Low-Rank Nonparametrics For Gridded Precipitation Estimation

by

Gregory Benton

Department of Applied Mathematics

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Applied Mathematics

2018

This thesis entitled:
Low-Rank Nonparametrics For Gridded Precipitation Estimation
written by Gregory Benton
has been approved for the Department of Applied Mathematics

_____

William Kleiber

_____

Brian Zaharatos

_____

Ben Livneh

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Benton, Gregory (M.S., Applied Mathematics)

Low-Rank Nonparametrics For Gridded Precipitation Estimation

Thesis directed by Prof. William Kleiber

Estimation of gridded precipitation is a major point of interest in climatological and hydrological research. Using a novel approach based around kernel density estimation we attempt to improve on a currently available estimators of gridded precipitation in both accuracy and understanding uncertainty in prediction. The method is constructed and validated using the United States Historical Climatology Network dataset covering the continental United States with sparse and irregular observation stations and accurate probability distributions that capture seasonally variance in the data are generated. Spatial estimates of local models at arbitrary locations, both in and out of the observational network, are analyzed and an accurate method using generalized additive models is developed. Finally a preliminary analysis of gridded estimation is discussed and serves as a motivation for further research.

Dedication

This work is dedicated to my parents, Jeff and Wendy Benton. Your gentle support and guidance along the way has made all the difference. Thank you for never giving up.

Acknowledgements

The list of people to whom I owe thanks is much longer than can be included here. Firstly, I would like to thank Prof. Will Kleiber, for being an inexhaustible resource through this whole process and for using his time and patience to introduce me to research. His enthusiasm for the project and for pursuing research goals has been contagious.

Along with Prof. Kleiber I would like to acknowledge the applied math community for fostering an incredible environment in which to grow and for being filled with students I am grateful to call my friends and faculty I have been honored to work with. Thank you to: Brian Zaharatos for giving me an opportunity to help instruct and for fostering my interests beyond math, Manuel Lladser for taking a fun side project radio show and getting excited about it, and Anne Dougherty for her expertise as my academic advisor, for putting in countless hours to curate an amazing department, and for funding this work through the NSF EXTREEMS grant DMS 1407340.

I would also like to thank the CUCRC for having open doors and fresh coffee and being an small oasis on a busy campus. I hope that I will someday be able to pass on a small portion of the support that was shown to me there. Additionally, I would like to thank the men and women of Jaywalker Lodge for guiding me back to school and showing me that my goals were worth pursuing.

Finally, I wish to express my endless gratitude for Ashley Flinn, for getting me to submit all of my applications on time and having a foolish level of belief in me - I wouldn't have gotten here without you.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Motivation

Precipitation is a major factor in the study of climatology and has immense importance as a social and economic driver, with some regions of the United States obtaining over 75% of drinking water from high alpine runoff [5]. The particular issue of estimating high alpine water availability has been approached by a number of methods including spatial interpolation of local precipitation models and more recently airborne laser altimetry (lidar) has served as a method for empirical data collection [4, 6]. Issues have arisen for both of these methods as spatial interpolation can be dependent on a dense set of collection site with which to start from and lidar collection is unfeasible for collection at the watershed scale which can be 100's of square miles in size. These issues with common methodologies motivate the need for precipitation models that can be accurately estimated across fine-resolution grids. Examining the locations of sites in the United States Historical Climatology Network (USHCN) in Figure 1.1 (the dataset used here) it is clear that these grids must be able to estimated from a sparse sampling of points.

Overall precipitation modeling presents a number of challenges, but due to the importance of reliable and accurate understanding of water resources as a driving factor in events such as wild fires, droughts, and floods it has long been studied from a statistical perspective. Among these challenges are the difficulties in modeling precipitation with any parametric distributions, and although some success has been made with mixed distributions many past successful models have revolved around the use of nonparametric models, namely kernel density estimators [9] [16].

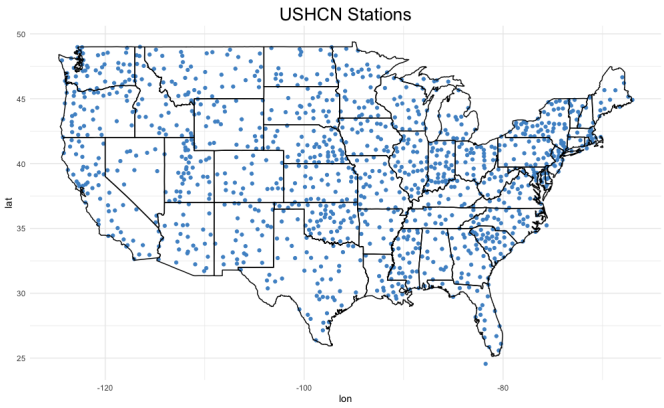

Figure 1.1: A spatial map of where data is recorded in the USHCN

Additionally precipitation recording sites are typically unavailable in climatologically significant regions, such as at high altitudes or remote mountains meaning that for inference to be made about these regions models must be estimated at spatial locations outside of the original observational network.

The natural issue is that parametric models that can be captured in a small number of significant parameters, and are thus ideal for spatial estimation, suffer in terms of accuracy, and more accurate nonparametric models for local estimation may not have a form that is possible to be estimated at new spatial locations [9]. This means that a new method of estimation is necessary that is able to combine the strength of local predictions of nonparametric models and the spatial significance of parameters that are observed in parametric models.

Comparison to Alternative Products

There are a number of extant climatology products for which near-continuous estimates are able to be generated, with two of the most popular being PRISM and TOPOWX [4, 12]. A shortcoming with these two products, and most others in the space of statistical climatology, is that spatial estimates are generated only as point estimates. There is a calculation of uncertainty in the estimates, but few current products in this domain are able to produce estimates of probability distributions at arbitrary spatial locations.

By generating estimates of densities rather than point estimates of precipitation we are able to generate a much broader range of statistics, including all quantiles. One area in which the use of this difference in estimation may be apparent is in estimating the risk for extreme precipitation, i.e. determining the 99% quantile at an arbitrary location. Given point estimates this is a challenging task; however given a full estimate for the distribution of precipitation this problem becomes rather straightforward.

Much of the work outlined here is ultimately done in the specific pursuit of this goal of generating distributions at arbitrary locations. The inferential advantages of having full distributions rather than point estimates are significant and we hope that this increased flexibility and utility of the model will provide new avenues of precipitation estimation for locations outside of observational networks.

Outline

Here we propose and develop a novel "low-rank kernel density estimator" for estimation of daily precipitation in the continental United States that combines the useful inclusion of spatially, temporally, and climatologically significant parameters while achieving the accuracy that has been previously been found with nonparametric estimators. Parameters for this model are estimated at the daily scale and regressed using a small number of harmonic terms leading to local models for which the parameters are significant both spatially and temporally.

With the local model in place an exploration of methods (including linear models, generalized additive models, and kriging) for spatial estimation at new locations is conducted and each is analyzed. Ultimately a spatial model using generalized additive models is constructed and the accuracy of spatial estimates using hold-one-out cross validation is discussed. This allows local models to be estimated at arbitrary locations for any point in the continental United States.

One major focus of precipitation studies has been the estimation and simulation of gridded precipitation, i.e. generating simulations at a fine grid of points covering a domain. Using the spatial model mentioned above, gridded simulation is covered briefly, and is included primarily as

a motivation for further research that we hope will stem from this project.

In conclusion the aims of the project are evaluated, a few directions for further research are covered, and the potential applications of the model as it stands are discussed.

1.1 The Data

The data presented here are provided by the United States Historical Climatology Network (USHCN) [11]. Data are recorded at 1218 precipitation stations distributed across the continental United States and are taken as daily measurements in units of 0.01 inches. We use data sourced over all days of the year (leap days excluded) during the years from 1900 through 2014. Data are flagged with quality and measurement controls to ensure accuracy. The quality flags include flags for failures in spatial and temporal consistency, and duplication and gap checks among a number of other outlier controls (more information available through the USHCN [11]). The measurement flags indicate the type of measurement recorded as well as indicating the presence of a "trace" precipitation event that was non-zero but below the 0.01 inch threshold of recording positive precipitation.

1.1.1 Quality Control

Prior to processing the data are scrubbed of any observations that failed any of the quality control metrics included in the dataset. The number of removed points for each station is shown in Figure 1.2, there are a small number of outliers but typically the number of points removed does not significantly change the number of observations in the data. This is a small enough change to not dramatically alter the data, and ensures that all observations are reflective of true precipitation.

1.1.2 Trace Precipitation

Among the measurement flags is the encoding of trace precipitation events. The threshold cutoff to record any positive precipitation is 0.01 inches and anything below that registers as 0 inches [11]. If there is any precipitation at all these 0 inch recordings can be accompanied by a trace flag indicating that some precipitation between 0 and 0.01 inches (recorded as 1 in units of

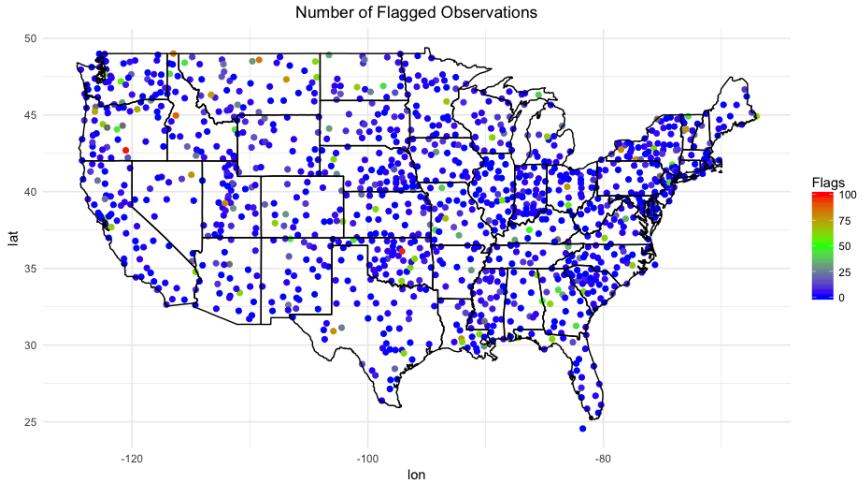

Figure 1.2: Number of observations removed for quality control from each recording site.

hundredth inches in the data) did occur. With only this information on the data we impute all trace precipitation events as independent Uniform(0,1) (in hundredth inches) random variables. The number of trace events per station is spatially structured and can be upwards of 10% of the total number of observations for any given recording site (Figure 1.3).

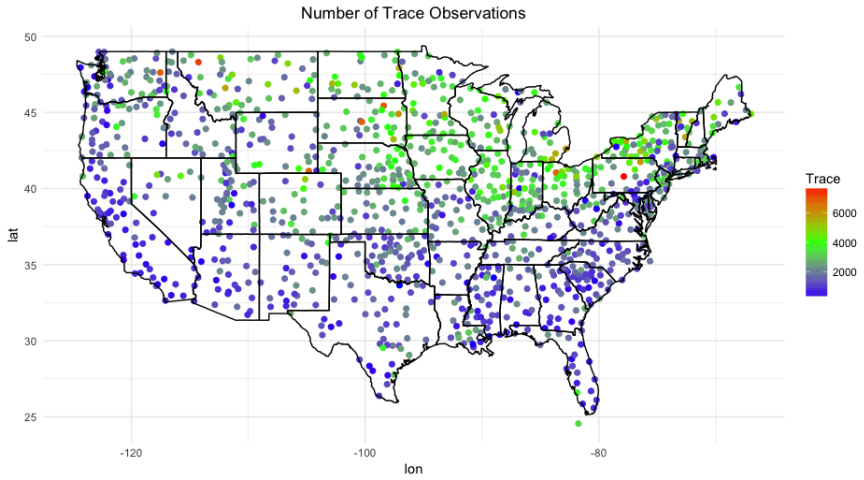

Figure 1.3: Number of trace precipitation events per recording site.

This imputation is motivated as a simpler version of more complex data correction processes that are common in precipitation literature [17, 18]. Without any knowledge of the distribution of trace precipitation, using uniform random variables to impute data requires the least number of assumptions to be made while still correcting for these important precipitation events.

From Figure 1.3 it is clear that these trace events are quite common for some regions of the United States, and in total reflect a major factor in precipitation as a climatological process. Constructing a model without the inclusion of these trace events would greatly reduce the accuracy of any simulations as they relate to true precipitation beyond the collection scale of the data.

A Note on Notation

Much of the work presented here revolves around generating estimators of precipitation for a "given day" of the year. In this context we assume that there are no larger trends in the data and that all observations of precipitation observed at a location for a day of the year are from the same distribution regardless of the year in which they were observed. Therefore all references to precipitation observed on a given day generally refers to all instances of precipitation on that day of the year across all recorded years of observation.

Chapter 2

Parametric Precipitation Models

The model presented here utilizes a more complex and intricate non-parametric density function than common and well-defined parametric distributions, thus to show the necessity of such a model we first discuss the drawbacks associated with using parametric distributions to model precipitation, namely the inability to fully capture local climatology and inaccuracies upon simulation.

There are a number of current precipitation models which are based on well-known parametric probability distributions with one of the most common choices of distributions for positive precipitation being the gamma, which is focused on here as the predominant candidate for a parametric precipitation model [9]. Dependent on the shape (α) and rate (β) parameters the density of a gamma random variable is defined as

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{2.1}$$

in which $\Gamma(\cdot)$ is the standard gamma function. The gamma distribution is flexible, and when x is used to represent log-precipitation the gamma distribution is heavy-tailed, which is necessary for accurate estimators of precipitation [16].

2.1 Maximum Likelihood Estimates

Proposed gamma distribution estimators are produced in multiple ways, all dependent on maximum likelihood estimation. Maximum likelihood estimators (MLEs) are produced by choosing the parameters of a distribution such that the likelihood of sampling the observed data (in this case

strictly positive precipitation) is maximized. Here two approaches to generating MLEs for daily precipitation for gamma distributions are examined and ultimately rejected.

2.1.1 Linear Models of Maximum Likelihood Parameters

The `R` library `fitdistrplus` allows for easy computation of maximum likelihood estimators (MLEs) for various distributions given a set of data. Using this library estimates for the rate and shape parameters for gamma distributions are generated at all stations for each day of the year. The seasonal dependence exhibited by these parameter estimates, seen in figure 2.1 motivate the use of a linear regression models based on sine on cosine terms. In order for the parameter estimates to be guaranteed to be non-negative (as is needed for both the rate and shape parameters of gamma distributions) linear regression is performed on the logarithm of the MLEs for both parameters. This method allows for exponentiation of the predicted log-valued parameter, generating a positive estimate for the true parameter; this is an extremely useful tool when strictly positive estimates of parameters are needed and will be used repeatedly throughout the development of this model. The assumed form of the log parameters is then

$$log(\alpha) = \sigma_0 + \sigma_1 \sin\left(\frac{2\pi t}{365}\right) + \sigma_2 \cos\left(\frac{2\pi t}{365}\right) + \sigma_3 \sin\left(\frac{4\pi t}{365}\right) + \sigma_4 \cos\left(\frac{4\pi t}{365}\right) + \epsilon \tag{2.2}$$

$$log(\beta) = \sigma_5 + \sigma_6 \sin\left(\frac{2\pi t}{365}\right) + \sigma_7 \cos\left(\frac{2\pi t}{365}\right) + \sigma_8 \sin\left(\frac{4\pi t}{365}\right) + \sigma_9 \cos\left(\frac{4\pi t}{365}\right) + \epsilon \tag{2.3}$$

where $t = 1, 2, \ldots, 365$ represents the day of the year, and ϵ is assumed to be a mean zero normally distributed random variable. These regression models allow estimates of $\boldsymbol{\sigma}$ to be generated using least-squares regression and reduce the number of coefficients to a few meaningful values, which is necessary when producing a model that can be predicted at new locations for which observations have not been recorded since all parameters will need to be predicted spatially.

The daily resolution MLEs and associated regression coefficients can be rapidly generated for all locations however upon sampling from the distributions produced it is clear that the gamma distributions are not representative of the empirical distributions of the data. This is shown for randomly chosen days for multiple recording stations.

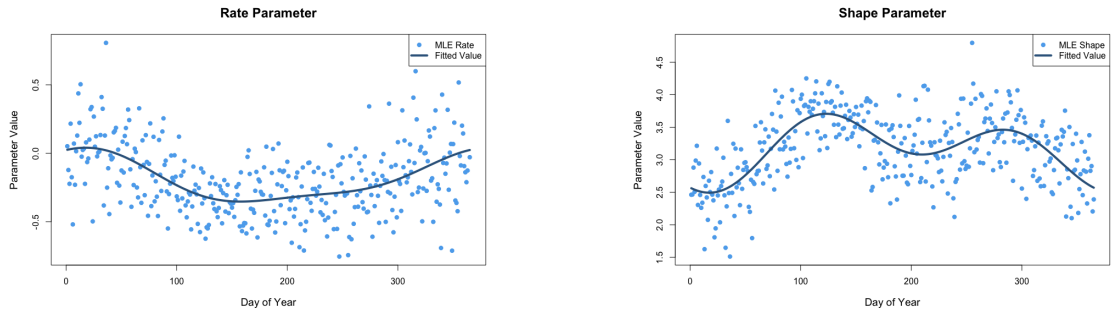

Figure 2.1: Parameter MLEs and fitted regression values for Boulder, CO

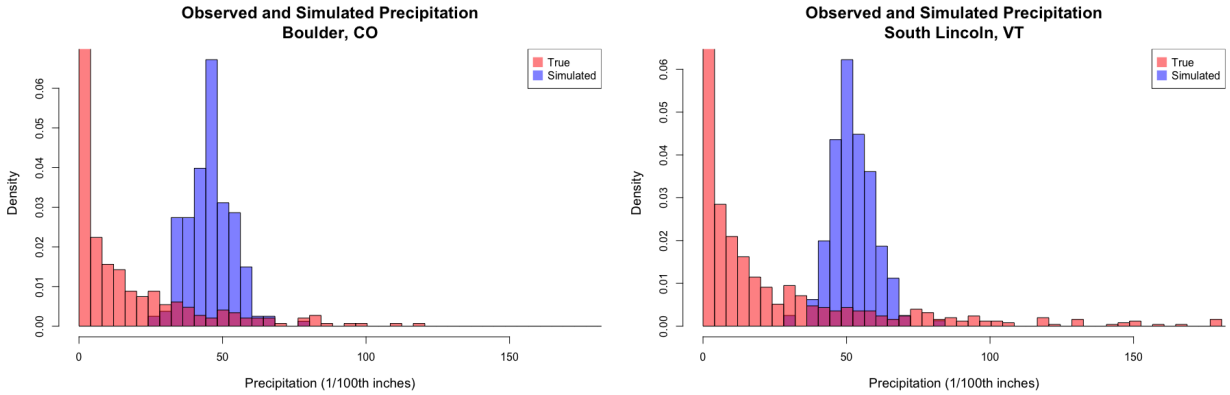

Figure 2.2: Simulated and Observed Precipitation for Boulder, CO and South Lincoln, VT

The estimators resulting from this method produce distributions whose means often agree with observed data, but whose distributions are not representative of the observed process.

2.1.2 Nonlinear Optimization of Log-Likelihood

As an alternative to estimating regression coefficients from maximum likelihood estimates of parameters to gamma distributions we attempt to estimate the full array of σ coefficients in equations 2.2 and 2.3 simultaneously by minimizing the negative log-likelihood using built in non-linear minimization functions in `R` [13]. In this method we assume that the optimal rate and shape parameters for a gamma distribution take the form

$$\alpha = \exp\left\{\sigma_0 + \sigma_1 \sin\left(\frac{2\pi t}{365}\right) + \sigma_2 \cos\left(\frac{2\pi t}{365}\right) + \sigma_3 \sin\left(\frac{4\pi t}{365}\right) + \sigma_4 \cos\left(\frac{4\pi t}{365}\right)\right\} \tag{2.4}$$

$$\beta = \exp\left\{\sigma_5 + \sigma_6 \sin\left(\frac{2\pi t}{365}\right) + \sigma_7 \cos\left(\frac{2\pi t}{365}\right) + \sigma_8 \sin\left(\frac{4\pi t}{365}\right) + \sigma_9 \cos\left(\frac{4\pi t}{365}\right)\right\} \tag{2.5}$$

and the negative log-likelihood of observing the recorded precipitation data given rate and shape parameters α and β is minimized over the vector of coefficients σ simultaneously. This has the benefit over the method described in section 2.1.1 that all necessary coefficients are found at once rather than needing to calculate MLEs of α and β and perform linear regressions.

The estimators produced in this fashion are more closely matched to the data, however there are still substantial shortcomings that prevent moving forward with this approach. Figure 2.3 shows histograms of simulated and observed data for a given day of the year. The two main areas of shortcoming seen in these plots are the inability of the estimators to capture the inclusion of trace precipitation events and the tails of the gamma distributions that cannot capture the heavy-tail of precipitation [16].

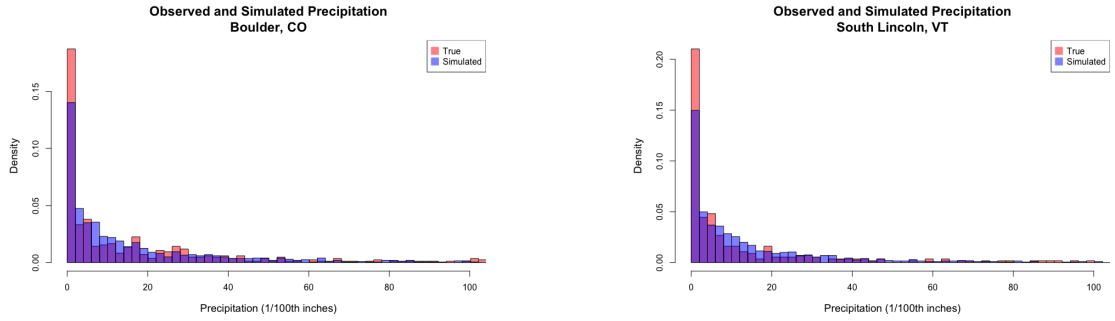

Figure 2.3: Simulated and Observed Precipitation for Boulder, CO and South Lincoln, VT

This parametric exploration serves as an anchoring for the research and to increase familiarity with the data. Furthermore, with the shortcomings of the most common parametric models identified the development of a novel model for estimating precipitation is motivated and justified.

Chapter 3

Locally Generated Precipitation Model

The first step in the development of framework for large-scale precipitation estimation and simulation is to develop a local model that accurately reflects the observed data and is comprised of meaningful parameters that can be used to predict estimators at new spatial locations. We propose the use of a "low-rank kernel density estimator", a modified kernel density estimator (KDE) constructed using a reduced number of weighted kernels rather than placing one kernel function at each data point. Here we refer to the location of these kernel functions as "knots". This low-rank estimator provides increased computational efficiency over traditional kernel density estimation and is constructed with statistically and climatologically significant parameters which make the development of a spatial model of these estimated distributions in later chapters a tractable problem. An outline of kernel density estimators and their formulation here is covered in appendix A.

The model in developed in this chapter estimates only positive observations of precipitation transformed to log-space. Thus, unless otherwise noted, the units for precipitation moving forward are in log $1/100^{th}$ inches. This transformation follows from past precipitation models that make the same transformation and is particularly useful in handling extreme precipitation observations and tail densities as the transformation forces extreme events occur nearer to the centers of the estimated distributions [9].

3.1 Low-Rank Kernel Density Estimation

The form of the proposed kernel density estimator using a small number of knots takes the form,

$$f(x) \propto \sum_{k=1}^{n} \sigma_k K\left(\frac{x - x_k}{h}\right). \tag{3.1}$$

In this formulation we use the following definitions:

- $x \sim$ quantity to be estimated (log $1/100^{th}$ inches),
- $n \sim$ number of knots used,
- $\sigma_k \sim$ weight attached to the k^{th} knot,
- $K(\cdot) \sim$ the Gaussian kernel function,
- $x_k \sim$ the k^{th} knot,
- $h \sim$ the bandwidth parameter.

The discussion of how these parameters are estimated and which are dependent on space and time in which ways and which are independent follows below. Note that a brief discussion of the selection of kernel function and bandwidth parameters is included in appendix A.

The generation of a low-rank KDE is a primarily a two-fold problem: determining the location and number of optimal knots and the optimal weights to attach to these knots (as is shown below there are other details to handle but these two are of principal importance). We claim that the optimal combination of weights and knots are those that minimize the integrated quadratic distance between the proposed cumulative distribution function (CDF) of the estimator and the empirical cumulative distribution function as determined by the data [15]. Thus, with $\hat{F}(\cdot)$ as the proposed estimator and $F_{emp}(\cdot)$ as the empirical CDF we attempt to minimize

$$d_{IQ}(\hat{F}, F_{emp}) = \int_{-\infty}^{\infty} (\hat{F}(t) - F_{emp}(t))^2 dt. \tag{3.2}$$

The empirical CDF is calculated in the standard way, where

$$F_{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\boldsymbol{x_i} \leq \boldsymbol{t}}. \tag{3.3}$$

There is an underlying chicken-and-egg problem here; portions of the model must be developed (i.e. formulation of the weights and bandwidth parameters) before the sets of knots to be used can be robustly selected, however sets of knots must be selected upon which estimates of these parameters can be generated (since a weight is necessarily "attached" to a specific knot). Some trial and error and experimentation with fully formed kernel density estimators as described in equation A.1 allows us to make a reasonable estimate as to a set of knots to be broadly applied to all stations and days of the year to begin constructing the model in full. It is found that choosing 9 knots evenly spaced between 0 and 6 in the log $1/100^{th}$ inch precipitation space is a dense enough set of points that high degrees of accuracy can be achieved, but is small enough in number that computations are not hindered. Upon building a model that satisfies the basic needs of accuracy and efficiency we return to the issue of knot selection (section 3.1.2) to determine a method of choosing optimal low-rank knots.

3.1.1 Model Construction

This section outlines the procedure that is followed to generate the final form of the estimator used in further aspects of the project. Following this discussion the model is restated formally and the pipeline of methods that could be used to generate similar estimators for any application is explained.

3.1.1.1 Bandwidth Parameter

As described in appendix A.1, we employ pre-made methods for determining the bandwidth for a given day and location via minimizing the asymptotic mean integrated squared error. To do this we take 15 day rolling windows of data (one week to before and after the date being examined) for all observed years and calculate the optimal bandwidth, h_{amise} as the bandwidth that will

be used in the low-rank approximation to the KDE. These optimally computed values are highly seasonally structured, and thus least-squares linear regression on a small number of harmonic functions can be used to capture a substantial amount of the variation and greatly reduces the parameter space of the model over the course of a year. It is assumed that the optimal bandwidth parameters for a given location behave according to the following:

$$h_{amise}(t) = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{365}\right) + \beta_2 \cos\left(\frac{2\pi t}{365}\right) + \beta_3 \sin\left(\frac{4\pi t}{365}\right) + \beta_4 \cos\left(\frac{4\pi t}{365}\right) + \epsilon. \tag{3.4}$$

In 3.4 t represents the day of the year and ϵ represents a Gaussian error term. This form allows estiamtes of the β parameters to be calculated using least-squares regression. Examples of the resulting optimally calculated bandwidths and fitted bandwidths from the regression model above are shown in Figure 3.1.

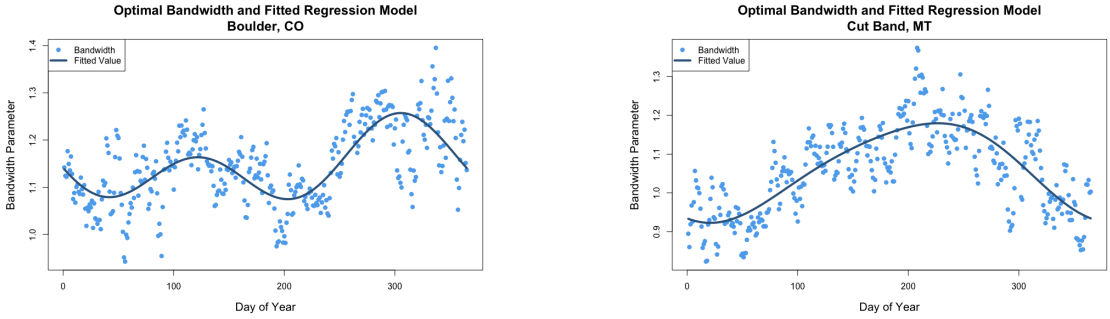

Figure 3.1: Optimally calculated bandwidth and fitted values from regression model

Not only does this regression provide a great reduction in the number or parameters required for each recording location, but the regression coefficients from 3.4 exhibit good spatial structure which is necessary for generating estimators at new locations once the local models are completed. Note that these are not exceptionally good regression predictions, however the models are not exceptionally sensitive to bandwidths and the variation over the course of a year is not large enough to cause unacceptable errors.

3.1.1.2 Optimal Weights

With temporary knots in place and bandwidth parameters calculated we move to determining the optimal weights, the σ_k terms in equation 3.1. Using the built-in `R` method `optim` we minimize 3.2 using a quasi-Newton iterative method under the constraint $\sigma_k \geq 0 \quad \forall k$, ensuring that the resulting PDF is nonnegative, for each day and each location (more specific information on the optimization method can be found under `optim` in the `R` documentation)[13]. This is one of the most computationally expensive components to the development of the model, thus some simplifications are necessary to compute optimal weights for each location.

We produce PDFs at the daily resolution for each location, thus 365 sets of weights must be produced for each recording station. As with the bandwidth parameters in section 3.1.1.1 we consider 15 day rolling windows of data with which to construct the empirical CDF. Performing this optimization for a single station and examining one of the weights over the course of the year it is evident that there is strong seasonal structure which can be captured in a few regression terms, this is shown for the first weight for the Boulder, CO recording station in Figure 3.2.

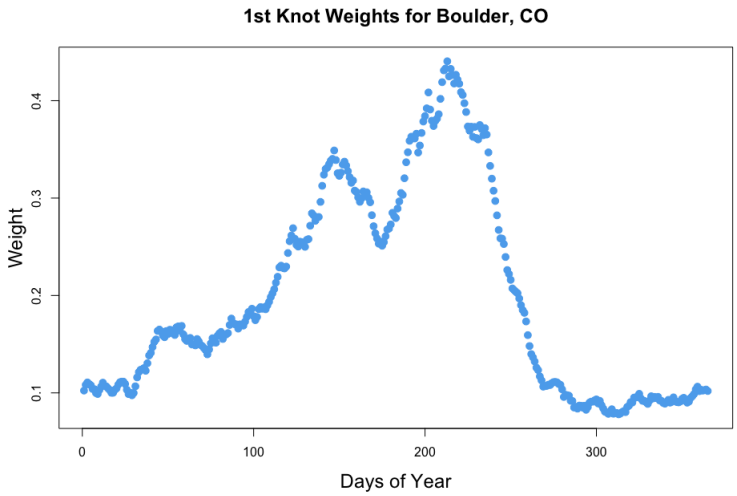

Figure 3.2: Optimal weights for the first knot for Boulder, CO

This structure motivates simplifications to both increase the speed with which models can be constructed for new recording locations, and another decrease in the parameter space that will

be necessary in the development of an efficient spatial model. By capturing the seasonality of the weights in a simple regression model the optimal weights need to be calculated for only a sparse sampling of days over the course of the year, thus decreasing the time to compute the full set of weights for a location and allowing the full year of weights for each knot to be contained in a small set of meaningful regression parameters rather than 365 weight terms.

Additionally the weights are constrained to be positive, so we form the regression on the logarithm of the parameters; this accentuates the the seasonal variance and typically leads to slightly better model fits as determined by R^2 values. Calculating weights for only every fifth day and assuming a seasonal model on all the weights according to,

$$log(\sigma) = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{365}\right) + \beta_2 \cos\left(\frac{2\pi t}{365}\right) + \beta_3 \sin\left(\frac{4\pi t}{365}\right) + \beta_4 \cos\left(\frac{4\pi t}{365}\right) + \epsilon \tag{3.5}$$

where t is the day of the year and ϵ is a Gaussian error term leads to the result shown in Figure 3.3 where the β terms are calculated via least-squares regression.

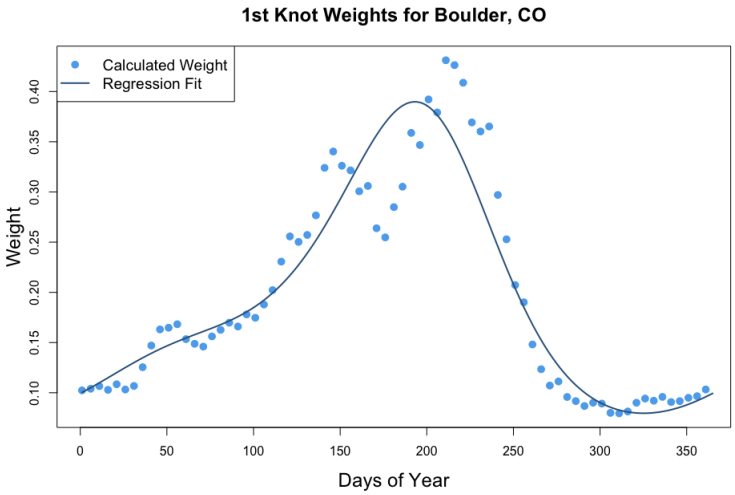

Figure 3.3: Sparsely sampled optimal weights and regression line (exponentiated fitted values) for the first knot for Boulder, CO

This procedure is done for all 1218 recording locations, and with the bandwidth regression coefficients calculated (equation 3.4) we have preliminary estimates for distributions of positive precipitation for all recording locations for all days of the year. These estimators are generally

very accurate, however discussion of this is avoided here as more alterations will be made prior to final forms of the estimators being decided upon and analyzed. With the model parameters determined and methods of estimation in place the process of selecting the optimal sets of knots for each location can begin in earnest.

3.1.2 Selection of the Knots

With preliminary estimates for all space-time combinations generated, we move to considering the quadratic divergence (equation 3.2) as a function of both the knots and weights in equation 3.1. As is shown below the choice of knots is a much less sensitive problem than that of selecting the values of the weights. This turns out to be fortunate since there is a far less useful formulation for determining knots than there is for determining weights given sets of knots.

The optimal weights for each knot can only be found once the optimal knots are fixed. With knots fixed and optimal weights found the quadratic divergence (equation 3.2) between the theoretical estimate of the CDF and the empirical distribution function as generated by the data is calculated for each day of the year. Data to calculate the empirical CDF are taken from a 15 day window centered at the day being investigated over all years of observation (i.e. if the date of investigation is July 20^{th} we include data from July 13^{th} through July 27^{th} from all years of observation). Without a well defined "cost function" with which to optimize knots we resort to proposing sets of knots, building estimators, and taking the set of knots that minimizes the quadratic divergence between the estimator and the empirical CDF. Through investigation and trial and error we find reasonable estimates for the location of the knots in log-precipitation space ($1/100^{th}$ inches) to form initial proposal sets of knots with which to fit models for a sampling of locations.

3.1.2.1 Number of Knots

Initially an acceptable number of knots to include in the model is determined using the following procedure. For a variety of potential values for the total number of knots in the model

(4, 5, 6, or 7), potential sets of knots are constructed (approximately 7 sets of equispaced knots with differing values for the maximum knot), the optimal weights that minimize the divergence in equation 3.2 for each set of knots are found, and the average divergence over the course of a year for each potential model is stored. Then for each station and each potential number of knots we extract the model with the minimum average divergence over the course of a year.

Upon comparison we see that the number of knots does not have a major impact on the accuracy of the model as quantified by quadratic divergence. A boxplot comparison of quadratic divergences by the number of knots used is shown in Table 3.1. Seen in this Figure is that using 4 knots produces slightly less accurate models, but there is little increase in accuracy moving to more complex models. In the interest of accuracy and computational efficiency we move forward using 5 knots for all stations.

Number of Knots	Mean Quadratic Divergence
4	511.33
5	447.76
6	433.40
7	438.70

Table 3.1: Mean yearly divergence for a sampling of stations by number of knots

3.1.2.2 Location of Knots

With the number of knots chosen we turn to determining the optimal sets of knots across all spatial locations. This is a computationally challenging problem and a number of simplifying assumptions are taken here; improvement in the choice of knots for each station could be a potentially rich inroad for future development of this model. To make the problem computationally tractable we assume that the knots will be equally spaced in the domain, thus, given that 5 knots are used, we can construct a full set of knots given just two values: the last knot, and the distance between the first and last knots.

We begin by assuming that knots do not vary over time. Since a set of knots must be proposed,

then a model built around it and evaluated for accuracy, letting the knots vary seasonally would increase the space of potential sets of knots too greatly to be computationally feasible. Thus we turn to first determining if the knots should be fixed distances apart or if different stations should have knots spread different distances apart.

Since calculating divergence for a a number of proposed knot combinations is a time intensive process, a random sample of 200 stations is taken and divergence is calculated for sets of knots. The proposed sets of knots are all comprised of 5 equispaced knots (as per section 3.1.2.1) with maximal knots of 2, 2.5, 3, 3.5, 4, 4.5, and 5, and with the distances between the maximum and minimum knots of 3, 3.5, 4, 4.5, and 5. There are 7 options for maximum knots, and 5 options for total spreads, giving 35 sets of knots per station total. It is found that the value of the maximum knot has a much more significant impact on the goodness of fit as determined by quadratic divergence than the overall spread between knots.

What is found is that by averaging the best case divergence for each potential spread of knots (i.e. the spacing between the largest and smallest knots) over all stations the average divergences ranges between about 550 and 470. However, when averaging the best case divergence for each potential maximum knot over all stations the divergences range between almost 4700 and about 500. This, along with the computational issues that come with permitting an additional degree of freedom in allowing the spread of the knots to vary spatially, motivates fixing the spread of the knots for all recording stations and reduce the calculation to only computing the value of the maximum knots for each location.

From these samplings it is found that using a distance of 5 units of log-precipitation between the maximum and minimum knots produces the best results, so moving forward all local models are built using 5 equispaced and spatially varying knots $\{x_1(s), x_2(s), x_3(s), x_4(s), x_5(s)\}$ such that $x_5(s) - x_1(s) = 5$. Thus all that must be computed to determine the full set of knots for a given location is the value of the maximum knot.

Calculating Maximum Knots

An ad hoc method is used to calculate the values of the maximum knots, which is the only

parameter that is needed since knots are fixed by location and the spread of the knots is constant. Even with the simplifying assumptions from above there is not currently a useful formulation to properly optimize this parameter in terms of distribution matching, therefore again sets of knots are proposed and evaluated, and the best of these sets are taken. From inspecting the results of the above process it is found that the span of the best maximum knots can typically be bound between 1.5 and 5.5 (in units of log-precipitation), thus the sets of proposed knots are taken using maximum knots in this space.

Using five proposal sets models are constructed and evaluated for accuracy (in terms of quadratic divergence, equation 3.2) for all stations sampled every 5 days through the year. From here it is easy to say which of these proposed sets of knots is optimal for a given location, however this essentially discretizes the maximum knot parameter since it can only be drawn from one of five potential values.

To combat this a smoothing process is applied to each station in which a linear regression model of the discrete maximum knots from above is constructed without the station's parameter and a new value is predicted using the model. The predictions take the form

$$\hat{y} = \beta_0 + \beta_1 \cdot \text{lon.} + \beta_2 \cdot \text{lat.} + \beta_3 \cdot \text{lon.} \times \text{lat.} + \beta_4 \cdot \text{elev.} + \beta_5 \cdot \text{slope} + \beta_6 \cdot \text{aspect} \tag{3.6}$$

in which the β's come from the standard least-squares regression model. Note that there is further discussion on the specific formulation of the elevation, slope, and aspect values in chapter 4. This produces both smooth and continuous estimates of the maximum knots, which is how precipitation is expected to behave. From these estimates of maximum knots full sets of knots are constructed, and all weighting terms and least-squares seasonal regressions according to equation 3.5 are recalculated for the final time.

This is not an ideal process, however this is a novel issue in estimation and there is no established method of solving for the knot parameters in 3.1 efficiently. With this in mind it is worth pointing out that the selection of the knots is far from the most important consideration in this process and the accuracy of this method (discussed below) is impressive and motivates that

this formulation of the knot parameters is sufficient to move forward with model development.

To clarify before moving forward a few plots are displayed. Shown in Figure 3.4 is the update from discrete draws to continuous estimates.

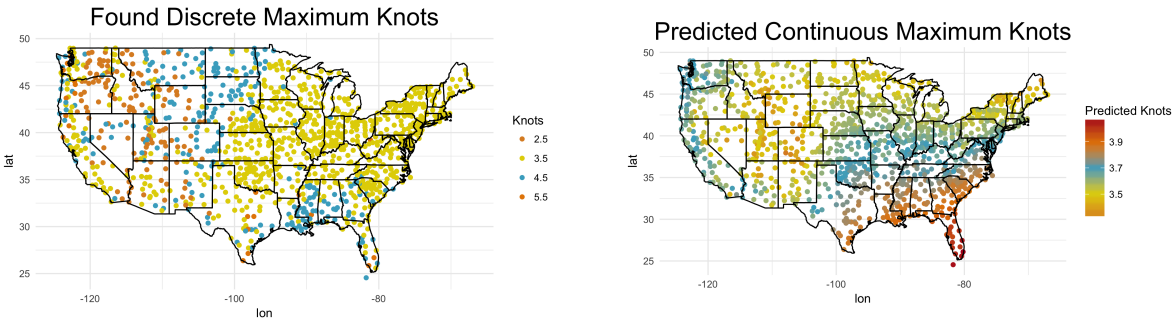

Figure 3.4: The initially found knots drawn from a discrete set of options (right) and the updated continuous knot estimates (left). Note the scale changes between plots to make differences more visible on the continuous scale.

These plots serve to clarify the process through which the maximum knot estimates are found, and although this process is not ideal the knots generated are capable of highly accurate estimation and the maximum knots exhibit significant spatial structure which is useful in the development of the spatial model in chapter 4.

3.2 Bandwidth Alteration

Upon simulation from the model outlined above, with optimal weights and optimal knots chosen, we see substantial over-estimation of density in the tails of the distributions. This issue arises as an artifact of estimating log-precipitation then transforming back to standard precipitation space; the tail density is dominated by the greatest knot in the low-rank KDE (equation 3.1). With the standard bandwidth chosen and a Gaussian kernel function the tail of the log-precipitation esti-

mator decays too slowly and this issue is exaggerated upon exponentiation to the full precipitation space.

This is most readily seen in quantile-quantile plots comparing simulations drawn from the proposed estimators and data taken from 15 day rolling windows of data. Examples of these quantile-quantile plots are shown for random locations and random days in Figure 3.5.

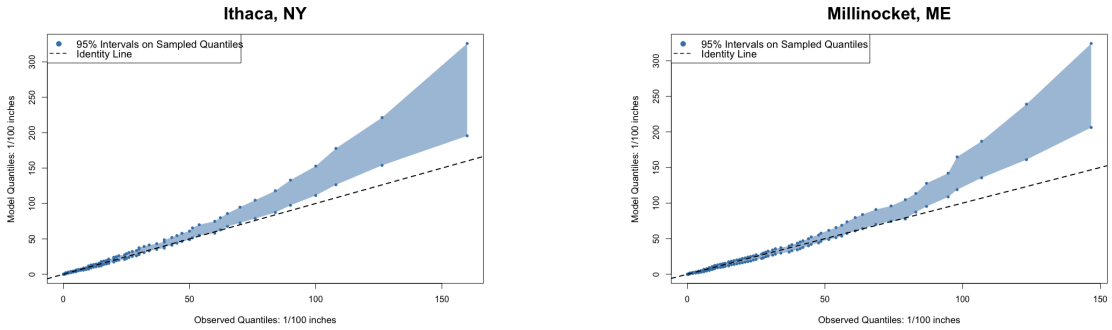

Figure 3.5: Quantile-quantile plots showing over-simulation of large-scale precipitation.

In these Figures the shaded region represents the area covered by 95% of quantiles of simulated precipitation taken from the modeled estimator. This trend is consistently seen across recording stations, and causes the mean simulated precipitation to be over-estimated. A comparison of observed and simulated means for the same stations shown in Figure 3.5 is shown in Figure 3.6.

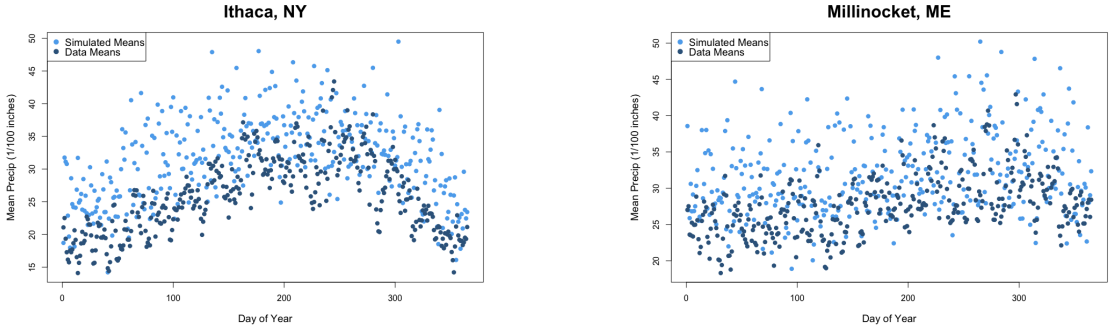

Figure 3.6: Mean estimates by day exhibiting over-simulation of precipitation in the model.

This motivates decreasing the magnitude of the bandwidth used to calculate the kernel function at the largest knot. Seeing Figure A.1 it is seen that decreasing the magnitude of this bandwidth

will cause the density in the tail of the distribution to decay more quickly and will remedy the issues seen in Figures 3.5 and 3.6. Thus a term is added into the bandwidth formulation that acts as a decay term on the terminal bandwidth to remedy the issues seen in the tail of the distribution. The new form of the estimator becomes

$$f(x|s,t,\boldsymbol{\eta}) \propto \sum_{k=1}^{5} \sigma_k(s,t) K\left(\frac{x - x_k(s)}{\eta_k(s,t)h(s,t)}\right), \tag{3.7}$$

where all the definitions from equation 3.1 hold, and $\eta_1, \ldots, \eta_4 = 1$ and η_5 is a decay applied to the bandwidth of maximum knot kernel function.

3.2.1 Calculating Bandwidth Decay

Motivated by the behavior seen in Figure 3.5 the integrated distance between quantiles of the data and the low-rank kernel density estimator is used to calculate the decay term applied to the bandwidth of the largest knot. The optimal weight for the last bandwidth is then

$$\boldsymbol{\eta}_{opt}(s,t) = \underset{\eta>0}{\operatorname{argmin}} \int_0^1 (\hat{F}^{-1}(x|s,t,\boldsymbol{\eta}) - F^{-1}_{emp}(x|s,t))^2 dx \tag{3.8}$$

where $\hat{F}^{-1}(\cdot)$ is the quantile function of the low-rank estimator in equation 3.7 and $F^{-1}_{emp}(\cdot)$ is the estimated quantile function as calculated from the empirical CDF of observed data over a 15 day rolling window of data. $\boldsymbol{\eta}$ represents the vector $\{\eta_1, \ldots, \eta_5\}$ where, as above, $\eta_1, \ldots, \eta_4 = 1$ and η_5 is the decay multiplier of the last bandwidth.

Calculating the values of η is costly, and we follow the same practice in section 3.1.1.2 in which values are calculated for a sparse sampling over the course of a year and a linear regression over a four harmonic terms is performed. This leads to a decrease in model parameters and an increase in the spatial significance of terms that will be predicted at new locations. We assume the decay terms follow the model,

$$\eta(s,t) = \beta_0(s) + \beta_1(s)\sin(\frac{2\pi t}{365}) + \beta_2(s)\cos(\frac{2\pi t}{365}) + \beta_3(s)\sin(\frac{4\pi t}{365}) + \beta_4(s)\cos(\frac{4\pi t}{365}) + \epsilon \tag{3.9}$$

where s and t represent spatial location and time of year respectively and ϵ is a Gaussian error

term. This allows the β parameters to estimated by least-squares linear regression. These models are typically extremely accurate, an example of which is shown in Figure 3.7.

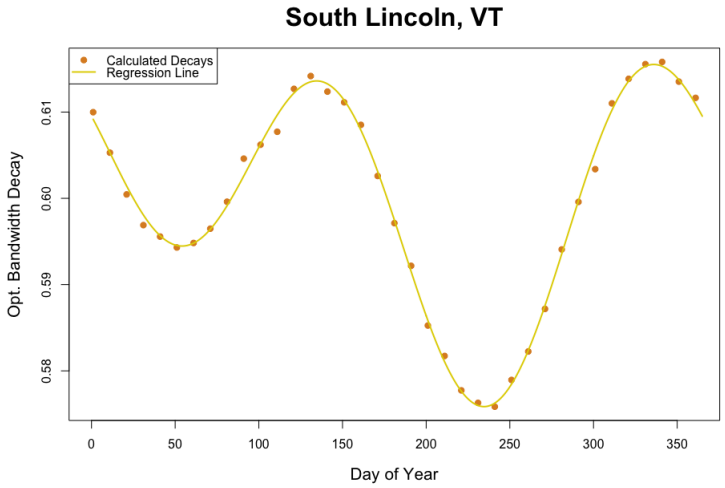

Figure 3.7: The calculated decay terms (shown as points) and the associated regression line from equation 3.9

With these parameters calculated for all recording stations the local models the development of the local models is complete, and the results of including these decay terms in the estimators are shown and discussed in section 3.4.

3.3 Final Form of the Estimator

With all decisions made regarding the calculation of the weighting terms, and the knots, and the inclusion of a decay term in the bandwidth the proposed form of the estimator then become,

$$f(x|s,t) \propto \sum_{k=1}^{5} \sigma_k(s,t) K\left(\frac{x - x_k(s)}{\eta_k(s,t)h(s,t)}\right), \tag{3.10}$$

where s represents a spatial location and t represents time (in terms of day of the year), and the parameters are as follows:

- $\sigma_k(s,t)$ is the weight attached to the k^{th} knot, seasonally and temporally dependent, calculated from regression coefficients,
- $x_k(s)$ is the k^{th} knot for location s, calculated from the maximal knot,

- $\eta_k(s,t)$ is the decay attached to the k^{th} bandwidth (only the 5^{th} is not 1), varying over space and time and calculated from regression coefficients,

- $h(s,t)$ is the bandwidth for a given location and date, calculated from regression coefficients.

It is also worth noting at this point that we do not include a discussion of the normalization constant for this estimator. For simulation from these estimators (which is ultimately the primary use) it suffices to numerically calculate a CDF from an estimated and normalize so that the CDF has 1 as its maximum value. From there simple inverse CDF methods can be used to draw samples.

Recap of Procedure

To clarify the process with which to construct this model for future reference this is the basic outline of the procedure undertaken:

- $h(s,t)$: Calculate the bandwidth parameters for all days and all locations (section A.1) and fit regression models (equation 3.4) to estimate all bandwidths for all days with only 5 regression parameters at a given station. Section 3.1.1.1

- $x_k(s)$: Here the number of knots and the spread between the knots is held constant for all locations and for any given station the knots are held constant for all days. Thus the knots are fixed in the log-precipitation domain based on the location of the maximum knot by considering potential models and selecting the one that minimizes the integrated quadratic divergence (equation 3.2). Section 3.1.2.2

- $x_k(s)$: Use these found discrete samplings of knots and hold-one-out regression models to generate continuous and spatially smooth estimates of the maximum knot values at all locations. Section 3.1.2.2

- $\sigma_k(s,t)$: With the final values of the knots for each station fixed calculate optimal weights (quadratic divergence minimizers) for each day and each location, fitting 5 parameter regression models to the logarithm of the weights (equation 3.5). Section 3.1.1.2

- $\eta_k(s,t)$: By minimizing the divergence of the quantiles of the estimators and the empirical quantiles of the data calculate a decay for the bandwidth attached to the last kernel function to prevent over-estimation of tail probabilities and form 5 parameter regressions according to equation 3.9. Section 3.2

3.4 Accuracy Verification

This model is capable of robust local estimation that captures a significant amount of climatological information and is highly accurate upon resampling. Shown below are comparable Figures to 3.5 and 3.6, generated for the same recording stations, but for the model after the inclusion of a dampened bandwidth in the last kernel function.

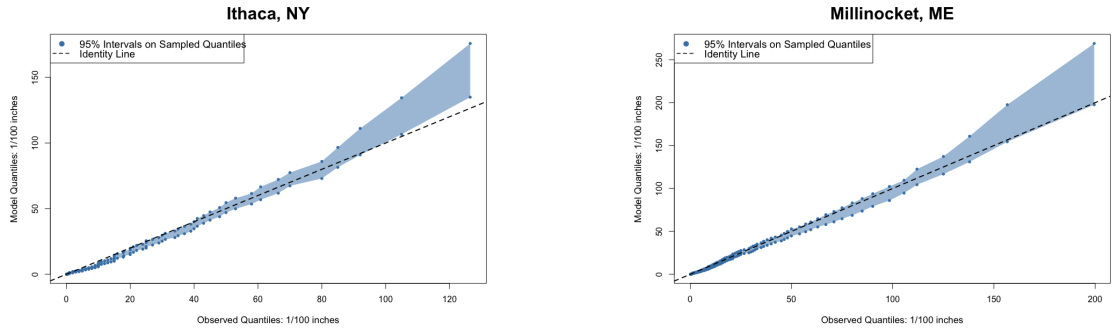

Figure 3.8: Quantile-quantile plots comparing the model to observed positive precipitation (95% confidence intervals shaded).

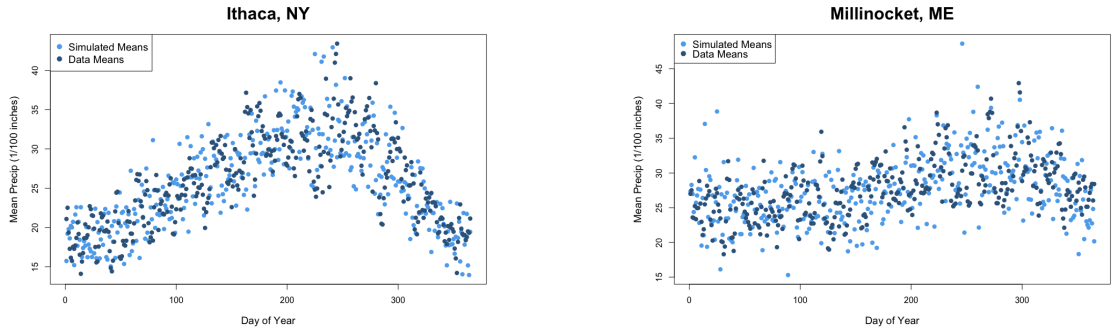

Figure 3.9: Mean estimates by day of both the model and observed positive precipitation.

It is clear from these Figures that the adjustment of the terminal bandwidth corrects the

biases in the model seen in Figures 3.5 and 3.6. Examining all quantile-quantile plots, for all locations and all days we find that over 85% of the empirical quantiles fall within the 95% intervals of quantiles of the estimators. Plots showing quantile comparisons of climatologically diverse stations for a set of days over the course of the year are shown in Figure 3.10. This Figure shows that not only are local estimates of climatology accurate, but there is a high degree of seasonality seen in precipitation data that is captured by the estimators. With acceptable accuracy at both the daily and yearly resolution we move forward in developing a spatial model for which localized models can be predicted at new spatial locations.

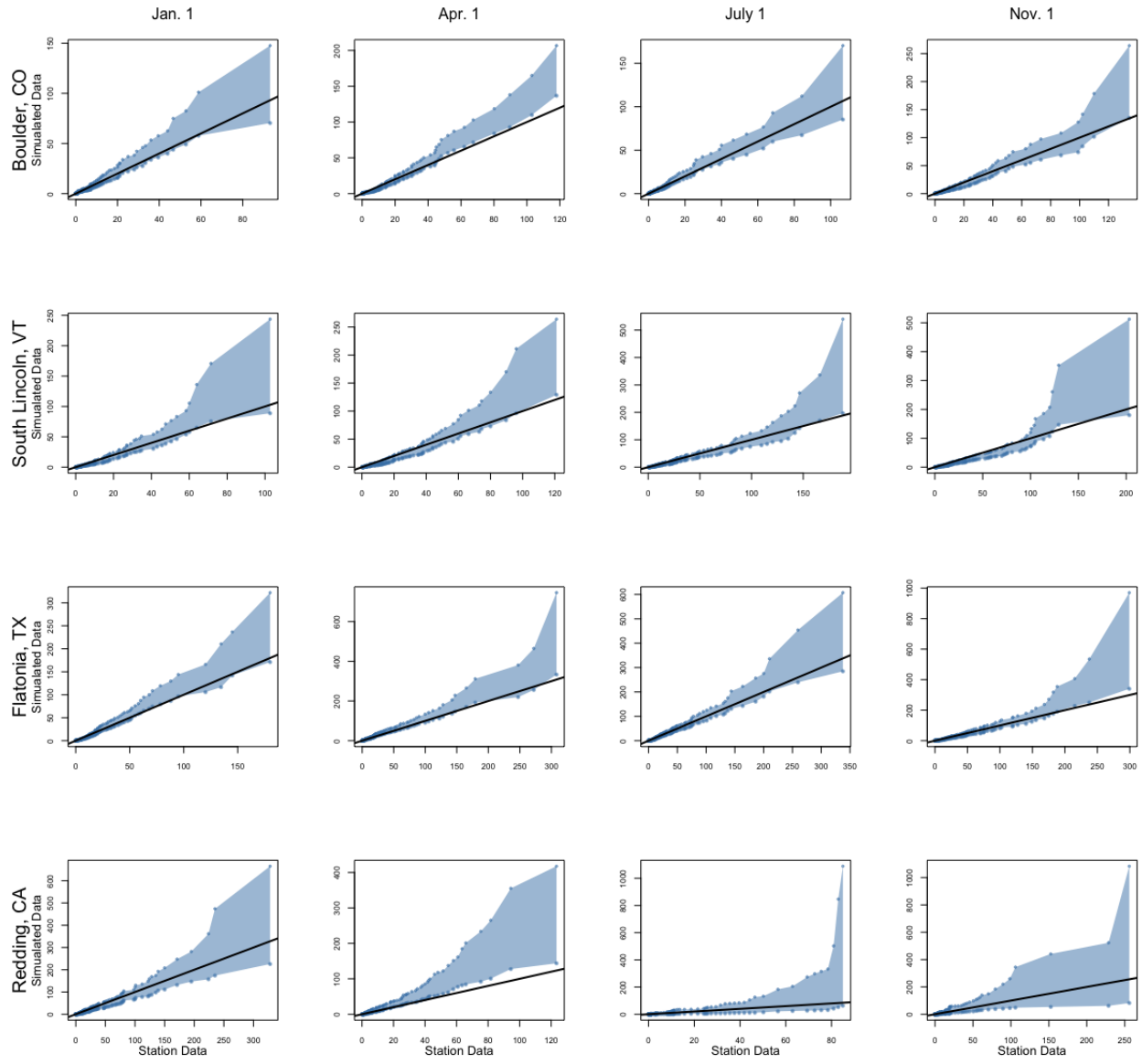

Figure 3.10: This plots show a comparison of estimated to observed quantiles for a diverse set of stations through varying seasons of the year. There are still outlying points in the tails for some space-time pairs, however overall we see that the 95% bounds on the estimators typically cover almost all of the observed quantiles. (Quantiles are in 1/100th inches)

Chapter 4

Predicting Local Models at New Locations

With a model for producing individual estimators on the daily scale for each recording location developed using parameters exhibiting significant spatial structure, a sufficient spatial model is constructed to predict distributions of positive precipitation at new locations for which data have not been recorded. For the construction of these models there are a number of selections that must be made, including the modeling of the underlying mean trends, the covariance functions and behavior, and how these will be used to predict parameters, and thus distributions, at new locations. Multiple spatial models are constructed and examined, ultimately resulting in the selection of using generalized additive models (GAMs) to model parameters and excluding the use of any covariance function or kriging process.

We begin here with a brief discussion of the predictor variables used in the regression models below, explaining the heuristics behind the methods used and pointing to further reading along the way.

4.1 Predictor Variables and Assumptions Made

As can be seen below in Figure 4.1 there is significant structure to the model parameters that is able to be captured in a variety of linear models. The constraints on the selection of variables upon which regression can be performed is that the information must be available at locations for which data have never been observed. This makes geophysical traits such as longitude, latitude, elevation, slope, and aspect ideal predictors as these are data that are readily available across the

United States.

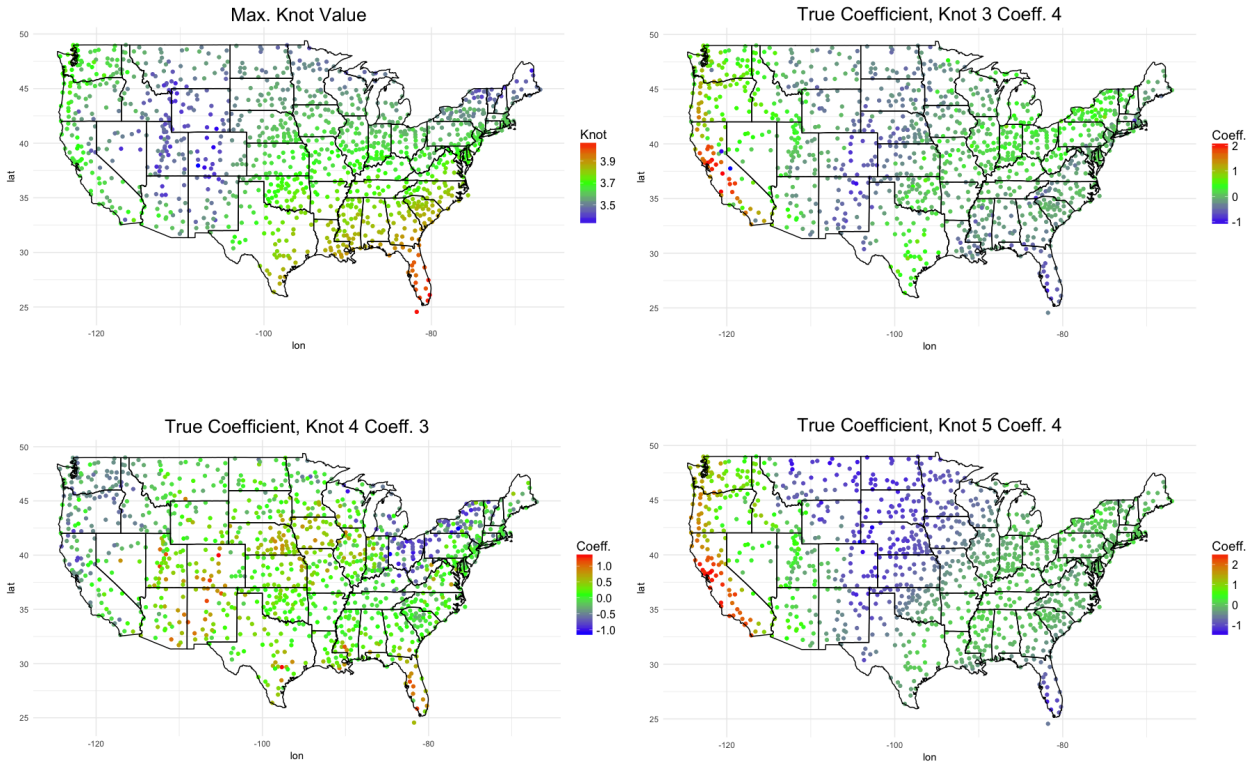

Figure 4.1: Examples of spatial plots of the maximum knot parameters (top left) and weight parameter regression coefficients.

Current precipitation models make use of these features alone and in combinations, specifically in the case of mountainous domains, and here the same protocol is followed [4]. Latitude and longitude of maintained stations are provided in the USHCN dataset, while elevation is taken from the United States Geological Survey's Global 30 Arc-Second Elevation dataset (GTOPO30, data available from the U.S. Geological Survey) and is used to calculate slope and aspect. Elevation, slope, and aspect are retrieved at the 1km resolution which must be upscaled to produce meaningful results as related to precipitation. Shown in Figure 4.2 we upscale to a 25km resolution, which captures sufficient information about the location while being on a large enough scale to be meaningful on the scale of precipitation.

This upscaling becomes an important consideration when the aspect of terrain (direction

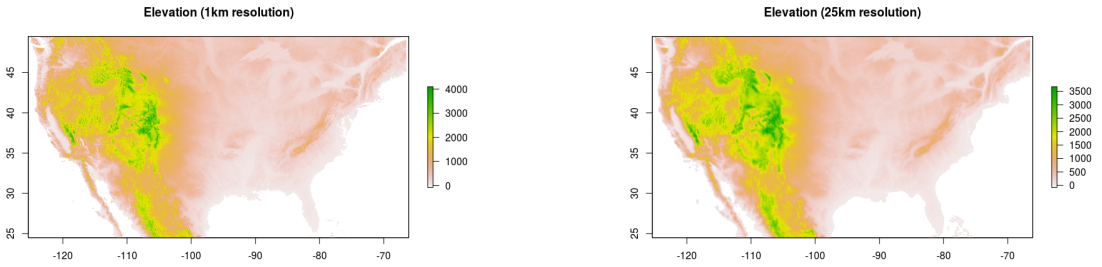

Figure 4.2: Elevation in meters at a 1km resolution (left) and 25km resolution (right)

faced in radians) is considered as a regression parameter. With a resolution too fine a map of aspect over the United States is extremely rough and does not reflect the portion of aspect that is relevant when considering precipitation, namely orographic effects in mountainous regions [3]. Aspect over the United States can be seen in Figure 4.3 as it is calculated from elevation at both the 1km and 25km resolutions. The upscaled plot (25km resolution) shows aspect that reflects prominent mountain ranges in the United States which are major drivers of precipitation.

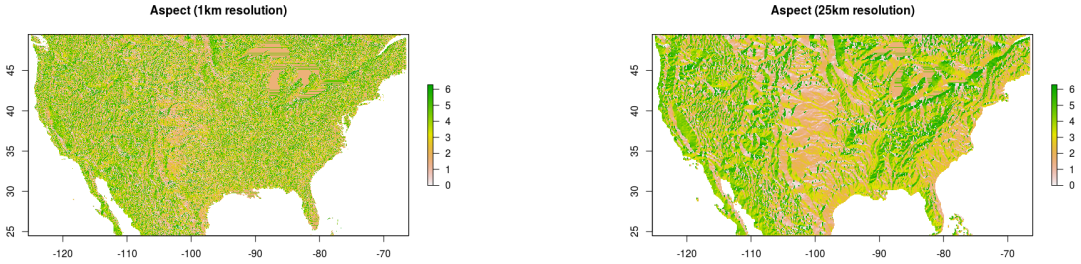

Figure 4.3: Aspect in radians from south, calculated from elevation at a 1km resolution (left) and 25km resolution (right)

The standard physiographic predictors employed here are then: longitude, latitude, and the pairwise product (longitude × latitude), as well as elevation, slope, aspect, and the three associate pairwise products. One additional predictor is also included; the distance from the Gulf Coast of Mississippi (discussed in section 4.1.1). For the duration of this chapter we use the notation

$$X = [\text{lon., lat., lon. } \times \text{ lat., elev., slope, aspect, elev. } \times \text{ slope,} \\ \text{elev. } \times \text{ aspect, slope } \times \text{ aspect, distance from gulf}], \tag{4.1}$$

for the matrix of predictor variables to be used in regression and additive models. As can be seen in Figure 4.1 multiple model parameters exhibit radial symmetry as distance increases from the northern Gulf of Mexico. This is consistent with literature relating to the "warming hole" as a driving force of precipitation.

4.1.1 The Warming Hole

Recent climatological research has shown a strong relationship between sulfate aerosol release and both cooling trends in temperature and variation in seasonal distributions of precipitation over the United States [10]. As sulfate aerosol is released in larger quantities in the eastern United States this trend is exacerbated regionally and is most densely observed surrounding the Gulf of Mexico. These findings are consistent with the model parameters shown in Figure 4.1 where many of the parameters that dictate seasonal behavior of the estimated distributions of positive precipitation show symmetry radially outward from this region.

To examine these trends in the model parameters more robustly we look specifically at the 200 stations nearest to the gulf coast of Mississippi and apply two-dimensional interpolation to the fields of model parameters and look for extrema in the Gulf of Mexico, examples of the generated interpolating surfaces with highlighted extrema can be seen in Figure 4.4. This gives estimates for the centers of the observed radial symmetry, from which distances to each station can be calculated and used as predictor variables in the regression models to follow below. Doing this process for each of the model parameters corresponding to the weights of the knots and averaging across the locations of the found extrema produces an estimate for the center of radial symmetry, or the gulf coast predictor, seen in Figure 4.5, which is consistent with the literature on the warming hole in the United States [10].

4.2 Kriging

The first attempt to produce a robust spatial model capable of accurately estimating distributions at new locations is to perform kriging. This first attempt at producing spatial estimates

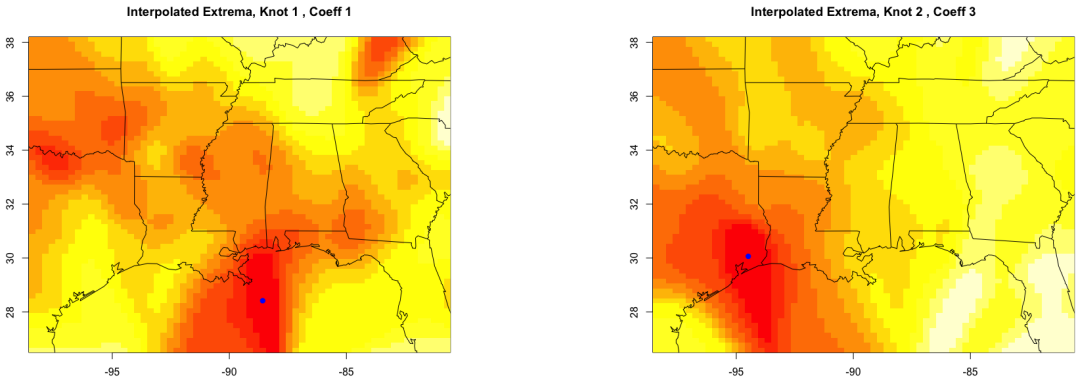

Figure 4.4: Interpolated fields of model parameters with extrema highlighted in blue, these serve as estimates for the center of radial symmetry from which distance is measured to be used as a predictor variable.

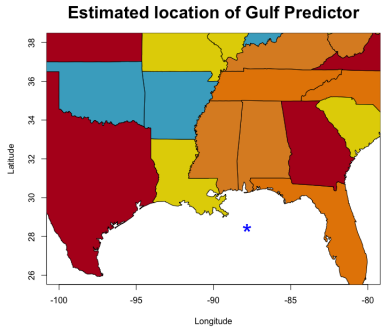

Figure 4.5: Average Location of the extrema found over interpolating fields.

uses a linear regression model (with the predictor variables from equation 4.1) to capture the underlying mean trend in the data, then kriging is performed on the residuals of the regression model via employment of binned semivariograms.

4.2.1 Kriging Residuals Using Binned Semivariograms

The assumed form of the predicted parameter at a new location is

$$y(s_0) = \mu(s_0) + z(s_0) \tag{4.2}$$

in which $\mu(\cdot)$ represents the mean trend of the process, and $z(\cdot)$ is a mean zero gaussian process with some non-trivial covariance function $cov(s_1, s_2)$ [7]. We can consider $\mu(\cdot)$ as the product of

a regression model, and $z(\cdot)$ as the residuals of said process, which are mean zero. Predictions at location s_0 are formed by summing the predicted mean function at s_0 and an estimate for the kriged residual $\hat{z}(s_0)$, that is

$$\hat{y}(s_0) = \mu(s_0) + \hat{z}(s_0) = \mu(s_0) + \Sigma_0^T \Sigma^{-1} \boldsymbol{z}. \tag{4.3}$$

The matrices Σ_0 and Σ are the covariance matrices where Σ_0 is $n \times 1$ with entry i being $cov(y(s_i), z(s_0)) = cov(s_i, s_0)$ and Σ is $n \times n$ and is the covariance matrix between all observed locations $s_1, \ldots, s_n$. The product $\Sigma_0^T \Sigma^{-1}$ describes the "kriging weights" which are applied to $\boldsymbol{z}$ to produce new estimates.

Thus with an estimate of the mean trend (equation 4.4) and a model for $z(\cdot)$ (equation 4.5), all that is needed is an estimate of the underlying covariance function and kriging (equation 4.3) can be performed.

The regression model assumption is that the model parameters follow the form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.4}$$

in which X is the matrix of predictors defined in equation 4.1, $\boldsymbol{\beta}$ are the regression weights, and $\boldsymbol{y}$ is the vector of the model parameters for each spatial location, and $\boldsymbol{\epsilon}$ is a vector of error terms. The mean trend is then defined by $\boldsymbol{\mu} = X\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ represents the standard least squares regression coefficients, and the residuals,

$$\hat{\boldsymbol{w}} = \boldsymbol{y} - \boldsymbol{\mu}, \tag{4.5}$$

should be mean zero with nontrivial covariance.

We build a covariance model by fitting a variogram function (from which a covariance function follows) to a binned semivariogram (equation 4.8) for all model parameters, and use these to predict the parameters at new locations via kriging. The two variogram functions employed here are the exponential,

$$\gamma(r) = \sigma^2(1 - e^{-\frac{r}{a}}), r > 0 \tag{4.6}$$

and the Matern,

$$\gamma(r) = \sigma^2(1 - \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{r}{a})^{\nu}\kappa_{\nu}(\frac{r}{a})), r > 0. \tag{4.7}$$

In 4.6 and 4.7 r is the distance between spatial locations, σ^2 is the variance for any individual station, a is a range parameter determining how correlated the process is over space, ν is the smoothness of the process, $\Gamma(\cdot)$ is the Gamma function, and $\kappa_\nu(\cdot)$ is a modified Bessel function of the second kind [7].

The form of the binned semivariogram is,

$$\hat{\gamma}(r) = \frac{1}{2|N(r)|} \sum_{\|s_i - s_j\| \in N(r)} (\hat{w}(s_i) - \hat{w}(s_j))^2 \tag{4.8}$$

where $N(r)$ is a neighborhood of radius r (i.e. a bin around r), and $|N(r)|$ is the number of points in that neighborhood. Intuitively, the values of $\hat{\gamma(r)}$ serve to represent how decoupled parameters get as they become further apart in space, as this is the idea behind spatial covariance we model variogram functions (and associated covariance functions) off of these approximations.

Variogram functions are fit by minimizing a weighted sum of squared errors between the binned semivariogram and variogram function evaluated at the bin centers. The parameter sets, $\boldsymbol{\theta}$, that define the variogram and covariance functions are selected as,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{|N(r_k)|}{\gamma(r_k|\boldsymbol{\theta})} (\hat{\gamma}(r_k) - \gamma(r_k|\boldsymbol{\theta}))^2 \tag{4.9}$$

where $\gamma(r_k|\boldsymbol{\theta})$ is a variogram function with parameters $\boldsymbol{\theta}$ and $\hat{\gamma}(r_k)$ is the estimate drawn from the binned semivariogram, both evaluated at binned distance r_k. This formulation puts increased weight on bins with both more observations ($|N(r_k)|$ term) and at smaller distances ($\frac{1}{\gamma(r_k|\boldsymbol{\theta})}$ term) [7].

Figure 4.6 shows spatial plots of the residuals once the mean trend is removed from two randomly chosen model parameters.

Looking at Figure 4.6 we see that there is still a substantial amount of spatial structure left in the residuals that has not been sufficiently captured by the mean model. Using the linear models of equation 4.4 as a basis point we attempt to form a more sophisticated estimate of the mean trends of model parameters $\boldsymbol{\mu}$ in equation 4.5 using generalized additive models.

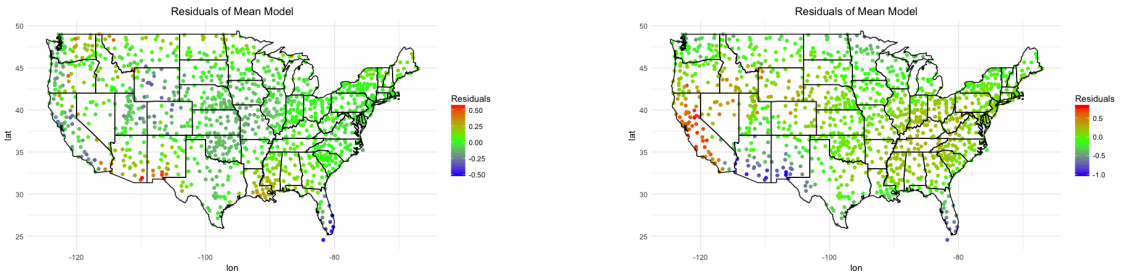

Figure 4.6: Example of a spatial plot of residuals for randomly chosen model parameters.

4.2.1.1 Generalized Additive Models

Generalized additive models (GAMs), as used here, are likelihood based regression models in which the regressors are formulated as smoothed functions of the predictor variables. Formally we estimate $\boldsymbol{\mu}$ by

$$\boldsymbol{\mu} = s_0 + \sum_{j=1}^{P} s_j(X_j) \tag{4.10}$$

where $s_j(\cdot)$ are smooth functions; here we use a range of first to third degree splines. For further reading on the theory behind generalized additive models see Hastie and Tibshirani, 1990 [8]. In principle this gives a much greater amount of flexibility to the models of the mean trends and allows us to more sufficiently capture the smaller scale variability that is seen in the residuals of Figure 4.6.

The trends seen in Figure 4.6 indicate that an increase in flexibility surrounding longitude and latitude terms, as well as the secondary geophysical predictors like slope and elevation would lead to overall decrease in residuals. Using this as motivation we construct a generalized additive model using smoothing splines for all predictor variables. Observing the p-tests for significance of the predictors we find that the majority of the predictors are significant a the standard 0.05 level across all parameter regressions and that all are significant in at least some cases. In the interest of generality we accept the use of this model for all parameters. The average p-test values across all parameters for each third degree smoothed predictor is shown in table 4.1.

Using the intercept term attached to the weight of the largest knot, since this is a major driver of tail probabilities and mean precipitation as a whole, the significance values upon cross

Predictor Variable	Average F-Test Value
Longitude	0.0052
Latitude	0.0140
Latitude × Longitude	0.0227
Elevation	0.0579
Slope	0.1230
Aspect	0.1011
Elevation × Slope	0.2031
Elevation × Aspect	0.1710
Slope × Aspect	0.1945
Gulf Parameter	≈ 0

Table 4.1: Average F-Test values across all model parameters for the GAM predictors.

validation at each station are recorded for all parameters and a selection are displayed in Figure 4.7. What is seen in these plots is that even for predictors that are not significant in average have predictive significance for a number of stations and thus they are kept in the model for generating spatial predictions.

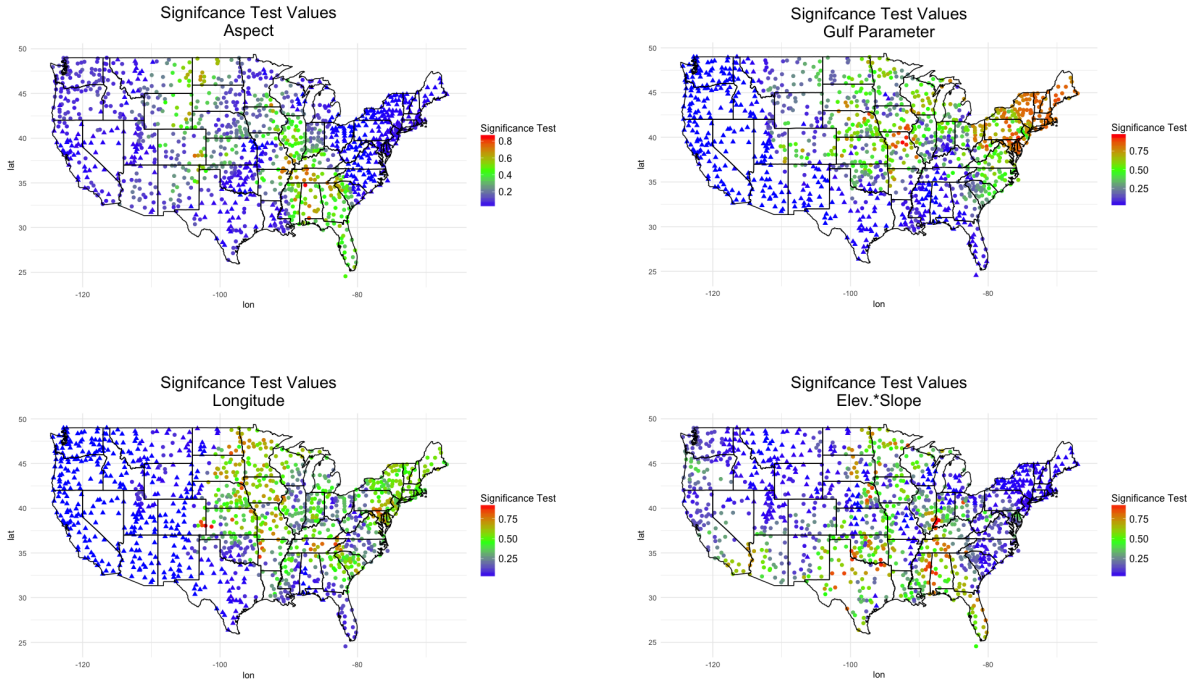

Figure 4.7: Examples of spatial plots of significance tests of GAM predictors, note that all points that are displayed as triangles are significant at the standard 0.05 level.

The specific model used here is then the same as 4.10 where the $s(\cdot)$ functions are cubic smoothing splines and the predictors are the same as those in 4.1. Moving forward all references to the mean model refer to this GAM formulation of the model.

Back to Variogram Estimation

With the coefficients in the variogram function estimated we construct a covariance function with which kriging is performed according to equation 4.3. We first re-estimate all model parameters for each location according to hold-one-out cross validation. Looking at the kriging residuals in Figure 4.8 we see that there is still spatial structure that is failing to be captured by the kriging model. Additionally the matrix computations involved in equation 4.3 are particularly expensive, and looking towards generating estimates across gridded domains any meaningful resolution would become challenging to compute in a reasonable amount of time.

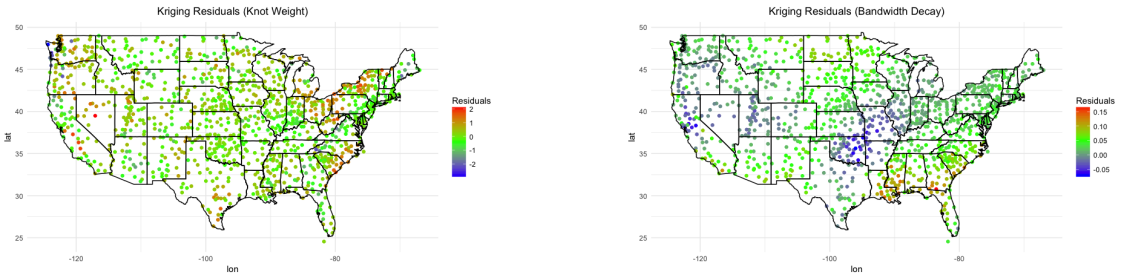

Figure 4.8: Residuals for a randomly chosen weight coefficient (left) and a randomly chosen bandwidth decay coefficient (right)

4.2.2 Non-Stationary Kriging

By considering the data as non-stationary (i.e. non-constant variance over space) we can perform the same procedure outlined in section 4.2.1 but only use information from a subset of neighboring stations in the generation of both the linear model of the mean trend (equation 4.4) and the binned semivariogram (equation 4.8). This leads to not only improved prediction and increased capturing of spatial trends, but dramatically reduced compute times needed to produce estimates.

In choosing the number of neighboring stations used to produce hold-one-out cross validation we look at both the R^2 values of the linear models of the mean trends and the prediction errors of these means. Figure 4.9 shows how mean R^2 and mean prediction errors across all hold-one-out cross validations change according to the number of neighboring stations used to build the mean model trend.

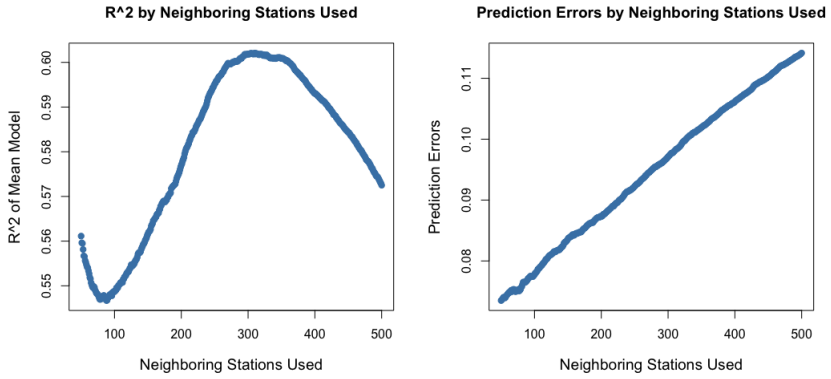

Figure 4.9: R^2 (left) and absolute prediction error (right) according to number of neighboring stations used to build mean trend model.

We see that in terms of variance explained the optimal number of stations to use is about 300, however prediction errors decrease using smaller numbers of neighbors. Given the bias towards smaller sample sizes for increased computational performance and the small gains and losses seen in these statistics (note the scales in Figure 4.9) we move forward using the nearest 200 stations to perform a localized version of kriging and check the accuracy of this model using hold-one-out cross validation on all stations. Note that all measurements are done in the great circle distance as found using longitude and latitude coordinates.

Accuracy of Non-Stationary Kriging

The corresponding plots to Figure 4.8 for this localized method are displayed in Figure 4.10. It is immediately clear that improvements have been made in capturing the spatial structure of the model parameters. Unfortunately even with the increase in captured information the products of non-stationary kriging are insufficient estimators and do not accurately reflect the data. An

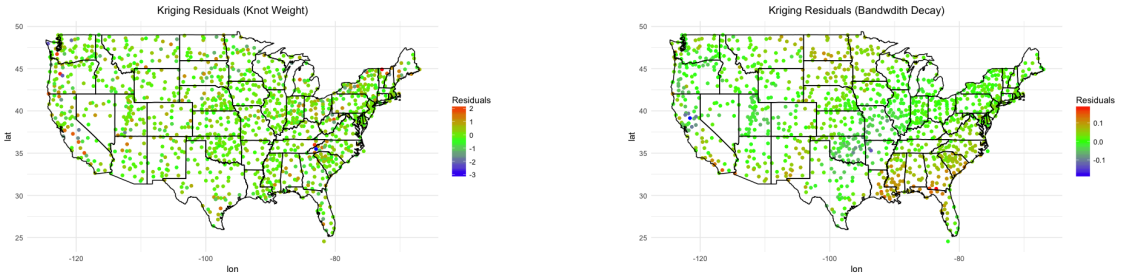

Figure 4.10: Residuals for a randomly chosen weight coefficient (left) and a randomly chosen bandwidth decay coefficient (right) for localized (non-stationary) kriging.

overlayed comparison of quantile-quantile plots of both the local model and the predicted model via kriging against the observed data for randomly selected space-time pairs is shown in Figure 4.11. This figure shows that alterations will need to be made before we can move forward to gridded estimation.

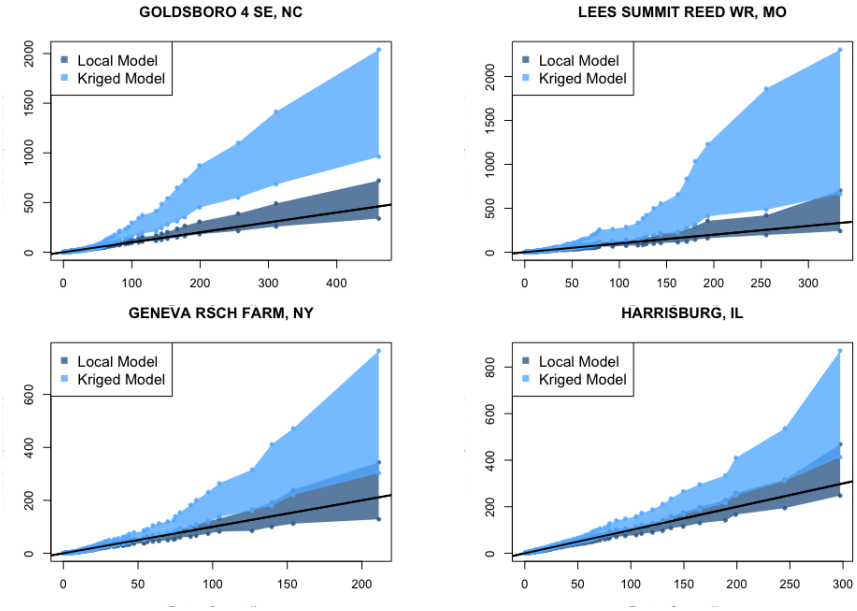

Figure 4.11: Quantile-Quantile plots of local and kriged models for randomly sampled locations and days (units of 1/100th inches).

4.2.3 Maximum Likelihood Estimation of Covariance

Noting the inaccuracies associated with kriging using binned semivariograms and exponential covariance functions we make two modifications in an attempt to generate more accurate spatially predicted estimators: using maximum likelihood estimation of the covariance function, and using a Matern covariance (equation 4.11). The Matern covariance is analgous to the variogram function in equation 4.7, and is defined by the distance r between two locations as

$$\gamma(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\frac{r}{a})^{\nu} \kappa_{\nu}(\frac{r}{a}), r > 0 \tag{4.11}$$

with the same definitions as equation 4.7. Note that a Matern with smoothness (ν) of 1/2 is equivalent to an exponential, thus we are just using a more generalized assumption about the underlying covariance of the process.

Since we expect the residual terms from 4.4 to form a mean zero gaussian process with some associated covariance we can estimate the parameters ν and a in equation 4.11 by selecting those that would produce the greatest likelihood of observing the found residuals. Similar to the process in section 2.1.2 we minimize the negative log-likelihood using nonlinear minimization techniques built into `R` [13].

The resulting process is not only more computationally efficient than the methods outlined in section 4.2.1 but the predicted estimators via kriging are much closer to the local estimates developed in chapter 3. Figure 4.12 shows a similar set of plots to Figure 4.11, and it is clear that the kriged estimators are in much closer agreement with the established local estimators with the use of maximum likelihood estimates and non-stationary kriging.

What we see at this point is that much of the processing done to produce these spatial estimates is unnecessary. Typically it is seen that the prediction of the mean of the process is many orders of magnitude larger than the kriged residual coming from any of the methods described here. In a standard case we may have a mean model prediction that is on the order of 10^{-1} or 10^{-2} and a residual predicted through kriging that is $< 10^{-10}$.

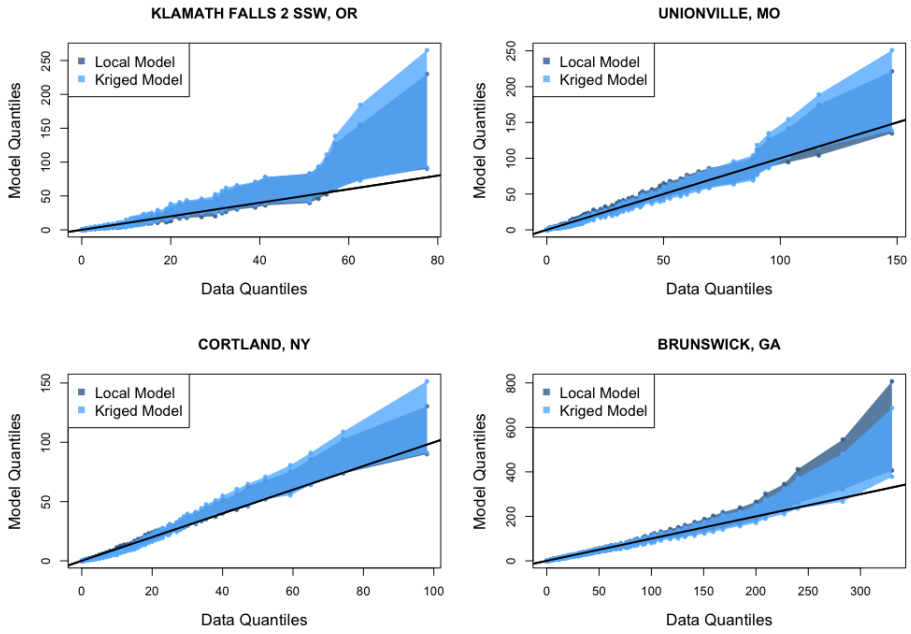

Figure 4.12: Quantile-Quantile plots of local and kriged models for randomly sampled locations and days (units of 1/100th inches).

4.3 Predictions Without Kriging

Having exhausted kriging as a potential method for spatial prediction, generating spatial predictions using only the generalized additive models outlined in section 4.2.1.1 is investigated. Provided that we cannot glean further information through more robust statistical models such as kriging, and that predictive inference does not suffer from simplifying the spatial model, we gain significant benefits through this reduction in complexity in terms of both model simplicity and computational efficiency.

First let us discuss a brief refresher to clarify this final form of the method for spatial predictions of model parameters. We wish to predict the model parameters outlined in chapter 3 at some new location l_0. After rejecting kriging as a productive method of forming this estimator GAMs alone are used to generate spatial models of each parameter. As discussed in 4.2.2 we see gains in the accuracy of estimations by assuming a non-stationary model and using only a number of nearest neighbors, here decided to be the nearest 200 recording stations.

Thus for some parameter at location l_0 with associated predictor variables $x_1, \ldots, x_n$ (geophysical traits of location l_0) we have the mean prediction of that parameter as $\hat{\mu}(l_0)$ is estimated as

$$\hat{\mu}(l_0) = \sum_{i=1}^{n} s_i(x_i) \tag{4.12}$$

where $s_i(\cdot)$ are found as the smoothing functions in the GAM as constructed using the 200 nearest stations to l_0, and x_i are the predictor variables at those locations.

The issue then becomes the lack of small scale variability of predictions. Using only GAMs produces estimates that are far "smoother" spatially than we know precipitation behaves by looking at spatial plots of both data and the optimal model parameters. To combat this a small amount of noise is added to the predicted parameters upon calculation. This gives a small element of randomness to the model for a given space-time pair each time it is constructed, accounting for the over-certainty that accompanies using GAMs as the only avenue for spatial prediction.

The process for adding noise to parameter predictions is as follows: perform hold-one-out cross validation for a parameter at all locations to get the initial estimates of parameters $\hat{\boldsymbol{\mu}}$ (denoted as such because after randomness is added this is considered the mean of the prediction), then assuming the same form of non-stationarity from the mean model generation above estimate the standard deviation for a given location based on the residuals of the spatially predicted parameters $\hat{\sigma}(l_0)$, then the new parameter p_0 for a given location becomes $p_0(l_0) = N(\hat{\mu}(l_0), \hat{\sigma}(l_0)^2)$. This noise represents the uncertainty of prediction and upon repeated sampling of parameters according to this procedure, models generated from these parameter sets have a broader range of potentially simulated precipitation leading to an increase of agreement in the comparison of observed quantiles and 95% confidence intervals of simulated quantiles.

Estimates of standard deviations are calculated via the standard method for sample variance,

$$\hat{\sigma}(l_0) = \left[\frac{\sum_{i=1}^{200} (x_i - \bar{x})^2}{N - 1} \right]^{1/2} \tag{4.13}$$

in which the x_i's are sampled from the nearest 200 stations to l_0 and the mean of these is $\bar{x}$.

We validate this method using hold-one-out cross validation on all recording stations then move to generating local models for gridded sets of locations across the United States. Initially we look at a plot similar to Figure 3.10 but with 95% confidence intervals for spatially predicted models superimposed on the locally generated models. This is shown in Figure 4.13, in which it is clear that the predicted models accurately reflect the locally generated models and have an overall tendency to capture the observed quantiles of the data over the two weeks surrounding the indicated dates.

With a local model for climatology in place (chapter 3) and a spatial model capable of estimating these local models at new locations developed in this chapter the next step is to move towards estimation of precipitation across a grid of points covering the United States. This is a complex problem and is by no means covered in full here, rather a cursory investigation is discussed and used as motivation for further applications of this model and its potential deployment as a predictive tool.

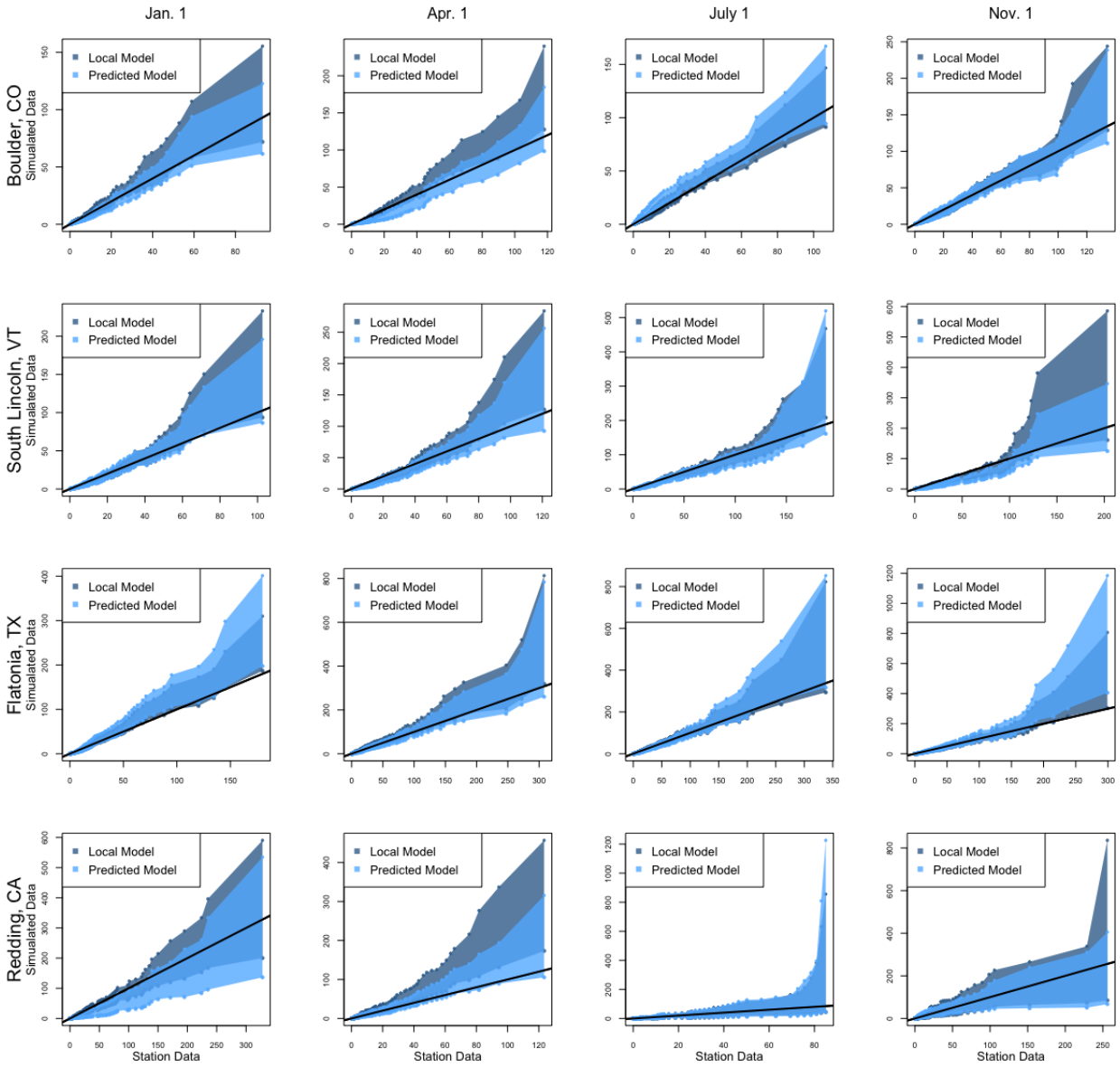

Figure 4.13: This plots show 95% confidence intervals of quantiles for both locally generated models and hold-one-out predicted models against the quantiles of data observed over two week windows centered at the indicated dates (units in 1/100th inches).

Chapter 5

Assessment

A major benefit of a model such as this is the ability to estimate the precipitation climatologies at locations for which recordings have never been taken. Chapter 4 developed the methods for this prediction primarily using cross validation, but only examined recording locations. In this chapter a brief examination of the ability of the product of Chapters 3 and 4 to estimate climatologies at new location with particular focus on generating gridded estimation of precipitation, ultimately moving in the direction of simulating precipitation events across a near-continuous gridded set of points over the United States.

5.1 Gridded Model Estimates

Before large scale precipitation events can be simulated a set of gridded model estimates for the parameters of the estimator outlined in equation 3.10. Note that for any reasonable resolution of grid this is computationally expensive, but is highly parallelizable reducing the burden greatly. For investigation a grid of points spaced at every half degree of latitude and longitude are generated with estimates produced at each point; estimates are generated in the same way that hold-one-out cross validation seen in section 4.3.

With these estimates, and working from the assumption that the spatially estimated model parameters are reflective of local climatology, inference about precipitation processes can be drawn from from parameter estimates, i.e. if a large maximum knot is predicted at a given location, this implies that heavy precipitation is more likely at this location than others for which the maximum

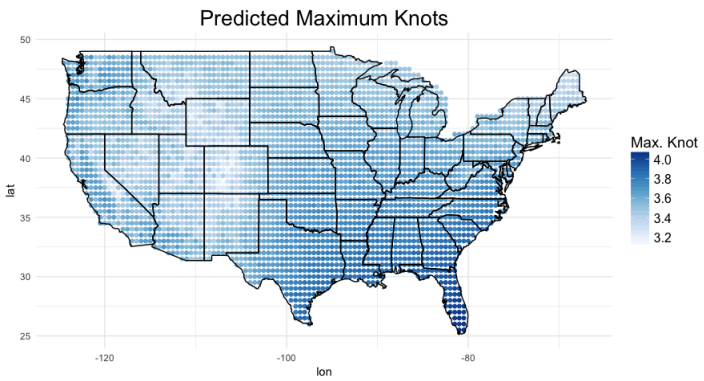

Figure 5.1: Predicted maximum knots for gridded points

knot is small. For reference to this specific idea Figure 5.1 shows the predicted maximum knots for the set of gridded points. As a sort of sanity check this figure is not only consistent with the findings of this work, but what is intuitively known about precipitation: that heavy precipitation events are more common in places like Louisiana and Florida and less likely in high elevation and arid climates such as the Rocky Mountains and New Mexico.

As a point of comparison to other gridded precipitation products we include median simulated precipitation on January first and July first. These results are displayed in Figure 5.2 and when compared to established gridded precipitation products are typically well aligned [4].

Using the sets of parameter estimates PDFs can be generated for all gridded points for a given day of the year. This enables patterns and structure in the estimated distributions to be seen over space and is displayed for January 1 in Figure 5.3. The curves displayed on the map in this figure represent the PDFs estimated at each location (note that the PDFs have been truncated for clarity). From this it is seen that there are regions for which precipitation has an overall "flatter" distribution and there are regions with much more pronounced "spikes" in density with faster decay into the tail densities. Comparing this to the median simulated precipitation on January 1 in Figure **??** it is seen that the areas with slower decay into the tail densities (west coast and the Gulf of Mexico) observe higher median precipitation upon simulation. This is expected, but further

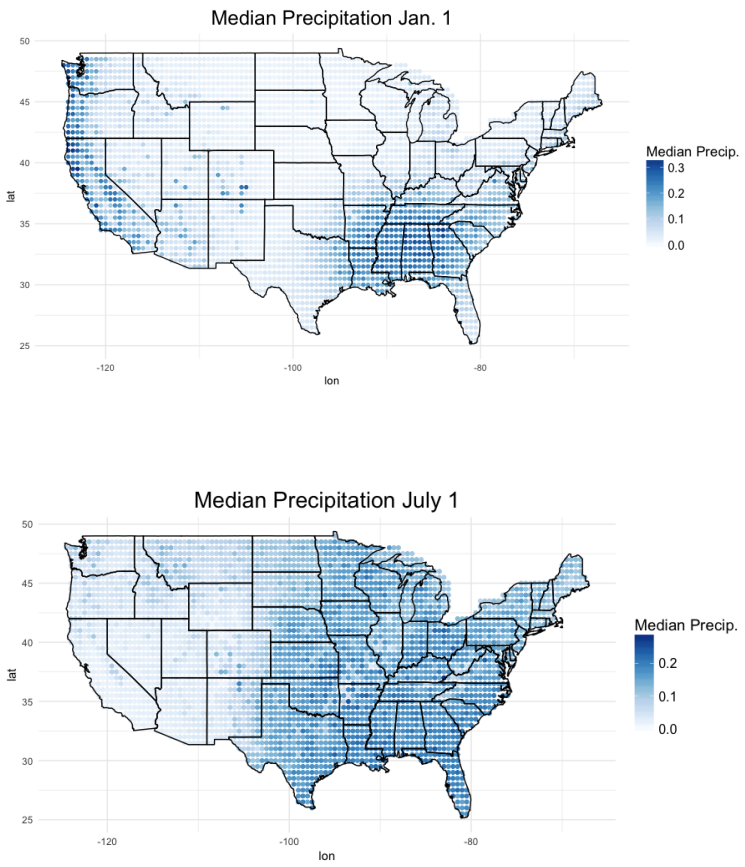

Figure 5.2: Median simulated precipitation for January 1 (top) and July 1 (bottom).

indicates the use of having full estimated densities for gridded sets of locations rather than point estimates alone.

5.2 Simulation of Precipitation

With gridded estimates of distributions of precipitation determined a major area of interest can begin to be investigated: the simulation of fields of precipitation for the United States. This process or the model outlined here is not complete and is therefore not discussed in great detail, it is merely discussed to give a general sense of the direction of the project and where this model may be taken with further research. The general outline of the process by which simulation of gridded

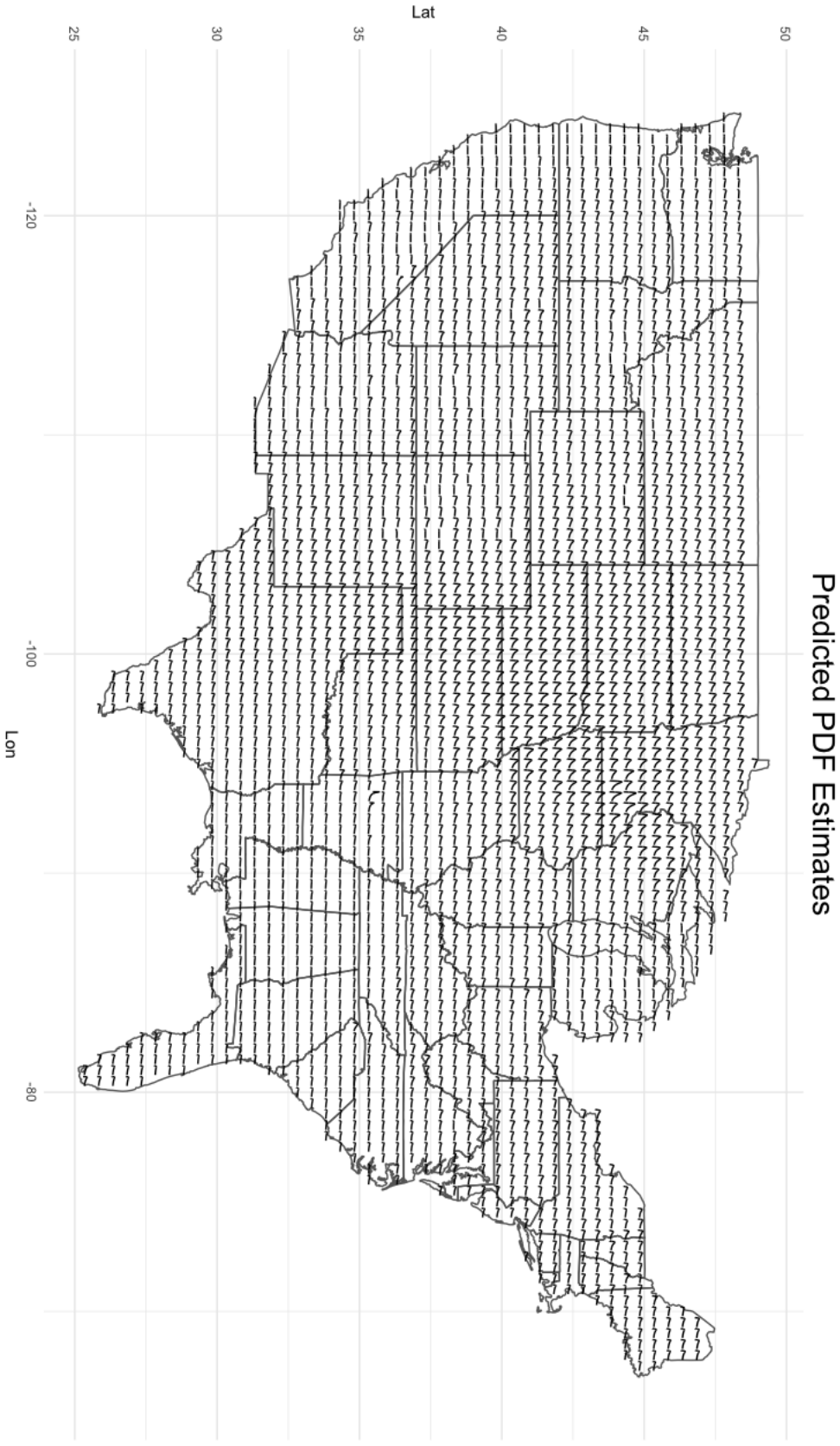

Figure 5.3: Estimates of precipitation distributions for gridded points

precipitation is simulated is:

(1) Use a probit model and censoring field to determine a set of stations for which positive precipitation occurs

(2) Simulate a mean zero Gaussian process (the precipitation field) across stations with positive precipitation

(3) Transform this Gaussian process to positive precipitation according to estimated density functions.

Thus it is necessary to determine the probability of observing positive precipitation for a given location on a given day of the year according to a probit model and a covariance model with which the censoring and precipitation fields can be generated.

5.2.0.1 Determining Covariance

As mentioned in Chapter 4 the exponential covariance function is common in geostatistical applications such as precipitation and is employed here as a method of determining the covariance of the above Gaussian process, Z [1]. Formally, the covariance between two points in the field is estimated as,

$$\text{cov}(Z(s_1), Z(s_2)) = exp\left\{-\frac{\|s_1 - s_2\|}{a}\right\} \tag{5.1}$$

in which a is a range parameter that serves to indicate how quickly the covariance between two points decays with distance (i.e. larger ranges mean that points far away are more correlated), and $Z(s_i)$ is the observation of the standard normal random process $Z(\cdot)$ at location s_i. Note that here the norm represents distance between points s_1 and s_2 as calculated by the great circle distance.

To calculate this range parameter, a, maximum likelihood estimates (MLEs) are employed in the same fashion as seen in Chapter 4. To generate these estimates the data must be transformed to normality. The transformation from precipitation Y to normal random variables is,

$$z = \Phi^{-1}(\hat{F}_{s,t}(log(y))) \tag{5.2}$$

and the back transformation from normality to precipitation is done as

$$y = \exp\left\{\hat{F}_{s,t}^{-1}(\Phi(z))\right\}. \tag{5.3}$$

In both 5.2 and 5.3 z is an observation of a standard normal random variable, y is an observation of precipitation, Φ is the CDF of a standard normal, and $\hat{F}_{s,t}$ is the estimated CDF of precipitation for location s and day t.

What these transformations are useful for is that for any single day of observation (not accumulated over all 115 years of recording, just a single day of a single year), all positive observations are transformed to normality and a maximum likelihood estimate for a in 5.1 can be generated. This process is repeated for all 365 days of all 115 years and estimates of range parameters for all 41975 single days of observation are found and thus for any randomly selected date over the 115 years of observation a Gaussian field with covariance reflective of the data can be generated.

5.2.1 Probit Model and Censoring Field

To determine the probability of observing positive precipitation a probit model is used. Probit modeling is a method of regression similar in scope to logistic regression in that it serves to model the probability of some binary event, in this case the observation of positive precipitation. The specific model for a given location is such that

$$P(Y(s,t) = 1) = \Phi(X\boldsymbol{\beta}(s)) \tag{5.4}$$

where Y is the event that positive precipitation has occurred, $\beta(s)$ is the vector of regression coefficients, X is the predictor matrix, and Φ represents the CDF of the standard normal distribution [1]. The βs are estimated using generalized linear regression in `R` with the probit link function, $\Phi^{-1}(p)$, and where the i^{th} row of the predictor matrix X is of the form,

$$\begin{aligned} X_i = \Bigg[&\sin\left(\frac{2\pi t}{365}\right), \cos\left(\frac{2\pi t}{365}\right), \sin\left(\frac{4\pi t}{365}\right), \cos\left(\frac{4\pi t}{365}\right), \\ &\sin\left(\frac{6\pi t}{365}\right), \cos\left(\frac{6\pi t}{365}\right), \sin\left(\frac{8\pi t}{365}\right), \cos\left(\frac{8\pi t}{365}\right) \Bigg] \end{aligned} \tag{5.5}$$

where t is day of the year of recording i and the response variable is either 1 or 0 indicating observation of positive precipitation for that recording. Examining the empirically calculated probabilities of positive precipitation by day (i.e. the number of positive observations divided by the total number of observations) for a number of locations indicates that the inclusion of the additional harmonics in equation 5.6 as compared to similar models discussed in Chapter 3 and 4 was necessary and even the higher order terms are typically found to be statistically significant.

Once the probit coefficients $\boldsymbol{\beta}(s)$ are found for all locations in the observational network the same process of spatial prediction with GAMs from Chapter 4 can be applied and thus precipitation probabilities can be calculated at arbitrary locations, including on a grid.

To determine the stations for which positive precipitation is observed upon simulation for day t_0, let $\mu(s) = X(t_0)\boldsymbol{\beta}(s)$ from equation 5.4 where $X(t_0)$ is defined as,

$$X(t_0) = \left[\sin\left(\frac{2\pi t_0}{365}\right), \cos\left(\frac{2\pi t_0}{365}\right), \sin\left(\frac{4\pi t_0}{365}\right), \cos\left(\frac{4\pi t_0}{365}\right), \right. \\ \left. \sin\left(\frac{6\pi t_0}{365}\right), \cos\left(\frac{6\pi t_0}{365}\right), \sin\left(\frac{8\pi t_0}{365}\right), \cos\left(\frac{8\pi t_0}{365}\right) \right] \tag{5.6}$$

i.e. the predictor variables for day t_0, and $\boldsymbol{\beta}(s)$ is the vector of probit coefficients for location s. Furthermore let Z be a mean zero Gaussian process (the covariance of which is discussed above) with realizations $Z(s)$ at all gridded locations $\boldsymbol{s}$. Following Berrocal et al. (2009) the Gaussian field $W(s) = \mu(s) + Z(s)$ is computed at all locations $s \in \boldsymbol{s}$, and is referred to here as the censoring field and serves to determine whether or not precipitation at a given location $Y(s)$ is either zero or positive. Positive precipitation is simulated for all locations where the censoring field is greater than 0, that is,

$$Y(s) = \begin{cases} 0 & W(s) \leq 0 \\ > 0 & W(s) > 0. \end{cases} \tag{5.7}$$

5.2.2 Simulation of Positive Precipitation

With a covariance model (section 5.2.0.1) and method for determining which stations observe positive precipitation (section 5.2.1), positive precipitation can be simulated. This is done by

simulating field W as described above, then simulating a new Gaussian field Z' at all locations for which $W > 0$ from equation 5.7, then applying the transformation 5.3 to all points in Z' gives sets of observations in full (not log) precipitation space.

Figure 5.4 shows plots in which each row signifies a randomly chosen day of the year (one observation, not aggregated over a rolling window), and the first column is observed precipitation and the second column is simulated precipitation on a grid. What is significant about these figures is that the simulations are on the same scale as the observed data, and that the ratio of number of locations for which the data/simulations are positive to the number of locations for which data/simulations are zero is approximately the same across observations and simulations. These plots are included to serve as motivation that these simulations are comparable to observed precipitation upon simple inspection and that further analysis and development of gridded estimation of precipitation could lead to robust and useful results.

From this chapter we see that although not covered in full here this work serves as an effective framework for the development of a robust gridded precipitation product for which a high degree of confidence can be placed in local estimates.

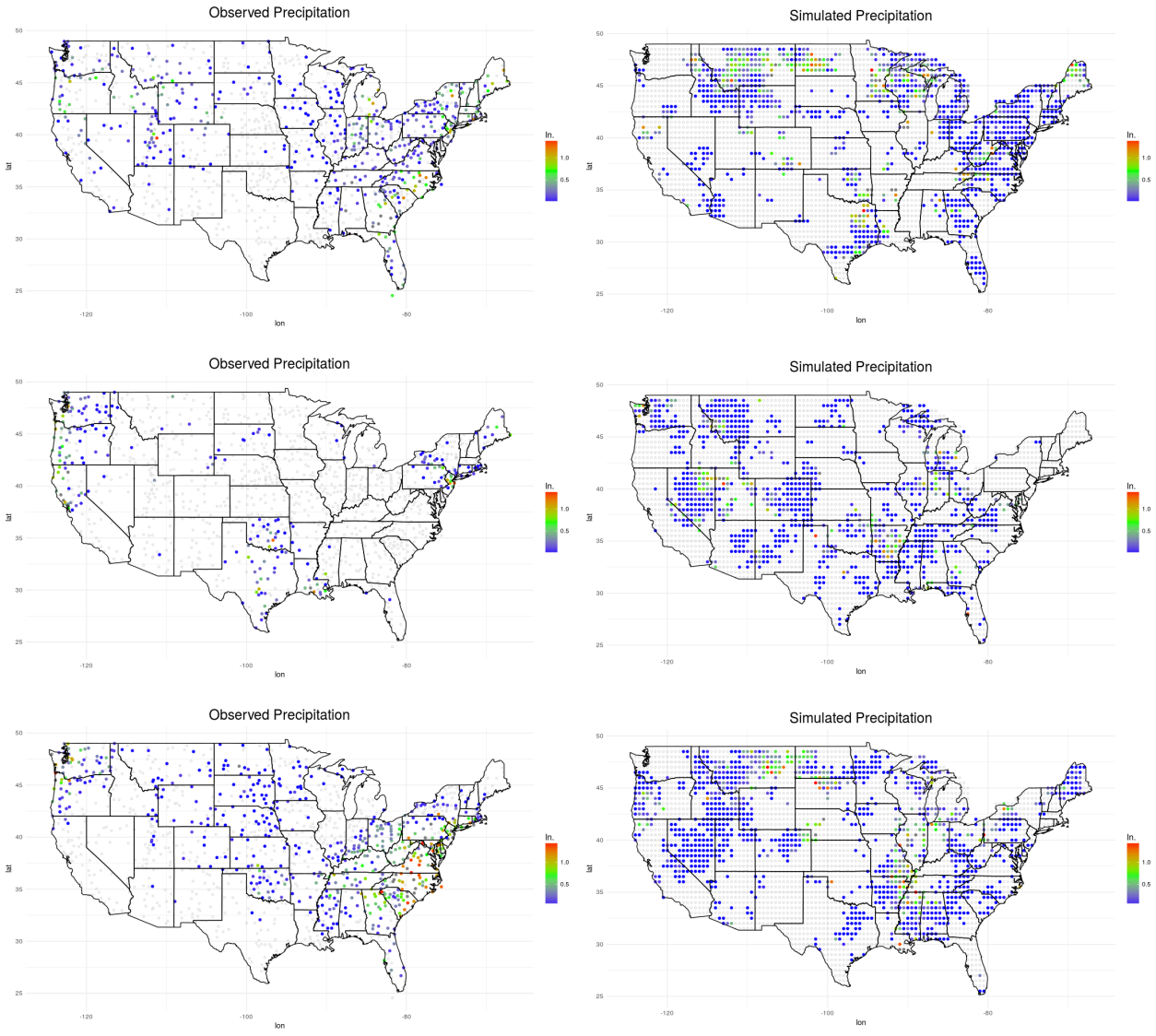

Figure 5.4: Samples of observed data (left column) and simulated gridded precipitation (right column) where rows indicate same randomly chosen days of the year.

Chapter 6

Conclusion and Further Research

The aim of this work is to develop an accurate and interpretable method for generating local estimators of precipitation climatology based on in situ measurements and predict these estimators at arbitrary locations across the United States including gridded sets of points. Many of the goals and milestones aimed for have been reached and the resulting models for local estimation and spatial estimation of local estimates are both accurate and interpretable.

Chief among the developments made here is the validation of the use of low-rank kernel density estimators as modeling tools; this is a new approach but the accuracy obtained on data as notoriously difficult to model as precipitation in chapter 3 serves as an indicator of the estimator's ability to capture and recreate trends in data effectively. Not only is the estimator accurate but it has the necessary property of being able to be estimated across any spatial domain for which data has been observed relatively nearby, making it a promising tool for future climatological research.

We believe that low-rank kernel density estimators will provide a new modeling tool that could be broadly applied to new domains, and that the work done here will serve as a useful foundation for further inquiries in the development of gridded precipitation estimation products.

6.1 Further Research

This work contains a number of novel approaches to precipitation modeling as well as the employment of a new "low-rank kernel density estimator", as such there are components that have room for development and are numerous avenues through which this work can be built upon and

investigated further.

Semi-Parametric Model

The first of these potential sources for improvement, which is currently being investigated, is to transform the non-parametric estimator seen in chapter 3 into a mixed distribution in which the portion of the distribution corresponding to negative log-precipitation is governed by an negative exponential distribution. This arises from the common result that the logarithm of a uniform random variable is a negative exponential. if we consider Y to be the negative logarithm of $U \sim unif(0,1)$, representing precipitation it is shown that

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(-\log(U) < y) = P(U > e^{-y}) \\ &= 1 - P(U < e^{-y}) = 1 - F_{unif(0,1)}(e^{-y}) \end{aligned} \tag{6.1}$$

and differentiating yields

$$f_Y(y) = f_{unif(0,1)}(e^{-y})e^{-y} = e^{-y} \tag{6.2}$$

giving that Y is an exponential random variable. Using this result and the process of randomly imputing trace precipitation in $(0,1)$ where trace events were observed we find that the log-precipitation values less than 0 will directly follow an exponential random variable reflected over the y-axis.

Since this formulation only effects the negative log-precipitation portion of the estimator much of the work done in chapter 3, such as knot selection and bandwidth decay, carries over for positive log-precipitation. Using the model from chapter 3 but truncating the knots to those residing along the positive log-precipitation axis and implementing a simple exponential mixture model has led to positive preliminary results.

The form of the PDF of the estimator for log-precipitation in this model is

$$\hat{f}(x) = f(x|s,t) \propto \sum_{x_k>0} \sigma_k(s,t) K\left(\frac{x - x_k(s)}{\eta_k(s,t)h(s,t)}\right) + \lambda e^{\lambda x} \cdot \mathbb{1}_{x \leq 0}. \tag{6.3}$$

Figure 6.1 shows the difference between this model and the model developed through the body of this thesis. Note the difference in smaller quantiles of the distribution. This motivates

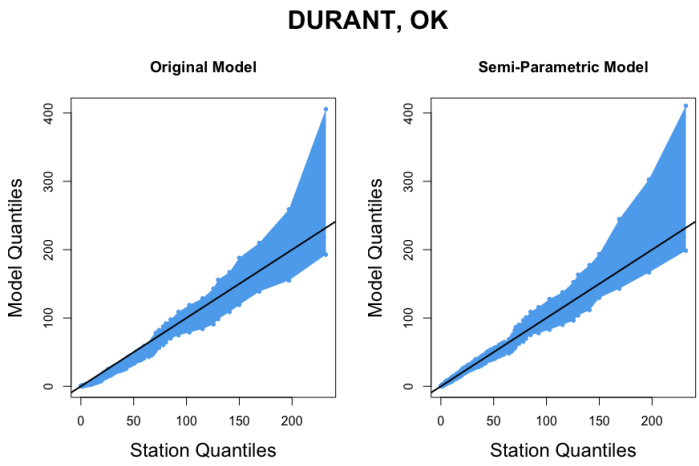

Figure 6.1: An example of the change seen when moving to a semi-parametric distribution, the particular interest here is in the smaller quantiles of the distribution where the semi-parametric model shows improved matching. Units are in $1/100^{th}$ inches of precipitation

that there is ground to be gained here and that with further time improvements could be made by adopting a semi-parametric mixed model.

Knot Selection

Selecting the knots is a particularly challenging component in developing low-rank approximations to kernel density estimators. It is referenced in section 3.1.2 that in order to move forward a somewhat ad-hoc approach became necessary for this model, however with further research it is possible that a computationally tractable way of computing the optimal set of knots could be developed. Given the accuracy shown for applications of a model like this for precipitation and the computational efficiency of simulation and prediction once the model is developed there is certainly potential for low-rank kernel density estimators, and ironing out the details of how to generate estimators efficiently would be useful in many domains.

Bibliography

[1] Veronica J Berrocal, Adrian E Raftery, Tilmann Gneiting, et al. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. The Annals of Applied Statistics, 2(4):1170–1193, 2008.

[2] Noel Cressie. Kriging nonstationary data. Journal of the American Statistical Association, 81(395):625–634, 1986.

[3] Christopher Daly, Ronald P Neilson, and Donald L Phillips. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology, 33(2):140–158, 1994.

[4] Christopher Daly, Melissa E Slater, Joshua A Roberti, Stephanie H Laseter, and Lloyd W Swift. High-resolution precipitation mapping in a mountainous watershed: ground truth for evaluating uncertainty in a national precipitation dataset. International Journal of Climatology, 37(S1):124–137, 2017.

[5] SF Daly, R Davis, E Ochs, and T Pangburn. An approach to spatially distributed snow modelling of the sacramento and san joaquin basins, california. Hydrological Processes, 14(18):3257–3271, 2000.

[6] Jeffrey S Deems, Thomas H Painter, and David C Finnegan. Lidar measurement of snow depth: a review. Journal of Glaciology, 59(215):467–479, 2013.

[7] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. Handbook of Spatial Statistics. CRC press, 2010.

[8] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. Wiley Online Library, 1990.

[9] Upmanu Lall, Balaji Rajagopalan, and David G Tarboton. A nonparametric wet/dry spell model for resampling daily precipitation. Water Resources Research, 32(9):2803–2823, 1996.

[10] Gerald A Meehl, Julie M Arblaster, and Grant Branstator. Mechanisms contributing to the warming hole and the consequent us east–west differential of heat extremes. Journal of Climate, 25(18):6394–6408, 2012.

[11] M J Menne, C N Williams, and R S Vose. United states historical climatology network daily temperature, precipitation, and snow data. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 2015.

[12] Jared W Oyler, Ashley Ballantyne, Kelsey Jencso, Michael Sweet, and Steven W Running. Creating a topoclimatic daily air temperature dataset for the conterminous united states using homogenized station data and remotely sensed land skin temperature. International Journal of Climatology, 35(9):2258–2279, 2015.

[13] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[14] David W Scott. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, 2015.

[15] Thordis L Thorarinsdottir, Tilmann Gneiting, and Nadine Gissibl. Using proper divergence functions to evaluate climate models. SIAM/ASA Journal on Uncertainty Quantification, 1(1):522–534, 2013.

[16] M Vrac and P Naveau. Stochastic downscaling of precipitation: From dry events to heavy rainfalls. Water Resources Research, 43(7), 2007.

[17] Daqing Yang, Barry E Goodison, Shig Ishida, and Carl S Benson. Adjustment of daily precipitation data at 10 climate stations in alaska: Application of world meteorological organization intercomparison results. Water Resources Research, 34(2):241–256, 1998.

[18] Daqing Yang, Douglas Kane, Zhongping Zhang, David Legates, and Barry Goodison. Bias corrections of long-term (1973–2004) daily precipitation data over the northern regions. Geophysical Research Letters, 32(19), 2005.

Appendix A

Kernel Density Estimators

Kernel density estimators (KDEs) are a nonparametric modeling tool produced as the sum of a number of basis functions, which is then normalized to integrate to 1 to form a valid probability density function. For a given kernel function $K(\cdot)$ and data observations x_i, $i = 1, 2, \ldots, n$ the resulting KDE of the density at x, $\hat{f}(x)$ is generated as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{A.1}$$

where n is the number of kernel functions summed over, the x_i values are the locations of the kernel functions, and h represents the bandwidth parameter. The bandwidth parameter dictates the sensitivity to distance between the observations (x_i) and the point of evaluation x, a higher bandwidth will result in estimators that fill in space between observations more substantially and a lower bandwidth will represent a closer fit to the observations. An example of kernel density estimation and the effects of the bandwidth parameter can be seen below in figure A.1.

A.0.1 Function and Bandwidth Selection

The two necessary selections that must be made here are the choice of kernel function and bandwidth parameter. The kernel function is of less importance than the selection of the bandwidth parameter, and following in similar fashion to established nonparametric precipitation models we employ the Gaussian kernel function for the entirety of the model development here [9]. The

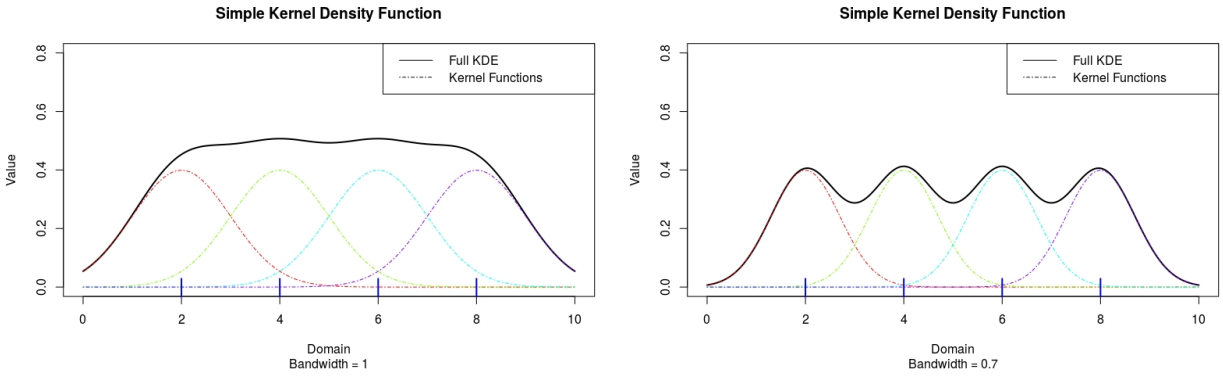

Figure A.1: Example of kernel density estimation and the effect of the bandwidth

Gaussian kernel is defined below in equation A.2.

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-x_i}{h})^2} \tag{A.2}$$

The bandwidth parameter selection is of significantly more importance in the development of a KDE and must be chosen carefully [9]. A standard method, employed for the remainder of the development of this model, is to select the bandwidth that minimizes the asymptotic mean integrated squared error of the estimator [14]. This results in an involved calculation involving the estimated density, the kernel function, and the order of the derivative of the data to matched, and is well documented and widely available [14]. Here we utilize the `kedd` package in `R`, allowing for easy computation of the optimal bandwidth h_{amise} given a Gaussian kernel and a set of data.