus of the spatially explicit studies described above is often to develop the most accurate estimates of precipitation without necessarily understanding the role of various predictors or formally testing hypotheses about them. Schabenberger and Gotway (2005) note that the use of predictive variables is not the primary focus of many of these types of studies: predictive variables

are often used to “account for a spatially varying mean and to avoid bias.” In many cases, there is no intention of interpreting the relationships between the predictive variables and the dependent variable, or their significance.

When hypothesis testing and understanding are of primary interest, Schabenberger and Gotway (2005) identify spatial regression as a form of data analysis “where the focus is on modeling and understanding the mean function.” [Emphasis added.] Understanding is gained when significant variables that play a role in precipitation patterns are identified through hypothesis testing. However, as shown in Table 2-1, recent studies designed to test hypotheses regarding multiple predictors and their influence on rainfall rely heavily on OLS regression. Since OLS regression does not account for spatial autocorrelation, results are biased in the presence of spatially autocorrelated data.

More recently, spatially explicit regression techniques have begun to appear in climate-related literature; however, each of the applications is somewhat limited, either by the method used, the covariates selected, or the overall application (hypothesis testing or estimation). Further, none of the approaches have characterized endogenous relations, such as those between precipitation and vegetation. Accounting for these feedback mechanisms is expected to produce improved predictions of precipitation over space and refinements in the ability to separate sources of variability for the purpose of hypothesis testing.

Spatial regression may thus provide a solution to multiple problems of interest in the climate research domain: it allows for improved understanding through testing of hypotheses related to multiple independent variables while accounting for spatial

autocorrelation and it provides a method for estimating climate data in data sparse regions. This dissertation addresses these issues by experimenting with spatial regression techniques, existing kriging algorithms, and by developing a kriging approach that accommodates simultaneous relations through theoretical innovation.

From a theoretical perspective, much work has been done to mathematically compare and categorize generalized least-squares approaches such as kriging and spatial regression (e.g., Christensen 1991; Cressie 1993; Christensen 1996; Schabenberger and Gotway 2005). However, this work has not been extended to simultaneous equations spatial regression. Consequently, it is not known whether and to what extent recent developments in spatial regression reach beyond kriging capabilities.

Thus, the gaps in the literature are both empirical and theoretical. This dissertation will address these gaps as described in Chapter 3.

Table 2-1					
Author(s)	**Methods**	**Application/Location and Context**	**Predictors**	**Dependent variable**	**Results**
Methods relating precipitation to elevation					
Goovaerts (1999a, 2000)	OLS regression Thiessen polygons Inverse square distance OK, SKlm, KED, CK	Estimation / The Algarve, Southern Portugal, Atlantic Ocean to south and west	Elevation	Average monthly and annual precipitation	SKlm outperformed KED, CK, and univariate methods (cross-validation); OK outperformed OLS regression when r<0.75.
Goovaerts (1999b)	OLS regression SKlm, KED, CK	Estimation / The Algarve, Southern Portugal,	Elevation (average of values at 4 discrete points in a 1 square kilometer cell)	Erosivity data averaged on a monthly and annual basis	CK outperformed other methods (cross-validation)
Pardo-Iguzquiza (1998)	Thiessen polygons OK, CK, KED	Estimation / Guadalhorce river basin in southern Spain	Elevation	Mean annual rainfall over 20 year period	KED outperformed others methods (cross-validation)
Arora *et al.* (2006)	OLS regression	Hypothesis testing / Chenab basin, western Himalayas	Elevation Distance to lowest station	Seasonal and annual precip; grouped by mountain range, windward and leeward side	Different models generally resulted for different mountain ranges and for windward/leeward sides
Diodato (2005)	IDW, OK, CK	Estimation / Benevento Province, S. Italy, Mediterranean reion, variable topography	Elevation Smoothed elevation (DEM) Vegetation cover factor Topographic index	Average seasonal and annual precipitation	Highest correlation between topographic index and average annual precipitation (r^2=0.542)
Kyriakidis *et al.* (2004)	1) time series at stations 2) spatial regression of coeffs. with elevation and atmospheric data 3) CK of residuals 4) reconstruction of trend coefficients 5) conditional stochastic simulation	Estimation / Northern California coastal region, characterized by complex topography	Elevation Large-scale specific humidity from NCEP/NCAR reassessment	Daily precipitation	Applicability of the method was demonstrated; the method was able to reproduce the spatiotemporal characteristics of observed rainfall measurements

Table 2-1 (cont'd)					
Author(s)	**Methods**	**Application**	**Predictors**	**Dependent variable**	**Results**
Methods relating precipitation to elevation					
Marquinez *et al.* (2003)	OLS regression (backwards stepwise regression)	/ central area of the Cantabrian Coast, northern Spain	Distance from each station to coastline Distance from each station to relative west Elevation Avg. elevation per sub-basin Average slope per sub-basin	Dry season Wet season Annual	Adjusted correlation coefficients ranged between 0.58 and 0.67
Daly, *et al*. 1994	OLS regression over individual topographic facets	Estimation / Willamette River Basin, Oregon	Elevation	Averaged on a monthly and annual basis	PRISM exhibited the lowest cross-validation bias and absolute error when compared to kriging, detrended kriging, and CK
Oettli and Camberlin (2005)	OLS regression (forward stepwise regression) Cross-validation Calculation of estimates at gridpoints Calculation of residuals at gridpoints with stations Kriging of residuals (cubic interpolation)	Hypothesis testing and estimation / East Africa, southern Kenya, northeastern Tanzania	Topo. principal components For 35 different scaling windows: average and median elevation, standard deviation, amplitude, skewness, and kurtosis, slope Geographical locators (lat, long, distance from Lake Victoria)	Averaged on a monthly and annual basis	See text
Anders et al. (2006)	Regression model that accounts for spatial lags: utilizes OLS coefficients for the independent variables, and OLS coefficients for the spatial lags of those variables	Hypothesis Testing / Himalayas	Surface saturation vapor pressure (Vp), slope (S), relative elevation (E), product of slope and Sv (Svp), product of slope and relative elevation (Se); all variables were standardized	Average annual precipitation generated from 4 yrs of TRMM data (1998-2001)	Best 1 parameter model uses Vp; best 2 parameter model uses Vp and Svp; mean average error is not substantially decreased by adding more parameters

Table 2-1 (cont'd)					
Author(s)	**Methods**	**Application**	**Predictors**	**Dependent variable**	**Results**
Methods relating precipitation to elevation					
KeifferWeisse and Bois, 2001	Multivariate OLS regression(forward stepwise regression) with kriging of residuals	Estimation / French Alps	Regional variables such as X and Y coordinates, distance to the Mediterranean; local variables such as elevation, smoothed elevation, exposure parameters, and slope	Heavy rainfall amounts (10-yr and 100-yr rainfall events) at time steps ranging from 1 hr to 24 hrs.	Multivariate coefficients of determination ranging from 0.77 for hourly data decreasing to 0.57 for daily 100-yr data
Hutchinson and Bischof, 1983	Laplacian smoothing splines	Estimation / Hunter Valley, New South Wales	Latitude Longitude Elevation	Rainfall averaged on seasonal and annual basis	Analysis is objective and explicit; prior record standardization not required; each surface is valid for entire catchment; surfaces are consistent apart from data points; method provides percent predictive error estimates
Hutchinson, 1998a	Thin plate smoothing splines	Estimation / Swiss Alps	X and Y coordinates	One day of rainfall data	Estimates show good agreement with withheld data; short-range correlation dealt with by removing one point from close data pairs; square-root transformation of the data improved estimates; higher order splines were found to perform less well.
Hutchinson, 1998b	Thin plate smoothing splines	Estimation / Swiss Alps	X and Y coordinates Elevation E and N components of the unit normal vector to represent slope and aspect	One day of rainfall data	Analysis confirmed the importance of incorporating topographic variables

Table 2-1 (cont'd)					
Author(s)	**Methods**	**Application**	**Predictors**	**Dependent variable**	**Results**
Methods relating precipitation to elevation					
Barros and Lettenmaier, 1993	Process-based physical approach utilizing a 4D Lagrangian model	Estimation		Seasonal and annual runoff data Point estimates of monthly precipitation from snow courses and low-elevation precipitation gauges	Average areal precipitation was reproduced with errors in the range of 10-15%.
Spatially explicit statistical methods relating vegetation and climatic variables					
Propastin, et al. (2006)	OLS regression, Geographically Weighted Regression (GWR)	Hypothesis testing /	Dependent variable: NDVI Independent variable: Precipitation	Summed 10-day rainfall for each of 17 years by land cover class; gridded maps created from 9 climate stations using inverse distance weighting and resized to pixel resolution of NDVI	OLS r^2 ranged from 0.36 to 0.67; GWR r^2 averaged 0.88; much reduced spatial autocorrelation in residuals of GWR.
Ji and Peters (2004)	OLS model selection with ridge regression, integration of spatial variability using variogram method with mixed linear models (restricted maximum likelihood procedure)	Hypothesis testing / Northern and central U.S. Great Plains	Dependent var.: NDVI Independent vars.: precipitation, potential evapotranspiration (PET), daily max and min air temperature, soil temperature, solar irradiation	Averaged NDVI within a 10 km weather station buffer, by land cover class	Precipitation and PET were most significant variables; stronger correlations for grassland than cropland; r^2 of 46% for grassland, 24% for cropland.

Table 2-1 (cont'd)					
Author(s)	**Methods**	**Application**	**Predictors**	**Dependent variable**	**Results**
Spatially explicit statistical methods relating precipitation, vegetation, and topography					
Hession (2011)	Local OK, Universal Kriging, Universal Kriging with Instrumental Variables, and Spatial Regression	Hypothesis testing and estimation	Distance to Lake Victoria, distance to Indian Ocean, elevation, northern and eastern components of the unit normal vector, surface curvature, NDVI	Total monthly precipitation representing 4 seasons for 2 years	LOK performed best on the basis of RMSE, most consistent significant predictor was distance to water body, followed by NDVI and elevation.

IDW-Inverse distance weighting; OK-Ordinary kriging; SKlm-Simple kriging with varying local means; KED-Kriging with an external drift, CK-Cokriging

Chapter 3

Statistical Theory and Derivations

This dissertation will incorporate both theoretical and empirical research. The theoretical component is described here through the development of a theoretical mapping between geostatistical kriging methods and spatial regression techniques, all of which belong to the family of generalized least-squares regression algorithms. The theoretical mapping is followed by the development of an extension to universal kriging which accounts for simultaneity between two endogenous variables, such as that which occurs between precipitation and vegetation previously as described in Chapter 2.

3.1 Background

Prior to mapping theoretical similarities and differences between kriging and spatial regression techniques, each technique is presented in detail to provide a basis for the theoretical work.

3.1.1 Kriging

Many forms of kriging have been developed to predict attribute values at unsampled locations, including univariate techniques that consider only one variable (i.e., the variable of interest) and multivariate techniques that utilize secondary information in the form of independent variables or covariate data.

All forms of kriging belong to a family of generalized least-squares regression algorithms and can be derived via the basic linear regression estimator $Z^*(\mathbf{u})$ defined as (Goovaerts 1997):

$$Z^*(\boldsymbol{u}) - m(\boldsymbol{u}) = \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha(u)[Z(u_\alpha) - m(u_\alpha)] \quad (1)$$

where $\lambda_\alpha(\boldsymbol{u})$ is the weight assigned to the individual realization $z(u_\alpha)$ of the random variable (RV) $Z(u_\alpha)$. The expected values of RVs $Z(\boldsymbol{u})$ and $Z(u_\alpha)$ are represented by $m(\boldsymbol{u})$ and $m(u_\alpha)$. The number of locations used in the estimate at location u is represented by $n(u)$. Further, all varieties of kriging are derived with the goal of minimizing the error variance

$$\sigma_E^2(u) = Var\{Z^*(u) - Z(u)\} \quad (2)$$

with the constraint $E\{Z^*(u) - Z(u)\} = 0$ (Goovaerts 1997). The model used to represent $m(u)$ distinguishes between the forms of kriging.

Several forms of kriging are summarized below using a common notation in which $m(u)$ is represented more generally by $\boldsymbol{\mu}$. A summary of the common notation used throughout the remainder of this chapter is provided in Table 3-1. In general, vectors and matrices are represented in bold type, whereas constant terms are not bolded.

Table 3-1. Summary of common notation.

$\boldsymbol{y}$	a vector of observations of the dependent variable at *n* locations
$\mu, \boldsymbol{\mu}$	a constant, the mean value at a location, assumed to be known in simple kriging; bolded, $\boldsymbol{\mu}$ represents a vector of means at n locations
$\boldsymbol{e}$	a vector of error terms at ***n*** locations, with mean ***0*** and covariance $\boldsymbol{\Sigma}$
$\boldsymbol{\Sigma}$	the *n* x *n* covariance matrix of $\boldsymbol{e}$
$\boldsymbol{\sigma}$	a *n* x 1 vector of covariances between $\boldsymbol{y}$ at a single location to be predicted through kriging and the measured data at *n* locations
$\mathbf{1}$	a *n* x 1 vector of 1's
$\boldsymbol{X}$	a *n* x (*p*+1) matrix of measured values for *p* covariates at *n* locations preceded by a column of ones
$\boldsymbol{\beta}$	a (*p*+1) x *n* vector of regression coefficients
$\boldsymbol{x}$	a (*p*+1) x 1 vector of measured values for *p* covariates at a single locations to be predicted preceded by an entry of one (1)
$\boldsymbol{\beta}_{gls}$	a (*p*+1) x *n* vector of regression coefficients derived through generalized least squares
ρ	a constant representing the autoregressive coefficient under the spatial autoregressive (SAR)
$\boldsymbol{W}$	a *n* x *n* spatial weights matrix used in spatial regression
$\boldsymbol{Wy}$	a *n* x 1 vector representing the spatially lagged dependent variable
$\boldsymbol{\varepsilon}$	for the SAR model, a *n* x 1 vector of independent and identically distributed error terms at ***n*** locations, with a mean of ***0*** and variance σ^2; for the spatial error model (SEM), a *n* x 1 vector of spatially autocorrelated error terms
$\boldsymbol{\beta}_{SAR}$	a (*p*+1) x *n* vector of regression coefficients for the SAR model
$\boldsymbol{I}$	a *n* x n identity matrix
λ	a constant representing the autoregressive coefficient for the error terms under the SEM
$\boldsymbol{u}$	a *n* x 1 vector of independent and identically distributed error terms at ***n*** locations, with a mean of ***0*** and variance σ^2
$\boldsymbol{\beta}_{SEM}$	a (*p*+1) x *n* vector of regression coefficients for the SEM model
$\boldsymbol{\beta}_{SAC}$	a (*p*+1) x *n* vector of regression coefficients for the SAC model
$\boldsymbol{C}$	the *n* x *n* covariance matrix of $\boldsymbol{e}$ used in the derivation of simple kriging estimators (same as $\boldsymbol{\Sigma}$)
$\mathbf{c}$	a *n* x 1 vector of covariances between $\boldsymbol{y}$ at a single location to be predicted (same as $\boldsymbol{\sigma}$)
$\boldsymbol{K}, \boldsymbol{K}_{\boldsymbol{uk}}$	the (*n* + *p* + 1) x (*n* + *p* + 1) augmented matrix used in the derivation of universal kriging estimators
$\boldsymbol{k}, \boldsymbol{k}_{\boldsymbol{uk}}$	a (*n* + *p* + 1) x 1 augmented vector used in the derivation of universal kriging estimators

1) The theoretical model underlying **simple kriging** is presented below in matrix notation:

$$\boldsymbol{y} = \mu + \boldsymbol{e},\ \boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma}) \tag{3}$$

where ***y*** is a vector of observations measured at *n* locations, ***e*** is a vector of error terms at *n* locations, $\boldsymbol{\Sigma}$ represents the covariance matrix of the vector ***e*** and $\boldsymbol{\Sigma}$ is known. Furthermore, the mean μ is assumed to be constant and known.

The optimal linear predictor at an individual location developed under generalized least squares is

$$\hat{y}_{sk} = \mu + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{4}$$

and the simple kriging variance is

$$\sigma_{sk}^2 = \sigma^2 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} \tag{5}$$

where σ^2 = Var[*y*] at a single location is a constant and $\boldsymbol{\sigma}$ is a vector of covariances between *y* at a single location to be predicted and the observed data at *n* locations (from Schabenberger and Gotway 2005, pp. 223-224).

2) The theoretical model serving as the basis for **ordinary kriging** is stated below.

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{e},\ \boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma}) \tag{6}$$

where μ is constant but unknown and $\boldsymbol{\Sigma}$ is known.

The generalized least-squares form of the ordinary kriging predictor is

$$\hat{y}_{ok} = \hat{\mu} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu}) \tag{7}$$

and the ordinary kriging variance is

$$\sigma_{ok}^2 = \sigma^2 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + {(1 - \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma})^2}/{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \tag{8}$$

where σ^2 = Var[y] (Schabenberger and Gotway 2005, pp. 226-227; from Cressie 1993, p. 123). As previously noted, ordinary kriging estimators are developed with the assumption that $\boldsymbol{\mu}$ is constant but unknown; however, local neighborhoods can be established throughout which $\boldsymbol{\mu}$ is constant, relaxing the assumption that $\boldsymbol{\mu}$ is constant through the entire study area. This form of kriging is known as local ordinary kriging.

3) **Universal kriging** (UK; Schabenberger and Gotway 2005), also known as kriging with an external drift (KED; Goovaerts 1997), is based on an underlying theoretical model that allows for a varying mean throughout an area of interest. The mean is modeled by a general linear regression model that holds for both the data and the unobservables and incorporates *p* covariates or explanatory variables:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e},\ \ \boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma}) \tag{9}$$

$$y = \boldsymbol{x\beta} + e \tag{10}$$

This model represents a linear mean function and a spatially correlated error process. It is assumed that the data and the unobservables are spatially correlated with a variance-covariance matrix $\boldsymbol{\Sigma}$, Cov[$\boldsymbol{y}$, y] = $\boldsymbol{\sigma}$, and Var[y] = σ_0.

The regression coefficients $\boldsymbol{\beta}$ are estimated using generalized least squares as follows:

$$\widehat{\boldsymbol{\beta}}_{gls} = (\boldsymbol{X'\Sigma^{-1}X})^{-1}\boldsymbol{X'\Sigma^{-1}y}, \tag{11}$$

resulting in the following best linear unbiased predictor of y:

$$\hat{y}_{uk} = \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{gls} + \boldsymbol{\sigma'\Sigma^{-1}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{gls}) \tag{12}$$

with kriging variance

$$\sigma_{uk}^2 = \sigma_0 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + (\boldsymbol{x}' - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\mathbf{X}) \cdot (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \cdot (\boldsymbol{x}' - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\mathbf{X})' \quad (13)$$

(from Schabenberger and Gotway 2005, pp. 241-242). A similar approach, regression kriging, also known as kriging after detrending (Goovaerts, 1999b), is described by Hengl, Heuvelink, and Stein (2003) and is compared to UK/KED.

3.1.2 Spatial regression

Three forms of spatial regression models provide the basis for the modeling of different forms of spatial autocorrelation: 1) the **spatial lag model**, also known as a spatial autoregressive (SAR) model; 2) the **spatial error model** (SEM); and 3) the **general spatial model** (SAC). The SAR model accounts for the presence of spatial autocorrelation in the dependent variable, but assumes spatial independence of the error terms. The SEM allows for spatial dependence of the error terms, which may occur through the omission of a spatially varying covariate. If neither form of spatial autocorrelation is present, an aspatial OLS regression model may be used. If both forms of spatial autocorrelation are present, the SAC model may be employed. Each of these spatial regression models is presented below. More detailed descriptions are provided by Anselin (2006), Anselin and Bera (1998), LeSage (1998), and LeSage and Pace (2009).

1) **SAR model:** The SAR model incorporates a spatial lag operator and can be theoretically represented as in (14) with its corresponding data generating process (DGP; from LeSage and Pace 2009) shown in (15):

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \tag{15}$$

where $\boldsymbol{y}$ is a vector of dependent variable observations, ρ is the autoregressive coefficient, $\boldsymbol{W}$ is a weights matrix, $\boldsymbol{W}\boldsymbol{y}$ is a spatially lagged dependent variable, $\boldsymbol{X}$ is a matrix of observations of the independent variables, $\boldsymbol{\beta}$ is a vector of coefficients for the regression model, and $\boldsymbol{\varepsilon}$ is a vector of independent and identically distributed error terms.

As described by LeSage and Pace (2009), estimates for $\boldsymbol{\beta}$, σ^2, and ρ can be obtained through maximum likelihood estimation. A simplified approach is described in which a log-likelihood function that is "concentrated" with respect to $\boldsymbol{\beta}$ and σ^2 is developed. First, closed-form solutions for $\boldsymbol{\beta}$ and σ^2 are derived and substituted into the log-likelihood function, yielding a log-likelihood that is concentrated with respect to $\boldsymbol{\beta}$ and σ^2. This reduces the maximum likelihood to "a univariate optimization problem in the parameter ρ" and obviates the need to simultaneously solve the first order conditions for all three parameters. The estimators for $\boldsymbol{\beta}$ and σ^2 obtained using the concentrated log-likelihood function are identical to the form achieved by the full log-

likelihood function (Davidson and MacKinnon 1993, pgs. 267-269). The estimate for the parameter ρ is not achievable in closed form and must be derived computationally.

The resulting estimator for $\boldsymbol{\beta}$ under the SAR model is:

$$\widehat{\boldsymbol{\beta}}_{SAR} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \hat{\rho}\boldsymbol{W})\boldsymbol{y}. \quad (16)$$

Predicted values of $\boldsymbol{y}$ can be estimated completely at the n sampled locations: using Haining's (1990) terminology, the complete estimation referred to here can be developed as the sum of the trend, signal, and noise (residuals) components, since the values of the response variable are known at the sampled locations (Bivand 2009). Thus, the complete estimator is:

$$\hat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}\boldsymbol{\varepsilon} \quad (17)$$

obtained from the DGP or, replacing $\boldsymbol{\varepsilon}$ with the residual term,

$$\hat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR}) . \quad (18)$$

Predicted values of $\boldsymbol{y}$ at unsampled locations, however, can be obtained using the trend component only of the SAR model:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR} \quad (19)$$

Note that the dimensions of $\widehat{\boldsymbol{y}}$ and $\boldsymbol{X}$ in (19) reflect the number of locations at which the response variable $\boldsymbol{y}$ will be estimated. Since no observations of the response variable are available at unsampled locations, the signal component cannot fully reflect the spatial variation (Bivand 2009). The following "feasible signal component" has been suggested by Bivand 2009:

$$\rho \boldsymbol{W}\boldsymbol{y} = \rho \boldsymbol{W}(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta} \tag{20}$$

2) **SEM model**: The SEM allows for spatial autocorrelation in the residual term. The theoretical form of the SEM follows (21) and (22), with a DGP of (23):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{21}$$

$$\boldsymbol{\varepsilon} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{u} \tag{22}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \lambda \boldsymbol{W})^{-1}\boldsymbol{u} \tag{23}$$

where in this case $\boldsymbol{\varepsilon}$ is a vector of spatially autocorrelated error terms, λ is the autoregressive coefficient for the error terms, and $\boldsymbol{uu}$ is a vector of independent and identically distributed normal error terms.

Estimates for $\boldsymbol{\beta}$, σ^2, and λ can be obtained through maximum likelihood estimation, using the concentrated log-likelihood approach previously described

(LeSage and Pace 2009). Similar to the parameter ρ in the SAR model, λ is not achievable in closed form and must be derived computationally.

The maximum likelihood estimator for $\boldsymbol{\beta}$ using the concentrated log-likelihood under the SEM is

$$\widehat{\boldsymbol{\beta}}_{SEM} = (\boldsymbol{X}'(\boldsymbol{I} - \lambda\boldsymbol{W})'(\boldsymbol{I} - \lambda\boldsymbol{W})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \lambda\boldsymbol{W})'(\boldsymbol{I} - \lambda\boldsymbol{W})\boldsymbol{y} \quad (24)$$

Estimation of **y** at the *n* sampled locations can be accomplished using (from Bivand 2009):

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM} + \left(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W}\right)^{-1}\boldsymbol{u} \quad (25)$$

obtained from the DGP or, replacing $\boldsymbol{u}$ with the residual term,

$$\widehat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})^{-1}\,(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM}) + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM} \quad (26)$$

At unsampled locations, Bivand (2009) describes estimation of $\boldsymbol{y}$ using the trend component only of the SEM model:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM} \quad (27)$$

The signal component is set to zero because the spatial smoothing process in the SEM model is expressed only in terms of the error term $\boldsymbol{u}$ as shown in (25), for which the expectation is zero. Again, the dimensions of $\widehat{\boldsymbol{y}}$ and $\boldsymbol{X}$ in (27) reflect the number of locations at which the response variable $\boldsymbol{y}$ will be estimated.

3) **SAC model**: The SAC model incorporates both forms of spatial autocorrelation through a spatial lag term and a spatially correlated error structure, and is represented in (28) and (29). The DGP for this model is shown in (30)

$$\boldsymbol{y} = \rho \boldsymbol{W_1 y} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{28}$$

$$\boldsymbol{\varepsilon} = \lambda \boldsymbol{W_2 \varepsilon} + \boldsymbol{u} \tag{29}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho \boldsymbol{W_1})^{-1} \boldsymbol{X\beta} + (\boldsymbol{I} - \rho \boldsymbol{W_1})^{-1} (\boldsymbol{I} - \lambda \boldsymbol{W_2})^{-1} \boldsymbol{u} \tag{30}$$

where again $\boldsymbol{\varepsilon}$ is a vector of spatially autocorrelated error terms and $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are weights matrices, ρ is the autoregressive coefficient for the dependent variable, λ is the autoregressive coefficient for the error terms, and $\boldsymbol{u}$ is a vector of independent and identically distributed normal error terms.

Estimates for $\boldsymbol{\beta}$, σ^2, ρ, and λ can be obtained through maximum likelihood estimation, using the concentrated log-likelihood approach described above (LeSage and Pace 2009). However, in the case of the SAC model, a bivariate optimization in

two parameters (ρ and λ) is required. Neither parameter is achievable in closed form; therefore, both must be derived computationally.

The maximum likelihood estimator for $\boldsymbol{\beta}$ using the concentrated log-likelihood under the SAC is:

$$\widehat{\boldsymbol{\beta}}_{SAC} = (\boldsymbol{X}'(\boldsymbol{I} - \lambda\boldsymbol{W})'(\boldsymbol{I} - \lambda\boldsymbol{W})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \lambda\boldsymbol{W})'(\boldsymbol{I} - \lambda\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{y} \quad (31)$$

Estimation of $\boldsymbol{y}$ at the n sampled locations can be accomplished using (from Bivand 2009):

$$\boldsymbol{y} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1}\left(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2}\right)^{-1}\boldsymbol{u} \quad (32)$$

obtained from the DGP or, replacing $\boldsymbol{u}$ with the residual term,

$$\boldsymbol{y} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1}\left(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC}) \quad (33)$$

At unsampled locations, estimation of $\boldsymbol{y}$ can be completed using the trend component only of the SAC model:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC} \quad (34)$$

The dimensions of $\widehat{\boldsymbol{y}}$ and $\boldsymbol{X}$ in (34) reflect the number of locations at which the response variable $\boldsymbol{y}$ will be estimated. Similar to the SAR model, no observations of the response variable are available at unsampled locations; consequently, the signal component cannot fully reflect the spatial smoothing process. However, since the first term in (33) is similar to that of (18), the corresponding formula for the SAR model, the "feasible" signal component described by Bivand (2009) might also be considered:

$$\rho \boldsymbol{W}\boldsymbol{y} = \rho \boldsymbol{W}(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{X}\boldsymbol{\beta} \tag{35}$$

A **decision process** recommended by Anselin (2005) may be used to select between OLS, SAR, and SEM models for analysis of the data. In summary, the process involves completing an aspatial OLS regression analysis and calculating diagnostics for spatial dependence. The OLS results may be relied upon only if the assumptions underlying OLS are not violated. First, the lagrange multiplier (LM) statistics for both a spatial lag and spatial error model are tested for significance. If neither LM is significant, the OLS results may be used. If one of LM statistics is significant, the corresponding model should be used to evaluate the data (e.g., if the LM statistic for only the spatial lag model is significant, the spatial lag model should be used to evaluate the data). If both of the LM statistics are significant, robust LM statistics should be tested for significance and the appropriate model selected.

In the case that both robust LM statistics are significant, generalized spatial modeling (SAC) may be used to account for this more complex form of spatial autocorrelation (LeSage, 1998).

3.1.3 Summary of theoretical models and estimators

Theoretical models are summarized in Table 3-2 for both kriging and spatial regression models. Data generating processes (DGP) are also provided for the spatial regression models (LeSage and Pace 2009). A summary of estimators for each theoretical model follows (Table 3-3).

Note in Table 3-2 that the kriging models all incorporate spatially correlated error terms. Conversely, the spatial regression methods incorporate spatial autocorrelation directly in the model statement, resulting in uncorrelated error terms. Universal kriging attempts to incorporate spatial autocorrelation into the modeling step through the use of the generalized least squares estimate of $\boldsymbol{\beta}$ shown in (11). However, methods used to estimate the variance-covariance matrix $\boldsymbol{\Sigma}$, are not exact; consequently, it is expected that the error terms retain some spatial autocorrelation. These observations illustrate the following point made by Schabenberger and Gotway (2005):

> "The fact that we consider such very different models for modeling (and predicting) spatial data is due to the adage that, 'one modeler's fixed effect (regressor variable) is another modeler's random effect (spatial dependency.' Historically, estimation and prediction in models for spatial data started at the two extremes: regression models with uncorrelated errors (statistics) and correlated errors with a constant mean (geostatistics)."

Fortunately, much progress has been made in both directions, resulting in an ever-smaller gap between these approaches. This chapter will demonstrate the similarities and differences between the geostatistical and spatial regression approaches at present.

Table 3-2. Summary of theoretical models underlying kriging and spatial regression approaches.

Method	Theoretical Model	Distribution of Error Terms
Geostatistical Kriging Methods		
Simple Kriging	$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$	$\boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ is assumed to be constant and known
Ordinary Kriging	$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{e}$	$\boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma})$, μ is assumed to be constant and unknown
Universal Kriging	$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$	$\boldsymbol{e} \sim (\mathbf{0}, \boldsymbol{\Sigma})$
Spatial Regression Methods		
SAR Model	$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$	$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$
	DGP: $\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{X\beta} + (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}$	
SEM Model	$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} = \lambda\boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{u}$	$\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$
	DGP: $\boldsymbol{y} = \boldsymbol{X\beta} + (\boldsymbol{I} - \lambda\boldsymbol{W})^{-1}\boldsymbol{u}$	
SAC Model	$\boldsymbol{y} = \rho\boldsymbol{W_1}\boldsymbol{y} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} = \lambda\boldsymbol{W_2}\boldsymbol{\varepsilon} + \boldsymbol{u}$	$\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$
	DGP*: $\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W_1})^{-1}\boldsymbol{X\beta} + (\boldsymbol{I} - \rho\boldsymbol{W_1})^{-1}(\boldsymbol{I} - \lambda\boldsymbol{W_2})^{-1}\boldsymbol{u}$	

Table 3-3. Summary of kriging and spatial regression estimators.

Method	Estimators of y ($\boldsymbol{y}$)	Estimation Method
Geostatistical Kriging Methods		
Simple Kriging	$\hat{y}_{sk} = \mu + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$	Var[$\boldsymbol{y}$] = $\boldsymbol{\Sigma}$, Cov[$\boldsymbol{y}$, y] = $\boldsymbol{\sigma}$: populated using a variogram model; μ and $\boldsymbol{\mu}$: assumed known
Ordinary Kriging	$\hat{y}_{ok} = \hat{\mu} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})$	Var[$\boldsymbol{y}$] = $\boldsymbol{\Sigma}$, Cov[$\boldsymbol{y}$, y] = $\boldsymbol{\sigma}$: populated using a variogram model; $\hat{\mu}$: estimated by LS
Universal Kriging	$\hat{y}_{sk} = \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{gls} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{gls})$ $\widehat{\boldsymbol{\beta}}_{gls} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$	Var[$\boldsymbol{y}$] = $\boldsymbol{\Sigma}$, Cov[$\boldsymbol{y}$, y] = $\boldsymbol{\sigma}$: populated using a variogram model; $\widehat{\boldsymbol{\beta}}_{gls}$: estimated by GLS
Spatial Regression Methods		
SAR Model	$\widehat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1} \cdot (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR})$ $\widehat{\boldsymbol{\beta}}_{SAR} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \hat{\rho}\boldsymbol{W})\boldsymbol{y}$	$\widehat{\boldsymbol{\beta}}_{SAR}$: concentrated ML; $\hat{\rho}$: concentrated ML, numerical estimation
SEM Model	$\widehat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM}) + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM}$ $\hat{\beta}_{SEM} = \left(\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})\boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})' \cdot (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})\boldsymbol{y}$	$\widehat{\boldsymbol{\beta}}_{SEM}$: concentrated ML; $\hat{\lambda}$: concentrated ML, numerical estimation
SAC Model	$\widehat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC} + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})^{-1} \cdot (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2})^{-1} \cdot (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC})$ $\widehat{\boldsymbol{\beta}}_{SAC} = \left(\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})\boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})' \cdot (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})(\boldsymbol{I} - \hat{\rho}\boldsymbol{W})\boldsymbol{y}$	$\widehat{\boldsymbol{\beta}}_{SAC}$: concentrated ML; $\hat{\rho}$, $\hat{\lambda}$: concentrated ML, numerical estimation

The spatial regression estimators for $\widehat{\boldsymbol{y}}$ are all shown in their complete form in Table 3-3, which may be used when making predictions at previously sampled locations only. When predicting at unsampled locations, only the trend components are estimated (i.e., $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAR}$, $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SEM}$, $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{SAC}$)**.** The signal (i.e., the spatial smoothing) component is not reflected; however, each of the $\boldsymbol{\beta}$ terms incorporate spatial autocorrelation terms: ρ (SAR), λ (SEM), ρ and λ (SAC). This is a limitation spatial regression when interpolating at unsampled locations: since only the trend component is estimated, the predictions are likely overly smooth.

3.2 Theoretical mapping between statistical approaches

Theoretical work presented in this section involves the development of a classification system for kriging and spatial regression techniques, identifying mathematical similarities and distinctions between them. An important objective of this dissertation is to consider the theoretical correspondence between these two estimation paradigms.

As previously noted, kriging and spatial regression algorithms both belong to the family of generalized least-squares estimators. Generalized least-squares estimators (GLSEs) can be presented in a general linear regression model, such as the following linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{36}$$

where $\boldsymbol{\varepsilon}$ has a covariance structure of $\boldsymbol{\Sigma}$.

The Gauss-Markov estimator (GME) for $\boldsymbol{\beta}$ is of the form

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \tag{37}$$

and is the best linear unbiased estimator (BLUE) if $\boldsymbol{\Sigma}$ is known (Kariya and Kurata 2004). Since $\boldsymbol{\Sigma}$ is usually not known, calculation of the GME is not possible. In this case, the GLSE $\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\Sigma}}\right)$ can be calculated as above using $\widehat{\boldsymbol{\Sigma}}$ in place of $\boldsymbol{\Sigma}$. The estimated $\widehat{\boldsymbol{\Sigma}}$ must be positive definite. Additionally, many models use alternate formulations for $\boldsymbol{\Sigma}^{-1}$, including

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{I} + \lambda\boldsymbol{M}) \tag{38}$$

where $\boldsymbol{M}$ is a known square matrix (Kariya and Kurata 2004).

Note that the theoretical models for simple kriging, ordinary kriging, and universal kriging, respectively, are similar. They are restated here for convenience.

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e},\ \boldsymbol{e}\sim(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{3}$$

$$\boldsymbol{y} = \mu\boldsymbol{1} + \boldsymbol{e},\ \boldsymbol{e}\sim(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{6}$$

$$\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{e},\ \boldsymbol{e}\sim(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{9}$$

Each equation comprises a term for the mean, which may or may not vary over space, and a covariance structure for the error terms. GLSEs for all three forms of kriging can also be presented using the following notation:

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\lambda}}'\boldsymbol{y} \tag{39}$$

where $\boldsymbol{y}$ is a vector of the n observations of the variable of interest, $\hat{\boldsymbol{\lambda}}$ is a vector of estimated kriging weights, and $\hat{\boldsymbol{y}}$ is the kriging estimate at a given unsampled location. The kriging weights are generally calculated using the inverse of the full covariance matrix, which is augmented in the case of ordinary and universal kriging, multiplied by a vector of covariances between observed locations and the location to be predicted. For simple kriging, kriging weights are estimated as:

$$\hat{\boldsymbol{\lambda}}_{sk} = \boldsymbol{C}^{-1}\boldsymbol{c} \tag{40}$$

where $\boldsymbol{C}$ is the covariance matrix between all pairs of sampled locations, and $\boldsymbol{c}$ is the vector of covariances between each of the sampled locations and the location to be estimated. The resulting vector (40) can be substituted in (39) to calculate the simple kriging estimate.

The kriging weights for universal kriging, a multivariate form of kriging, are more complicated and are calculated using the following formula and augmented matrices (where matrix dimensions and row/column locations are shown in subscripts):

$$\hat{\boldsymbol{\lambda}}_{augmented} = \boldsymbol{K}_{uk}^{-1}\boldsymbol{k}_{uk} \tag{41}$$

where

$$\hat{\boldsymbol{\lambda}}_{augmented} = \begin{pmatrix} \lambda_{UK1} \\ \vdots \\ \lambda_{UKn} \\ \phi_1 \\ \vdots \\ \phi_{p+1} \end{pmatrix} \text{ or } \hat{\boldsymbol{\lambda}}_{augmented} = \begin{pmatrix} \boldsymbol{\lambda}_{UKn,1} \\ \boldsymbol{\phi}_{p+1,1} \end{pmatrix}, \tag{42}$$

$$\boldsymbol{k}_{uk} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \\ 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \text{ or } \boldsymbol{k}_{uk} = \begin{pmatrix} \boldsymbol{c}_{n,1} \\ \boldsymbol{x}_{p+1,1} \end{pmatrix}, \text{ and} \tag{43}$$

$$\boldsymbol{K}_{uk} = \begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 & x_{n1} & \cdots & x_{np} \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x_{11} & \cdots & x_{1n} & 0 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{p1} & \cdots & x_{pn} & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{44}$$

which can also be written in a more condensed form as a partitioned matrix:

$$\boldsymbol{K}_{uk} = \begin{bmatrix} \boldsymbol{C}_{n,n} & \boldsymbol{X}_{n,p+1} \\ \boldsymbol{X}'_{n,p+1} & \boldsymbol{0}_{p+1,p+1} \end{bmatrix} \tag{45}$$

Note that $\boldsymbol{c}_{n,1}$ and $\boldsymbol{C}_{n,n}$ are the same as $\boldsymbol{c}$ and $\boldsymbol{C}$ in Equation (40).

To calculate $\hat{\boldsymbol{\lambda}}_{augmented}$, $\boldsymbol{K}_{uk}$ must first be inverted. As shown by Theil (1971, pp. 17-19), a nonsingular matrix $\boldsymbol{D}$ is first defined as:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{R}_1 \\ \boldsymbol{R}_1' & \boldsymbol{Q}_1 \end{bmatrix} \tag{46}$$

where $\boldsymbol{P}_1$ and $\boldsymbol{Q}_1$ are non-singular symmetric matrices. Then $\boldsymbol{D}^{-1}$ can be written as:

$$\boldsymbol{D}^{-1} = \begin{bmatrix} \boldsymbol{P}_1^{-1} + \boldsymbol{P}_1^{-1}\boldsymbol{R}_1(\boldsymbol{Q}_1 - \boldsymbol{R}_1'\boldsymbol{P}_1^{-1}\boldsymbol{R}_1)^{-1}\boldsymbol{R}_1'\boldsymbol{P}_1^{-1} & -\boldsymbol{P}_1^{-1}\boldsymbol{R}_1(\boldsymbol{Q}_1 - \boldsymbol{R}_1'\boldsymbol{P}_1^{-1}\boldsymbol{R}_1)^{-1} \\ -(\boldsymbol{Q}_1 - \boldsymbol{R}_1'\boldsymbol{P}_1^{-1}\boldsymbol{R}_1)^{-1}\boldsymbol{R}_1{}'\boldsymbol{P}_1^{-1} & (\boldsymbol{Q}_1 - \boldsymbol{R}_1'\boldsymbol{P}_1^{-1}\boldsymbol{R}_1)^{-1} \end{bmatrix} \tag{47}$$

Since $\boldsymbol{Q}_1 = \boldsymbol{0}$, $\boldsymbol{K}_{uk}{}^{-1}$ can be written as:

$$\boldsymbol{K}_{uk}{}^{-1} = \begin{bmatrix} \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1} & \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \\ (\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1} & -(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \end{bmatrix} \tag{48}$$

Substituting Equations (43) and (48) into Equation (41) yields:

$$\hat{\boldsymbol{\lambda}}_{augmented} = \begin{bmatrix} \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1} & \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \\ (\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1} & -(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \end{bmatrix} \cdot \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{x} \end{pmatrix} \tag{49}$$

Since only $\boldsymbol{\lambda}_{UKn,1}$ from (42) is necessary for calculating the universal kriging estimate, the top row of the augmented matrix multiplied by vector $\begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{x} \end{pmatrix}$ can be rewritten as:

$$\hat{\boldsymbol{\lambda}}_{UK} = (\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}\,)^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1})\boldsymbol{c}\; + \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x} \quad (50)$$

The lower half of $\hat{\boldsymbol{\lambda}}_{augmented}$ in (42) contains the Langragian multipliers necessary to ensure that the results are unbiased and are not directly used in calculating the universal kriging estimate.

After much matrix algebra, it can be shown that

$$\hat{\boldsymbol{\lambda}}_{UK} = \boldsymbol{C}^{-1}\{\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}\,)^{-1}(\boldsymbol{x} - \boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{c})\} \quad (51)$$

and

$$\hat{\boldsymbol{\lambda}}_{UK}{}' = \{(\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}\,)^{-1}(\boldsymbol{x} - \boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{c})\}'\boldsymbol{C}^{-1} \quad (52)$$

Substituting (52) into (39) gives the universal kriging estimate as

$$\hat{\boldsymbol{y}} = \{(\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X}\,)^{-1}(\boldsymbol{x} - \boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{c})\}'\boldsymbol{C}^{-1}\boldsymbol{y} \quad (53)$$

which can also be written

$$\hat{\boldsymbol{y}} = \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{y} - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{y} + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{y} \quad (54)$$

and simplified as

$$\hat{\boldsymbol{y}} = \boldsymbol{\lambda}_{SK}{}'\boldsymbol{y} - \boldsymbol{\lambda}_{SK}{}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{x}\hat{\boldsymbol{\beta}}_{GLS} \quad (55)$$

or

$$\hat{\boldsymbol{y}} = \boldsymbol{\lambda}_{SK}{}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}) + \boldsymbol{x}\hat{\boldsymbol{\beta}}_{GLS} \quad (56)$$

In words, it can be shown that universal kriging reduces to simple kriging of the residuals from a generalized least-squares analysis plus the generalized least-squares estimate at the location of interest.

Estimators for the spatial regression models summarized in Equations (18), (26), and (33) can also be written in the form of Equation (56). Recall that these complete estimators are written for estimation at previously sampled locations, where information on the response variable is available. For example, the SEM model can be written in the form:

$$\hat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SEM}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SEM} \quad (57)$$

and the SAR model can be written as:

$$\hat{\boldsymbol{y}} = (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SAR}) + (\boldsymbol{I} - \hat{\rho}\boldsymbol{W})^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{SAR}\,. \quad (58)$$

where $\hat{\lambda}$ and $\hat{\rho}$ are constants, and $\boldsymbol{W}$ is a neighbor matrix. The estimated coefficients for universal kriging, the SEM model, and the SAR model are:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{y} \quad (59)$$

$$\widehat{\boldsymbol{\beta}}_{SEM} = \left(\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})\boldsymbol{X}\right)^{-1}\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W})\boldsymbol{y} \quad (60)$$

$$\widehat{\boldsymbol{\beta}}_{SAR} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \hat{\rho}\boldsymbol{W})\boldsymbol{y}. \quad (61)$$

Equations (56), (57), and (58) can be generalized as:

$$\hat{\boldsymbol{y}} = \boldsymbol{M}_1\,(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\,) + \boldsymbol{M}_2\boldsymbol{X}\widehat{\boldsymbol{\beta}} \quad (62)$$

Furthermore, (59), (60), and (61) above can be generalized as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M}_3\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_4\,\boldsymbol{y} \quad (63)$$

Table 3-4 summarizes the contents of the matrices in (62) and (63) for universal kriging, the SEM model, and the SAR model.

Table 3-4. Summary of multipliers for each estimation type.

Estimation Type	$\boldsymbol{M}_1$	$\boldsymbol{M}_2$	$\boldsymbol{M}_3$	$\boldsymbol{M}_4$
Universal kriging	$\boldsymbol{\lambda}_{SK}{}' = \boldsymbol{c}'\boldsymbol{C}^{-1}$	$\boldsymbol{I}$	$\boldsymbol{C}^{-1}$	$\boldsymbol{C}^{-1}$
SEM model	$(\boldsymbol{I} - \lambda\boldsymbol{W})^{-1}$	$\boldsymbol{I}$	$(\boldsymbol{I} - \lambda\boldsymbol{W})' \cdot (\boldsymbol{I} - \lambda\boldsymbol{W})$	$(\boldsymbol{I} - \lambda\boldsymbol{W})' \cdot (\boldsymbol{I} - \lambda\boldsymbol{W})$
SAR model	$(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$	$(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$	$\boldsymbol{I}$	$(\boldsymbol{I} - \rho\boldsymbol{W})$

3.3 Modeling of feedback simultaneity in a spatial setting

The spatial regression models presented in Section 3.1.2 have been expanded upon in the field of spatial econometrics to allow for the modeling of feedback simultaneity between two endogenous variables, such as precipitation and vegetation, in systems of multiple equations (Rey and Boarnet 2004). The use of spatially-explicit simultaneous equations to evaluate and understand feedback simultaneity is described in the Section 3.3.1. Estimators using spatial two-stage least squares (S2SLS) are also presented. In Section 3.3.2, concepts from the development of spatially-explicit regression models with feedback simultaneity are merged with the universal kriging technique (Section 3.1.1) to develop a new approach in multivariate kriging that incorporates feedback simultaneity between two endogenous variables. This new approach extends the ability to model and describe simultaneous feedback effects to the prediction of attribute values at unsampled locations in a manner that is both spatially explicit and reflective of feedback simultaneity between the dependent variable and a predictive variable.

3.3.1 Simultaneous equations spatial regression model

Recent developments in the field of econometrics allow for an expansion into systems of multiple, simultaneous equations (Kelejian and Prucha, 2004; Rey and Boarnet, 2004). Rey and Boarnet (2004) present a taxonomy of spatial econometrics models in simultaneous equations systems which allow for feedbacks between two endogenous variables, spatial lags for one or both variables, and various forms of cross-lags between both terms. The case which incorporates both feedback simultaneity and spatial simultaneity is described in more detail as follows (corrections made to equation 1 in Rey and Boarnet, 2004, p. 103):

$$\boldsymbol{y}_1 = \boldsymbol{X}\boldsymbol{\beta}_1 + \gamma_{21}\boldsymbol{y}_2 + \rho_{21}\boldsymbol{W}\boldsymbol{y}_2 + \rho_{11}\boldsymbol{W}\boldsymbol{y}_1 + \boldsymbol{\varepsilon}_1 \qquad (64)$$

$$\boldsymbol{y}_2 = \boldsymbol{X}\boldsymbol{\beta}_2 + \gamma_{12}\boldsymbol{y}_1 + \rho_{12}\boldsymbol{W}\boldsymbol{y}_1 + \rho_{22}\boldsymbol{W}\boldsymbol{y}_2 + \boldsymbol{\varepsilon}_2 \qquad (65)$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the n x 1 vectors of observations for each dependent variable, $\boldsymbol{X}$ is an n x k matrix of observations for k independent variables associated with the k x 1 vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, γ_{21} and γ_{21} provide for feedback simultaneity between the dependent variables, $\boldsymbol{W}$ is an n x n spatial weights matrix, ρ_{11} and ρ_{22} are the spatial autoregressive lag terms, ρ_{21} and ρ_{12} are spatial cross-regressive terms, and $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are error terms with the following properties:

$$Cov\left[\boldsymbol{\varepsilon}_{1,i}, \boldsymbol{\varepsilon}_{2,i}\right] = 0, \text{ for all } i,$$

$$Cov\left[\boldsymbol{\varepsilon}_{l,j}, \boldsymbol{\varepsilon}_{l,j}\right] = 0, \text{ for all } i \neq j, \text{ and } l = 1,2,$$

$$Cov\left[\boldsymbol{\varepsilon}_{1,i}, \boldsymbol{\varepsilon}_{2,j}\right] = 0, \text{ for all } i \neq j.$$

This system of equations can be expressed in matrix notation as follows:

$$\boldsymbol{Y\Gamma = WYP + XB + E} \tag{66}$$

where $\boldsymbol{Y} = [y_1, y_2]$ is a vector of endogenous (dependent) variables, $\boldsymbol{X}$ is a matrix of exogenous or independent variables, $\boldsymbol{B} = (\beta_1, \beta_2)$, $\boldsymbol{E} = (\varepsilon_1, \varepsilon_2)$,

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}.$$

Estimators for this system of equations have been proposed and initially evaluated by Rey and Boarnet (2004). A spatial two stage estimator can be obtained following the steps below:

1. Calculate the predicted values for the endogenous variable on the right hand side (RHS) of the equation by running OLS regression on one or more of the exogenous variables used in predicting the endogenous variable on the left hand side (LHS) plus at least one exogenous variable not used in the prediction of the LHS endogenous variable.
2. Calculate predicted values for the lagged endogenous values on the RHS (i.e, $\boldsymbol{Wy_1}$ and/or $\boldsymbol{Wy_2}$) as above.

3. Replace each of the endogenous variables on the RHS with their predicted values, then estimate the parameters of the equation using OLS regression.

An aspatial two stage estimator can be obtained by omitting the lagged terms and step 2 above.

For example, the S2SLS estimators of the parameter vector θ_1 ($\theta_1{}' = [\beta_1', \gamma_{21}, \rho_{11}]$) for the first equation in the system of equations incorporating a feedback and two spatial lag terms (Rey and Boarnet 2004, Table 5.1, Model 13) can be estimated in matrix terms as:

$$\hat{\theta}_{S2SLS} = (Z_1{}'Z_1)^{-1}Z_1 y_1 \qquad (67)$$

where $Z_1 = \left[X^*, \hat{y}_{2,} W y_1\right]$, X^* is the matrix of exogenous variables excluding the additional exogenous variable(s) not used to estimate y_1 (i.e., the instrumental variable), $\hat{y}_2 = Q y_2$, $W\hat{y}_1 = WQ y_1$, and $Q = X(X'X)^{-1}X'$ (which is also commonly known as the hat matrix). The matrix X is the full matrix of exogenous variables, or the matrix X^* above with the instrumental variable. The variables $\hat{y}_2$ and $W\hat{y}_1$ are the instrumented variables or the instruments.

To test for the presence of feedback simultaneity, it is most straightforward to start with an aspatial OLS model. First, fit an OLS model after calculating an instrument for y_2 (assuming that feedback simultaneity is present between y_1 and y_2). For

example, the S2SLS estimators of the parameter vector θ_1 (${\theta_1}' = [\beta_1', \gamma_{21}]$) can be estimated in matrix terms as:

$$\hat{\theta}_{S2SLS} = ({Z_1}'Z_1)^{-1}Z_1 y_1 \quad (68)$$

where $Z_1 = [X^*, \hat{y}_2]$, X^* is the matrix of exogenous variables excluding the additional exogenous variable(s) not used to estimate y_1, $\hat{y}_2 = Qy_2$, and $Q = X(X'X)^{-1}X'$. The matrix X is the full matrix of exogenous variables. Next, fit an OLS model for y_1 treating the second endogenous variable y_2 as an exogenous variable (assuming no feedback simultaneity).

$$\hat{\theta}_{OLS} = X^+({X^+}'X^+)^{-1}{X^+}'y_1 \quad (69)$$

where $X^+ = [X^* \quad y_2]$. The two models can be compared using the Hausman Test (Hausman 1978).

3.3.2 An extension to universal kriging

This dissertation extends the multivariate kriging technique known as universal kriging (UK; Schabenberger and Gotway 2005), also known as kriging with an external drift (KED; Goovaerts 1997). Although UK can provide considerable improvement over univariate forms of estimation, applications to date neglect critical relationships between

the kriged variables (e.g., precipitation) and selected covariates. In particular, feedbacks, or simultaneity, between variables of interest are not considered by existing kriging methods, a shortcoming in the kriging tool kit.

The presentation of universal kriging in Section 3.2 itself to extension through the incorporation of simultaneity. The formulation of the universal kriging estimator shown in (41) through (51) evidently assumes a lack of relationship between variables in the $\boldsymbol{X}$ matrix, and the dependent variable, *y*. Given that the covariation structure is typically taken to be purely spatial, possible co-variation introduced by simultaneity bias is neglected in predicting $\hat{y}$. The research presented herein considers solutions to (56) that account for endogenous relationships among the variables, specifically precipitation and vegetative cover, by extending universal kriging through an instrumental variables approach. Specifically, an instrument is included in the multivariate regression model used to estimate the mean process at each location. For example, let X^* be made up of three independent variables (elevation (e), distance from the coast (d), and a measure of vegetation (v)): $\boldsymbol{X}^* = [1 \quad \boldsymbol{X_e} \quad \boldsymbol{X_d} \quad \boldsymbol{X_v}]$. The feedback or simultaneity between the dependent variable precipitation and vegetation can be modeled using an instrumental variable (e.g., a variable related to vegetation but less so to precipitation, such as soil type (s)). Let

$$\hat{\boldsymbol{y}}_2 = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_2 \tag{70}$$

where $= [\boldsymbol{X}^{*} \quad \boldsymbol{X_s}]$. The instrumented variable $\hat{y}_2$ can be substituted for $\boldsymbol{X_v}$ in the original matrix to correct for problems due to simultaneity, giving a new data matrix, $\boldsymbol{X}^{**} = [\mathbf{1} \quad \boldsymbol{X_e} \quad \boldsymbol{X_d} \quad \widehat{\boldsymbol{y}}_2]$. $\boldsymbol{X}^{**}$ is then substituted for $\boldsymbol{X}$ in (70), and $\widehat{\boldsymbol{y}}$ is predicted as before. This extended form of universal kriging is referred to as universal kriging with instrumental variables (UKIV).

This dissertation implements UKIV in the case study (Chapter 5). As described in Chapter 5, no simultaneity was identified based on testing using the OLS approach described in Section 3.3.1 and the Hausman test. However, UKIV is retained in the case study as a purely theoretical development.

Chapter 4

Case Study

4.1 Study area description

The case study area is located in eastern Africa (Figure 4-1), falling mainly in the East African country of Kenya and overlapping into northern Tanzania. The geographic coordinates of the study area range from 34.6° to 39.1° E longitude and 3.7° S to 1.7° N latitude, falling in the center of the tropical belt, which ranges in latitude from the Tropic of Cancer in the northern hemisphere at approximately 23.4° N to the Tropic of Capricorn in the southern hemisphere at 23.4° S. Although the tropics are often thought of as hot and humid, due to their position near the equator and the tropical rain belt,

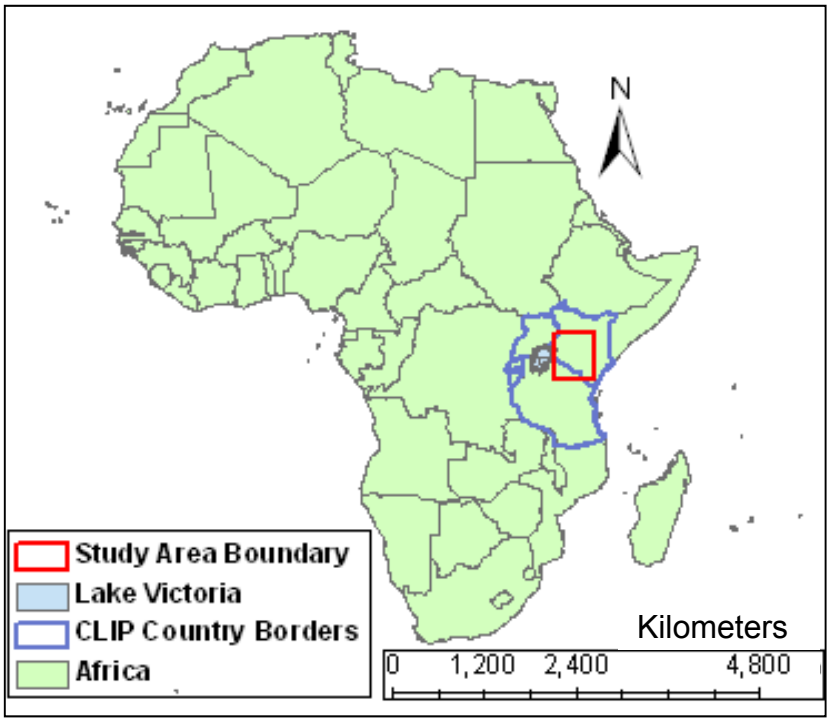

Figure 4-1. Location of the case study area within East Africa (CLIP region) and Africa.

other features within the tropics (e.g., topography, continental and regional scale winds) influence regional climates, resulting in conditions that range from arid to humid (Stock, 2004).

The study area contains Mount Kenya to the northeast, and much of the Kenya Highlands to the west (Figure 4-2). Elevations in this study area range from 195 to 5,778 meters and average 1,291 meters with a standard deviation of 649 meters. Substantial differences in local terrain occur across the region, from Mount Kenya, the second highest peak in Africa, to the Great Rift Valley, which cuts through the center of the region and Lake Victoria, bordering the western edge of the region. The size of the study area was chosen to be large enough to observe spatial variability in precipitation

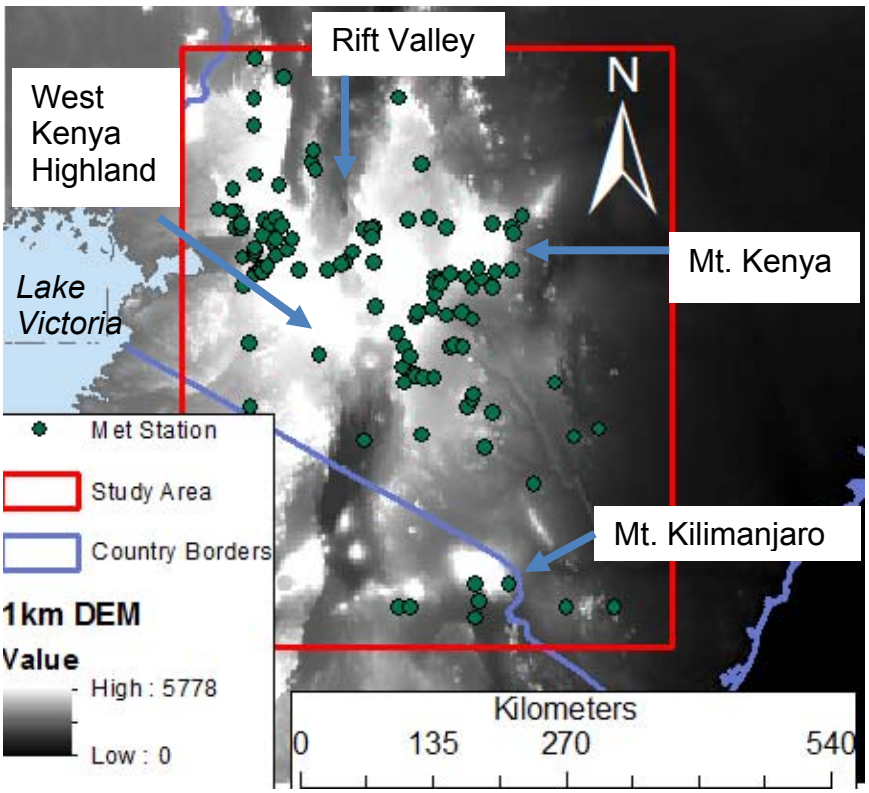

Figure 4-2. Meteorological station locations and the study area boundary are shown above on a map of elevation (1 km). Major topographic features in the study area are labeled.

and to allow for the evaluation of predictive variables at multiple scales. Furthermore, the sample size of approximately 120 precipitation stations is adequate for statistical analysis.

Equatorial East African rainfall seasonality is dominated by the “long rains” (March through May) and “short rains” (October through December) associated with the strong atmospheric convergence of the passing ITCZ (Hastenrath *et al*., 1993; Stock, 2004). This seasonal pattern in rainfall is illustrated on Figure 4-3. The long rains “provide more rainfall than the ‘short rains’ and have a lower interannual variability” (Camberlin and Okoola, 2003), but the short rains’ start is more predictable. Outside of the long and short rains coincident with the passing of the ITCZ (i.e., January and February, June through September), precipitation is typically localized convective rainfall or, in the highland areas, stratiform rainfall during the cooler months (Ng’ang’a, 1992). Thus, the spatial scale of precipitation varies from large-scale during the long rains and short rains to mesoscale during the drier months, driven by processes at their respective scales. The months chosen for statistical analysis (i.e., January, April, August, and November), therefore, represent the seasons in East Africa and the corresponding forms of precipitation. Furthermore, independent variables related to topography, such as elevation, were evaluated for predictive ability at two scales, 1 km and 9 km, in an effort to capture the spatial scale over which precipitation occurs at each season.

Higher rainfall occurs at higher elevations in part due to cooler temperatures and also due to the mountains acting as barriers to moisture-bearing winds. The slopes of the surfaces interact with elevation (and also the direction of the moisture-bearing

winds) such that steeper slopes extract more precipitation. Easterly flow from the Indian Ocean is the dominant source of moisture for this region, particularly during the short rains (Black *et al*., 2003). Since the relatively dry northeasterly and southeasterly monsoonal winds are weaker at these transition times, onshore moisture transport is stronger (Nicholson, 1996). Consequently, during the dry seasons generally wetter conditions occur near Lake Victoria immediately west of the study area, and the Indian Ocean coast to the east of the study area. While Lake Victoria acts a moisture source for local convective rainfall in the surrounding highland areas, the coastal climate differentiates itself from the highland climate due to small-scale diurnal convection from the land/sea breeze (Camberlin and Planchon, 1997). Blocked by the Rift Valley slopes, the western parts of East Africa receive moist westerly flows from the Congo basin. Correspondingly, potential explanatory variables in the statistical analysis included elevation (1 km and 9 km resolutions), a term that combined information from slope and aspect (1 km and 9 km resolutions), distance from Lake Victoria and distance from the Indian Ocean.

4.2 Data

Monthly precipitation is the dependent variable in this analysis. Monthly precipitation data (mm) from approximately 120 meteorological stations obtained from the Department of Meteorology, Government of Kenya were used in this analysis. Meteorological station locations are shown on Figure 4-2. Meteorological stations are generally located in areas of higher population; therefore, areas of high elevations (i.e., greater than 2,500 m) and low elevations (i.e., less than 1,000 m) are not well represented.

Precipitation data for this analysis were chosen to represent the four seasons in equatorial East Africa (i.e., the dry season in December, January, and February, when the ITCA is in its southernmost position; the long rainy season in March, April, May, and June, when the ITCZ is overhead; the cool dry season in July and August, when the ITCA is in its northernmost position; and the short rainy season in September, October, and November, when the ITCZ is again overhead) for two different years. The individual months of January, April, August, and November were chosen as the most representative month of each of the respective seasons (Figure 4-3).

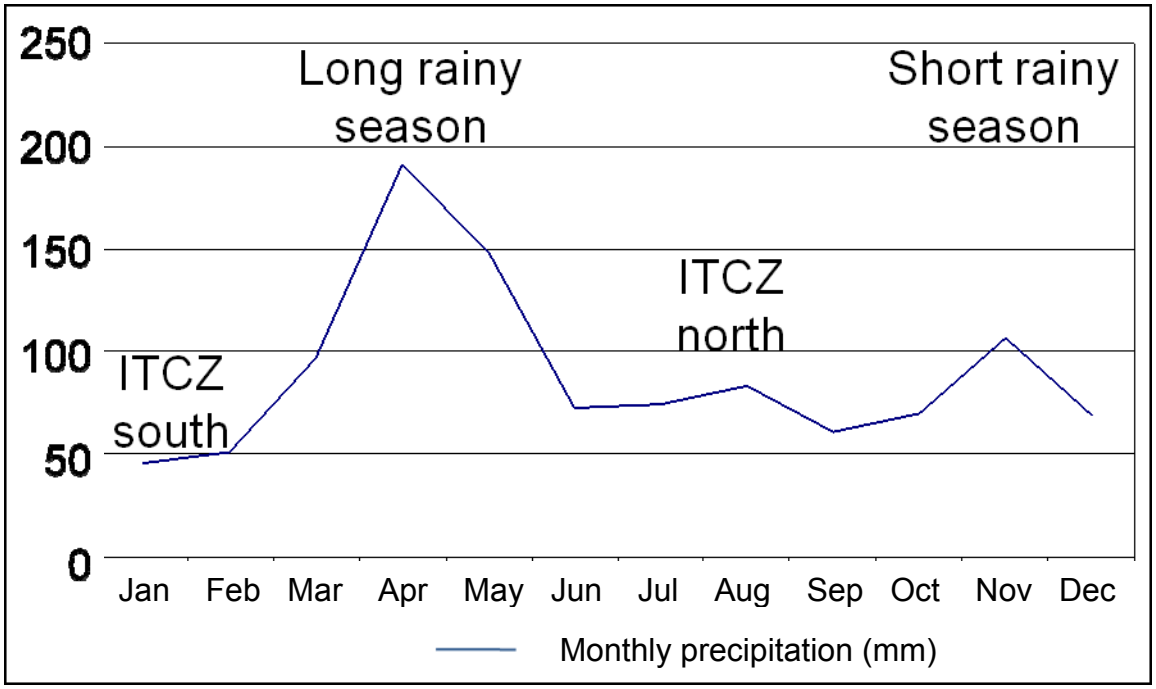

Figure 4-3. Long-term average monthly precipitation (mm) from 1926 to 1998 averaged over all meteorological stations within the study area. The position of the intertropical convergence zone (ITCZ) relative to the study area is labeled at various points in the year.

Precipitation data from 1984 and 1985 were chosen since these years represent the overlap between years in which the largest numbers of meteorological stations were

measured (Figure 4-4) and the years in which remotely sensed vegetation data were available (beginning in July 1982). Furthermore, the years to be evaluated were chosen to represent a typical year (1985) and an atypical year (1984). Figure 4-5 illustrates monthly precipitation for the years 1982 through 1985 plotted with long-term monthly averages (dashed lines).

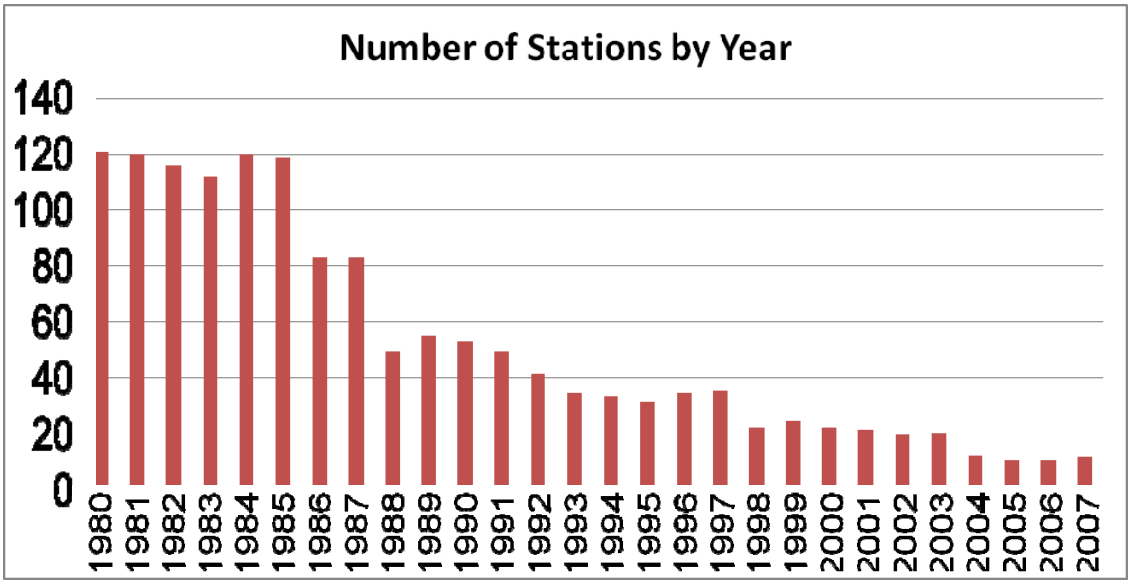

Figure 4-4. Summary of the number of meteorological stations measured by year within the study area. The number of stations measured decreases after 1985 and continues to rapidly decline through 2007. This decline is probably due to a combination changes in monitoring locations or, more likely, lack of access to more recent meteorological station data due to financial restrictions.

The year 1985 appears to be closest to the long-term average. The differences between monthly averages for each month and the long term averages were quantified by calculating average squared differences for each year as the sum of the differences squared for that year divided by n (12). This value was lowest for 1985 (51.8). The average sum of squared differences was highest for 1984 (288.4), with the largest

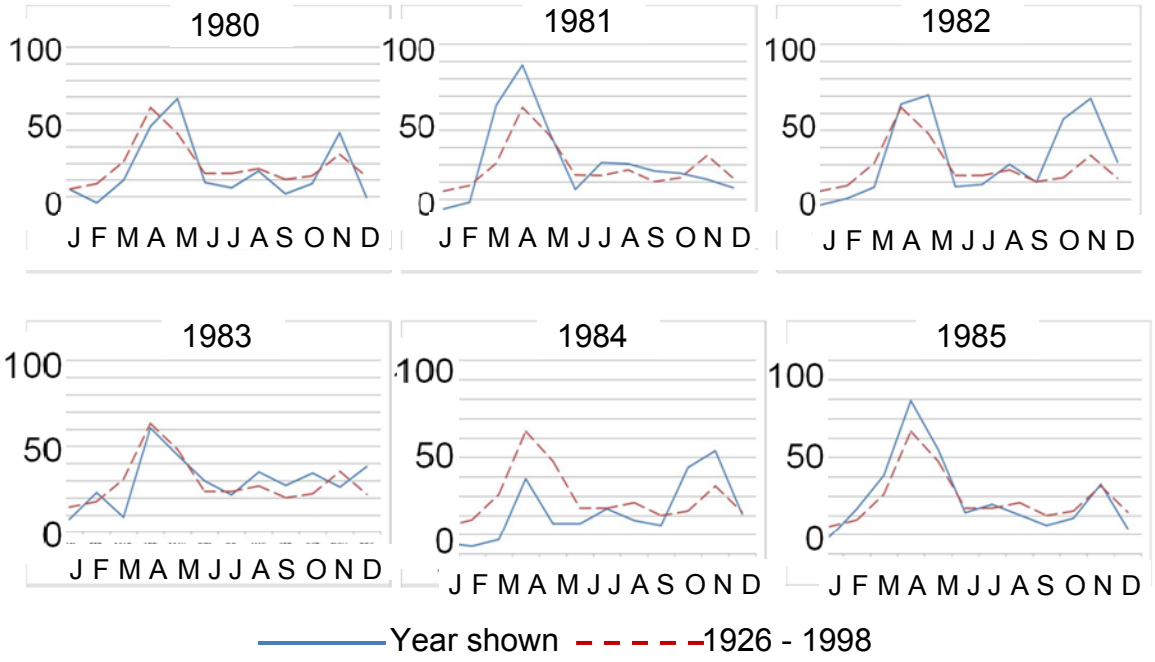

Figure 4-5. Monthly precipitation for the years 1980 through 1985 plotted with long-term monthly averages (dashed lines). The average sum of squared differences was lowest for 1985 (51.8). The average sum of squared differences was highest for 1984 (288.4), with the largest numbers of months falling below the long-term average. Precipitation amounts in 1984 are lower than average for most of the year.

numbers of months falling below the long-term average.

Independent variables included two distance measures: minimum distance of each meteorological station from the Indian Ocean coastline and distance of each station from the centroid of Lake Victoria. Since precipitation was not linearly related to either of these distance measures, dummy variables for each measure were established to represent “distance bands,” within which precipitation levels were similar. Distance bands of 0 to 300 km, 300 to 450 km, 450 to 600 km, and greater than 600 km were created as a measure for distance to Lake Victoria. Distance bands of 0 to 500 km and greater than 500 km were used to represent distance to the Indian Ocean. All distance bands were chosen based on inspection of precipitation versus distance plots; the cutoff

distances were generally observed to be indicative of a change in the relationship between precipitation and distance from water body.

Other independent variables were a vegetation index, elevation, and derivatives of elevation (measures of the degree to which a slope faces north and east, and the curvature of the slope) at scales of 1 km and 9 km. The scale of 9 km was chosen to evaluate the scale over which convective rainfall occurs. Furthermore, this scale is consistent with the findings of Hession and Moore (2010) and Sharples et al. (2005); Sharples et al. (2005) identify an optimal topographic scale of dependence of around 6-8 km. Sharples et al. also indicate that for scales of over 30 km, analyses incorporating elevation yield results similar to analyses based on longitude and latitude alone.

Elevation at each station location was estimated using the SRTM 30 arc second Digital Elevation Model (DEM) raster, shown on Figure 4-2. At locations near the equator, 30 arc seconds is close to 1 km resolution. Average elevation at a 9 km resolution was calculated for each raster cell using the nine cells centered on that cell and the FOCALMEAN function in ArcGIS (ESRI, 2006). Elevation data within a large buffer surrounding the study area were included to avoid edge effects or biased results within the study area.

The eastern and northern components of the unit normal vector were used to represent the effects of aspect and slope on precipitation (Hutchinson 1998b). These components were calculated at both scales (1 km and 9 km) as follows:

$$p = \cos(\alpha) \sin(\theta)$$

$$q = \sin(\alpha) \sin(\theta)$$

where α is the angle of aspect in degrees and θ is the angle of the slope in degrees. The values of p and q represent aspect scaled by the steepness of the slope. That is, these values are largest in magnitude on the steepest slopes and approach zero in flat areas. In addition, curvature was calculated in ArcGIS (ESRI, 2006) at the scales of 1 km and 9 km to evaluate whether curvature has an impact on precipitation patterns.

Vegetation was represented by one of the first vegetation indices created, the normalized difference vegetation index (NDVI; Rouse et al 1974). NDVI is a ratio between (Near infrared - Red) and (Near infrared + Red). Absorption patterns between the red (0.55-0.68 μm) and near infrared (0.73-1.1 μm) portions of the spectrum provide an indicator of vegetation amount and vigor. The ratio is sensitive to the difference between near infrared and red, with increasing chlorophyll concentration and green leaf vegetation density increasing NDVI value. Theoretically, NDVI ranges between -1 and +1, although values typically range between 0 for bare ground to 0.7 for lush dense vegetation. NDVI data were obtained from the Global Inventory Modeling and Mapping Studies (GIMMS) dataset, a record of bimonthly NDVI beginning in July 1982. Average monthly NDVI data was calculated from the two bimonthly measurements in each month. A one-month lag was chosen as an optimal lag time to identify correlations between precipitation and vegetation (Brunsell, 2006; Richard and Poccard, 1998; Wang *et al*.,2003).

For the purpose of prediction using universal kriging with instrumental variables, one independent variable expected to influence NDVI but not used to predict precipitation patters was chosen. This variable was a measurement of soil pH, specifically the pH in the water found in the interstitial spaces of the soil. These data

were obtained from the Harmonized World Soil Database (FAO/IIASA/ISRIC/ISS-CAS/JRC 2009), developed in a collaborative effort between the Food and Agriculture Organization of the United Nations (FAO), the International Institute for Applied Systems Analysis (IIASA), the ISRIC-World Soil Information, the Institute of Soil Science – Chinese Academy of Sciences (ISSCAS), and the Joint Research Centre of the European Commission (JRC).

Data for each of the independent variables were extracted from the grid cells overlapping meteorological station locations and combined with precipitation data at each respective station using the Intersect Point Tool in Hawth's Analysis Tools (Beyer, 2002), an add on to ArcGIS. Table 4-1 summarizes the independent variables considered in the statistical analyses.

For comparison to results presented herein, the CRU TS 3.0 data generated by the Climatic Research Unit of the University of East Anglia (CRU) were obtained. The CRU TS 3.0 is a recent global data set, representing the years 1901 to 2006, based on the methodology developed for CRU TS 2.1 (Mitchell and Jones 2005). The CRU TS 2.1 data were generated by interpolating climate data onto a regular 0.5 degree grid following New et al (2000). The CRU TS 3.0 data were combined with the remaining data by extracting from the 0.5 degree grid to sampled locations, as described above.

4.3 Methods

Four methods were used to generate estimated precipitation surfaces for the case study area: local ordinary kriging (LOK), universal kriging (UK), a newly developed extension to universal kriging that incorporates simultaneity between precipitation and vegetation using an instrumental variable approach (UKIV), and spatial regression

modeling (SpReg). Each of these methods is described in detail in Chapter 3. LOK was selected as the only univariate method for estimating precipitation (i.e., only precipitation data were utilized in the analysis). The remaining methods are all multivariate, and incorporate data from independent variables that are correlated with precipitation in an effort to improve predictions at unsampled locations. For the multivariate methods, initial model selection was conducted using ordinary least squares (OLS) regression. The independent variables identified as predictors for precipitation in the selected months were then incorporated in the multivariate kriging analyses as well as the spatial regression analyses. The spatial regression techniques were also used for the purpose of hypothesis testing (Chapter 5) in a demonstration of their “added value”: spatial regression techniques are useful not only for prediction at unsampled locations, but may also be used to improve understanding and test for significant predictors of the spatial distribution of precipitation.

Table 4-1. Summary of independent variables used in OLS regression analysis. Abbreviated names are shown as well.

Independent Variable	Scale	Abbreviated Name	Notes
Distance to Lake Victoria (km)		dist2lv	
		d2lv300	= 1 if dist2lv <= 300 km, 0 otherwise
		d2lv450	= 1 if dist2lv > 300 and <= 450 km, 0 otherwise
		d2lv600	= 1 if dist2lv > 450 and <= 600 km, 0 otherwise
			if all 3 indicator variables are 0, then dist2lv > 600 km
Distance to Indian Ocean (km)		dist2coast	
		d2c500	= 1 if dist2coast <= 500 km, 0 otherwise
Elevation (m)	1 km	dem1km	
	9 km	dem9km	
Combined measure of apsect and slope: measure of "northness"	1 km	p1km	p = cos(α) sin(θ), where α is aspect angle and θ is slope angle
(unitless)	9 km	p9km	p = cos(α) sin(θ)
Combined measure of apsect and slope: measure of "eastness"	1 km	q1km	q = sin(α) sin(θ)
(unitless)	9 km	q9km	q = sin(α) sin(θ)
Surface curvature	1 km	curv1km	
(profile; unitless)	9 km	curv9km	
NDVI	8 km	ndvi842	Average NDVI for February 1984
(unitless)	8 km	ndvi845	Average NDVI for May 1984
	8 km	ndvi849	Average NDVI for September 1984
	8 km	ndvi8412	Average NDVI for December 1984
	8 km	ndvi852	Average NDVI for February 1985
	8 km	ndvi855	Average NDVI for May 1985
	8 km	ndvi859	Average NDVI for September 1985
	8 km	ndvi8512	Average NDVI for December 1985

4.3.1 Model selection

Monthly precipitation data for four months (January, April, August, and December) in 1984 and 1985 were evaluated in eight independent analyses over space. Initial model selection was performed using OLS regression. Although automated selection procedures such as stepwise regression are commonly used in model selection, recent studies have identified many pitfalls in these approaches such as model misspecification and inaccurate results due to multicollinearity and confounding between independent variables. An alternative approach for developing and evaluating a set of candidate models described by Burnham and Anderson (1998) was used. Central to this approach is abandoning the use of automated variable selection methods in favor of careful *a priori* specification of candidate models (hypotheses) of particular interest. A number of candidate models were then compared directly and ranked based on an information theoretic criterion, the Akaike Information Criterion (AIC; Akaike, 1974). This approach, therefore, provided a means to rank the relative strength of models leading to an understanding of their predictive power and uncertainties. Importantly, this methodology is not susceptible to the instabilities of variable selection procedures previously mentioned.

Development of statistical models in this way requires careful thought and engagement between discipline experts and statisticians. Variables identified as possible predictors of spatial patterns in precipitation were selected in collaboration with experts in East African climatology and vegetative land cover, then tested in several regression model formulations.

As shown on Table 4-2, precipitation amounts for each month were evaluated using four OLS regression model formulations, or candidate models. The candidate models were developed to include the following sets of independent variables: (1) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 1 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean; (2) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 9 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean; (3) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 1 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean and interaction terms between distance to Lake Victoria (dist2lv) and the first two indicator variables (i.e., d2lv300 and d2lv450); and (4) average monthly NDVI for the subsequent month, each elevation term (dem, p, q, and curv) at a 9 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean and distance to the Indian Ocean and interaction terms between dist2lv and the first two indicator variables (i.e., d2lv300 and d2lv450).

The AIC value is shown for each model formulation on Table 4-2. Lower AIC values indicate better model fits. Each model is ranked based on AIC values; models with a ranking of 1 were selected for subsequent use in each multivariate prediction technique. Final models are summarized in Appendix 1.

Monthly precipitation values were transformed to normality using a normal score transformation (Perttunen and Stuckman, 1990; Wu *et al*., 2006) prior to developing the

candidate models above. Transformations of the dependent variable are commonly used in regression modeling to stabilize variance and improve normality (Gibbons, 1994; Hutchinson, 1998a; Hutchinson, 1998b; Neter *et al*., 1990; Schabenberger and Gotway, 2005; Sharples *et al*., 2005), particularly for variables that exhibit asymmetrical distributions, such as precipitation. Precipitation data are more often right-skewed in distribution since the range of possible values is restricted by a lower bound of zero. This issue may be exacerbated during dry months, when precipitation amounts are lower and more often equal to zero.

The normal score transformation is a monotonic transformation that replaces the observed value with a "typical" value from the standard normal distribution that corresponds to the same order statistic, or ranked value. Thus, the normal score transform ensures that the assumptions of normality and equal variances are met (Perttunen and Stuckman, 1990). Back-transformation of the results, allowing for interpretation in the original data scale, is relatively straightforward (Wu *et al.*, 2006).

The normal score transformation was completed using an R function (Shortridge 2010), modified to allow for right skewed distributions such as precipitation. In summary, this R function (and the normal score transform used for this application) calculates normal scores for a vector of values such as precipitation at multiple locations. First, the order statistics for the vector of values are determined

Table 4-2. Summary of OLS regression models ranked by AIC.

Model	Dep. Variable	Independent Variables	AIC	Rank
1	Jan 84 Precip Nscores	ndvi842, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	293.9	(4)
2	Jan 84 Precip Nscores	ndvi842, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	292.6	(3)
3	Jan 84 Precip Nscores	ndvi842, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	291.4	(2)
4	Jan 84 Precip Nscores	ndvi842, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	288.4	(1)
1	Apr 84 Precip Nscores	ndvi845, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	291.3	(3)
2	Apr 84 Precip Nscores	ndvi845, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	282.0	(1)
3	Apr 84 Precip Nscores	ndvi845, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	294.0	(4)
4	Apr 84 Precip Nscores	ndvi845, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	284.3	(2)
1	Aug 84 Precip Nscores	ndvi849, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	229.1	(4)
2	Aug 84 Precip Nscores	ndvi849, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	228.9	(3)
3	Aug 84 Precip Nscores	ndvi849, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	223.7	(1)
4	Aug 84 Precip Nscores	ndvi849, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	224.3	(2)
1	Nov 84 Precip Nscores	ndvi8412, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	277.2	(2)
2	Nov 84 Precip Nscores	ndvi8412, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	273.5	(1)
3	Nov 84 Precip Nscores	ndvi8412, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	281.2	(4)
4	Nov 84 Precip Nscores	ndvi8412, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	277.4	(3)

Table 4-2 (continued). Summary of OLS regression models ranked by AIC.

1	Jan 85 Precip Nscores	ndvi852, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	288.4	(2)
2	Jan 85 Precip Nscores	ndvi852, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	296.6	(4)
3	Jan 85 Precip Nscores	ndvi852, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	279.8	(1)
4	Jan 85 Precip Nscores	ndvi852, dem9km, p9km, q9km, curv9km,d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	291.0	(3)
1	Apr 85 Precip Nscores	ndvi855, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	311.7	(1)
2	Apr 85 Precip Nscores	ndvi855, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	315.5	(3)
3	Apr 85 Precip Nscores	ndvi855, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	312.6	(2)
4	Apr 85 Precip Nscores	ndvi855, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	317.1	(4)
1	Aug 85 Precip Nscores	ndvi859, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	218.9	(3)
2	Aug 85 Precip Nscores	ndvi859, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	214.7	(1)
3	Aug 85 Precip Nscores	ndvi859, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	220.7	(4)
4	Aug 85 Precip Nscores	ndvi859, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	216.6	(2)
1	Nov 85 Precip Nscores	ndvi8512, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	268.2	(3)
2	Nov 85 Precip Nscores	ndvi8512, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	264.4	(1)
3	Nov 85 Precip Nscores	ndvi8512, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	271.5	(4)
4	Nov 85 Precip Nscores	ndvi8512, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	268.0	(2)

Differences of less than 3 between AIC values are not considered significant.

(i.e., the data are sorted and ranked from smallest to largest). The normal score for each observation is then found as the "typical" value of the *k*th smallest number from a standard normal distribution (i.e., a normal distribution with a mean of zero and a standard deviation of one) (Perttunen and Stuckman, 1990). Results are back-transformed to the original data scale by linear interpolation between data. Values above the highest observed datum are extrapolated: (1) the mean of the original data is calculated and values above the mean identified, (2) the standard deviation of those values is calculated, (e) if the mean score plus the score of the maximum value multiplied by the standard deviation calculated in step 2 is less extreme than the actual maximum, the actual maximum is used instead. No extrapolation was completed for low values since the lower end of the output distribution was constrained to zero.

4.3.2 Consideration of other candidate OLS models

Many other models were considered prior to selecting the candidate models summarized in Table 4-2. For example, models using the continuous distance measures (dist2lv and dist2coast) were developed; however, AIC values were consistently higher (worse) for these models. In addition, inconsistent results were obtained due to multicollinearity between the continuous distance measures and other variables such as NDVI.

Elevation cutoffs were developed to allow for potentially non-linear relationships between precipitation and elevation and to account for the relatively small number of precipitation stations at higher and lower elevations. Cutoffs of 1,000 meters and 2,500 meters were used. These variables were not significant in various modeling attempts

due, in large part, to the significance of the distance “bands” that were developed which also indirectly stratified rain stations by elevation.

In addition, models were attempted without transforming the monthly precipitation data. These models consistently yielded large numbers of negative predicted values for precipitation and diagnostic plots that confirmed the presence of heteroskedasticity (i.e., unequal variances) and non-normal residuals, thus violating the underlying assumptions of the regression techniques being used.

4.3.3 Prediction of spatial patterns in monthly precipitation

The following four methods were used to estimate precipitation surfaces for precipitation in the months of January, April, August, and December, in 1984 and 1985:

- Local ordinary kriging with nine nearest neighbors
- Universal kriging
- Universal kriging with an instrumental variable
- Spatial regression (weights matrices developed using nine nearest neighbors)

Local ordinary kriging and spatial regression both require specification of the number of nearest neighbors to be incorporated in the analysis. A weighting scheme incorporating nine nearest neighbors was selected, consistent with the work of New et al (2000) and Piper and Stewart (1996). New et al (2000) used eight nearest neighbors to interpolate climate data anomalies for mapping purposes. Piper and Stewart (1996) used between five and ten neighboring stations to interpolate climate data.

Universal kriging with instrumental variables requires calculation of an instrument for NDVI, the variable that is believed to be simultaneously related with precipitation. In other words, a feedback simultaneity is believed to exist between precipitation and vegetation (NDVI). An instrument representing NDVI, to be used in place of NDVI in the UKIV analysis, was calculated using OLS regression, with the following independent variables: the soil pH variable described in Section 4.2, Data, and a subset of independent variables also used in the prediction of precipitation patterns, i.e., distance to Lake Victoria, distance to the Indian Ocean coast, and elevation at a resolution of 9 km. Universal kriging with instrumental variables (UKIV) was included in the case study as a purely theoretical development, however, since formal testing for simultaneity between precipitation and the 1 month-lagged vegetation measure (NDVI) did not identify simultaneity for any month in 1985, the year selected to represent a typical precipitation year in East Africa. Formal testing was completed using the Hausman test (Hausman 1978).

Maps for estimated precipitation in April 1985 are shown on Figure 4-6. The remaining maps are provided in Appendix 1. Table 4-3 provides a summary of predicted values generated by each method. For comparison purposes, a summary of CRU TS 3.0 data is also provided on Table 4-3.

Predicted precipitation patterns are similar in all four maps in April 1985 (Figure 4-4), the first wet season of the year in the study area. The ITCZ is positioned directly overhead the study area during this month. Peak estimated rains for the month (approximately 600 mm) are depicted over Mount Kilimanjaro and, in particular, Mount Kenya. Drier regions (< 50 mm) appear along the eastern side of the study area, and

dry regions also appear generally north/northwest of Mount Kenya, and southwest of Mount Kenya. The LOK map illustrates streak- like features jetting out from the central portion of the study area where rain stations are located; this is an effect of the extrapolation tendencies of LOK and nearest neighbor configuration.

The effects of the distance bands, which were significant for all three cutoffs (i.e., 300 km, 450 km, and 600 km from Lake Victoria), are evident in the maps created using universal kriging, universal kriging with simultaneity, and spatial regression. The distance bands allowed for decreasing precipitation at locations further from Lake Victoria in a stepwise fashion. It was hoped that the continuous distance measure would result in a smoother surface; however, distance bands were necessary due to the complicated, non-linear relationship between distance from Lake Victoria and amount of precipitation and the confounding between the continuous distance measure and other independent variables (e.g., elevation and NDVI). In addition, elevation at a resolution of 1 km was a significant predictor of April 1985 precipitation, which is evident in the increased rainfall amounts occurring at higher elevations (e.g., Mount Kenya, Mount Kilimanjaro, and the highlands of western Kenya). Figure 4-7 shows a map derived using universal kriging but excluding the dummy variables used to create the distance band; no “bulls-eye” effect is evident, demonstrating that this effect is due to the use of the distance bands.

Table 4-3. Summary of predicted precipitation amounts for each prediction method.

Month	Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Jan-84	Observed	0	1.45	7.8	20.2	31.8	129.8
	OK	0.8118	7.268	15.89	20.7	30.06	74.45
	UK	0	0	0	8.464	3.818	143.3
	UKIV	0	0	1.156	9.487	4.1	149.8
	SpReg	0	0	0	9.587	3.5	158.5
	CRU	2.82	6.69	8.49	9.518	11.58	19.49
Apr-84	Observed	3	67.4	96.7	138.2	199	821.1
	OK	5.936	69.34	119.8	130.2	181	728.4
	UK	3	37.31	57.35	64.69	75.31	821.1
	UKIV	3	44.65	67.57	79.56	86.43	821.1
	SpReg	3	58.46	71.46	70.12	86.59	821.1
	CRU	9.87	16.51	19.77	20.28	22.54	43.53
Aug-84	Observed	0	2.613	20.81	54.3	78.34	283.7
	OK	0.117	4.475	13.05	25.36	26.06	226.5
	UK	0	0	0	17.23	10.87	283.7
	UKIV	0	0	0	11.18	5.86	283.7
	SpReg	0	0	0	17.74	14.02	283.7
	CRU	0.2	3.1	5.99	7.256	12.55	16.44
Nov-84	Observed	0	87.45	131	170.4	202.3	767.2
	OK	49.8	99.63	162.8	183.4	245.9	529
	UK	0	55.5	87.28	88.27	116.2	767.2
	UKIV	0	55.5	82.15	84.96	108.6	767.2
	SpReg	0	73.02	95.24	94.19	119.2	767.2
	CRU	0.29	1.017	1.85	2.477	3.71	8.6
Jan-85	Observed	0	3.7	13.3	27.55	44.37	147.3
	OK	1.262	10.68	15.72	18.01	19.78	86.83
	UK	0	1.411	4.5	8.378	10.02	147.3
	UKIV	0	3.403	5.011	9.124	11.3	147.3
	SpReg	0	3.035	4.902	8.857	10.41	147.3
	CRU	0.43	1.03	1.385	2.17	2.89	11.79

Table 4-3 (continued). Summary of predicted precipitation amounts for each prediction method.

Month	Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Apr-85	Observed	3.2	175.4	258.9	251.9	335.8	543
	OK	12.79	147.3	187.5	191.9	233.5	455.4
	UK	3.2	55.15	175	157.1	241.2	543
	UKIV	3.2	87.16	196.1	195.3	267.3	543
	SpReg	3.2	149.8	189.7	170.9	221.9	543
	CRU	3.82	8.385	10.81	13.26	15	56.31
Aug-85	Observed	0	2.71	29.03	63.14	109.1	246.1
	OK	0.1317	7.015	20.24	38.65	62.5	189.2
	UK	0	2.02	4.764	28.68	37.18	246.1
	UKIV	0	5.2	10.8	39.21	72.73	246.1
	SpReg	0	2.566	5.613	22.17	25.44	246.1
	CRU	0	2.91	11.3	8.353	11.69	14.68
Nov-85	Observed	21.3	67.45	102.1	121.1	140.9	561.1
	OK	28.23	87.27	112.4	118.9	144.1	426.9
	UK	21.3	45.29	63	70.33	91.12	561.1
	UKIV	21.3	49.29	76.03	76.5	94.53	561.1
	SpReg	21.3	44.82	61.55	63.61	82.81	561.1
	CRU	4.21	7.08	8.96	14.47	21.05	44.6

Note: UK, UKIV, and SpReg all used normal score transformations of average monthly precipitation values. Values were transformed back to the original scale using a back transform function.

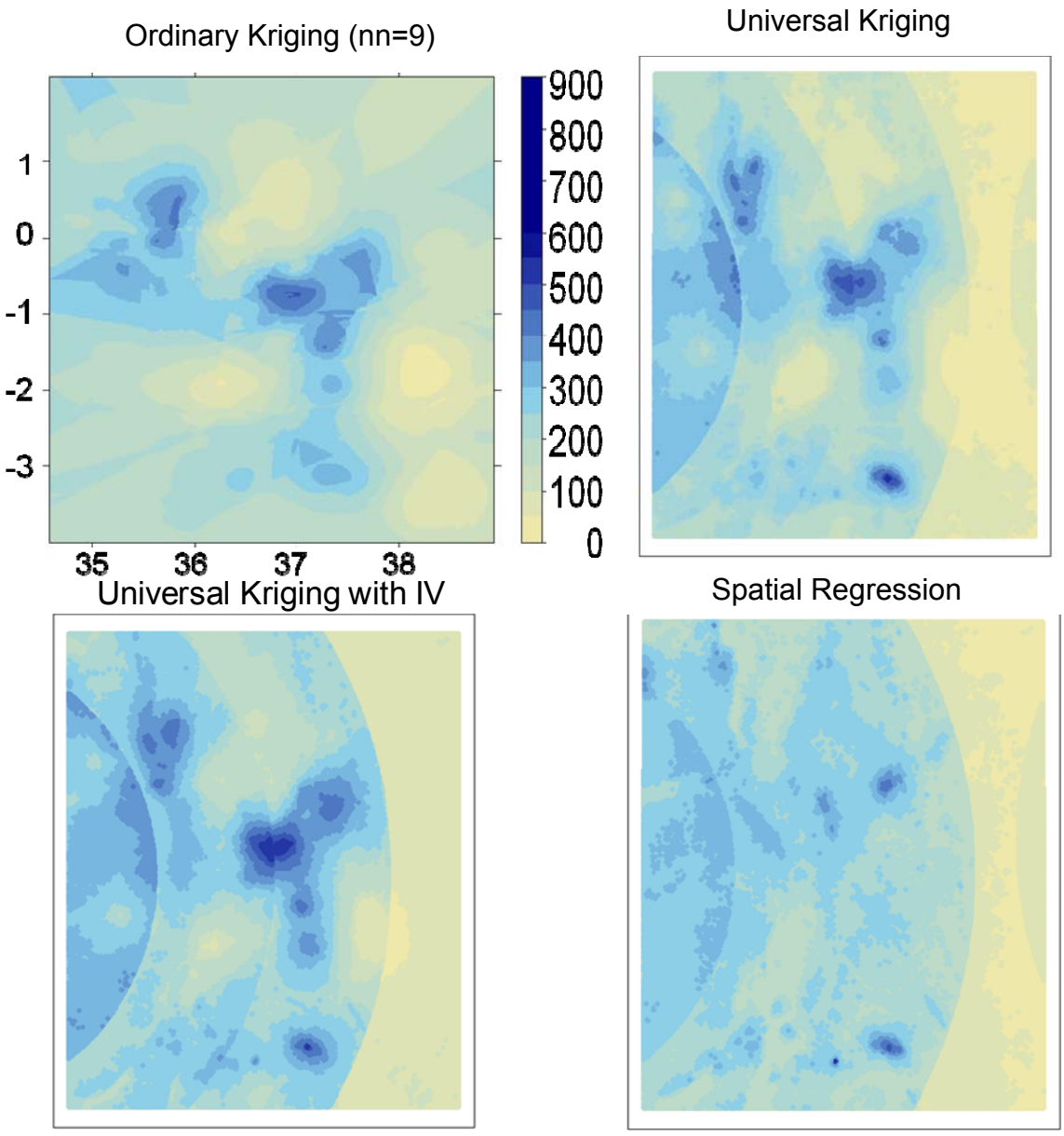

Figure 4-6. Precipitation maps for April 1985 generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

Universal Kriging

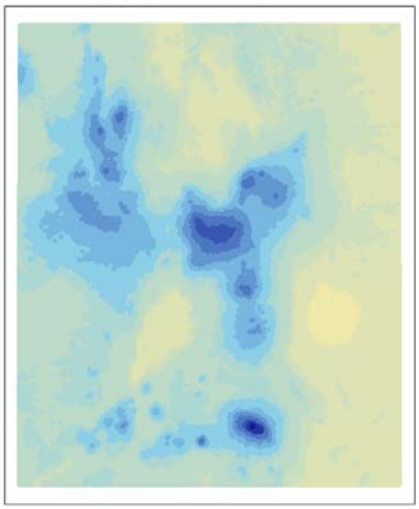

Figure 4-7. Precipitation map for April 1985 generated using UK without distance bands.

Estimated precipitation maps for all of the months evaluated in this dissertation are provided in Appendix 1. Based on these maps, the following observations were made.

January is the month during which the ITCZ is at its most extreme southerly position; consequently, it is a dry month near the equator. All four of the **January 1984** precipitation maps show the highest levels of precipitation on the western edge of the study area, nearest Lake Victoria. Peaks shown in the eastern central and southern portions of the OK map, corresponding with high elevation locations (i.e., Mount Kenya and Mount Kilimanjaro), are also evident on the UK map. These spatial patterns in precipitation are consistent with the expectation that rains occurring in the dry seasons are mainly localized convective rainfall near Lake Victoria or stratiform rainfall in the highland areas (Ng'ang'a 1992). Maps for UK, UKIV, and OLS regression (Figure 1,

Appendix 1) appear very similar. These three models were all based on the same OLS model which incorporated categories for distance from Lake Victoria with interaction terms for the < 300 km and 300 to 450 km categories. The scales for these maps are also similar, extending to approximately 150 mm.

The range of precipitation amounts is much higher in **April 1984**, extending to 728 mm in the OK map, and to approximately 820 mm in the remaining maps. This is expected since the ITCZ is directly over the study area at this time of year. The OK map also depicts precipitation across the study area, although in the remaining maps it appears that the rainfall is concentrated mostly in the west, near Lake Victoria. This is partly the result of the categorization scheme used in these map scales, since the lowest category extends to approximately 55 mm. The OK map shows a strong peak at the location of Mount Kilimanjaro. This peak also occurs on the UK and UKIV maps.

Precipitation maps for **August 1984** once again show a generally lower range of precipitation amounts, ranging up to 280 mm. The ITCZ is at its northernmost position in August, causing the second dry season of the year in the study area. Similar to January 1984, higher levels of precipitation in August 1984 occur mainly in the western portion of the study area near Lake Victoria and in the Western Highlands. Higher precipitation levels are also seen over the peaks of Mounts Kenya and Kilimanjaro. The rings shown in the UK, UKIV, and OLS regression maps illustrate the interaction between elevation effects within distance band nearest Lake Victoria. The second distance band, from 300 to 450 km from Lake Victoria, illustrates the effect of location within that band, which is decreased precipitation. A decreasing trend with elevation is evident within the second distance band. The overall spatial pattern of precipitation in

August 1984 is also consistent with the expectation that dry season rains are mainly localized convective rainfall near Lake Victoria or stratiform rainfall in the highland areas (Ng'ang'a 1992).

In **November 1984**, the ITCZ is over the study area again, providing the second rainy season, also known as the "short rainy season" in the region. Precipitation amounts range up to approximately 770 mm in the UK, UKIV, and spatial regression maps, and up to 530 mm in the OK map. Precipitation is generally present in elevated amounts across the study region, with the highest amounts in the eastern central portion of the area. The spatial patterns in the November 1984 precipitation maps, illustrating more widespread areas of higher precipitation, are consistent with the larger-scale effects of the passing ITCZ. The map based on spatial regression appears somewhat smoother than the kriging-based maps.

The precipitation maps for **January 1985** appear similar to the January and August 1984 maps than the January 1984 maps; the range of precipitation is closer to that of January 1984 (i.e., maximum predicted rainfall amounts reach approximately 150 mm). As it has been shown, precipitation is predominantly in the western portions of the study area in the dry seasons, consistent with the mesoscale effects of Lake Victoria and of increased elevation. Although not observed in January 1985, the interaction between the distance to Lake Victoria terms in the multivariate models and elevation are evident, similar to August 1984.

Similar to April 1984, precipitation maps for **April 1985** depict the higher rainfall levels that occur during the long rainy season ranging up to approximately 550 mm. Increased levels of rainfall are evident throughout most of the study area, consistent

with the large-scale effects of the passing of the ITCZ. The highest levels of rainfall are observed over Mounts Kenya and Kilimanjaro; this pattern is most evident in the OK, UK, and UKIV maps. The spatial regression map appears smoother in spatial distribution than the kriging-based maps, with less differentiation between areas of peak rainfall and areas of lower rainfall amounts.

Precipitation amounts predicted for **August 1985** are similar in spatial distribution. All four maps depict higher rainfall levels in the northwest of the study region, ranging up to approximately 250 mm. All three of the multivariate prediction methods (UK, UKIV, and spatial regression) show more detail in the central and south central portions of the study area, where precipitation amounts appear higher than in the OK map. These spatial patterns in precipitation are consistent with the mesoscale effects of Lake Victoria and of higher elevations expected during the dry seasons in East Africa. The streaking features common in maps generated by LOK are clearly present.

November 1985 precipitation patterns are similar in distribution to the November 1984 rainfall amounts, although generally lower overall. Although 1985 was observed to be a more typical year, 1984 was atypical in that rainfall amounts were lower than usual in the early part of the year and higher than most years in the later part of the year. Precipitation amounts range to approximately 430 mm (LOK map) to approximately 560 mm (UK, UKIV, and OLS regression maps). The highest precipitation amounts are shown in the southeastern portion of the study area over Mount Kilimanjaro, extending northward towards Mount Kenya. The effect of Lake Victoria is also evident in all four maps.

The bimodal nature of precipitation throughout the year along the equator in East Africa is also evident in Table 4-3, which provides a summary of predicted precipitation amounts by prediction method. Summary statistics of predicted precipitation amounts for all four prediction methods illustrate two wet seasons and two dry seasons each year. There is a tendency of the multivariate methods (i.e., UK, UKIV, and regression methods) to predict larger proportions of zeroes, particularly in the dry months and higher maximum predictions, resulting in more highly right-skewed rainfall amounts. LOK generally yields a smaller range of precipitation amounts: the lower percentiles are generally higher for LOK, and the maximum values are lower.

CRU precipitation estimates are consistently lower than estimates generated herein (Table 4-3). This is expected since global data sets are often too smooth and do not reflect the more extreme values. In some months, CRU results correspond to lower percentiles of the distribution more closely (e.g., January 1984, April 1984, August 1984, January 1985, and August 1985). Months with particularly low CRU values include April and November of 1984, and April and November of 1985, representing the wet seasons in the study area. It appears that the CRU data greatly underestimate precipitation in wet East African seasons.

4.3.4 Comparison and evaluation of map accuracy

Accuracy of the precipitation maps was evaluated by comparing root mean square errors (RMSEs) for each the four maps generated at each time interval. RMSEs are summarized in Table 4-4.

Table 4-4. Comparison of root mean squared errors

	OK	UK	UKIV	Sp. Reg.	CRU
Month	rmse	rmse	rmse	rmse	rmse
Jan-84	***20.162***	24.368	20.840	20.529	26.461
Apr-84	***74.707***	166.342	99.637	103.903	161.403
Aug-84	***34.903***	82.620	39.360	38.558	79.831
Nov-84	***101.027***	175.267	122.138	119.139	212.218
Jan-85	***23.669***	37.768	26.888	24.781	40.682
Apr-85	***89.125***	270.690	90.644	96.990	267.046
Aug-85	***34.798***	89.018	39.504	35.705	84.766
Nov-85	73.881	130.970	72.749	***67.992***	130.234

Root mean square errors are generally lowest for LOK, the univariate method of estimation. Spatial regression provided the lowest RMSE for November 1985 precipitation estimates. In general, RMSE values are closest in dry seasons, particularly January 1984, January 1985, August 1985. Spatial regression yielded lower RMSEs than either UK method in four of the eight months.

RMSE values are consistently the highest for the CRU data, particularly during the wet seasons. In some cases, the CRU RMSE is twice as high as the others, indicating the high amount of error from corresponding observations collected at meteorological stations in the study area.

Since RMSEs represent overall map accuracy with a single number, precipitation map accuracy was also evaluated using maps of standardized residuals to evaluate for spatial patterns in error terms. Large positive residuals (i.e., zscore > 2) are indicated with red dots and large negative residuals (i.e., zscore < -2) are shown with blue dots. A map of standardized residuals is shown in Figure 4-7 for **April 1985**.

In April 1985, two areas of significant residuals are generally evident: in the northwest of the study area in the western Kenya highlands, and in the center of the study area near Mount Kenya. In most cases, both high and low significant residuals are observed in the western Kenya highlands, indicating that there is a large amount of variability in precipitation in this region. This may be due, in part, to the highly variable topography in the area. The UK model underestimated precipitation in a few locations in the center of the study area, while the SpReg model overestimated precipitation in this region. The UKIV model has the fewest significant standardized residuals, indicating a generally improved model fit. The RMSE for the UKIV model (29.79) is very close to the smallest RMSE (29.71) achieved by the OK model.

Standardized residual maps for all of the months evaluated in this dissertation are provided in Appendix 1. Based on these maps, the following observations were made.

Patterns of significantly high or low residuals are somewhat similar across prediction methods and months being evaluated. In **January 1984**, several significant negative residuals are shown for the OLS method. A cluster of low residuals is present in the northwest of the maps for all four methods, although a high residual is shown very near to at least one low residual in all cases, indicating highly variable rainfall amounts in the region. Large positive residuals are indicative of the model underestimating actual precipitation; large negative residuals occur when the model overestimates actual precipitation.

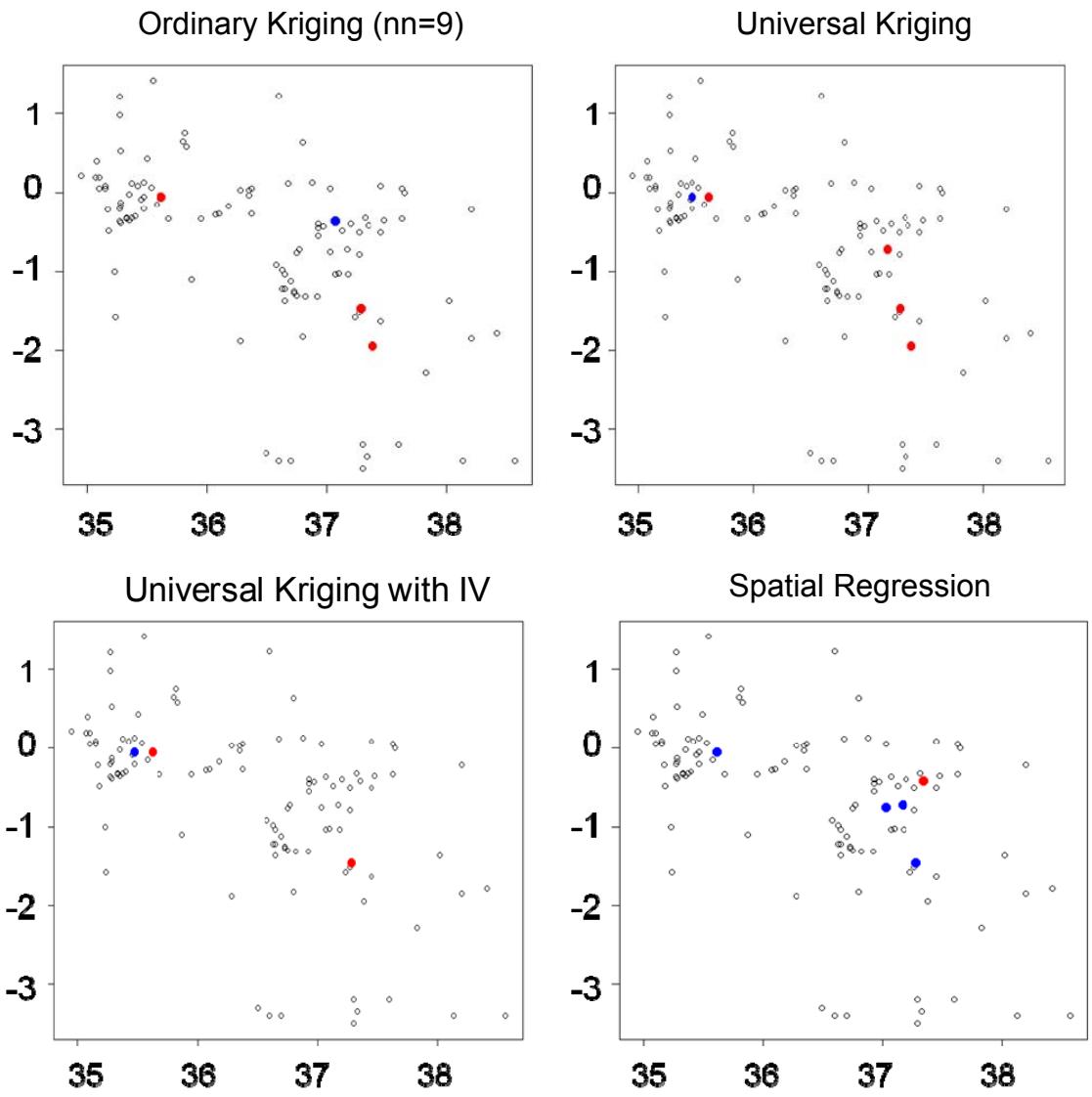

Figure 4-8. Maps of Significant Cross Validation residuals (|z-score|>2) in April 1985 for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right). Significantly high residuals are shown in red, and significantly low residuals are shown in blue.

Significant residuals are present in the central/eastern regions of the study area in **April 1984**. UK residuals depict a cluster of high residuals east of Mount Kenya, and another cluster of high residuals in northern Tanzania, indicating that the UK model overestimated precipitation in these locations.

In general, there are more significantly low residuals than high residuals in **August 1984**. UK and UKIV residual maps show four significantly low residuals in the center of the study region; however, in most cases they are surrounded by non-significant residuals. Another area in which significant residuals is present is the in along the western edge of the study area. Both positive and negative residuals appear in this region, indicating high amounts of rainfall variability over space.

UK and UKIV residual patterns are similar to each other in **November 1984**, with negative residuals in the center of the study area (indicating that the models overestimate in these locations) and positive residuals present in the northwest and southeast, where the model underestimates precipitation. Significant residuals extend to the northwest and southeast regions of the study area. Significant residuals are mainly negative on the LOK and SpReg maps, indicating that the models generally overestimate precipitation in these areas.

There are fewer significant residuals in **January 1985** than in January 1984. LOK predicted values significantly underestimate actual precipitation amounts in several locations. UK and UKIV maps illustrate a general scattering of significant positive and negative residuals, with the negative residuals falling in more central areas.

In **August 1985**, negative residuals are shown generally in the center of the study region, with the exception of LOK, for which two high residuals are shown in the central study area. This indicates that UK, UKIV, and SpReg overestimate precipitation at a cluster of locations in the center of the study region, while LOK is more likely to underestimate precipitation is this region.

In **November 1985**, there are generally more negative than positive residuals; however, in this month the positive residuals are generally interspersed with the negative residuals in the central portion of the study area, indicating higher precipitation variability in November 1985.

Maps of CRU residuals for 1985 (the “typical” year) were also plotted to evaluate for spatial patterns in the residuals (Figure 4-8). From these plots, it can be seen that there are large numbers of significantly high residuals (|standardized residual or zscore| > 2), generally in clusters. High residuals indicate locations where the CRU modeled values significantly underestimate precipitation observed at the corresponding meteorological station. Again, this is expected due to the overly-smoothed nature of global climate model estimates of climatic variables. The significant residuals are generally clustered, in large part, because the global model estimates do not take into account local features that influence precipitation on a local scale, such as topography and surface water bodies. Because these features are correlated over space, error terms would also be expected to be correlated over space.

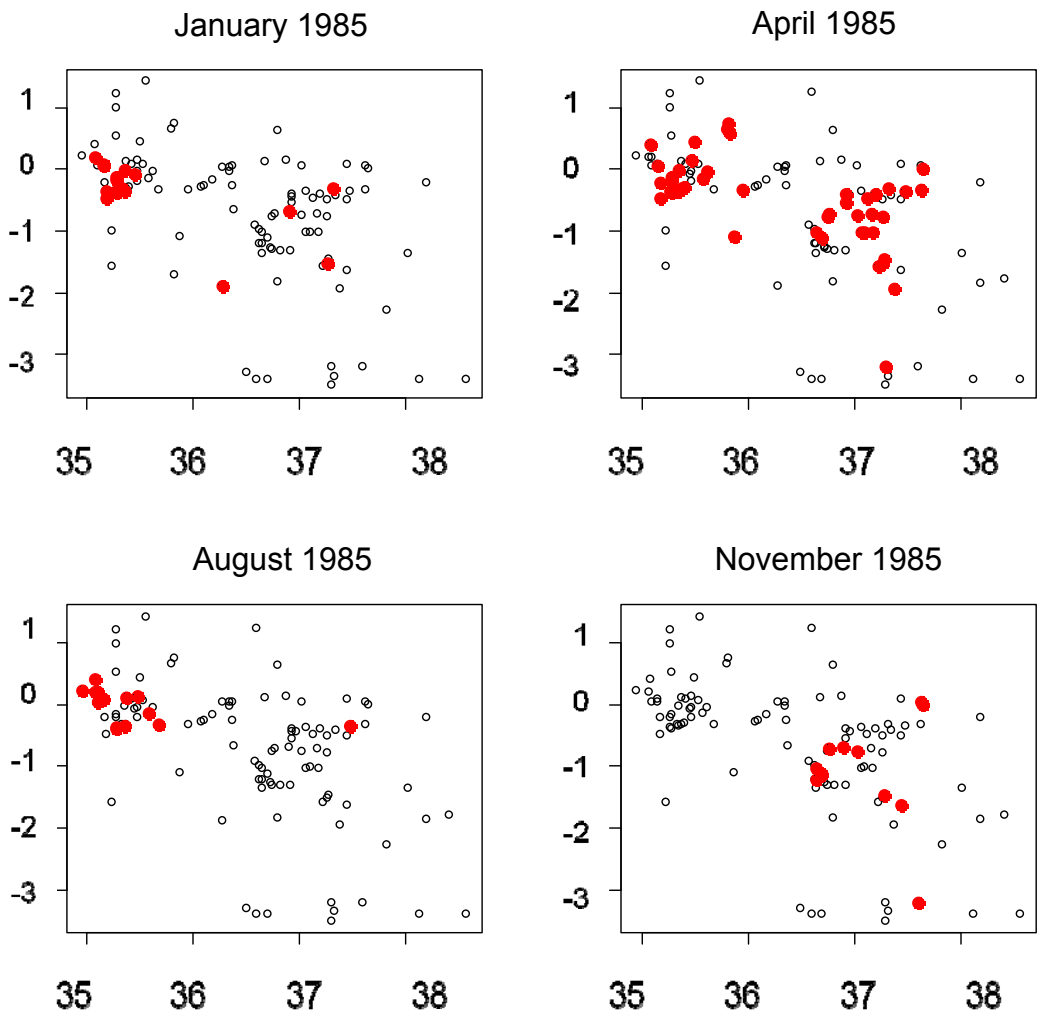

Figure 4-9. Maps of CRU residuals in 1985. Significantly high residuals (zscore > 2) are shown in red, and significantly low residuals (zscore < -2) are shown in blue.

In identifying an optimal model, LOK performed the best on the basis of minimum RMSE, a single map-wide measure. This is result is consistent with the findings of Goovaerts (1999a, 2000), in which a local form of simple kriging (simple kriging with varying local means) outperformed multivariate techniques when applied to the mountainous terrain in The Algarve, Southern Portugal.

The multivariate techniques were close runners up in some months, and had the lowest RMSE in November 1985, in which the spatial regression RMSE was the lowest. As shown on Figure 4-5, the least typical year in terms of monthly rainfall amounts was 1984 compared to all other years from 1980 through 1985, with droughts occurring in the early part of the year, and surplus rain in the later months of the year. It could be surmised from these results that multivariate methods of interpolation perform better in typical years than in atypical years, in which other drivers of precipitation may contribute to monthly precipitation patterns.

Although many studies have identified the importance of including topographic variables such as elevation and its derivatives in predicting precipitation patterns over space (Arora et al. 2006; Daly et al. 1994; Diodato 2005; Goovaerts 1999a, 1999b, 2000; Hutchinson 1998b; Hutchinson and Bischof 1983; KeifferWeisse and Bois 2001; Kyriakidis et al. 2004; Marquinez et al. 2003; Oettli and Camberlin 2005; Pardo-Iguzquiza 1998), it can also be hypothesized that local forms of univariate kriging perform better than multivariate techniques in areas of highly variable terrain due to the difficulty in modeling the complexity of precipitation patterns with multivariate models in these regions.

Chapter 5

Hypothesis Testing

5.1 Introduction

This chapter makes use of the regression modeling results described in Chapter 4 for the purpose of hypothesis testing in order to gain an improved understanding of the significant predictors of monthly precipitation patterns over space within the case study area. Significant variables in predicting the spatial patterns of precipitation are identified through a combination of a regression model selection process, in which competing models with varying selections of independent variables are compared and selected on the basis of the Akaike Information Criterion (AIC, Akaike 1974), and formal testing of the regression coefficients within each selected model. One additional test, the Hausman Test, is presented and used to identify whether significant simultaneity can be documented between precipitation and vegetation (Hausman 1978).

5.1.1 Study area

The position of the study area in East Africa, as well as its proximity to Lake Victoria and the Indian Ocean are illustrated in Figure 5-1. Understanding of these relative locations is importance since explanatory variables for describing variation in monthly rainfall include distance to Lake Victoria and distance to the Indian Ocean.

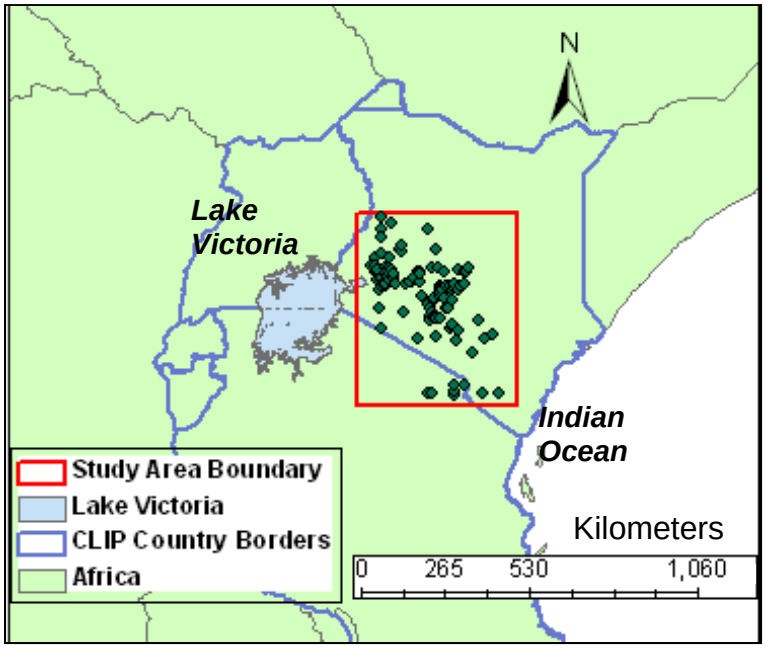

Figure 5-1. Location of the case study area within East Africa (CLIP region). Meteorological stations (green diamonds) are shown in proximity to Lake Victoria and the Indian Ocean.

5.1.2 Identification of dependent variable and selection of representative data

Monthly precipitation (mm) represents the dependent variable in this analysis. Precipitation data from roughly 120 meteorological stations obtained from the Department of Meteorology, Government of Kenya were used. Meteorological station locations are shown on Figure 5-1.

Monthly precipitation was evaluated through eight different regression analyses; the data for these analyses were chosen to represent the four seasons (i.e., the dry season in December, January, and February; the long rainy season in March, April, May, and June; the cool dry season in July and August; and the short rainy season in September, October, and November) in eastern Africa for two different years. The individual months of January, April, August, and November were chosen as the most

representative month of each of the seasons described above within the case study area (Figure 5-2).

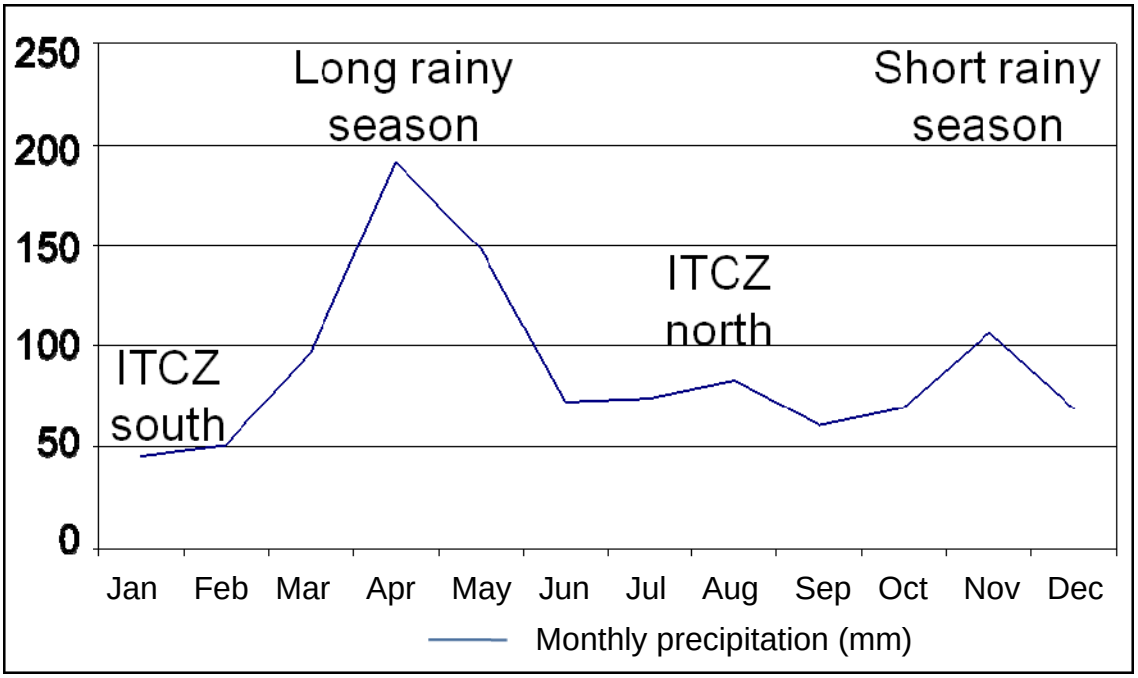

Figure 5-2. Long-term average monthly precipitation (mm) collected from 1926 to 1998 averaged over all meteorological stations within the study area. The position of the intertropical convergence zone (ITCZ) relative to the study area is labeled at various points in the year.

The years 1984 and 1985 were chosen for this analysis since these years represent the overlap between years in which the largest numbers of meteorological stations were measured (approximately 120 stations) and the years in which remotely sensed vegetation data were available (beginning in July 1982). Furthermore, the years to be evaluated were chosen to represent a typical year (1985) and an atypical year (1984). Additional detail regarding selection of years for analysis is provided in Chapter 4.

5.1.3 Identification of independent or descriptive variables

Distance from major water bodies (i.e., Lake Victoria and the Indian Ocean) was used to evaluate monthly precipitation in East Africa. As described in Chapter 4, categorical “distance bands” were used in lieu of the linear distance measures since precipitation was not linearly related to either of these measures. Dummy variables were established to identify areas within which precipitation levels were similar. Distance bands of 0 to 300 km, 300 to 450 km, 450 to 600 km, and greater than 600 km were created as a measure for distance to Lake Victoria. Distance bands of 0 to 500 km and greater than 500 km were used to represent distance to the Indian Ocean.

Other potential explanatory variables included NDVI (a vegetation index; Rouse et al. 1974), elevation, and derivatives of elevation (i.e., measures of the degree to which a slope faces north and east, and the curvature of the slope) at scales of 1 km and 9 km. The scale of 9 km was chosen to evaluate the scale over which convective rainfall occurs. Furthermore, this scale is consistent with the findings of Hession and Moore (2010) and Sharples et al. (2005); Sharples et al. (2005) identify an optimal topographic scale of dependence of around 6-8 km.

Elevation at each station location was estimated using the SRTM 30 arc second Digital Elevation Model (DEM) raster, shown on Figure 5-3. A measure of 30 arc seconds is approximately 1 km near the equator. Average elevation at a 9 km resolution was also calculated using elevation at surrounding pixels, as described in Chapter 4.

The influence of aspect and slope on precipitation was evaluated using the eastern and northern components of the unit normal vector (Hutchinson 1998). These

values characterize aspect scaled by the steepness of the slope. Results are largest in magnitude on the steepest slopes and approach zero in flat areas. In addition, curvature was calculated in ArcGIS (ESRI, 2006) to evaluate whether curvature has an impact on precipitation patterns. A summary of all potential explanatory variables used in this study is provided in Table 5-1.

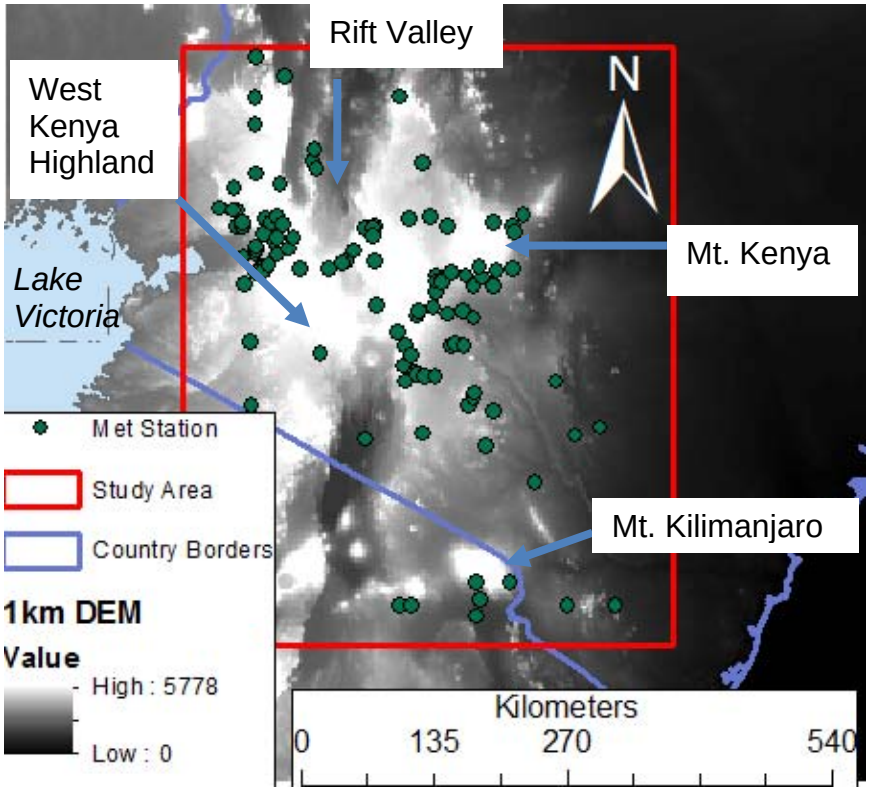

Figure 5-3. Map of elevation (1 km resolution) with meteorological station locations and study area boundary are shown above. Topographic features within the study area are labeled.

Table 5-1. Summary of independent variables used in regression analysis and hypthosis testing. Abbreviated names are shown as well.

Independent Variable	Scale	Abbreviated Name	Notes
Distance to Lake Victoria (km)		dist2lv	
		d2lv300	= 1 if dist2lv <= 300 km, 0 otherwise
		d2lv450	= 1 if dist2lv > 300 and <= 450 km, 0 otherwise
		d2lv600	= 1 if dist2lv > 450 and <= 600 km, 0 otherwise
			if all 3 indicator variables are 0, then dist2lv > 600 km
Distance to Indian Ocean (km)		dist2coast	
		d2c500	= 1 if dist2coast <= 500 km, 0 otherwise
Elevation (m)	1 km	dem1km	
	9 km	dem9km	
Combined measure of apsect and slope : measure of "northness"	1 km	p1km	p = cos(α) sin(θ), where α is aspect angle and θ is slope angle
(unitless)	9 km	p9km	p = cos(α) sin(θ)
Combined measure of apsect and slope: measure of "eastness"	1 km	q1km	q = sin(α) sin(θ)
(unitless)	9 km	q9km	q = sin(α) sin(θ)
Surface curvature	1 km	curv1km	
(profile; unitless)	9 km	curv9km	
NDVI	8 km	ndvi842	Average NDVI for February 1984
(unitless)	8 km	ndvi845	Average NDVI for May 1984
	8 km	ndvi849	Average NDVI for September 1984
	8 km	ndvi8412	Average NDVI for December 1984
	8 km	ndvi852	Average NDVI for February 1985
	8 km	ndvi855	Average NDVI for May 1985
	8 km	ndvi859	Average NDVI for September 1985
	8 km	ndvi8512	Average NDVI for December 1985

5.2 Regression model selection and testing

Initial regression model selection for precipitation in each month (i.e., January, April, August, and November) in 1984 and 1985 was conducted using OLS regression and the recommended approach of Burnham and Anderson (1998). OLS regression was used in this step since it is likely to identify more significant variables than spatially explicit regression models: in the presence of spatial autocorrelation, precision of the OLS estimates tends to be overstated, resulting in elevated estimates of the OLS coefficients and values such as the coefficient of determination (R^2).

All regression modeling was completed using a normal score transformation of the precipitation data (Chapter 4). Once an OLS model formulation was selected for each month in consideration, these models were further scrutinized to determine if they could be improved through spatial regression modeling (Anselin, 2006; LeSage 1998, LeSage and Pace 2009). This process is described in more detail in Chapter 2 of this dissertation. In Chapter 4, the final selections of regression models were utilized in all three of the multivariate methods used to generate predicted precipitation maps.

In this chapter, the final selection of regression models was further evaluated through hypothesis testing. The hypotheses to be tested are specified in Section 5.2.1 along with the basis for the hypotheses. Results of hypothesis and model interpretation are presented in Section 5.2.2.

5.2.1 Statement of hypotheses

A summary of the independent variables tested through regression modeling is provided in Table 5-1. This section provides a brief description of the independent

variables and their expected relationships with precipitation followed by a formal statement of hypothesis.

Distance measures (i.e., distance to Lake Victoria and distance to the Indian Ocean) were used to model synoptic and global scale processes (Figure 5-1). The distance variables, which were categorized to represent "distance bands," are expected to represent a complex interaction of different local, regional, and mesoscale processes that include the lake/sea breeze near Lake Victoria in the western portion of the study area and orographic effects in the highland areas (Camberlin and Planchon 1997, Ng'ang'a 1992). In general, higher levels of precipitation are expected to occur closer to the large water bodies, particularly in dry seasons when mesoscale effects dominate. It is also expected that the role of these factors will vary according to season.

Hypothesis 1

$H1_0$: Precipitation amounts do not vary with distance from water bodies.

$H1_1$: Precipitation amounts decrease as distance from water bodies increases.

Elevation (Figure 5-3) and its derivatives are expected to reflect mesoscale processes related to orographic precipitation. Increasing precipitation is expected to occur with increasing elevation (Arora *et al*., 2006; Spreen, 1947). This hypothesis was evaluated at the scales of 1 and 9 km.

Hypothesis 2

$H2_0$: Precipitation amounts do not vary with elevation.

$H2_1$: Precipitation amounts increase as elevation increases.

The eastern and northern components of the unit normal vector were used to represent the effects of aspect and slope on precipitation (Hutchinson, 1998). The largest values of *p* and *q* occur on the steepest slopes; the lowest values occur on peaks, valley floors, or generally flat areas such as savannas. Since *p* and *q* incorporate both slope and aspect, the effects of both can be evaluated. The direction of these effects can also be evaluated without reference to the prevailing wind field (Hutchinson, 1998). Steeper slopes provide stronger orographic lifting; thus, increasing values of *p* and *q* are expected to be associated with higher rainfall (locally, at least; Buytaert *et al*., 2006).

Hypothesis 3

$H3_0$: Precipitation amounts do not vary with the eastern and northern components of the unit normal vector, representing slope and aspect.

$H3_1$: Precipitation amounts increase with increasing values of the eastern/northern components of the unit normal vector.

The above relationship, or the significance of this relationship, is expected to vary by season. It is also anticipated that the spatial scale at which these processes occur

will vary by season due to different synoptic forcings, wind patterns, and other seasonal variations (e.g., long rains and short rains) (Ng'ang'a, 1992).

Curvature provides a measure of the degree to which a surface is convex or concave. It is calculated as the slope of the slope (the second derivative of the surface). High values of curvature represent upwardly convex surfaces, such as a hilltop. Conversely, low values of curvature correspond with upwardly concave surfaces, such as a valley bottom. Since higher curvature would generally occur at the relative higher elevations (e.g., at peaks), higher precipitation is expected to correspond with increased curvature. Furthermore, curvature was found to be less correlated with other independent variables than elevation; consequently, it is less likely to be confounded with other variables and may be a more effective predictor of precipitation.

Hypothesis 4

$H4_0$: Precipitation amounts do not vary with curvature.

$H4_1$: Precipitation amounts increase with increasing curvature.

Increased NDVI values (measured in the subsequent month) are expected to correlate with increased precipitation in the month being evaluated. This would be expected since increased precipitation causes increased greenness in most natural landscapes (Rodriguez-Iturbe and Porporato, 2004).

Hypothesis 5

$H5_0$: Precipitation amounts do not vary with NDVI.

$H5_1$: Precipitation amounts increase with increasing NDVI.

As previously noted, significance of the above variables and their corresponding hypotheses is expected to vary from month to month according to seasonal variability in the factors that affect spatial patterns in precipitation, such as position of the ITCZ, the relative importance of elevation and orographic precipitation in dry seasons versus rainy seasons, and the relative importance of constant sources of moisture, such as Lake Victoria.

Furthermore, the two years evaluated in this dissertation, 1984 and 1985, were chosen to represent atypical (1984) and typical (1985) in terms of monthly rainfall averages compared to long-term monthly averages. In particular, early months in 1984 were characterized by drought conditions; later months experienced higher precipitation than long term averages. The impact of this distinction on causal factors of precipitation patterns is unknown; therefore, no hypothesis is stated. Similarities and differences between these years were noted for future hypothesis development.

5.2.2 Results and interpretation

Initial model selection was performed using OLS regression, as described in Chapter 4. OLS regression was used in this step since it is likely to identify more significant variables than spatially explicit regression models: in the presence of spatial autocorrelation, precision of the OLS estimates tends to be overstated, resulting in

elevated estimates of the OLS coefficients and values such as the coefficient of determination (R^2).

A number of candidate models were compared and ranked based on an information theoretic criterion, the Akaike Information Criterion (AIC; Akaike, 1974). This approach provided a means to rank the relative strength of the OLS models. Regression models with the lowest AIC value were identified with a rank of (1) and selected for further hypothesis testing. Through this model selection process, some independent variables were excluded from further consideration.

As illustrated in Table 5-2, precipitation amounts for each month were evaluated using four OLS regression model formulations, or candidate models. The candidate models were developed to include the following sets of independent variables: (1) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 1 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean; (2) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 9 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean; (3) average monthly NDVI for the subsequent month, the elevation term and its derivatives (i.e., dem, p, q, and curv) at a 1 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean and interaction terms between distance to Lake Victoria (dist2lv) and the first two indicator variables (i.e., d2lv300 and d2lv450); and (4) average monthly NDVI for the subsequent month, each elevation term (dem, p, q, and curv) at a 9 km scale, each indicator variable representing distance to Lake Victoria and distance to the Indian Ocean and

distance to the Indian Ocean and interaction terms between dist2lv and the first two indicator variables (i.e., d2lv300 and d2lv450).

Results of initial model selection are summarized in Table 5-2. The AIC value is shown for each model formulation: lower AIC values indicate better model fits. Each model is ranked based on AIC values; models with a ranking of (1) were selected for subsequent use in each multivariate prediction technique.

The initial model formulations selected for each month and year were further tested using spatially explicit methods to determine if use of the OLS is appropriate, or if a spatially explicit model is required. The three spatially explicit regression models considered (SEM, SAR, and SAC) are described in detail in Chapter 3 of this dissertation, as well as the decision process for selecting the appropriate model type.

Table 5-2. Summary of OLS regression models ranked by AIC.

Model	Dep. Variable	Independent Variables	AIC	Rank
1	Jan 84 Precip Nscores	ndvi842, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	293.9	(4)
2	Jan 84 Precip Nscores	ndvi842, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	292.6	(3)
3	Jan 84 Precip Nscores	ndvi842, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	291.4	(2)
4	Jan 84 Precip Nscores	ndvi842, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	288.4	(1)
1	Apr 84 Precip Nscores	ndvi845, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	291.3	(3)
2	Apr 84 Precip Nscores	ndvi845, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	282.0	(1)
3	Apr 84 Precip Nscores	ndvi845, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	294.0	(4)
4	Apr 84 Precip Nscores	ndvi845, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	284.3	(2)
1	Aug 84 Precip Nscores	ndvi849, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	229.1	(4)
2	Aug 84 Precip Nscores	ndvi849, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	228.9	(3)
3	Aug 84 Precip Nscores	ndvi849, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	223.7	(1)
4	Aug 84 Precip Nscores	ndvi849, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	224.3	(2)
1	Nov 84 Precip Nscores	ndvi8412, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	277.2	(2)
2	Nov 84 Precip Nscores	ndvi8412, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	273.5	(1)
3	Nov 84 Precip Nscores	ndvi8412, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	281.2	(4)
4	Nov 84 Precip Nscores	ndvi8412, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	277.4	(3)

Table 5-2 (continued). Summary of OLS regression models ranked by AIC.

1	Jan 85 Precip Nscores	ndvi852, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	288.4	(2)
2	Jan 85 Precip Nscores	ndvi852, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	296.6	(4)
3	Jan 85 Precip Nscores	ndvi852, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	279.8	(1)
4	Jan 85 Precip Nscores	ndvi852, dem9km, p9km, q9km, curv9km,d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	291.0	(3)
1	Apr 85 Precip Nscores	ndvi855, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	311.7	(1)
2	Apr 85 Precip Nscores	ndvi855, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	315.5	(3)
3	Apr 85 Precip Nscores	ndvi855, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	312.6	(2)
4	Apr 85 Precip Nscores	ndvi855, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	317.1	(4)
1	Aug 85 Precip Nscores	ndvi859, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	218.9	(3)
2	Aug 85 Precip Nscores	ndvi859, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	214.7	(1)
3	Aug 85 Precip Nscores	ndvi859, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	220.7	(4)
4	Aug 85 Precip Nscores	ndvi859, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	216.6	(2)
1	Nov 85 Precip Nscores	ndvi8512, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500	268.2	(3)
2	Nov 85 Precip Nscores	ndvi8512, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500	264.4	(1)
3	Nov 85 Precip Nscores	ndvi8512, dem1km, p1km, q1km, curv1km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	271.5	(4)
4	Nov 85 Precip Nscores	ndvi8512, dem9km, p9km, q9km, curv9km, d2lv300, d2lv450, d2lv600, d2c500, d2lv300*dist2lv, d2lv450*dist2lv	268.0	(2)

Differences of less than 3 between AIC values are not considered significant.

The regression models selected for each month/year are summarized in Table 5-3. For each model, the regression technique used is given as well as the list of independent variables tested, their estimated coefficients, and measures of statistical significance. Since regression modeling was conducted using standardized precipitation data based on the n-score transformation (see Chapter 4 for additional detail), estimated coefficients are interpreted as the average change in standardized precipitation for each unit change in the independent variable. Significance at the levels of 0.1%, 1%, 5%, and 10% are indicated. Further discussion of significant variables is based on a 10% level of significance.

In **January 1984**, significant predictors of precipitation included NDVI (p=0.0259), distance to Lake Victoria (distances less than 300 km [p=0.0618] and between 300 and 450 km [p=0.0198]) and distance to the Indian Ocean (distances less than 500 km [p=0.0002]). As expected, locations closer to Lake Victoria were found to experience increased rainfall. Regression coefficients for the indicator variables are interpreted as the increase in mean expected rainfall (standardized) over that expected in the greatest distance category (i.e., greater than 600 km) when located within the given interval. For example, locations within 300 km of Lake Victoria were expected to receive approximately 10 additional standardized rainfall units (SRU) than locations greater than 600 km away. In this way, the greatest distance category is used as a baseline for evaluating the remaining distance categories. Locations between 300 and 450 km from Lake Victoria were expected to receive almost 4 more standardized rainfall units than locations more than 600 km away. Since the interaction term between d2lv450 (indicating a location of between 300 and 450 km from Lake Victoria) and the

continuous distance measure dist2lv was significant (p=0.0221) with a negative coefficient, rainfall was shown to decrease with increasing distance in this interval, as would be expected. Locations within 500 km of the Indian Ocean were expected to receive 1.2 additional SRU than further locations.

In **January 1984**, NDVI was negatively associated with precipitation (p=0.0259), indicating that increased NDVI corresponded with decreased precipitation. This result was not as expected. A possible explanation is the relatively low coefficient of determination (R^2=0.3) obtained for the January 1984 OLS regression model, indicating that only 30% of the variation in rainfall over space was explained by the model. Addition of one or more other explanatory variables may improve model interpretations.

Table 5-3. Summary of regression modeling results.

Month	Method	Independent Variable	Estimate	St. Error	t value	Pr(>|t|)	
Jan-84	OLS	(Intercept)	-1.3530	0.6059	-2.2330	0.0278	*
		ndvi842	**-1.5500**	0.6857	-2.2610	**0.0259**	*
		dem9km	0.0003	0.0003	1.1930	0.2358	
		p9km	-0.2401	0.1573	-1.5260	0.1302	
		q9km	0.0468	0.1590	0.2950	0.7690	
		curv9km	-50.6300	93.3400	-0.5420	0.5887	
		d2lv300	**10.2200**	5.4080	1.8890	**0.0618**	.
		d2lv450	**3.7850**	1.5980	2.3690	**0.0198**	*
		d2lv600	0.6514	0.5481	1.1890	0.2375	
		d2c500	**1.1610**	0.2960	3.9210	**0.0002**	***
		I(d2lv300 * dist2lv)	-0.0271	0.0194	-1.3990	0.1649	
		I(d2lv450 * dist2lv)	**-0.0081**	0.0035	-2.3260	**0.0221**	*
Apr-84	SARlag	(Intercept)	-1.9160	0.4924	-3.8910	0.0001	***
		ndvi845	**0.7846**	0.4462	1.7584	**0.0787**	**.**
		dem9km	0.0001	0.0002	0.5272	0.5981	
		p9km	-0.0479	0.1241	-0.3864	0.6992	
		q9km	0.2024	0.1256	1.6123	0.1069	
		curv9km	**-136.4866**	72.5866	-1.8803	**0.0601**	**.**
		d2lv300	**1.4931**	0.4961	3.0094	**0.0026**	**
		d2lv450	**0.9401**	0.4808	1.9552	**0.0506**	**.**
		d2lv600	**1.0393**	0.4441	2.3404	**0.0193**	*
		d2c500	**0.3932**	0.1891	2.0787	**0.0376**	*
		Rho	**0.6040**	0.1017	5.9358	**0.0000**	***
Aug-84	OLS	(Intercept)	-2.1604	0.4720	-4.5770	0.0000	***
		ndvi849	0.7530	0.5121	1.4710	0.1446	
		dem1km	**0.0006**	0.0002	3.2120	**0.0018**	**
		p1km	-0.0632	0.1242	-0.5090	0.6119	
		q1km	0.1014	0.1205	0.8420	0.4020	
		curv1km	-3.9075	6.7899	-0.5750	0.5663	
		d2lv300	-1.5345	4.1207	-0.3720	0.7104	
		d2lv450	**3.3980**	1.2118	2.8040	**0.0061**	**
		d2lv600	-0.1001	0.4035	-0.2480	0.8046	
		d2c500	**0.4681**	0.2371	1.9740	**0.0511**	**.**
		I(d2lv300 * dist2lv)	0.0116	0.0148	0.7880	0.4327	
		I(d2lv450 * dist2lv)	**-0.0076**	0.0026	-2.8860	**0.0048**	**
Nov-84	SARlag	(Intercept)	-1.4962	0.6446	-2.3211	0.0203	*
		ndvi8412	-0.1582	0.6878	-0.2300	0.8181	
		dem9km	0.0001	0.0002	0.4865	0.6266	
		p9km	**-0.3114**	0.1415	-2.2008	**0.0278**	*
		q9km	-0.0247	0.1390	-0.1776	0.8591	
		curv9km	66.7779	81.0492	0.8239	0.4100	
		d2lv300	**1.4101**	0.5406	2.6082	**0.0091**	**
		d2lv450	0.7143	0.5338	1.3382	0.1808	
		d2lv600	**1.2727**	0.4849	2.6246	**0.0087**	**
		d2c500	**0.7044**	0.2504	2.8133	**0.0049**	**
		Rho	0.2500	0.1629	1.5348	0.1248	

Table 5-3. Summary of regression modeling results.

Month	Method	Independent Variable	Estimate	St. Error	t value	Pr(>|t|)	
Jan-85	SARlag	(Intercept)	-0.7867	0.6318	-1.2450	0.2131	
		ndvi852	0.5605	0.5799	0.9665	0.3338	
		dem1km	-0.0002	0.0002	-0.7656	0.4439	
		p1km	0.0418	0.1515	0.2757	0.7828	
		q1km	**0.4667**	0.1441	3.2398	**0.0012**	**
		curv1km	12.2468	8.7368	1.4018	0.1610	
		d2lv300	-4.8020	4.5302	-1.0600	0.2891	
		d2lv450	**3.3218**	1.5306	2.1702	**0.0300**	*
		d2lv600	0.2569	0.5711	0.4498	0.6529	
		d2c500	0.3788	0.2664	1.4220	0.1550	
		I(d2lv300 * dist2lv)	0.0218	0.0161	1.3515	0.1765	
		I(d2lv450 * dist2lv)	**-0.0072**	0.0033	-2.1659	**0.0303**	*
		Rho	**0.2970**	0.1613	1.8406	**0.0657**	**.**
Apr-85	SARlag	(Intercept)	**-2.0913**	0.6588	-3.1744	**0.0015**	**
		ndvi855	-0.2445	0.7459	-0.3277	0.7431	
		dem1km	**0.0005**	0.0002	2.5010	**0.0124**	*
		p1km	0.1716	0.1474	1.1637	0.2446	
		q1km	0.2458	0.1510	1.6278	0.1036	
		curv1km	-6.1177	8.3879	-0.7294	0.4658	
		d2lv300	**1.5083**	0.5984	2.5206	**0.0117**	*
		d2lv450	**1.0736**	0.5685	1.8883	**0.0590**	**.**
		d2lv600	**1.1885**	0.5218	2.2777	**0.0227**	*
		d2c500	0.2507	0.2290	1.0947	0.2736	
		Rho	**0.6349**	0.1050	6.0476	**0.0000**	***
Aug-85	SARlag	(Intercept)	-1.0773	0.4340	-2.4820	**0.0131**	*
		ndvi859	**0.9904**	0.4516	2.1929	**0.0283**	*
		dem9km	**0.0005**	0.0002	3.0242	**0.0025**	**
		p9km	0.0794	0.1078	0.7373	0.4610	
		q9km	**0.2018**	0.1082	1.8647	**0.0622**	**.**
		curv9km	-9.1431	64.7750	-0.1412	0.8878	
		d2lv300	0.1276	0.4247	0.3005	0.7638	
		d2lv450	-0.3907	0.3993	-0.9787	0.3277	
		d2lv600	-0.4021	0.3687	-1.0904	0.2755	
		d2c500	-0.0420	0.2040	-0.2057	0.8370	
		Rho	**0.4170**	0.1329	3.1382	**0.0017**	**
Nov-85	OLS	(Intercept)	-2.7490	0.6493	-4.2330	**0.0001**	***
		ndvi8512	**2.2120**	0.5926	3.7330	**0.0003**	***
		dem9km	-0.0004	0.0002	-1.4950	0.1383	
		p9km	-0.0934	0.1640	-0.5690	0.5705	
		q9km	-0.0210	0.1629	-0.1290	0.8976	
		curv9km	**177.2000**	94.1500	1.8820	**0.0629**	**.**
		d2lv300	**2.2380**	0.6147	3.6410	**0.0004**	***
		d2lv450	**1.6200**	0.6146	2.6360	**0.0098**	**
		d2lv600	**1.5390**	0.5507	2.7940	**0.0063**	**
		d2c500	**0.9699**	0.2356	4.1160	**0.0001**	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Bolded variables are significant.

April 1984 precipitation was evaluated using a spatial lag model with the same independent variables as those in the January 1984 model except for the interaction terms. A model without interaction terms was found to provide a better fit to the April 1984 precipitation amounts during initial model selection, indicating that the interaction terms did not improve model fit. All of the additive indicator variables for distance were significant and positive, indicating increased rainfall compared to the furthest distance category. Locations closest to Lake Victoria are expected to receive approximately 1.5 more than locations more than 600 km from Lake Victoria (p=0.0026) in terms of SRU; locations between 300 and 450 km from Lake Victoria are expected to receive 0.94 additional SRU (p=0.0506), and locations between 450 and 600 km should receive 1.04 SRU more than the furthest distance category (p=0.0376). These results indicate that distance from Lake Victoria plays a significant role in precipitation amounts even during the long rainy season when the ITCZ is overhead. Profile curvature at a scale of 9 km was marginally significant (p=0.0601), with a negative coefficient. This finding does not support the hypothesis that increasing precipitation would occur with increasing curvature for this particular month.

Elevation at a scale of 1 km was found to be a significant predictor of **August 1984** precipitation (p=0.0018), with a positive coefficient. This indicates that precipitation increases with increasing elevation at the estimated rate of 0.0006 SRU per meter, or 6 mm per kilometer. Significance of the elevation term in the short dry season confirms the hypothesis that local scale, orographic precipitation occurs during this season. Precipitation was also found to be approximately 3.4 SRU greater in locations between 300 and 450 km from Lake Victoria (p=0.0061) than the

most distant locations (greater than 600 km). The interaction term was also significant (p=0.0048) and negative, indicating that precipitation amounts decrease as distance from Lake Victoria increases within this distance band, as expected. Locations within 500 km of the Indian Ocean were expected to receive 0.47 SRU additional rainfall on average than more distant locations (p=0.0511).

In **November 1984**, the eastern component of the unit vector was a significant predictor of precipitation, indicating that east facing slopes were expected to receive 0.31 SRU less rainfall than west facing slopes (p=0.0278). The ITCZ was overhead during this time. Furthermore, generally higher precipitation was observed in the eastern portion of the study area. This would indicate that increased precipitation should occur on the east-facing slopes. However, the grid of the eastern component of the unit vector at a spatial resolution of 9 km appears highly variable and pixilated in this region. It is possible that some meteorological stations in this region are on west-facing slopes; however, this is not clear. In addition, higher rainfall amounts were shown to coincide with distances of less than 300 km from Lake Victoria (1.4 SRU, p=0.0091), between 450 and 600 km from Lake Victoria (1.3 SRU, p=0.0087), and less than 500 km from the Indian Ocean (0.7 SRU, 0.0049) compared to greater distances. Again, this illustrates that distance from surface water bodies has a significant impact on precipitation amounts during rainy seasons (in this case, the short rainy season).

Distances from Lake Victoria between 300 and 450 km were shown to receive 3.3 SRU more in **January 1985** than locations greater than 600 km, on average (p=0.0300). These amounts were shown to decrease as distance from Lake Victoria increased in this distance band (p=0.0303), as expected. Furthermore, north-facing

slopes were shown to receive more rainfall than south-facing slopes (p=0.0012). Inspection of the measured data for this month indicated that the greatest amount of precipitation fell in the northwest of the study area, on the north- and west-facing slopes of the highland mountains. Elevation at a spatial resolution of 1 kilometer was not a significant predictor of precipitation in the study area during this dry season (p=0.7828); it is possible that orographic precipitation occurred at scales other than that considered herein, or that other causal factors were involved.

Similar to April 1984, all three components of the distance to Lake Victoria measure were significant predictors of precipitation in **April 1985** (p=0.0117, p=0,0590, and 0.0227, respectively, for each distance band). All of the coefficients were positive, indicating that the distance bands closer to Lake Victoria received more rainfall than the furthest distance category (greater than 600 km), as expected. In addition, elevation at a scale of 1 km was a significant predictor of precipitation, with an additional 0.0005 SRU per meter (or 5 SRU per km) of elevation expected.

In **August 1985**, NDVI was identified as a significant predictor of precipitation (p=0.0283): higher NDVI values corresponded with increased rainfall, as expected. Elevation at a scale of 9 km was also significantly (p=0.0025) and positively related to precipitation, with an additional 0.0005 SRU of precipitation per meter of elevation expected. The northern component of the unit vector was also significant in August 1985 (p=0.0622) with a positive coefficient. This indicates that higher precipitation amounts were expected on north-facing slopes. Inspection of the measured precipitation data indicated that higher rainfall amounts were observed in the northwestern portion of the study area, similar to January 1985 in which the northern

component was also significant and positive. Furthermore, the ITCZ was in its northernmost position during this month.

All of the distance bands for Lake Victoria and the Indian Ocean were significant and positive in **November 1985**. Precipitation within 300 km of Lake Victoria was expected to be 2.2 SRU higher on average than precipitation more than 600 km from the lake (p=0.0004). Locations between 300 and 450 km from the lake were expected to receive 1.6 SRU more (p=0.0098), while locations between 450 and 600 km were expected to receive 1.5 SRU more than locations beyond 600 km from Lake Victoria (p=0.0063). Locations within 500 km of the Indian Ocean were expected to receive almost 1 SRU above more distant locations, on average (p=0.0001). NDVI was also a significant predictor of precipitation in November 1985 (p=0.0003): precipitation was shown to increase with increasing NDVI, as expected. Curvature at the 9 km scale was positively related to precipitation (p=0.0629), indicating that increased precipitation corresponded with upwardly convex surfaces (hilltops, representative of higher elevation), as expected.

In general, distance measures were the most common significant predictors of precipitation in the study area during both rainy and dry seasons. The significant influence of distance measures or general geographic locators (e.g., latitude and longitude) was also identified by Arora et al. (2006), Marquinez et al. (2003), Oettli and Camberlin (2005), KeifferWeisse and Bois (2001); Hutchinson and Bischof (1983), and Hutchinson (1998a, 1998b). In some cases, an interaction term between the indicator variables and the continuous distance measure was able to refine the predictions, allowing for decreasing precipitation as distance from Lake Victoria increased within a

distance band. It is interesting to note that this decreasing trend occurred only within the 300 to 450 km distance band, and only during dry seasons. The only exception is August 1985, in which no distance terms were significant.

NDVI was a significant predictor in four of the eight months considered (i.e., January 1984, April 1984, August 1985, and November 1985). The relationship was positive in all but one month. These results are difficult to interpret and are not well understood, since there was not a clear seasonal pattern to the significance of NDVI.

Elevation and its derivatives were significant only sporadically. Elevation at a 1 km resolution was significant in August 1984 and April 1985. Elevation at a 9 km resolution was significant in August 1985. However, significance of the an elevation term in the months of August (i.e., the second dry season) supports the hypothesis that orographic rainfall is a significant contributor to precipitation in this dry season. Oettli and Camberlin (2005) found similarly that mean elevation was not highly correlated to precipitation amounts except, perhaps, in relatively smaller study areas in East Africa due to the “considerable inconsistency in the elevation–rainfall relationship, which only applies to either very small or very large space scales.”

The eastern component of the unit vector was significant in November 1984 only, whereas the northern component was significant in both January and August 1985. Significance of this term in January and August 1985 can be understood based on inspection of the observed precipitation data: higher rainfall amounts were observed in the northern/northwestern portions of the study area in these months. Furthermore, the ITCZ was in its northernmost position in August 1985.

Curvature was significant but negative in April 1984, which was not expected. However, curvature was significant and positive in November 1985, indicating that higher precipitation amounts occurred at hilltops (or local areas of higher elevation), as expected.

With regard to model type, OLS models were adequate for the modeling of precipitation in January 1984, August 1984, and November 1985. Spatial lag models provided better fits to precipitation data in the remaining months. No obvious seasonality was evident in model selection based on these findings.

Comparison of results for 1984 and 1985 indicates similarities in the significance of distance variables: increased precipitation was observed in the 300 to 450 km distance band in January 1984 and 1985 (the first dry season), with decreasing precipitation in this band as distance from Lake Victoria increases. This pattern was also observed in August 1984 (the second dry season), although it was not detected in August 1985. In addition, nearly all of the distance categories were significant in all rainy seasons (i.e., April 1984, August 1984, April 1985, and August 1985). This pattern indicates that significantly increased precipitation is most likely observed in the range of 300 to 450 km from Lake Victoria during the dry seasons. Significance of all distance terms in the rainy seasons indicates that rainfall is generally widespread at this time, with the exception of the most distant locations from water bodies (e.g., greater than 600 km from Lake Victoria.

Elevation terms were significant predictors of variability of rainfall over space in both August 1984 and August 1985, the second dry season of each year, supporting the

likelihood of the occurrence of orographic precipitation during these dry seasons. However, elevation was not significant in either January 1984 or January 1985. Precipitation in January 1984 was lower than the long-term average for this month. January 1985 was also slightly lower than the long term average; this could explain reduced precipitation of all forms, including orographic, in these months.

Observed similarities between significant predictors of 1984 and 1985 monthly precipitation over space would suggest that causal factors do not vary significantly between years that are typical and atypical in terms of monthly precipitation amounts. However, additional study is required to test this hypothesis.

5.3 Hausman test for simultaneity

Formal testing for simultaneity between monthly precipitation and the 1 month-lagged vegetation measure (NDVI) was conducted for each month in 1985, the year selected to represent a typical precipitation year in East Africa.

Formal testing was completed using the Hausman test (Hausman 1978). The Hausman test compares two regression models, a model that incorporates potential simultaneity through the use of an instrumental variable and the same model without the instrumental variable, to test the following hypothesis:

Hypothesis 6

$H6_0$: Simultaneity does not exist between monthly precipitation and vegetation (represented by a 1-month lagged NDVI).

$H6_1$: Simultaneity does exist between monthly precipitation and vegetation.

Essentially, the Hausman identifies whether a regression model formulated under the alternative hypothesis is significantly better than the regression model under the null hypothesis. If so, then significant simultaneity is concluded to be present.

Formal testing of the above hypothesis failed to identify the presence of significant simultaneity for any month in 1985. This result may be due, in part, to a limitation in available data and an over-simplification of the lag between increased precipitation and increased NDVI. A one-month lag was used to model the relationship between precipitation and NDVI based on citations in the literature. However, due to the highly variable landscapes and land cover types of East Africa, it is likely that the lag between precipitation and vegetation varies over space according to vegetation type and other factors. Improved results could be obtained by allowing for a variable lag time according to land cover type.

Chapter 6

Conclusions and Future Research

In the context of understanding the impacts of global climate change and the development of effective adaptation strategies, the need for higher resolution climate data is clear. Although the relatively coarse-scale data simulated by global models can be useful for evaluating global climate trends, they are not sufficient for regional evaluations of climatic patterns, particularly in the presence of high landscape variability such as that which occurs in eastern Africa. This concern has been identified by individual researchers as well as Intergovernmental Panel on Climate Change (Boko et al. 2007), who state that finer-scale climate data are necessary for decision-makers. Furthermore, development of successful adaptation strategies requires higher-resolution climate data and the combined efforts of localized, community-based efforts (Brooks et al. 2006, Thornton et al. 2009).

This dissertation is devoted to gaining a more complete understanding of existing techniques for developing finer-scale data by theoretically mapping similarities and differences between two statistical paradigms: kriging and spatial regression; developing improvements to those methods by expanding the capabilities of universal kriging to incorporate feedback simultaneity that may occur between the dependent variable of interest, rainfall in this study, and a variable such as vegetation; and applying these techniques to a case study set in East Africa in hopes that it will be useful for improved understanding of climate change and its impacts, and for further development of adaptation strategies to climate change in East Africa.

6.1 Findings and conclusions

Precipitation maps were plotted using ordinary kriging (LOK), universal kriging (UK), universal kriging with an instrumental variable (UKIV) and a selected spatial regression model for each month evaluated as part of this dissertation. The maps primarily illustrated the effects of elevation, with maximum values of monthly precipitation generally occurring on or near Mts. Kenya and Kilimanjaro, and distance from Lake Victoria, particularly in the dry seasons. In the dry seasons represented by January and August, the precipitation maps illustrated that the highest levels of precipitation occur on the western edge of the study area, nearest Lake Victoria, and in areas of high elevation (i.e., Mount Kenya and Mount Kilimanjaro). These spatial patterns in precipitation are consistent with the expectation that rains occurring in the dry seasons are mainly localized convective rainfall near Lake Victoria or stratiform rainfall in the highland areas (Ng'ang'a 1992). Precipitation maps for the months representing the wet seasons, April and November, higher levels of precipitation were observed to occur more uniformly over the study area. This is expected since the ITCZ is directly over the study area during these months, influencing precipitation on a large scale.

A comparison of map accuracies indicated that root mean squared errors (RMSEs) were consistently lower for LOK, with the exception of November 1985 in which the UK estimate had the lowest RMSE. However, there were many close runners up which included one or more of the other model types (UK, UKIV, and/or spatial regression) depending on the month. Maps of standardized error terms were also reviewed for overall number of significant error terms and spatial patterns in the error

terms. Clusters of standardized residuals were common in the west Kenya highlands and in the vicinity of Mt. Kenya. However, the clusters in the northwest often included high and low values, indicating the presence of highly variable precipitation amounts over space.

In April 1985, the standard error map for UKIV (the newly derived method) indicated that the number and configuration of significant standard errors were better than that for LOK. This confirms that UKIV can perform similarly to other standard methods of estimation, and perhaps better in the presence of significant measurable simultaneity.

Two prominent approaches to generating higher resolution climate data were identified: downscaling modeled values from generalized circulation models (GCMs) and statistical techniques that utilize measured climate data. Although the former is common, due to the availability and completeness of modeled data sets, the latter has been shown to provide more accurate regional predictions in East Africa. Data from the CRU TS 3.1 global data set were mapped in the East African study area and compared to the maps described above. The CRU TS 3.1 data were found to substantially underestimate precipitation amounts in the region. Furthermore, large contiguous areas of significantly low estimates illustrated spatial patterns in the error terms of the CRU data set, indicating its shortcomings in reflecting local geographic features.

Hypothesis testing was conducted to better understand the system of precipitation in the study region by identifying significant explanatory variables for describing the spatial patterns in precipitation. Spatially explicit regression models were used where diagnostic testing indicated that OLS models were inappropriate. In

general, distance measures were the most common significant predictors of precipitation in the study area during both rainy and dry seasons. In some cases, an interaction term between the indicator variables and the continuous distance measure was able to refine the predictions, allowing for decreasing precipitation as distance from Lake Victoria increased within a distance band; this decreasing trend occurred only within the 300 to 450 km distance band, and only during dry seasons, however. OLS models were adequate for the modeling of precipitation in January 1984, August 1984, and November 1985. Spatial lag models provided better fits to precipitation data in the remaining months. Consequently, no obvious seasonality was evident in model selection based on these findings.

Comparison of results for 1984 (an atypical year with drought conditions in the early part of the year) and 1985 (a typical year) indicates similarities in the significance of distance variables and elevation, in particular. Observed similarities between significant predictors of 1984 and 1985 monthly precipitation suggest that causal factors do not vary significantly between atypical and typical years in terms of monthly precipitation amounts. However, additional study is required to test this hypothesis.

6.2 Gaps in the literature and contributions to address them

A gap exists that is not completely filled by the variety of statistical techniques historically used to evaluate climate data. The focus of the many spatially explicit precipitation studies has often been to develop the ‘best’ estimates of precipitation, not to understand the rainfall system in the region under study. To better understand systems of precipitation, aspatial ordinary least squares regression techniques have

often been used, although results are biased when spatial patterns in the data are not considered. Spatial regression consists of several regression techniques that are not only spatially explicit, but can also be used to improve understanding of variables that influence precipitation patterns. Spatially explicit regression techniques have begun to appear in climate-related literature; however, each of the applications is somewhat limited, either by the method used, the covariates selected, or the overall application (hypothesis testing or estimation). Further, none of the approaches have characterized endogenous relations, such as those between precipitation and vegetation. This dissertation addresses this gap by 1) including spatial regression as a potential technique for estimation of precipitation at unsampled locations, 2) developing an extended version of universal kriging that explicitly represents and incorporates endogenous relations between vegetation and precipitation into estimations at unsampled locations, and 3) conducts hypothesis testing of significant factors in explaining precipitation patterns in a spatially explicit manner.

6.2.1 Theoretical and methodological contributions

This dissertation has provides a theoretical summary of kriging and spatial regression techniques, ultimately demonstrating that one set of equations can be used to represent both kriging and spatial regression methods. The differences between the set of equations for universal kriging, the SEM spatial regression model, and the SAR spatial regression model are summarized in Table 3-2; differences between these approaches are restricted to the spatial weights matrices used to derive the regression coefficients, and the spatial weights applied to the model residuals. Thus, this

dissertation has successfully identified the similarities between kriging approaches and spatial regression techniques. Furthermore, the differences between these approaches, contained within the spatial weights matrices, have been specified.

The theoretical portion of this dissertation also included the development of a modified universal kriging algorithm that accounts for complex feedbacks between precipitation and a covariate, in this case vegetation, by incorporating the simultaneity between these variables. This was accomplished through an instrumental variables approach. First, an instrumental variable was created by regressing vegetation (i.e., NDVI) against a subset of the variables included in the precipitation model, plus one variable that was correlated with NDVI but not precipitation. This variable was soil pH. Predicted values for NDVI were generated then substituted for the observed NDVI data in the universal kriging model used to predict precipitation. Formal testing for simultaneity between precipitation and vegetation did not confirm the occurrence of simultaneity feedback; however, the new method of universal kriging with instrumental variables was retained in the case study to further develop the application of this method.

The new and innovative technique for spatial analysis of data developed in this dissertation can be applied in any context when the goal is to understand spatial patterns in any continuous data, to predict data at unsampled locations, and/or to understand the relationship between a variable of interest and other factors that may influence the spatial distribution of that variable. This technique lends itself to evaluation of many types of data, including data used in ecological evaluations, environmental contamination studies, deforestation studies, geological and mining

studies, and econometric studies. Furthermore, these techniques are applicable to evaluating data collected anywhere in the world, potentially improving the ability to study regional patterns using these multivariate techniques.

6.2.2 Contributions to Regional Geography of Africa and Geography as a whole

The results and findings of this dissertation add to the knowledge and understanding of the physical mechanisms of precipitation in East Africa as well as spatial patterns of precipitation within the study area. In addition, once more recent precipitation data are available, updated and improved maps of spatial patterns in precipitation within the study domain can be generated, contributing to Physical Geography and Regional Geography of Africa.

Results of this dissertation will contribute to work conducted by Human-Environment researchers such as high-resolution climate impact assessments (Mearns et al.,1999, 2001a, 2001b) which may be used in the evaluation of coupled human natural systems and formulation of associated policy.

6.3 Limitations

This dissertation has achieved many of the goals established at the outset; however, looking to the future, improvements can be made. This section describes some of the limitations of the work presented herein and suggestions for future improvements.

6.3.1 Limitations in data and modeling

As noted in Chapter 4, meteorological stations are generally located in areas of higher population, and do not fully represent the range of elevations in the study area. The representativeness of station elevations and the impact that this has on the representation of precipitation in the region is considered here. Figure 6-1a summarizes elevation for the entire study area; for comparison Figure 6-1b summarizes elevation for the meteorological stations only. The maximum elevation of a meteorological station of 2,773 m falls well below the maximum elevation for the study area of 5,778 m; the minimum elevation for a meteorological station of 454 m is somewhat closer to the minimum elevation in the region of 195 m. Inspection of the histograms also indicates that the locations of the meteorological stations over-represent elevations in the range of roughly 1,300 m to 2,500 m. The most notable under-representation occurs in the elevations below 1,300 m; more than 50% of the study area falls below 1,300 m. The statistical approaches considered herein are based on an assumption that the data used in the analysis are representative of the variable of interested (i.e., precipitation). To accurately represent precipitation in the study area, some form of randomization would be necessary in the placement of meteorological stations. This is clearly not the case, resulting in a bias in the results of this dissertation. While it is unlikely that a randomized and fully representative distribution of meteorological stations will be available in the near future, solutions to this problem can be addressed by better stratifying the study area based on elevation, or possible through the use of declustering techniques.

A one-month lag was used to model the relationship between precipitation and NDVI based on citations in the literature. It is likely that the lag between precipitation and vegetation varies over space according to vegetation type and other factors. A surface of spatially varying lag times could be developed in which the lag time depends on factors such as land cover type; however, limitations in the temporal resolution of vegetation data would have to be considered. For example, NDVI data are available in biweekly time steps only from the Global Inventory Modeling and Mapping Studies (GIMMS).

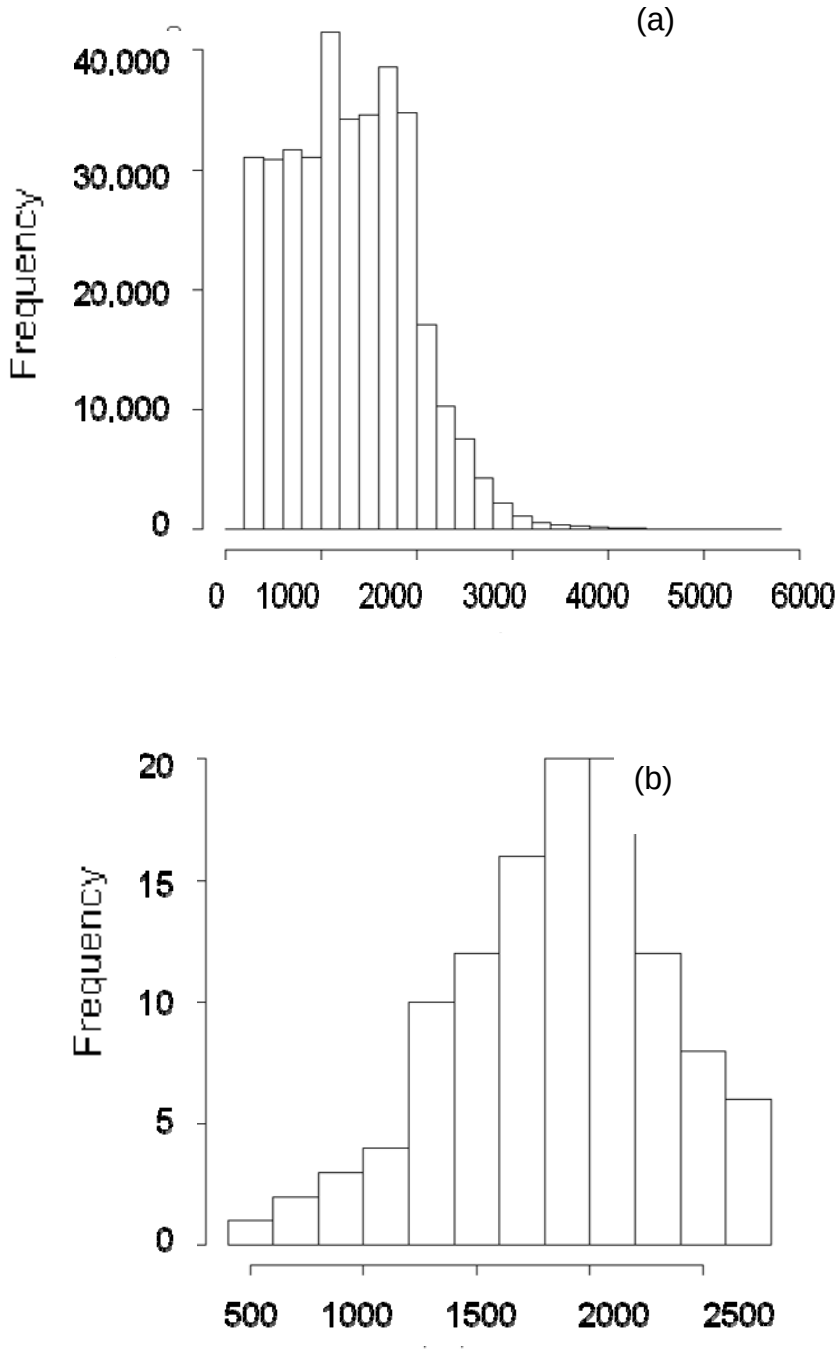

Figure 6-1. Comparison of elevations throughout the study area (a). Histogram of elevations at meteorological station locations (b). Elevations were obtained from the SRTM 30 arc second (approximately equivalent to 1 km at the equator) Digital Elevation Model.

Elevation measurements were not available for every meteorological station. Consequently, elevation was estimated for each station using a 1 km DEM. This increases the uncertainty in the effect of elevation on precipitation. In addition, elevation was considered at spatial resolutions of 1 km and 9 km based on previous work cited in the literature. While this scale may be the best for representation of localized effects on precipitation due to the presence of large water bodies and variable topography, it is my opinion that elevation should be considered in other scales or represented using other factors. For example, variables such as aspect and curvature likely impact precipitation patterns at scales much larger than 9 km (e.g., when identifying the windward and leeward side of a mountain). Also, the distance bands used in categorizing distance from Lake Victoria were highly correlated with a categorical elevation variable that was initially developed and not used due to multicollinearity. This suggests that the distance bands were acting as a proxy for average elevation. This observation is supported by the fact that the first distance band generally lines up with the West Kenya Highlands, the second distance band generally corresponds to the Rift Valley, and the third corresponds to higher elevations to the east including Mounts Kenya and Kilimanjaro.

6.3.2 Limitations in statistical methodology

Use of the spatial regression techniques to estimate precipitation at unsampled locations is currently limited: when predicting at unsampled locations, only the trend components were included**.** The signal (i.e., the spatial smoothing) component is not estimated, even though each of the $\boldsymbol{\beta}$ terms incorporate spatial autocorrelation terms.

Potential estimators for the signal component have been suggested, but not yet evaluated for performance. This limitation results in overly smoothed predictions using this method.

6.4 Future research

Future work is needed to evaluate the effect of modifications to the methods used in this dissertation. For example, rain station locations were generally clustered in the vicinity of higher population areas; consequently, the observed data may not have adequately represented less populated (or unpopulated) areas at higher elevations or in the savanna regions. Use of stratification or declustering techniques may be useful to provide a better representation of precipitation over space.

Analysis of more recent data is also important for generating current finer-scale precipitation estimates, and for identifying changes in significant predictive variables through hypothesis testing. Consequently, efforts should be made to obtain more recent, spatially complete data from meteorological stations in East Africa.

Additional work can be done to improve the results of the UKIV model. Incorporating a spatially varying lag term may help in identifying any feedback simultaneity that occurs between precipitation and vegetation. Improved regression results through consideration of additional variables for use in predicting patterns in NDVI (i.e., in generating the instrumented variable) may improve overall predictive ability of the UKIV method. Further, additional research in the mechanisms of precipitation may identify additional variables for inclusion in regression models, potentially providing improved results.

Lastly, additional research is necessary to expand the methods presented herein to evaluate trends in both space and time.

The statistical methods used in this dissertation are broadly applicable to many types of data, including environmental, economic, epidemiological, among others. The ability to incorporate simultaneity between multiple endogenous variables when creating predicted surfaces of key variables will be of practical benefit. For example, many contaminants of concern in environmental investigations are be related endogenously (e.g., concentrations of various metals vary together naturally or due to similar mechanisms of environmental release). In some cases, however, one or more variables may be more costly to measure, such as dioxin concentrations in soil or blood levels. The ability to utilize information on this endogeneity may reduce sampling and analytical costs, improve study accuracy, and lead to more efficient and effective environmental remediation.

In the context of climate studies, the methods used herein are broadly applicable to other regions of the world. Understanding and predicting spatial patterns in East African precipitation is highly complex due to changing seasonal patterns and widely varying topography in this region of the world. The univariate statistical method of LOK proved to be most effective for mapping precipitation patterns in East Africa due to the complexity of causal factors and their influence on precipitation patterns in the region. However, multivariate, spatially explicit statistical methods are expected to prove more effective in other regions of the world with less variable seasonality and topography over space, such as the North America plain states, or larger, more spatially contiguous moisture sources, such as the Congo Basin in Central Africa. Furthermore, the

combined approaches of spatially explicit regression modeling and interpolation methods yield not only improved prediction surfaces, but also better understanding of the factors that influence precipitation patterns over space.

`

Appendix

APPENDIX Case Study Results

Table A-1. Summary of regression modeling results.

Month	Method	Independent Variable	Estimate	St. Error	t value	Pr(>|t|)	
Jan-84	OLS	(Intercept)	-1.3530	0.6059	-2.2330	0.0278	*
		ndvi842	**-1.5500**	0.6857	-2.2610	**0.0259**	*
		dem9km	0.0003	0.0003	1.1930	0.2358	
		p9km	-0.2401	0.1573	-1.5260	0.1302	
		q9km	0.0468	0.1590	0.2950	0.7690	
		curv9km	-50.6300	93.3400	-0.5420	0.5887	
		d2lv300	**10.2200**	5.4080	1.8890	**0.0618**	.
		d2lv450	**3.7850**	1.5980	2.3690	**0.0198**	*
		d2lv600	0.6514	0.5481	1.1890	0.2375	
		d2c500	**1.1610**	0.2960	3.9210	**0.0002**	***
		I(d2lv300 * dist2lv)	-0.0271	0.0194	-1.3990	0.1649	
		I(d2lv450 * dist2lv)	**-0.0081**	0.0035	-2.3260	**0.0221**	*
Apr-84	SARlag	(Intercept)	-1.9160	0.4924	-3.8910	0.0001	***
		ndvi845	**0.7846**	0.4462	1.7584	**0.0787**	**.**
		dem9km	0.0001	0.0002	0.5272	0.5981	
		p9km	-0.0479	0.1241	-0.3864	0.6992	
		q9km	0.2024	0.1256	1.6123	0.1069	
		curv9km	**-136.4866**	72.5866	-1.8803	**0.0601**	**.**
		d2lv300	**1.4931**	0.4961	3.0094	**0.0026**	**
		d2lv450	**0.9401**	0.4808	1.9552	**0.0506**	**.**
		d2lv600	**1.0393**	0.4441	2.3404	**0.0193**	*
		d2c500	**0.3932**	0.1891	2.0787	**0.0376**	*
		Rho	**0.6040**	0.1017	5.9358	**0.0000**	***
Aug-84	OLS	(Intercept)	-2.1604	0.4720	-4.5770	0.0000	***
		ndvi849	0.7530	0.5121	1.4710	0.1446	
		dem1km	**0.0006**	0.0002	3.2120	**0.0018**	**
		p1km	-0.0632	0.1242	-0.5090	0.6119	
		q1km	0.1014	0.1205	0.8420	0.4020	
		curv1km	-3.9075	6.7899	-0.5750	0.5663	
		d2lv300	-1.5345	4.1207	-0.3720	0.7104	
		d2lv450	**3.3980**	1.2118	2.8040	**0.0061**	**
		d2lv600	-0.1001	0.4035	-0.2480	0.8046	
		d2c500	**0.4681**	0.2371	1.9740	**0.0511**	**.**
		I(d2lv300 * dist2lv)	0.0116	0.0148	0.7880	0.4327	
		I(d2lv450 * dist2lv)	**-0.0076**	0.0026	-2.8860	**0.0048**	**
Nov-84	SARlag	(Intercept)	-1.4962	0.6446	-2.3211	0.0203	*
		ndvi8412	-0.1582	0.6878	-0.2300	0.8181	
		dem9km	0.0001	0.0002	0.4865	0.6266	
		p9km	**-0.3114**	0.1415	-2.2008	**0.0278**	*
		q9km	-0.0247	0.1390	-0.1776	0.8591	
		curv9km	66.7779	81.0492	0.8239	0.4100	
		d2lv300	**1.4101**	0.5406	2.6082	**0.0091**	**
		d2lv450	0.7143	0.5338	1.3382	0.1808	
		d2lv600	**1.2727**	0.4849	2.6246	**0.0087**	**
		d2c500	**0.7044**	0.2504	2.8133	**0.0049**	**
		Rho	0.2500	0.1629	1.5348	0.1248	

Table A-1. Summary of regression modeling results (Continued).

Month	Method	Independent Variable	Estimate	St. Error	t value	Pr(>|t|)	
Jan-85	SARlag	(Intercept)	-0.7867	0.6318	-1.2450	0.2131	
		ndvi852	0.5605	0.5799	0.9665	0.3338	
		dem1km	-0.0002	0.0002	-0.7656	0.4439	
		p1km	0.0418	0.1515	0.2757	0.7828	
		q1km	**0.4667**	0.1441	3.2398	**0.0012**	**
		curv1km	12.2468	8.7368	1.4018	0.1610	
		d2lv300	-4.8020	4.5302	-1.0600	0.2891	
		d2lv450	**3.3218**	1.5306	2.1702	**0.0300**	*
		d2lv600	0.2569	0.5711	0.4498	0.6529	
		d2c500	0.3788	0.2664	1.4220	0.1550	
		I(d2lv300 * dist2lv)	0.0218	0.0161	1.3515	0.1765	
		I(d2lv450 * dist2lv)	**-0.0072**	0.0033	-2.1659	**0.0303**	*
		Rho	**0.2970**	0.1613	1.8406	**0.0657**	**.**
Apr-85	SARlag	(Intercept)	**-2.0913**	0.6588	-3.1744	**0.0015**	**
		ndvi855	-0.2445	0.7459	-0.3277	0.7431	
		dem1km	**0.0005**	0.0002	2.5010	**0.0124**	*
		p1km	0.1716	0.1474	1.1637	0.2446	
		q1km	0.2458	0.1510	1.6278	0.1036	
		curv1km	-6.1177	8.3879	-0.7294	0.4658	
		d2lv300	**1.5083**	0.5984	2.5206	**0.0117**	*
		d2lv450	**1.0736**	0.5685	1.8883	**0.0590**	**.**
		d2lv600	**1.1885**	0.5218	2.2777	**0.0227**	*
		d2c500	0.2507	0.2290	1.0947	0.2736	
		Rho	**0.6349**	0.1050	6.0476	**0.0000**	***
Aug-85	SARlag	(Intercept)	-1.0773	0.4340	-2.4820	**0.0131**	*
		ndvi859	**0.9904**	0.4516	2.1929	**0.0283**	*
		dem9km	**0.0005**	0.0002	3.0242	**0.0025**	**
		p9km	0.0794	0.1078	0.7373	0.4610	
		q9km	**0.2018**	0.1082	1.8647	**0.0622**	**.**
		curv9km	-9.1431	64.7750	-0.1412	0.8878	
		d2lv300	0.1276	0.4247	0.3005	0.7638	
		d2lv450	-0.3907	0.3993	-0.9787	0.3277	
		d2lv600	-0.4021	0.3687	-1.0904	0.2755	
		d2c500	-0.0420	0.2040	-0.2057	0.8370	
		Rho	**0.4170**	0.1329	3.1382	**0.0017**	**
Nov-85	OLS	(Intercept)	-2.7490	0.6493	-4.2330	**0.0001**	***
		ndvi8512	**2.2120**	0.5926	3.7330	**0.0003**	***
		dem9km	-0.0004	0.0002	-1.4950	0.1383	
		p9km	-0.0934	0.1640	-0.5690	0.5705	
		q9km	-0.0210	0.1629	-0.1290	0.8976	
		curv9km	**177.2000**	94.1500	1.8820	**0.0629**	**.**
		d2lv300	**2.2380**	0.6147	3.6410	**0.0004**	***
		d2lv450	**1.6200**	0.6146	2.6360	**0.0098**	**
		d2lv600	**1.5390**	0.5507	2.7940	**0.0063**	**
		d2c500	**0.9699**	0.2356	4.1160	**0.0001**	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Bolded variables are significant.

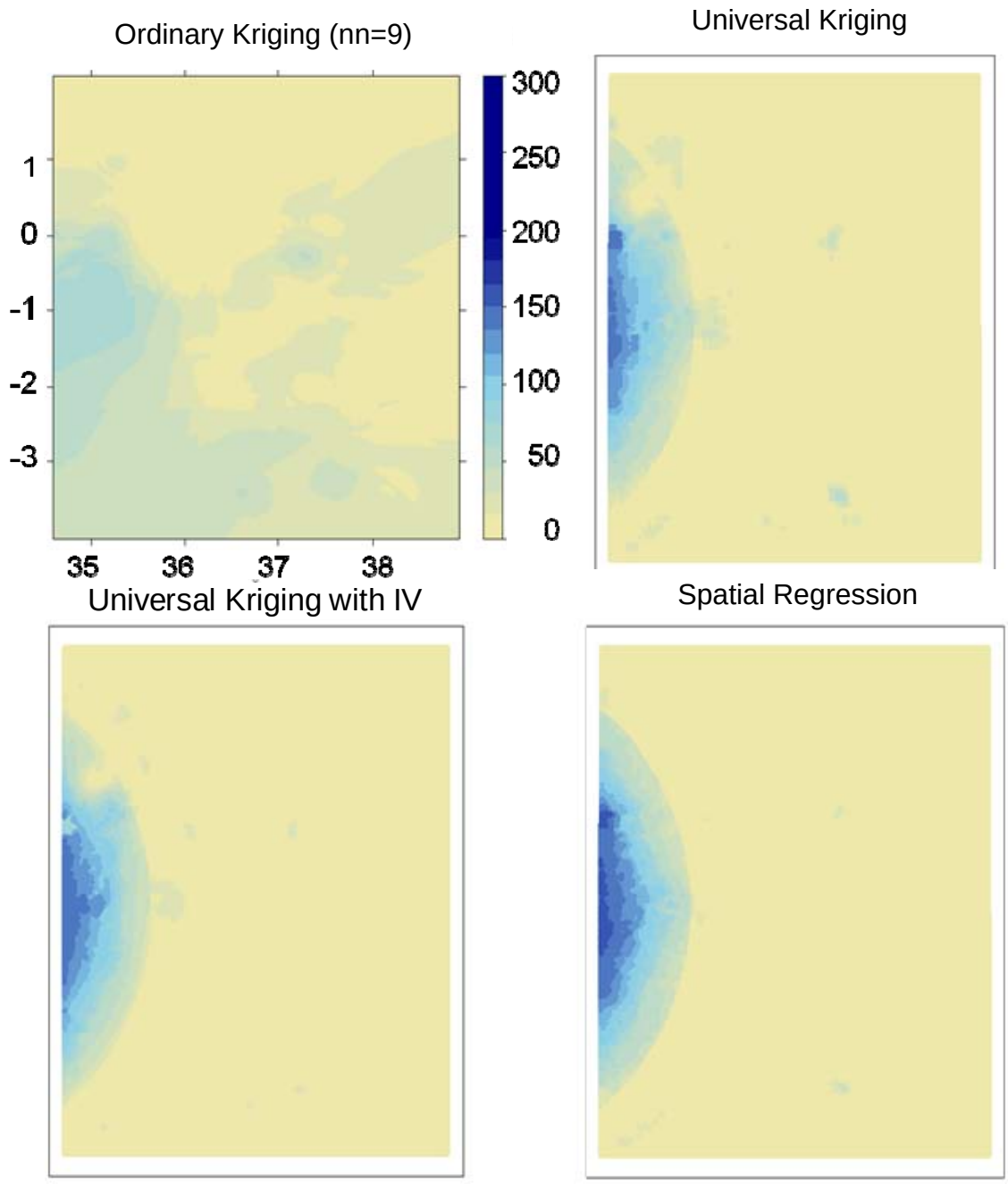

Figure A-1. January 1984 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

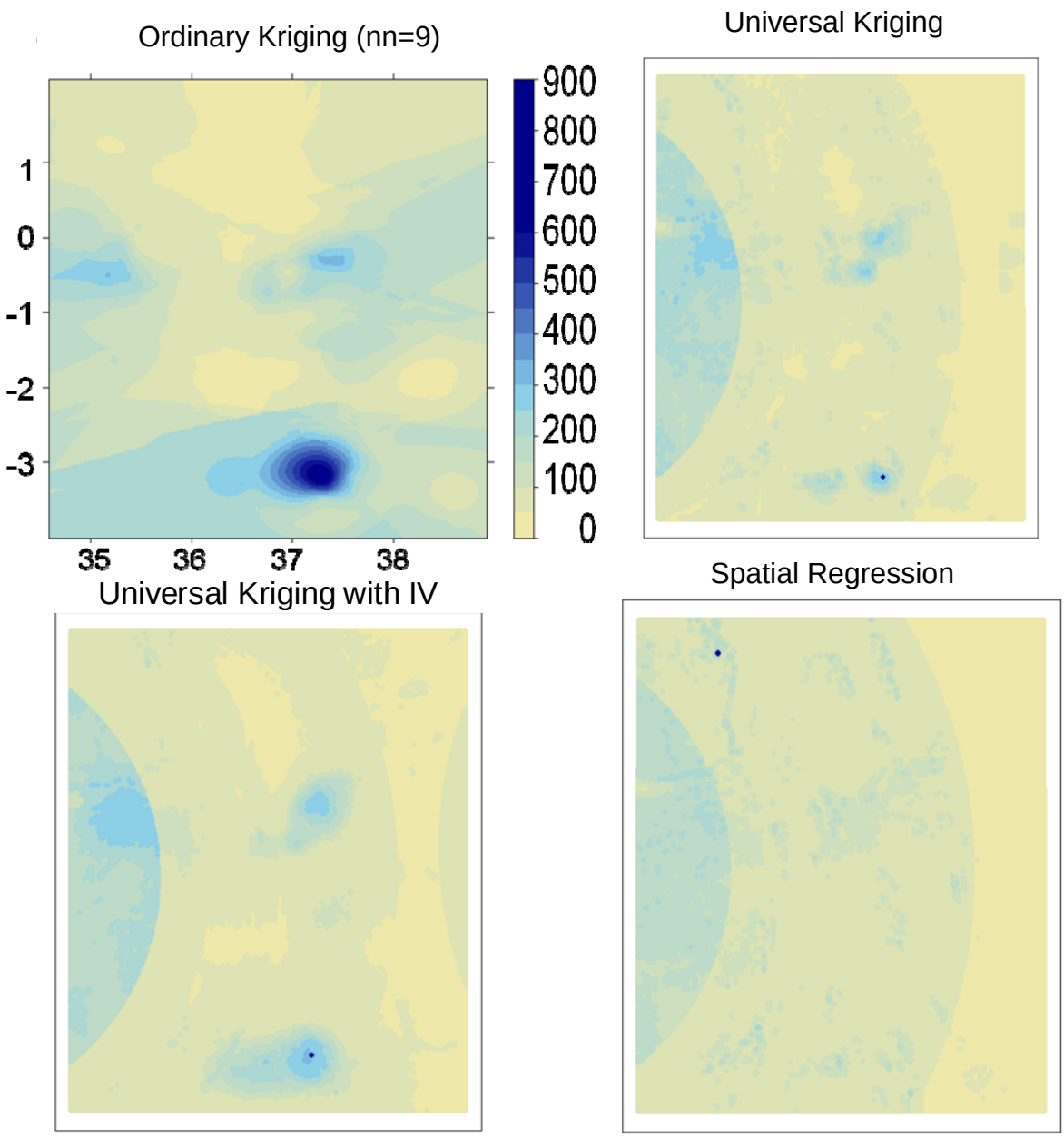

Figure A-2. April 1984 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

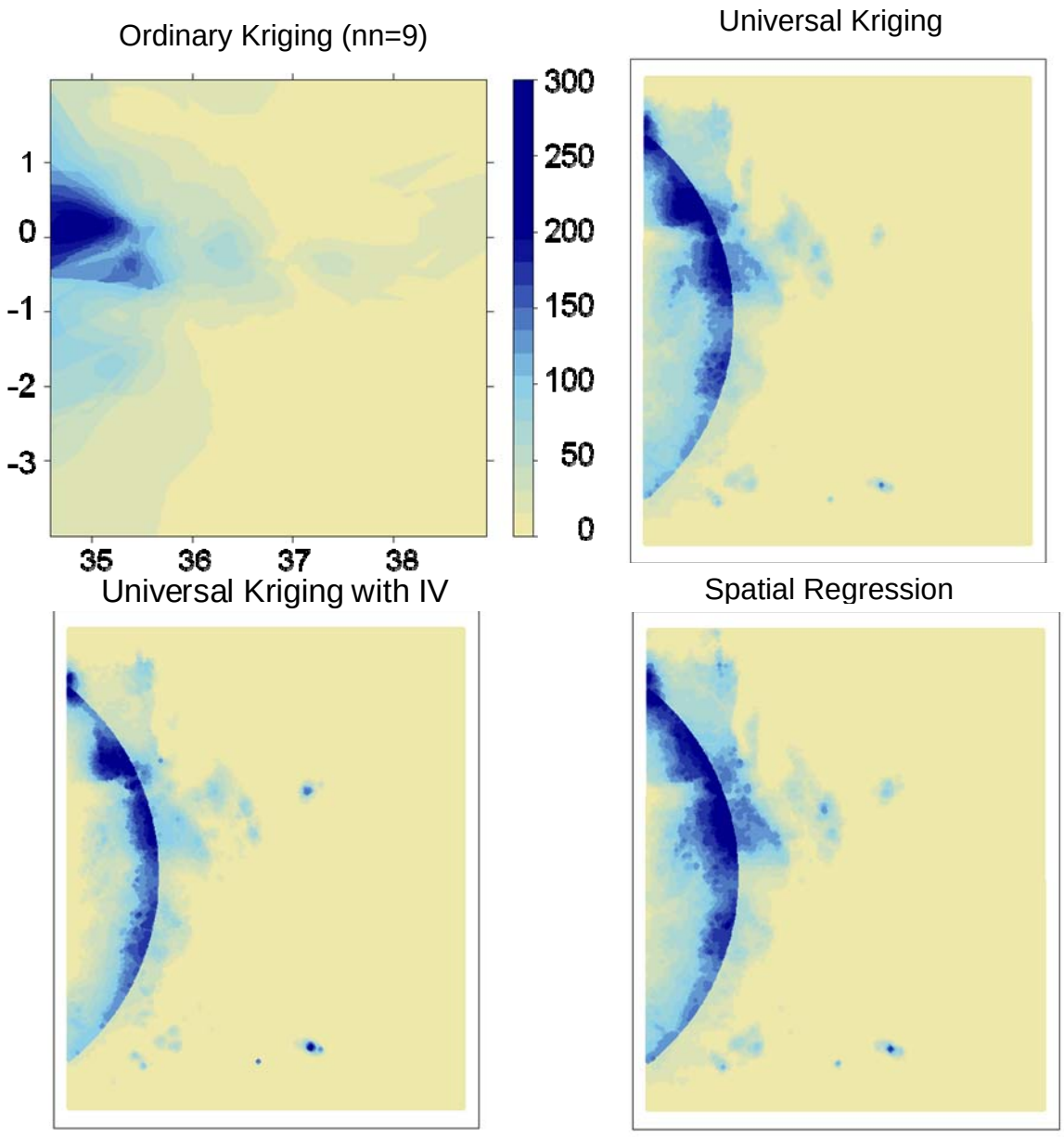

Figure A-3. August 1984 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

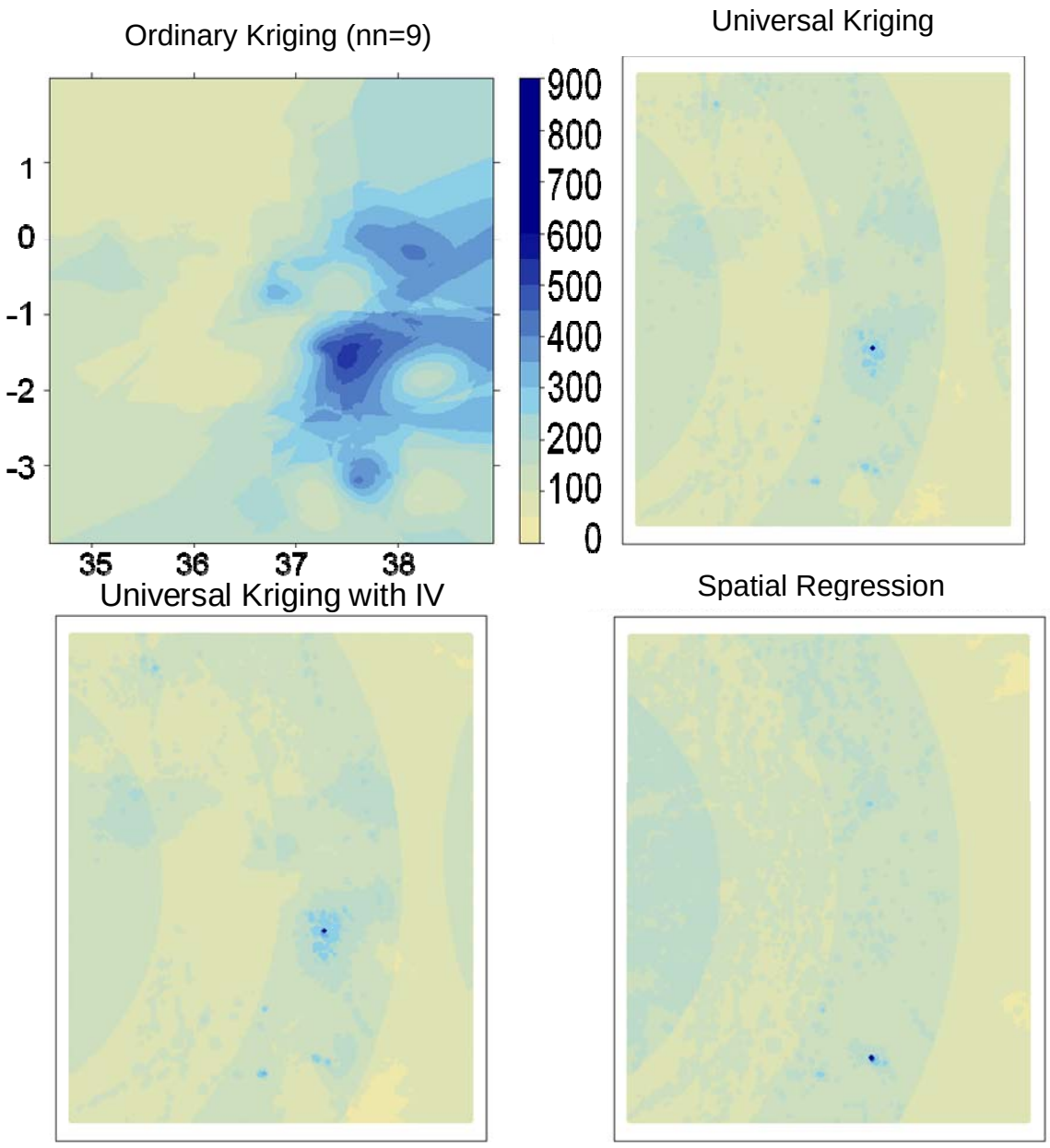

Figure A-4. November 1984 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

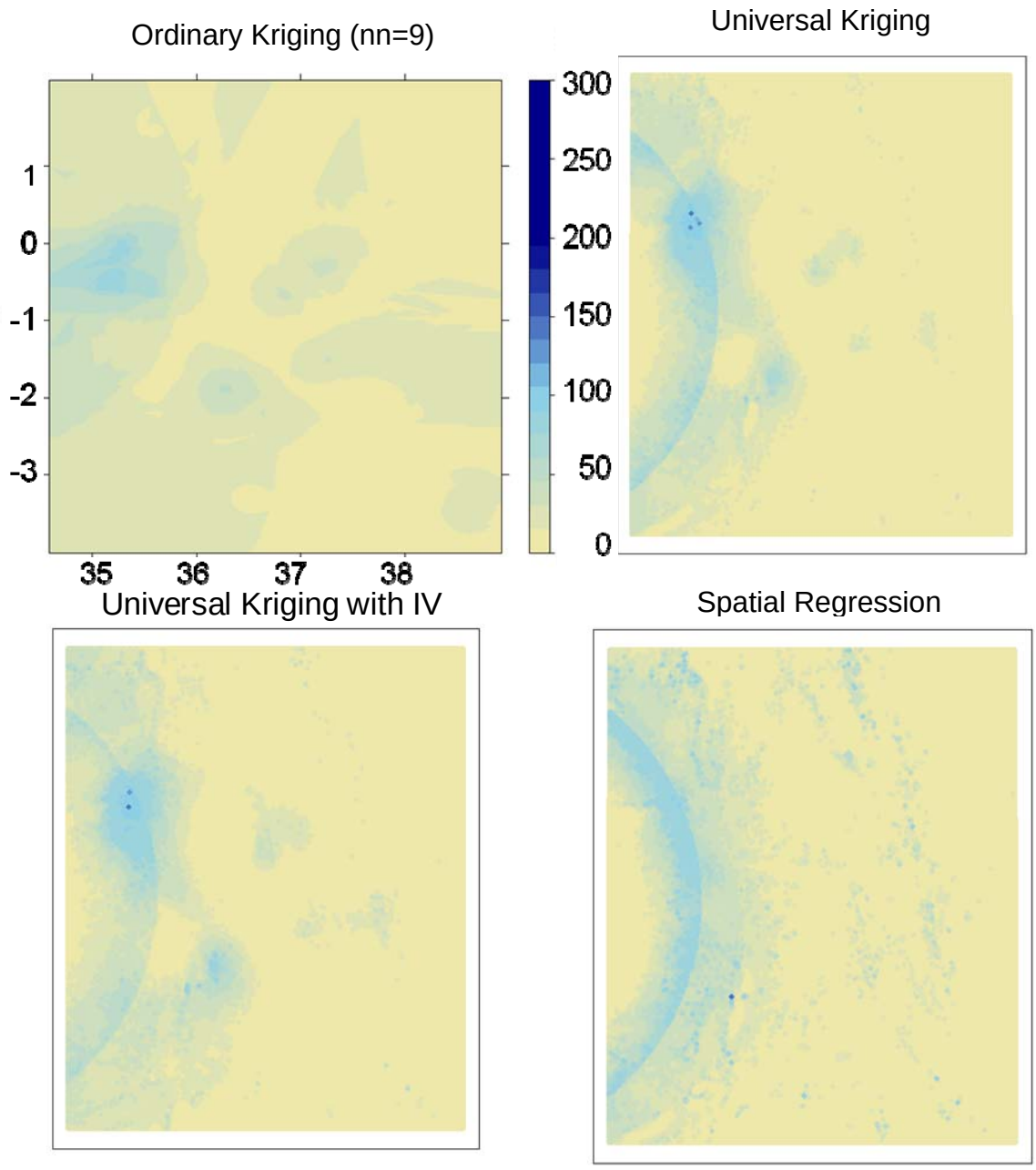

Figure A-5. January 1985 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

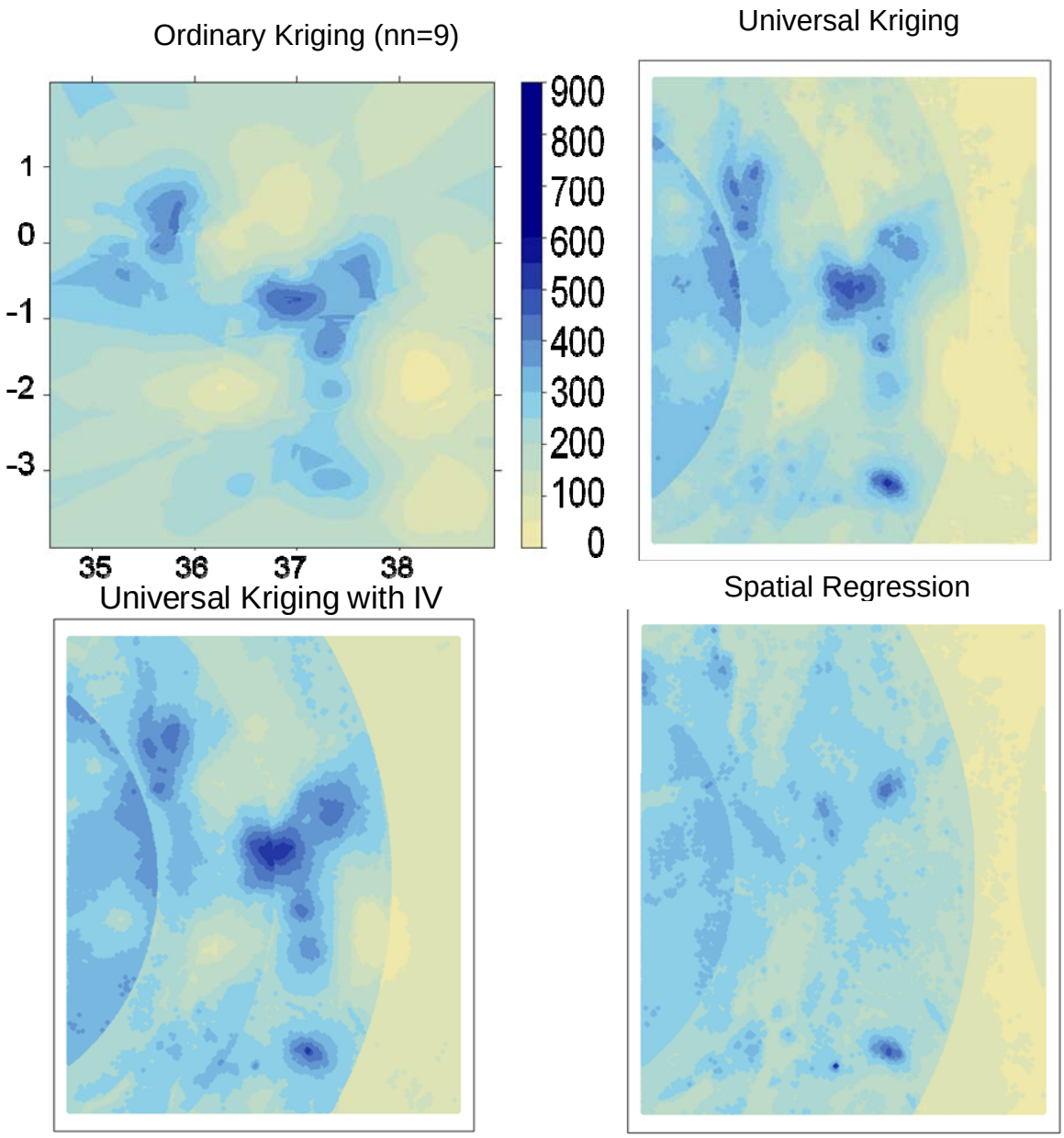

Figure A-6. April 1985 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

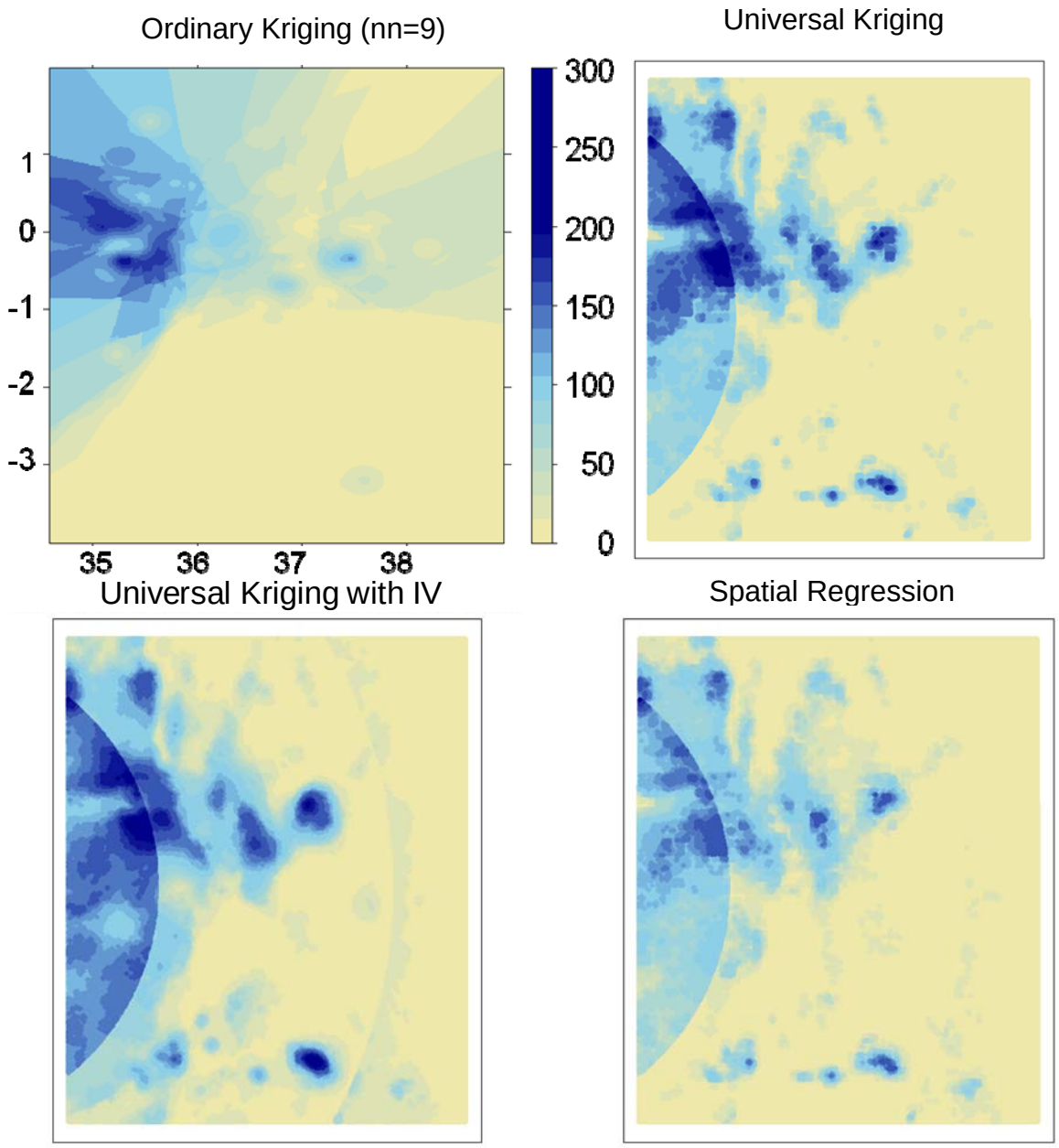

Figure A-7. August 1985 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

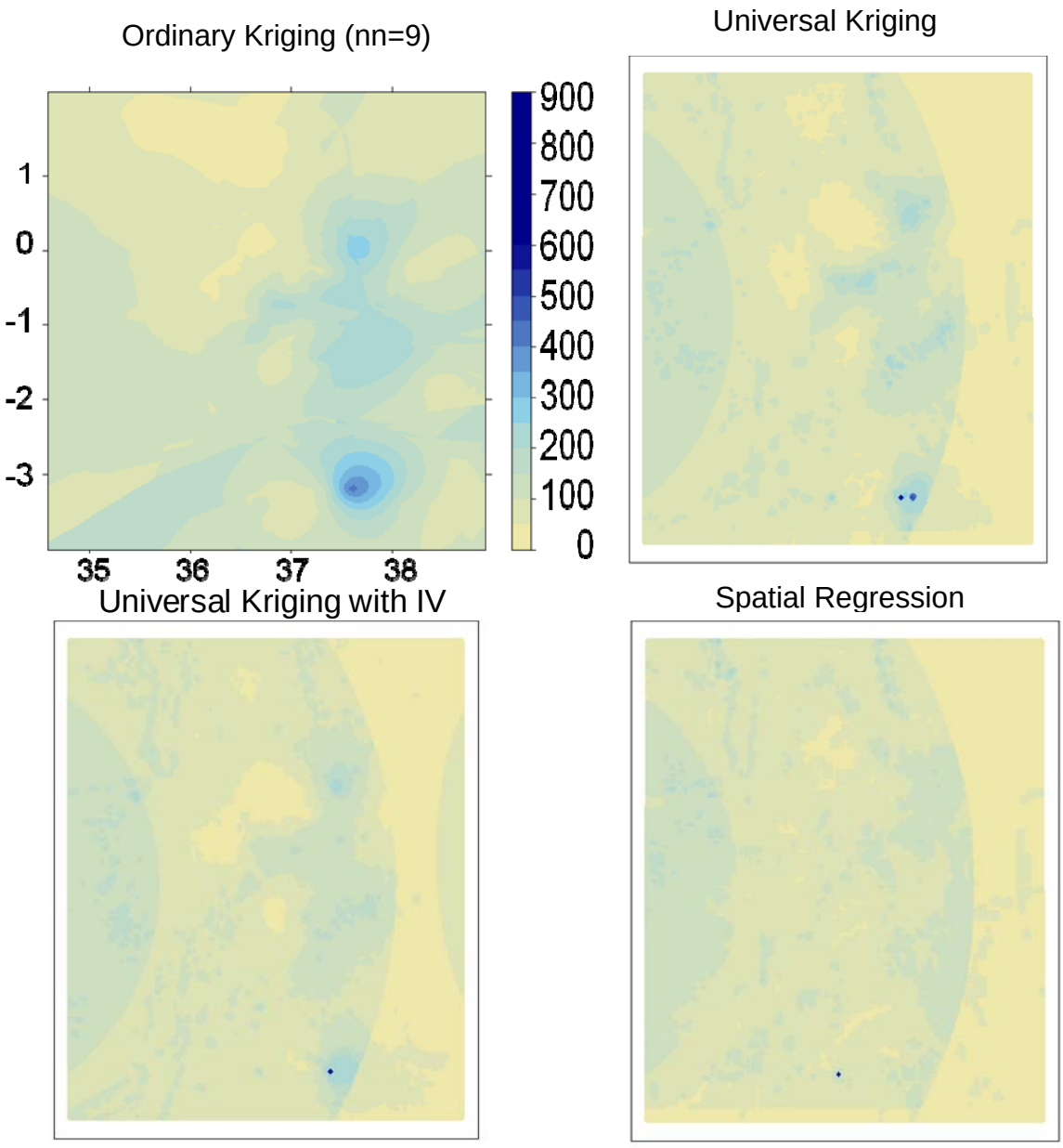

Figure A-8. November 1985 Average monthly precipitation maps generated using LOK (top left), UK (top right), UKIV (bottom left), and regression techniques (either ordinary least squares or spatial lag models, as indicated; bottom right).

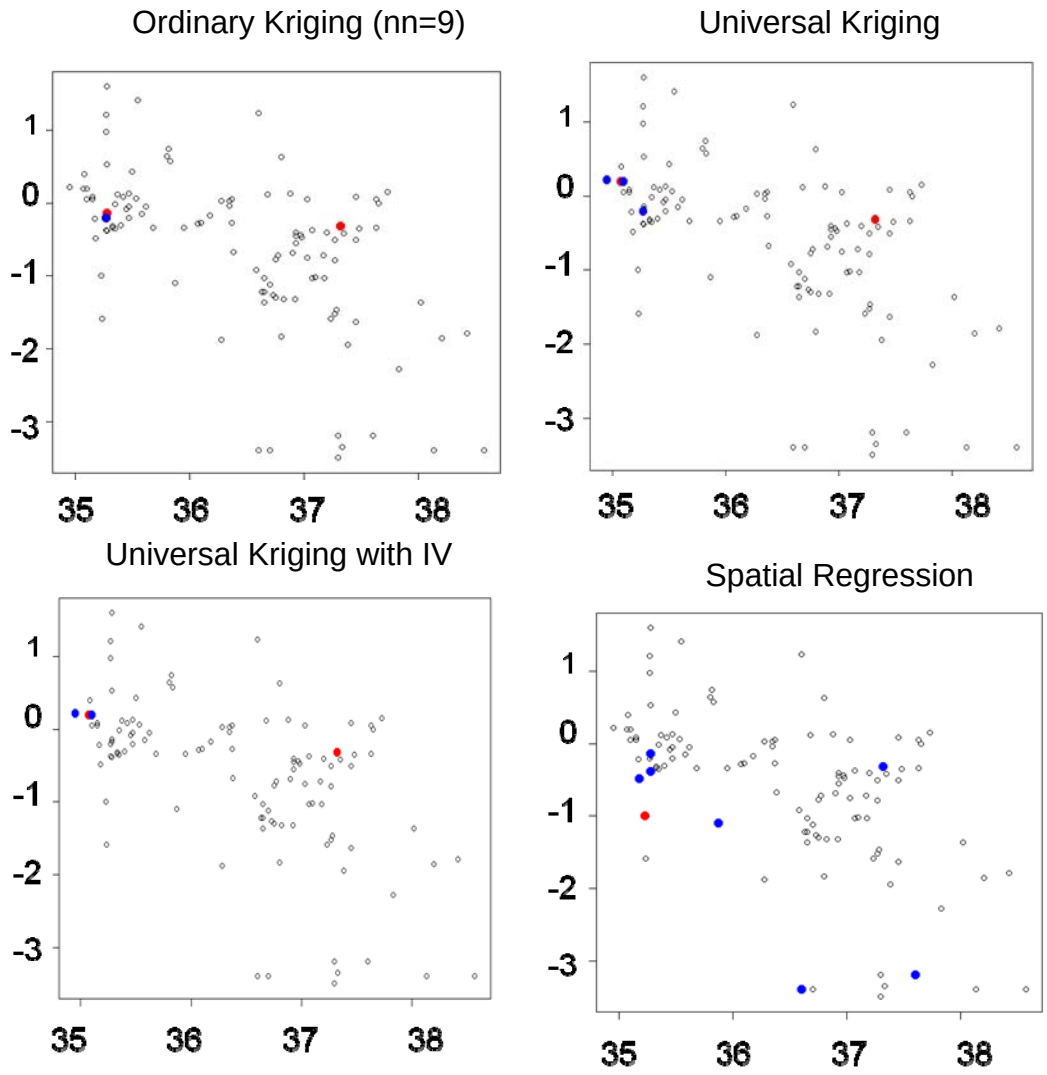

Figure A-9. January 1984 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

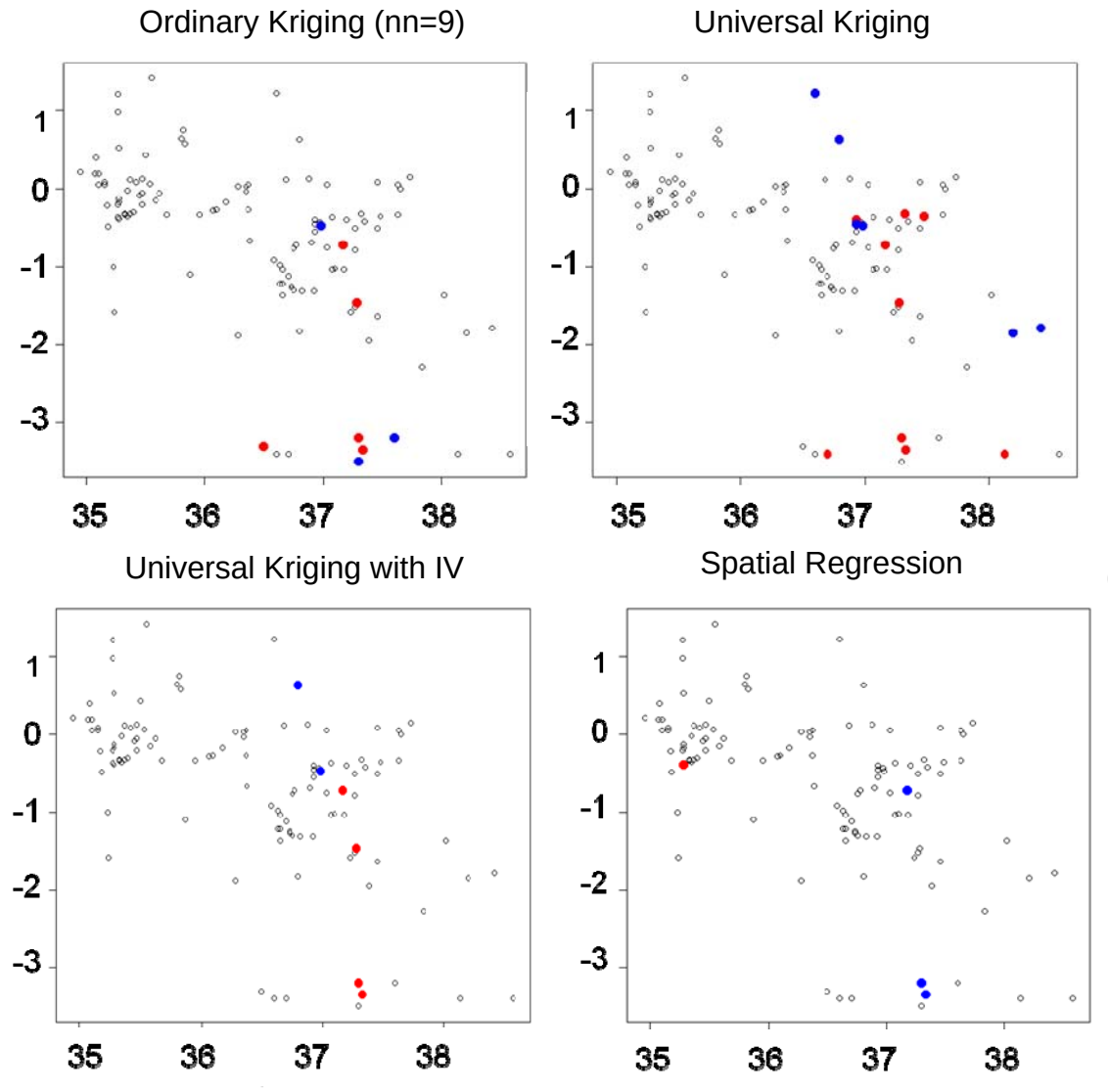

Figure A-10. April 1984 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

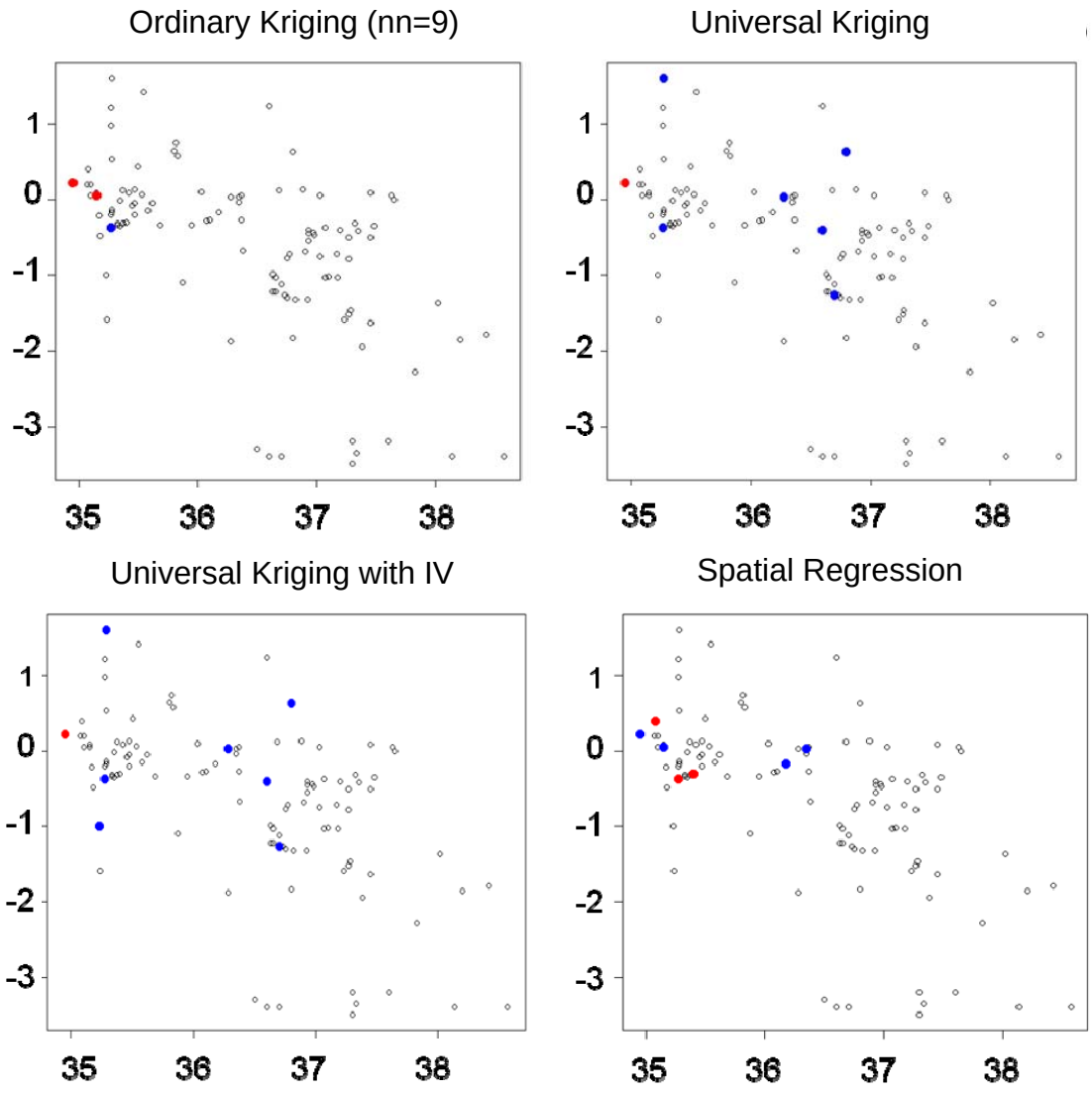

Figure A-11. August 1984 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

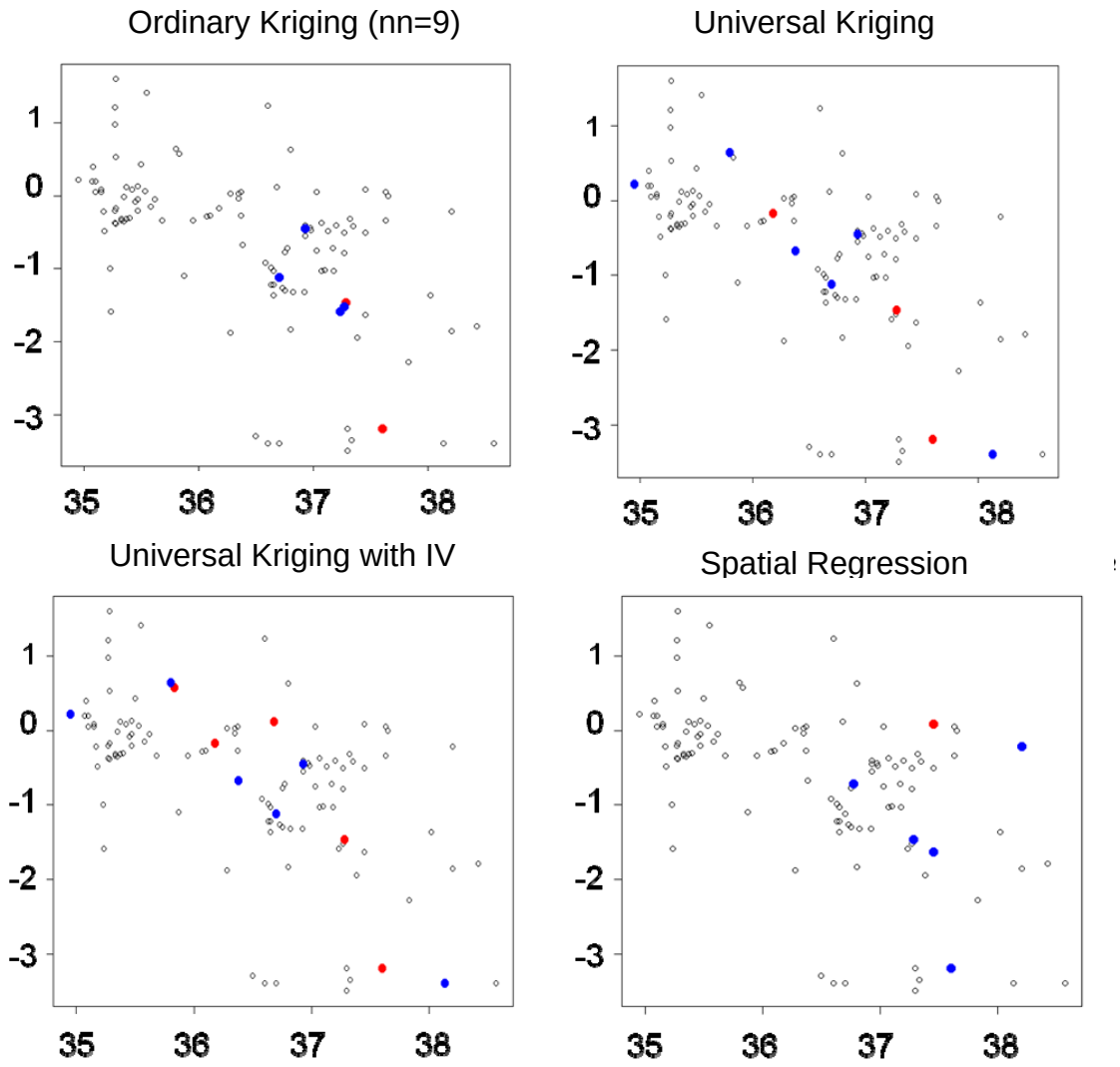

Figure A-12. November 1984 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

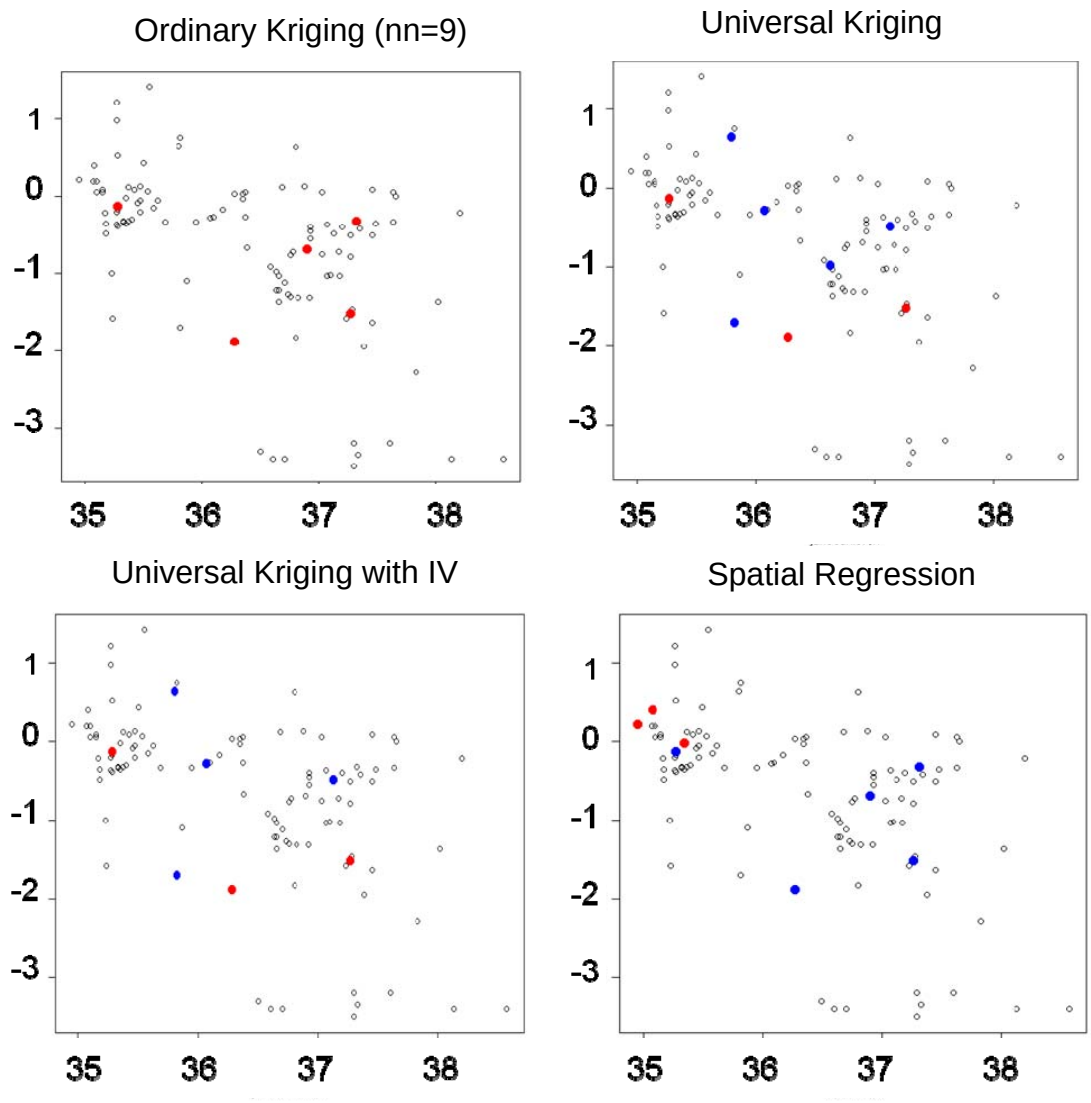

Figure A-13. January 1985 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

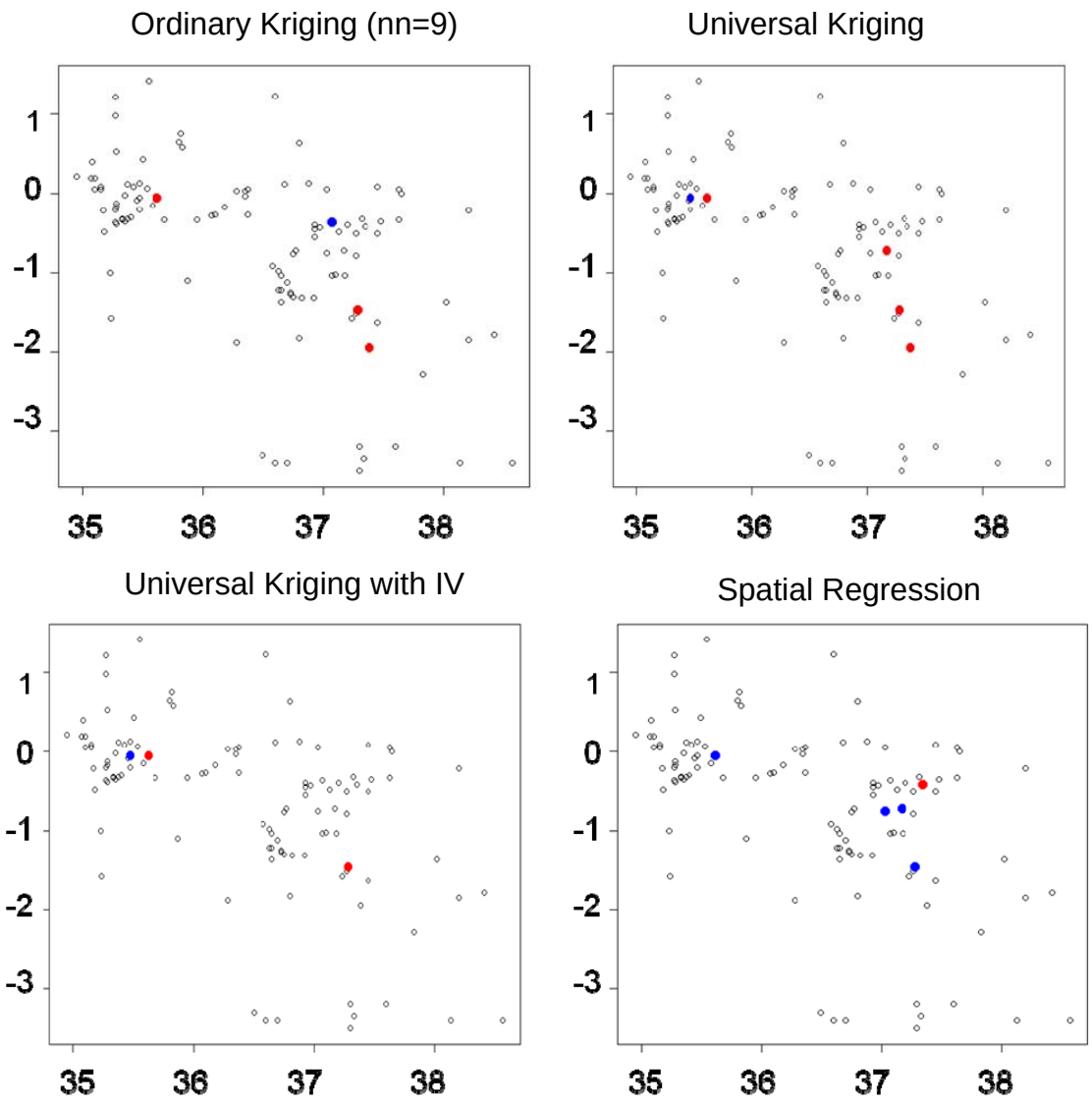

Figure A-14. April 1985 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

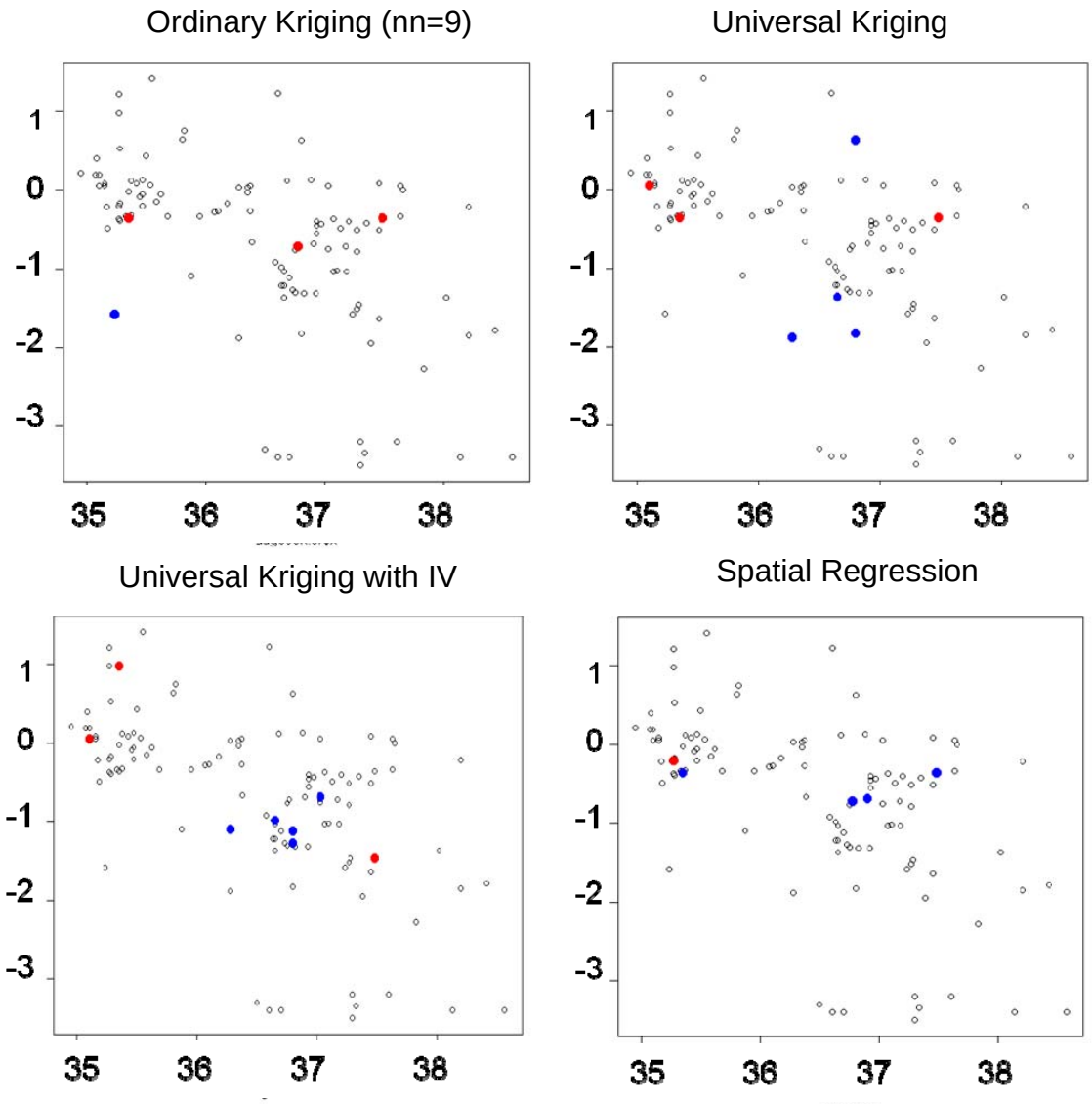

Figure A-15. August 1985 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

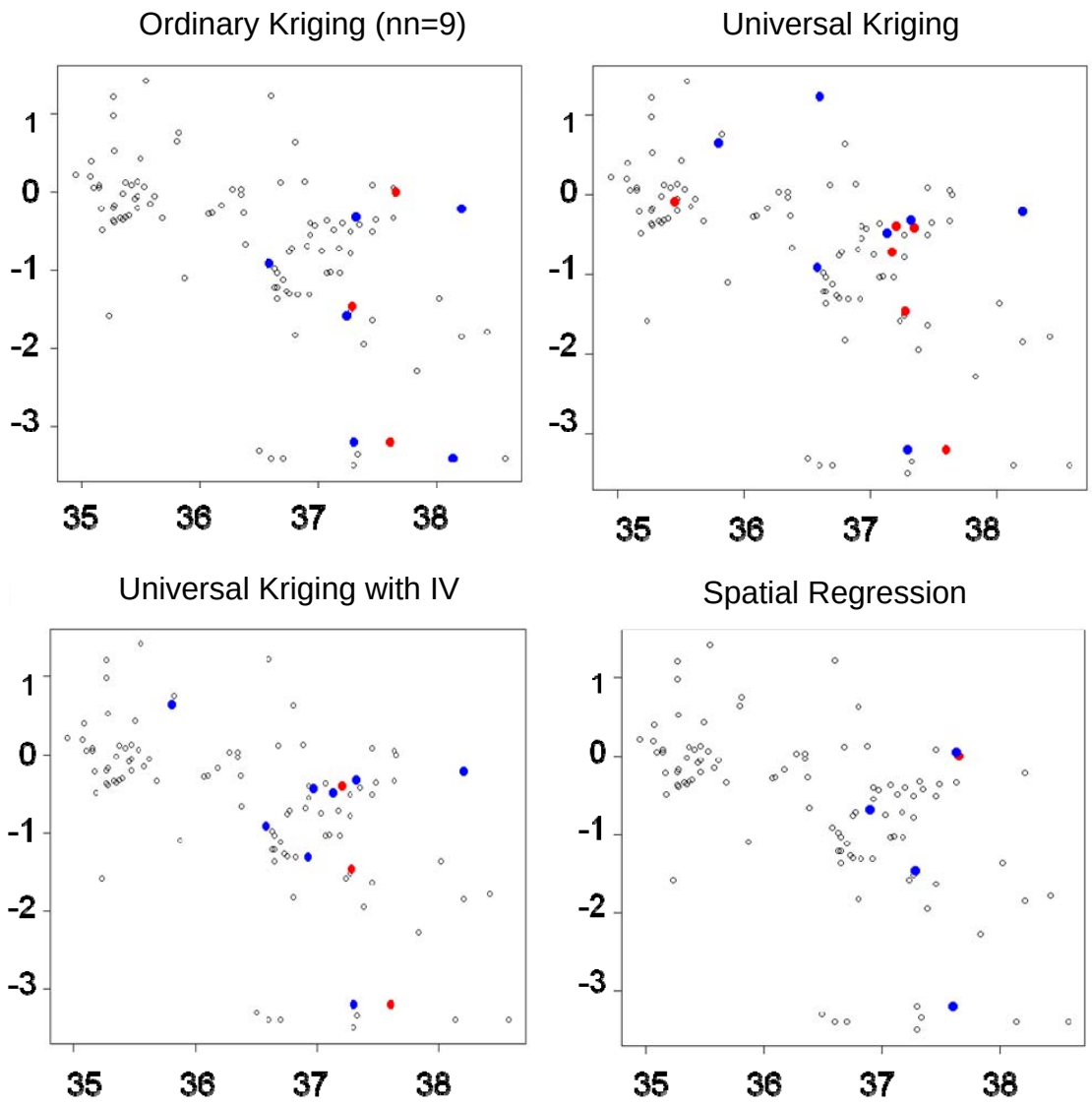

Figure A-16. November 1985 Maps of significant cross validation residuals (|z-score|>2) for LOK (top left), UK (top right), UKIV (bottom left), and regression residuals (for regression model; bottom right).

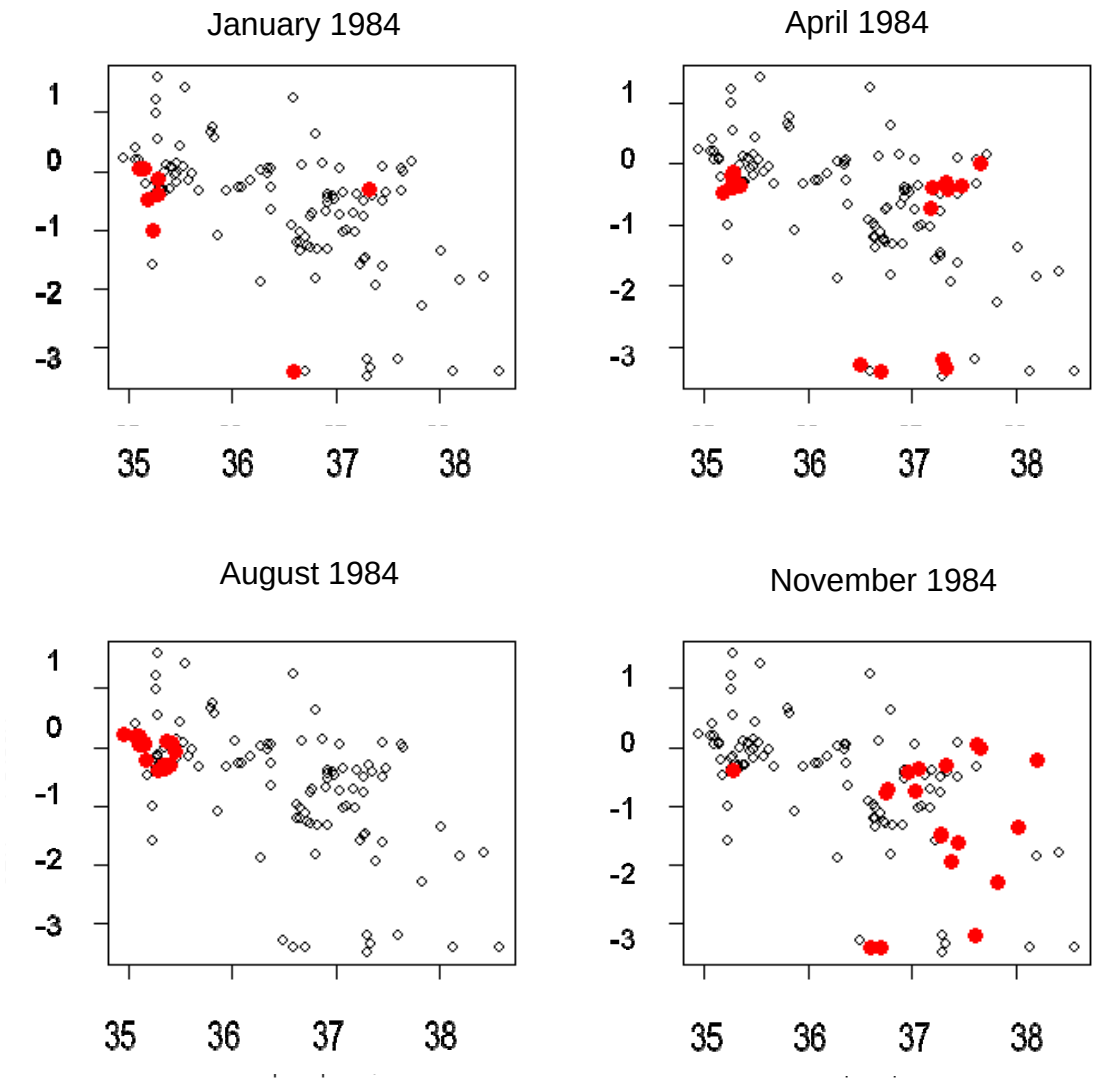

Figure A-17. Maps of CRU residuals in 1984. Significantly high residuals (zscore > 2) are shown in red, and significantly low residuals (zscore < -2) are shown in blue.

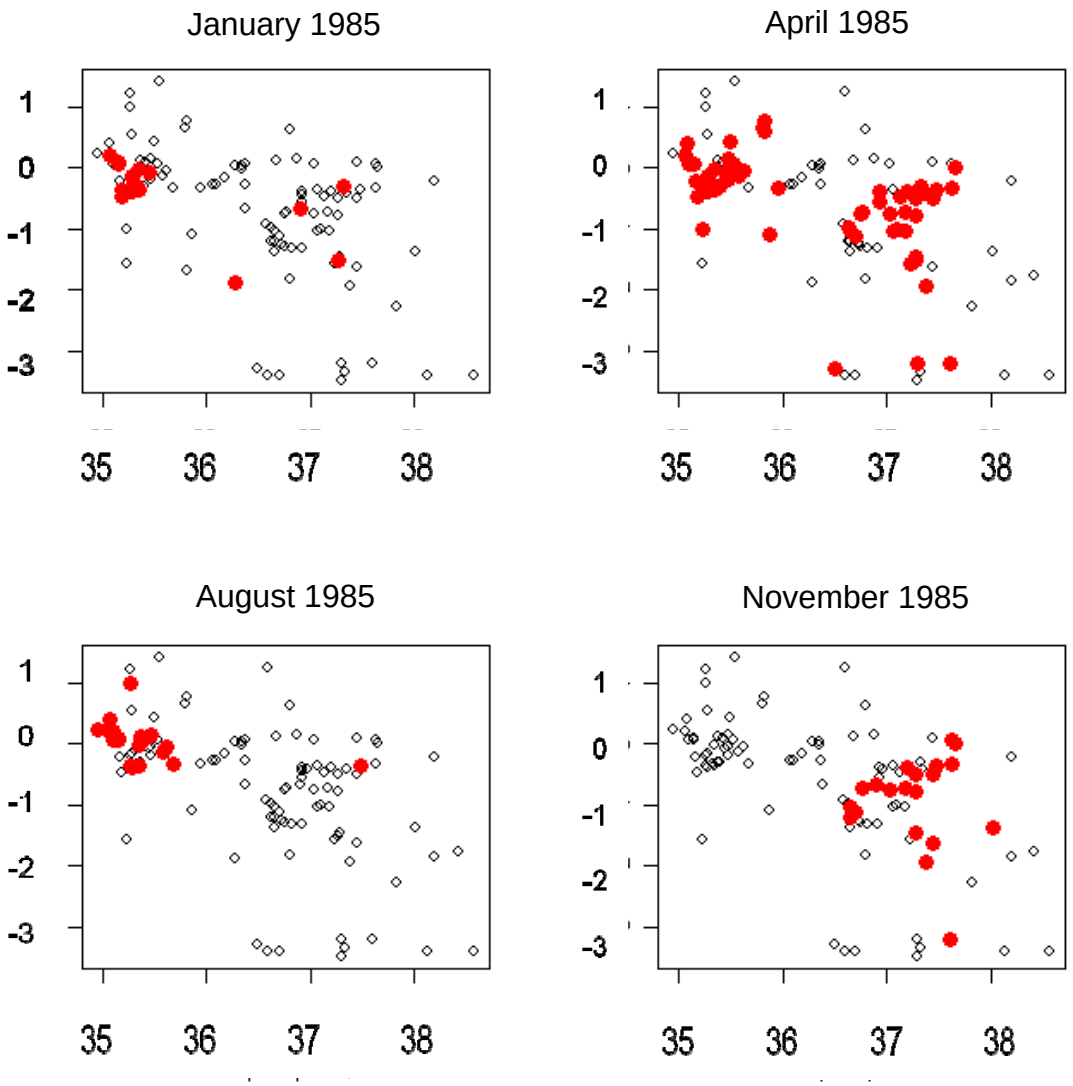

Figure A-18. Maps of CRU residuals in 1985. Significantly high residuals (zscore > 2) are shown in red, and significantly low residuals (zscore < -2) are shown in blue.

Bibliography

BIBLIOGRAPHY

Ahrens, C. 2007. *Meteorology today: An introduction to weather, climate, and the environment*. 8th ed. Pacific Grove, CA: Brooks/Cole-Thompson Learning.

Akaike, H. 1974. New look at statistical-model identification. Ieee Transactions on Automatic Control AC19: 716-23.

Anders, A.M., G.H. Roe, B. Hallet, D.R. Montgomery, N.J. Finnegan and J. Putkonen. 2006. Spatial patterns of precipitation and topography in the himalaya. In *Tectonics, climate, and landscape evolution*: Geological Society of America.

Anselin, L. 1988. A test for spatial auto-correlation in seemingly unrelated regressions. Economics Letters 28: 335-41.

---. 2005. Exploring spatial data with geoda: A workbook. Urbana-Champaign, IL.

---. 2006. Spatial regression. Urbana-Champaign, IL.

Anselin, L.A.B.A. 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In *Handbook of applied economic statistics*, 237-89. New York: Marcel Dekker.

Arora, M., P. Singh, N.K. Goel and R.D. Singh. 2006. Spatial distribution and seasonal variability of rainfall in a mountainous basin in the himalayan region. Water Resources Management 20: 489-508.

Augustin, N., D. Borchers, M. Mugglestone and S. Buckland. 1996. Regression methods with spatially referenced data. Aspects of Applied Biology 46: 67-74.

Bailey, T.C. and T.C. Gatrell. 1995. *Interactive spatial data analysis*. Essex, England: Prentice Hall.

Barros, A.P. and D.P. Lettenmaier. 1993. Dynamic modeling of the spatial distribution of precipitation in remote mountainous areas. Monthly Weather Review 121: 1195-214.

Beyer, H.L. 2004. Hawth's analysis tools for arcgis. http://www.spatialecology.com/htools.

Black, E., J. Slingo and K.R. Sperber. 2003. An observational study of the relationship between excessively strong short rains in coastal east africa and indian ocean sst. Monthly Weather Review 131: 74-94.

Boko, M., I. Niang, A. Nyong, C. Vogel, A. Githeko, M. Medany, B. Osman-Elasha, R. Tabo and P. Yanda. 2007. Africa. In *Climate change 2007: Impacts, adaptation and vulnerability. Contribution of working group ii to the fourth assessment report of the intergovernmental panel on climate change*, 433-67. Cambridge UK: Cambridge University Press.

Brooks, N., W.N. Adger and P.M. Kelly. 2005. The determinants of vulnerability and adaptive capacity at the national level and the implications for adaptation. Global Environmental Change 15: 151-63.

Brunsell, N.A. 2006. Characterization of land-surface precipitation feedback regimes with remote sensing. Remote Sensing of Environment 100: 200 - 11.

Burnham, K.P. and D.R. Anderson. 1998. *Model selection and inference: A practical information-theoretic approach*. New York.: Springer Verlag.

Buytaert, W., R. Celleri, P. Willems, B.D. Bi´Evre and G. Wyseure. 2006. Spatial and temporal rainfall variability in mountainous areas: A case study from the south ecuadorian andes. Journal of Hydrology 329: 413-21.

Caldas, M., R. Walker, E. Arima, S. Perz, S. Aldrich and C. Simmons. 2007. Theorizing land cover and land use change: The peasant economy of amazonian deforestation. Annals of the Association of American Geographers 97: 86-110.

Camberlin, P., S. Janicot and I. Poccard. 2001. Seasonality and atmospheric dynamics of the teleconnection between african rainfall and tropical sea-surface temperature: Atlantic vs. Enso. International Journal of Climatology 21: 973-1005.

Camberlin, P. and R.E. Okoola. 2003. The onset and cessation of the “long rains” in eastern africa and their interannual variability. Theoretical and Applied Climatology 75: 43-54.

Camberlin, P. and O. Planchon. 1997. Coastal precipitation regimes in kenya. Geografiska Annaler Series A: Physical Geography 79A: 109-19.

Campbell, D., J. Olson, J. Andresen, J. Qi, R. Glew, D.-Y. Kim and N. Moore. 2007. Dynamic interactions among people, livestock, and savanna ecosystems under climate change. A proposal submitted to the national science foundation, biocomplexity program: Coupled natural and human systems. .

Campbell, D.J. 1999. Response to drought among farmers and herders in southern kajiado district, kenya: A comparison of 1972-1976 and 1994-1996. Human Ecology 27: 377-416.

Campbell, D.J., J. Andresen, J. Olson, B. Pijanowski and J. Qi. 2002. An integrated analysis of regional land-climate interactions. A proposal submitted to the national science foundation, biocomplexity program: Coupled natural and human systems. .

Christensen, R. 1991. Ed. Fienberg, S. and I. Olkin. *Linear models for multivariate, time series, and spatial data Springer texts in statistics*. New York: Springer-Verlag New York, Inc.

---. 1996. Ed. Casella, G., S. Fienberg and I. Olkin. *Plane answers to complex questions: The theory of linear models Springer texts in statistics*. Second ed. New York: Springer-Verlag New York, Inc.

Cressie, N.A.C. 1993. Ed. Barnett, V., R.A. Bradley, N.I. Fisher, J.S. Hunter, J.B. Kadane, D.G. Kendall, A.F.M. Smith, S.M. Stigler, J.L. Teugels and G.S. Watson. *Statistics for spatial data Wiley series in probability and mathematical statistics*. Revised ed. New York: John Wiley & Sons, Inc.

Daly, C., R.P. Neilson and D.L. Phillips. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33: 140-58.

Davidson, O., K. Halsnaes, S. Huq, M. Ko, B.Metz, Y. Sokona and J. Verhagen. 2003. The development and climate nexus: The case of sub-sahara African Climate Policy 3: 97-113.

Dekker, S.C., M. Rietkerk and M.F.P. Bierkens. 2007. Coupling microscale vegetation–soil water and macroscale vegetation–precipitation feedbacks in semiarid ecosystems. Global Change Biology 13: 671-78.

Dickenson, R.E., R.M. Errico, F. Giorgi and G.T. Bates. 1989. A regional climate model for western united states. Climatic Change 31: 273-304.

Diodato, N. 2005. The influence of topographic co-variables on the spatial variability of precipitation over small regions of complex terrain. International Journal of Climatology 25: 351-63.

Esri. 2006. Esri arcmap 9.2. Copyright 1999-2006 ESRI Inc.

Fao/Iiasa/Isric/Iss-Cas/Jrc. 2009. Harmonized world soil database version 1.1: FAO/IIASA/ISRIC/ISS-CAS/JRC, FAO, Rome, Italy and IIASA, Laxenburg, Austria. .

Fauchereau, N., S. Trzaska, Y. Richard, P. Roucou and P. Camberlin. 2003. Seasurface temperature co-variability in the southern atlantic and indian oceans and its connections with the atmospheric circulation in the southern hemisphere. International Journal of Climatoloty 23: 663-77.

Gibbons, R. 1994. *Statistical methods for groundwater monitoring.* New York: John Wiley & Sons.

Giorgi, F. 1990. Simulation of regional climate using a limited area model nested in a general circulation model. Journal of Climate 3: 941-63.

---. 2006. Regional climate modeling: Status and perspectives. J. Phys. IV France 139: 101-18.

Giorgi, F., P.H. Whetton, R.G. Jones, J.H. Christensen, L.O. Mearns, B. Hewitson, H.V. Storch, R. Francisco and C. Jack. 2001. Emerging patterns of simulated regional climatic changes for the 21st century due to anthropogenic forcings Geophysical Research Letters 28: 3317-20.

Goovaerts, P. 1997. *Geostatistics for natural resources evaluation*. New York, Oxford: Oxford University Press.

---. 1999. Performance comparison of geostatistical algorithms for incorporating elevation into the mapping of precipitation. In *4th International Conference on GeoComputation*. Mary Washington College, Fredericksburg: Virginia, USA.

---. 1999. Using elevation to aid the geostatistical mapping of rainfall erosivity. Catena 34: 227-42.

---. 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology 228: 113-29.

Gottschalk, T.K., K. Ekschmitt, S. Isfendiyaroglu, E. Gem and V. Wolters. 2007. Assessing the potential distribution of the caucasian black grouse tetrao mlokosiewiczi in turkey through spatial modelling. Journal of Ornithology 148: 427-34.

Gumperts, M., J. Graham and J. Ristaino. 1997. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variation on disease presence. Journal of Agricultural Biology and Environmental Statistics 2: 131-56.

Haining, R.P. 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge: Cambridge University Press.

Hastenrath, S., A. Nicklis and L. Greishar. 1993. Atmospheric-hydrospheric mechanisms of climate anomalies in the western equatorial indian ocean. Journal of Geophysical Research 98: 20219-35.

Hausman, J.A. 1978. Specification tests in econometrics. Econometrica 46: 1251-72.

Hengl, T., G.B.M. Heuvelink and A. Stein. 2003. Comparison of kriging with external drift and regression-kriging. http://www.itc.nl/library/Academic_output/.

Hession, S.L. and N. Moore. 2010. A spatial regression analysis of the influence of topography on monthly rainfall in east africa. International Journal of Climatology.

Huffer, F.W. and H.L. Wu. 1998. Markov chain monte carlo for autologistic regression models with application to the distribution of plant species. Biometrics 54: 509-24.

Hulme, M., R. Doherty, T. Ngara, M. New and D. Lister. 2001. African climate change: 1900-2100. Climate Research 17: 145-68.

Hulme, M. and M. New. 1997. Dependence of large-scale precipitation climatologies on temporal and spatial gauge sampling. Journal of Climate 10: 1099-113.

Hulme, M., T.J. Osborn and T.C. Johns. 1998. Precipitation sensitivity to global warming: Comparison of observations with hadcm2 simulations Geophysical Research Letters 25: 3379-82.

Hutchinson, M.F. 1998. Interpolation of rainfall data with thin-plate smoothing splines i: Two dimensional smoothing of data with short range correlation. Journal of Geographic Information and Decision Analysis 2: 152-67.

---. 1998. Interpolation of rainfall data with thin-plate smoothing splines ii: Analysis of topographic dependence. Journal of Geographic Information and Decision Analysis 2: 168-85.

Hutchinson, M.F. and R.J. Bischof. 1983. A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the hunter valley, new south wales. Australian Meteorological Magazine 31: 179-84.

Ji, L. and A.J. Peters. 2004. A spatial regression procedure for evaluating the relationship between avhrr-ndvi and climate in the northern great plains. International Journal of Remote Sensing 25: 297-311.

Jones, P.G.A.P.K.T. 2003. The potential impacts of climate change on maize production in africa and latinamerica in 2055. Global Environmental Change 13: 51-59.

Journel, A.G. and C.J. Huijbregts. 1978. *Mining geostatistics*. London: Academic Press.

Journel, A.G. and M.E. Rossi. 1989. When do we need a trend model in kriging? Mathematical Geology 21: 715-39.

Kariya, T. and H. Kurata. 2004. *Generalized least squares*. West Sussex, England: John Wiley & Sons Ltd.

Kates, R.W. 2000. Cautionary tales: Adaptation and the global poor. Climatic Change 45: 5-17.

Kazembe, L.N. 2007. Spatial modelling and risk factors of malaria incidence in northern malawi. Acta Tropica 102: 126-37.

Kazembe, L.N. and J.J. Namangale. 2007. A bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in malawi. European Journal of Epidemiology 22: 545-56.

Kelejian, H.H. and I.R. Prucha. 2004. Estimation of simultaneous systems of spatially interrelated cross sectional equations. Journal of Econometrics 118: 27-50.

Khan, M.S., P. Coulibal and Y. Dibike. 2006. Uncertainty analysis of statistical downscaling methods. Journal of Hydrology 319: 357-82.

Kim, Y. and G. Wang. 2007. Impact of vegetation feedback on the response of precipitation to antecedent soil moisture anomalies over north america. Journal of Hydrometeorology 8: 534-50.

Krige, D.G. 1951. A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical, Metal. and Mining Society of South Africa 52: 119-39.

Kyriakidis, P.C., J. Kim and N.L. Miller. 2001. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. Journal of Applied Meteorology 40: 1855-77.

Kyriakidis, P.C., N.L. Miller and J. Kim. 2004. A spatial time series framework for simulating daily precipitation at regional scales. Journal of Hydrology 297: 236-55.

Lesage, J.P. 1998. Spatial econometrics. Toledo, OH.

Lesage, J.P. and R.K. Pace. 2009. *Introduction to spatial econometrics*: Chapman and Hall/CRC 2009.

Lobell, D.B., M.B. Burke, C. Tebaldi and M.D. Mastrandrea. 2008. Prioritizing climate change adaptation needs for food security in 2030. Science 319: 607.

Lobell, D.B. and C.B. Field. 2008. Estimation of the carbon dioxide (co2) fertilization effect using growth rate anomalies of co2 and crop yields since 1961. Global Change Biology 14.

Lyon, S.W., F.D. F, D.J. Gochis, N.A. Brunsell, C.L. Castro, F.K. Chow, Y. Fan, D. Fuka, Y. Hong, P.A. Kucera, S.W. Nesbitt, N. Salzmann, J. Schmidli, P.K. Snyder, A.J. Teuling, T.E. Twine, S. Levis, J.D. Lundquist, G.D. Salvucci, A.M. Sealy and M.T. Walter. 2008. Coupling terrestrial and atmospheric water dynamics to improve prediction in a changing environment. Bulletin of the American Meteorology Society 89: 1275-79.

Majugu, A.W. and S.A.K. Magezi. 1985. Enso based seasonal to inter-annual rainfall variability over uganda in particular and eastern africa in genera.

Marquinez, J., J. Lastra and P. Garcia. 2003. Estimation models for precipitation in mountainous regions: The use of gis and multivariate analysis. Journal of Hydrology 270: 1-11.

Matheron, G. 1963. Principles of geostatistics. Economic Geology 58: 1246-66.

Mccarthy, J.J. 2001. *Climate change 2001: Impacts, adaptation, and vulnerability Intergovernmental panel on climate change. Working group ii.* Cambridge: Press Syndicate of the University of Cambridge.

Mearns, L.O., W. Easterling, C. Hays and D. Marx. 2001. Comparison of agricultural impacts of climate change calculated from high and low resolution climate model scenarios: Part i. The uncertainty due to spatial scale. Climate Change 51: 131-72.

Mearns, L.O., F. Giorgi, P. Whetton, D. Pabon, M. Hulme and M. Lal. 2003. Guidelines for use of climate scenarios developed from regional climate model experiments. In *DDC of IPCC TGCIA*.

Mearns, L.O., M. Hulme, T.R. Carter, R. Leemans, M. Lal and P. Whetton. 2001. Climate scenario development In *Climate change 2001: The scientific basis. Contribution of working group i to the third assessment report of the ipcc*, 583-638. Cambridge: Cambridge University Press.

Mearns, L.O., T. Mavromatis, E. Tsvetsinskaya, C. Hays and W. Easterling. 1999. Comparative responses of epic and ceres crop models to high and low resolution climate change scenarios. Journal of Geophysical Research 104: 6623-46.

Méndez-Barroso, L.A., E.R. Vivoni, C.J. Watts and J.C. Rodríguez. 2009. Seasonal and interannual relations between precipitation, surface soil moisture and vegetation dynamics in the north american monsoon region. Journal of Hydrology 377: 59-70.

Miller, J., J. Franklin and R. Aspinall. 2007. Incorporating spatial dependence in predictive vegetation models. Ecological Modelling 202: 225-42.

Mitchell, T.D. and P.D. Jones. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology 25: 693-712.

Moore, N., G. Alagarswamy, B. Pijanowski, P. Thornton, B. Lofgren, J. Olson, J. Andresen, P. Yanda and J. Qi. In review. East african food security as influenced by future climate change and land use change at local to regional scales. Climatic Change.

Moore, N., E. Arima, R. Walker and R.R. Da Silva. 2007. Uncertainty and the changing hydroclimatology of the amazon. Geophysical Research Letters 34: -.

Moore, N., B. Lofgren, J. Andresen, J. Olson, D. Ray, B. Pijanowski and N. Torbick. 2006. Simulations of climate variability resulting from projected land cover change in east africa. In *American Association of Geographers Annual Meeting*. Chicago, IL.

Moore, N., B. Lofgren, J. Andresen, B. Pijanowski and J.M. Olson. 2005. Projected changes in precipitation variability and distribution due to land cover change in east africa. In *American Geophysical Union (AGU) Fall Meeting*. San Francisco, CA.

Mukabana, J.R. and R.A. Pielke. 1996. Investigating the influence of synoptic-scale monsoonal winds and mesoscale circulations on diurnal weather patterns over kenya using a mesoscale numerical model. Monthly Weather Review 124: 224-43.

Munga, S., N. Minakawa, G. Zhou, E. Mushinzimana, O.O.J. Barrack, A.K.Githeko and G. Yan. 2006. Association between land cover and habitat productivity of malaria vectors in western kenyan highlands. Am. J. Trop. Med. Hyg. 74: 69-75.

Mutai, C.C. and M.N. Ward. 2000. East african rainfall and the tropical circulation/convection on intraseasonal to interannual timescales. Journal of Climate 13: 3915-39.

Neter, J., W. Wasserman and M.H. Kutner. 1990. *Applied linear statistical models: Regression, analysis of variance, and experimental designs*. Third ed. Homewood, IL 60430 Boston, MA 02116: Richard D. Irwin, Inc.

New, M., M. Hulme and P. Jones. 2000. Representing twentieth-century space–time climate variability. Part ii: Development of 1901–96 monthly grids of terrestrial surface climate. Journal of Climate 13: 2217-38.

Ng'ang'a, J.K. 1992. The climate and meteorology of nairobi region, kenya. African Urban Quarterly 7: 6-12.

Nicholson, S.E. 1996. A review of climate dynamics and climate variability in east africa. In *The limnology, climatology and paleoclimatology of the east african lakes*. Amsterdam: Gordon and Breah Publishers.

Notaro, M. and Z. Liu. 2008. Statistical and dynamical assessment of vegetation feedbacks on climate over the boreal forest. Climate Dynamics 31: 691-712.

Notaro, M., Y. Wang, Z. Liu, R. Gallimore and S. Levis. 2008. Combined statistical and dynamical assessment of simulated vegetation-rainfall during the mid-holocene. Global Change Biology 14: 347-68.

Oba, G., N.C. Stenseth and W.J. Lusigi. 2000. New perspectives on sustainable grazing management in arid zones of sub-saharan africa. Bioscience 50: 35-51.

Oba, G., R.B. Weladji, W.J. Lusigi and N.C. Stenseth. 2003. Scale-dependent effects of grazing on rangeland degradation in northern kenya: A test of equilibrium and non-equilibrium hypothesis. Land Degradation & Development 14: 83-94.

Oettli, P. and P. Camberlin. 2005. Influence of topography on monthly rainfall distribution over east africa. Climate Research 28: 199-212.

Olson, J.M., S. Misana, D.J. Campbell, M. Mbonile and S. Mugisha. 2004. The spatial patterns and root causes of land use change in east africa, lucid project working paper 47.

Ord, K. 1975. Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70: 120-26.

Pandey, D.N., A.K. Gupta and D.M.Anderson. 2003. Rainwater harvesting as an adaptation to climate change. Curr. Sci. India 85: 46-59.

Pardo-Iguzquiza, E. 1998. Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. International Journal of Climatology 18: 1031-47.

Pascual, M., J.A. Ahumada, L.F. Chaves, X. Rodó and M. Bouma. 2006. Malaria resurgence in the east african highlands: Temperature trends revisited. P. Natl. Acad. Sci. 103: 5829-34.

Perttunen, C.D. and B.E. Stuckman. 1990. The rank transformation applied to a multivariate method of global optimization. IEEE Transactions on Systems, Man and Cybernetics 20: 1216 - 20.

Piper, S.C. and E.F. Stewart. 1996. A gridded global data set of daily temperature and precipitation for terrestrial biosphere modeling. Global Biogeochemical Cycles 10: 757-82.

Propastin, P., N. Muratova and M. Kappas. 2006. Reducing uncertainty in analysis of relationship between vegetation patterns and precipitation. In *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* Lisbon, Portugal.

Rey, S.J. and M.G. Boarnet. 2004. A taxonomy of spatial econometric models for simultaneous equations systems. In *Advances in spatial econometrics: Methodology, tools and applications*, 99-119. Heidelberg, New York: Springer Berlin.

Richard, Y., N. Fauchereau, I. Poccard, M. Rouault and S. Trzaska. 2001. 20th century droughts in southern africa: Spatial and temporal variability, teleconnections with oceanic and atmospheric conditions. International Journal of Climatology 21: 873-85.

Richard, Y. and I. Poccard. 1998. A statistical study of ndvi sensitivity to seasonal and interannual rainfall variations in southern africa International Journal of Remote Sensing 19: 2907 - 20.

Rodriguez-Iturbe, I. and A. Porporato. 2004. *Ecohydrology of water-controlled ecosystems: Soil moisture and plant dynamics*. Cambridge [UK], New York: Cambridge University Press.

Rogers, J.C., J.A. Winkler, D.R. Legates and L.O. Mearns. 2003. Climate. In *Geography in america at the dawn of the 21st century*. New York: Oxford University Press.

Rouse, J.W., R.H. Haas, J.A. Schell, D.W. Deering and J.C. Harlan. 1974. Monitoring the vernal advancement and retrogradation (greenwave effect) of natural vegetation. Type iii final report., 371. Greenbeld, MD.

Roy, S.B. 2009. Mesoscale vegetation-atmosphere feedbacks in amazonia. Journal of Geophysical Research 114.

Salathe, E.P.J., P.W. Mote and M.W. Wiley. 2007. Review of scenario selection and downscaling methods for the assessment of climate change impacts on hydrology in the united states pacific northwest. International Journal of Climatology 27: 1611-21.

Sankaran, M., N.P. Hanan, R.J. Scholes, J. Ratnam, D.J. Augustine, B.S. Cade, J. Gignoux, S.I. Higgins, X.L. Roux, F. Ludwig, J. Ardo, F. Banyikwa, A. Bronn, G. Bucini, K.K. Caylor, M.B. Coughenour, A. Diouf, W. Ekaya, C.J. Feral, E.C. February, P.G.H. Frost, P. Hiernaux, H. Hrabar, K.L. Metzger, H.H.T. Prins, S. Ringrose, W. Sea, J. Tews, J. Worden and N. Zambatis. 2005. Determinants of woody cover in african savannas. Nature 438: 846-49.

Schabenberger, O. and C.A. Gotway. 2005. Ed. Carlin, B.P., C. Chatfield, M. Tanner and J. Zidek. *Statistical methods for spatial data analysis Texts in statistical science*. Boca Raton, FL: Chapman & Hall/CRC.

Scholes, R.J. and R. Biggs. 2004. Ecosystem services in southern africa: A regional assessment

In *Council for Scientific and Industrial Research.* Pretoria, South Africa.

Schreck, C.J. and F.H.M. Semazzi. 2004. Variability of the recent climate of eastern africa. International Journal of Climatology 24: 681-701.

Searle, S.R. 1982. *Matrix algebra useful for statistics*. New York: Wiley.

Sharples, J.J., M.F. Hutchinson and D.R. Jellett. 2005. On the horizontal scale of elevation dependence of australian monthly precipitation

Journal of Applied Meteorology 44: 1850-65.

Shortridge, A.M. 2010. R code for nscore transformations.

Smucker, T.A. and B. Wisner. 2008. Changing household responses to drought tharaka, kenya: Persistence, change, and challenge. Disasters 32: 190-215.

Spreen, W.C. 1947. A determination of the effect of topography upon precipitation. Transactions of the American Geophysical Union 28: 285-90.

Stock, R.F. 2004. *Africa south of the sahara: A geographical interpretation*. 2nd ed. New York: Guilford Press.

Storch, H.V., H. Langenberg and F. Feser. 2000. A spectral nudging technique for dynamical downscaling purposes. Mon. Wea. Rev. 128: 3664-73.

Tarhule, A. and P.J. Lamb. 2003. Climate research and seasonal forecasting for west africans: Perceptions, dissemination, and use. B. Am. Meteorol. Soc. 84: 1741-59.

Theil, H. 1971. *Principles of econometrics*. New York: Wiley.

Thomas, C.D., A. Cameron, R.E. Green, M. Bakkenes, L.J. Beaumont, Y.C. Collingham, B.F.N. Erasmus and M.F.D. Siqueira. 2004. Extinction from climate change. Nature 427: 145-48.

Thornton, P., T. Owiyo, R. Kruska, M. Herrero, P. Kristjanson, A. Notenbaert, N. Bekele, A. Omolo, P. Jones, V. Orindi, A. Adwerah, B. Otiende, S. Bhadwal, K. Anantram, S. Nair and V. Kumar. 2006. Mapping climate vulnerability and poverty in africa. Report for the u.K. Department for international development.

Thornton, P.K., P.G. Jones, G. Alagarswamy and J. Andresen. 2009. Spatial variation of crop yield response to climate change in east africa. Global Environmental Change 19: 54-65.

Torbick, N.M., B.L. Becker, S.L. Hession, J. Qi, G.J. Roloff and R.J. Stevenson. 2010. Assessing invasive plant infestation and disturbance gradients in a freshwater wetland using a giscience approach. Wetlands Ecology Management.

Walker, R., E. Moran and L. Anselin. 2000. Deforestation and cattle ranching in the brazilian amazon: External capital and household processes. World Development 28: 683-99.

Wang, J., P.M. Rich and K.P. Price. 2003. Temporal responses of ndvi to precipitation and temperature in the central great plains, USA. International Journal of Remote Sensing 24: 2345-64.

Wang, Y., M. Notaro, Z. Liu, R. Gallimore and J.E. Kutzback. 2008. Detecting vegetation-precipitation feedbacks in mid-holocene north africa from two climate models. Climate of the Past 4: 59-67.

Wangui, E.E. 2003. Links between gendered division of labour and land uses in kajiado district, kenya. Lucid working paper 23.

---. 2004. Links between gendered division of labor and land use in oloitokito division, s.E. Kajiado district, kenya. In *Department of Geography*. East Lansing, MI: Michigan State University.

Weisse, A.K. and P. Bois. 2001. Topographic effects on statistical characteristics of heavy rainfall and mapping in the french alps. Journal of Applied Meteorology 40: 720-40.

Wisner, B. 2004. Assessment of capability and vulnerability. In *Vulnerability: Disasters, development and people*, 183-93. London: Earthscan.

Wu, J., W.A. Norvell and R.M. Welch. 2006. Kriging on highly skewed data for dtpa-extractable soil zn with auxiliary information for ph and organic carbon. Geoderma 134: 187-99.

Zeng, N. and J. Yoon. 2009. Expansion of the world's deserts due to vegetation-albedo feedback under global warming. Geophysical Research Letters 36.