University of California

Los Angeles

Regime Based Clustering for the Modeling of Two-Dimensional Vector Fields

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Mark Hiroshi Nakamura

2014

Abstract of the Dissertation

Regime Based Clustering for the Modeling of Two-Dimensional Vector Fields

by

Mark Hiroshi Nakamura

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2014

Professor Mark Handcock, Chair

A two-dimensional dynamic vector field is any system in which spatially separated values move over or through a diverse terrain. Examples of this include wind fields, ocean currents, gaseous systems, and the movement of people. With the exponential growth of these dynamic data sets, via environmental sensing satellites, smart cameras, and GPS enabled cell phones, there is an immediate need for new techniques.

Modeling dynamic vector fields not only entails measuring the spatial dependence between locations and variables, but also capturing the independent markers that signal directional flow changes. In order to capture this, the objective of this new technique was to break down the modeling process into several steps. Given that spatial dependence changes with directional flow, it is the first objective to cluster the training data set of vectors into several distinct flow regimes. The following step is to then build separate spatially dependent models for predicting the value's strength and direction within each regime. By first grouping the data into clusters, we achieve subsets of data with spatial dependence structures that are more homogenous, reducing overall variance or mixed signals in our modeling training.

The process and interpretation of regime based cluster modeling will be demonstrated in the statistical downscaling of two-dimensional ten meter wind fields over the diverse coastal and mountainous terrain of Southern California. Southern California's coastal mountains provide the perfect example of topography changing spatial dependencies with the change of the overall wind direction. This statistical downscaling process helps local regions prepare for and understand possible climate changes and leverages the growing number and importance of Global Climate Models.

The dissertation of Mark Hiroshi Nakamura is approved.

Alan Yuille

Rick Paik Schoenberg

Alex Hall

Mark Handcock, Committee Chair

University of California, Los Angeles

2014

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank all my scientific advisors that have made this research possible. I would like to thank Mark Handcock for his guidance, advice, and support. I would like to thank Alex Hall for providing me with an opportunity. I would then like to thank all the people in my personal life who have supported me, Andrea, my parents, family and friends.

Vita

2008	B.A. (Mathematics and Economics), UCSD, San Diego, California.
2008	B.S. (Psychology), UCSD, San Diego, California.
Summer 2011	M.S. (Statistics), UCLA, Los Angeles, California.
Winter 2012	Graduate Student Researcher, (JPL) Jet Propulsion Laboratory, Pasadena, California.
Winter 2013	Lecturer Stats 10, Department of Statistics, UCLA.
Fall 2013	Lecturer Stats xl10, UCLA Extension.
Winter 2014	Lecturer Stats 10, Department of Statistics, UCLA.
2010–present	Teaching Assistant, Department of Statistics, UCLA.
2010–present	Graduate Student Consultant, Statistical Consulting Center, UCLA.

CHAPTER 1

Introduction

1.1 Statement of the Problem

The objective of this research is to build a predictive spatial model for two-dimensional dynamical systems. I will use the words *dynamical* and *flow* to represent spatial systems that move throughout time. This implies that predictions will be made not only for a specific value (e.g. strength, intensity, etc.) but also for the value's direction in movement. This natural movement is not only caused by the relationships between several variables but they are also heavily influenced by the physical setting, or topographic information. The covariates can give us the source and strength of the push, but topographic barriers redirect and channel these forces. These systems are more often than not deterministic but are still very complex processes and can be represented as vector fields.

Studying and modeling environmental systems provides researchers with a unique opportunity. Environmental systems can have well known processes, but the complexity of the inter-relationships and the physical setting makes the predictions difficult. In these studies, it is in my opinion that the objective of the statistician is to balance the simplification of the system while still capturing the natural properties of the relationships, so the models can be understandable and informative. In addition, these two-dimensional flows also provide statisticians the opportunity to further develop the relationship between spatial statistics and directional statistics.

Modeling dynamic vector fields not only entails measuring the dependence between locations and variables, but also capturing the independent markers that signal directional flow changes. The objective of this new technique is to break down the modeling process into several steps. Given that spatial dependence structure changes with directional flow, it is the first objective to cluster the training vector data set into several distinct flow regimes. With these flow regimes established, the subsequent vector groupings or regimes contain more homogenous spatial dependence structures. With these smaller homogenous data sets, or regimes, you are then able to build separate models for predicting value's strength and direction. These models are capable of explaining more system variance because they are no longer receiving mixed signals about spatial dependence structures. For example, in southern California cool winds that are generated over the ocean and come onshore, southwesterly, are much different from dry winds generated over the high desert that channel through the mountain passes towards the ocean, northeasterly.

In the model building process, previous approaches aimed to avoid circular statistics and represent each flow vector as two numerical responses. Each vector was projected onto both the x and y axis proving two lengths on constant directions. Subsequently two models are created separately, one for the projection onto the x axis and the other for the y axis projection. It is my belief that this technique does not optimally leverage the spatially dependent information. My approach is to treat the vector as polar coordinates. Creating one model to predict the direction of the vector and a separate model to predict the size or magnitude of the vector. I believe treating vectors as polar coordinates will better leverage the regime clustering data separation. For example, if we were to predict the wind speed at the campus of UCLA, once we determine that the day belongs in the offshore Santa Ana regime, we know the source of the wind is generated to the east and the strength of that wind should be highly related to the strength of the

covariates over the high desert.

In the model building process, a regime clustering technique was developed that is an adaptation to hierarchical clustering for vectors. After optimal clustering was performed, several model types were tried in the process of finding an accurate and robust fit for both magnitude and directional predictions. In the end, a transformed linear model performed best for magnitude predictions and a kernel based circular regression fit best for directional predictions.

1.2 Motivation

One obvious example of a dynamic vector flows is two-dimensional wind fields. Winds plays a major roll in southern California wildfires, energy sustainability, and pollution diffusion. Obtaining more informed predictions of future winds allows communities and businesses to be able to evolve and adapt to new environmental challenges.

A main motivation for the need of new vector models has come with the rise of climate modeling and prediction. This major subdivision of climatology has grown along with the strength and ability of computing. For a long time climatologists have studied the physical equations that govern our atmosphere and ocean, but only till recently have our computers gotten to the point to be able to accurately model our earth's vast atmosphere and oceans with these equations. These "dynamical" computer models are called Global Climate Models (GCMs), or they are also known as General Circulation Models (see figure 1.1). There is a loosely estimated number of twenty current GCMs and growing. By having several models from different research groups predict future climates, we end up with a distribution of predictions, which in turn helps us understand the probabilities and variance in our predictions of possible future climate.

Not only do these "dynamical" programs allow us to get a multitude, or distri-

bution, of predictions but they also allow us to control or experiment with different possible levels of future carbon emission, known as Representative Concentration Pathways (RCP) scenarios. This allows climate researchers to understand the different impacts of changing environmental policy. These global climate models can give us a important predictions of future climates on a macro global level, but fail to give predictions at high enough resolutions to inform local governments on a micro level. This is main driver for the motivation of this modeling scheme. The resolution and intricacy of the predictions are held back by the expensive computational costs associated with modeling these enormously complex systems. Even though the GCM projections cannot give us the information interpretable at a local level they still remain the main driver for predictions made at the higher local resolution. Statistical models allow us to leverage our knowledge of local dynamics and perturb or fill in the blanks of these global predictions.

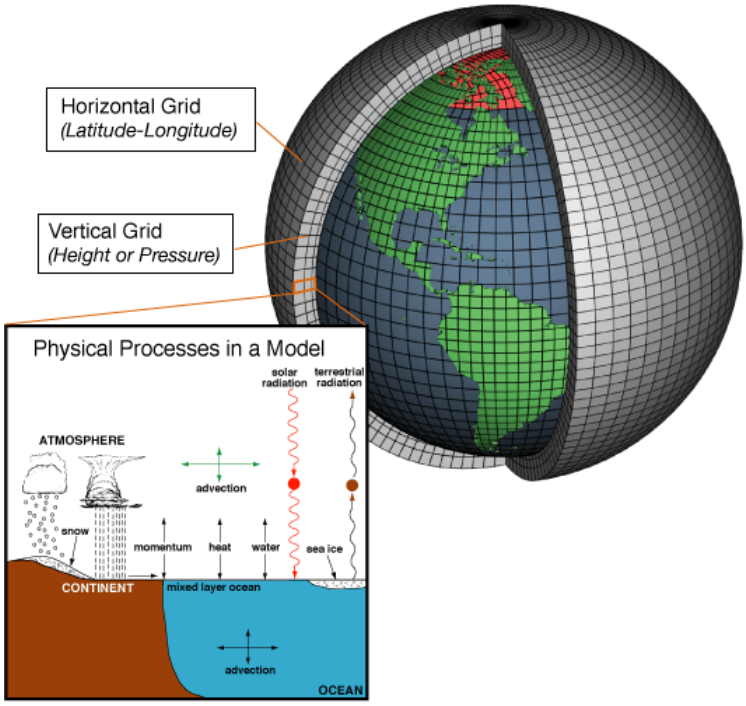


Figure 1.1: NOAA's rendering of a Global Climate Model

The statistical models used to obtain higher resolution predictions of the local

climate dynamics are known as *statistical downscaling* techniques in climatology. "Statistical downscaling is based on the view that the regional climate is conditioned by two factors: the large scale climate state, and the regional/local physiological features... From this perspective, regional or local climate information is derived by first determining a statistical model which relates large-scale climate variables (or 'predictors') to regional and local variables (or 'predictands')" Wilby et al. (2004) [29]. Once you have an adequate *statistical* model of the local climate's spatial and temporal correlations, predictions can be made by interpolating GCM future climate predictions. These statistical models have an advantage of being accessible and much less computationally expensive. In addition, since the computation cost is relatively low, we can use several GCMs output to get a local distribution of predictions for future climate.

The methods of regime cluster modeling for two-dimensional vector flows can be used as a new statistical downscaling approach that is simple, but based in climate dynamics. The hopes are that this technique can make high-resolution long term climate predictions accessible to all local governments and businesses. In addition, once models are fit, you can easily downscale several different GCMs as well as different RCP scenarios to get a distribution of high resolution predictions.

1.3 Notation

The data that will be used in this thesis will be observations evenly distributed in lattice structures in the two-dimensional space. These two-dimensional lattices occur at evenly spaced time intervals, more specifically I will be working with daily averages. This modeling technique does not need an evenly spaced structure of observations and can be applied with uneven or random spatial distribution. Wind vectors will be represented by $\psi = (\phi, x)$ with ϕ representing the direction of the

vector and x representing the magnitude of the vector. In addition, we will use:

$$\mathbf{\Psi^t} = \{\psi_1^t, \psi_2^t, ..., \psi_n^t\} \tag{1.1}$$

to represent a vector of wind observations, one lattice, of size n at time t. We can also represent one location's wind directions, say location a across times 1 to t by:

$$\phi_a = \{\phi_a^1, \phi_a^2, ..., \phi_a^t\} \tag{1.2}$$

The wind speed magnitude can also be represented by itself for one location, say the same location a across times 1 to t by:

$$x_a = \{x_a^1, x_a^2, ..., x_a^t\} \tag{1.3}$$

Similarly, a non-directional vector of $n$observations (e.g. temperature, pressure, etc) at time t will be represented by:

$$\mathbf{\Gamma^t} = \{\gamma_1^t, \gamma_2^t, ..., \gamma_n^t\} \tag{1.4}$$

CHAPTER 2

Literature Review

Statistical downscaling has been a major tool for climatologists to understand and estimate local impacts of possible future climate changes. The statistical downscaling of near-surface winds is a relatively new area of research. The downscaling techniques of precipitation and temperature are more developed and talked about in this area of research.

For my literature review I am going to first introduce some circular statistics and regression models that predict circular responses, which are needed for the understanding of directional research. Then I will introduce the general statistical downscaling techniques commonly applied to more common variables such as temperature. Followed by statistical downscaling techniques specific for wind angle, wind magnitude, or both angle and magnitude.

2.1 Circular Statistics

To be able to review literature concerning wind predictions, we must first be able to understand and talk about circular statistics. Well known measures of center (median, mean, etc.) and spread (variance, standard deviation, etc.) can not be directly applied to circular variables, such as wind direction. Even more complex are circular regression methods, models in which the response variable is a directional observation. I will start by reviewing some common statistics surrounding circular variables then move on to describing some of the most common circular

regression models.

2.1.1 Measures of Center and Spread

Since we will be using directional observations, we will be referring to a mean direction for many calculations. Given a vector of directions at a location a, $\phi_a = \{\phi_a^1, \phi_a^2, ..., \phi_a^t\}$, we will use the Jammalamadaka and SenGupta (2001) [14] mean direction formula. This is calculated for a vector of directions, ϕ_a, by treating each directional observation as a unit vector. Then the unit vectors are added component wise and the resultant direction is the mean direction, $\bar{\phi_a}$. This is found can be found by defining

$$\mathbf{R} = \left(\sum_{i=1}^{n} \cos(\phi_a^i), \sum_{i=1}^{n} \sin(\phi_a^i) \right) = (C,\, S) \tag{2.1}$$

and we will also define R to be some measure of resultant length

$$R = ||\mathbf{R}|| = \sqrt{C^2 + S^2} \tag{2.2}$$

Now the mean direction $\bar{\phi_a}$ is given by the quadrant-specific inverse of the tangent

$$\bar{\phi_a} = \arctan^*(S/C) = \begin{cases} \arctan(S/C) & \text{if } C > 0, S \geq 0 \\ \pi/2 & \text{if } C = 0, S > 0 \\ \arctan(S/C) + \pi & \text{if } C < 0 \\ \arctan(S/C) + 2\pi & \text{if } C \geq 0, S < 0 \\ \text{undefined} & \text{if } C = 0, S = 0 \end{cases} \tag{2.3}$$

We can also use $\mathbf{R}$ to find our measure of spread or dispersion for unimodal data. The resultant length R conveys to us how concentrated the angles are towards the mean direction. For example, if all n vectors point in the same direction, then R should be the same size as n. Therefore $(n - R)$ cab be thought of as one equivalent to sample variance for angles.

Jammalamadaka and SenGupta (2001) [14] also define two useful measures for distance between two angles. The distance between any two angles ϕ_α and ϕ_β can

be found by selecting the smaller of the two arc lengths between ϕ_α and ϕ_β:

$$d_\phi(\phi_\alpha, \phi_\beta) = \min(|\phi_\alpha - \phi_\beta|,\ 2\pi - |\phi_\alpha - \phi_\beta|) \tag{2.4}$$

In addition to looking at the smallest arc between two angles, the authors go on to define another measure of circular distance as:

$$d_{\phi 2}(\phi_\alpha, \phi_\beta) = 1 - \cos(\phi_\alpha - \phi_\beta) \tag{2.5}$$

This method gives the largest separation, an arc of π, a distance score of 2 and the smallest score, for two directions at the same angle, a score of 0. Both methods of distance between two angles will be used in this dissertation.

The mean direction of a circular response will be handled differently from finding the average of a set of vectors, or vector averaging. For some methods in this paper, vectors will be averaged and we will want the average vector to be weighted by the length of input vectors. Figure 2.1 visually displays the vector averaging method. Vector averaging takes all vectors and adds them together component wise. The resulting vector angle is the average angle and the resultant length divided by n is the magnitude of the average vector. As a note, finding the resultant angle, θ_1 is quadrant specific.

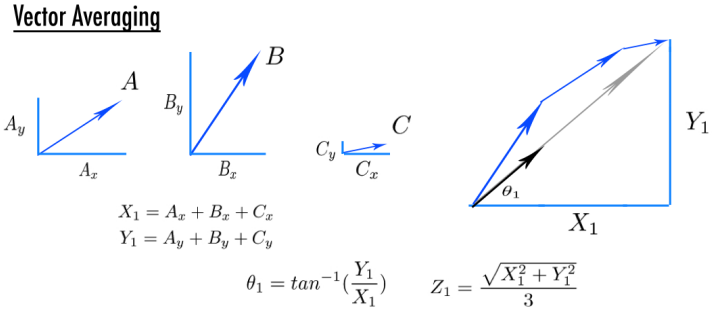


Figure 2.1: Vector Averaging Methods

2.1.2 Strength of Relationship

While using circular variables in your analysis it is also important to define the correlation between one circular variable and another and to also define the correlation of one circular variable to one linear variable.

Jammalamadaka and SenGupta (2001) [14] define circular-circular correlation, between to vectors a and b as:

$$r_{\phi_{a,b}} = \frac{\sum_{i=1}^{t} \sin(\phi_a^i - \bar{\phi}_a) \sin(\phi_b^i - \bar{\phi}_b)}{\sqrt{\sum_{i=1}^{t} \sin^2(\phi_a^i - \bar{\phi}_a)} \sqrt{\sum_{i=1}^{t} \sin^2(\phi_b^i - \bar{\phi}_b)}} \tag{2.6}$$

This is an adaptation of Pearson's correlation coefficient and returns values between 0 and 1.

Mardia (1976) [19] defines a correlational coefficient for a linear variable, x, and a circular variable, ϕ. The idea is to find multiple correlations between x and the components $(\sin\phi, \cos\phi)$ corresponding to the variable ϕ.

$$r_{x,\phi}^2 = \frac{r_{xc}^2 + r_{xs}^2 - 2r_{xc}r_{xs}r_{cs}}{1 - r_{cs}^2} \tag{2.7}$$

where

$$\begin{aligned} r_{xc} &= corr(x, \cos\phi), \\ r_{xs} &= corr(x, \sin\phi), \\ r_{cs} &= corr(\cos\phi, \sin\phi) \end{aligned} \tag{2.8}$$

If x is normally distributed and independent of ϕ then it can be shown that

$$\frac{(n-3)r^2}{1-r^2} \sim F_{2,n-3} \tag{2.9}$$

2.2 Circular Response Prediction Models

Circular statistics is a relatively new research field. Large gains in the predictive modeling of a circular response were made in the early 1990's. The only

downside was that those early models only allowed for one type of independent predictor, linear or circular. Sarma and Jammalamadaka (1993) [25] created a circular prediction model for circular independent variables. On the other hand, Fisher and Lee (1992) created a model that predicted circular responses for a set of linear predictors. Later attempts were made by Lund (1999) [17] to model circular responses with a set of linear covariates along side one circular covariate. Later, Lund (2002) [18] suggests a new form of model that allows for both multiple linear and circular covariates. In this section we will review all three methods: circular-circular, circular-linear, circular-circular/linear.

2.2.1 One Circular Covariate (Circular-Circular)

Sarma and Jammalamadaka (1993) [25] propose a model to predict a circular response, ϕ_β, using one circular predictor, ϕ_α. The idea of this model is to fit a trigonometric polynomial of ϕ_α against the sine and cosine of ϕ_β. Fitted values of ϕ_β are then obtained by taking the inverse tangent of the predicted values of the $\sin(\phi_\beta)$ divided by the predicted values of the $\cos(\phi_\beta)$.

To be more specific, the vector corresponding to ϕ_β is predicted by the conditional expectation of $e^{i\phi_\beta}$ given ϕ_α:

$$\begin{aligned} E(\cos\phi_\beta|\phi_\alpha) &= g_1(\phi_\alpha) \\ E(\sin\phi_\beta|\phi_\alpha) &= g_2(\phi_\alpha) \end{aligned} \tag{2.10}$$

from which ϕ_β will be predicted as:

$$\mu(\phi_\alpha) = \hat{\phi_\beta} = \begin{cases} \tan^{-1}(\frac{g_2(\alpha)}{g_1(\alpha)}) & \text{if } g_1(\alpha) \geq 0 \\ \pi + \tan^{-1}(\frac{g_2(\alpha)}{g_1(\alpha)}) & \text{if } g_1(\alpha) \leq 0 \\ \text{undefined} & \text{if } g_1 = g_2 = 0 \end{cases} \tag{2.11}$$

Where $\mu(\phi_\alpha)$ represents the conditional mean direction of ϕ_β given ϕ_α. Approximations of g_1 and g_2 can be computed by trigonometric polynomials of degree

m and random errors ϵ, with mean zero. The idea is to fit two trigonometric polynomials that predict the sin and cos of the response variable.

$$\begin{aligned}\hat{g_1}(\phi_\alpha) = \cos\phi_\beta &= \sum_{k=0}^{m}(A_k\cos(k\phi_\alpha) + B_k\sin(k\phi_\alpha)) + \epsilon_1 \\ \hat{g_2}(\phi_\alpha) = \sin\phi_\beta &= \sum_{k=0}^{m}(C_k\cos(k\phi_\alpha) + D_k\sin(k\phi_\alpha)) + \epsilon_2\end{aligned} \tag{2.12}$$

The order m of the trigonometric polynomial is determined by testing the significance of the $m+1$ terms in both regression models. If neither model needs the $m+1$ terms, then an order of m is used. Estimation of the parameters, (A_k, B_k, C_k, D_k), is done so in a least squares method. In addition to $B_0 = D_0 = 0$, the LS estimates of the parameters are written below

$$\begin{aligned}\hat{A}_0 &= \frac{1}{n}\sum_{i=1}^{n}\cos\phi_\beta^i \\ \hat{A}_j &= \frac{1}{n}\sum_{i=1}^{n}\cos\phi_\beta^i\ \cos\ j\phi_\alpha^i \\ \hat{B}_j &= \frac{1}{n}\sum_{i=1}^{n}\cos\phi_\beta^i\ \sin\ j\phi_\alpha^i \\ \hat{C}_0 &= \frac{1}{n}\sum_{i=1}^{n}\sin\phi_\beta^i \\ \hat{C}_j &= \frac{1}{n}\sum_{i=1}^{n}\sin\phi_\beta^i\ \cos\ j\phi_\alpha^i \\ \hat{D}_j &= \frac{1}{n}\sum_{i=1}^{n}\sin\phi_\beta^i\ \sin\ j\phi_\alpha^i\end{aligned} \tag{2.13}$$

An example of this regression model can be seen with some real wind angle data. Figure 2.2 plots the training data of the independent angle against the dependent angle in black and the model predictions in red.

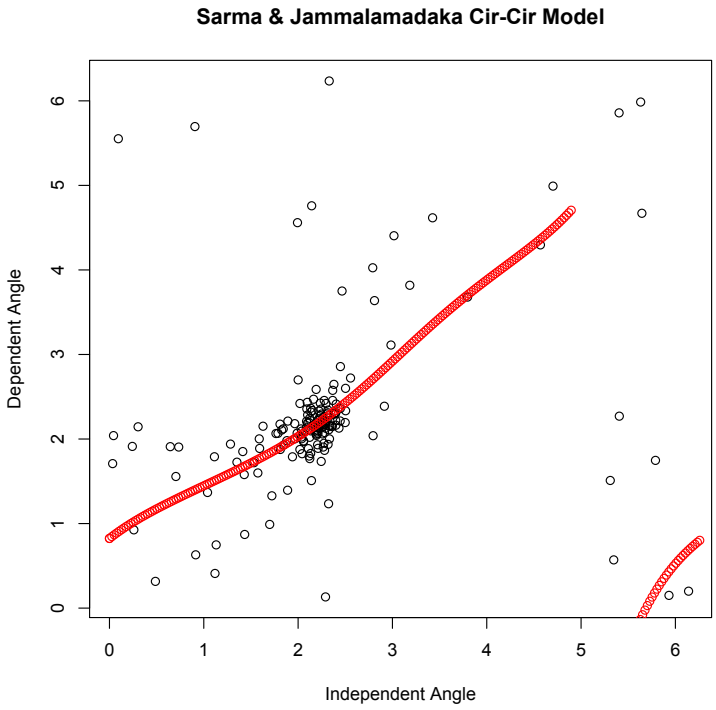


Figure 2.2: An Example Circ-Circ Sarma and Jammalamadaka Model

Another circular prediction model using just one circular response was proposed by DiMarzio et al. (2012) [4]. Specifically they propose a non-parametric smoothing solution that aims to estimate, by local averaging, a conditional mean response. This process is somewhat similar to Sarma and Jammalamadaka's [25] in that they want to solve for the arc-tangent of the ratio between the first trigonometric moments of the response variable.

To preform the process of smoothing they introduce similar statistics to the previous paper [25]. Given the vectors of the response ϕ_β and predictor ϕ_A, we are trying to find a function m such that

$$E[1 - \cos(\phi_\beta - m(\phi_A)] \tag{2.14}$$

is minimized. For any $\phi_\alpha \in$ the domain of ϕ_A, let $m_1(\phi_\alpha) = E[\sin(\phi_\beta)|\phi_A = \phi_\alpha]$

and $m_2(\phi_\alpha) = E[\cos(\phi_\beta)|\phi_A = \phi_\alpha]$. Then the sample statistics become

$$\begin{aligned} \hat{g}_1(\phi_\alpha) &= \frac{1}{n}\sum_{t=1}^{n}(\sin(\phi_\beta^t)W(\phi_A^t - \phi_\alpha)) \\ \hat{g}_2(\phi_\alpha) &= \frac{1}{n}\sum_{t=1}^{n}(\cos(\phi_\beta^t)W(\phi_A^t - \phi_\alpha)) \end{aligned} \tag{2.15}$$

where W being a local kernel weight, is derived so that the ratio $\frac{\hat{g}_1(\phi_\alpha)}{\hat{g}_2(\phi_\alpha)}$ is asymptotically unbiased for $\frac{g_1(\phi_\alpha)}{g_2(\phi_\alpha)}$. Therefore then making the estimator for the regression function at ϕ_α is

$$\hat{m}(\phi_\alpha) = tan^{-1}(\frac{\hat{g}_1(\phi_\alpha)}{\hat{g}_2(\phi_\alpha)}) \tag{2.16}$$

The same data used in figure 2.2 was used to visually show the DiMarzio et al. method. This smoothed model can be seen in figure 2.3 and contrasted to figure 2.2.

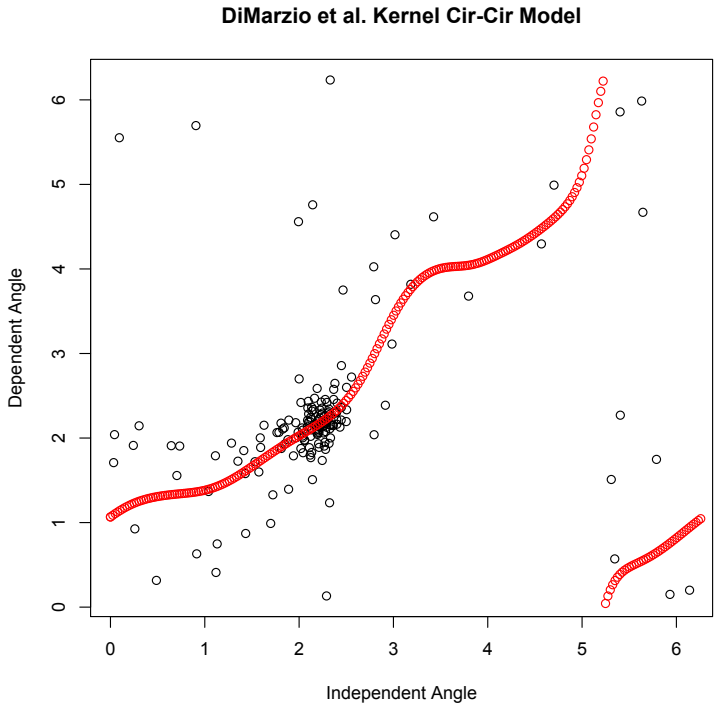


Figure 2.3: An Example Circ-Circ DiMarzio et al. Model

2.2.2 Linear Covariates (Circular-Linear)

Another sub-genre of circular models aim to predict the mean direction of a circular response with a link function of a linear independent variable. The most commonly used model was derived by Fisher and Lee (1992) [6]. It is of importance to note that the dependent circular response is assumed to be from a von Mises $VM(\mu_i, \kappa)$ distribution of n observations. The von Mises distribution can be thought of as the normal distribution extended to circular responses. The distributions are unimodal and symmetric with μ and $\frac{1}{\kappa}$ analogous to μ and σ^2, the center and spread.

To specify the model, we assume that the mean direction μ_i is related to the explanatory variable $\mathbf{X}_i$ by the regression equation

$$\mu_i = \mu + g(\boldsymbol{\beta}^{'}\mathbf{X_i}) \equiv \mu + g(\beta_1 X_1 + \cdots + \beta_k X_k) \tag{2.17}$$

The function g is our *link function* which is going to map the real line to the circle. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ contains all the regression coefficients that need to be estimated. Fisher and Lee go through several options for the link function, but we shall use the one most common to our other methods

$$g(u) = 2tan^{-1}(u) \tag{2.18}$$

The next step is to iteratively solve for the parameters $\mu, \boldsymbol{\beta}, \kappa$. Defining the following quantities

$$\mu_i = \sin(\theta_i - \mu - g(\boldsymbol{\beta}^{i}\mathbf{x}_i)) \tag{2.19}$$

$$\mathbf{u}^{'} = (u_1, \ldots, u_n) \tag{2.20}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{'} \\ \ldots \\ \mathbf{x}_n^{'} \end{pmatrix} \tag{2.21}$$

$$\mathbf{G} = \text{diag}(g^{'}(\boldsymbol{\beta}^{'}\mathbf{x}_1), \ldots, g(\boldsymbol{\beta}^{'}\mathbf{x}_n)) \tag{2.22}$$

$$S = \sum_{i=1}^{n} \sin(\theta_i - g(\boldsymbol{\beta}^{'}\mathbf{x}_i))/n \tag{2.23}$$

$$C = \sum_{i=1}^{n} \cos(\theta_i - g(\boldsymbol{\beta}^{'}\mathbf{x}_i))/n \tag{2.24}$$

$$R = \sqrt{S^2 + C^2} \tag{2.25}$$

The log likelihood used to estimate the parameters is

$$-n\log I_0(\kappa) \; + \; \kappa \sum_{i=1}^{n} \cos(\theta_i - \mu - g(\boldsymbol{\beta}^{i}\mathbf{x}_i)) \tag{2.26}$$

Where $I_0(\kappa)$ is the modified Bessel function of the first kind and order p. The solutions to the equations are then

$$\mathbf{X}^{'}\mathbf{G}\mathbf{u} = 0 \tag{2.27}$$

$$R\sin\hat{\mu} = S \tag{2.28}$$

$$R\cos\hat{\mu} = C \tag{2.29}$$

$$A_1(\hat{\kappa}) = R \tag{2.30}$$

where $A_1(\hat{\kappa}) = I_1(\kappa)/I_0(\kappa)$. To iterate, start with an initial $\hat{\beta}$ and solve for $S, C,$ and R. We then use these in an iteratively reweighed least squares method to solve for an updated $\hat{\beta}^*$. With the updating equations

$$\mathbf{X}^{'}\mathbf{G^2}\mathbf{X}(\hat{\boldsymbol{\beta}}^{*} - \hat{\boldsymbol{\beta}}) = \mathbf{X}^{'}\mathbf{G^2}\mathbf{y} \tag{2.31}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ and $y_i = u_i/A_1(\kappa)g^{'}(\hat{\boldsymbol{\beta}}^{\prime}\mathbf{x_i})$.

Large sample variances and covariances of the regression coefficient estimates can be found in the paper [6].

2.2.3 Linear and Circular Covariates (Circular-Cir/Lin)

In his paper, Lund (2002) [18], forgoes a traditional regression function and instead uses a tree-based regression method. This allows for both the use of circular

and linear covariates. In addition, the structure of the tree is interpretable and provides insight into the relationship between the circular response and the varied independent variables. Lund's methods aims to partition the training data set of circular responses into a tree of binary splits. Let us define $\mathcal{L}$ to be the set of all training circular observations. For the independent variables, we will define Φ be the set of circular responses and $\mathbf{X}$ be the set of linear and categorical responses. And let us define $\Psi = (\Phi, \mathbf{X})$ to be the set of all predictors.

"Construction of the tree begins by including all observations of $\mathcal{L}$ in the root node, the top of the tree. The data points in the root node are separated into two groups, the left and right daughter nodes, according to a binary split on one of the predictors in Φ. Roughly, the split is selected so that the two daughter nodes are as homogeneous as possible with respect to the response variable, This process is repeated for subsequent nodes in the tree until there is sufficient homogeneity, or a minimum number of observations is obtained in a node. Nodes that do not undergo a split are called terminal nodes. All observations falling into the same terminal node are assigned the same predicted value for the response variable" Lund (2002) [18].

To be more specific we will use the sample mean direction of the observed circular values in a terminal node to be the assigned predicted value of landing in that node. In the determination of node homogeneity, we will use the angular distance function $d(\phi_\alpha, \phi_\beta) = 1 - \cos(\phi_\alpha - \phi_\beta)$ to measure the distance between two angles.

"In particular, suppose the learning sample consists of N observations, and denote the sample mean direction of observations in node t by $\bar{\theta}(t)$. If node t is non-terminal, let the left and right daughter nodes of t be denoted by t_L and t_R. Finally, let S be the set of all possible splits of node t, and s be an element of S. The decrease in node homogeneity given by split s can be measured by

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R) \tag{2.32}$$

where

$$R(t) = \frac{1}{N} \sum_{\theta_i \in t} [1 - \cos(\theta_i - \bar{\theta}(t))] \tag{2.33}$$

Using the above definitions, the best split s^* of node t is taken such that" Lund (2002) [18]

$$\Delta R(s^*, t) = \max_{s \in S} \Delta R(s, t) \tag{2.34}$$

This means that the best split has two daughter nodes with the smallest summed homogeneity. To split based on a linear variable, x, we could say one daughter has all $x \leq a$ and the other daughter node has all $x > a$ for some real number a. For a independent circular variable α, a way to think of splitting would be to say that one daughter node has values in the arc (α_1, α_2) and the other daughter has all values outside that arc.

All other ideas of regression trees apply to this method too. For example, we want to ensure that the tree is not overly sensitive to the training data set. To do this several methods like bootstrapping and cross-validation can be employed.

2.3 General Statistical Downscaling Models

Now that we have established a background in circular statistics, we can review popular statistical downscaling methods as well as specific statistical downscaling methods of wind fields. In regards to general statistical downscaling methods, the simplest and one of the earliest ways researchers obtained regional climate projections is called the Change Factor Method. It works by trying to "apply coarse-scale climate change projections to a high resolution observed climate baseline (Wilby et al. [29])." For example, you could take a historical high resolution data set for your area of prediction and then add on to all observed values the mean climatological change observed in a GCM.

In more recent research, I noticed there were two different approaches or

schools of thought to general statistical downscaling methods that were currently being employed in the climate research field. One of the statistical downscaling's main focus is to model temporal relationships for each specific prediction point while the other approach's main focus is to model the entire prediction area's spatial relationships as a whole one, one time step at a time.

2.3.1 Temporal Methods

2.3.1.1 Regression Methods

The first general approach I will review are temporal methods. The temporal approach takes advantage of the fact that GCMs, RCMs and weather stations provide more than adequate amounts of data over time, at time steps as low as every hour. The simplest models in this area are called Regression Methods. They build a linear regression model for every high resolution prediction point using the closest local neighbors, or the highest correlated neighbors, from the coarser resolution.

This method has been used to predict mean rainfall and temperature over Oregon (Kim et al., 1984 [15]), the continental United States (T. Hoar and N. Nychka, 2008 [11]), and the mean daily temperatures over central Europe (R. Huth, 1999 [13]). This type of modeling shows how a high-resolution local point's climate relates over time, daily or seasonally to its low resolution area averaged local neighbors. If your dependent data set is inherently linked to your dependent data set, as in our case, then this creates techniques that preform well because they create a unique model for each prediction point. This can be very powerful because climate states are very anisotropic. Contrastly these models have limitations in that they ignore the spatial correlation of the area of interest and how that spatial correlation may change with time. This would affect the change in weighting of neighbors' cells with the overall climate state. This may come most into play

when trying to predict wind fields or precipitation.

2.3.1.2 Weather Pattern Approaches

Other more advanced regression methods attempt to leverage spatial information by making relationships conditional on the current weather state or pattern. These regression methods are Weather Pattern approaches and Stochastic Weather Generators. Weather Pattern classification approaches group days into a discrete number of weather states. The grouping can either be done by applying a cluster analysis (B.C. Hewitson and R.G. Crane, 2002 [9]), principal component analysis (D. White et al., 1991[27]), canonical correlation analyses (D. Gyalistras et al., 1994 [8]) or can also be done subjectively. "Having selected a classification scheme it is then necessary to condition the local surface variables, such as precipitation, on the corresponding (daily) weather patterns. This is accomplished by deriving conditional probability distributions for the observed data" (R.L. Wilby and T.M.L. Wigley, 1994 [30]). These models are similar in the sense that they create a unique linear regression model for every point but they remain open to the fact that those local relationships depend on the overall weather state.

2.3.1.3 Constructed Analogues Method

The Constructed Analogues method is similar to Weather Pattern approaches in the sense that it attempts to make current time predictions based on a linear regression of historical days with a similar spatial distribution. However it is very different in the sense that it does not create a model for each prediction point, but creates a single model for all points in each time slice. Constructed Analogues was produced by H. G. Hidalgo et al. (2008) [10] to predict daily precipitation and average temperature patterns for the contiguous United States. Hidalgo et al. states their method is"based on the premise that an analogue for a given coarse-scale

daily weather (target) pattern (for example, from a general circulation model simulation) can be constructed by combining the weather patterns for several days (predictors) from a library of previously observed patterns. In this application the analogue pattern is constructed at coarse scale, but a similar construction can be made using a companion library of high-resolution patterns using the same days as the coarse-scale predictors. Thus, a fine resolution downscaled estimate is created for the given pattern for that particular day." To preform this method procedurally, you need a collection of corresponding high and low resolution historical data sets. Then the historical low-resolution times with the most similarity in climate pattern to your coarse-resolution data set are linearly regressed upon the GCM coarse-resolution current prediction time. This regression will give the highest weights to days most similar in distribution to you current time. Then you use the same linear combination of weights on their corresponding high resolution times to make a constructed high-resolution climate analogue. This method's strengths are that the weather patterns produced are dynamically consistent with the high-resolution topography. However, its drawback is assuming that future climate distributions are the exact same as past historical distributions. The second drawback is that a historical time slice may be very similar to you current time in a specific area, like the mountains, but be very different in another, like the desert, but that all points within that one analogue receive the same weight.

2.3.1.4 Stochastic Weather Generators

The second group of models that condition local probabilities on overall weather states are Stochastic Weather Generators. "Richardson's (1981) WGEN model is the most commonly used for climate impact studies: this was originally designed to simulate daily time-series of precipitation amount, maximum and minimum temperature, and solar radiation for the present climate. Rather than being conditioned by circulation patterns, all variables in the Richardson model are sim-

ulated on precipitation occurrence. At the heart of all such models are first- or multiple-order Markov renewal processes in which, for each successive day, the precipitation occurrence (and possibly amount) is governed by outcomes on previous days (R.L. Wilby and T.M.L. Wigley, 1994 [30])". One way this method has been applied in the literature is as a nonhomogeneous hidden Markov model (NHMM). B. C. Bates et al. (1998) [1] used the NHMM to predict atmospheric and precipitation variables over South-West Western Australia with success. This Richardson model can be written out for daily precipitation occurrence as a two-state Markov chain:

$$p_{01} = Pr\{\text{precipitation on day } t | \text{no precipitation on day } t-1\} \tag{2.35}$$

$$p_{11} = Pr\{\text{precipitation on day } t | \text{precipitation on day } t-1\} \tag{2.36}$$

Then the unconditional wet day probablitites (π_w) and the lag-1 autocorrelation (r) are:

$$\pi_w = \frac{p_{01}}{1 + p_{01} - p_{11}} \tag{2.37}$$

$$r = p_{11} - p_{01} \tag{2.38}$$

When stochastic weather generators were used in R.L. Wilby et al. (1998) [31], daily precipitation amounts were modeled as independent gamma variates, with probability density:

$$f(x) = (x/\beta)^{a-1} \frac{e^{-x/\beta}}{\beta\Gamma(a)} \tag{2.39}$$

All these temporal regression methods perform well in predicting mean climate states in the literature. This is expected since linear regression methods aim to keep residuals low by predicting the mean state. This has its drawbacks though. This implies, and has been seen in the literature (R.L. Wilby et al. (2004) [29]), to miss extreme events or daily highs and lows. This is somewhat troubling since the overall aim of GCMs and downscaling in general is to predict unnaturally high or low climatological events that will occur in the near future.

2.3.2 Spatial Methods

2.3.2.1 Inverse Distance Weighting

The second general approach makes spatial correlation the key model building relationship, but these models rely much less on the availability of temporal relationships. These techniques are rooted in the Geostatistical framework of statistics and hold the key theory that points closer to the prediction area will be more correlated than points further away. The techniques in this family look at the spatial correlation of all the points in one time slice and interpolate the climatic variable as a weighted average of surrounding points. Assigning weights to each point based on the relationship of distance and correlation. The simplest and first methods of these types were Inverse Distance Weighting methods (IDW). These IDW methods are simple and assign the highest weights to the closest neighbors in a linear manner ignoring anisotropy. The most commonly used form of IDW is Inverse Distance Squared Weighting (IDSW).

Another offshoot of IDSW is the Gradient Plus Inverse Distance Squared method (GIDS). This method combines multiple linear regression with IDSW. First, a multiple linear regression of your prediction variable is fit to latitude, longitude, and elevation to remove first order trends. Then you can use this first order gradient to help weight the point's neighbors. GIDS was developed by I. Nalder and R. Wein (1998) [23] and it and IDSW were used to predict monthly averages of temperature and precipitation over the western Canadian forest. IDW was also use by G. Tabios and J. Salas (1985) [26] to predict annual precipitation over Nebraska and Kansas. Similar to the GIDS method are Spline methods. These Spline methods fit polynomials to the set of observed values to provide a smooth surface that passes through all observed points with as little curvature as possible. Then high resolution points can be predicted by using the value predicted by the spline at that location. These techniques have the obvious problem

of over smoothing. This is especially exacerbated by the fact that GCM values represent area averages. Therefore using a GCM prediction at a specific point to create a smooth surface not only smoothes over the highly topographical varied land of southern California, but it also smoothes again by using averages as point predictions.

2.3.2.2 Kriging Methods

Kriging methods are considered in the spatial prediction methods but are more sophisticated in their estimation of anisotropy and their estimates of error at prediction points. Kriging methods hinge on the estimation of the variogram $2\gamma(.)$ or semivariogram $\gamma(.)$. The variogram attempts to model the spatial variance between two points dependent only on the distance separating the two points. Kriging is a linear least squares estimation with all points in one time being used to estimate and model the variogram. Then the variogram is used to assign weights to observed values to predict high resolution un-observed values through linear combination:

$$\hat{Z} = \sum_{i=1}^{N_v} w_i X \tag{2.40}$$

Where the weights $\mathbf{W} = \{w_1, w_2, ..., w_{N_v}, \lambda\}$ are found by solving the system:

$$\mathbf{W} = \mathbf{\Gamma}^{-1}\gamma \tag{2.41}$$

Where λ is the Lagrange multiplier and $\gamma = (\gamma(s_0 - s_1), \gamma(s_0 - s_2), ..., \gamma(s_0 - s_{N_v}), 1)'$ and

$$\mathbf{\Gamma} = \begin{cases} \gamma(s_i - s_j) & i = 1, 2, ..., n, \;\; j = 1, 2, ..., n \\ 1 & i = n+1, \;\; j = 1, .., n \\ 1 & j = n+1, \;\; i = 1, .., n \\ 0 & i = n+1, \;\; j = n+1 \end{cases} \tag{2.42}$$

Since kriging methods were developed first for ore mining, these method assume first order stationarity, which is almost never a good assumption for dynamic

climatic variables. However, more advanced methods of kriging (e.g. universal kriging and de-trended kriging) have been developed to help with this assumption. These methods are similar to Spline methods in that they first model the first order trend and then model the remaining residuals. Co-kriging is another method in the kriging family, but it uses the relationships with additional correlated climatic variables to help improve predictions. For example N. Diodato (2005) [5] uses terrain elevation data and a topographic index to help predict annual and season precipitation for a complex mountainous region in southern Italy (Benevento province). One major advantage of using kriging techniques is the availability of off the shelf computer packages ready to compute all the different forms of kriging in a quick and efficient manner. Papers like I. Nalder and R. Wein (1998) [23] and G. Tabios and J. Salas (1985) [26] were able to try several kriging methods for the same area to find the most consistent model.

In order to help with the over smoothing problem associated with kriging, Area-to-Point Kriging methods have recently been developed to handle the fact that GCM products are not points but instead area averages. P. Kyriakidis (2004) [16] developed this method and is closely related to the idea of block or point-to-area kriging.

Given these two very different approaches, the spatial and temporal, I believe the temporal approach will be more applicable to wind fields. The overall complexity of topographic area and changing direction of influence leads spatial methods to be inadequate. In addition to building models for each specific point, we will allow the spatial dependence structure to change with wind regime change, similar to weather pattern approaches.

2.4 Wind Specific Downscaling Models

Wind models are a specific sub-division of climatological downscaling methods because of their inherent circular nature. Most statistical downscaling methods found have sought to avoid circular statistics all together and instead treat each wind vector as a projection onto both the x and y axis. Then separate models are built to predict the resulting x projection length and y projection length. Other methods in which wind vectors are represented as polar coordinates are less investigated. In this section we will start with the review methods which predict just wind magnitudes, avoiding direction all together. Followed by reviewing models which predict magnitude and angle through projection based methods. Finally, we will review papers which predict the angle of the wind direction alone.

2.4.1 Wind Speed Magnitude Prediction Methods

Some papers are only interested in the prediction of wind speeds. These models may be helpful to understand overall turbulence, but fail to detect Santa Ana winds or correct orientation for wind farms. The dependent variable in their models is the length of the wind vector at the high-resolution location. In W. de Rooy and K Kok (2004) [3] they aim to apply a regression based technique to wind speeds at several weather stations in the Netherlands. The idea of their method is to decompose the total error of the predictions into two parts: representation mismatch (RM) and a large-scale model error ($\bar{\text{M}}$E). The hopes of this decomposition is to first address the physical differences in the topography of the low-resolution GCMs and the prediction locations with the representation mismatch and secondly to account for the large-scale model error by linearly regressing the new RM local estimates on the corresponding observations. Specifically to address the physical differences in low-resolution and high-resolution topography, the authors take the low-resolution wind speed estimates from higher in the atmosphere and

then use high-resolution land roughness estimates to account for local speed disturbances. They then use this new local estimate as a independent variable in the regression to predict the high-resolution location.

This method has the benefit of adding actual local based topography which improves covariates strength of relationship to the high-resolution winds. But, this method does not account for the change in spacial dependence with the change in regime. The new independent variable, created by RM is the same for all times and wind regimes. In addition the location of the independent variable is always assigned to the closest local estimate, which may not be best related.

2.4.2 Projection Based Methods

Projection based downscaling aims to predict both wind direction and magnitude but in an indirect way. The model does this by taking the high resolution wind vector at ten meters and projecting it onto the x-axis (u_{10}) and y-axis (v_{10}). Then modeling the relationship between the linear covariates and u_{10} and v_{10} separately. Projection based methods typically do not think of prediction error as the sum of direction and magnitude error, but as the error coming from the x-axis projection and error coming from the y-axis projection. This makes model performance analysis less directly interpretable.

Projection based methods were found to be the most common form of the statistical downscaling of winds for future climate change studies. I will review several papers that use this method, but differ in the modeling types and covariate selection. This projection based method is used to predict sea-surface winds in the subarctic northeast Pacific Ocean (A. Monahan 2011 [22]), wind fields over southern France (T. Salameh et al. 2008 [24]), sea-surface winds off of Peru (K. Goubanova et al. 2011 [7]), and tropical ocean surface winds (C. Wikle et al. 2001 [28]).

A. Monahan (2011) [22] is an example of the simplest projection based method. The aim of the paper is to predict sea surface winds in the subarctic northeast Pacific off the coast of Canada. Monahan keeps his model fairly simple by predicting wind components u, v, and magnitude for high-resolution buoys from linear regressions of principal components from the low-resolution u, v, magnitude, and temperature variables at different pressure levels.

T. Salameh et al. (2008) [24] use a similar projection based method to predict near-surface winds over southern France but add some complexity in the model selection. The prediction area of this paper is similar to southern California in that it contains coastal regions and mountainous areas making it quite complex. To account for the spatial complexity, they first look into subsetting their data into weather regimes. To do this they group daily geopotential height anomalies at 500 hPa (Z500) from the ERA40 re-analysis data set , over a domain covering the north east Atlantic ocean. They then take the Z500 observations and perform principal component analysis on the re-analysis data. They end up retaining the first ten eignenvectors, which explain 85% of the variance, and then perform k-means clustering upon them to get the final four regimes. After some investigation they decide that the weather regime clustering is not helpful and will not use it for their model.

The model they decide upon to make predictions for u_{10} and v_{10} are Generalized Additive Models (GAM) that take the sum of spline functions applied to different covariates X_j

$$\begin{aligned} u_i &= \sum_{j=1}^{p} f_{i,j}^{u}(X_j) + \epsilon_{u_i} \\ v_i &= \sum_{j=1}^{p} f_{i,j}^{v}(X_j) + \epsilon_{v_i} \end{aligned} \tag{2.43}$$

where i corresponds to the high-resolution location, j indicates an independent variable, p is the number of independent variables, u_{10} and v_{10} are the wind

components at location i, $f^u_{i,j}$ and $f^v_{i,j}$ are spline functions, and ϵ_{u_i} and ϵ_{v_i} are the respective model errors. For the spline functions, they choose piecewise third-order polynomial functions. The covariates used in the modeling process are: surface pressure gradient, 500 hPa relative voracity, near-surface pressure gradient, low-level winds at 925 and 850 hPa, and geostrophic wind at 700 hPa.

The results of their technique show that the model can explain one component accurately, while the other is much lower. The component which is well explained switches from location to location. It is hard to say if the lack of performance in one component was due to not clustering or separately modeling the u and v components. The paper also does not say how the location of the covariates was chosen. For some covariates it does not matter because they are gradients, which are a combination of several locations.

The third example of projection based methods is by K. Goubanova et al. 2011 [7]. They use a principal component based linear regression to predict sea-surface winds over the Peru-Chile upwelling region. To be more specific the want to predict the projections of the wind anomalies, which is the wind vectors components with the seasonal mean subtracted from them. The model is based on multiple linear regression (MLR) and uses low-resolution 10m wind components and sea level pressure as covariates.

In this paper the predictor-predictand relationship is made in the principal component space. Both independent and dependent variables will be eigen vectors. A covariance matrix form the the high-resolution dependent wind components, u and v, has PCA performed on it and the first ten eigen vectors are retained, PCQ_i for $i = 1, \ldots, 10$. The independent covariates are represented by the first twenty eigen vectors resulting from the correlation matrix containing sea level pressure, and u and v components from a re-analysis data set. The resulting model is

$$PCQ_i = \sum_{j=1}^{20} \alpha_{i,j} PCN_j + \epsilon \tag{2.44}$$

where $\alpha_{i,j}$ are the regression coefficients, PCN_j are the PCs of the low-resolution variables, and ϵ is the model error.

In this paper they also add some complexity, by allowing the location range of the low-resolution variables to change with prediction location. But, they do not allow the location of the domain to change with regime change.

In a somewhat related projection based method, C. Wikle et al. (2001) [28] predict tropical ocean surface winds. This paper is a little different in that they are looking to combine two disparate wind data sources to make one more complete wind data set. What is common to our problem is the structure and nature of these data sets. One of the data sets comes from high-resolution satellite observations. While the second wind data set comes from a low-resolution re-analysis data set. Wikle et al. aim to combine these two data sources by performing a hierarchical Bayesian spatiotemporal model on the u and v wind components.

This model takes the form of three stages. Stage 1 is the "Data Model", $[data|process, \theta_1]$. The purpose of stage 1 is to model the measurement error. Stage 2 is the "Process Model", $[process|\theta_2]$. The purpose of stage 2 is to formulate a joint probability model for the wind processes of u and v. They do this by decomposing the wind process into three physically interpretable components.

$$\begin{aligned} u_t &= \mu_u + u_t^E + \check{u}_t \\ v_t &= \mu_v + v_t^E + \check{v}_t \end{aligned} \tag{2.45}$$

where μ_u and μ_v are spatial means, u_t^E and v_t^E are component contributions from the thin-fluid approximation, and $\check{u}_t$ and $\check{v}_t$ represent small scale motions. Stage 3 is the estimate of the "Prior on Parameters", $[\theta_1, \theta_2]$ To estimate the posterior distribution of all the unknowns in this process, Wikle et al. use the MCMC method of a Gibbs sampler to estimate them.

Even though this paper's aims are different from ours, we can see some similarities and can learn from their diversity of thought.

2.4.3 Wind Angle Prediction Methods

In a new approach U. Lund (2006) [18] aims to predict wind angles, but with both linear and angular variables as covariates. The methods of this tree-based regression are reviewed in section 2.3.3. In his paper Lund aims to predict wind directions at PointSan Luis from covariates measured at Point Conception. Both locations are buoys off the central coast of California. The mixture of linear and circular covariates used are: wind direction, wind speed, barometric pressure, temperature, and time of day. Time of day is treated as a circular response.

Lund's method has the advantage of being able to utilize both linear and circular variables. Which may be a big advantage depending on the relationship on the high-resolution wind angle and the covariates. It also somewhat mimics segmenting the data into regimes. When the tree splits the data, it does so by selecting more homogeneous daughter nodes. The drawbacks of this method may be over fitting and over simplification. Also, decisions in tree splits need to be done so binomially, which may not actually be the case, when a more linear or non-linear relationship may be present.

CHAPTER 3

Data: Source and Motivation

The modeling process explained in this research uses observations at low resolution as the covariates for prediction, these observations can either be evenly spaced as a lattice or unevenly distributed in space. For the example in this work, observations come from a reanalysis data set. A reanalysis data set is a data set created by taking physically observed climate data and then interpolating it onto a system of grids. These grids of data are then used to drive a three-dimensional dynamical model in order to obtain the remaining unobservable climate variables.

Models can be built with covariates that are only observations by using the methods in these paper, but the motivation for the application calls for the use of a reanalysis data set.

The dependent data set used in this research is also latticed and comes from the dynamical downscaling of the independent reanalysis data set. By modeling the dynamical downscaling process, we ensure the climate dynamics of our statistical model is representative of true local dynamics.

Dynamical downscaling is the other approach to obtaining high-resolution local climate predictions. Dynamical downscaling, relative to statistical downscaling, is computationally expensive and technical. Dynamical downscaling works by nesting multiple dynamical models, this somewhat keeps the overall computational costs down by restricting the inner dynamical model to the confines of a local prediction range, this process is also known by the name limited-area Regional Climate Models (RCM). Then the inner nested model is initialized and main-

tained by forcing its local boundaries with predictions from the lower resolution GCM. Typically you will nest up to three dynamical models, each inner model having a higher resolution and a smaller prediction range. After the largest nested dynamical model is initialized at its boundaries with the predicted GCM data, its resulting predictions are then used to force the boundaries of the next inner most model and so on. Figure 3.1 is an example of dynamical nesting over the Sierra Nevada Mountain Range. The outer most resolution is 36 kilometers (km) followed by 12 km for the middle model, and then 4 km for the inner most model. You could see how this process consumes a lot of computing power, dynamically downscaling the climate three times consecutively.

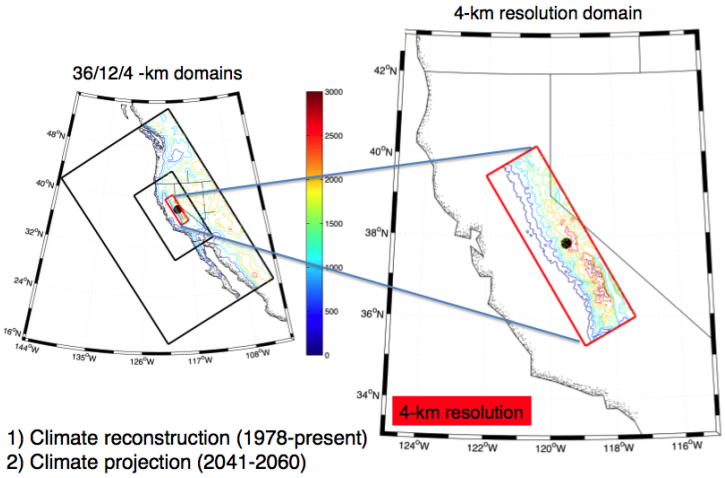


Figure 3.1: Dynamical Model Nesting

The *dynamical* downscaling process we will be trying to statistically model is a complex process. As you may recall the overall process starts by forcing the boundaries of the local outer-most nested dynamical model with the data from a global climate model. We first start with a data set that we believe accurately represents the southern California weather variability. This data set is the NCEP North American Regional Reanalysis (NARR) [20]. Once we obtain twenty years of NARR data, 1981-2000, which can be downloaded from NCEP's website, we can

use these predictions to drive the dynamical Weather Research and Forecasting (WRF) [21] model to get three levels of a higher-resolution representation of past climate.

The last layer of complexity is the presence of three forcing events on dynamical models before we get our final high resolution product (see figure 3.2). To get the future predictions the initial forcing is done with the NARR perturbed data set, as explained above, and this is a transition from 32 km to 18 km resolution. Then, only the the data created by the initial dynamical downscaling that surrounds the boundaries of the next model will be used to force the next nested dynamical model, this is a transition from 18 km to 6 km. This is done one more time to achieve the highest resolution and smallest prediction area over southern California, this is a final transition from 6 km to 2 km resolution.

In the creation of my statistical downscaling technique for this research, I will not attempt to model all three transitions from lower resolution to a higher resolution. Instead I will try to estimate this process by looking only at the very first and last data sets created by this process. I will train my model by looking at the relationship between the historical NARR data set and the resulting WRF domain 3 (2 km) historical representation. We will test the performance of model by training our model using the first ten years of data, 1981-1990, and then compare our predictions with the last WRF future predictions made at a resolution of 2 km, from 1991-2000. You may be able to explain more variance by modeling every transition, but this may defeat the purpose of statistical downscaling to reduce computation and complexity.

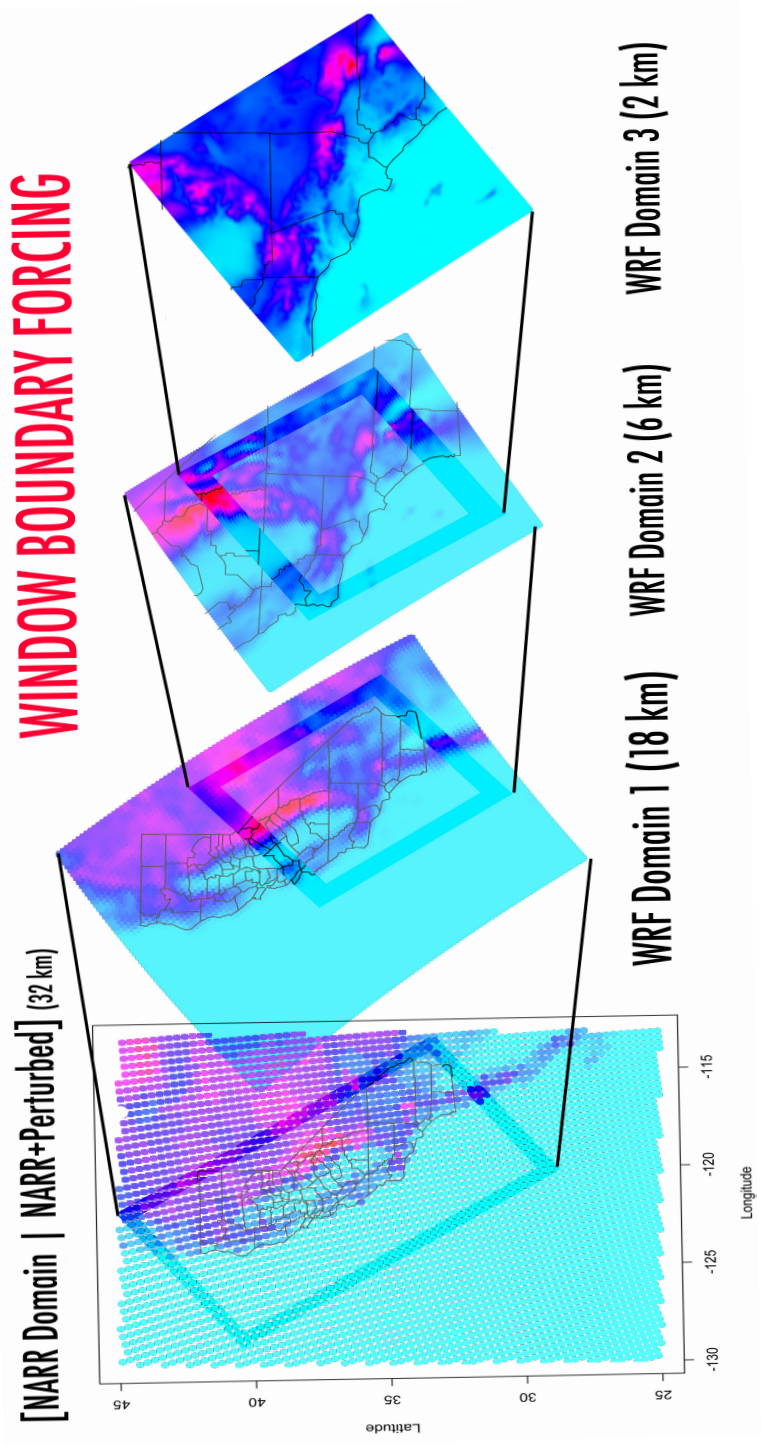


Figure 3.2: Dynamical Model Forcing Example

CHAPTER 4

Methods

The motivation for this new modeling technique was to simulate the physical process of local wind dynamics. Therefore the backbone of this new technique is the adaptation to changes in influence. I feel adaptation needs to be at the forefront because each individual prediction location is unique. The prediction of each location is not only unique when compared to other points in the two-dimensional space, but each location is different even within itself. Because of local climate dynamics and topography, within one location, dependence structures change with time and the direction of the major flow. This is especially true for coastal and mountainous regions, both of which make up the southern California landscape.

In order for this technique to properly adapt to each changing wind regime, we must first be able to accurately group days into distinct wind regimes. After we group or cluster the training period into several distinct wind regimes, then distinct models can be trained within each regime. These distinct models allow the locations and the strength of influence of each of the independent variables to change from regime to regime. This allows the model to adapt to the physical influences that channel and redirect winds from different directions.

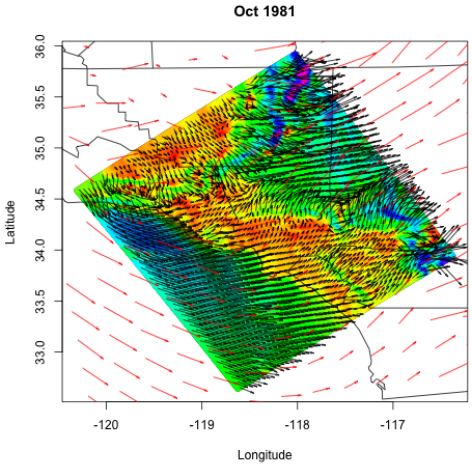
We will use a transformed linear models to predict the wind's magnitude within each regime and we will use a circular kernel smoothing regressions to model the wind's direction within each regime. Therefore, to predict the wind speed and direction for one location there will be a total of six models. Three models

for predicting magnitude and three models for predicting direction. The three models represent the three different wind regimes of southern California (Conil and Hall, 2006 [2]). See figure 4.2 for visualizations of these three regimes. In these figures, the red vectors represent the low resolution predictor covariate vectors and the black vectors represent the corresponding high resolution predictand wind field. Figure 4.1a represents the typical cool onshore breezes, figure 4.1b displays mild wind days, the most common wind regime in southern California, and 4.1c represents offshore hot winds created by highly dense air over the high desert.

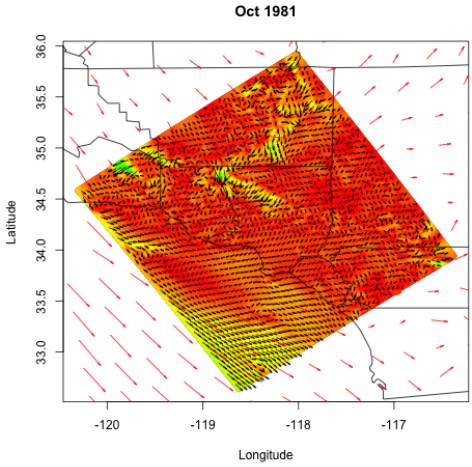
By changing the time period of your training data set, you can determine how specific your models are. For example, you can create one regime based cluster modeling group for the entire year, capturing all seasons. On the opposite end of the spectrum, you can do stepwise cluster modeling for each specific month of the year. Tuning each model set to the specifics of the particular month. For this paper, the methods are built upon the data from the month of October. I chose to use October data because it is within the Santa Ana wind season and has daily averages that resemble all three distinct regimes of southern California, hence giving us a diverse and rich data set.

4.1 Step 1: Vector Clustering

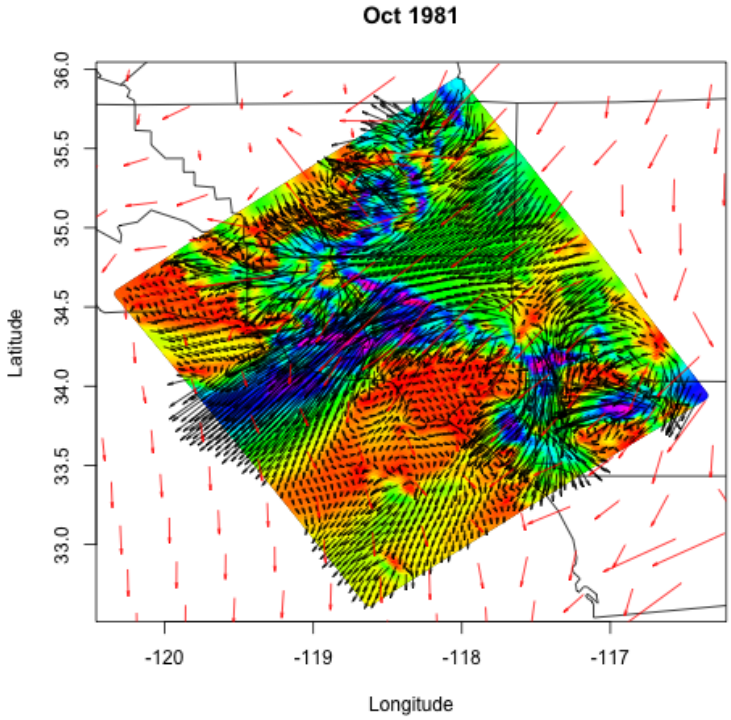
As stated above, the first step of this vector prediction technique is to be able to group similar days together to represent the major flow regimes. The process starts by collecting the model training data vectors that represents the daily general conditions surrounding the prediction location. We will then cluster these days, or representative vectors, into three groups. This is because in southern California there are three major wind regimes, with two major directional flows [2]. The two-directional flows are an onshore wind and an offshore wind referred to as Santa Ana winds. These Santa Ana winds represent one regime (fig 4.1c) and the



(a) Onshore Wind Regime



(b) Mild Wind Regime



(c) Santa Ana Wind Regime

Figure 4.1: Southern California's Three Wind Regimes

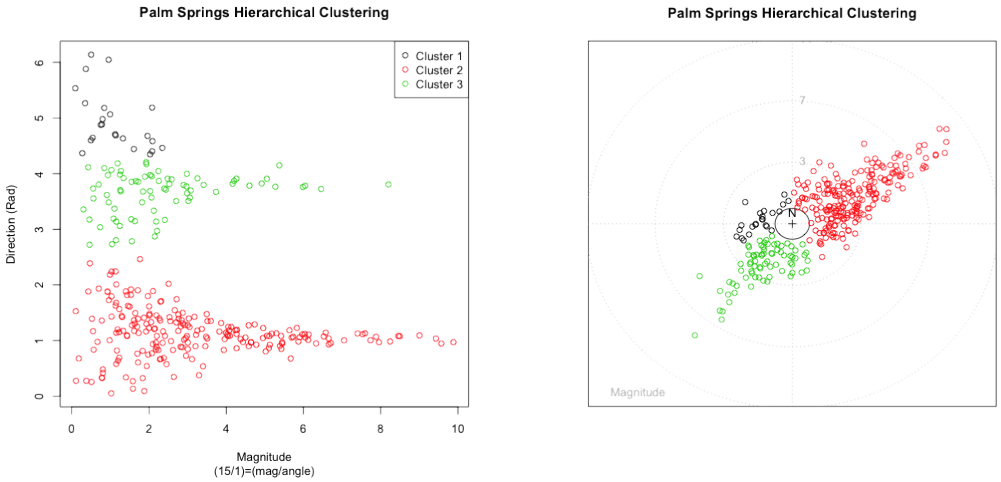
onshore flow is broken into two regimes, one for strong days (fig 4.1a) and another for mild days (figure 4.1b). This method can be applied to locations or scenarios with more or less than three wind regimes. To do so this method can be adapted by changing the cut off point in the hierarchical clustering tree, creating more or less clustered groups/regimes.

In order to classify individual days as Santa Ana's, strong onshore winds, or mild onshore winds there was a need to create a new vector clustering method. This vector clustering method was adapted from the fundamentals of the bottom up hierarchical clustering methods with complete linkage. A bottom up hierarchical clustering method, starts with each observation being its own unique cluster of size one. Then the first two closest clusters and grouped into a new higher level cluster. This clustering continues up the tree until we have a complete structure with one group of all points at the top. Distance between two clusters will be measured with the complete linkage method. This means that the distance between two clusters is measured by the maximum distance between any two points within both clusters.

In order to adapt a hierarchical clustering method for vectors we must adjust how distance is measured between vectors. Since vectors have a direction and magnitude, $\psi = (\phi, x)$, you can not apply typical methods of measuring distance, like Euclidean distance (4.1), which are typically used for hierarchical modeling.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.1}$$

To help understand how we will measure distance between vectors it will help first to be able to plot them. Vectors can be plotted in two ways. One way is with the standard Cartisian coordinate system, displaying the vector's magnitude on the x-axis and the vector's direction on the y-axis (fig 4.2a). The other way is to display the vectors circularly with magnitude displayed as distance from the center of the diagram and direction of the vector is displayed as the circular placement



(a) Cartisian Vector Plotting (b) Circular Vector Plotting

Figure 4.2: Plotting Southern California's Wind Regimes

of the observation (fig 4.2b). Cartisian plotting helps distinguish regimes, but it is misleading in displaying distance on the y-axis. As the points on the bottom of the graph, around zero radians, are close to the points on the top of the graph, around 2π.

Now that we can visualize these distinct regimes, we still need a way of measuring the distance between vectors, let us call them $\psi_1 = (\phi_1, x_1)$ and $\psi_2 = (\phi_2, x_2)$, in order to cluster them properly. To adjust for the circular nature of the vector's direction we will use d_ϕ to represent angular distance between any two directions ϕ_1 and ϕ_2:

$$d_\phi = \min(|\phi_1 - \phi_2|, 2\pi - |\phi_1 - \phi_2|) \tag{4.2}$$

We can measure the difference or distance between any two magnitudes, d_x, by subtraction:

$$d_x = (x_2 - x_1) \tag{4.3}$$

After establishing a way to measure the angular and magnitude distances between vectors, we can adapt the common Euclidean distance formula 4.1 to fit our needs. Using the distance between the angles of two vectors instead of the

distance between y projections, $(y_1 - y_2)$. We then get a formula to measure the distance between vectors $\psi_1 = (\phi_1, x_1)$ and $\psi_2 = (\phi_2, x_2)$ to be:

$$\begin{aligned} d_v &= \sqrt{(x_1 - x_2)^2 + \min(|\phi_1 - \phi_2|, 2\pi - |\phi_1 - \phi_2|)^2} \\ &= \sqrt{d_x^2 + d_\phi^2} \end{aligned} \tag{4.4}$$

The last thing to consider is scale. Since magnitude and angle are measured on completely different scales, it would be unfair to leave values as is in our distance formula 4.4. For example, if we were to measure angles in degrees, from 0 to 360, these angular distances would overshadow the distances between magnitudes, in which a large wind magnitude is 10 units. Magnitude and angular distance would not be contributing equally to the distance formula. Therefore, we will first standardize both variables magnitude and direction. Let x_{std} and ϕ_{std} represent standardize values or magnitudes and directions respectively. In future sections it will be assumed that magnitude and angle vectors are standardized before clustering, leaving off the sub notation of "std".

$$x_{std} = \frac{(x - \bar{x})}{\sigma_x}, \qquad \phi_{std} = \frac{(\phi - \bar{\phi})}{\sigma_\phi} \tag{4.5}$$

Now that are magnitudes and angles are standardized we can assign weights to each distance in our formula 4.4. Assigning weights gives us one more tool to further tune clustering for model performance. Assigning a larger weight to angle (w_ϕ) emphasizes grouping wind regimes on direction, while placing a larger weight on magnitude (w_x) separates winds into mild, med, and strong wind clusters no dependent on their direction. We will address finding the optimal weighting ratio in the coming sections. This gives us our **final** vector distance measure, d_ψ, formula:

$$\begin{aligned} d_\psi &= \sqrt{w_x(x_2 - x_1)^2 + w_\phi[\min(|\phi_1 - \phi_2|, 2\pi - |\phi_1 - \phi_2|)^2]} \\ &= \sqrt{w_x d_x^2 + w_\phi d_\phi^2} \end{aligned} \tag{4.6}$$

There are several points to address and think of when creating the optimal cluster groupings. Because in theory, the better the cluster assignment represents the different regimes, the more homogenous the individual spatial dependence structures are. Reducing the noise in the spatial dependence structure results in stronger prediction models that are more physically based on the influences of geography on wind. Now that we have established a method for clustering vectors, the three explorations we will undertake in this chapter to improve cluster predictions are: "What vector should we use to represent each average daily flow?", "What weighting scheme should be used when determining distance between two vectors?", and "Does one weighting ratio fit all?".

4.1.1 Point Representation or Neighbor Schemes

Each point in the plots above, figures 4.2a and 4.2b, represent a daily average. But where did that data come from. How do we find a vector that represents the specific wind regime or flow direction around our prediction point? It would be optimal to use the actual high-resolution wind vectors, for the point you want to predict, to cluster days. But if clustering is done upon these high-resolution predicted winds, then when you want to predict future winds, what are you going to use to assign the day into a regime? You can't use the high-resolution wind vector because that is exactly what you want to predict.

Therefore, we are to use low-resolution independent vectors as a representations of what wind regime is occurring at and around the prediction location. The question then we must answer are: Does the closest low-resolution wind vector best represent the local point? What about an average wind vector of nearest neighbors? Can one location cluster days for all prediction points?

To answer these questions, in this section we will try several of the possible representations and choose the most robust method. The different daily vector

representations for a point we will try are: using the closest low-resolution wind vector (NARR 1), averaging the two closest low-resolution wind vectors (Flow 2), averaging the three closest vectors (Flow 3), up to averaging the five closest wind vectors (Flow 5) to your prediction point. Figure 4.3 illustrates how different clustering assignment can play out depending on representation choice. The graphics within 4.3 are attempts to represents the daily average 10m winds for a point in coastal Malibu. The top left figure is the actual daily averages for the high-resolution location. The following figures represent the 6 different schemes tried. The 310 points shown in each graphic are 10 years of October data (31x10) spanning from 1981 to 1990.

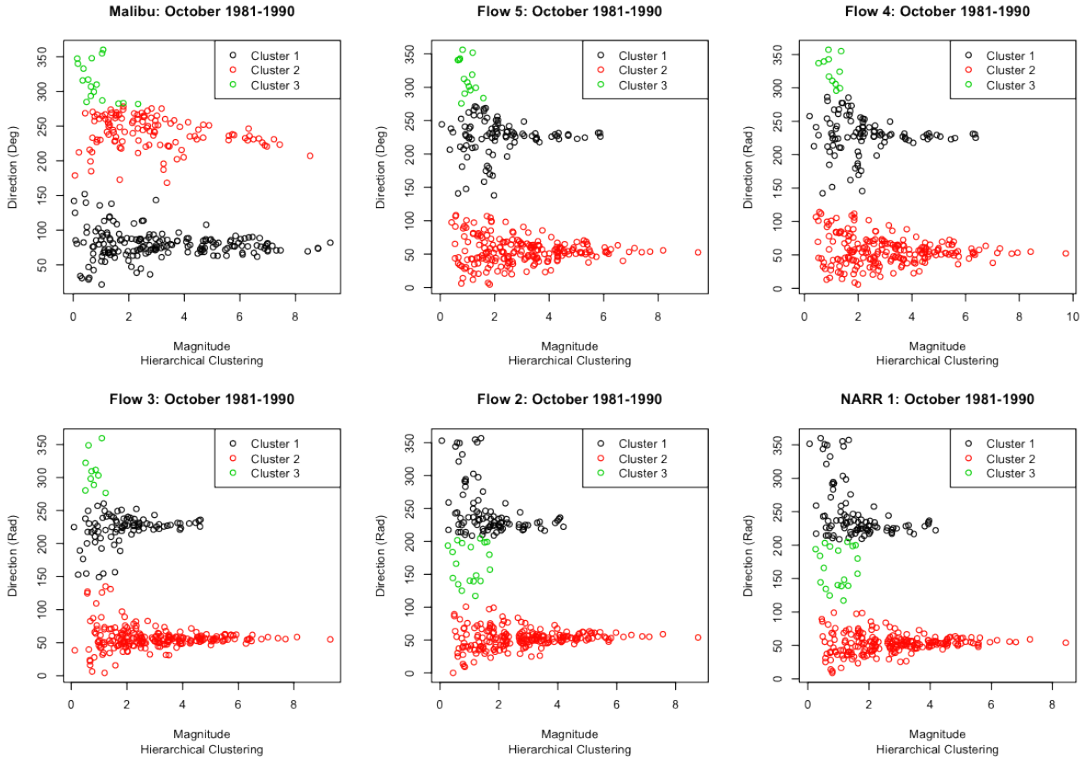


Figure 4.3: Different Malibu Clustering Outcomes for changing Neighbor Schemes

In order to find a robust method, five high-resolution local points were chosen to represent the varying southern California landscape. These points not only vary in spatial location, but also in elevation, terrain, proximity to the coast, and

mountainous surroundings. The first location chosen, Malibu, is a coastal point which lies in a main channel for the Santa Ana winds. The second location, Big Bear Mountain, is the highest peak in our prediction window. The third location, a point in the High Desert near Edwards Air Force base outside of Palmdale, is a desert landscape with little change in topography but a source of Santa Ana winds. The fourth location, Palm Springs, is also a desert landscape but it lies in a valley behind the San Jacinto Mountains and is in the south eastern corner of our prediction window. The last location, was a point in the Central Valley, it is the closest to the edge of the prediction boundary and lies below Bakersfield. These locations are seen in figure 4.4 along with some addition points that will be used in future analysis (note: county outlines are slightly shifted north of real locations).

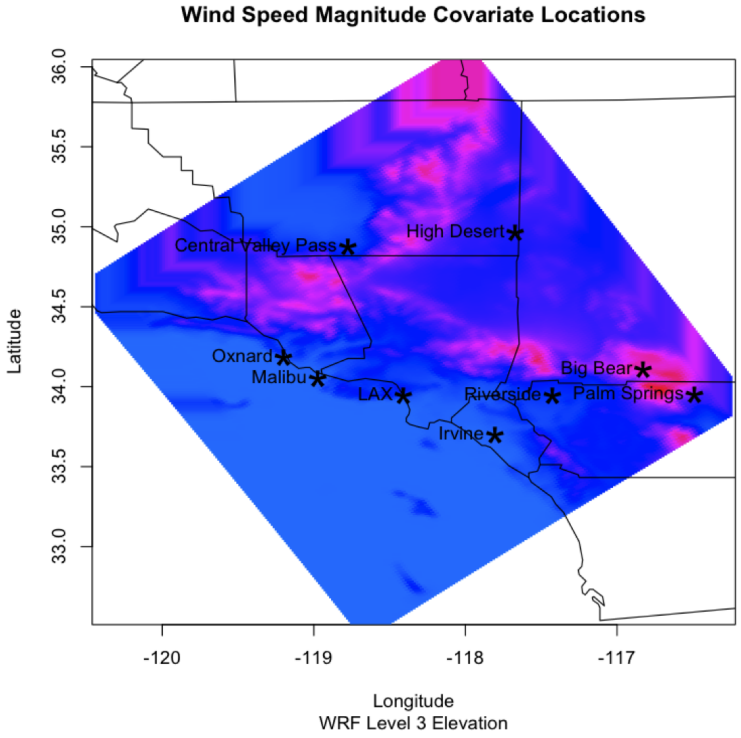


Figure 4.4: Representative Locations shown on High Resolution Elevation Map

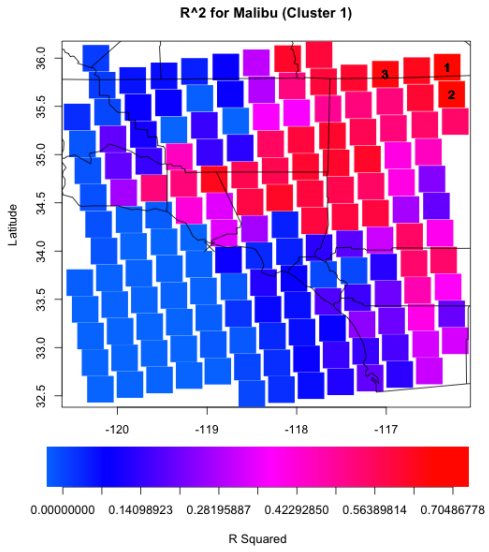
For each of the locations, a Pearson's correlation coefficient was calculated

between the high-resolution wind speed magnitude and each of the low-resolution wind speed magnitudes in the observed area. The way in which clustering methods were judged was based on a weighted average of squared correlation values, or r-squared values. Each location was given an individual score for each clustering method. More specifically, within each cluster, the three best correlations were found between the high-resolution location's magnitude and the low-resolution covariates' magnitudes. Then, these top three location's R^2 were averaged, giving us one r-squared value for each cluster. Since each cluster is made of a different amount of days, when we average the three clusters' values to get one value, we will weight each cluster's R^2 directly depending on the number of days that are in each cluster. The more days, the higher the weight it receives.

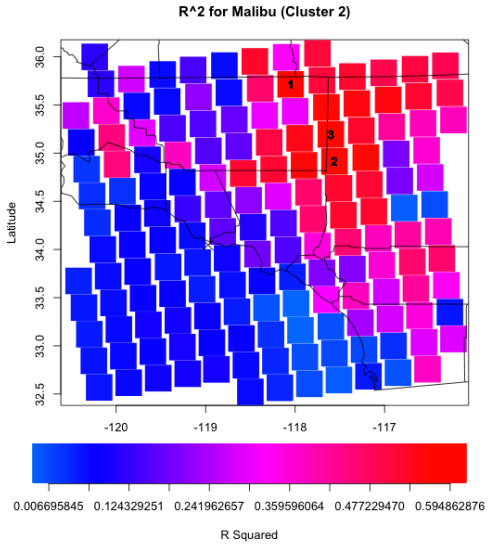
To demonstrate this process, we will choose the Malibu location as an example. Figure 4.5 shows plotting of theR^2 values for Malibu and it's covariates for the three clustered regimes under the NARR1 neighbor scheme. This figure displays how strong these spacial dependencies change with wind regime and labels the three largest R^2 values. Table 4.1 displays the correlations squared as well as the final overall weighted score for the NARR1 scheme. We can then repeat this process for the other four neighbor representation methods: Flow2, Flow3, Flow4, and Flow5.

NARR1	$R^2[1]$	$R^2[2]$	$R^2[3]$	Avg.
Cluster 1	0.705	0.703	0.689	0.699
Cluster 2	0.595	0.591	0.591	0.592
Cluster 3	0.684	0.650	0.650	0.661
Weighted Avg.				**0.627**

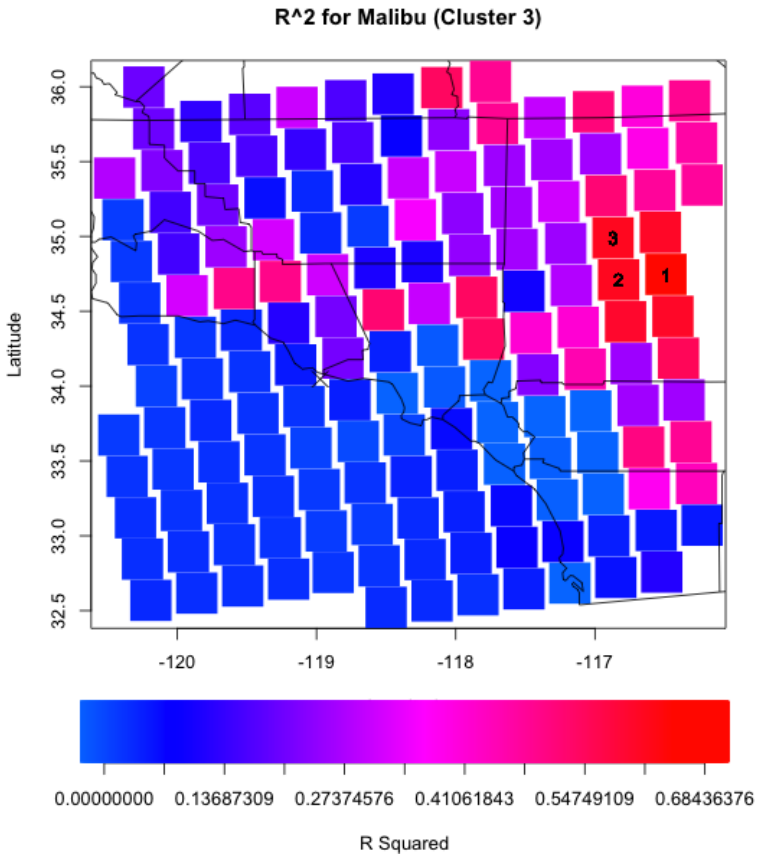
Table 4.1: Correlational summary table for Malibu using NARR1 neighbor representation



(a) Santa Ana Regime



(b) Strong Onshore Regime



(c) Mild Onshore Regime

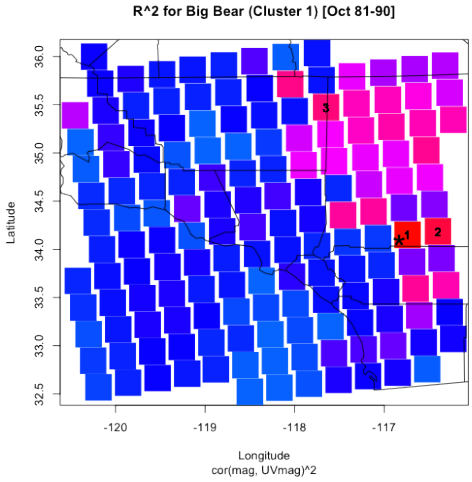
Figure 4.5: R-Squared values between Low-Res Magnitudes and Local Malibu Magnitudes

In the end, each of the five high-resolution locations will get five weighted scores representing each of the five neighboring representation schemes. Table 4.2 displays these results. You can see that the weighted average result from table 4.1 of (0.627) is the bottom left entry in table 4.2.

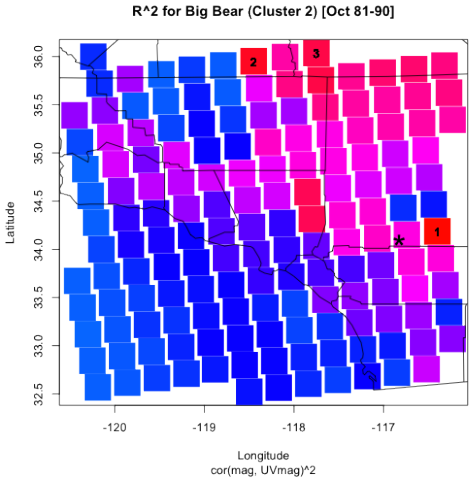
	Malibu	Big Bear	High Desert	Palm Springs	Central Valley	**Avg.**
Flow 5	0.618	0.627	0.600	0.570	0.440	**0.571**
Flow 4	0.620	0.650	0.601	0.586	0.456	**0.583**
Flow 3	0.604	0.630	0.601	0.579	0.431	**0.569**
Flow 2	0.630	0.631	0.599	0.620	0.430	**0.582**
NARR1	0.627	0.632	0.599	0.671	0.425	**0.591**

Table 4.2: Weighted Summary Scores For All Five Locations and Neighboring Schemes

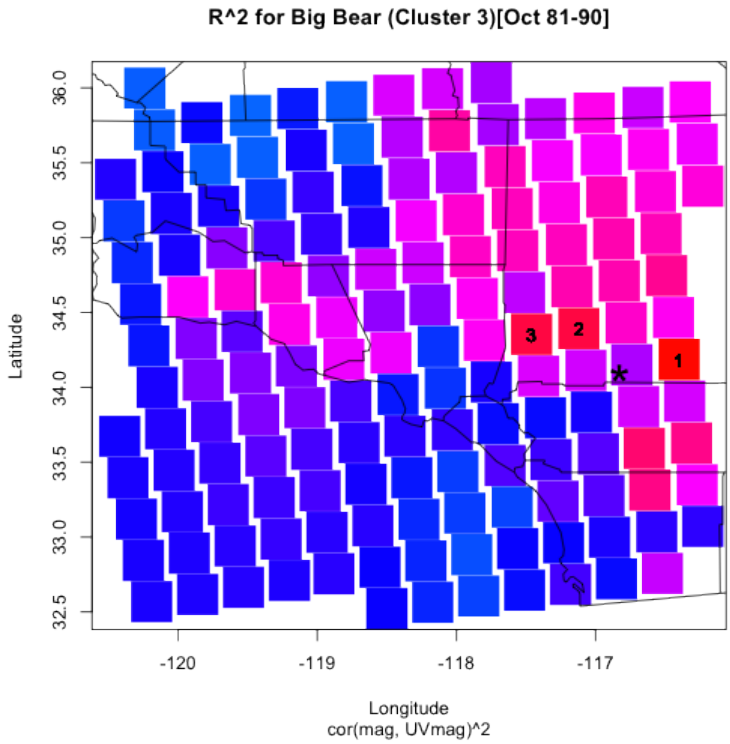
Table 4.2 shows us that all five methods perform well but we see a trend of decreasing returns as we average more neighbors to represent the local point. In the end, the NARR1 neighboring scheme, using the closest low resolution neighbor, is the most robust measure. This could be due to it's high performance across the board, including Palm Springs location which is relatively low for the other methods. We can plot the spatial correlation maps for the remaining locations: Big Bear (figure 4.6), High Desert (figure 4.7), Palm Springs (figure 4.8), and the Central Valley (figure 4.9). These maps display the spatial complexity as well as reinforce the differences in dependence structures with directional regime change. As a reminder, each box is colored by the correlation between the high-resolution magnitude, marked with an "x", and the low-resolution location of the box.



(a) Mild Onshore Regime

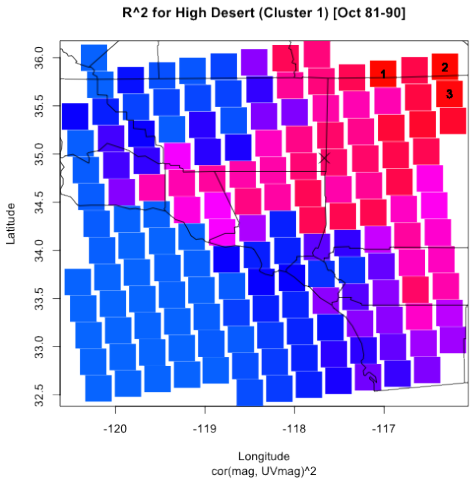


(b) Strong Onshore Regime

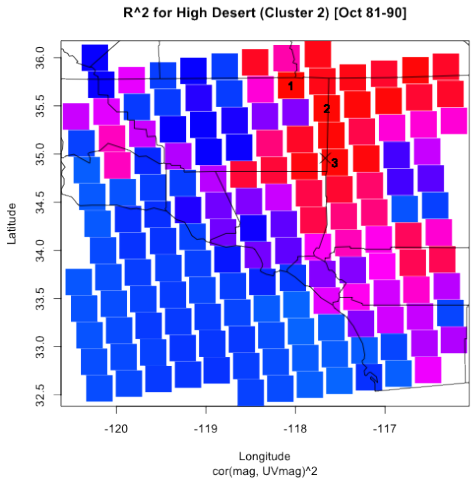


(c) Santa Ana Regime

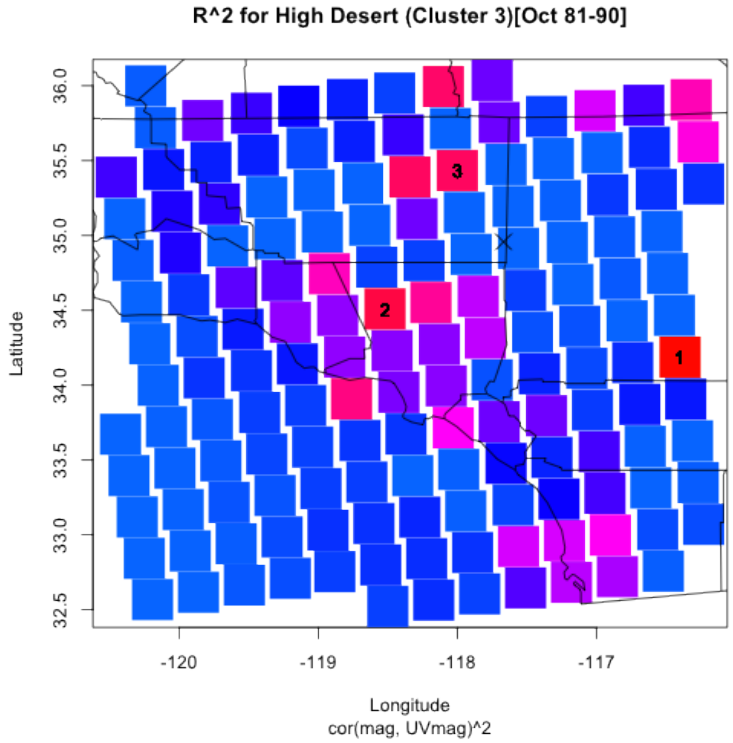
Figure 4.6: R-Squared values between Low-Res Magnitudes and Local Big Bear Mountain Magnitudes



(a) Santa Ana Regime

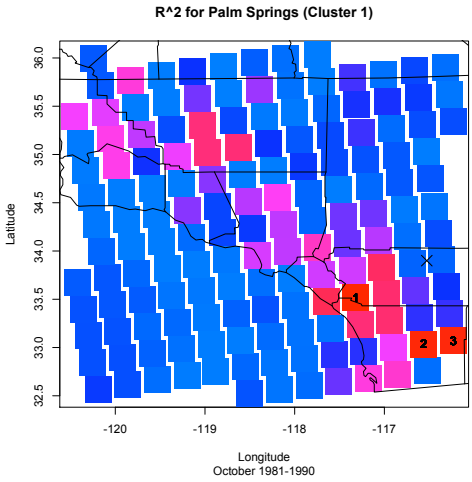


(b) Strong Onshore Regime

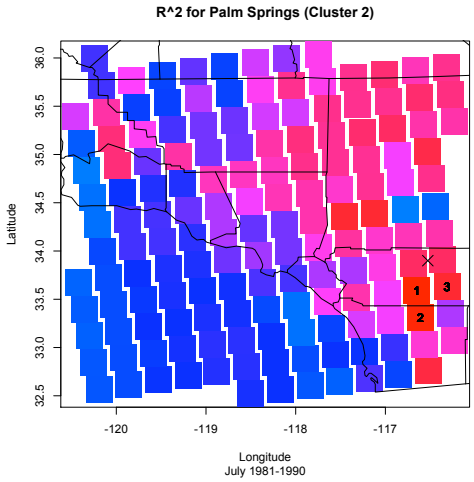


(c) Mild Onshore Regime

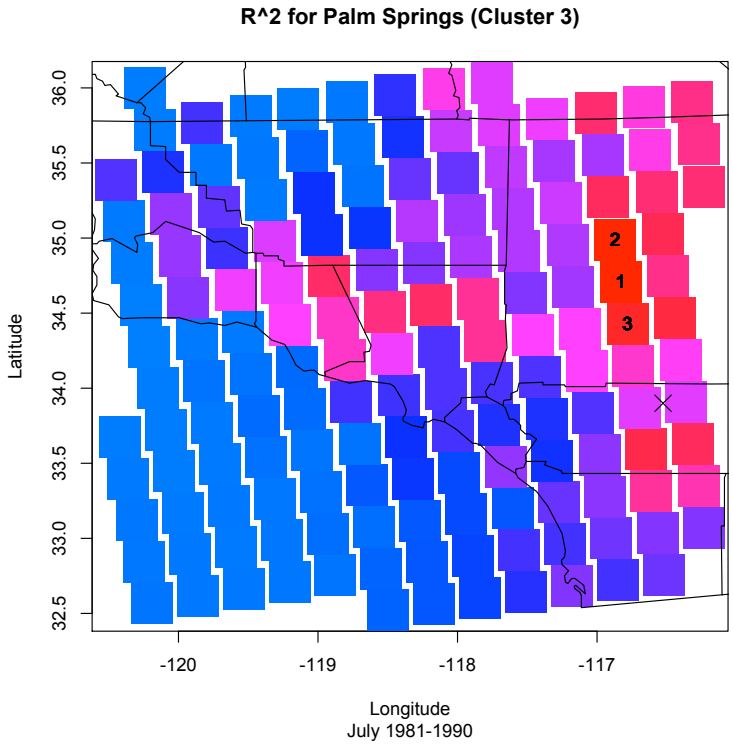
Figure 4.7: R-Squared values between Low-Res Magnitudes and Local High Desert Magnitudes



(a) Mild Onshore Regime

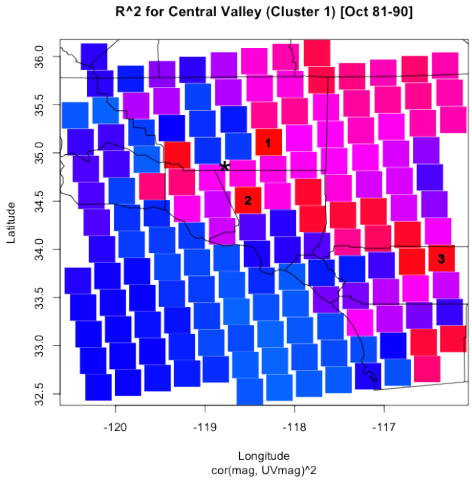


(b) Strong Onshore Regime

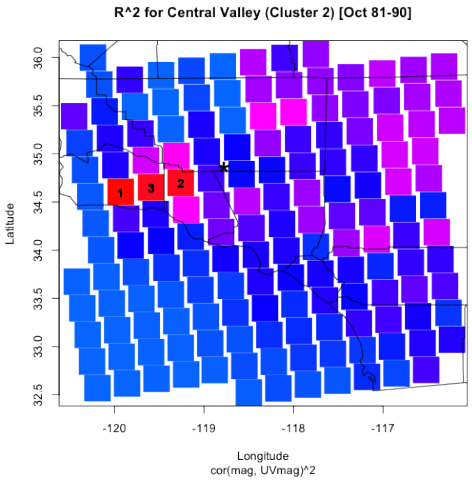


(c) Santa Ana Regime

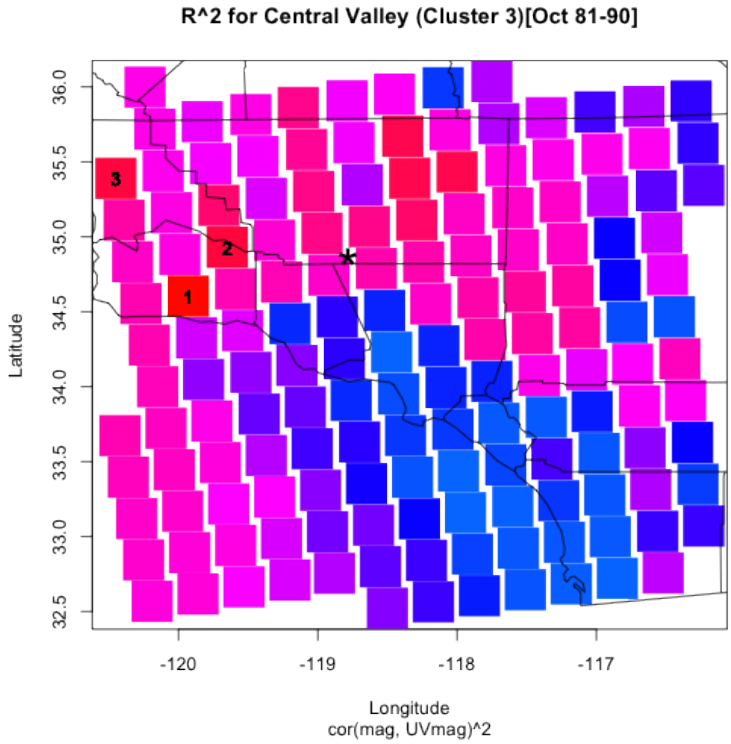
Figure 4.8: R-Squared values between Low-Res Magnitudes and Local Palm Springs Magnitudes



(a) Santa Ana Regime



(b) Strong Onshore Regime



(c) Mild Onshore Regime

Figure 4.9: R-Squared values between Low-Res Magnitudes and Local Central Valley Magnitudes

4.1.2 Distinguishing the Optimum Clustering Ratio

Now that we have determined that the closest low-resolution neighboring vector (NARR1) does the optimal job in representing the daily averages of the high-resolution location, the next problem of scale/weighting needs to be addressed. Circular distance and magnitude are measured on different scales. Therefore, to combine both of these measures in a Euclidean fashion we will standardize both variables first, direction and magnitude. Then, we will apply a weight to each distance measurement. As a reminder, this gives us our final vector distance formula:

$$d_{\psi} = \sqrt{w_1(x_2 - x_1)^2 + w_2 * \min(|\phi_1 - \phi_2|, 2\pi - |\phi_1 - \phi_2|)^2} \tag{4.7}$$

Assigning a higher weight to angle measurements (w_2) places a higher importance on clustering winds together that share the same direction (fig 4.2b). While assigning a higher weight to magnitude (w_1) clusters vectors together by strength. For instance, this clusters mild, medium, and strong wind days together, regardless of direction. Therefore a brief a priori study needs to be conducted to assess which weighting ratio will provide optimal dependence structures. Figure 4.10 goes through displaying several different weighting ratios and their effect on clustering outcomes. In these figures and future figures we will refer to the weights as weighting ratios with the ratio comprised of [w_1:w_2].

Predictions for a wind's magnitude have a different optimal weighting ratio than predictions made for a wind's direction. The reason for this is that we want there to be variance in the dependent variable while still maintaining consistent signals from the independent variable group. In essence, if we want to predict magnitude, we would think to cluster winds belonging in the same regime, i.e. going in the same direction. This insures that the low-resolution covariate locations are correctly identified as well as providing a wide range of magnitudes within that direction for prediction. While for predicting direction, we would think we

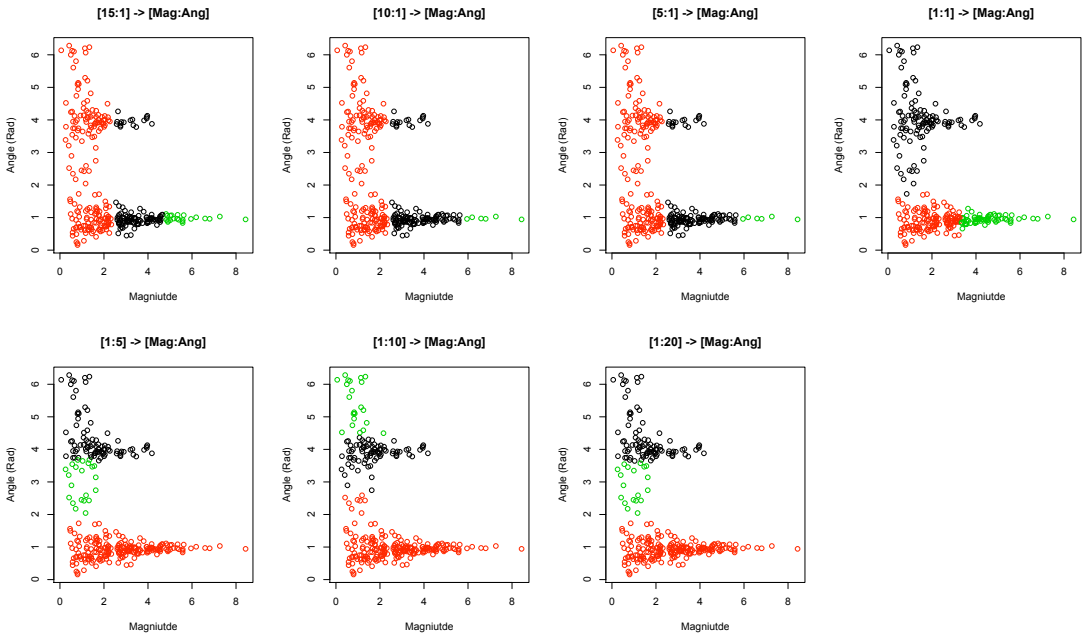


Figure 4.10: Vector Clustering with Varying Weighting Ratios

may want to cluster daily wind averages with more emphasis on magnitude. This groups mild winds, medium winds, and strong winds together regardless of direction (as seen in [10:1] in figure 4.10). Therefore we will perform this empirical exercise twice, once for the optimal magnitude weighting ratio and once for the optimal directional weighting ratio.

What will be common across both is the empirical ratio selection process. First, I had randomly selected 500 prediction locations to test various weighting ratios. Then for each location I will cluster regimes by five different ratio levels. Within each weighting ratio, we will find the three highest correlated values for each cluster. Using these top three relationships, we will find a single weighted correlation squared, R^2, value for each point and weighting ratio. The process of solving for the weighted R^2 value is similar to the process we performed in optimal neighbor representation section. But, instead of changing the vectors we cluster upon, we will change the weights that go into formula (4.7).

4.1.2.1 Magnitude Clustering Ratio

The five different weighting ratios, $[w_{mag}{:}w_{angle}]$, that were chosen to sample were: [15:1], [10:1], [1:1], [1:10], and [1:15]. These ratios were selected to test on a larger sample of points from a previous quicker glance at performances from 15 different weighting ratios.

When we plot the weighted correlation squared value for each locations, figures 4.11, we get to see an obvious pattern which validates are original hypothesis. In these figures relationships between high-resolution prediction magnitudes and low-resolution independent magnitudes improve as weights start to move away from favoring magnitudes to even weighting ratios and improve even more when ratios start favor angles. The returns appear to remain constant after the [1:10] ratio.

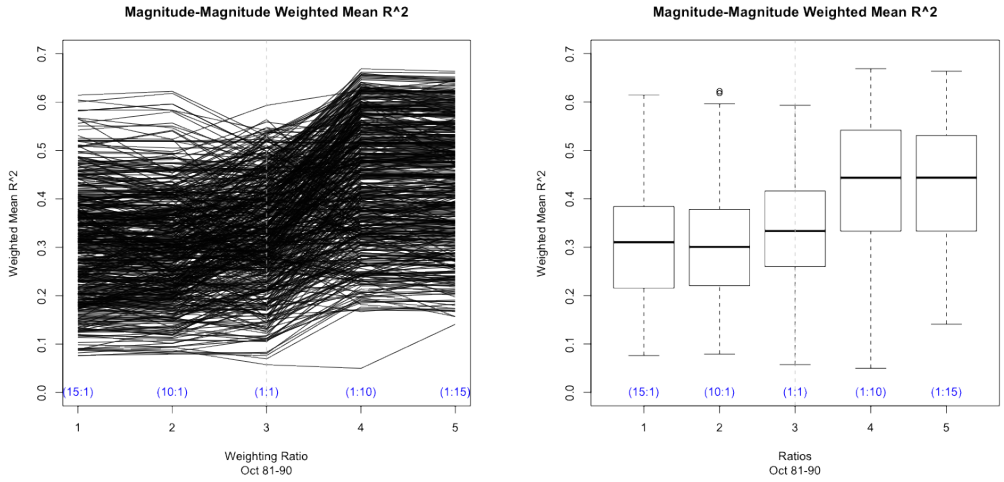


Figure 4.11: Weighted R^2 Values for 500 Randomly Sampled Locations

Since returns seem to remain constant after the [1:10] ratio, simplicity leads us to choose this for our magnitude prediction modeling process.

We can also plot the weighted R^2 values for each point under the [1:10] ratio to see if there is any locational problems. We can see in figure 4.12 that there does

seem to be some spatial pattern to success of the [1:10] weighting ratio. Coastal points and central valley points seem to not perform as well as high desert and ocean locations. A possible future improvement for this method would be to find the optimal weighting ratio for each prediction location, instead of using one general weighting ratio for all magnitude predictions.

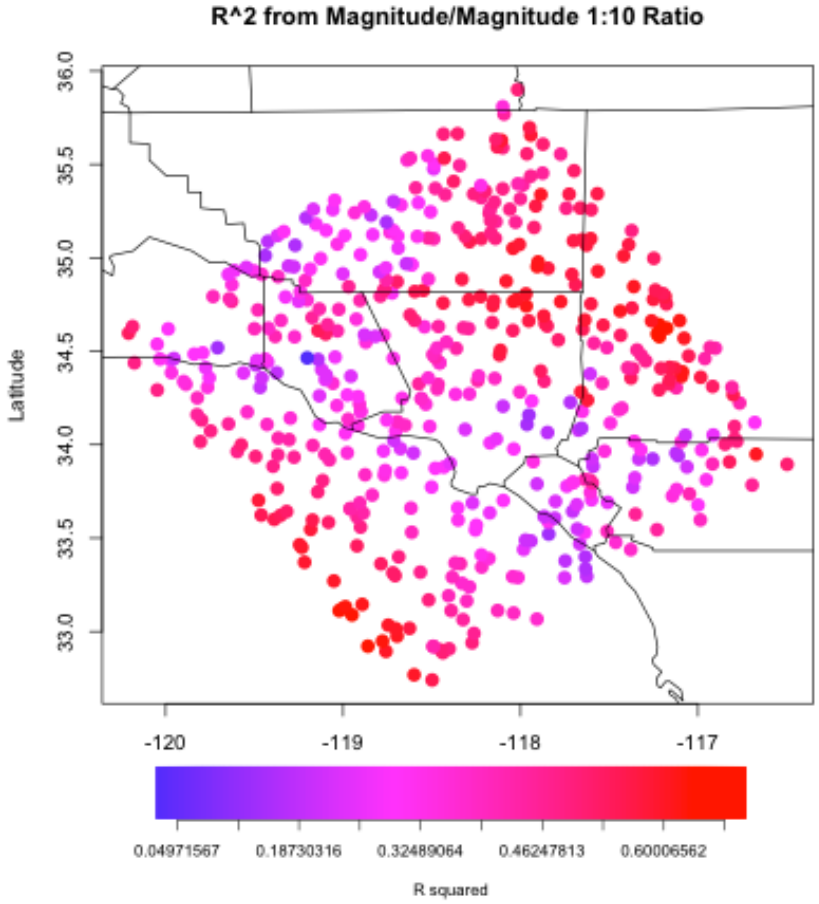


Figure 4.12: Spatial Plotting of Weighted R^2 Values

4.1.2.2 Directional Clustering Ratio

We will repeat the same process for directional prediction. The only difference will be the way we measure the strength of a relationship between vectors of angles/directions. To measure the strength of a relationship between two vectors

of angles we can not use the typical Pearson's Correlation Coefficient as we have done with vectors of magnitudes. This is because the directional value 359° is close to 2° but linear methods interpret them as far apart. Therefore, we will use Jammalamadaka and Sarma's (2001) [14] circular correlation coefficient (r_ϕ) to measure the strength of relationship between two angular vectors a and b for a time 1 to t:

$$r_{\phi_{a,b}} = \frac{\sum_{i=1}^{t} \sin(\phi_a^i - \bar{\phi_a}) \sin(\phi_b^i - \bar{\phi_b})}{\sqrt{\sum_{i=1}^{t} \sin(\phi_a^i - \bar{\phi_a})^2} \sqrt{\sum_{i=1}^{t} \sin(\phi_b^i - \bar{\phi_b})^2}} \tag{4.8}$$

When we apply the circular correlation coefficient to the high-resolution wind directions for our 500 random locations and the low-resolution independent wind directions we see a similar patter, but in reverse.

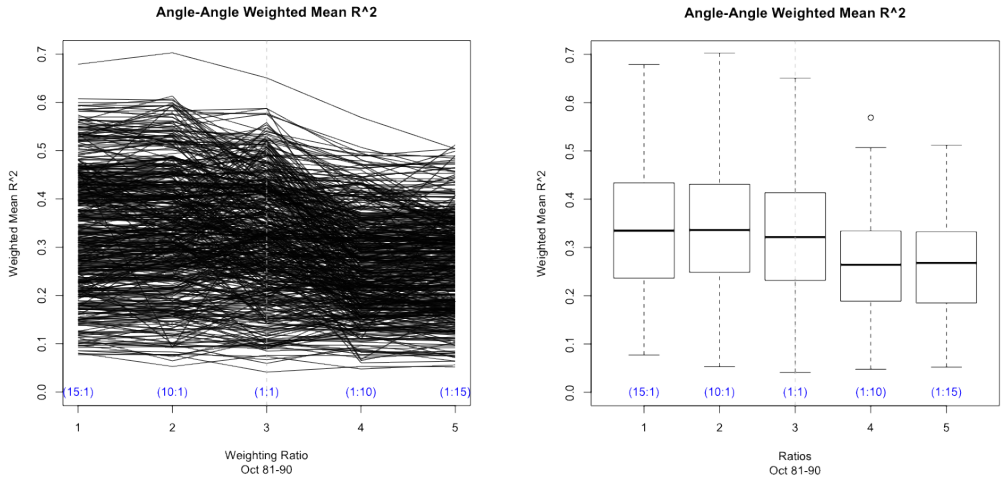


Figure 4.13: Weighted R^2 Values for 500 Randomly Sampled Locations

Figure 4.13 shows us that relationships become stronger as we move towards ratios that favor weighting magnitude over direction. Again, we also see diminishing returns after the [10:1] ratio. This again reinforces are idea that clustering days in mild, medium, and strong magnitude days, regardless or direction, provides us

with dependent variable sets with more variation for directional prediction. Therefore, we will choose [10:1] as are optimal clustering ratio for directional prediction modeling.

4.1.3 Vector Clustering Conclusions

In conclusion, the vector clustering process will be the first step before the model building process for both magnitude and directional predictions. The aim of the clustering process is to provide the models with increased variance in the dependent variable while sub-setting the training data into more homogeneous spatial dependent structures.

For each prediction location, we will cluster days into three regimes using the hierarchical clustering method, with complete linkage, and the distance formula adaptation (d_ψ). The vectors which we will cluster, will be the closest low-resolution observations because they have been observed to best represent the daily averages across multiple prediction location types. For magnitude predictions, the weights that go into the calculation of d_ψ will be: $w_1 = 1$ and $w_1 = 10$. For directional predictions, the weights that go into the calculation of d_ψ will be the inverse of the magnitude's ratio: $w_1 = 10$ and $w_1 = 1$.

Improvements on these clustering methods can be unique neighbor representation schemes for different locations and unique weighting ratios for each prediction location. A simpler improvement could be to create vector outlines of locations that receive different ratios (i.e. the central valley).

One more clustering decision needs to be made. After the modeling process has been completed and we need to predict a wind angle or magnitude for a location; the new daily average vector, ($NARR1_{future}$), needs to be assigned to a cluster in order to apply the correct regime prediction model. Future points, or days, will be assigned to the closest cluster. Distance between a new point and

clusters will be determined in the complete-linkage method, where the distance assigned is the maximum distance between the new point and all points within the cluster. Therefore, the assigned cluster will be the one with the minimum maximum distance.

4.2 Step 2: Model Building

Once we have completed our vector clustering we can build our prediction models. A model is built for each regime allowing the covariates to change with each cluster. Specifically, covariates can move spatially and independent variables can change across regimes or clusters. For instance, when predicting winds over UCLA, sea level pressure over the ocean may be most influential for onshore winds but air density over the desert may be most influential for Santa Ana wind days.

Now that we have optimized the segmentation of the data in the previous sections; unique models need to be built with each segment of the data. In order to build the best predictive models we need to ask: What are the differences between models that best predict magnitudes and directions? And after we decide which model type best represents our response variables in both distribution and nature, we can then ask ourselves: What covariates are the most influential for both direction and magnitude prediction? Through conversation with climatologists, I came up with a list of eight major independent variables that would effect or be related to high-resolution wind vectors. We can investigate the statistical strength of these independent variables for both high-resolution magnitude and directional prediction. These low-resolution independent variables are:

- Mag - Wind Magnitude
- Ang - Wind Direction
- Rho - Air Density

- Dpt - Dew Point
- Hpbl - Planetary Boundary Layer Height
- Lhtfl - Latent Heat Flux
- Prmsl - Pressure Reduced to Mean Sea Level
- Shtfl - Sensible Heat Flux

Because of the difference in nature between magnitude and direction dependent variables, we will perform both steps, model and covariate selection, separately in the following subsections.

4.2.1 Magnitude Model Exploration

To select the best model for the prediction of wind speed magnitude, we must first think of what a distribution of wind speed magnitudes would typically look like. In the example data set and in general, wind speeds can not be smaller than zero and a large wind speeds are above ten. This creates distributions that are right skewed and capped at zero. Because of the non-normal nature of the response variable, we will try two different approaches and choose the more robust method. The two model types we will test are: transformed linear models (with different transformation functions) and a generalized linear model. The transformed linear model allows us to apply a function to the response variable in hopes to make the right skewed distribution more normal.

The transformation that was found to be most effective in a random sample of dependent wind speed magnitude vectors was a fourth root. Therefore, instead of the response variable being a vector of magnitudes at location a for a time t: $x_a = \{x_a^1, x_a^2, \ldots, x_a^t\}$ it will now be $\sqrt[4]{x_a} = \{\sqrt[4]{x_a^1}, \sqrt[4]{x_a^2}, ..., \sqrt[4]{x_a^t}\}$. Since the model is built to predict fourth roots of wind speeds, we must remember to back transform to get our final prediction. For a fourth root transformation, by

using the raw fourth moment of the Gaussian distribution the back transform for a predicted $\hat{x}$ is $(\hat{x}^4 + 6\hat{x}\sigma^2 + 3\sigma^4)$. This formula is based on a Taylor series expansion of the expected value of a fourth root. The only concern for this model is that it may make negative wind speed predictions. To correct for this, we will replace any negative wind speed prediction with a prediction of zero.

Another way to correct for the negative predictions is to avoid them all together. This can be done with a generalized linear model with an inverse gamma link. The reason we chose the gamma link is because the gamma distribution visually represents the distribution of wind speeds best.

Given both modeling types, we will apply both to a random set of locations and choose the most robust model.

Next, we want to explore the selection of the covariates that we will enter into our magnitude prediction models. Since there is a total of eight possible covariates, with many possible locations, using all eight covariates with three locations each would overpower some of the smaller regimes training data sets. Therefore we will perform some covariate strength exploration with the representative locations we had chosen in our neighboring scheme tests, plus some new additional locations (figure 4.4). This covariate filtering stage will help reduce the number of testing covariates.

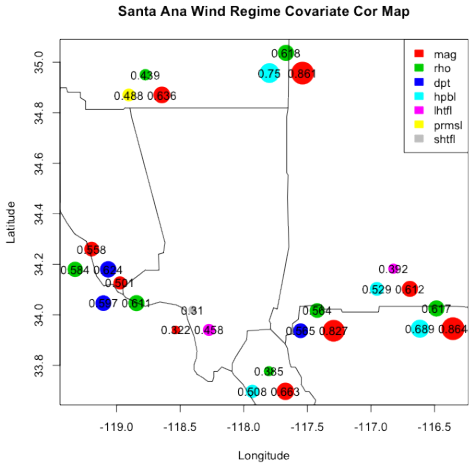
For each of the eight locations in figure 4.4 clustering of 310 days was performed by the methods selected in previous sections and then correlations between the response wind magnitude and the possible low-resolution covariates were found and summarized again with a weighted R^2 value. Then for each location the three strongest covariates were selected and plotted within each regime, with the size of the plotting point representing the strength of correlations. In addition to plotting the strength with size, the actual weighted R^2 values is also printed on top of the plotting point. Results can be seen in figure 4.19 below.

We can see in figure 4.19 that within one location, covariates change across regimes. We can also make observations across all points but within regimes. It appears that specific regimes are better suited to specific covariates. Table 4.3 summarizes these relationships, seen in figure 4.19, across regime by three categorizations of location.

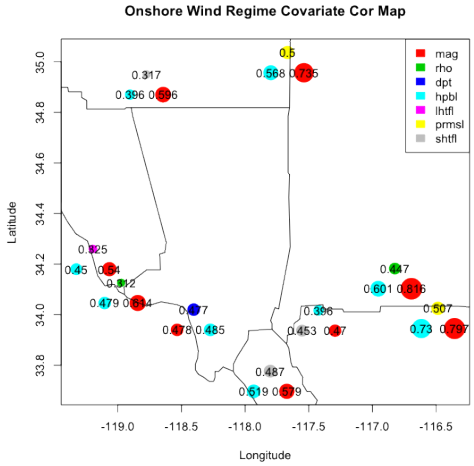
	Santa Ana	Mild Onshore	Strong Onshore
Coastal	Mag,Rho,Dpt	Mag,Hpbl,Dpt	Mag,Hpbl
Inland	Mag,Rho,Hpbl	Mag,Hpbl,Ltfl	Mag,Hpbl,Prmsl
Mountain	Mag,Hpbl,Ltfl	Mag,Hpbl,Ltfl	Mag,Hpbl,Rho

Table 4.3: Covariate Strength Across Regime and Location Type

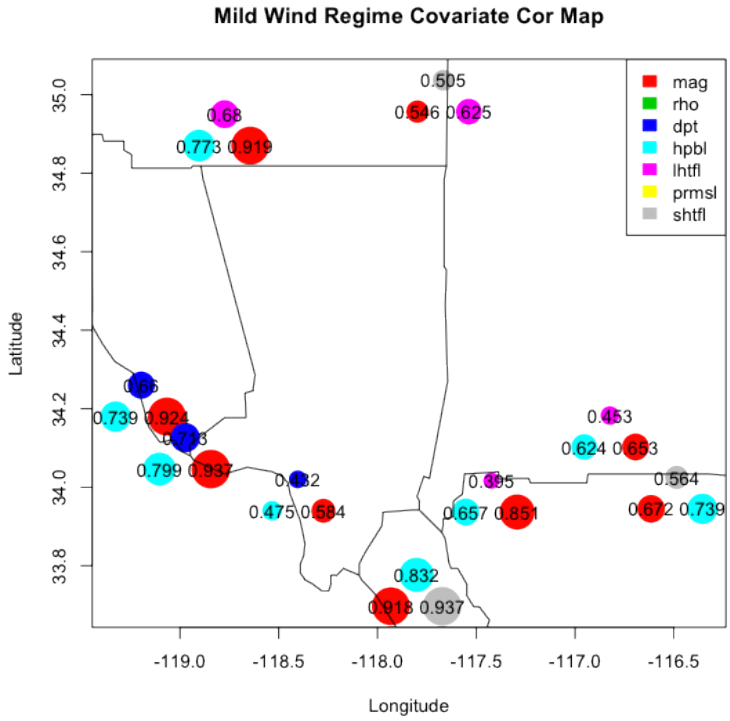
From the figure and table we can limit our initial set of eight covariates down to four. The four we will use for model testing are: Wind Speed Magnitude (Mag), Planetary Boundary Layer Height (Hpbl), Air Density (Rho), and Dew Point (Dpt). Some obvious improvements to the model would be to find the best fitting covariates for each individual prediction point. Another possible improvement would be to reduce the initial set of covariates with principal component analysis and then regress high-resolution magnitudes on the first couple principal components that explain the most variance.



(a) Santa Ana Regime



(b) Strong Onshore Regime



(c) Mild Onshore Regime

Figure 4.14: Magnitude Covariate Search

4.2.2 Magnitude Model Selection

Now that we have decided upon models and covariates to test, we can set up a factorial type experiment to find the most robust method. To perform the model selection process we will use ten years of training data, 1981-1990, and tens years of testing data, 1991-2000. This data only spans the month October, which has 31 days, so their will be 310 observations in each set for each prediction point. We will also randomly choose one fourth of the total WRF high-resolution locations to test all models on. This gives us a total of 4288 prediction locations to test all model types on.

For each of these models, when a covariate is chosen to be placed in the model, we will use the three best locations for that specific variable. For instance, if we use Hpbl to predict wind magnitude over UCLA within the Santa Ana regime, we will find the three strongest low-resolution relationship locations and use those as three covariates within the model.

Since we looked at all the covariates individually in the selection process, we will also want to test if multiple covaritates add any new information or is a simple model the preferred model. Therefore we try both models, transformed and generalized linear models, and vary the amount of covariates within the models. First using all four covariates: Mag, Hpbl, Rho, and Dpt. Then trying both models with only two covariates: Mag and Hpbl. Then lastly trying both model types with only one covariate: Mag. As a reminder, each covariate entails three independent variables from different spatial locations. Therefore, the first set of models will have a total of twelve covariates, the next a total of six, and the last a total of three. This will give us a total of six different models to test:

To measure the ability of the model to predict we will look at two factors. The first will be the correlation between the prediction and the actual wind speed magnitude. The second will be the root mean square deviation, $RMSD$, of the

	Mag,Hpbl,Rho,Dpt	Mag,Hpbl	Mag
Generalized Linear Model	**GLM2**	**GLM3**	**GLM4**
Transformed Linear Model	**RLM2**	**RLM3**	**RLM4**

Table 4.4: Naming Conventions of Testing Models

prediction and the actual. For high-resolution location a with predictions spanning from time 1 to t, $RMSD_a$ will be calculated as:

$$RMSD_a = \sqrt{\frac{\sum_{i=1}^{t}(x_a^i - \hat{x_a^i})^2}{t}} \tag{4.9}$$

The results for the 4288 random locations can be seen in figures 4.15 through 4.18. There are two overall trends we can see in these graphics. First, is that overall the transformed linear model perfumes better than the generalized linear model counterpart with the same exact covariates. Secondly, the simpler the model gets in regards to covariates, the better the predictions get. This can be seen as you scan from left to right in the figures, the overall performance increases. Combining these two findings together we can conclude that the transformed linear model, with a fourth root transformation, using only magnitude as a covariate type, will provide the most robust model. In the bottom right hand panel of figure 4.16 we can see the typical correlation between the actual magnitude and the predicted magnitude for these 4288 points is around 0.70. We can also see that for the most robust model, RLM4, in the bottom right panel of figure 4.18, that the typical $RMSD$ is around 2.

Even though RLM4 does a good job across the board, there does appear to be some spatial structure to the error. We can see that the Central Valley, eastern Los Angeles County, and off the coast of Orange County contain the most prediction error. This may be due to the one model fits all mentality.

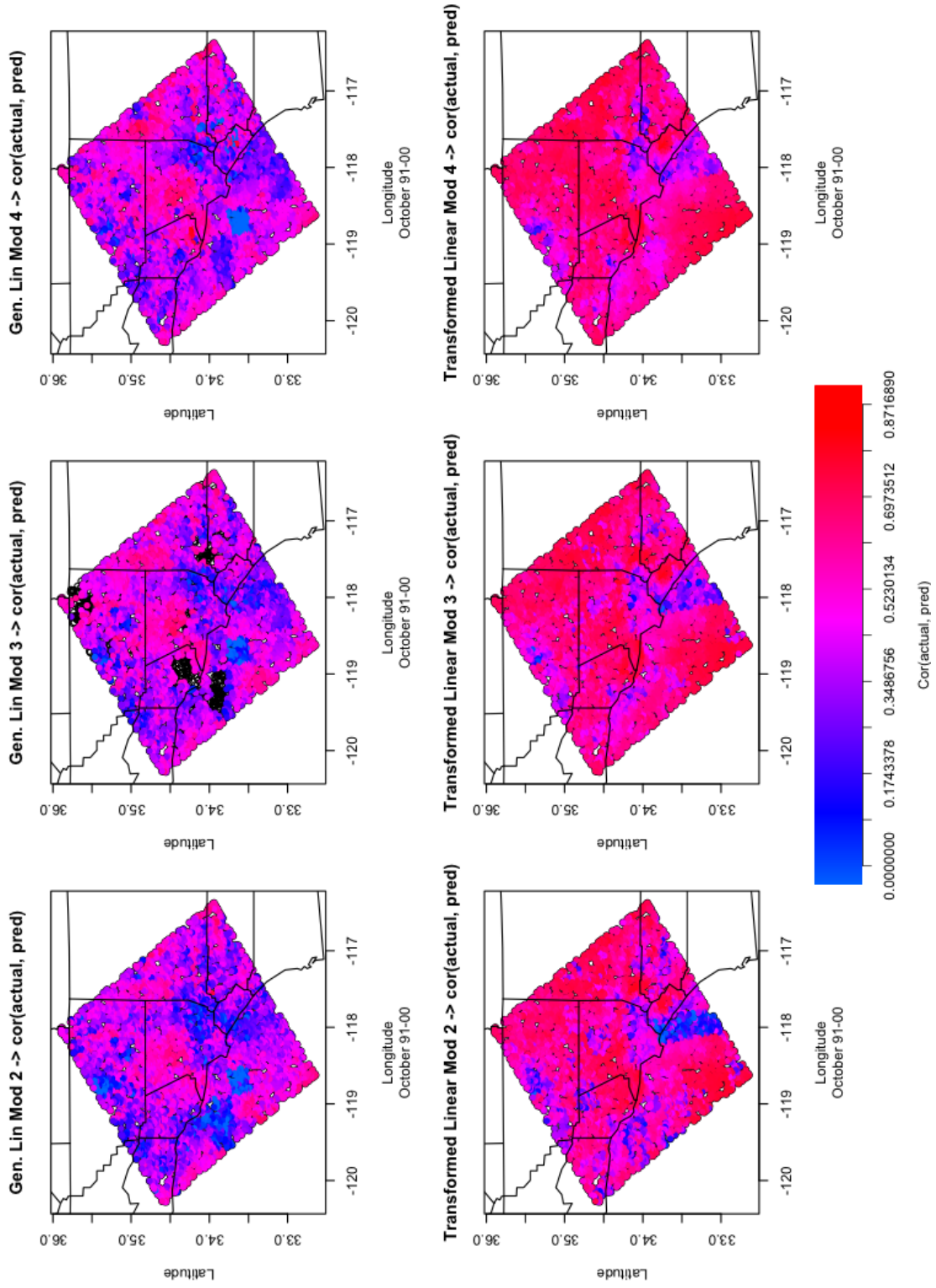


Figure 4.15: Spatial Plotting of the Cor(Actual,Prediction) across Model Types

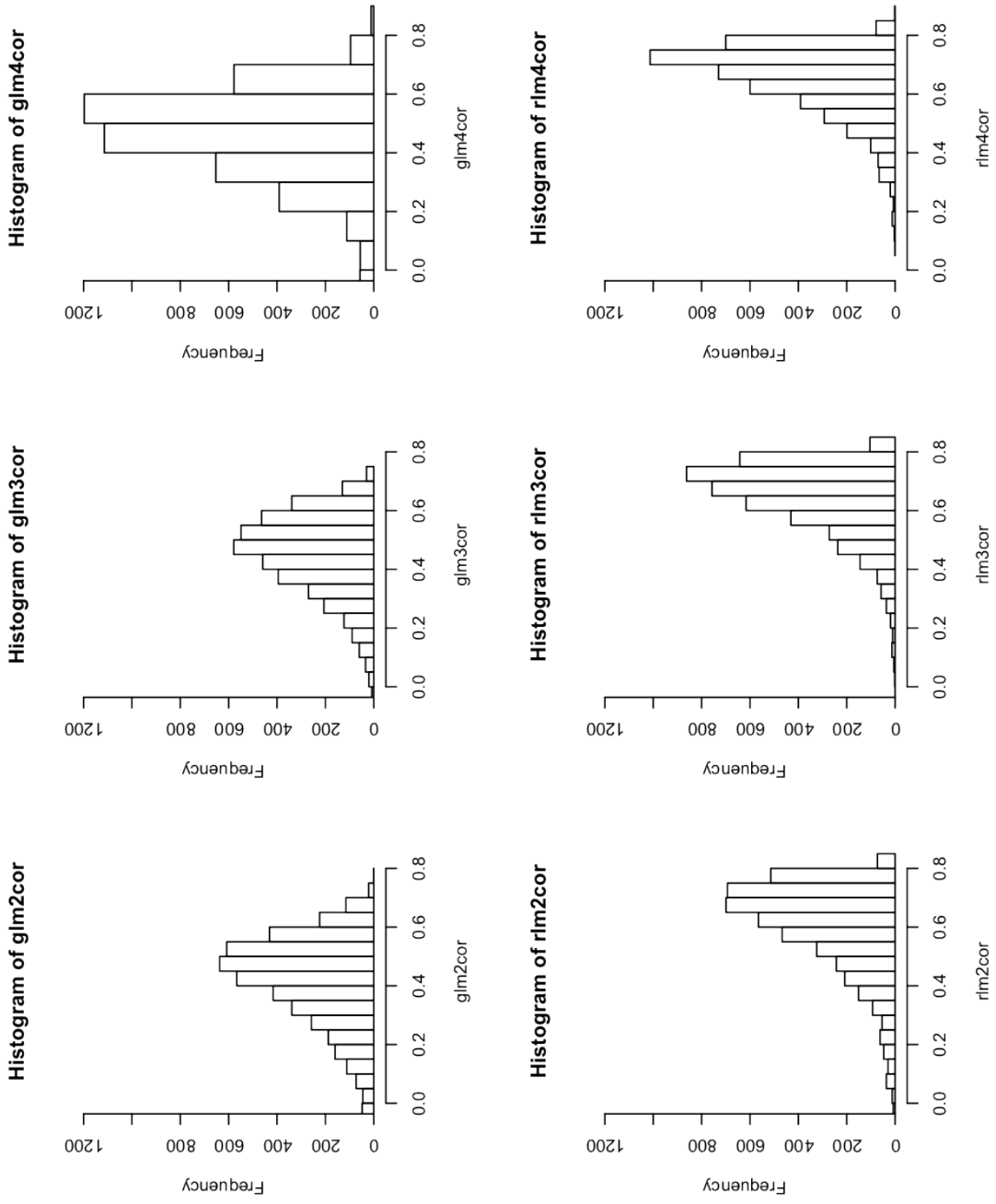


Figure 4.16: Histograms of the Cor(Actual,Prediction) across Model Types

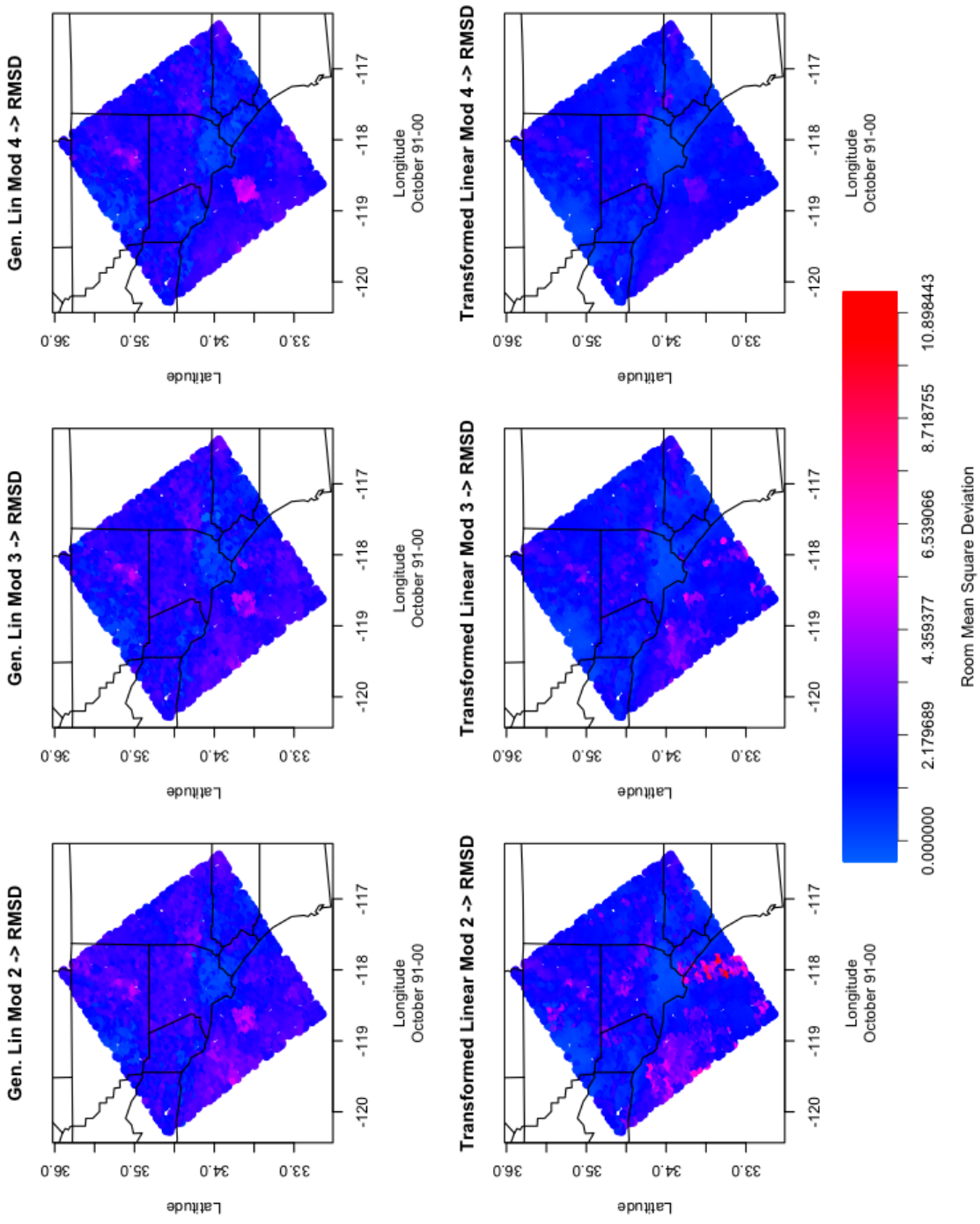


Figure 4.17: Spatial Plotting of the RMSD(Actual,Prediction) across Model Types

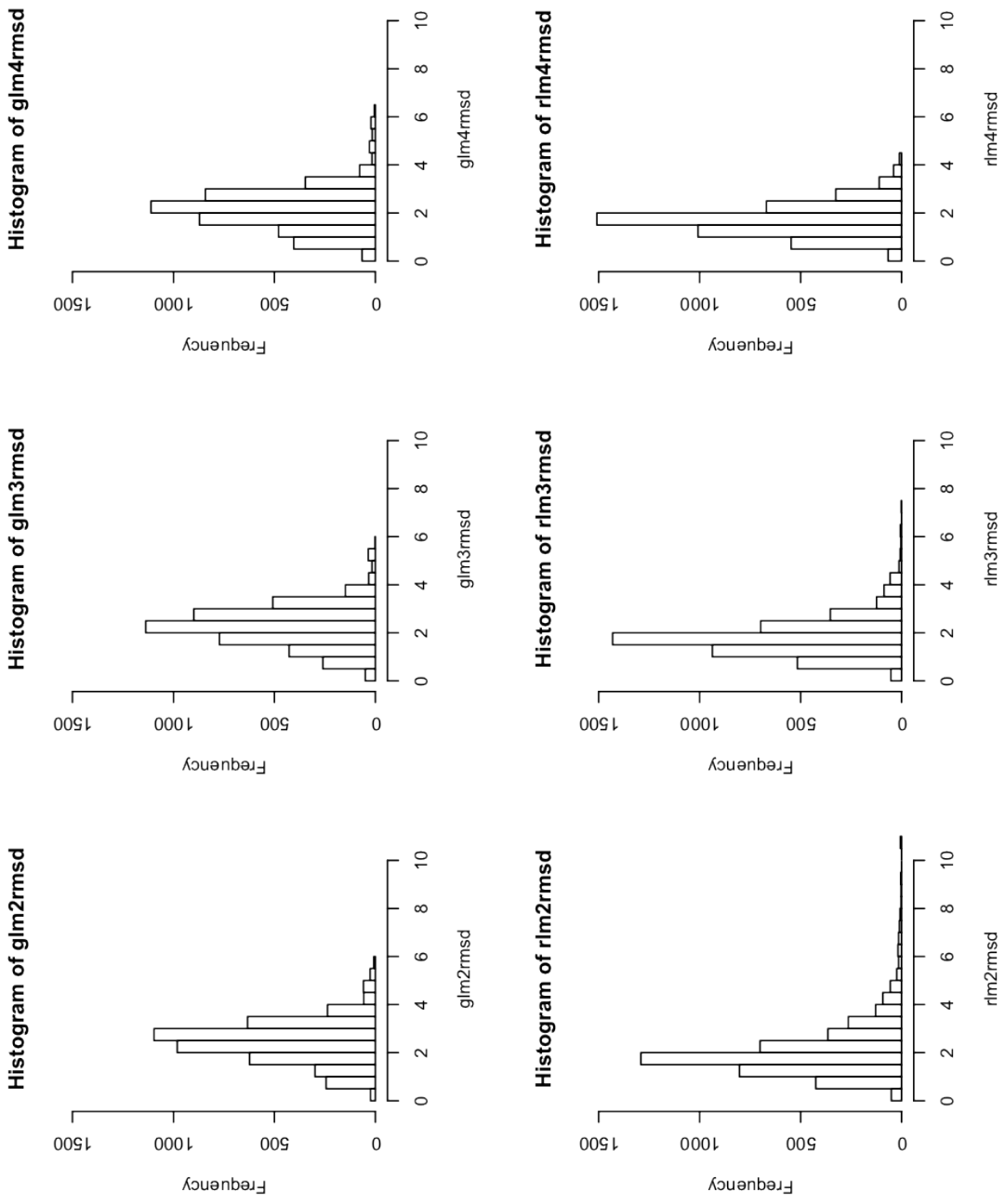
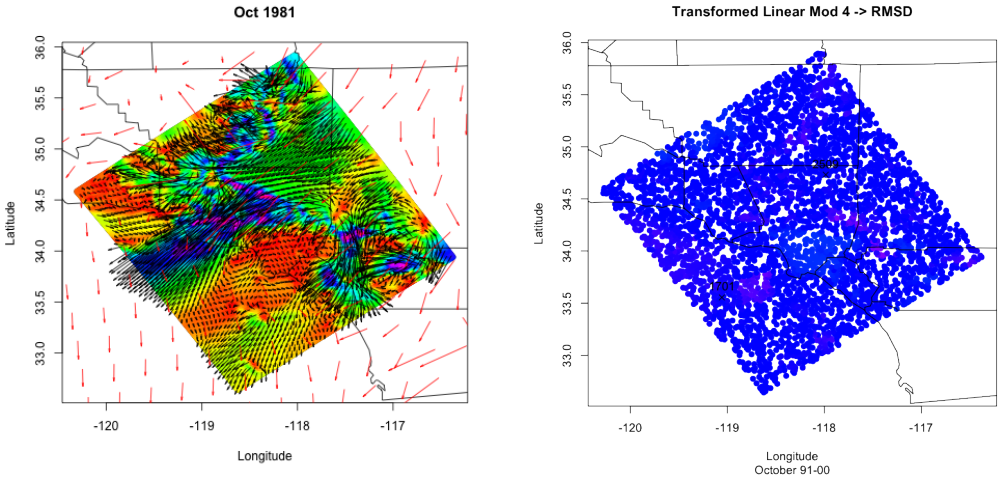


Figure 4.18: Histograms of the RMSD(Actual,Prediction) across Model Types

Another important aspect of model validity is the ability to predict Santa Ana wind magnitudes. Santa Ana winds create a large problem in southern California in regards to wildfires. For this statistical downscaling model to be useful, it

should be able to give accurate future projections of Santa Ana events and their respective magnitudes.

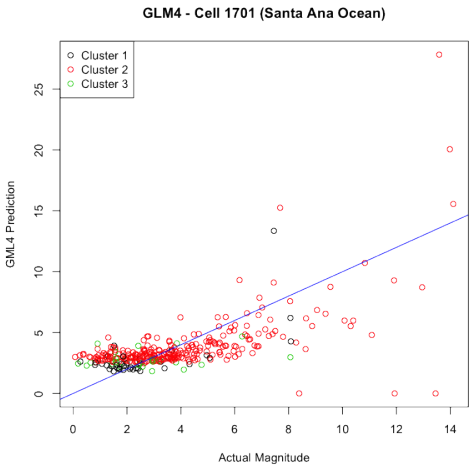
For this, I have chosen two random points that are within strong Santa Ana areas to investigate the differences in the models' ability to predict these uncommon events. We have seen in the investigations above that the simple models, with only three low-resolution magnitude covariates, perform best. Therefore we will compare predictions from the GLM4 and RLM4 models for these two Santa Ana investigation points. Figure 4.19b plots the locations of these two points and its accompanying figure 4.19a shows where they lie in a strong Santa Ana day.



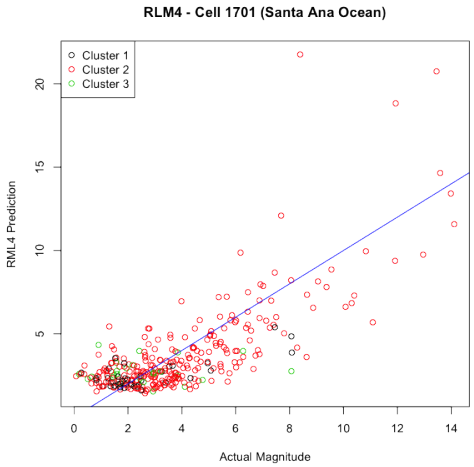
(a) Santa Ana Wind Regime (b) Location of Investigation Points

Figure 4.19: Santa Ana Magnitude Model Verification

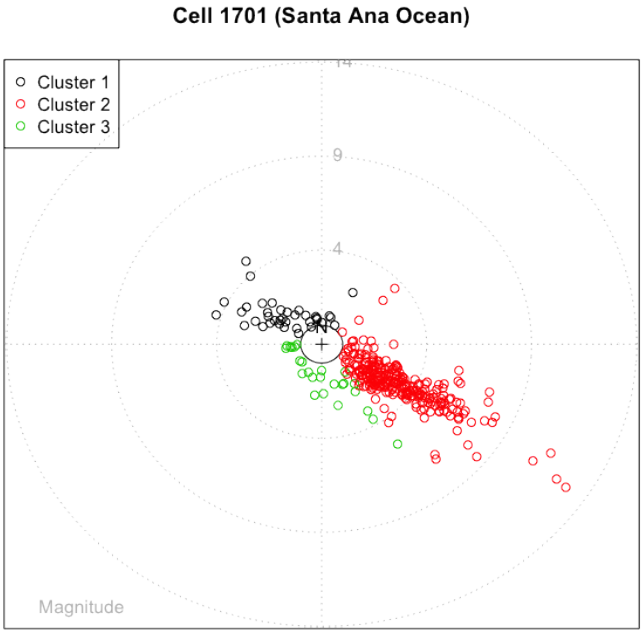
The first point, numbered 1701, is located in the ocean outside of the mountain range that channels the Santa Ana winds. The second point, numbered 2509, is located in the high desert. This is the location of the source of the Santa Ana winds. For both of these points, we will plot the clustering of the testing data set to display wind regimes and wind strength. Alongside the clustering, we will plot the actual wind magnitude against the predicted wind magnitude colored by the respective wind regime for both models.



(a) GLM4 Predictions Vs. Actual

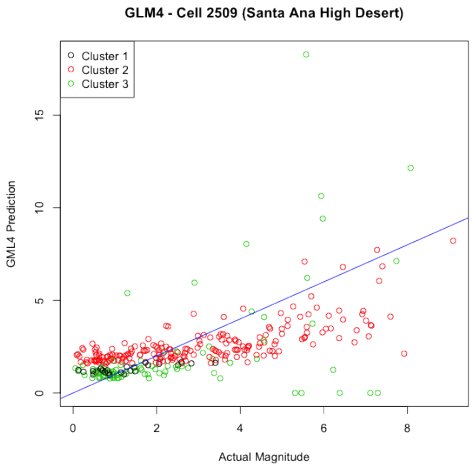


(b) RLM4 Predictions Vs. Actual

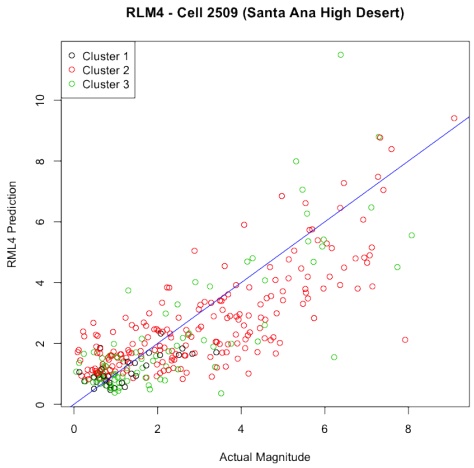


(c) Point 1701Testing Data

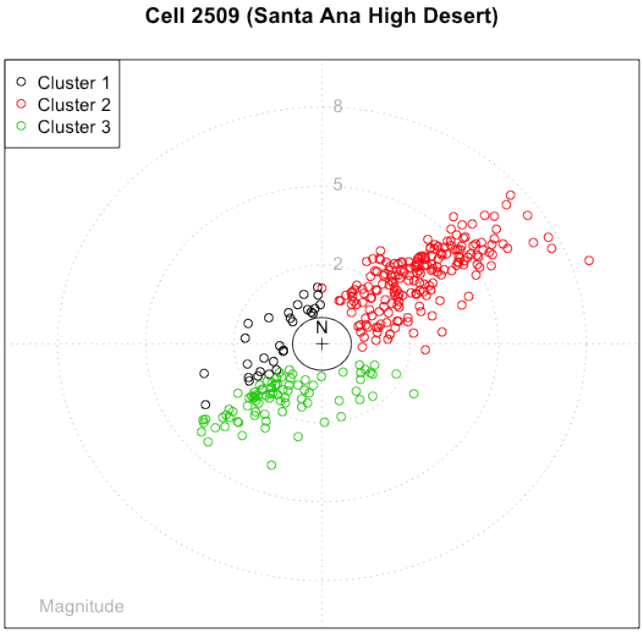
Figure 4.20: Point 1701 Santa Ana Investigation



(a) GLM4 Predictions Vs. Actual



(b) RLM4 Predictions Vs. Actual



(c) Point 2509Testing Data

Figure 4.21: Point 2509 Santa Ana Investigation

We can see that for both points, number 1701 and 2509, the transformed linear model RLM4 outperforms the generalized linear model GLM4. In both

figures 4.20a and 4.21a the generalized linear model, with a Gamma link function, under predicts strong winds and overs predicts mild winds. This creates an overall flat looking plot. The blue line in these plots represents perfect predictions, $x = \hat{x}$, or the line $y = x$. While in the figures 4.20b and 4.21b we can see the points are more appropriately scattered around the "perfect" blue line.

When looking at specific Santa Ana performance, we can see that for point 2509 in figure 4.21c that the green plotting points represent the Santa Ana regime. When we compare the performance of green prediction points in figures 4.21a and 4.21b we see again that the transformed model outperforms the generalized model.

In conclusion, we have seen through experimentation that the transformed linear model is the best and most robust model for magnitude predictions. After clustering is performed on each prediction location's training data set, a RLM4 model will be built for each corresponding subset of data. In these transformed linear models, the three low-resolution magnitude locations with the strongest relationship will be chosen as the covariates within the model. This allows the locations of influence to change with overall wind pattern or regime.

4.2.3 Directional Model Exploration

Given the circular nature of wind directions, normal linear modeling techniques are not sufficient. We will look at three different circular dependent model types that were discussed in the literature review for possible usage. These three models will vary in model class as well as accepted independent variables, circular or linear. The first model we will explore is a regression tree-based method by Lund [18] which allows for multiple linear and circular independent variables. The second model we will explore is the classical circular-circular model by Sarma and Jammalamadaka [25]. This model allows for only one circular independent variable. The third model we will explore is also a circular-circular model, one

circular independent variable, and is the DiMarzio et al. [4] non-parametric kernel smoothing method.

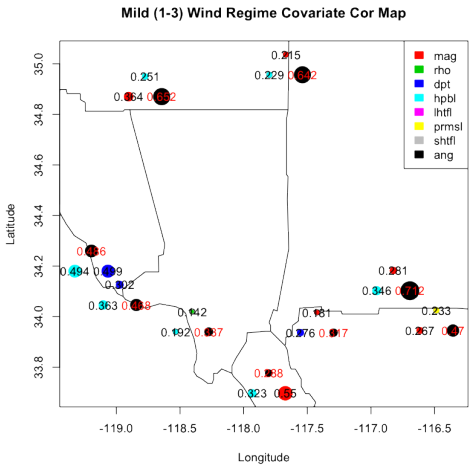
Similar to the magnitude model exploration process, we will want to find a paired down list of possible covariates to try inside of our prediction models. We will use the same list of representative locations (figure 4.4) to observe the strength of the covariates as well as how these might change accross regimes.

For each of the eight locations in figure 4.4 clustering of 310 days was performed by the methods selected in previous sections and then correlations between the response wind angle and the possible low-resolution covariates were found and summarized again with a weighted R^2 value. Both a circular-linear and circular-circular correlation coefficients were used when appropriate (Section 2.2.2). Then for each location the three strongest covariates were selected and plotted within each regime, with the size of the plotting point representing the strength of correlations. In addition, the actual weighted R^2 values is also printed on top of the plotting point. The results can be seen in figure 4.22 as well as summarized in table 4.5.

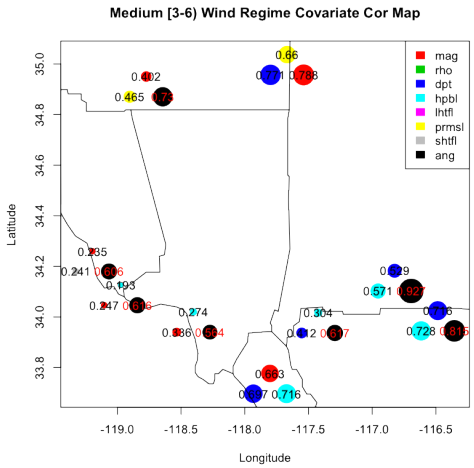
Remember that for wind angle prediction an opposite weighting scheme is used in the regime clustering process. This heavy weighting of magnitude creates regimes that represent mild winds (magnitudes $\approx$ 1-3), medium wind speeds (magnitudes $\approx$ 3-6), and strong winds (magnitudes $>$ 6).

	Mild Winds	Medium Winds	Strong Winds
Coastal	Angle,Hpbl	Angle,Mag,Hpbl	Angle,Mag
Inland	Angle,Mag,Hpbl	Angle,Dpt	Angle
Mountain	Angle,Hpbl,Mag	Angle,Dpt,Hpbl	Angle,Mag,Dpt

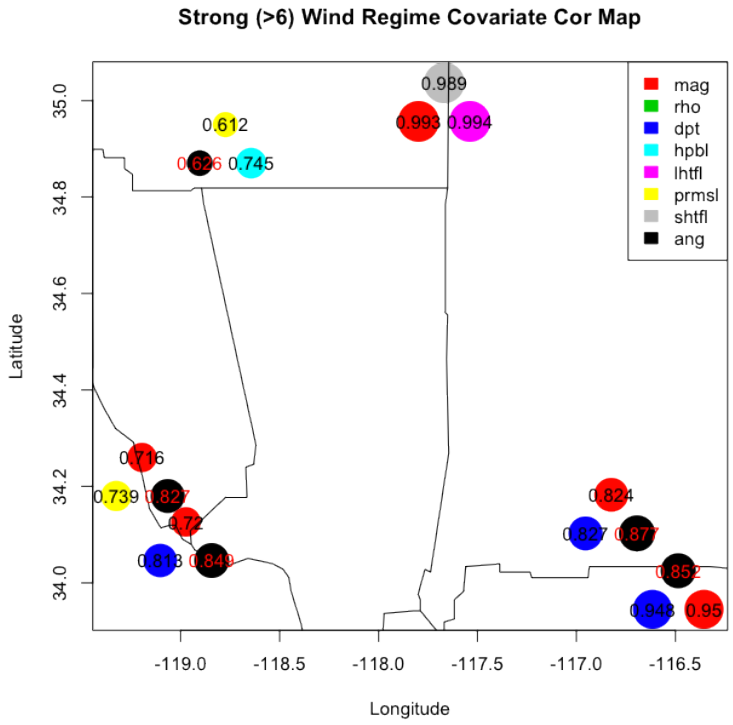
Table 4.5: Angle Covariate Strength Across Regime and Location Type



(a) Mild Wind Regime



(b) Medium Wind Regime

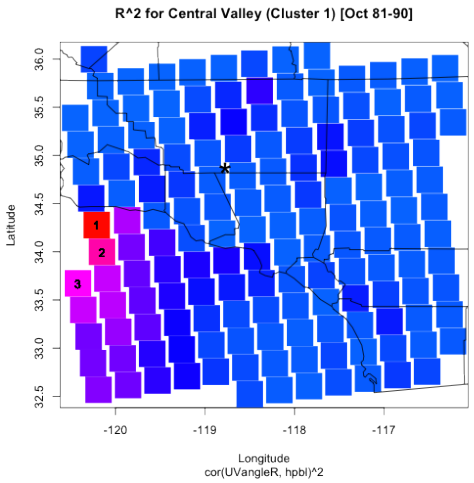


(c) Strong Wind Regime

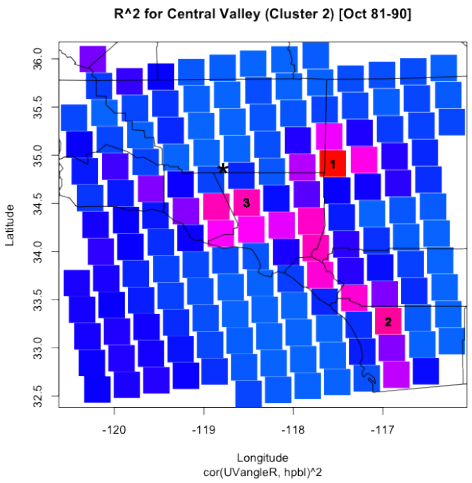
Figure 4.22: Directional Covariate Search

Figure 4.23 displays how a a more prominent directional covariate, Hpbl, spatial structure changes with regime change. Each plot within figure 4.23 displays

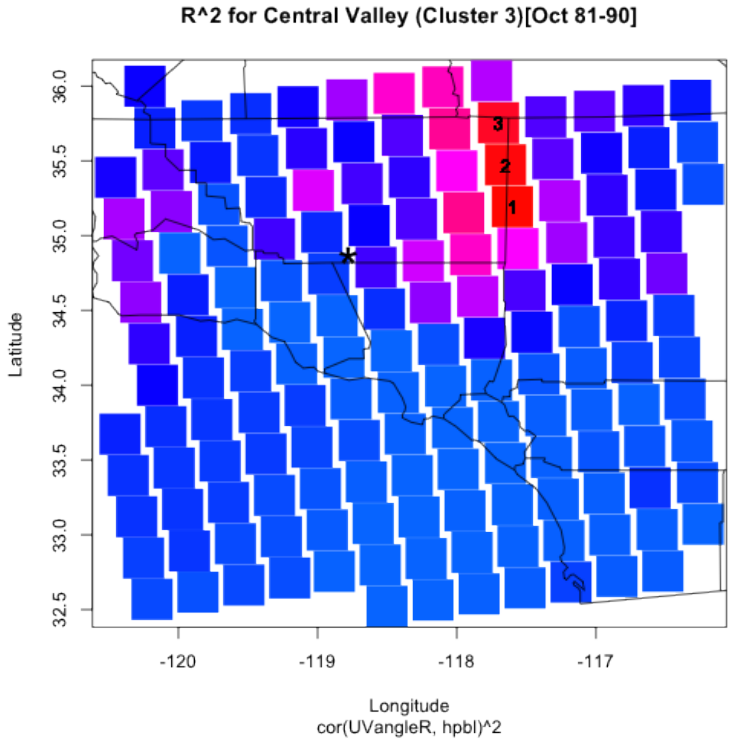
the strength of circular-linear correlation between Hpbl and high-resolution wind angles at the specified central valley location.



(a) Cluster 1



(b) Cluster 2



(c) Mild Onshore Regime

Figure 4.23: Cluster 3

From figure 4.22 and table 4.5 we can see that low-resolution wind angle is the most robust of the independent variables. Other prominent independent variables that were discovered are Planetary Boundary Layer Height (Hpbl), Wind Speed Magnitutde (Mag), and Dew Point(Dpt).

4.2.4 Directional Model Selection

From the three models types chosen to explore, circular-regression trees, are the only model that allows for linear covariate use. In the process of coding this model, I used a few test locations to measure the progress and came back with similar findings. It seems as though the problem of over-fitting was reoccurring from point to point. The first binary split usually does a good job in separating Santa Ana winds from onshore winds. But, subsequent splits seem to only separate one or two points at a time. An example of this can be seen in figure 4.24. We can see from this example, that every subsequent split that occurs after the first, only separates one to two points at a time.

These results were not surprising, as this is a common complaint about regression trees. This leads me to contemplating not using this model across the 4288 testing locations. In additional defense of dropping this possible model type, we observed in the magnitude model selection, that magnitude was the best independent covariate and simplicity was preferred. These two conclusions together lead me to only pursue more intensive model testing on the two circular-circular model types.

Therefore we will make predictions for the same 4288 randomly sampled locations used in the magnitude model selection process using both circular-circular regression models and compare results. In the end, choosing the more robust circular prediction method.

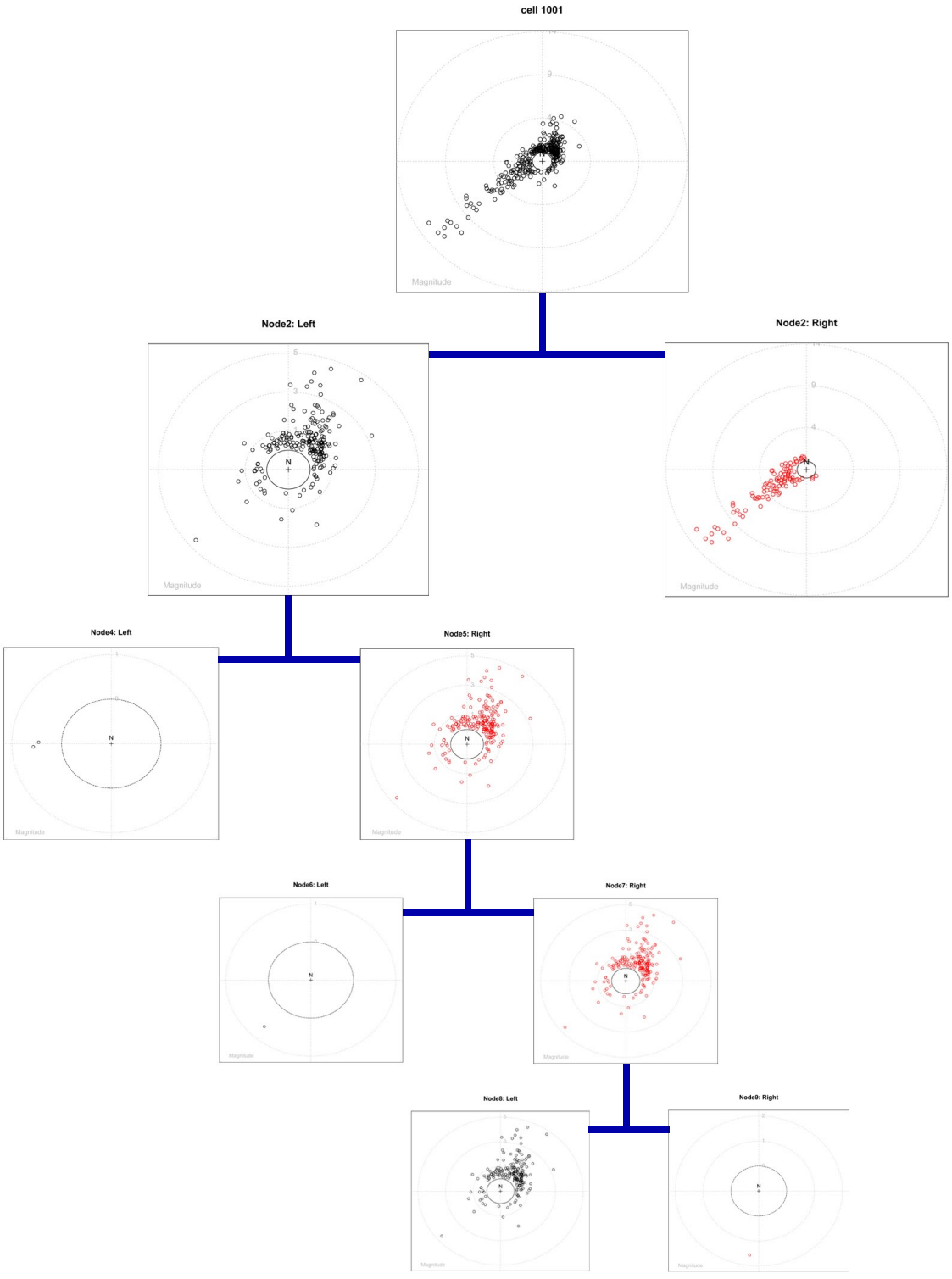


Figure 4.24: An Example of Circular Tree-Based Regression

The first model, which we will call **Model 1**, is Sarma and Jammalamadaka's classic model. The second model, which we will call **Model 2**, is the DiMarzio et al. kernel regression smoothing model. Both models only allow for the use of one circular covariate. To recall from the previous magnitude model selection process, the best performing prediction model used only three low-resolution magnitude locations as covariates. This conclusion led me to want to test an additional idea. In addition to using the best correlated location to make predictions using Model 1 and Model 2. I will also make predictions with the second and third best related low-resolution wind directions. Then, I will make a final weighted prediction, averaging these three separate models, weighting them based on their strength of circular-circular correlation. The first getting the largest weight and so on. For example the weight of prediction one, w_1, will be fraction of the sum of the top three correlational coefficients squared and summed.

$$w_1 = \frac{r_1^2}{r_1^2 + r_2^2 + r_3^2} \tag{4.10}$$

The upside to this weighted prediction is that we can have three separate location's wind angle influence the prediction location. The downside to this idea is that it does not account for covariance between the three models' respective covariates. These predictions we will call **Weighted Model 1** and **Weighted Model 2**. This will give us a total four angular prediction methods to compare results from.

	One Circular Covariate	Top 3 Weighted Average
Sarma and Jammalamadaka	**Model 1**	**Model 1 Weighted**
DiMarzio et al.	**Model 2**	**Model 2 Weighted**

Table 4.6: Naming Conventions of Wind Angle Prediction Models

To measure the ability of the model we will look at the actual wind angle and compare it to the predicted wind angle. For each prediction location, there will be 310 predictions coming from ten years of October data, 1991-2000. The first

statistic we will compare is the absolute prediction error, e_ϕ. We compare the models predicted wind angle, $\hat{\phi}$, to actual wind angle, ϕ, across all locations and times.

$$e_\phi = \min(|\phi - \hat{\phi}|, 2\pi - |\phi - \hat{\phi}|) \tag{4.11}$$

Figure 4.25 shows histograms of all 1,329,280 errors for each model (4288 locations x 310 predictions). While table 4.7 shows the summary statistics for each of the four distributions.

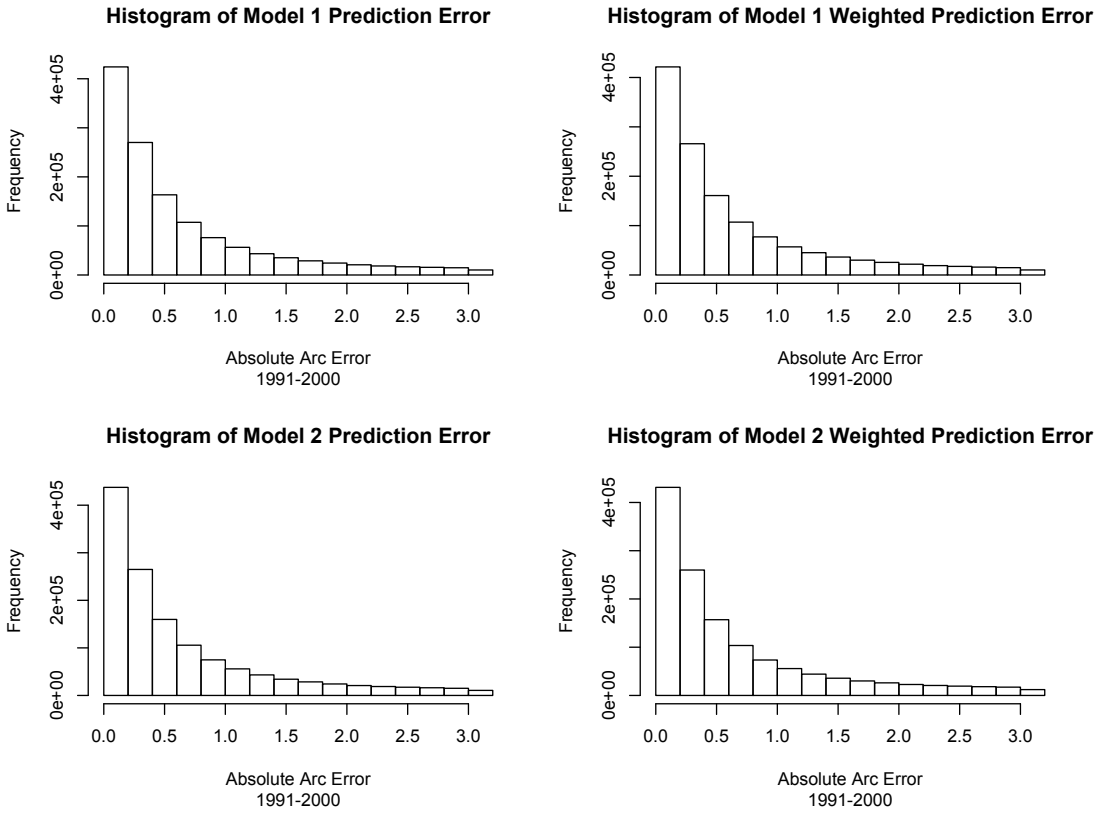


Figure 4.25: Histograms of Model Error

There are a few obvious observations that can be made by looking at figure 4.25 and table 4.7. First, the four models are not that different in their ability to predict wind angle, at least not as much difference as there was in the magnitude prediction models. Given that the distribution of errors our highly right skewed,

	Min	1st Qt.	Median	Mean	3rd Qt.	Max	NA
Model1	0	0.149	0.370	0.642	0.869	π	3640
Model1 Weighted	0	0.150	0.377	0.652	0.894	π	3640
Model2	0	0.143	0.363	0.640	0.865	π	1976
Model2 Weighted	0	0.146	0.373	0.665	0.912	π	1976

Table 4.7: Summary Statistics of Absolute Error for Each Model

the typical value will be measured with the median of the distribution. Comparing all four distributions, we can see they all have a typical error of around 0.37 radians, which is about 21 degrees. The second observation is that the weighted prediction does worse than it's singular counterpart. Again, simplicity wins. The third observation is that there were about half the errors in the kernel smoothing models compared to the classical Model 1 types. This should be taken into account when comparing future statistics.

Next, we will look at the absolute prediction errors by prediction location. This allows us to see if there are specific spatial areas that suffer or do specifically well under each model. The statistics we will use to compare prediction locations are the circular correlation between the actual wind angle and the predicted angle, the median absolute prediction error, and the root mean squared error ($rmse_\phi$). Here the angular root mean square error will be defined as

$$RMSE_\phi = \sqrt{\frac{\sum_{i=1}^{n}[\min(|\phi_i - \hat{\phi}_i|, 2\pi - |\phi_i - \hat{\phi}_i|)]^2}{n}} \tag{4.12}$$

Figure 4.26 shows us the spatial mapping of the median error for each of the randomly chosen 4288 points. This gives us some idea of the typical error we should expect at each location across all regimes. Figure 4.27 shows us the histograms of the median error and table 4.8 displays the summary statistics of the 4288 medians for each model.

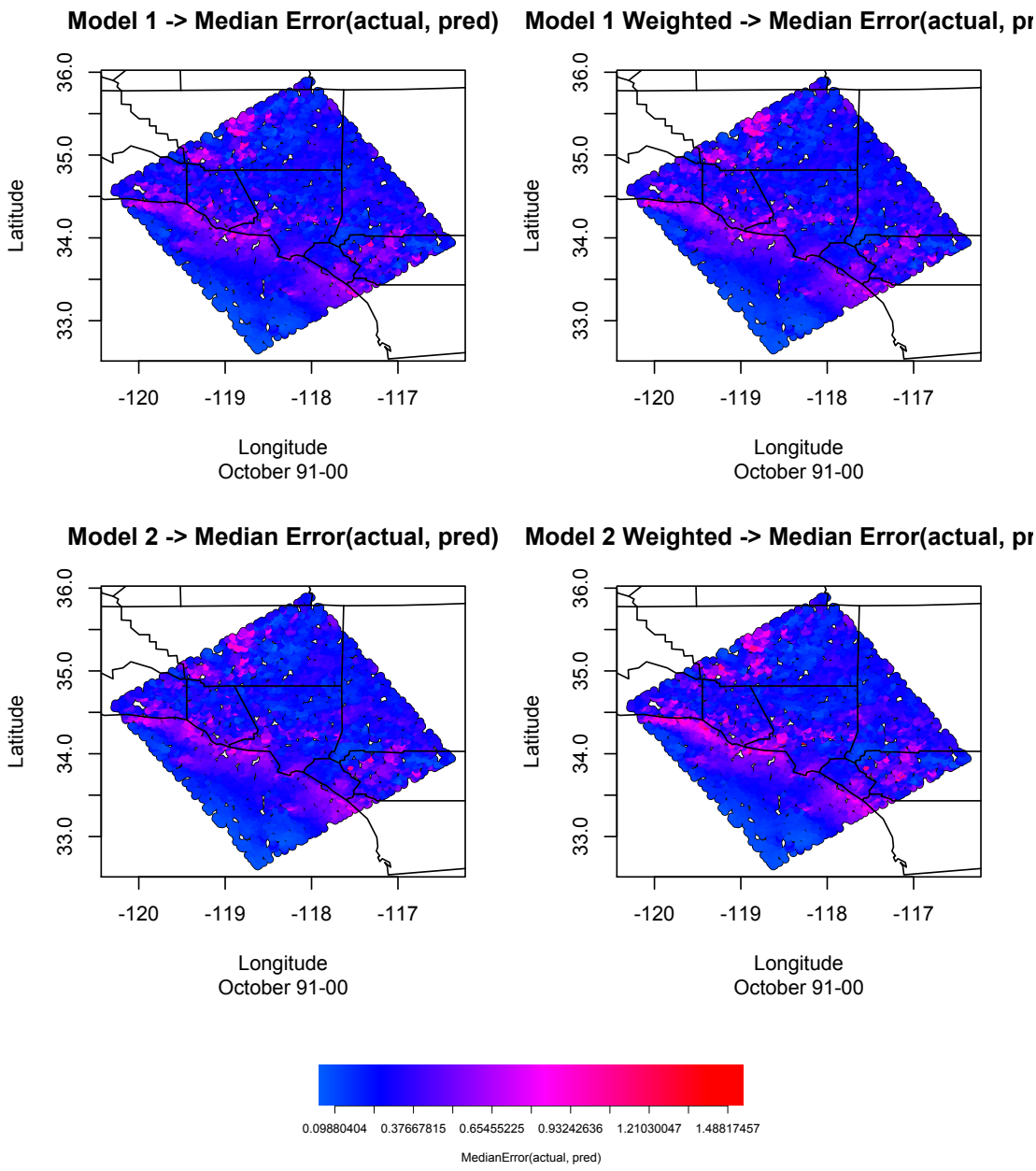


Figure 4.26: Median Angular Error by Prediction Location

	Min	1st Qt.	Median	Mean	3rd Qt.	Max	N
Model1	0.131	0.283	0.364	0.397	0.485	1.127	4288
Model1 Weighted	0.127	0.285	0.376	0.406	0.499	1.351	4288
Model2	0.099	0.276	0.359	0.391	0.476	1.147	4288
Model2 Weighted	0.108	0.278	0.366	0.406	0.495	1.488	4288

Table 4.8: Summary Statistics for the Median Angular Error for Each Location (Radians)

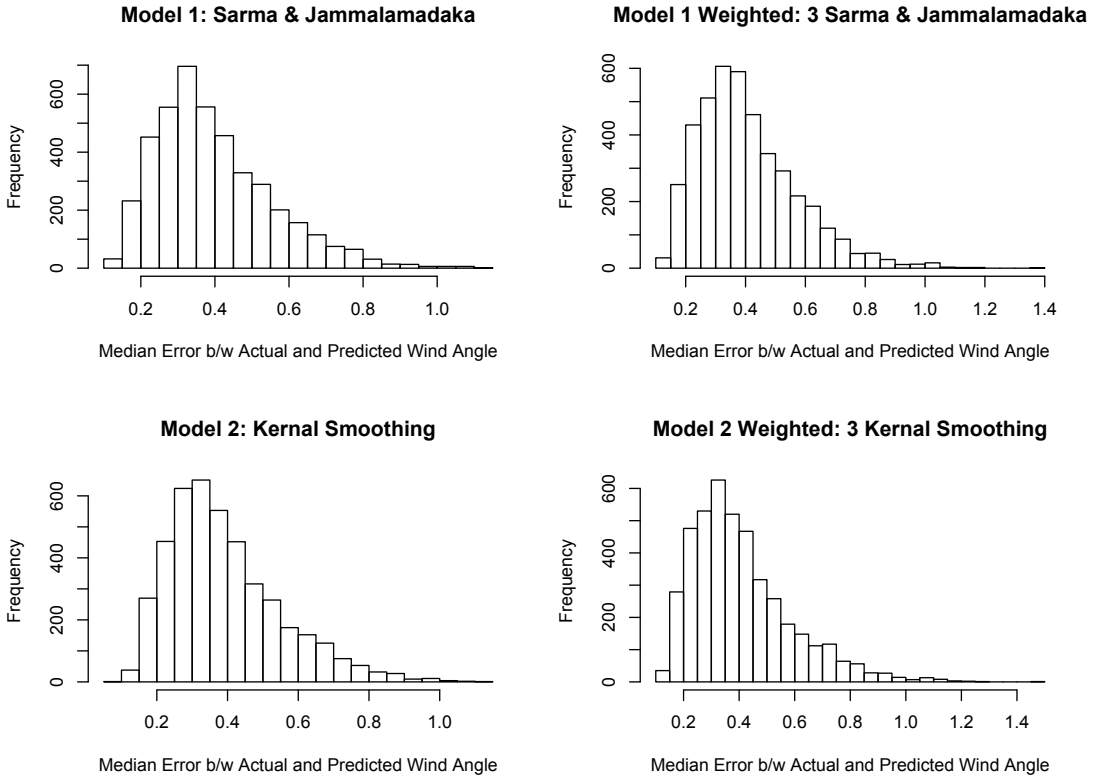


Figure 4.27: Histograms of Median Angular Error

The next group of figures will show us the another statistic that will define model performance across prediction points. Figure 4.28 shows us the spatial mapping of the circular-circular correlation's between the actual wind angle and

the predicted wind angle for each of the randomly chosen 4288 points. While 4.29 shows us the histograms of the correlations and table 4.9 displays the summary statistics of the correlations for each model.

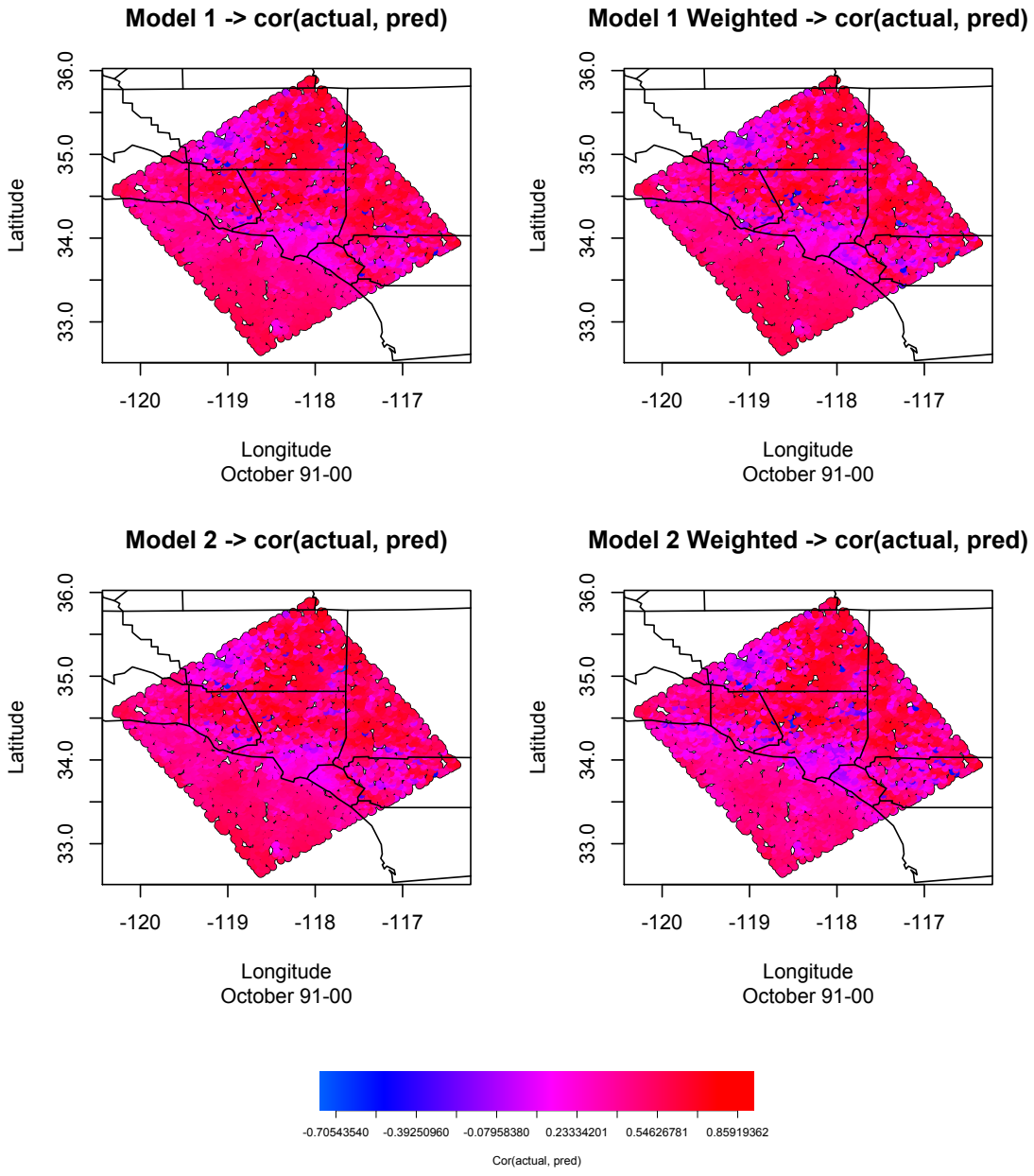


Figure 4.28: Correlations of Actual Wind Angle and Predictions

	Min	1st Qt.	Median	Mean	3rd Qt.	Max	N
Model1	-0.682	0.422	0.550	0.520	0.651	0.851	4288
Model1 Weighted	-0.590	0.403	0.545	0.508	0.651	0.858	4288
Model2	-0.705	0.419	0.543	0.517	0.648	0.850	4288
Model2 Weighted	-0.549	0.353	0.502	0.476	0.634	0.859	4288

Table 4.9: Summary Statistics for the Cor(Actual, Predicted) for Each Location

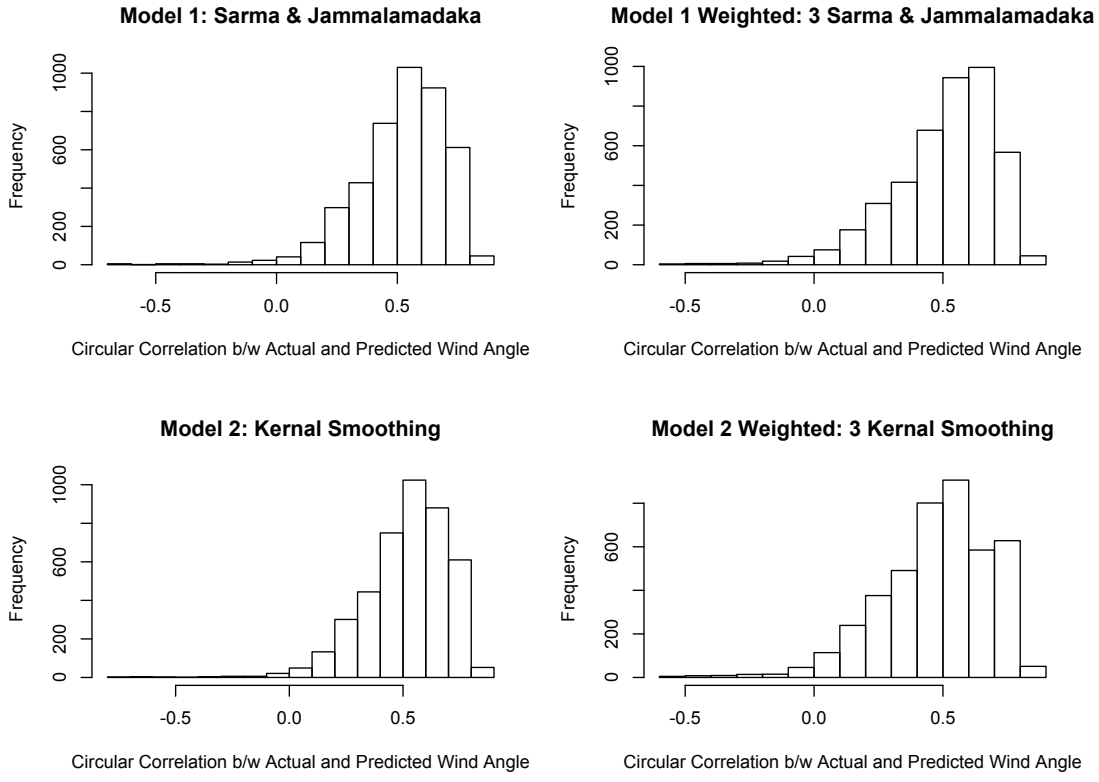


Figure 4.29: Histograms of Location Cor(Actual, Predicted)

Figure 4.30 shows us the spatial mapping of the circular root mean square error between the actual wind angle and the predicted wind angle for each of the 4288 points. This statistic will also give us a look at the "typical" error but will also give a higher weight to large prediction errors. Figure 4.31 shows us the

histograms of the RMSE's and table 4.10 displays the summary statistics of the RMSE's for each model.

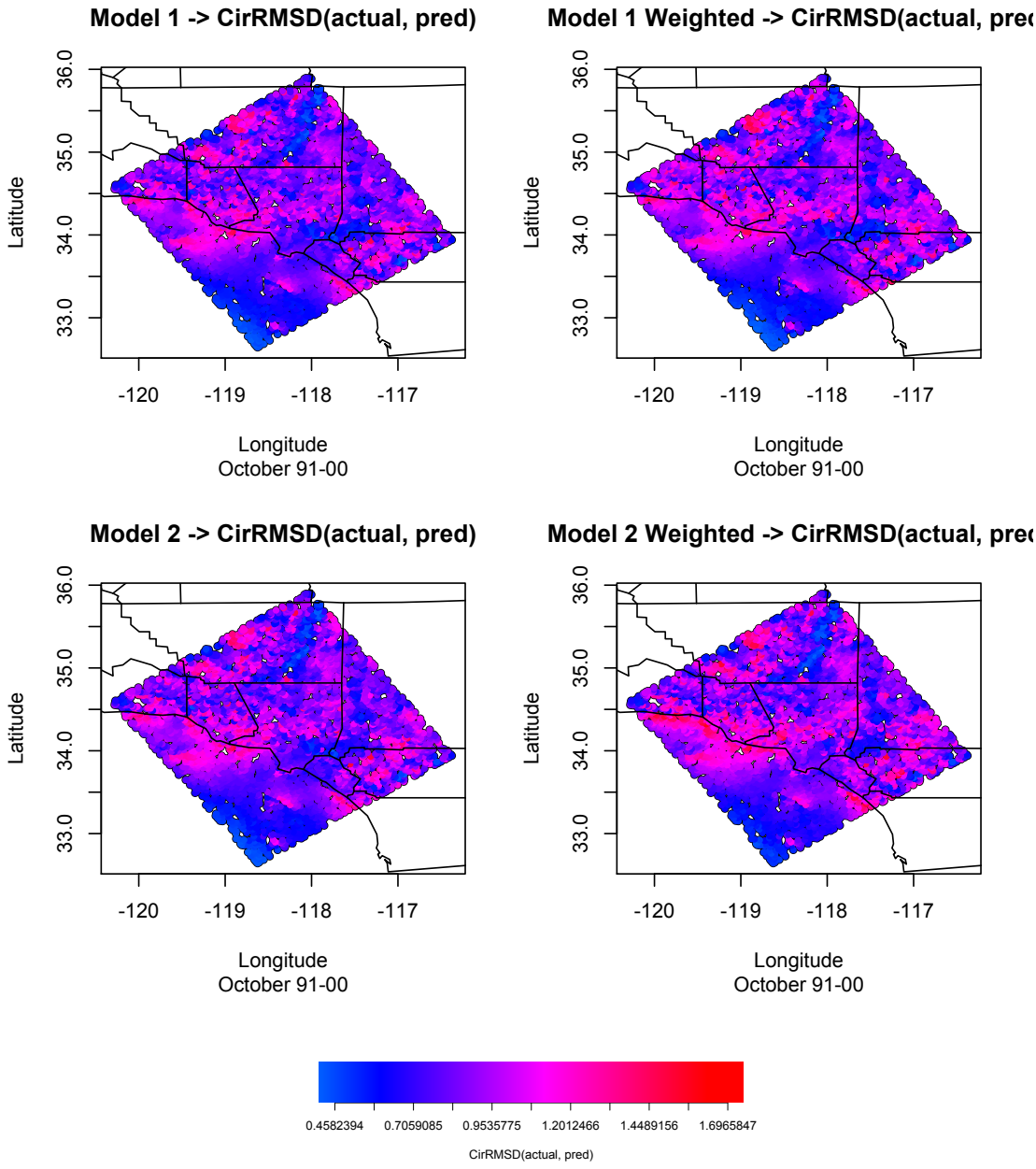


Figure 4.30: RMSE of Actual Wind Angle and Predictions

	Min	1st Qt.	Median	Mean	3rd Qt.	Max	N
Model1	0.460	0.790	0.921	0.929	1.057	1.589	4288
Model1 Weighted	0.4458	0.785	0.933	0.938	1.082	1.697	4288
Model2	0.462	0.793	0.923	0.933	1.063	1.572	4288
Model2 Weighted	0.469	0.812	0.955	0.966	1.103	1.665	4288

Table 4.10: Summary Statistics of the RMSE's Across Locations

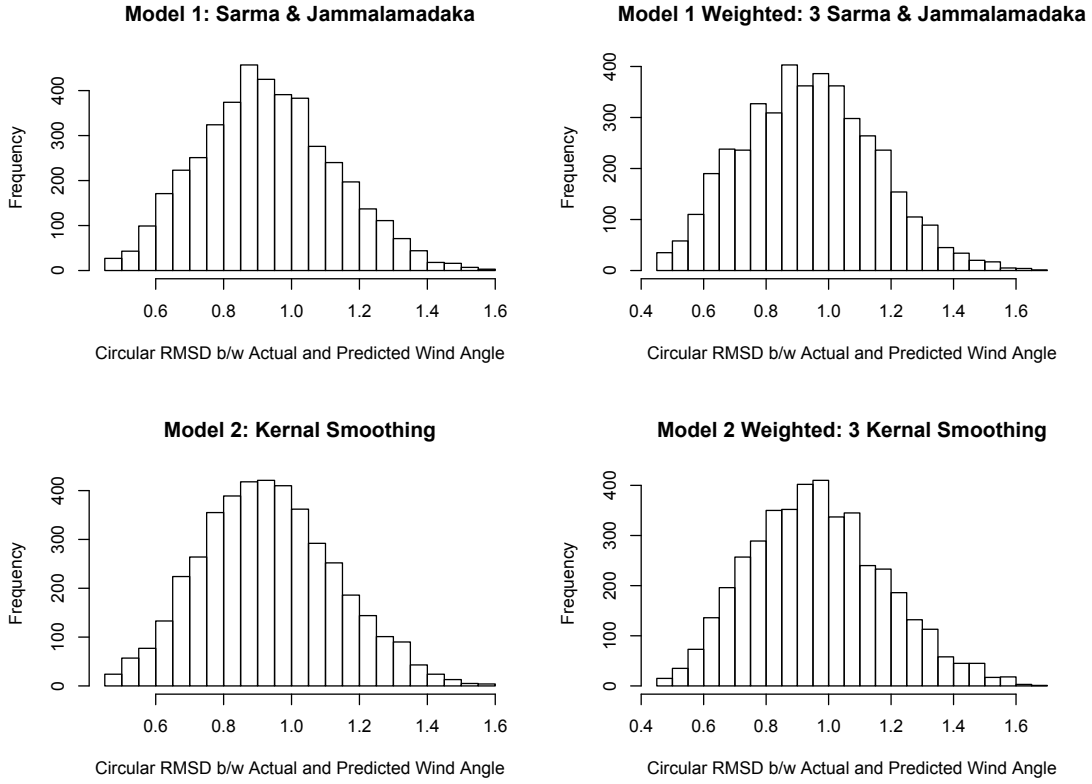


Figure 4.31: Histograms of Location's RMSE's

Several observations can be made by looking at the statistical breakdown of each of these models. First, we again see that the weighted models do worse than their simple counterpart across all statistics. Secondly, Model 1 and Model 2 again perform similarly across all three statistics. When referring to the median error

by location, Model 2 performs the best. But, when referring to the RMSE and the Cor(Actual,Prediction), Model 1 outperforms Model 2. These differences are not large though across all three statistics.

We can also make observations about locational performance. We can see in figure 4.26 that the coastal points contain more error than others. This is somewhat expected because of the overall topographic complexity and larger variance of observations. We can also see in figure 4.26 that similar errors are occurring over the Central Valley. When looking at the correlations in figure 4.28 we can see that lower correlations occur at low elevation land points. This includes Los Angeles and the Central Valley, but not over the high desert.

Given the similarity in model performance and that there were almost twice as many model errors in Model 1 compared to Model 2, we will choose Model 2 as our final angular prediction model.

4.2.5 Model Conclusions

After much experimentation it was found that using a transformed linear regression, with a fourth root transform and three unique magnitude covariate locations, performs best for wind speed magnitude predictions. For each prediction location, this gives us a transformed model for each of the three magnitude regimes: Santa Ana, onshore, and mild.

For our second model type, it was also found that a kernel smoothing circular-circular regression, with one low-resolution's wind direction as a covariate, performed best for wind angle predictions. For each prediction location, this gives us a circular kernel model for each of three directional regimes: mild, medium, and strong winds.

Taken together, this means that for each vector prediction, angle and magnitude, only four covariates will be used: three magnitude locations and only one

directional location. I believe that achieving high prediction performance with such simplistic models attests to the strength of the unique regime clustering process.

CHAPTER 5

Predictions

Now that we have both wind speed magnitude and angle predictions for the Octobers of 1991-2000, we can look at measures of performance combining the two. One important measure of a statistical downscaling method for wind fields is it's ability to capture rare and potentially damaging events, like Santa Ana winds.

5.1 Santa Ana Predictions

Santa Ana winds are of heightened importance to the southern California region because of their possibility for aiding destruction. Southern California has been experiencing a prolonged draught period and the Santa Ana wind's dry conditions and wind speeds are ideal for fanning wildfires.

In a paper by Hughes and Hall (2010) [12] they define a new metric, called the Santa Ana Index, which captures when a Santa Ana event is occurring along with the strength of the event. To create the index, they place a bounding box in the largest gap which channels the Santa Ana flow towards the ocean. The box boundaries are $(33.8° < Latitude < 34.3°N)$ and $(-119.4° < Longitude < -118.8°E)$. This box can be seen highlighted in green in the figure 5.1. For each day, all wind vectors within the box are projected onto the off-direction of 225°, the expected Santa Ana flow direction, and then averaged to get the final indexed value. They define a "Santa Ana day" as any day where the index is greater than five.

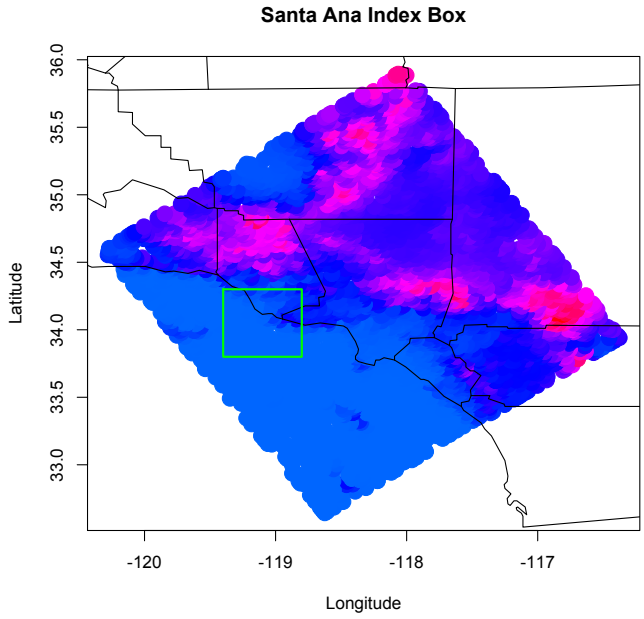


Figure 5.1: Santa Ana Index Bounding Box

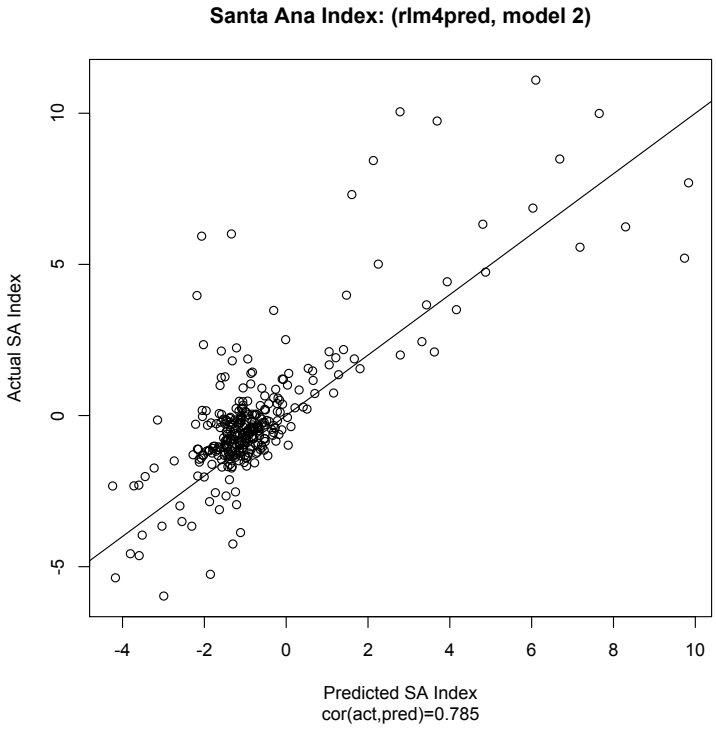


Figure 5.2: Santa Ana Index: October 1991-2000

Figure 5.2 plots all 310 Santa Ana indices against our predicted Santa Ana index for this ten year time span. For these predictions, I used the top performing transformed linear model to predict magnitudes and the kernel smoothing circular regression for wind direction predictions. A Pearson's correlation coefficient of 0.785 was calculated between the actual index and the predicted index. Figure 5.3 plots the overall ten year distributions of Santa Ana indices between the actual and the predicted. By looking at both figures 5.2 and 5.3, we can see that the model does a good job in predicting the overall distribution, but there remains a slight overall trend of under-prediction by the model.

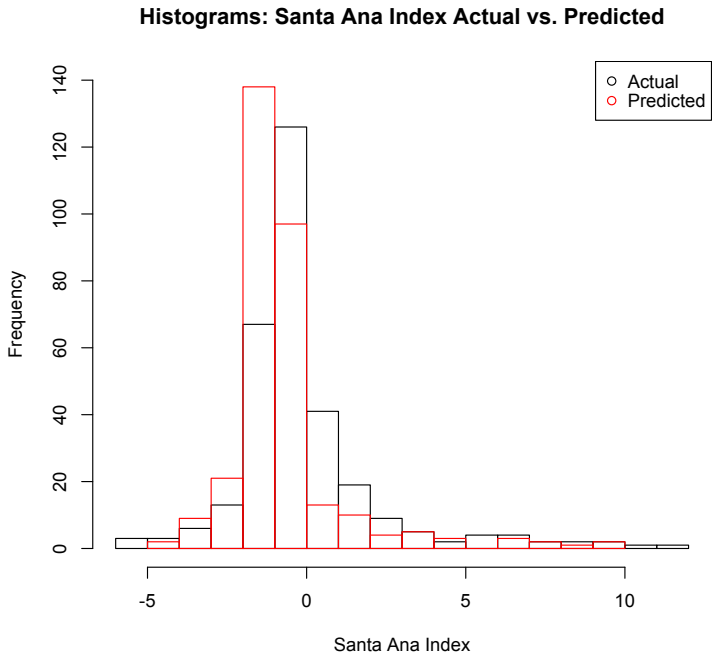
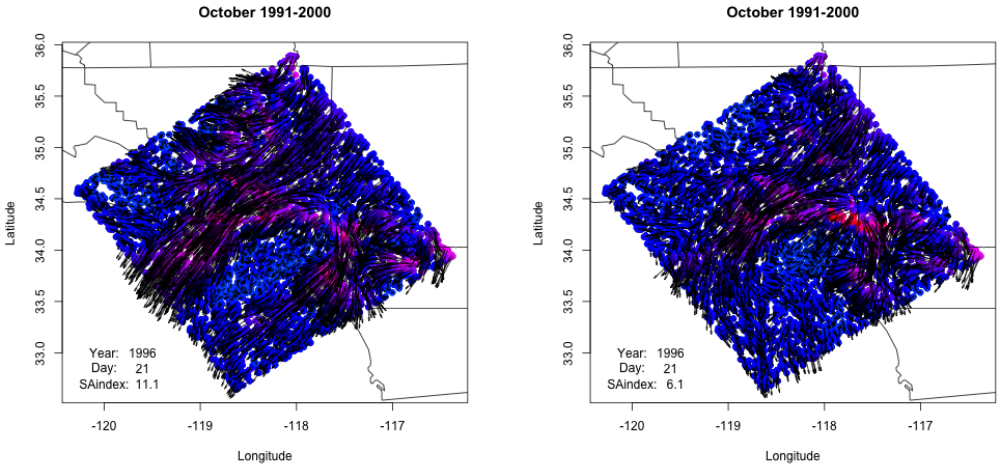


Figure 5.3: Santa Ana Index: Actual vs. Predicted

Of the sixteen days with actual Santa Ana indexes greater than five, we can compare the two-dimensional maps of the actual wind fields and the predicted wind fields. The first comparison is of the strongest observed Santa Ana index which occurred on October 26th of 1996, seen in figure 5.4. The left hand side represents the actual wind field, with a Santa Ana index of 11.1, while the right

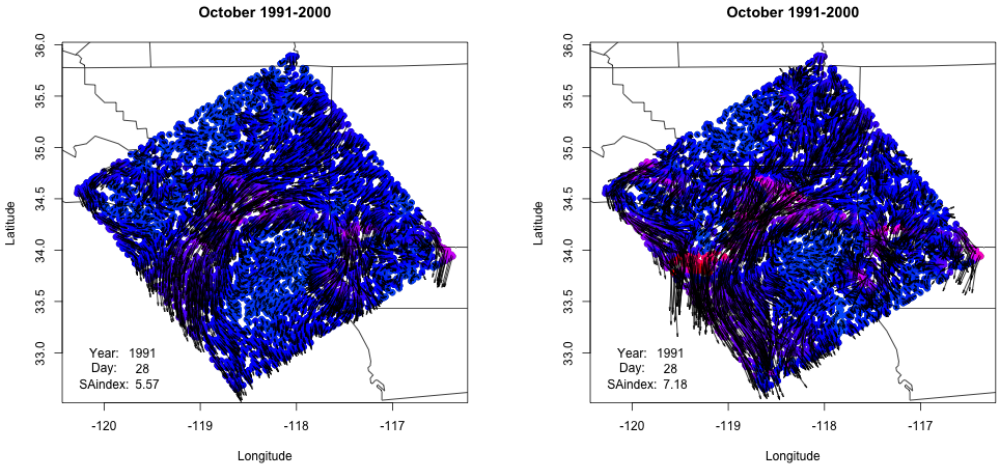
hand side displays the models predicted wind field, with a Santa Ana index of 6.1.



(a) Actual Wind Field (b) Predicted

Figure 5.4: Strongest Observed Santa Ana

On the lower end, a mild Santa Ana event occurred on October 28th of 1991 that registered a Santa Ana index of 5.57. Figure 5.5 again shows the actual wind field versus the predicted wind field.



(a) Actual Wind Field (b) Predicted

Figure 5.5: Mild Santa Ana Event

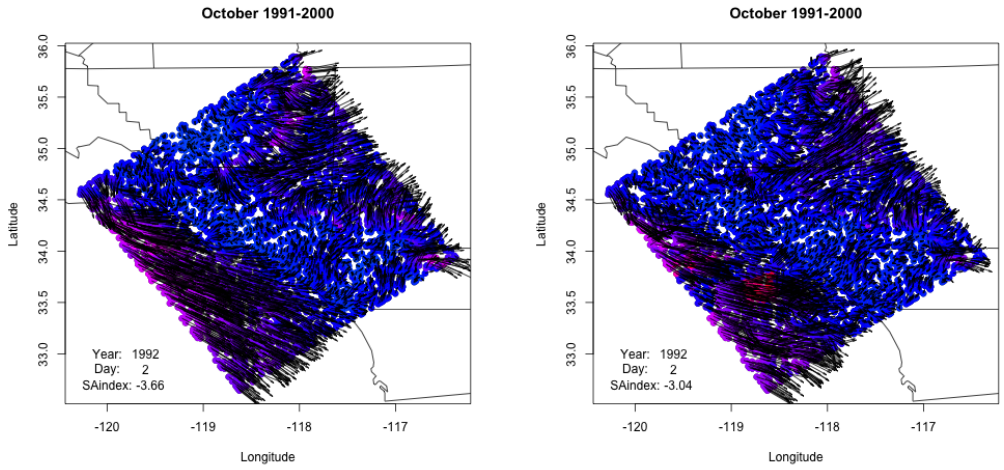
These figures 5.4 and 5.5 show that the regime based cluster modeling is able to predict the wind field to an accurate degree. The model is able to pick up the local nuances created by high-resolution topography in a Santa Ana event. In the strong Santa Ana event, figure 5.4 we can see the Santa Ana generates from the High Desert and has large effects over the Central Valley and Orange County. The model is able to detect and recreate that, but just not to the desired intensity. In contrast, the more mild event, figure 5.5 has strong winds near the coast that curve towards the southern direction over the open ocean. We can see the model detects this along with the Orange County coastal winds, but just with a little more intensity.

5.2 Onshore Predictions

We can do the same sort of spatial analysis for onshore winds. It can be assumed that some onshore wind events will have a larger negative Santa Ana index number, indicating a reverse in flow direction. A typical strong onshore event occurred on October 2nd in 1992 and is indicated in figure 5.6, with the actual Santa Ana index of -3.66 and a predicted index of -3.04.

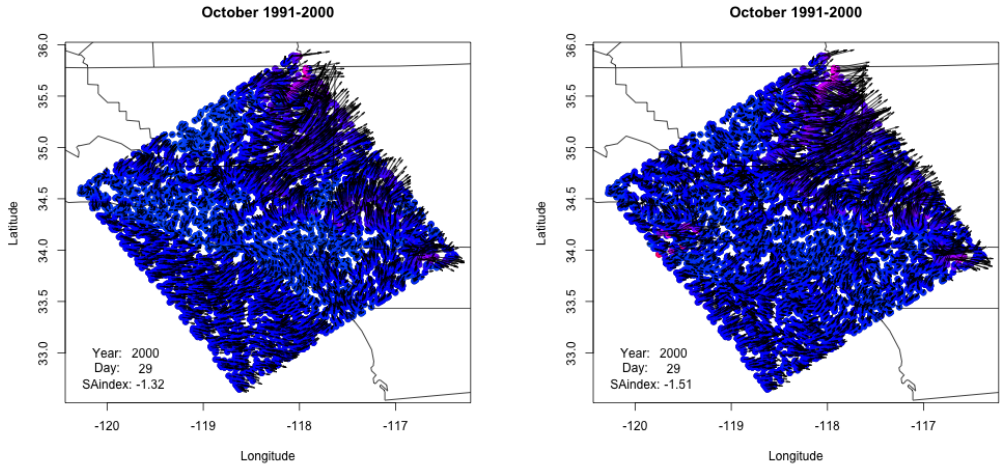
There are other onshore events where the wind generated over the ocean is not as strong, but strong winds are still blown onto the high desert. This phenomenon creates similar patterns over land but has a relatively small negative Santa Ana index. One example of this event occurred on October 29th of the year 2000 and can be seen in figure 5.7. The actual Santa Ana index is -1.32 and the predicted index is -1.51.

By comparing figures 5.6 and 5.7 we can see that the model not only can detect different wind regimes, but more subtle classifications within regimes. We can even see that in both figures, the model detects the strong pocket of winds in the small Palm Springs valley.



(a) Actual Wind Field (b) Predicted

Figure 5.6: Strong Onshore Event



(a) Actual Wind Field (b) Predicted

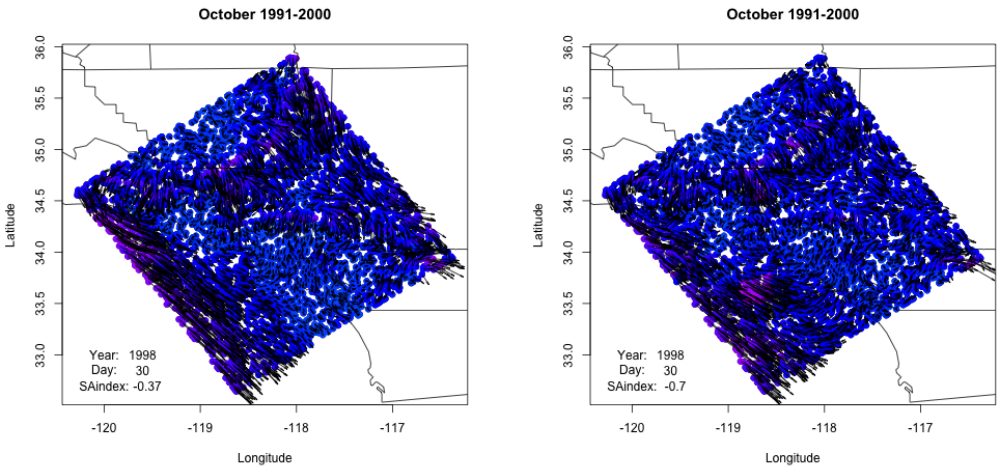
Figure 5.7: Mild Onshore with Strong High Desert Event

5.3 Mild and Other Predictions

Events with a small Santa Ana index may fall under the mild wind regime, but not all are "mild". For example, on October 30th of 1998 an actual Santa Ana

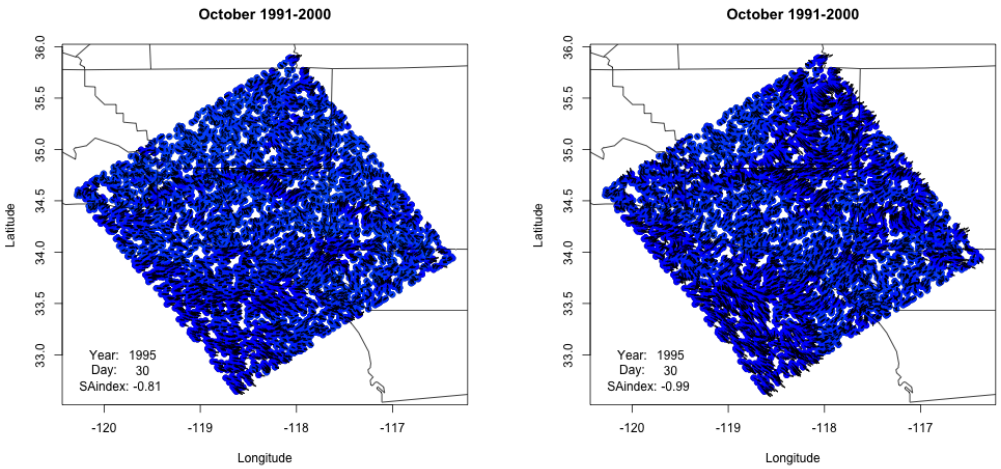
index was scored at -0.37, close to zero. But as we can see in figure 5.8, there is stronger northwesterly wind over the ocean and pouring out of the Central Valley.

Although, the model is also adept at predicting actual mild wind events. An actual mild wind regime along with the predicted can be seen in figure 5.9.



(a) Actual Wind Field (b) Predicted

Figure 5.8: "Mild" Event with Central Valley Pour



(a) Actual Wind Field (b) Predicted

Figure 5.9: Mild Event

CHAPTER 6

Discussion

The main purpose of this dissertation was to provide a new technique for modeling two-dimensional vector fields. It is important to note that even though the aim of the example was to predict high-resolution latticed grids. The same techniques can be used to predict just a few locations with historical data. The main strength of this method was the ability to capture the regime change and subsequent spatial dependent structure change. The models used to make predictions, after the unique clustering method was performed, were not innovative and even simple in their structure. I believe the simplicity of model type and limited number of covariates per model attests to the power of the data segmentation by flow direction and magnitude. Improvements to this method can be made in both the clustering process and the modeling process, including the covariate selection process and model selection type.

When performing the clustering process, we used the clustering ratio that was most robust across all prediction locations. For magnitude predictions this was a [1:10] ratio and for wind directional predictions this was a [10:1] ratio. The optimal, but most expensive computationally, would be to find the optimal clustering ratio for each individual prediction location. This may be preferred and even possible if the prediction locations are few (e.g. ocean buoys or weather stations). A more simple improvement would be to find optimal clustering ratios for land types or location type (i.e. coastal, mountain, desert, etc.).

When selecting covariates for the model building process, no account for co-

linearity between independent variables was considered (which is only applicable to the magnitude predictions because the angular prediction models only use one circular covariate). A simple improvement would be to perform principle component analysis on the covariate field of low-resolution magnitudes and then regress on a selected few principal components that account for the most variance in the independent data set.

In consideration to the actual models selected, regression models with circular response variables are still a relatively new statistical research topic. As time progresses and improvements in these models occur, the new circular models can easily be plugged into these methods and compared to previous model types.

Temporal changes in modeling can also be experimented with. For our model testing we selected a specific month, October, and modeled daily averages across ten years. Modeling can be month specific, like in this research, or even more general, grouping Santa Ana months together. This would create only two model processes to predict a whole year and would additionally significantly increase the amount of training data available.

CHAPTER 7

References

Bibliography

[1] Bryson C. Bates, Stephen P. Charles, and James P. Hughes. Stochastic downscaling of numerical climate model simulations. *Environmental Modeling and Software*, 13:325–331, 1998.

[2] Sebastien Conil and Alex Hall. Local regimes of atmospheric variability: A case study of southern california. *Journal of Climate*, 19(17):4308–4325, September 2006.

[3] Wim C. de Rooy and Kees Kok. A combined physical-statistical approach for the downscaling of model wind speed. *Weather and Forecasting*, 19:485–495, 2004.

[4] Marco DiMarzio, Agnese Panzera, and Charles Taylor. Non-parametric regression for circular responses. *Scandinavian Journal of Statistics - Theory and Applications*, 40:238–255, 2012.

[5] Nazzareno Diodato. The influence of topographic co-variables on the spatial variability of precipitation over small regions of complex terrain. *International Journal of Climatology*, 25:351–363, 2005.

[6] N. I. Fisher and A. J. Lee. Regression models for an angular response. *Biometrics*, 48(3):665–677, 1992.

[7] K. Goubanova, V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, and M. Vrac. Statistical downscaling of sea-surface wind over the peru-chile upwelling region: diagnosing the impact of climate change from the ipsl-cm4 model. *Climate Dynamics*, 36:1365–1378, 2011.

[8] D. Gyalistras, H. von Storch, A. Fischlin, and M. Beniston. Linking gcm-simulated climatic changes to ecosystem models: case studies of statistical downscaling in the alps. *Climate Research*, 4:167–189, 1994.

[9] B.C. Hewitson and R.G. Crane. Self-organizing maps: applications to synoptic climatology. *Climate Research*, 22:13–26, 2002.

[10] Hugo G. Hidalgo, Michael D. Dettinger, and Daniel R. Cayan. Downscaling with constructed analogues: Daily precipitation and temperature fields over the united states. *California Energy Commission, PIER CEC-500-2007-123; California Climate Change Center Report Series Number 2007-027*, pages 1–62, January 2008.

[11] Tim Hoar and Doug Nychka. Statistical downscaling of community climate system model (ccsm) monthly temperature and precipitation projections. *White Paper*, April 2008.

[12] Mimi Hughes and Alex Hall. Local and synoptic mechanisms causing southern california's santa ana winds. *Climate Dynamics*, 34(6):847–857, 2010.

[13] Radan Huth. Statistical downscaling in central europe: evaluation of methods and potential predictors. *Climate Research*, 13:91–101, 1999.

[14] S. Rao Jammalamadaka and A. SenGupta. *Topics in Circular Statistics*. World Scientific Publishing Co. Pte. Ltd., 2001.

[15] J. Kim, J. Chang, N. Baker, D. Wilks, and W. Gates. The statistical problem of climate inversion: Determination of the relationship between local and large-scale climate. *American Meteorological Society*, 112:2069–2077, October 1984.

[16] Phaedon C. Kyriakidis. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289, July 2004.

[17] Ulric J. Lund. Least circular distance regression for directional data. *Journal of Applied Statistics*, 26:723–733, 1999.

[18] Ulric J. Lund. Tree-based regression for a circular response. *Communications in Statistics - Theory and Methods*, 31(9):1549–1560, 2002.

[19] K. V. Mardia. Linear-circular correlation coefficients and rhythmmometry. *Biometrika*, 63:403–405, 1976.

[20] Fedor Mesinger, Geoff DiMego, Eugenia Kalnay, Kenneth Mitchell, Parry C. Shafran, Wesley Ebisuzaki, Dusan Jovic, Jack Woollen, Eric Rogers, ernesto H. Berbery, Michael B. Ek, Yun Fan, Robert Grumbine, Wayne Higgins, Hong Li, Ying Lin, Geoff Manikin, David Parrish, and Wei Shi. North american regional reanalysis. *Bulletin of the American Meteorological Society*, March 2006.

[21] J. Michalakes, J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock, and W. Wang. The weather research and forecast model: Software architecture and performance. *to appear in proceedings of the 11th ECMWF Workshop on the Use of High Performance Computing In Meteorology, 25-29 October 2004, Reading U.K. Ed. George Mozdzynski*, October 2008.

[22] Adam Monahan. Can we see the wind? statistical downscaling of historical sea surface winds in the subarctic northeast pacific. *Journal of Climate*, 25(5):1511–1528, 2012.

[23] Ian A. Nalder and Ross W. Wein. Spatial interpolation of climatic normals: test of a new method in the canadian boreal forest. *Agricultural and Forest Meteorology*, 92:211–225, July 1998.

[24] T. Salameh, P. Drobinski, M. Vrac, and P. Naveau. Statistical downscaling of near-surface wind over complex terrain in southern france. *Meteorology and Atmospheric Physics*, 103:253–265, 2009.

[25] Y.R. Sarma and S. Rao Jammalamadaka. Circular regression. In *Proceedings of the Third Pacific Area Statistical Conference*, pages 109–128, 1993.

[26] Guillermo Tabios and Jose Salas. A comparative analysis of techniques for spatial interpolation of precipitation. *American Water Resources Association: Water Resources Bulletin*, 21(3):365–380, June 1985.

[27] D. White, M. Richman, and B. Yarnal. Climate regionalization and rotation of principal components. *International Journal of Climatology*, 11:1–25, 1991.

[28] C. Wikle, R. Milliff, D. Nychka, and M. Berliner. Spatiotemporal hierarchical bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397, 2001.

[29] R.L. Wilby, S.P. Charles, E. Zorita, B. Timbal, P. Whetton, and L.O. Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods. *Supporting material of the Intergovernmental Panel on Climate Change (IPCC), prepared on behalf of Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA)*, August 2004.

[30] R.L. Wilby and T.M.L. Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548, 1997.

[31] R.L. Wilby, T.M.L. Wigley, D. Conway, P.D. Jones, B.C. Hewitson, J. Main, and D.S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34(11):2995–3008, November 1998.