Abstract

ALFARO CÓRDOBA, MARCELA. Variable Selection Methods with Applications to Atmospheric Sciences. (Under the direction of Montserrat Fuentes and Joseph Guinness.)

In a data rich world, the task of selecting variables that have a statistically significant contribution in the process of modeling, forecasting or emulating atmospheric behavior is both essential and challenging. In this dissertation we extend the variable selection methods available for hierarchical models and statistical emulators in the Bayesian context to accommodate spatial and spatial-temporal correlation, zero inflated count data and multivariate response. These methodological advances are motivated by and illustrated with atmospheric sciences applications.

Tropical cyclone and sea surface temperature data have been used in several studies to forecast the total number of hurricanes in the Atlantic Basin. Sea surface temperature (SST) and latent heat flux (LHF) are correlated with tropical cyclone occurrences, but this correlation is known to vary with location and strength of the storm. The objective of the first part of this dissertation is to identify features of SST and LHF that can explain the spatial-temporal variation of tropical cyclone counts, categorized by their strength. We develop a variable selection procedure for multivariate spatial-temporally varying coefficients, under a Poisson hurdle model (PHM) framework, which takes into account the zero inflated nature of the counts. The method differs from current spatial-temporal variable selection techniques by offering a dynamic variable selection procedure, which shares information among responses, locations, time and levels in the PHM context. Numerical studies are presented to analyze the median absolute deviation and the proportion of correctly estimated coefficients in several scenarios. The model is used to study the association between SST and LHF and the number of tropical cyclones of different strengths in 400 locations in the Atlantic Basin over the period of 1950-2013.

In the second part of this dissertation we present a fast alternative for spatially varying vari-
able selection, inspired in the concept of Bayesian model averaging, which is used to develop a statistical emulator for regional climate models (RCM). RCMs provide high-resolution climate simulations that are important to investigate variation in regional scale projections. We introduce a Bayesian statistical framework to emulate RCM output variables using output variables from Global Circulation Models (GCM) as covariates. This emulator is an alternative to the RCM runs, which are computationally expensive. The Bayesian RCM emulator has spatially varying coefficients with data-dependent priors to perform variable selection in each location. To that end, we model both the spatially varying coefficients and the posterior probability of the response as smooth spatial processes. The method is then applied to emulate precipitation from the CRCM regional model output from the North American Regional Climate Change Assessment Program (NARCCAP) using a set of variables from the CCSM global circulation model output. Furthermore, the prior weights are used to assess the usefulness of each covariate from the global model output, to emulate the regional model for each location of the NARCCAP data domain. We built the model on Summer data from 1970 to 1998 over a spatial domain covering the Western part of the continental United States and Canada and validate it with data from 1999-2000 over the same spatial domain.

The final part of the thesis presents a verification method for seasonal statistical hurricane forecasts. Each year, U.S. forecasters predict the number of seasonal named storms, hurricanes, and major hurricanes that will occur in the Atlantic Basin (including the Caribbean Sea and the Gulf of Mexico), as well as in the North East Pacific. With this verification we aim to assess the validity of forecasts and quantify their uncertainty, to improve forecasting and develop better hurricane preparedness strategies. We apply the procedure in order to evaluate the impact of using forecasted ENSO indices, as well as to compare the skill of forecasts that use information up to March versus up to May in the Atlantic Basin. The forecast and verification includes a variable selection procedure from the Frequentist perspective. The different options are evaluated using the data from climate indices and tropical cyclone data from 1951-2015.
Variable Selection Methods with Applications to Atmospheric Sciences

by
Marcela Alfaro Córdoba

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2017

APPROVED BY:

Ana-Maria Staicu
Lian Xie

Montserrat Fuentes
Co-chair of Advisory Committee

Joseph Guinness
Co-chair of Advisory Committee
Dedication

A mi abuela (in memoriam). A mi mamá y a mi hermana.
Biography

Marcela Alfaro Córdoba was born and grew up in San Ramón, Costa Rica. She started college at the University of Costa Rica, Sede de Occidente in 2000. She moved to San José, Costa Rica in 2003 to pursue her Bachelor of Science degree in Statistics. After graduating from UCR in 2007, she went to Iowa State University in Ames, IA where she earned her Masters in Science in Statistics in 2009. In the next two years, she worked as an invited lecturer in the Department of Statistics at the University of Costa Rica. She joined the Department of Statistics at North Carolina State University in 2011 to pursue a Ph.D. degree in Statistics.
Acknowledgements

I wish to express my gratitude to my co-advisors, Dr. Montse Fuentes and Dr. Joe Guinness for their constant support, positive energy, guidance and for reading my work over and over again. I thank Dr. Ana-Maria Staicu not only for being part of my committee but also for encouraging me to collaborate in new projects and explore new areas of research. I thank Dr. Fuentes, Dr. Guinness and Dr. Staicu for serving as an example of hard work and commitment with the practice of science and for being my role models during my graduate career.

I would like to thank the members of the NCSU research groups I regularly attended during my studies: STATMOS journal club, Environmental Statistics Lab and Dr. Lian Xie’s lab in MEAS. In particular, I would like to thank Dr. Lian Xie for teaching me about atmospheric phenomena, guiding me on how to communicate science to the media and also for his advice and insightful comments for my research.

I thank all of the faculty and my professors in the Stats Department at NCSU for providing a good support structure, an excellent academic training and a nice work environment. In particular, I thank Dr. Charles Smith for his nice treats during the long night shifts and Dr. Consuelo Arellano for being my support structure throughout the years. Also, I thank Dr. Jaqueline Hughes-Oliver, Dr. Kimberly Weems, and Dr. Marcia Gumpertz for opening spaces to students of diverse backgrounds and for making me feel welcome in the department since day one. I also thank all of the staff at NCSU, in particular Alison McCoy, Chris Waddell and Terry Byron for always helping me to solve problems.

With gratitude, I am pleased to acknowledge all of the funding sources that supported my research: Fondo de Incentivos from MICIT-Conicit Costa Rica, the University of Costa Rica, and STATMOS through Dr. Fuentes’ funding. I thank my professors in the University of Costa Rica who encouraged me to pursue this degree and supported my research plans. Specifically, M.Phil Oscar Hernández, Dr. Edgar Gutiérrez, Dr. Eiliana Montero and Master Fernando Ramírez for all their help and encouragement.
Part of my support network was provided by NCSU. I do not think I would have been able to finish my degree without the Recreation and Fitness programs and their amazing instructors in these years: Jeffrey, Holly, Natalie, Esdras, Ciara, Taylor and all of the yoga instructors. Also, the Graduate School Dissertation Institute and the Graduate Student Professional Workshop, that gave me tools and safe nets to feel confident that I could successfully finish my dissertation and start a promising professional career.

I am grateful for my support network here in Raleigh. In particular, I cannot thank enough to So Young, Markus and Diego for being my close family in the area during these years, and making graduate school a pleasant and happy process. I am grateful to have Matt, Cissy, Christie, Alejo, Santa, Anita, Héctor, Leo, Etna, Miguel, Víctor, Andrés and Meme becoming part of my extended family in these 6 years. A big part of this dissertation was done thanks to their encouragement, support, great food and all of the fun times spent together.

Last but not least, my infinite gratitude goes to my mom and sister for always being there. To my father, to George, and to my other moms and uncles of my blood and extended family in Costa Rica, thank you for providing that special place where I can always return to. Heartily thanks to Louli, Maria Paula, Caro H., Andre, Jeffrey, Nani, Caro F., Mónica, Cata, Massiel, Kristian, Esteban, Octavio, Levke and Dan for their support from the US and abroad.
List of Tables

Table 2.1 Mean acceptance rate for all parameters. ............................................. 22
Table 2.2 Median MAD per $\beta_A$ over locations and trimesters. All standard errors are less than 0.006. Proportion of correctly estimated $\beta_A = 0$ All standard errors are less than 0.005. NA in cases where we did not simulate zeroes. .................. 22
Table 2.3 Goodness of fit criterion DIC, and MSE overall, per location and per year for the full model (ST), spatial (SP) and independent (IN). DIC bar values (standard error in parenthesis) for models using different grid box sizes, compared to a model using climate indices. ................................................................. 24
Table 2.4 Posterior estimates for each response for the overall mean $\alpha_A$, the proportion $\pi_A$ of coefficients centered in $\alpha_A \neq 0$ and the variability $\sigma_A$ of coefficients centered in $\alpha_A$, for responses $Y_1 =$ Tropical Storms, $Y_2 =$ Hurricanes and $Y_3 =$ Major Hurricanes. Values with (*) have credible intervals that does not include zero. ................................................................. 26
Table 3.1 Descriptive statistics for the covariates. ............................................. 36
Table 3.2 Six settings with their respective sets for $v_k(s)$ and $Y$. ....................... 43
Table 3.3 Average MSE for $\beta$ and s.e in parenthesis, for each model and setting. . . . 44
Table 3.4 Average MAD for $\beta$ and s.e in parenthesis, for each model and setting. . . 45
Table 3.5 Average coverage for $\beta$ and s.e in parenthesis, for each model and setting. Nominal level is 95% ................................................................. 45
Table 3.6 Time to run 10000 MCMC iterations per model, N = 100, T = 12. ............. 46
Table 3.7 Timing comparison for all models. Comparisons are done on a Dell Optiplex 9020 with a 3.60GHz Intel i7-4790 processor with 32GB RAM, and the code is written in the R programing language (R Core Team [2013]). ...................... 49
Table 3.8 In-sample results for all emulators using different values of $J$ in the models that use basis decomposition representation. Standard errors are less than 0.02 for MSE, MAD and MLS. ................................................................. 50
Table 3.9 Out-of-sample results for all emulators using different values of $J$ in the models that use basis decomposition representation. Standard errors are less than 0.05 for MSE, MAD and MLS. ................................................................. 50
Table 4.1 Forecast groups that use statistical models exclusively, with their most recent reference. ................................................................. 55
Table 4.2 Possible variables and time domains. .................................................... 70
Table 4.3 Atlantic Basin tropical storms: five most selected variables using LASSO, per $F$. NINO variables are presented in bold letters. ................................................................. 73
Table 4.4 Median p-values for deviance goodness-of-fit test per response. Null hypothesis is the Poisson model is a good fit for the data. ................................................................. 75
Table A.1  Proportion of locations with estimated coefficients statistically different from zero (credible intervals do not include zero). From those coefficients different from zero (+) indicates that more than 60% of the locations have a positive coefficient for that specific trimester, response and score. In a similar way, (-) indicates that more than 60% of the locations have a negative coefficient for that specific trimester, response and score. Bold values are significant factors. Red values are those displayed in Figures 2.4, 2.5, A.3 and A.4.
List of Figures

Figure 2.1  HURDAT 6 hour track data from 1950 to 2013  
Figure 2.2  Definition of trimesters compared to the tropical cyclone season for TC data.  
Figure 2.3  Anomalies in the North Atlantic Basin during the first trimester for SST (left column) and LHF (middle column). Counts per grid box of tropical storms (right column). Years 1955, 1980, and 2013 randomly selected to illustrate data.  
Figure 2.4  Winter: SST Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.  
Figure 2.5  Spring: SST or LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.  
Figure 3.1  Coordinates for the Canadian Regional Climate Model (CRCM) and the Community Climate System Model (CCSM) in the North American Region. The area outlined in red is the selected domain for the analysis in Section 3.5.  
Figure 3.2  Response variable from the CRCM output and temperature (TREFHT) from the CCSM output.  
Figure 3.3  Elevation in meters in the North American Region. This variable was used as a driver to generate CRCM output, and although we do not include it in our model, it is highly correlated with several of the covariates we use over the domain of interest.  
Figure 3.4  Example of $p = 4$ fields for $f_k(s)$.  
Figure 3.5  Setting 2 (chessboard) and 3 (smooth) for $v_k(s)$.  
Figure 3.6  Trace plots for parameter $\sigma^2$ for (a) Model 2b and (b) Model 2c.  
Figure 3.7  Comparison between model 1b and model 2b using a random sample of locations and places from the out-of-sample data set. Black dots represent posterior medians, and the lines represent the 95% credible intervals, for each model.  
Figure 3.8  Weights $v_k(s)$ for each covariate. Areas in white represent values not significantly different from zero, i.e. areas where the respective covariate does not represent a significant driver for summer precipitation.  
Figure 4.1  Sliding windows: each row is a window $w$, displaying a training dataset ($n = nT$) in blue, test data ($n = nE$) in green and the final forecast year (2016) is displayed in red.  
Figure 4.2  Delimited regions for the forecast.
Figure 4.3 Cluster analysis results for 33 possible covariates for all the data, as an example. Strong red tones represent perfect positive correlation and strong blue tones represent perfect negative correlation. A dendrogram shows the grouped variables. ................................................................. 69

Figure 4.4 Process diagram for the verification method .............................................. 72

Figure 4.5 In the left column, the difference between climatology and MSE for Atlantic Basin tropical storm models $F$ are presented for each model, and for SWCV and LOOCV. The percentage of change between MSE for model $F$ and climatology is presented on top of each observation. Positive values mean that model $F$ has lower MSE than climatology. Right column present the results for $H_i = LS(F_{clim}, y_i) - LS(F_{LASSO}, y_i)$ for each Atlantic Basin tropical storm model and SWCV and LOOCV. Intervals not overlapping zero represent significant difference between LASSO and cluster models. Positive values indicate superior skill of the cluster model, compared to LASSO. . . 76

Figure 4.6 First column is the mean $H_i = LS(F_{LASSO}, y_i) - LS(F_{cluster}, y_i)$ by response. Intervals not overlapping zero (in red) represent significant difference between LASSO and cluster models. Positive values indicate superior skill of the cluster model, compared to LASSO. Second column is the mean $H_i = LS(F_{clim}, y_i) - LS(F_{LASSO}, y_i)$ by response. Intervals not overlapping zero (in red) represent significant difference between climatology and LASSO. Positive values indicate superior skill of LASSO, compared to climatology. . . 78

Figure A.1 MSE per location for each response: (a) Tropical Storms (TS), (b) Hurricanes (HU) and (c) Major Hurricanes (MH). ................................................................. 89

Figure A.2 MSE per year for each response: (a) Tropical Storms (TS), Hurricanes (HU) and Major Hurricanes (MH). ................................................................. 91

Figure A.3 Summer: SST or LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1. ................................................................. 92

Figure A.4 Fall: LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1. ................................................................. 93

Figure A.5 Example of simulation settings realizations for covariate 1. ......................... 93

Figure A.6 First four empirical orthogonal functions for sea surface temperature anomalies in the Atlantic Basin, per season. ................................................................. 94

Figure A.7 First four empirical orthogonal functions for latent heat flux anomalies in the Atlantic Basin, per season. ................................................................. 95

Figure A.8 Time series for each $\xi_{t,r}(t, w)$. 2 scores for SST are presented in the first row and 2 scores for LHF in the second row. ................................................................. 96
Figure A.9  Statistically significant factors per season ($w$), EOF ($r$) and covariate ($\ell$). First column represents SST coefficients, left column LHF coefficients. First row are coefficients that associate covariates with the occurrence of TS, second row with the occurrence of HU and third row are coefficients that associate covariates with the occurrence of MH.

Figure A.10 Statistically significant factors per season ($w$), EOF ($r$) and covariate ($\ell$). First column represents SST coefficients, left column LHF coefficients. First row are coefficients that associate covariates with the intensity parameter of TS, second row with the intensity parameter of HU and third row with the intensity parameter of MH.

Figure B.1 Percentage of times a Variable X is selected using LASSO models, per response and for SWCV. In total, there are 35 windows (repetitions of the model). Nino variables are highlighted with black lines.

Figure B.2 Percentage of times a Variable X is selected using cluster models, per response and for SWCV. In total, there are 35 windows (repetitions of the model). Nino variables are highlighted with black lines.
Chapter 1

INTRODUCTION

1.1 Motivation

Understanding and describing the uncertainty in atmospheric and oceanic behavior is one of the main goals of atmospheric sciences. Statistics has played an important role in every aspect of this goal, that can be divided into atmospheric statistical modeling and empirical statistical analysis (Solow [2003]). Statistical modeling and forecasting using data, and emulation of atmospheric models are methods that serve to quantify, evaluate, and make inferences about the climate in the face of uncertainty (Wilks [2011]). In a data rich world, the task of identifying variables that have a statistically significant contribution in describing atmospheric behavior is both essential and challenging. This dissertation focuses on two specific areas of study in atmospheric sciences: first, to gain better understanding of which factors drive tropical cyclone seasons in the Atlantic Basin, and to provide methodological tools to verify its forecast; and second, to propose a statistical framework to emulate and select drivers of regional climate models for the North American region.

Tropical cyclones (TC) are rotating systems characterized by strong winds and heavy rain that have caused numerous deaths and economic, agricultural and social damages in the past (Xie [2005]). A good understanding of their seasonal spatial distribution is crucial to develop
methods to model their locations and strengths over time. The scientific problem of interest addressed in Chapter 2 is whether space-time features of sea surface temperature (SST) and latent heat flux (LHF) are significantly associated with seasonal TC activity in specific locations. For this purpose, we develop a spatial-temporal variable selection method for hierarchical Bayesian models, which can accommodate spatial-temporal correlation, zero inflated count data, and a multivariate response.

In Chapter 3, we shift our focus to atmospheric modeling. General Circulation Models (GCM) are climate models that describe the large scale global atmospheric and oceanic dynamics. Because of their coarse resolution, GCMs provide limited information when assessing the impact of climate at regional scales. Regional Climate Models (RCM) are dynamic downscaling models that use atmospheric interactions in the GCM to develop a higher resolution output. They are useful to describe regional climate as well as to serve as an input for weather prediction models. However, they require orders of magnitude more computation time than statistical downscaling to compute equivalent scenarios (Wilby et al. [2000]). We present a novel statistical framework to emulate RCM output using GCM output, which can identify GCM drivers per location and that is computationally faster than its dynamical downscaling counterpart.

The last part of this thesis focuses on validation of seasonal TC forecasts. Quantifying and characterizing the uncertainty in seasonal tropical cyclone forecasts is essential for their proper statistical interpretation and further utilization. Forecasts are a result of many decisions such as which month-specific covariates to include, how to select them, and what probability distribution to use. Thus, it is of scientific interest to evaluate the impact of these decisions when constructing a forecast. While many authors like Gray et al. [1994], Elsner and Schmertmann [1994] and Klotzbach [2007], to name a few, have published statistical forecast for seasonal TC in the past, there is no reference for a statistical validation framework of the model. In Chapter 4, we provide a framework to validate statistical seasonal TC forecasts while taking into account their probabilistic nature, and also providing an option to incorporate variable selection in the
model building process.

In the following section, we detail well-established background concepts and methods that will be used throughout this dissertation.

1.2 Background

In spatial statistics, each point-level data with location index $s$ is assumed to vary continuously over $D$, a fixed subset of $\mathbb{R}^d$. Suppose that we observe data \{$Y(s_i) : s_i \in D$\} with index $i \in \{1, \ldots, N\}$ and define a latent spatial process $w$. A basic model for spatial processes is:

$$Y(s_i) = \mu(s_i) + w(s_i) + \varepsilon(s_i)$$

where $\mu(s_i)$ represents the mean structure, $w$ is a zero-centered Gaussian spatial process that describes the residual spatial association, and $\varepsilon(s_i)$ is an independent and identically distributed non-spatial residual, with mean zero and variance $\sigma^2$.

The spatial association is commonly represented using a valid correlation function $C(s_i, s_j, \Omega)$, where $\Omega$ is a set of spatial parameters. When $C$ is a function of the separation between sites $h = s_i - s_j$ then the model is stationary. Furthermore, if this association depends upon $h$ only through its length, then the model is isotropic. Examples of valid correlation functions include the Exponential, Gaussian and Matérn functions. Depending on the function we choose to model our data, a set of parameters $\Omega$ will be used to describe the spatial structure.

Spatial models can be extended when we observe $Y_t(s_i)$ in several discretized time points $t \in \{1, \ldots, T\}$, in which case the mean structure can be adjusted to vary with $t$ as

$$Y_t(s_i) = \mu_t(s_i) + w(s_i) + \varepsilon_t(s_i),$$

with $\mu_t(s_i) = \sum_{p=0}^{P} X_{tp}(s_i) \beta_p$, where $X_{tp}(s_i)$ is the observed $p$ covariate in time $t$ and location $s_i$, and $\beta_p$ is its respective regression coefficient. In certain applications, the assumption of a
regression coefficient being constant over space and time is not reasonable. For example, when modeling atmospheric phenomena we cannot assume that the temperature has a constant effect on precipitation over different locations with different altitudes and during different seasons. Gelfand et al. [2003a] formulates a spatial model with spatially varying coefficients, where the mean structure and the latent spatial vector are represented as the sum of non-spatial and spatial components: 

$$
\mu(s_i) + w(s_i) = \sum_{p=0}^{P} X_{tp}(s) \beta_p + X_{tp}(s) \beta_p(s).
$$

The spatially varying coefficients $\beta_p(s)$ are estimated as latent spatial processes which allow us to do spatial inference on the effects. The extension to spatially-temporally varying coefficients uses a latent spatial-temporal process instead of a latent spatial process to model each coefficient. A new challenge arises when the observations, locations and the covariates are numerous and we want to choose the optimal model, i.e. select the optimal combination of covariates, locations and seasons to model the response.

The surge of geo-referenced data sources and advances in computational methods in recent decades have motivated many theoretical and methodological developments in spatial statistics; see Cressie [2015], Wikle and others [2003], Gelfand et al. [2010], Diggle [2013] to name a few. Specifically, Bayesian methodology has had a rapid development in terms of more efficient MCMC sampling methods to estimate spatial models (Banerjee et al. [2014]). Bayesian spatial and spatial-temporal methods are flexible and convenient to simultaneously calculate coefficients with valid error estimates, as well as to incorporate the necessary regularization suitable for high dimensional modeling.

Hierarchical Bayesian models are appropriate in cases where we are interested in predictions of the spatial process, given that they can provide an intuitive way to “represent scientific knowledge in a series of conditional models, coherently linked together by simple probabilistic rules” (Wikle and others [2003]). When the inference in spatial statistics is mainly concerned with the regression coefficients, the use of a marginal model for that parameter can simplify the problem (Gelfand et al. [2010]). In Chapters 2 and 3 we assume a hierarchical Bayesian framework with spatially varying coefficients to model the data. In Chapter 2, we use the
marginal distribution of $\beta$ to propose a novel variable selection method to accommodate cases where $\beta$ varies in space and time. For that, we rely on stochastic search variable selection (SSVS) and the use of copulas to represent the spatial-temporal association. Then, in Chapter 3 we present a fast emulator that includes a method to perform spatially varying variable selection inspired in the concept of Bayesian model averaging (BMA).

1.3 Variable Selection for Bayesian Models

In the Bayesian framework, the variable selection problem can be viewed as a problem of parameter estimation: rather than searching for the single optimal model, we attempt to estimate the posterior probability of all models within the considered class of models. The advantage of using Bayesian variable selection methods is that they are equipped with natural measures of uncertainty, such as the posterior probability of each model and the marginal inclusion probabilities of the individual predictors (O’Hara et al. [2009]). MCMC methods can be used to sample the highest probability models, as opposed to fitting all possible models and comparing Bayes factors, where the number of potential models with $P$ predictors ($2^P$) quickly becomes intractable.

Stochastic Search Variable Selection (SSVS)

Stochastic Search Variable Selection (George and McCulloch [1993]) has become a popular tool for selecting among a large set of potential predictors. The method was first proposed in the context of linear models for Bayesian optimal model selection. Consider a linear model $Y_i = X_{i1}\beta_1 + \ldots + X_{iP}\beta_P + \varepsilon_i$, where $X_{ip}$ are $p = 1, \ldots, P$ covariates and $\varepsilon_i$ are normally distributed errors. Let $\beta_p$ be the coefficient on predictor $p$, and $\beta_p = \pi_p B_p$, where $\pi_p$ is a binary latent variable, and $B_p$ continuous. In other words, the $p$th covariate $X_p$ is included in the model only if $\pi_p = 1$.

The stochastic search model as described in George and McCulloch [1993] calculates the
posterior mean of a set of coefficients $\beta$ via Gibbs sampling, using the following full conditionals:

$$
p(\beta|\sigma, \pi, Y) \tag{1.1}
$$

$$
p(\sigma|\beta, \pi, Y) = p(\sigma|\beta, Y) \tag{1.2}
$$

$$
p(\pi_p|\beta, \pi(p), Y) = p(\pi_p|\beta, \pi(p)) \tag{1.3}
$$

The posterior mean for $\pi_p$ represents the posterior probability of the $p$th covariate being included in the model, as well as the joint posterior probability of any subset of $1, \ldots, p$. The sampling is straightforward since the linear model and the normally distributed errors ensures full conjugacy. When the model has a more complex structure, the mixture of normals formulation for the SSVS method can be more useful. In this case, the marginal distribution for $\beta_p$ is written in the following way:

$$
p(\beta_p) = (\pi_p)N(0, \sigma^2) + (1 - \pi_p)N(0, \sigma^2/C), \tag{1.4}
$$

where $\sigma^2$ controls the variability around the means of the mixed distribution and $C$ represents the ratio of variance for coefficients included and not included in the model. In Chapter 2 we propose to use SSVS for spatial-temporally varying coefficients using a copula to model its spatial-temporal structure.

**Bayesian Model Averaging (BMA)**

Bayesian model averaging is useful for model-building strategies that take account of model uncertainty, as it provides an alternative to variable selection methods. BMA was first proposed by Raftery et al. [1997] and can account for model uncertainty by taking a weighted average of models over a given model space.

Suppose $M$ is our model space, comprising a total of $L$ model structures. Each model $M_\ell$ is associated with a likelihood function $L(Y|\theta_\ell, M_\ell)$ where $Y$ is the observed data and $\theta_\ell$ are the parameters associated to model $M_\ell$. We denote a set of prior probability densities for parameters
\( \theta_\ell \) as \( p(\theta_\ell | M_\ell) \) in a general form, and assume it includes all priors for hyperparameters. The posterior distribution given a model is:

\[
p(\theta_\ell | Y, M_\ell) = \frac{L(Y|\theta_\ell, M_\ell)p(\theta_\ell | M_\ell)}{\int L(Y|\theta_\ell, M_\ell)p(\theta_\ell | M_\ell) d\theta_\ell}
\]  

(1.5)

where the integral in the denominator represents the marginal distribution of the dataset over all parameter values specified in model \( M_\ell \).

Now consider a coefficient \( \beta \). It follows that its marginal posterior distribution across all models is given by

\[
p(\beta | Y) = \sum_{\ell=1}^{L} p(\beta | M_\ell, Y) p(M_\ell | Y)
\]  

(1.6)

where

\[
p(M_\ell | Y) = \frac{p(Y | M_\ell) p(M_\ell)}{\sum_{q=1}^{L} p(Y | M_q) p(M_q)}
\]  

(1.7)

is the posterior model probability given the observed data. Here, \( p(Y | M_\ell) \) is the likelihood of \( Y \) given model \( \ell \) and \( p(M_\ell) \) is a prior for model \( \ell \). This proportion is interpreted as a weight that compares the goodness of fit for model \( \ell \) with the goodness of fit of the all the other models in our model space, given the observed data.

There is a relationship between the posterior model probability and the Bayes factor. The Bayes factor for comparing \( M_1 \) versus \( M_2 \) is defined as:

\[
B_{1,2} = \frac{p(M_1 | Y)}{p(M_2 | Y)}.
\]  

(1.8)

This equation can also be written as a wider comparison between model \( M_1 \) versus all other models, in the following way:

\[
p(M_\ell | Y) = \frac{B_{M_\ell} p(M_\ell)}{\sum_{q=1}^{L} B_{M_q} p(M_q)}
\]  

(1.9)

showing that we can estimate the posterior model probabilities by using estimates for Bayes factors.
Factors and vice versa. The Laplace method (Hoeting et al. [2000]) helps approximate \( p(Y|M_\ell) \) and yields BIC approximation under regularity conditions. In Chapter 3, we use the BIC approximation to calculate the BMA weights \( p(M_\ell|Y) \) and propose a spatially varying variable selection procedure for a statistical emulator that takes into account the spatial correlation in the response and the covariates.
Chapter 2

Multivariate Spatial-Temporal Variable Selection with Applications to Seasonal Tropical Cyclone Modeling

2.1 Introduction

A tropical cyclone is a rotating system characterized by a low-pressure center, strong winds, and heavy rain that can cause severe damage both out at sea and inland. Because of the numerous deaths and damages caused in the past (Xie [2005]), it is crucial to develop methods to model and predict the locations and strengths\(^1\) of storms. According to Keith and Xie [2009], a hurricane is an environmental heat engine driven by sensible and latent heating from the ocean. Measures of sea surface temperature (SST) and latent heat flux (LHF) values in certain areas and during some months of the year are correlated with cyclone activity (Gray et al. [1994], Xie

\(^1\)Throughout this dissertation we use the word strength to refer to the classification of the storm according to its strength. We use the word intensity to refer to the \(\lambda\) parameter in the Poisson hurdle model.
and Liu [2014] and Xie [2005]), and thus should be important variables to describe the number of tropical cyclones (TC) in space and time.

Numerous studies have described the relationship between SST and the number of TCs per season, with independent models per subregions in the North Atlantic Basin, delimited by $10^\circ N$ to $62^\circ N$ latitude and $10^\circ W$ to $110^\circ W$ longitude; see Figure 2.1. Lehmiller et al. [1997], Blake and Gray [2004], Xie [2005], Elsner and Jagger [2006], Keith and Xie [2009], Werner and Holbrook [2011] and Xie et al. [2014] use climate indices as covariates in models that predict the number of hurricanes in each of the following subregions: Caribbean Ocean, Gulf of Mexico and the rest of the North Atlantic Basin. While some of the climate indices are summaries of SST from specific geographical locations and months of the year, none of the models in the literature study all possible combinations of months and locations.

Previous investigations using spatial models have determined that it is possible to describe the spatial variation of hurricanes using a finer scale than the regional analyses discussed above. Elsner et al. [2000] show that the location of hurricanes on a grid can be modeled using spatial models and climate indices. Jagger et al. [2002] develop a space time autoregressive model and Hodges et al. [2014] use a Bayesian hierarchical spatial model, with total hurricanes per location as a response and climate indices per location as covariates. Jagger et al. [2002] investigate forecasting skill for the model when it predicts hurricane activity for each year in each location, but pointed out that their estimators failed to converge using a grid size smaller than 5 by 5 degrees.

One important feature of the TC data on a grid is that the number of storms in a given year is likely to be zero in most of the locations in the Atlantic. Since the probability of an occurrence is directly related to the expected number of occurrences in a year, we propose a model that appropriately models both parameters and their association. Neelon et al. [2013] propose the Poisson Hurdle Model (PHM) for such cases, as they warn that failing to account for this dependence in zero-inflated models may produce biased parameter estimates.

In this chapter, we model the spatial-temporal variation of the yearly number of three
strengths of TCs in the Atlantic Basin, and estimate their respective multivariate spatial-temporal association with SST and LHF. Simultaneously modeling the multivariate spatial-temporal process improves our ability to find features in the space-time covariates that are significantly associated with TC activity in specific locations. To this end, we propose a model to address the following challenges: (1) a multivariate response that is potentially zero inflated, (2) an association between response and covariates that varies over space and time, (3) presence of correlation among coefficients for different responses, locations and times. We use empirical orthogonal functions (EOFs) to create fields of SST and LHF features, and use their scores as covariates. We show that the multivariate spatial-temporal variation of TCs can be explained by spatial-temporal features of SST and LHF, using high spatial resolution and seasonal patterns in addition to the annual signal. We use a Bayesian PHM with spatial-temporal dependence, and add a prior that can facilitate the selection of areas with no association, taking into account other correlations among coefficients.

Often the effect of covariates on the response is assumed to be constant across locations and time. However, this assumption is inappropriate in our case as TC association with SST and LHF features varies with strength, location and time within a season (Hodges et al. [2014] and Xie and Liu [2014]). Thus, we use spatial-temporally varying coefficients (Gelfand et al. [2003b]) that can relax this assumption. Moreover, our proposed method applies variable selection techniques in order to identify areas with significant association in the spatial-temporal domain.

Many variable selection methods have been proposed for Bayesian models; for example, stochastic search variable selection (SSVS) from George and McCulloch [1993]. Smith and Fahrmeir [2007] and Chipman [1996] propose variations from SSVS, for single predictors and Scheel et al. [2013] propose to assume the coefficients of multiple predictors of interest as a product of a binary spatially dependent field and an independent Gaussian field, to perform spatial variable selection. Reich et al. [2010], Lum [2012] and Cai et al. [2013] describe SSVS specifically for spatially and spatial-temporally varying coefficients in Bayesian models. Re-
ich et al. [2010] use the spike and slab prior from Ishwaran and Rao [2005] and describe a computationally efficient way to estimate the model. Cai et al. [2013] propose a model that performs variable selection for spatial-temporally varying coefficients using a Dirichlet process prior and nonparametric techniques. Depending on the mixed distributions, the interpretation of these coefficients becomes very complicated and almost impossible. Boehm Vock et al. [2015] propose to use copulas to retain interpretability of the coefficient. We propose a new version of a copula prior that can create a continuous smooth prior surface and perform multivariate spatial-temporal variable selection.

The novelty in our method is a dynamic variable selection procedure, which shares information between responses, locations, time and levels in the PHM context. This skill is essential in cases where we have a limited number of variables available, with a strong dependence structure, as in TC forecast models, but also in many other areas. For example, in epidemiology studies, where one is interested in detecting location and time where covariates have significant effects on a zero inflated response; or in forestry or ecology where models for sparse associations of covariates and rare events in space and time are also common.

This chapter is organized as follows. The second section describes the data used in the application. A third section introduces the statistical model and its estimation details. The fourth section describes a simulation study that investigates the performance of the variable selection technique, and the fifth section presents the results from real data application of the model. Finally, the last section presents our discussion and conclusions.

2.2 Data

Our main source of data is HURDAT (HURicane DATa), a publicly available database that contains detailed information about all recorded TCs in the Atlantic Basin since 1851. We construct a grid cell count using the 6 hour track data from 1950 to 2013 presented in Figure 2.1. The polygon grids are created using rectangles of 2.5 by 2.5 degrees over the Atlantic Basin, and using the coordinates from other variables used in the model as centroids.
In meteorology, the 5 category Saffir-Simpson scale is used to describe hurricanes. In practice, forecasters report tropical cyclones in three strength groups according to said scale, which are:

- **Low**: Tropical storms. TCs with sustained winds of less than 33 m/s.

- **Mid**: Hurricanes of category 1 and 2. TCs with sustained winds between 33 and 49 m/s.

- **Strong**: Major hurricanes of category 3, 4 and 5. TCs with sustained winds of more than 50 m/s.

We use this classification to count the number of unique storms per strength and year in each grid box. These counts are the response variable in our model.

SST and LHF are used as covariates in our model. For SST, we use data from the High-resolution Blended Analysis (Kalnay *et al.* [1996]), which takes measurements from NOAA’s Advanced Very High Resolution Radiometer (AVHRR). SST is reported at a resolution of 2.5
LHF is defined as the flux of heat in watts per square meter (W/m²) from the Earth’s surface to the atmosphere, which is associated with evaporation of water at the surface and subsequent condensation of water vapor in the troposphere. We use data from the National Centers for Environmental Prediction (NCEP) Reanalysis (Kalnay et al. [1996]), with a 2.5 by 2.5 degree resolution. Both SST and LHF have daily observations available for the Atlantic Basin. We average the daily data over three-month periods (trimesters) that coincide with the definition of season from World Ocean Database (WOD) (Levitus (1983)), i.e. winter starting in January, spring in April, summer in July and fall in October. The relationship between trimesters $w$ and the tropical cyclone season is presented in Figure 2.2. We define $Z_\ell(s_i, t, w)$ as the average of covariate $\ell$, in trimester $w$, location $s_i$ and year $t$, and then anomalies $X_\ell(s_i, t, w)$ as departures from the temporal mean per trimester and location ($Z_\ell(s_i, w) = \frac{1}{T} \sum_t Z_\ell(s_i, t, w)$) in each variable. Anomalies for some selected years are presented in Figure 2.3.

Our data consist of observations for $T = 64$ years, $M = 4$ trimesters per year, $N = 400$ locations, $K = 3$ response variables (number of tropical cyclones of low, mid and strong strength), and $L = 2$ covariates: SST and LHF. Figure 2.3 presents maps of selected years as examples of the spatial locations in the Atlantic.
Figure 2.3: Anomalies in the North Atlantic Basin during the first trimester for SST (left column) and LHF (middle column). Counts per grid box of tropical storms (right column). Years 1955, 1980, and 2013 randomly selected to illustrate data.


2.3 Statistical Methods

In this section we describe our model for yearly counts of TCs per location and strength. The model simultaneously associates the different count distributions to covariates' features derived from anomalies.

2.3.1 Pre-processing

Since the association between SST and LHF and tropical cyclones is not necessarily limited to an association between measures in the same location, we seek multivariate spatial-temporal features to be used as covariates in an interpretative way, while still providing a practical fit to the data. We use Empirical Orthogonal Functions (EOFs) to describe said features. According to Lorenz [1956], EOFs are spatial components that display space-time modes of variability of a quantity over a region. Thus, each covariate anomaly, for each trimester $w$, $X_\ell(w) = X_\ell(\cdot, \cdot, w)$, where $X_\ell(\cdot, \cdot, w)$ is a vectorized matrix, is approximated by the principal component (PC) scores, calculated using singular value decomposition (SVD) of $X_\ell(w)$, for each $w$ and $\ell$. In this way, we can represent the anomalies as:

$$X_\ell(s_i, t, w) = \sum_{r=1}^{R} \xi_{\ell,r}(t, w) \phi_r(w, s_i) + \epsilon(s_i, t, w),$$

(2.1)

where $R$ is the number of scores used to explain more than 70% of its variation. EOFs $\phi_r(w, s_i)$ are defined for each trimester $w$, $\xi_{\ell,r}(t, w)$ are the correspondent PC scores for each covariate $\ell$, and trimester $w$, and $\epsilon(s_i, t, w)$ is the variation orthogonal to the first $R$ EOFs.

2.3.2 Model

Let $y_k(s_i, t)$ be the number of unique TCs in year $t$ of strength $k$ in location $s_i$. We model $y_k(s_i, t)$ using a hierarchical Poisson hurdle model (PHM), where the probability mass function
is defined as:

\[
P(y_k(s_i, t) = m|p_k(s_i, t), \lambda_k(s_i, t)) = \begin{cases} 
1 - p_k(s_i, t) & \text{if } m = 0 \\
p_k(s_i, t) \frac{\lambda_k(s_i, t)^m}{m! (\exp^{\lambda_k(s_i, t)} - 1)^m!} & \text{if } m > 0
\end{cases}
\] (2.2)

where \( p_k(s_i, t) \) is the probability of non-zero counts and \( \lambda_k(s_i, t) \) is the intensity parameter. Both \( p_k(s_i, t) \) and \( \lambda_k(s_i, t) \) are described as a logit and loglinear function of multivariate spatial-temporally varying coefficients of the covariates’ scores, respectively. Thus we have:

\[
\text{logit}(p_k(s_i, t)) = \sum_{\ell=1}^L \sum_{w=1}^M \sum_{r=1}^R \beta_{1,k,\ell,r}(s_i, w) \xi_{\ell,r}(t, w)
\]

\[
\log(\lambda_k(s_i, t)) = \sum_{\ell=1}^L \sum_{w=1}^M \sum_{r=1}^R \beta_{2,k,\ell,r}(s_i, w) \xi_{\ell,r}(t, w).
\] (2.3)

Since our model coefficients vary in space by trimesters and by strength they are able to explain the association between the space-time modes of variability in the covariates (LHF and SST) and both the probability of non-zero counts and the intensity, for every strength. This association is expected to be close to zero in many spatial locations and during some trimesters, depending on the TC strength. To this end, we propose a method that can identify where, when, and for which strength this association is different from zero. Furthermore, we present a prior for \( \beta \) that selects such cases and uses the correlation between responses, locations, time and levels in the PHM context to improve this estimation.

### 2.3.3 Multivariate Spatial-Temporal Prior

Let \( A \) be a combination of covariate \( \ell \), PC score \( r \), response \( k \) and level of PHM \( j \). Each coefficient \( \beta_A(s_i, w) \) is modeled using a Gaussian copula and a latent variable \( \theta_A(s_i, w) \). We define \( \theta_A(s_i, w) \sim \text{MVN}(0, \Sigma) \) and use \( \Sigma \) with a separable covariance structure with the following form:

\[
\Sigma_{\theta} = \Gamma_s \otimes \Gamma_w \otimes \Gamma_k \otimes \Gamma_j.
\] (2.4)
The term separable means that we can factor the covariance structure into purely spatial \((s)\), temporal \((w)\), multivariate \((k)\), and by PHM level \((j)\) covariance structures, which allows for computationally efficient estimation and inference. Each covariance \(\Gamma\) can be written as a correlation function \(C(h)\) where \(h\) is the Euclidean distance when the distance is defined in space \((s)\) or in time \((w)\), or the numerical absolute difference when the function is in terms of the ordered response strengths \((k = 1, 2, 3)\) or PHM levels \((j = 1, 2)\).

Then, we define \(\beta_A(s, w) = F_A^{-1}\{\Phi[\theta_A(s, w)]\}\) where \(\Phi\) is the cdf of a standard normal distribution and \(F_A\) is the marginal cdf of each \(\beta_A(s, w)\), distributed as a multivariate spike and slab mixture of normal densities:

\[
\beta_A \sim \pi_A N(\alpha_A, \sigma_A^2) + (1 - \pi_A)N(0, \sigma_A^2/C),
\]

where \(\pi_A\) represents the expected proportion of coefficients that are centered at \(\alpha_A\). In this way, coefficients are smoothed around \(\alpha_A\) or 0. \(\sigma_A^2\) controls variability around the means of the mixed distribution, and \(C\) represents the ratio of variance for coefficients centered in zero to centered in \(\alpha_A\).

A copula can flexibly model the marginal distributions while introducing dependence via multivariate normal distributions, creating a continuous smooth prior surface. The intuition in our case is that if a coefficient is close to zero in a specific location and time, nearby locations (in space and time) are unlikely to have large coefficients. Also, coefficients for adjacent categories from an ordered response and the two levels of a PHM are expected to be correlated, since the probability of having zero counts in a specific location and time is likely to be associated with its intensity.

For \(\pi_A\), we use a flat prior between 0 and 1. Other prior choices can give more weight to options like \(\pi_A = 0\): all locations and trimesters have a coefficient centered in zero for \(A\), or the opposite \(\pi_A = 1\). For \(\alpha_A\), the prior is a normal distribution with standard deviation 1, which allows the coefficients to be centered in values ranged from \([-4, 4]\). Previous work on mixed
models similar to Equation 2.5 has suggested that model fit can be sensitive to large values of $\sigma_A$ because it would be difficult to differentiate between coefficients centered at 0 from other values (Ishwaran and Rao [2005]). We use a Gamma(0.1,0.1) prior to encourage small values for $\sigma_A$.

We model the temporal and spatial covariance components using AR1 and exponential structures, respectively, while using a vague Wishart prior to model the multivariate covariance and the covariance between PHM levels. Hyperpriors for the spatial range parameter $r$, and the temporal correlation and variance $\rho_t$ and $\sigma_t$ should be informative, in the sense of restricting the range of possible parameter values to values which make sense in terms of the application.

To estimate the parameters in the model, we use a Markov chain Monte Carlo (MCMC) algorithm. Since we are using a mixed distribution as a prior for $\beta_A(s_i, w)$, we do not have conjugacy, and we need to use rejection sampling. The scheme requires factoring $24N \times 24N$ matrices that we simplify by using the Woodbury matrix identity on the Kronecker structure of $\Sigma_\theta$. In the simulation study in Section 2.4 we generate 15000 iterations, discard the first 5000 as burn-in, and thin the chains by keeping every fifth iteration.

### 2.4 Simulation Study

We study the performance of our model compared to two other competing models for variable selection. Our main focus is to describe the dynamic variable selection performance; thus we compare performance for varying degrees of spatial-temporal correlation in the coefficients $\beta_A(s_i, w)$. For that, we use the following statistics: median absolute deviation (MAD), ability to correctly identify null coefficients in space and time, and MSE for each coefficient estimated.

We generate 5 $\beta_A(s_i, w)$ under the following settings:

- **Data Setting 1**: Weak Temporal Correlation and Strong Spatial Correlation. Smooth surfaces are generated using linear combinations of latitude and longitude.

- **Data Setting 2**: Weak Spatial-Temporal Correlation. The true coefficients for each
location and trimester are exponentially spatially correlated with range = 2 km and temporally correlated with $\rho = 0.1$ for all $A$.

- **Data Setting 3:** Strong Spatial-Temporal Correlation. The true coefficients for each location and trimester are exponentially spatially correlated with range = 100 km and temporally correlated with $\rho = 0.9$ for all $A$.

Setting 1 is composed of smooth surfaces created using linear combinations of latitude, longitude, and $\alpha_A = \{2, 1.5, 1, 0.5, 0\}$, respectively. Figure 1 in A.5 shows an example realization. In this case $\beta_A(s_i, w)$ does not follow the same distribution as in any of our models. Settings 2 and 3 are generated using the copula transformation described in Section 2.3.3 and assuming 2.5 true, with $\alpha_A = \{2, 1.5, 1, 0.5, 0\}$, $\pi = \{1, 0.8, 0.5, 0.2, 0\}$, $\sigma^2_A = 1$ and $C = \infty$. They differ in the degree of spatial-temporal correlation. For each $A$ and each setting, we generate $M = 4$ time points and $N = 100$ locations.

Samples of $\xi_A(t, w) \sim N(0, 2^{-1/2})$ for $w = 1, 2, .., 4$ and $t = 1, 2, ..., 30$, with $A = 1, 2, .., 5$ are generated to simulate scores. We apply Gram-Schmidt orthonormalization to make them independent among $A$, i.e. $\xi_A(t, w) \perp \xi_A(t', w')$. Then, we use $\beta_A(s_i, w)$ from settings 1-3, to have $\beta^T \xi$, and use it to generate random samples of $Y(s_i, t)$ from a Poisson distribution: $\text{Pois}(\exp(\beta^T \xi))$. We generate $B = 50$ data sets.

We test the performance of three models under three different data settings. Since we concentrate on showing the advantages of sharing information between locations and times when performing variable selection, we simplify the response model to be $y(s_i, t) \sim \text{Pois}(\lambda(s_i, t))$, and use the second Equation in 2.3 to model the intensity parameter $\lambda(s_i, t)$ as a function of the scores $\xi_{\ell, r}(w, t)$. Here, $A$ represents scores for each covariate (component) that is used in our model. The three models we consider are our model (**Model 3**) and two other competing models for variable selection. They differ in the latent variable distribution used in the transformation for $\theta_A(s_i, w) \sim \text{MVN}(0, \Sigma_\theta)$:

- **Assumed Model 1:** $\beta_A(s_i, w)$ are independent across sites and times and follow the
distribution in 2.5. It does not share information, since \( \Sigma_\theta = \sigma^2_t \times I \) is just a diagonal matrix.

- **Assumed Model 2**: Spatial dependence in \( \beta_A(s_i, w) \) is induced with a Gaussian Copula. It shares information across sites, using \( \Sigma_\theta = S \otimes \sigma^2_t \times I \) where \( S \) represents an exponential spatial correlation matrix with range \( r \).

- **Assumed Model 3**: Spatial-temporal dependence in \( \beta_A(s_i, w) \) is induced using a Gaussian Copula. It shares information across sites and trimesters, using \( \Sigma_\theta = S \otimes T \), where \( S \) is defined as in Model 2 and \( T \) has an autoregressive 1 structure (AR1) with parameters \( \rho_t \) and \( \sigma^2_t \) for time.

In all cases, we use the same priors: \( \alpha_A \sim N(0, 2) \), \( \pi_A \sim \text{Beta}(1, 1) \), and \( \sigma^2_A \sim \text{InvGamma}(0.1, 0.1) \). We include Model 1 as a baseline to show the overall improvements in the estimation of \( \beta_A(s_i, w) \) when considering the spatial and the spatial-temporal structure that models 2 and 3 provide.

We compute posterior medians and credible intervals for each \( \beta_A(s_i, w) \) and posterior means for \( \pi_A \) and \( \alpha_A \). Median absolute deviation (MAD) is averaged over space, trimesters and iterations for each \( \beta_A \). Percentage of correctly estimated zeroes are calculated only for \( \beta_A \) and averaged over space, trimesters and iterations as well. MSE is also calculated and averaged over iterations for \( \alpha_A \) and \( \pi_A \).

MAD is computed in the following way, for each setting and model:

\[
\text{MAD}_A = \text{Median}_{BM} |\hat{\beta}_A(s_i, w) - \beta_A(s_i, w)|
\]

where \( B = 50 \) data sets, \( N = 100 \) locations and \( M = 4 \) time points.

Table 2.2 presents MAD for each \( \beta_A(s_i, w) \). Values were compared using credible intervals. We can see how the median absolute deviations for Model 3 are statistically smaller in most of the settings generated with strong correlation, as expected. In some cases, this improvement is reflected in a 40% decrease of MAD. Nevertheless, this difference is not reflected in cases where we have weak correlation and \( \alpha \) is generated as far from zero (\( \beta_1 \) and \( \beta_2 \)). When comparing
Table 2.1: Mean acceptance rate for all parameters.

<table>
<thead>
<tr>
<th>Weak T Correlation</th>
<th>Weak ST Correlation</th>
<th>Strong ST Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>M1</td>
<td>M2</td>
</tr>
<tr>
<td>α</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>π</td>
<td>0.31</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 2.2: Median MAD per $\beta_A$ over locations and trimesters. All standard errors are less than 0.006. Proportion of correctly estimated $\beta_A = 0$ All standard errors are less than 0.005. NA in cases where we did not simulate zeroes.

<table>
<thead>
<tr>
<th>Weak T Correlation</th>
<th>Weak ST Correlation</th>
<th>Strong ST Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>M1</td>
<td>M2</td>
</tr>
<tr>
<td>MAD</td>
<td>0.3073</td>
<td>0.4018</td>
</tr>
<tr>
<td>β₁</td>
<td>0.8121</td>
<td>0.7437</td>
</tr>
<tr>
<td>β₂</td>
<td>0.8457</td>
<td>0.7594</td>
</tr>
<tr>
<td>β₃</td>
<td>0.7434</td>
<td>0.8369</td>
</tr>
<tr>
<td>β₄</td>
<td>0.7184</td>
<td>0.7346</td>
</tr>
<tr>
<td>β₅</td>
<td>0.8941</td>
<td>0.8541</td>
</tr>
</tbody>
</table>

results for data generated with weak temporal correlation, MAD for our model is smaller than models 2 and 3 in all cases. MAD is statistically smaller using Model 3 in cases where the real $\alpha$ is equal or close to zero ($\beta₃$, $\beta₄$, and $\beta₅$).

As a way to evaluate the performance of our dynamic variable selection method, we calculate the proportion of correctly estimated $\beta_A = 0$. Our model detects a statistically higher proportion of real values equal to zero in all cases when the data is generated with weak and strong correlation. On the other hand, our model performs similarly to Model 2 in cases where we have weak temporal correlation or when $\alpha$ is far away from zero. Simulation results show an average increase of 13%, 38%, and 30% in the percentage of correctly detected zeroes, for each of the settings. The best case scenario of performance for our dynamic variable selection procedure is when data presents strong spatial-temporal correlation, and it is close to zero.

Intuitively, MAD result reflects the fact that our model (M3) benefits from using temporal
correlation to get a better fit when there is strong spatial-temporal correlation present, compared to the fit to weak spatial-temporal or temporal dependent data. The same benefit is reflected in the percentage of correctly detected zeroes, where sharing information in space and time gives the model more tools to correctly detect zeroes. Therefore, when there is strong spatial-temporal correlation present in the data, the simulation shows the need of a model that accounts for both space and time dependence.

Overall, our model shows an average increase of 13%, 38%, and 30% in the percentage of correctly detected zeroes for weak temporal, weakly and strongly correlated data, respectively. Also, on average our model has a MAD per $\beta_A$ 38% lower than the competing models for strongly correlated data, 24% lower for weakly correlated data, and 16% lower with weak temporal correlation.

2.5 Analysis of Tropical Cyclones in the Atlantic Basin

2.5.1 Model Comparison

We fit the full model (ST) from Section 2.3, and two other models (Spatial and Independent) to the tropical cyclone data described in Section 2.2. The difference between the full model (ST), spatial (SP) and independent (IN) model is the specification of $\Sigma_\theta$ in 2.4. In the SP model we use $\Sigma_\theta = \Gamma_s \otimes I_w \otimes I_k \otimes I_j$, and in the IN model we assume independence over all levels ($\Sigma_\theta = I_s \otimes I_w \otimes I_k \otimes I_j$). We perform the comparison in order to collect evidence of the need of the full model (or the lack thereof).

For each trimester $w$, we calculate anomalies $X_\ell(s_i, t, w)$ for both SST and LHF and then perform a SVD as described in Section 2.3. We obtain $R = 4$ EOFs with their respective scores $\xi_{\ell,r}(t, w)$ for each covariate $\ell$ and trimester $w$, to have 16 time series in total. The maps of the four resulting EOFs for each covariate and trimester are presented in the Appendix A. The obtained EOFs for each of the seasons and covariates together describe at least 70% of the variability. The estimated scores range from $-6$ to 6 in all cases and are centered close to zero.
Table 2.3: Goodness of fit criterion DIC, and MSE overall, per location and per year for the full model (ST), spatial (SP) and independent (IN). DIC bar values (standard error in parenthesis) for models using different grid box sizes, compared to a model using climate indices.

<table>
<thead>
<tr>
<th></th>
<th>DIC</th>
<th>TS</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>HU</td>
</tr>
<tr>
<td>Independent (IN)</td>
<td>183.1 (1.2)</td>
<td>3.16 (0.3)</td>
<td>2.51 (0.2)</td>
</tr>
<tr>
<td>Spatial (SP)</td>
<td>142.9 (1.1)</td>
<td>0.20 (0.01)</td>
<td>0.18 (0.02)</td>
</tr>
<tr>
<td>Spatial Temporal (ST)</td>
<td>136.2 (4.2)</td>
<td>0.13 (0.01)</td>
<td>0.12 (0.001)</td>
</tr>
<tr>
<td>200 grid boxes</td>
<td>136.9 (4.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400 grid boxes</td>
<td>136.2 (4.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>800 grid boxes</td>
<td>136.5 (4.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate indices</td>
<td>173.6 (5.8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the results subsection we describe these spatial patterns as features and relate them to the response with the statistically significant factors.

We compare models using deviance information criterion (DIC) and mean squared error (MSE), to evaluate goodness of fit of the model, but also their ability to estimate correctly the number of storms per location and per year. We use the MSE to measure the mean square distance between the expected posterior values and the observed values. The percentage of locations with no TCs in a given year in the Atlantic Basin ranges from 78% to 98% for low strength, from 83% to 99% for mid strength and from 89% to 100% for strong strength and the range of counts per location is from 0 to 4 storms. Table 2.3 presents all the goodness of fit measures for each model.

The full model (ST) is selected as the best model by DIC and MSE values for every response. We compare the observed with the expected values of the posterior distribution and obtain that the full model performance in terms of MSE per response is superior to the IN and SP models performance for all responses. Figures A.2 and A.1 in the Appendix show the MSE per location and year, respectively for the full model. Our model has a fairly superior performance per year compared to climatology (MSE per response is 0.14 for TC, 0.16 for HU and 0.03 for MH), with some exceptions like 1966, 1997 and 2005 for HU. In addition, the full model presents MSE
lower than 1 unit in most of the locations, with a unique exception in front of the African Coast, very close to the mid TC main development region. However, prediction is not the only focus of our model. We also use the model to identify potentially scientifically relevant associations between covariates and responses.

2.5.2 Sensitivity Study

We consider several choices of hyperpriors when fitting the full model. Also, we estimate the scores using three different rectangle sizes as grid boxes in space for SST and LHF measures. We evaluate the performance of the different modeling options using the Deviance Information Criterion (DIC).

We specify different hyperpriors for the spatial range parameter $r$, and the temporal correlation and variance $\rho_t$ and $\sigma_t$. Results were robust in terms of DIC when changing priors for $\rho_t$ and $\sigma_t$ from informative to non informative. For this analysis we use a uniform prior for $\rho_t$ and a Gamma(0.1,0.1) for $\sigma_t$. We observe that DIC is sensitive to a change in $r$; goodness of fit improves significantly when using a uniform distribution from 0 to 300 km, instead of 30 km or 3000 km as upper limits. This improvement could be explained by the typical ratio of a tropical cyclone eye, which approximately ranges from 3 to 300 km (NHC [2016]).

In addition to these hyperprior choices, we conduct a specific sensitivity analysis to compare the goodness of fit when using different rectangle grid box sizes to construct the SST and LHF scores, and at the same time compare it with a model that uses climate indices instead. The use of climate indices as base comparison is particularly interesting in this case, since those are the covariates that are typically used in TC forecast models (Xie et al. [2014]). DIC was nearly identical among the cases of using 200, 400 and 800 grid boxes to construct the SST and LHF scores, but it was higher in the model that uses climate indices instead of scores (Table 2.3). We set the rectangle size to be the same as the response (2.5 by 2.5 degrees) and use 400 grid boxes to calculate the SST and LHF scores. The first two scores per trimester are presented in Appendix A.
Table 2.4: Posterior estimates for each response for the overall mean $\alpha_A$, the proportion $\pi_A$ of coefficients centered in $\alpha_A \neq 0$ and the variability $\sigma_A$ of coefficients centered in $\alpha_A$, for responses $Y_1 =$ Tropical Storms, $Y_2 =$ Hurricanes and $Y_3 =$ Major Hurricanes. Values with (*) have credible intervals that do not include zero.

| $\alpha_A, r$ | SST, $p$ | LHF, $p$ | SST, $\lambda$ | LHF, $\lambda$ | $\pi_A, r$ | SST, $p$ | LHF, $p$ | SST, $\lambda$ | LHF, $\lambda$ | $\sigma_A, r$ | SST, $p$ | LHF, $p$ | SST, $\lambda$ | LHF, $\lambda$ |
|---------------|----------|----------|----------------|----------------|----------|----------|----------|----------------|----------------|----------|----------|----------------|----------------|
| $1$           | -0.14 (*) | 0.56     | 0.13           | 0.18 (*)       | 0.01     | -0.02    | 0.14     | -0.29         | 0.59 (*)       | 0.09     | 0.46 (*) | 0.02           | 0.30 (*)       |
| $2$           | -0.49 (*) | -0.48 (*)| -0.71 (*)      | -0.21 (*)      | -0.40 (*)| -0.36    | 0.09     | -0.31         | 0.17           | -0.33 (*)| 0.04     | -0.39 (*)      | 0.04           |
| $3$           | -0.76 (*) | 2.99 (*)  | -0.59 (*)      | 0.24 (*)       | -0.31    | 0.08     | -0.16    | -0.50         | -0.38          | -0.48 (*)| -0.18    | -0.76          |                |
| $4$           | -0.04 (*) | -0.34 (*)| -0.03          | -0.41 (*)      | 0.26     | 0.01     | 0.13     | 0.15          | -0.07          | 0.00     | -0.07   | 0.09           |                |

In order to improve MCMC convergence, we tune $C$ to give acceptance rates of around 30% or higher for most parameters. Convergence is monitored using trace plots of deviance and several representative parameters. For all the sensitivity analyses and the final analysis, we generate 20,000 samples and discard the first 5,000 as burn-in and thin the chains by keeping every fifth iteration.

2.5.3 Results

Table 2.4 presents the posterior estimates for each response for the overall mean $\alpha_A$, the proportion $\pi_A$ of coefficients centered in $\alpha_A \neq 0$ and the variability $\sigma_A$ of coefficients centered in $\alpha_A$. The proportion of locations where estimated coefficients are significantly different from zero for SST and LHF features are presented in Table A.1; significant difference is determined by whether or not credible intervals include zero.

We define a significant factor as an EOF for which more than 10% of the corresponding spatially varying coefficients for a given trimester $w$ are statistically different from zero (posterior credible interval does not include zero). We describe here those factors that have a sum of magnitudes greater than 20 for intensity and greater than 10 for probability of occurrence. SST features are the only significant factors of coefficients during winter ($w = 1$). These coefficients
Figure 2.4: Winter: SST Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.

are the most important in terms of determining predictors to use in the seasonal forecast, because they can be measured before the TC season starts (Figure 2.2). For example, the model indicates that warmer anomalies in SST over the mid TC main development region (in front of the African Coast) and in the Arctic during the winter are positively correlated with more major hurricanes in the Caribbean Sea and with more hurricanes around Florida, Cuba and the Mid Atlantic. Furthermore, the model shows that negative anomalies in SST over the Western Atlantic present a negative correlation with the occurrences of tropical storms over the coast of Florida and Cuba (see Figure 2.4).

TC seasonal variation is well described using SST and LHF features during spring and summer ($w = 2, 3$). While these covariates cannot be used for seasonal predictions, identifying their relationship with TC occurrence may provide guidance for future scientific studies. During spring, we have a significant factor that describes a positive correlation between positive anomalies in SST over the Arctic and the number of major hurricanes in the Gulf of Mexico and
the Caribbean Sea. Also, Figure 2.5 shows three significant factors that relate positive anomalies of LHF over the tropical Atlantic, Caribbean and Gulf of Mexico as positively associated with both the occurrence of tropical storms and their number of occurrences over the Atlantic.

During summer (Figure A.3 in the Appendix), we find an opposite pattern for SST anomalies in the Arctic, since they are negatively correlated with the number of hurricanes and major hurricanes over the Gulf of Mexico, around Cuba and over the Mid Atlantic. Furthermore, positive anomalies of LHF in front of the US East Coast are positively correlated with the occurrence of hurricanes in the same area. Finally, positive LHF anomalies over the Gulf of Mexico and the tropical Atlantic are negatively correlated with the occurrence of major hurricanes around the lower Antilles and with the number of tropical storms over the Gulf of Mexico, the Caribbean and the Mid Atlantic. Lastly, during the fall (Figure A.4 in the Appendix) all the statistically significant factors are related to LHF.

2.6 Discussion

In this chapter, we develop a statistical model that allows for dynamic variable selection. The multivariate spatial temporal smooth surface we use in the prior reflects the fact that categories, trimesters, locations, and levels of PHM can affect the association of SST and LHF with the response.

A simulation study shows that our model outperforms the independent and spatial models providing an average increase of 27% in the percentage of correctly detected zeroes, in all treatments. Also, on average our model has a MAD per $\beta_A$ 26% lower than the competing models. In some cases, this improvement is reflected in a 40% decrease of MAD. In the data application, we find that using 200, 400 or 800 grid boxes to construct the SST and LHF scores has no effect on model goodness of fit (DIC), but each is better than a traditional climate index model (Table 2.3). Also, our model has a significantly lower MSE values than the competing models.

This study provides a better understanding of how observed TC variations in space and
Figure 2.5: Spring: SST or LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.
time are related to variations in features of SST and LHF. No previous study has attempted to relate SST and LHF features to the variation of storms, using spatial-temporal models and variable selection techniques. Our results show that warmer SST anomalies during winter and spring over the mid TC main development region are associated with more major hurricanes in the Caribbean Sea and the Gulf of Mexico. The results of this study can be used to improve seasonal forecast models by using the proposed measures as seasonal predictors.
Chapter 3

Bayesian Regional Climate Model Emulator with Spatially Varying Variable Selection

3.1 Introduction

General Circulation Models (GCM) describe the large scale global atmospheric and oceanic dynamics. Computational constraints require coarse resolutions, and so GCMs provide limited information about regional scales. Regional Climate Models (RCM) are dynamic downscaling models that use atmospheric interactions in the GCM to develop a higher resolution output, which is influenced by a smaller scale of topographical features (Yuqing et al. [2004]). Downscaling is used for interpolating low resolution to high resolution predictor variables, under the assumption that “relationships can be established between atmospheric processes occurring at disparate temporal and/or spatial scales” (Wilby and Wigley [1997]). RCMs are useful to describe regional climate as well as to serve as an input for weather prediction models. However, as Wilby et al. [2000] state “RCMs are computationally demanding and require orders of magnitude more computer time than statistical downscaling to compute equivalent scenarios”.
Since climate models are simulators of the physical climate system (O’Hagan [2006]), we define an RCM statistical emulator as a statistical downscaling model for GCMs.

The North American Regional Climate Change Assessment Program (NARCCAP) is an international program that has as a main goal to generate climate scenarios for use in impacts research. (Mearns et al. [2009]). They have proposed and applied a dynamical downscaling technique that consists in embedding RCMs within GCMs to obtain higher resolution evaluations of the climate model over the domain of interest, using different combinations of RCMs and GCMs. Since this dynamical downscaling technique is computationally expensive, a fast approximation to perform sensitivity analysis or for making informed prior judgements relating the GCM drivers for an RCM is of high value.

In this chapter, we propose a statistical emulator of regional climate model output variables driven by GCM output variables. Our goal is to be able to use the GCM output to emulate the RCM output accurately, and to determine which set of variables from the GCM is the best possible combination to describe the RCM, depending on the location. To this end, we develop a spatially varying Bayesian model for smooth Model Output Statistics (MOS), with data-dependent weights for each covariate that also varies over space. We address the following challenges: (1) an association between response and covariates that varies over space, and hence a need to perform spatially varying variable selection, (2) a spatially smooth response, (3) a heavy computational burden given that the number of observations is big in this case, and (4) a need for a relatively fast emulation of the downscaling process. We use radial basis functions to represent and efficiently approximate the spatially varying regression coefficients, the BIC approximation for the Bayes factors used to estimate the data-dependent weights for each covariate in each location, and Vecchia’s approach to approximate the likelihood in the posterior sampling steps.

Previous work on statistical emulators from Castruccio et al. [2014], Overstall and Woods [2016] and O’Hagan [2006] are concentrated on climate models projections and on computer experiment output, but to our knowledge, there is no previous work on regional climate model
statistical emulators. The novelty of this work is twofold: first, we develop a spatially varying variable selection procedure useful for smooth model output statistics (MOS) that uses approximations to make it computationally efficient. Second, in terms of the application, our model can be used as a fast approximation of the RCM to emulate new scenarios in the RCM locations, given new values of covariates, and also to evaluate the relative importance of each combination of GCM output variables to describe RCM output. We show how our variable selection approach improves the statistical emulator performance, since it takes into account the uncertainty of the model selection (i.e. which set of covariates are included) using model averaging weights per location, and also models the spatial process for each coefficient and the response variable, separately.

The remainder of this chapter is divided into five sections. The second section presents the NARCCAP data, and the selection of covariates from GCMs and RCMs. Next, a third section presents the statistical methods used in this chapter while the fourth section presents a numerical simulation to demonstrate its properties. The fifth section outlines the data analysis for the NARCCAP data and the last section closes with a discussion of the results.

3.2 Data

Our main source of data is NARCCAP (Mearns et al. [2007]). Our data set consists of a response variable from the Canadian Regional Climate Model (CRCM) (Caya et al. [1995]) and seven covariates from the Community Climate System Model (CCSM), which is the global model embedded in the CRCM. The complete data sets can be accessed through: https://www.earthsystemgrid.org/project/NARCCAP.html. The two climate models have different resolutions and projections, as presented in Figure 3.1.

The global model resolution ($256 \times 128$ grid boxes across the globe) was subsetted to create a grid box system covering North America and consisting of $83 \times 33$ grid boxes with the latitude and longitude of the global variable at the center. Figure 3.2 presents the contrast between the spatial resolution from the response variable (RCM output) and that from the covariates.
Figure 3.1: Coordinates for the Canadian Regional Climate Model (CRCM) and the Community Climate System Model (CCSM) in the North American Region. The area outlined in red is the selected domain for the analysis in Section 3.5.
Finally, we choose to work with the Western part of North America, given that the region has a diverse topography, in terms of vegetation, climate and elevation (as shown in Figure 3.3).

The response variable is surface precipitation from the CRCM Regional Model, measured in kgm$^2$/s. Precipitation is usually presented in terms of mm, but we choose keep the model units and also to transform the response variable using a log scale, to improve normal assumption. We use several variables from the CCSM global model as covariates. Three of them are surface variables: sea level pressure (PSL) measured in Pa, reference height temperature (TREFHT) measured in K, and atmospheric specific humidity (QBOT); we use vertical velocity (OMEGA) measured in Pa/s, that represents a mid altitude variable (measured at 510.45 hPa), and lastly, we use two variables measured in the jet level winds: wind east-west (U) and wind north-south (V) both measured in m/s at 267 hPa. We average the daily data over the Summer periods.
Table 3.1: Descriptive statistics for the covariates.

<table>
<thead>
<tr>
<th>Units</th>
<th>log(PR)</th>
<th>PSL</th>
<th>TREFHT</th>
<th>OMEGA (510.45)</th>
<th>U (266.48)</th>
<th>V (266.48)</th>
<th>Qbot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>-13.82</td>
<td>100.949</td>
<td>273</td>
<td>-0.07888</td>
<td>-4.51</td>
<td>-7.540</td>
<td>-14.796</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>-12.20</td>
<td>101.399</td>
<td>287</td>
<td>-0.00963</td>
<td>10.04</td>
<td>0.557</td>
<td>-3.972</td>
</tr>
<tr>
<td>Median</td>
<td>-11.56</td>
<td>101.551</td>
<td>291</td>
<td>0.00732</td>
<td>15.12</td>
<td>3.762</td>
<td>-0.704</td>
</tr>
<tr>
<td>Mean</td>
<td>-11.62</td>
<td>101.674</td>
<td>292</td>
<td>0.01255</td>
<td>14.55</td>
<td>3.718</td>
<td>-1.664</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>-10.94</td>
<td>101.882</td>
<td>296</td>
<td>0.03098</td>
<td>19.44</td>
<td>6.660</td>
<td>0.448</td>
</tr>
<tr>
<td>Max.</td>
<td>-8.91</td>
<td>103.057</td>
<td>305</td>
<td>0.11237</td>
<td>32.57</td>
<td>16.438</td>
<td>7.481</td>
</tr>
</tbody>
</table>

(Jun 1, 1970 through Aug 31, 1999) in the following way:

\[ X_{kt}(s_i) = \frac{1}{91} \sum_{d=d_1}^{d_2} X_{ktd}(s_i) \]

for each covariate \(k\) in location \(s_i\), where \(d_1 = 152\) and \(d_2 = 152 + 91 - 1\) are the first and last days of the Summer for each year \(t\), respectively. In the same way \(Y_t(s_i) = \sum_{d=d_1}^{d_2} Y_{td}(s_i)\) for the response variable.

Our data consist of observations for \(T = 30\) years, \(N = 5,200\) locations, and \(P = 6\) covariates: PSL, TREFHT, QBOT, OMEGA, U500, and V500.

### 3.3 Statistical Model

Let \(Y_t(s)\) be RCM model output for location \(s\) from the RCM resolution and year \(t\). We model \(Y_t(s)\) as:

\[ Y_t(s) = \mu_t(s) + \varepsilon_t(s), \quad (3.1) \]

where \(\varepsilon_t \sim GP(0, C(h, \Omega))\) and \(C(h, \Omega)\) is the spatial covariance function that describes the smooth spatial process using distance \(h\) and a set of parameters \(\Omega\). We assume independence.
Figure 3.3: Elevation in meters in the North American Region. This variable was used as a driver to generate CRCM output, and although we do not include it in our model, it is highly correlated with several of the covariates we use over the domain of interest.
in time, although it could be incorporated into the structure of $\varepsilon$. We write $\mu_t(s)$ as:

$$
\mu_t(s) = \sum_{k=0}^{P} X_{kt}(s) \beta_k(s), \text{ with }
$$

$$
\beta_k(s) = v_k(s) f_k(s) = v_k(s) \sum_{j=1}^{J} \alpha_{kj} F_j(s)
$$

where $X_{kt}(s)$ contains the GCM output data on predictor $k$ at time $t$ in each location from GCM grid box containing $s$ from the regional model (RCM), and $X_{0t}(s) = 1$. Note that $X_{kt}(s)$ has the same value in all the RCM locations $s$ that are included in the same GCM gridbox.

Let $f_k(s)$ be a basis representation, and define $\Phi_{ii'} = \phi(s_i - s_{i'})$ given matrix $\Phi$. We use spectral value decomposition to write $\Phi = UDU^T$, where $D$ is diagonal and $U$ contains the respective $J$ eigenfunctions ordered by eigenvalue magnitude and evaluated at each location $s_i$. Let $U_j$ be the $j$th eigenvector of $\Phi$ and $F_j(s_i)$ be the $i$th element of $U_j$. We use thin-plate splines basis functions to construct $\Phi$, where $\phi(|s_i - s_{i'}|) = |s_i - s_{i'}|^2 \log |s_i - s_{i'}|/\eta$, and $\eta$ is a scale parameter. The advantages of thin-plate splines have to do with its flexibility to represent a myriad of spatial patterns given its piecewise smooth nature, as an alternative to the Gaussian radial basis that are infinitely smooth (Buhmann [2003]).

The smooth spatial process $\log(v_k) \sim GP(a_k, \Sigma_v)$ is a data-dependent weight, that describes how likely it is for covariate $k$ to be included in the model. The vector $v_k$ of length $N$ has a multivariate log normal smooth prior centered in the observed mean of the log weights, and a covariance matrix $\Sigma_v$; while $\alpha$ is a vector of coefficients to estimate, where each has a prior distribution $\alpha_{kj} \sim N(0, \sigma_k^2 \lambda_j^2)$, where $\lambda_j$ is the $j$ eigenvalue of the matrix constructed using radial basis functions. In this case, $\lambda_j$ can help to regularize the number of eigenfunctions used, since the value is decreasing and it approaches to zero when $j$ increases.

### 3.3.1 Estimation

We assume $v_k$, $\lambda_j$ and $F$ as pre-defined data-dependent quantities, thus the parameters to estimate in our model are only $\alpha_{kj}$, $\sigma_k^2$, and the set of parameters $\Omega$. In order to estimate the
pre-defined data-based weights, we use ideas of Bayesian model averaging (BMA). BMA was first proposed by Raftery et al. [1997] to account for model uncertainty by taking a weighted average of estimations coming from different models over a given model space. In this context, we call each of the $2^p$ possible combination of covariates a model structure $M_\ell$ and write the marginal posterior probability for the coefficient $\beta$ in location $s$ for covariate $k$ as:

$$p(\beta|Y, X) = \sum_{\ell=1}^{L} p(\beta|M_\ell, Y, X)p(M_\ell|Y, X)$$  

(3.4)

where $p(\beta|M_\ell, Y, X)$ is the likelihood of $\beta$ given the model specification $M_\ell$ and the observed data $\{Y, X\}$, and $p(M_\ell|Y, X)$ is the model probability given the observed data, or in other words, a weight that reflects the model $M_\ell$ goodness-of-fit provided the observed data. Hoeting et al. [2000] show that there is a straight relationship between the posterior model probability and the Bayes factor, and Kass and Raftery [1995] proved that BIC can be used to approximate the same quantity via Bayes factors. In this work, we use the BIC to approximate the posterior model probability $\hat{p}(M_\ell|Y, X)$, and write the weight for variable $k$ being included in the model as:

$$w_k = \sum_{\ell \in \{i: \beta_k \neq 0 \text{ in } M_i\}} p(M_\ell|Y, X) = \frac{BF_{M_\ell}}{\sum_h (BF_{M_h})}$$  

(3.5)

where $BF_{M_\ell} = \exp(-0.5BIC_{M_\ell})$.

We use the BMA package (Raftery et al. [2015]) to estimate $w_k(s)$ in each location and after that, we smooth each of the $k$ fields applying a simple local smoothing to $w_k(s)$, where we average the 5 closest observations (neighborhood $N_s$) to obtain:

$$v_k(s) = \frac{1}{5} \sum_{s' \in N_s} w_k(s')$$  

(3.6)

We represent the coordinates of the observations using radial basis functions and spectral decomposition, and then calculate $F_j(s)$ and $\lambda_j$, that we use to plug in in our model.

After obtaining all the pre-defined quantities, we perform a MCMC sampling scheme to
estimate the posterior distribution. The covariance function for the first layer of the hierarchical model is specified as exponential, and thus include two parameters: \( \Omega = \{ \sigma^2_x, \phi \} \). The hyperpriors are defined as: \( \sigma^2_b \sim IG(b, c) \), \( \sigma^2_{\varepsilon} \sim IG(d, e) \) and \( \phi \sim U(0, R) \). We use Gibbs sampling for \( \sigma^2_b \) and \( \sigma^2_{\varepsilon} \) and rejection sampling for all of the other parameters: \( \alpha_{kj} \) and \( \phi \). In order to evaluate the likelihood of \( Y \) in each MCMC iteration, we use Vecchia's approximation (Vecchia [1988]) as specified by Guinness [2016], to speed up the computation. The quantities \( b, c, d, e, R \) are fixed to have noninformative priors for \( \sigma_b \) and \( \sigma_{\varepsilon} \) and we use \( R \) to set a limit on the maximum range for the exponential covariance function.

**Vecchia’s approximation**

Vecchia’s approximation (Vecchia [1988]) is a way to reduce the computing time and physical memory needed to evaluate a likelihood function. Let \( Y \sim N(\mu, \Sigma_\theta) \) where \( \Sigma_\theta \) is a matrix with entries determined by a covariate function \( K \). The cost of evaluating the likelihood is in \( O(n^3) \) flops and the memory consumption is in \( O(n^2) \). If we have a permutation of the integers \( 1 : n \) \( \tau : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\} \) and a permuted vector \( y^\tau \) where \( y^\tau_i = y_{\tau(i)} \), then we can write the likelihood as:

\[
p_{\theta}(y_1, \ldots, y_n) = p_{\theta}(y_1) \prod_{i=2}^n p_{\theta}(y_i^\tau | y_1^\tau, \ldots, y_{i-1}^\tau).
\] (3.7)

Vecchia’s approximation is:

\[
p_{\theta, \tau, J}(y_1, \ldots, y_n) = p_{\theta}(y_1^\tau) \prod_{i=2}^n p_{\theta}(y_i^\tau | y_{j_{i1}}^\tau, \ldots, y_{j_{im_i}}^\tau)
\] (3.8)

where \( j_{i1}, \ldots, j_{im_i} \) is an increasing sequence of integers between 1 and \( i - 1 \), and \( y_{j_{i1}}^\tau, \ldots, y_{j_{im_i}}^\tau \) is a subset of \( y_1^\tau, \ldots, y_{i-1}^\tau \). The maximum minimum distance permutation was shown to provide good approximations when nearest neighbor conditioning sets were used. Guinness [2016] also provided an algorithm for grouping components of the likelihood approximation to increase speed and accuracy. We use this approximation to make the MCMC sampling from our emulator more computationally efficient.
3.4 Simulation Study

We study the performance of our model when estimating the spatially varying regression coefficients compared to a model that assumes the weights $v_k(s)$ are constant. Our main focus is to describe the ability of our model to estimate the coefficients accurately; thus we compare performance for varying degrees of spatial correlation in the weights, generating three different settings. For that, we use the MSE, the median absolute deviation (MAD) and the coverage for the coefficients $\beta_k(s) = v_k(s)f_k(s)$ in space.

We generate one set of $f_k(s)$ per simulation and 3 sets of $v_k(s)$ under different conditions, to have 3 settings as a result. In all settings we use a $10 \times 10$ spatial grid, with $N = 100$ locations and generate data for $X_k$ with $k = 1, 2, \ldots, 5$, each with a normal distribution $X_k \sim N(0, 2^{(1/k)})$ with $T = 12$ observations. Since the number of covariates is $p = 4$ the total number of possible models is $L = 2^p = 16$.

The set of $f_k(s)$ is generated as 4 independent spatially varying fields per simulation. Each of them is generated using a Matérn covariance function with different parameter values, as shown in Figure 3.4.

![Figure 3.4: Example of $p = 4$ fields for $f_k(s)$.](image)

The three options to generate $v_k(s)$ are (1) $v_k(s) = 1$ for all $s$ and $k$, (2) chessboard surfaces of $v_k(s)$ divided into 4 blocks that differ according to the covariate $k$, and (3) smooth surfaces that range from 0 to 1. For the chessboard case covariate $k = 1$ has weights of 0.5 in each of the
Northwest and Southeast corners, covariate $k = 2$ has weights of 0.5 in each of the Northeast and Southwest corners, and covariates $k = 3$ and 4 have weights of 0.25 on the South and North blocks respectively. For clarity, Figure 3.5 present spatial designs for $v_k(s)$, with $k = 1, 2, 3, 4$.

Then, we use the set of $f_k(s)$ and the two sets of $v_k(s)$, to have six combinations for $\beta_k(s) = f_k(s)v_k(s)$ and calculate $\mu(t) = \sum_{k=0}^{p} X_{tk}\beta_k(s)$ for each. We use $\mu(t)$ to generate random samples of $Y_t(s)$ using two options: one from a normal distribution centered in $\mu(t)$ and with $\sigma^2 = 1$ (independent samples over space), and the other one with $Y \sim GP(\mu, \Sigma)$, where $\Sigma$ is generated using an exponential covariance function with range 8 and variance 2.

Following the procedures outlined above, we generate $B = 80$ data sets in each of six settings, 3 assuming independence for $Y$ and 3 assuming a spatial covariance for $Y$, as described in Table 3.2.

We test the performance of five assumed models under six different data settings. The two models we consider are different versions of our model (Model 1 and Model 2) using two options to model the covariance of $Y$ and with and without incorporating the weights. Also,
Table 3.2: Six settings with their respective sets for $v_k(s)$ and $Y$.

<table>
<thead>
<tr>
<th>$Y \sim GP(\mu, \Sigma)$</th>
<th>$v_l(s)$</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma = \sigma^2 \varepsilon I$</td>
<td>Constant</td>
<td>Setting 1</td>
</tr>
<tr>
<td></td>
<td>Chessboard</td>
<td>Setting 2</td>
</tr>
<tr>
<td></td>
<td>Smooth</td>
<td>Setting 3</td>
</tr>
<tr>
<td>$\Sigma$ exponential</td>
<td>Constant</td>
<td>Setting 4</td>
</tr>
<tr>
<td></td>
<td>Chessboard</td>
<td>Setting 5</td>
</tr>
<tr>
<td></td>
<td>Smooth</td>
<td>Setting 6</td>
</tr>
</tbody>
</table>

we compare the results to the full Bayesian model described in Section 3.3 to compare the performance in terms of running time and how well it estimates the coefficients.

- **Model 1:** Fast Emulator assuming independence over $Y$'s.
  - (a) Assuming all $v_k(s) = 1$.
  - (b) Assuming $v_k(s)$ as a smooth surface, pre-estimated.

- **Model 2:** Fast Emulator modeling the covariance of $Y$.
  - (a) Assuming all $v_k(s) = 1$.
  - (b) Assuming $v_k(s)$ as a smooth surface, pre-estimated.
  - (c) Assuming $v_k(s)$ as a smooth surface.

For Models 1b and 2b we pre-estimate $v_k(s)$ to plug it into the Bayesian model, and for model 2c we include posterior sampling of $v_k(s)$ using a fixed exponential covariance. We use 10000 MCMC iterations, with a burn in period of 3000 and we thin the sample every 5 iterations. We compute posterior medians and credible intervals, mean standard error (MSE), median absolute deviation (MAD) and coverage for each set of coefficients and average them over covariates, space, and iterations. For a particular model and setting MAD is computed in the following way, for each setting and model:

$$\text{MAD} = \text{Median}_{i \in \{1,N\}, b \in \{1,B\}, k \in \{1,p\}} |\hat{\beta}^b_k(s_i) - \beta^p_k(s_i)| \quad (3.9)$$
where \( B = 80 \) data sets and \( N = 100 \) locations. For MSE and coverage the mean is also computed over data sets, locations and covariates.

Tables 3.3, 3.4 and 3.5 present MSE, MAD, and mean coverage for each set of covariates, according to the settings used to generate the data and the models assumed to estimate it. Values were compared using the empirical standard error and credible intervals. The MSE and MAD for the coefficients when assuming Model 2a and 2b are statistically smaller in all of the settings generated with correlated \( Y \)s, compared to those from Model 1a and 1b. When comparing results for models using pre-estimated weights for \( \beta \) (1b and 2b) versus those that do not, Setting 2 and 5 (chessboard with and without correlated \( Y \)s) have a statistically significant improvement in both MSE and MAD. Models that take the weights into account have up to a 60% decrease of MAD for \( \beta \) relative to those that assume constant \( v_s(k) \). Nevertheless, this difference is not reflected when comparing results for data generated with smooth weights and correlated \( Y \) when we do not model the variance of \( Y \) as spatial (setting 6 with 1a and 1b). This means that the fast emulator performs better or the same than the model that does not take into account the weights for \( \beta \), for data generated using chessboard-like, smooth weights or not using weights, if we model the variance of \( Y \) using a spatial process.

Table 3.3: Average MSE for \( \beta \) and s.e in parenthesis, for each model and setting.

<table>
<thead>
<tr>
<th>Setting</th>
<th>M1a</th>
<th>M1b</th>
<th>M2a</th>
<th>M2b</th>
<th>M2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting 1</td>
<td>0.178 (0.053)</td>
<td>0.179 (0.055)</td>
<td>0.397 (0.004)</td>
<td>0.397 (0.004)</td>
<td>0.831 (0.088)</td>
</tr>
<tr>
<td>Setting 2</td>
<td>0.623 (0.036)</td>
<td>0.382 (0.056)</td>
<td>0.655 (0.010)</td>
<td>0.439 (0.006)</td>
<td>1.449 (0.036)</td>
</tr>
<tr>
<td>Setting 3</td>
<td>0.341 (0.050)</td>
<td>0.312 (0.050)</td>
<td>1.127 (0.005)</td>
<td>1.154 (0.052)</td>
<td>1.186 (0.059)</td>
</tr>
<tr>
<td>Setting 4</td>
<td>0.328 (0.028)</td>
<td>0.331 (0.024)</td>
<td>0.100 (0.004)</td>
<td>0.100 (0.004)</td>
<td>0.859 (0.074)</td>
</tr>
<tr>
<td>Setting 5</td>
<td>1.655 (0.036)</td>
<td>0.505 (0.095)</td>
<td>0.427 (0.011)</td>
<td>0.191 (0.005)</td>
<td>1.471 (0.035)</td>
</tr>
<tr>
<td>Setting 6</td>
<td>1.207 (0.043)</td>
<td>1.207 (0.039)</td>
<td>0.143 (0.005)</td>
<td>0.035 (0.014)</td>
<td>1.478 (0.060)</td>
</tr>
</tbody>
</table>
Table 3.4: Average MAD for $\beta$ and s.e in parenthesis, for each model and setting.

<table>
<thead>
<tr>
<th>Setting</th>
<th>M1a</th>
<th>M1b</th>
<th>M2a</th>
<th>M2b</th>
<th>M2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.284 (0.042)</td>
<td>0.286 (0.043)</td>
<td>0.415 (0.004)</td>
<td>0.414 (0.003)</td>
<td>0.669 (0.031)</td>
</tr>
<tr>
<td>2</td>
<td>0.603 (0.009)</td>
<td>0.391 (0.047)</td>
<td>0.856 (0.006)</td>
<td>0.458 (0.004)</td>
<td>0.897 (0.011)</td>
</tr>
<tr>
<td>3</td>
<td>0.438 (0.009)</td>
<td>0.402 (0.009)</td>
<td>0.558 (0.003)</td>
<td>0.558 (0.017)</td>
<td>0.565 (0.013)</td>
</tr>
<tr>
<td>4</td>
<td>0.490 (0.021)</td>
<td>0.487 (0.018)</td>
<td>0.211 (0.004)</td>
<td>0.211 (0.004)</td>
<td>0.673 (0.028)</td>
</tr>
<tr>
<td>5</td>
<td>0.856 (0.009)</td>
<td>0.415 (0.037)</td>
<td>0.502 (0.006)</td>
<td>0.295 (0.004)</td>
<td>0.899 (0.010)</td>
</tr>
<tr>
<td>6</td>
<td>0.559 (0.008)</td>
<td>0.560 (0.007)</td>
<td>0.209 (0.003)</td>
<td>0.098 (0.007)</td>
<td>0.738 (0.015)</td>
</tr>
</tbody>
</table>

Table 3.5: Average coverage for $\beta$ and s.e in parenthesis, for each model and setting. Nominal level is 95%

<table>
<thead>
<tr>
<th>Setting</th>
<th>M1a</th>
<th>M1b</th>
<th>M2a</th>
<th>M2b</th>
<th>M2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.932 (0.012)</td>
<td>0.915 (0.011)</td>
<td>0.874 (0.007)</td>
<td>0.873 (0.006)</td>
<td>0.711 (0.010)</td>
</tr>
<tr>
<td>2</td>
<td>0.883 (0.003)</td>
<td>0.899 (0.008)</td>
<td>0.775 (0.008)</td>
<td>0.896 (0.006)</td>
<td>0.781 (0.011)</td>
</tr>
<tr>
<td>3</td>
<td>0.892 (0.006)</td>
<td>0.911 (0.003)</td>
<td>0.764 (0.006)</td>
<td>0.717 (0.007)</td>
<td>0.704 (0.008)</td>
</tr>
<tr>
<td>4</td>
<td>0.860 (0.008)</td>
<td>0.855 (0.007)</td>
<td>0.967 (0.006)</td>
<td>0.975 (0.006)</td>
<td>0.735 (0.012)</td>
</tr>
<tr>
<td>5</td>
<td>0.901 (0.001)</td>
<td>0.903 (0.007)</td>
<td>0.881 (0.005)</td>
<td>0.936 (0.006)</td>
<td>0.787 (0.011)</td>
</tr>
<tr>
<td>6</td>
<td>0.739 (0.003)</td>
<td>0.711 (0.003)</td>
<td>0.887 (0.008)</td>
<td>0.985 (0.008)</td>
<td>0.753 (0.012)</td>
</tr>
</tbody>
</table>

MAD and MSE results reflect the fact that our fast emulator benefits from using spatial correlation to get a better fit when there is spatial correlation in $Y$ present, compared to the fit to independent $Y$s. Also, the use of pre-estimated weights benefit the fit of the model by 20% – 60% in cases where the data is generated using spatially varying weights. Therefore, when there is spatial correlation present in the data and the covariates’ effect varies in space the simulation shows not only the need of a model that accounts for both the spatial correlation
in the coefficients and in the response, but also how our approach to estimate these two levels of spatial dependency works better than the simpler model to estimate the coefficients.

Table 3.6: Time to run 10000 MCMC iterations per model, \( N = 100, T = 12 \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1b) Independent ( Y ) / pre-estimate ( w )</td>
<td>21.6</td>
</tr>
<tr>
<td>(2b) Exponential covariance ( Y ) / pre-estimate ( w )</td>
<td>70.2</td>
</tr>
<tr>
<td>(2c) Exponential covariance ( Y ) / fix covariance ( w )</td>
<td>84.1</td>
</tr>
<tr>
<td>(Full Bayesian) Exponential covariance ( Y ) / complete model</td>
<td>324.0</td>
</tr>
</tbody>
</table>

Model 2c performs poorly in all the cases. Although it is not significantly slower per iteration than model 2b in this simulation, there is evidence to prefer model 2b even though the error in \( v \) is not incorporated into the model. This might be due to the overparametrization of model 2c and also to the difficulties to make all the parameters converge in this case (see Figure 3.6). The fourth option presented in Table 6 is not feasible given its running times, thus we did not included it. Furthermore, this simulation was done using \( N = 100 \) and, as we note in the next section, the computational cost for model 2c becomes a real burden when using \( N=5200 \) and in this case its benefits are not clear given its poor performance.

3.5 NARCCAP Data Analysis

In this section, the methodology from section 3.3 is applied to the data from section 3.2. The response in this case is precipitation (PR), and the final result is a high resolution climate map for the Western North American Region. In the application of interest, it is important to be able to (i) generate samples of the high resolution climate maps in a fast way (relative to the numerical models), and (ii) generate very similar variability patterns as in the original data set.

Feedback from the developers of NARCCAP data and some exploratory analysis allowed us
to identify six input variables as primary determinants for precipitation, which in the notation of Section 3.3 translates into $P = 6$ and $T = 30$. We use $N = 5200$ locations in Western North America, described in Section 3.2.

For the purpose of showing the advantages of the proposed model, we apply and compare the results of the following emulators:

- (a) Full Bayesian model with fixed covariance structure for the weights (Model 2c).
- (b) Two Stage Bayesian model with spatially dependent errors (Model 2b).
- (c) Two Stage Bayesian model with spatially independent errors (Model 1b).
- (d) A Gaussian Process without spatially varying coefficients nor variable selection, i.e. $\beta_k(s) = \beta_k$.

The gaussian process model is used as a reference model, since it is the most common statistical emulator for computer output (Overstall and Woods [2016]). In each model, all parameters were estimated for the response PR by their marginal posterior medians from an MCMC sample of size 10000. The number of MCMC iterations was determined using the lower bound for effective

![Figure 3.6: Trace plots for parameter $\sigma^2$ for (a) Model 2b and (b) Model 2c.](image)
sample size with a Monte Carlo error of 0.05 and for a 95% credible region on the estimates, proposed by Vats et al. [2017]. We discard the first 5000 as burn-in, and thin the chains by keeping every fifth iteration. We use the trace plots to check for convergence of all the parameters.

As part of our exploratory analysis, we ran several descriptive summaries to decide if it was necessary to include a temporal correlation structure for the response. The first summary per location we use is the ratio of the standard deviation of the residuals from two regression models with log precipitation as response and reference height temperature (TREFHT) as covariate: one with a lag 1 response as a covariate, and the other without the lagged term, to then obtain one ratio per each of the 5200 locations. The distribution of these ratios is centered in 1 and 50% of its values are between 0.990 and 1.018. Since the minimum value is 0.766 and the maximum is 1.020, the ratio does not present evidence of a lag 1 temporal component and thus we decide that including a temporal component does not improve our model. This temporal independence allows us to parallelize the approximation of the likelihood of $Y$ in time for each MCMC iteration, which is an advantage in practice.

We evaluate the performance of in-sample and out-of-sample estimations for each of the emulators using likelihood scores. For the out-of-sample evaluation, we select the last 2 years from the original data set and construct the emulator using the other 28 years. For each of those 2 years, we save the true RCM outputs at the 5200 locations, and evaluate likelihoods of those observations using the four emulators, and with three different values for $J$ for each of the three emulators that use basis functions (a,b,c), to evaluate the sensitivity (or the lack thereof) of the number of basis functions included in our emulators.

3.5.1 Results

The estimated time to run each of the models is presented in Table 3.7. Model 3b proves to be a computational burden when working with more than 5000 locations. Our results for this model are presented although they are not very meaningful, since the model did not converge
Table 3.7: Timing comparison for all models. Comparisons are done on a Dell Optiplex 9020 with a 3.60GHz Intel i7-4790 processor with 32GB RAM, and the code is written in the R programming language (R Core Team [2013]).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Model 1b</th>
<th>Model 2b</th>
<th>Model 2c</th>
<th>Gaussian Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 5200, T = 28 years</td>
<td>(J = 5,10,15)</td>
<td>(J = 5,10,15)</td>
<td>(J = 5,10,15)</td>
<td></td>
</tr>
<tr>
<td>Time to run 10000 MCMC</td>
<td>~14 hours</td>
<td>~22 hours</td>
<td>~32 hours</td>
<td>~20 hours</td>
</tr>
</tbody>
</table>

with the same conditions as the other models.

We use mean squared error (MSE), median average deviation (MAD) and mean (negative) likelihood score (MLS) to compare the performance of each of the four emulators when estimating the response. We define the likelihood score as:

\[
MLS = \frac{1}{N} \sum_{t=1}^{T} - \log p(y_t(s) | \hat{\alpha}, \hat{\sigma}^2, \hat{\Omega})
\]

where \(\hat{\alpha}, \hat{\sigma}^2, \) and \(\hat{\Omega}\) are the estimated parameters. We use \(t \in \{1, 28\}\) to construct the models when we have in-sample calculations and \(t \in \{29, 30\}\) when we have out-of-sample calculations.

For all of them, smaller values represent good estimations. Also, we use the proportion of observations where the credible interval includes the observed value and labeled it as \(Y \in CI\).

The results in Table 3.8 and Table 3.9 show how models 1b and 2b have a significantly better performance compared to the Gaussian emulator and model 3b for both the in-sample and out-of-sample data sets, as expected. The Gaussian emulator does not include spatially varying coefficients, thus its estimations have very wide credible intervals compared to those of models 1b and 2b. This explains why the model presents 100% of observed values inside the credible interval for both in-sample and out-sample results, but a very poor MLS.

Different values for \(J\) were used for models 1b, 2b and 3b, to evaluate how sensitive the estimations are to the number of eigenfunctions used to estimate the coefficients. The in-sample and out-of-sample results show almost no difference when using \(J = 10\) and \(J = 20\), which is expected, given that the values of \(\lambda_j^2\) for \(j \geq 8\) are almost zero. The result is not the same for \(J = 5\) which means that the model can be sensitive to having a number of eigenfunctions that
Table 3.8: In-sample results for all emulators using different values of $J$ in the models that use basis decomposition representation. Standard errors are less than 0.02 for MSE, MAD and MLS.

<table>
<thead>
<tr>
<th>Model</th>
<th>$MSE(Y)$</th>
<th>$MAD(Y)$</th>
<th>$Y \in CI$</th>
<th>MLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1b, $J = 5$</td>
<td>1.39</td>
<td>0.94</td>
<td>0.92</td>
<td>0.0621</td>
</tr>
<tr>
<td>Model 1b, $J = 10$</td>
<td>2.01</td>
<td>1.12</td>
<td>0.94</td>
<td>0.0689</td>
</tr>
<tr>
<td>Model 1b, $J = 20$</td>
<td>2.04</td>
<td>1.13</td>
<td>0.93</td>
<td>0.0695</td>
</tr>
<tr>
<td>Model 2b, $J = 5$</td>
<td>1.20</td>
<td>0.87</td>
<td>0.94</td>
<td>0.0365</td>
</tr>
<tr>
<td>Model 2b, $J = 10$</td>
<td>1.54</td>
<td>0.97</td>
<td>0.92</td>
<td>0.0378</td>
</tr>
<tr>
<td>Model 2b, $J = 20$</td>
<td>1.53</td>
<td>0.90</td>
<td>0.90</td>
<td>0.0381</td>
</tr>
<tr>
<td>Model 3b, $J = 5$</td>
<td>856.91</td>
<td>21.52</td>
<td>0.01</td>
<td>39.0274</td>
</tr>
<tr>
<td>Gaussian Model</td>
<td>2.44</td>
<td>1.34</td>
<td>1.00</td>
<td>0.1133</td>
</tr>
</tbody>
</table>

Table 3.9: Out-of-sample results for all emulators using different values of $J$ in the models that use basis decomposition representation. Standard errors are less than 0.05 for MSE, MAD and MLS.

<table>
<thead>
<tr>
<th>Model</th>
<th>$MSE(Y)$</th>
<th>$MAD(Y)$</th>
<th>$Y \in CI$</th>
<th>MLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1b, $J = 5$</td>
<td>1.36</td>
<td>0.93</td>
<td>0.92</td>
<td>0.8634</td>
</tr>
<tr>
<td>Model 1b, $J = 10$</td>
<td>1.94</td>
<td>1.10</td>
<td>0.95</td>
<td>0.9988</td>
</tr>
<tr>
<td>Model 1b, $J = 20$</td>
<td>2.05</td>
<td>1.21</td>
<td>0.96</td>
<td>1.0034</td>
</tr>
<tr>
<td>Model 2b, $J = 5$</td>
<td>1.06</td>
<td>0.81</td>
<td>0.89</td>
<td>0.8116</td>
</tr>
<tr>
<td>Model 2b, $J = 10$</td>
<td>1.84</td>
<td>1.08</td>
<td>0.92</td>
<td>0.8542</td>
</tr>
<tr>
<td>Model 2b, $J = 20$</td>
<td>1.95</td>
<td>1.15</td>
<td>0.91</td>
<td>0.8481</td>
</tr>
<tr>
<td>Model 3b, $J = 5$</td>
<td>858.71</td>
<td>21.54</td>
<td>0.01</td>
<td>1093.7231</td>
</tr>
<tr>
<td>Gaussian Model</td>
<td>2.29</td>
<td>1.29</td>
<td>1.00</td>
<td>2.9967</td>
</tr>
</tbody>
</table>

is lower than the value of $j$ for which $\lambda_j^2$ approaches zero. This could mean that if we use all the possible basis with $\lambda_j \neq 0$ we could be overfitting one or more coefficients..

Figure 3.7 contrasts RCM output values against corresponding predictions produced by each of the best emulators (1b and 2b), for a random sample of locations and years of the out-of-sample data set. For both emulators, the estimated (posterior median) output lies close to the true NARCCAP output, showing that they are emulating the true model quite accurately.

We choose to work with model 2b with $J = 5$ given its good performance for estimating the response in both in-sample and out-of-sample data sets. In terms of the interpretation, the most important parameter to interpret in this case is $v_k(s)$, since it can explain which covariates are
Figure 3.7: Comparison between model 1b and model 2b using a random sample of locations and places from the out-of-sample data set. Black dots represent posterior medians, and the lines represent the 95% credible intervals, for each model.

The main drivers for RCM output and where. Temperature (TREFHT) is an important driver for summer precipitation in the northwestern and northern region of the United States, as it is the specific humidity (QBOT) but over the Pacific Ocean, in front of the West Coast of the U.S. From the jet level winds, the most important is U (East-West stream), that has a significant driver effect over the Rocky Mountains in the U.S. and in front of the Canadian west coast. Lastly, the vertical velocity shows evidence of being a good driver for summer precipitation in the western part of Canada.
Figure 3.8: Weights $v_k(s)$ for each covariate. Areas in white represent values not significantly different from zero, i.e. areas where the respective covariate does not represent a significant driver for summer precipitation.

3.6 Discussion

This chapter focuses on two objectives. First, we develop a statistical RCM emulator using GCM output, incorporating a spatial variable selection method to the Bayesian hierarchical
model. Second, we consider the use of various versions of our emulator to model RCM output for summer precipitation in the Western part of the North American Region, using NARCCAP data. A discussion of the modeling advantages and restrictions implied by the four alternative emulators suggested that two stage Bayesian model with spatially dependent errors (Model 2b from Section 3.4) approach was the most satisfactory in terms of both variable selection and prediction performance. The full Bayesian model option presented several restrictions and no advantages, given its long running time and poor results.

Our approach proves to be computationally feasible and flexible enough to accommodate complex spatial structures. We show how using a small number of basis representations for each covariate can give the flexibility to the model to represent these complex spatial structures. Furthermore, we show how the use of data-dependent weights can facilitate the shrinkage of the resulting spatially varying coefficients, while keeping a feasible computing time for the model.

The evidence we present shows that the two stage emulator can form the basis for successful emulation of dynamic downscaling climate models, which in practice is an important tool for solving problems such as uncertainty analysis, sensitivity analysis and calibration of regional climate models. However, there remain some directions for further research, such as the incorporation of a time structure in the response covariance in a computationally feasible way, in order to tackle the problem of monthly or daily emulation of RCMs.
Chapter 4

Verification of Statistical Seasonal Tropical Cyclone Forecast

4.1 Introduction

Quantifying and characterizing the uncertainty in forecasts is essential for their proper statistical interpretation and further utilization by public and private decision makers (government officials, insurance companies, agriculture-based business, among others). Forecasts are a result of many decisions such as which month-specific covariates to include, how to select them, and what probability distribution to use. Thus, an evaluation of the impact of these decisions when constructing a forecast is needed. This chapter aims to provide a framework to validate and evaluate the model and method decisions while taking into account the probabilistic nature of the forecast, by using a probability distribution as opposed
Table 4.1: Forecast groups that use statistical models exclusively, with their most recent reference.

<table>
<thead>
<tr>
<th>Forecast Group</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Storm Risk (UCL)</td>
<td>Saunders and Lea [2005]</td>
</tr>
<tr>
<td>North Carolina State University</td>
<td>Xie et al. [2014], Keith and Xie [2009]</td>
</tr>
<tr>
<td>Coastal Carolina University</td>
<td>Yan et al. [2015]</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>Koar et al. [2012]</td>
</tr>
<tr>
<td>University of Arizona</td>
<td>Davis et al. [2015]</td>
</tr>
<tr>
<td>Instituto Meteorológico Nacional (México)</td>
<td>N/A</td>
</tr>
<tr>
<td>Instituto Meteorológico (Cuba)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

After the VI International Workshop on Tropical Cyclones (ITWC-VI) in 2006, the agreement among forecast groups was “to develop guidelines for the development and validation of the tropical cyclone forecasts” (Camargo et al. [2010]). The World Meteorology Organization (WMO) has published suggested verification skill scores and general guidelines for all long-range forecasts (LRF) (WMO [2002]). Also, WMO encourages the use of probabilistic forecasts, given that they suggest the use of proper scores (as defined in Gneiting and Katzfuss [2014]) to evaluate all LRFs.

The Seasonal Hurricane Predictions website (Caron and Klotzbach [2016]) presents seventeen forecast groups that release their hurricane seasonal outlook for the Atlantic Basin every year. Seven of them use statistical forecasts exclusively and are presented in Table 4.1. Commonalities among all the groups with statistical forecasts are the use of storm counts as the predictand and measures of sea surface temperatures and climate indices as predictors.

Most of the literature on statistical seasonal hurricane forecasts have a section on verification, where they evaluate their forecast based on methods like cross-validation (CV) or jackknife. For example, Saunders and Lea [2005] use a 5-year block cross-validation hindcast, following the recommendations from Elsner and Schmertmann [1994] and WMO [2002]. They eliminate 5-year blocks of observations at a time, then hindcast those values and compare them with the observed ones using Spearman rank correlation. Leave-one-out cross-validation (LOOCV) and ranked probability skill score are used by Xie et al. [2014] and Keith and Xie [2009] to compare
their forecast with climatology. Yan et al. [2015] also uses LOOCV, and calculate the root mean square error and the ranked skill score for the same purpose.

Kozar et al. [2012] and Davis et al. [2015] use the data for two independent cross-validation experiments to evaluate their forecast skill. Kozar et al. [2012] split the data first into two equal parts and uses the first half to calibrate the model, then forecasts the second half. The second experiment is a one-year validation methodology, where they predict the number of storms in one year, using all previous data. For both experiments, they use the coefficient of determination $R^2$, the coefficient of efficiency, and the reduction of error as scores. Davis et al. [2015] on the other hand, divide the data set into two periods: 1950-1980 and 1981-2013 to test the robustness of their forecast. In the first period they use future data to fit a model and predict past values, as a sensitivity test, and for the second part they use only past data to forecast each year. They use mean absolute error and root mean square error to compare their statistical model prediction capabilities with respect to the climatology.

Research from Jagger and Elsner [2010], Lehmiller et al. [1997] and Klotzbach [2007] are not included in Caron and Klotzbach [2016] as groups with statistical seasonal hurricane forecasts, but they have proposed verification techniques for statistical forecasts in the literature. Elsner and Jagger [2006] first use a Bayesian model to forecast and maximum a posteriori as hindcast skill, and in a later paper they use 11-fold cross-validation to compare their consensus model with other options in terms of mean square error, ranked probability and Brier (quadratic) score, as well as the logarithmic score (Jagger and Elsner [2010]). Klotzbach [2007] uses hindcast cross-validation and warns that the practice of evaluating hindcasts using future data to predict the past “should only be used as an upper bound measure for skill”. In this case, $R^2$ and MSE are used to measure the forecast skill. Lehmiller et al. [1997] provide a different perspective and use multivariate discriminant analysis to forecast location and occurrence of storms in the Atlantic Basin. They use proportion of correct forecasts in cross-validation to verify their forecasts.

An important part of constructing statistical forecasts is the variable selection method. Most of the literature in statistical seasonal hurricane forecasts use a small set of covariates when
constructing their statistical model (see for example Elsner and Jagger [2006], Davis et al. [2015] and Keith and Xie [2009]). They select the covariates based on their expert opinion, either from previous evidence of association or from scientific hypotheses. Xie et al. [2014] and Yan et al. [2015] use the forward and backward approach, to perform variable selection in a relatively small (9) set of covariates. Jagger and Elsner [2010] on the other hand, incorporate Bayesian model averaging techniques as an alternative for variable selection, where they average over all possible models constructed using different combinations of five covariates. None of these studies take into account all possible month-specific variables that have been shown to have an association with tropical cyclones, mainly due to the curse of dimensionality, in which there are more covariates than observations and thus, insufficient degrees of freedom to estimate the full model.

The main objective of this chapter is to propose a detailed guideline to verify statistical seasonal hurricane forecasts that takes into account only past data, and acknowledges its probabilistic nature. In order to highlight its importance, the verification method is compared to other methods that use hindcast (i.e. without making the difference between the past and future data) to verify forecasts. Furthermore, to illustrate the verification method, a set of forecasting models are constructed first and then evaluated. Specifically, these models aim to predict the distribution of tropical storms (TS), hurricanes (HR) and major hurricanes (MH) counts to form or pass through particular areas of the Atlantic Basin during a specific hurricane season (1 June to 30 November). Moreover, the use of LASSO and clustering as variable selection procedures to construct the forecasts is proposed and the need of incorporating them into the cross-validation procedure is discussed. These variable selection procedures for tropical cyclone forecast can handle a set of covariates larger than the number of observations.

The rest of this chapter is organized into four sections. The second section presents the definitions and general notations for statistical seasonal hurricane forecasts. In a third section the tropical cyclone data and possible predictors for the analysis are presented, as well as a description of the methodology used in the proposed verification procedure and the construction
and variable selection of the SHFs. The fourth section describes the numerical results in terms of selected predictors using the proposed variable selection method as well as the corresponding model verification. The last section closes with conclusions and discussion.

4.2 Definitions

4.2.1 Statistical Seasonal Hurricane Forecasts

Gneiting and Katzfuss [2014] define a probabilistic forecast as a “forecast in the form of a probability distribution over future quantities or events.” There are several ways to represent a probabilistic forecast, but Gneiting and Ranjan [2013] recommend to write it as a predictive cumulative distribution function (CDF), given its “unified treatment of all real-valued predictands.” These real-valued predictands can be density forecasts, mixed discrete-continuous predictive distributions, probability mass functions for count data, and probability forecast of a dichotomous event. The forecast $F$ can be constructed without observed data, i.e. assuming a probability distribution, or using observed data to estimate a probability distribution.

A statistical seasonal hurricane forecast (SHF) is defined as a probabilistic forecast for the total counts of tropical cyclones to form in a particular area, where the probability distributions are estimated from past data. According to Klotzbach et al. (2010) and Camargo et al. (2007) most of the statistical seasonal hurricane forecasts use regression models. The main difference between them is the response that is used. Saunders and Lea [2005] use the accumulated cyclonic energy (ACE) as a continuous response, while all other authors use the number of tropical cyclones (tropical storms, hurricanes, major hurricanes, landfalling storms) as a discrete response. In the latter case, the distribution assumed is discrete, specifically, a probability mass function for count data.

Consider $n$ years of data $\{y_i, x_{i1}, \ldots, x_{ip}\}$, where $y_i$ is a storm count in year $i$, and $x_{i1}, \ldots, x_{ip}$ are covariate values (e.g. sea surface temperature indices in year $i$). In the forecast literature, climatology is used as a reference to evaluate the efficacy of various forecast methods. In this
chapter, climatology is defined in two ways: (1) \( \hat{F}_1 = \text{Poisson}(\bar{y}) \), where \( \bar{y} = \sum_{i=1}^{n} y_i / n \), and (2) \( \hat{F}_2 \) as the empirical CDF of \( y_1, \ldots, y_n \). The second option is useful for shedding some light on the correctness of the Poisson distribution assumption. Climatology forecasts do not make use of covariates that vary from year to year. In order to incorporate these covariates, the mean storm count is modeled as

\[
\log \mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j.
\]

The dataset can be used to provide parameter estimates \( \hat{\beta}_0, \ldots, \hat{\beta}_p \), and then the forecast distribution for year \( n + 1 \) is simply \( F_3 = \text{Poisson}(\hat{\mu}_{n+1}) \), where

\[
\log \hat{\mu}_{n+1} = \hat{\beta}_0 + \sum_{j=1}^{p} x_{n+1,j} \hat{\beta}_j.
\]

When there are many covariates available, it is often preferred to set some of the parameter estimates exactly to zero, effectively eliminating certain covariates from affecting the forecast. A careful study of this variable selection procedure is one of the contributions of this chapter.

Compared to other seasonal hurricane forecasts like the dynamical or blended, statistical forecasts have the advantage of being extremely fast to compute, simple to implement and interpret, and also that they provide a measure of variability, due to their probabilistic nature.

### 4.2.2 Verification Methods for Statistical Seasonal Hurricane Forecasts

According to Wilks [2011] a “forecast verification involves measures of the relationship between a forecast or a set of forecasts, and the corresponding observation(s) of the predictand.” Given the nature of the statistical model, using the same data to construct the statistical forecast and also verify it, will cause underestimation of errors (James et al. [2013]). Therefore, when proposing a verification method for a statistical forecast, there are two procedures that need to be addressed: first, how to measure the relationship between the forecast and the observed values, and second, how to separate the data into training and test sets and thus calculate and
evaluate the forecast using independent data. The definitions for each of the procedures are as follows.

**Scoring Rules**

A scoring rule function measures the accuracy of probabilistic predictions. According to Gneiting and Katzfuss [2014], a scoring rule $S(F,y)$ assigns a numerical score to each pair of $(F,y)$, where $F$ represents a probabilistic forecast and $y$ is the realized value. In the meteorology literature, many authors rely on the root mean squared error or simply on a binary measure of the confidence interval including or not the real value as distance measures (e.g. hit rate, linear error probability space, probability of detection, false alarm rate, among others). The problem with a binary measure in this case, is that ignores the forecast variability.

Gneiting and Katzfuss [2014] define a scoring rule to be proper if the expected score of an observation drawn from the distribution $G$ is minimized using the probabilistic forecast $G$, rather than any other $F \neq G$. In other words, if $S$ is a proper scoring rule, and if observations $Y$ are drawn from $G$, and our forecast distribution is $F$, then $E(S(G,Y)) \leq E(S(F,Y))$. The scoring rule is strictly proper if the inequality is strict whenever $F \neq G$.

WMO [2002] recommends the relative operating characteristic curve (ROC) skill score, ranked probability skill score (RPSS) or likelihood skill scores (Harte and Vere-Jones [2005]) to validate statistical seasonal hurricane forecast, depending on the nature of the predictand. The ROC skill score is only useful for evaluating binary outcomes. RPSS and the likelihood skill score have the ranked probability score and the logarithmic score (LS) as their associated scoring rules, respectively. The LS is a proper scoring rule (Gneiting and Katzfuss [2014]), and the ranked probability has a modified version called continuous ranked probability score for evaluating probabilistic forecasts, which is also a proper scoring rule (Gneiting and Katzfuss [2014]).

The LS is a proper score that heavily penalizes when an event forecasted as a low probability event actually occurs. Thus, using LS has the advantage of being able to acknowledge
the probabilistic nature of the forecast, compared to MSE, MAE and other scalar accuracy measures. Compared to the continuous ranked probability score the LS is simpler to calculate, and Gneiting and Raftery [2007] recommended it for evaluating forecasts that are generated by probabilistic models, given its relationship with the likelihood ratio and the concept of Bayes factors.

The logarithmic score $\text{LS}(F, y)$ is used in this study to compare the forecast value from $F$ with the observed value $y_i$, using the following formula:

$$\text{LS}(F, y_i) = -\log [F(y_i | \hat{\mu}_{i-1}) - F(y_i - 1 | \hat{\mu}_{i-1})]$$ (4.1)

where $\hat{\mu}_{i-1}$ is the estimated parameter calculated using the previous $i - 1$ observations. The likelihood skill score is defined as:

$$\bar{H} = \frac{1}{n} \sum_{i=1}^{n} \left[ \text{LS}(\hat{F}, y_i) - \text{LS}(F, y_i) \right]$$ (4.2)

where $\hat{F}$ is the climatology defined in the previous section. The likelihood skill score recommended by the WMO is represented as $\exp(\bar{H})$ and is equivalent to the log probability gain (see entropy score in Harte and Vere-Jones [2005]). Positive values of $H$ or in the same way values of $\tilde{p}$ greater than 1 indicate superior skill than climatology for model $F$.

**Cross-validation and sliding window cross-validation**

In a data-rich situation, James et al. [2013] recommended splitting the observed data into training, validation and testing sets, to be able to characterize the forecast uncertainty properly. The same authors encourage the use of “efficient sample re-use, such as cross-validation or bootstrap” in cases where the data is limited. Cross-validation is a method that partitions a sample of data into complementary subsets, one to fit the model and the other to test it. These partitions can be done by splitting the data into $K$ equal-sized parts (K-fold) to fit the model using $K - 1$ parts of the data, and calculate and test the forecast using the remaining part.
The procedure is repeated for \( k = 1, 2, \ldots, K \) to have \( K \) estimates of an evaluation measure. When \( K \) is equal to the sample size, the cross-validation is called leave-one-out cross-validation (LOOCV). In all cases the final goal is to have several repetitions of the estimated measure, to construct an empirical distribution and hence an approximation of its variability.

Elsner and Schmertmann [1994] describe cross-validation for assessing forecast skill and warn about two common errors: “false cross-validation” and “serial correlation.” The first one refers to the use of information from the test set when fitting the model, and the second refers to the presence of observations in the training data set that are “especially informative about the omitted predictand, but unavailable in a real forecast situation” (Elsner and Schmertmann [1994]). The first problem commonly arises when performing variable selection procedures in a preliminary process that uses all the data to select predictors, and then partition the data to validate the final model. The second problem might be present when using k-fold CV, since it is possible that the use of future years can contribute to a better forecast when in reality the information will not be available in a real forecast situation.

Also, since calculating a scoring rule or any other statistic using only one sample is not useful to quantify uncertainty, a verification method needs not only to have a training set and a testing set, but also repetitions of said estimation. A verification process that uses a sliding window creates the repetitions needed and avoids the common cross-validation errors, since the method respects the temporal order nature of the data.

In a sliding window cross-validation (SWCV), each window is composed by \( n_T \) training observations (years) and a \( n_E = 1 \) testing observation (year). The \( i \)th testing set is \((y_i, x_{i1}, \ldots, x_{ip})\), with training set \((y_{i-n_T}, x_{i-n_T,1}, \ldots, x_{i-n_T,p}), \ldots, (y_{i-1}, x_{i-1,1}, \ldots, x_{i-1,p})\). Thus, the training set for each observation consists of the data in the immediately preceding \( n_T \) years, as shown in Figure 4.1. The model is fitted using \( n_T \) years and the scoring rule is calculated using the forecast of the \( n_E \) observation in each window of the data. The window incrementally advances across all possible values of \( i \) and, at each new year, the scoring rule is calculated using the forecast from the model constructed using \( n_T \) training observations. The procedure
Figure 4.1: Sliding windows: each row is a window $w$, displaying a training dataset ($n = n_T$) in blue, test data ($n = n_E$) in green and the final forecast year (2016) is displayed in red.

is repeated for $i = 1, \ldots, w$. In each window, a forecast value from $F_i$ is compared with the observed value $y_i$ using $LS(F_i, y_i)$. In this way, a statistical seasonal hurricane forecast can be constructed using each window of data and evaluated with the subsequent year.

Sliding window cross-validation allows repeating the exercise $w$ times, each of them with the same sample size, using the historical data. SWCV respects the chronological order of the data and gives a more realistic approximation of the true forecast procedure. On the other hand, it reduces the sample size, since the data is split into groups of smaller windows. Nevertheless, the advantages of having several samples to measure variability and having realistic forecast scenarios can overcome the small sample size disadvantage in this case. In the next section, the data and specific methods used in this chapter are presented.

### 4.3 Application to TC in the Atlantic Basin

The goal is to compare two climatology definitions and two cross-validation types, and then use the best combination to evaluate several model options for statistical seasonal hurricane forecasts, using tropical storm data from the Atlantic Basin. To illustrate the verification method, a set of forecasting models are constructed first and then evaluated in terms of the following modeling decisions: type of variable selection method, Outlook release month and the
starting set of variables considered.

4.3.1 Data

The historical tropical cyclone (TC) counts are obtained from the National Hurricane Center (NHC) HURDAT best-track data map available at: http://www.nhc.noaa.gov/pastall.shtml. TC counts are determined by region and then further categorized by the peak strength within each region according to the Saffir-Simpson hurricane wind scale. The three regions considered are the Atlantic Ocean, the Caribbean Sea, and the Gulf of Mexico (Figure 4.2). The Atlantic Ocean regions (ATL) stands for the whole Atlantic TC basin. The Caribbean Sea region (CAR) is enclosed by the West Indies and the coast of Central America from the East coast of the Yucatan Peninsula to Venezuela. The Gulf of Mexico region (GOM) is bordered by the Gulf Coast of the United States (from the Southern tip of Florida to Texas) to Mexico and the Northern edge of the Yucatan Peninsula and the Northwestern coast of Cuba. This best track data for the Atlantic basin contains data since 1851. However, only data after 1951 are used in this analysis because of the large uncertainties due to non reliable sources in the earlier data (Xie et al. [2014]). Subtropical and extra-tropical storms, as well as storms outside of the 1 June-31 November hurricane season, are filtered out. In total, there are 65-years (1951-2015) of observations for all variables, with a total of 698 tropical cyclones.

Various climate factors are considered as predictors of TC counts for an upcoming hurricane season. These candidate predictors include SST-related climate indices, El Niño Southern Oscillation (ENSO) related indices, atmospheric and teleconnection indices, as well as parameters in the Main Development Region (MDR) for Atlantic hurricanes.

NOAA Earth System Research Laboratory’s Physical Division (http://www.esrl.noaa.gov/psd/data/climateindices/list) compiles the information from NOAA sources, including Climate Prediction Center and Atlantic Oceanographic and Meteorological Laboratory. Atlantic SST-related climate indices such as AO (Atlantic Oscillation), NAO (Northern Atlantic Oscillation), AMM (Atlantic Meridional Mode), AMO (Atlantic Multi-decadal Oscilla-
Figure 4.2: Delimited regions for the forecast
tion), TNA (Tropical Northern Atlantic), TSA (Tropical Southern Atlantic) and DM (Atlantic Dipole Mode) are included as possible predictors and collected from such sources.

Pacific SST-related climate indices are also included. PNA (Pacific North American Index), WP (Pacific Warm-pool), WHWP (Western Hemisphere Warm Pool), SOI (Southern Oscillation Index), EPO (East Pacific/North Pacific Oscillation), and PDO (Pacific Decadal Oscillation) are listed in the aforementioned sources. Also, a group of SST indices from the correspondent Niño regions is listed (NINO12, NINO34, NINO3, NINO4), MEI (Multivariate ENSO Index), CENSO (Bivariate ENSO) and TNI (Trans-Niño Index). Other indices from the same sources are Solar Flux and QBO (Quasi-Biennial Oscillation).

Several measures from the main development region (MDR, 10-20N, 80-20W) are included as climate indices. These include monthly averaged sea-level pressure (MDRSLP), SST (MDRSST), and vertical wind shear (MDRVWS, wind shear between 200 and 850 mb). Also, 200 and 850 mb pressure level u and v wind (m/s) measures from the same region are included (MDRU200, MDRU850, MDRV200, and MDRV850). All the MDR indices are derived from the NCEP/NCAR Reanalysis I dataset by http://www.esrl.noaa.gov/psd/data/timeseries/

Global (GGST), north-hemisphere (NGST), and south-hemisphere (SGST) monthly Land-Surface Air and Sea-Surface Water Temperature Anomalies, based on the GISS Surface Temperature Analysis (GISTEMP), are also used as candidate predictors. Data is obtained from GISS (NASA) and is available from http://data.giss.nasa.gov/.

The surface latent heat flux (LHF) derived from the NCEP/NCAR reanalysis data is used to compute Empirical Orthogonal Functions (EOFs). Winter data is used to calculate anomalies and then construct EOFs. In total 34 indices with monthly measures per year are considered as predictors. Since many of these variables are almost perfectly correlated with each other, in each window a dendrogram analysis was carried out to define a list of 10 possible covariates. Definitions and more specific sources for all indices can be found in Appendix B.
4.3.2 Verification Procedure

The data is divided into 35 windows to use the sliding window procedure described in Section 4.2.2. The number of windows and the size of each of them is calculated in a way that the sample size for both windows and years within windows is not smaller than 30. A small sensitivity study was performed to see how the window size changed the results, with no significant change.

As a first step, the model is constructed using data from 1951-1980, to then use it to forecast 1981. Since there are 65 years in total and the window size is 30, the model construction can be repeated 35 times, moving the window by one year, so that the year 2015 can be forecasted using the last window of data, as shown in Figure 4.1. The log score \( \text{LS}(F, y) \) is calculated using the formula specified in Equation 4.1 and then \( \bar{H} \) is calculated with the number of windows as the sample size \( n \).

The statistical forecast \( F \) with higher \( \bar{H} \) will be then chosen to forecast 2016. In order to illustrate the verification method, a set of the forecasting models are constructed first and then evaluated using the proposed method. Leave-one-out cross-validation is used as a comparison since it is the most commonly used method, and will be labeled from now on as:

- **SWCV.** Proposed sliding window cross-validation method, described in Section 4.2.2.
- **LOOCV.** Regular leave-one-out cross-validation is used as baseline comparison. Again, future data is used to forecast the past.

All the standard error (SE) calculations were approximated using bootstrap with \( B = 10000 \) iterations and resampling \( H_i \) for each method. The bootstrap includes an implicit assumption of independent samples, which could be violated in this application, so we also computed standard errors under an AR(1) model for \( H_i \), and the results did not change significantly, so we used the bootstrap throughout.

The statistical forecasts estimate the distributions of the tropical storms (TS), hurricanes (HR) and major hurricanes (MH) counts to form or pass through particular areas of the Atlantic Basin during a hurricane season (1 June to 30 November). In the next section, the SHF
construction and variable selection procedure is presented.

4.3.3 SHFs Construction and Variable Selection

A generalized linear model LASSO regularization is used (James et al. [2013]). The model assumes that the logarithm of the intensity parameter $\mu$ is linearly related to the climate indices, and performs a variable selection procedure using a shrinkage parameter to minimize a penalized likelihood. LASSO is useful to handle cases where there are more covariates than observations. In this case, LASSO is especially useful since 35 covariates multiplied by the number of months observed (up to 7) is in many cases greater than the number of years observed. Nevertheless, depending on the window of data, the correlation between pairs of covariates can be almost perfect, in which case LASSO has limitations. Hence, a hierarchical clustering algorithm (hclust, from R Core Team [2013]) is applied before applying LASSO in each window, to select one variable for each of the $k = 10$ clusters constructed using distances from the covariates correlation matrix. The number of clusters $k = 10$ is selected to ensure $k < n$ (i.e. enough residual degrees of freedom left) and also since using different values of $k \in \{10, n-2\}$ did not change the results significantly. Figure 4.3 presents the results of a cluster analysis with all 65 years, as an example. The algorithm selects 10 clusters and then chooses the variable with highest correlation with the response in each cluster, in each window separately. Hastie et al. [2009] highlights the importance of incorporating the complete algorithm, including variable selection, in cross-validation, arguing that “in a multistep modeling procedure, cross-validation must be applied to the entire sequence of modeling steps.” (p.246). Therefore, both the cluster and LASSO procedures are applied in each window.

Nine sets of month-specific predictors were used to construct three different model options for the forecasts. These sets are a combination of time domains for the climate indices (covariates) and groups of covariates: Core group, NINO group and LHF group, with the objective of having different scenarios for the outlook release date (March, May or March with ENSO JAS (July, August, September)). The NINO group comprises all ENSO related climate indices, the LHF
Figure 4.3: Cluster analysis results for 33 possible covariates for all the data, as an example. Strong red tones represent perfect positive correlation and strong blue tones represent perfect negative correlation. A dendrogram shows the grouped variables.
Table 4.2: Possible variables and time domains.

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_1$ = March Outlook</th>
<th>$F_2$ = May Outlook</th>
<th>$F_3$ = March Outlook + JAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Domain</td>
<td>2 months (Jan-Feb)</td>
<td>4 months (Jan-Feb + JAS)</td>
<td>4 months (Jan-April)</td>
</tr>
<tr>
<td>NINO</td>
<td>NINO12 - NINO3 - NINO23 - NINO4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LHF</td>
<td>LHF EOF WINTER (1 to 5)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The group includes the EOFs constructed using LHF winter data, and the Core group contains all other indices (see Table 4.2). The goal in here is to investigate whether or not using forecasted covariates improves the storm forecasts (March vs March with ENSO) and/or if a forecast released at the beginning of the season is better than one released in April using forecasted data (May vs March ENSO).

In summary, the following models are used to make a comparison using the verification method, for each response:

$F_{1X}$: March Outlook using core variables ($F_{1B}$), core and NINO variables ($F_{1N}$), or core, NINO and LHF variables ($F_{1L}$). The model is constructed using data from January and February. This outlook can be ready by April 1st, when the first forecast is released.

$F_{2X}$: May Outlook using core variables ($F_{2B}$), core and NINO variables ($F_{2N}$), or core, NINO and LHF variables ($F_{2L}$). The model is constructed using data from January to April. This outlook can be ready by June 1st, when the forecast is updated at the beginning of the season.

$F_{3X}$: March Outlook and ENSO JAS using core variables ($F_{3B}$), core and NINO variables ($F_{3N}$), or core, NINO and LHF variables ($F_{3L}$). The model is constructed using data from January and February, plus the ENSO values from July, August, and September. This model uses future data since the NINO information is not available until the beginning of the hurricane season, but it is used to evaluate the possibility of equating the future NINO values’ forecast skill with other variables. This outlook can be ready by April 1st.
using forecasted JAS NINO data.

For each of the training sets $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$, a Poisson distribution with intensity parameter $\mu_i$ was used to model the counts: $Y_i \sim \text{Poisson}(\mu_i)$ with: $\log(\mu_i) = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j$. Here, it is of interest to minimize the following function (objective function):

$$\min_{\beta_0, \beta} \frac{1}{N} \ell(\beta | X, Y) + \gamma \left( \sum_{j=1}^p |\beta_j| \right)$$

where $\ell(\beta | X, Y)$ is the likelihood function,

$$\ell(\beta | X, Y) = \sum_{i=1}^N (y_i (\beta_0 + \sum_{j=1}^p x_{ij}\beta_j) - \exp(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j))$$

and $\gamma$ is the LASSO shrinkage parameter and $\beta_0$ is the intercept; $\beta_j$ and index $j$ are the regression coefficients and the selected indices, respectively, which are specific to each region and strength category. All covariates are scaled based on the correspondent window values.

For clarity, the procedure can be described in the following steps (see also Figure 4.4):

1. Divide the data into $w$ windows, depending on the cross-validation method ($w = 35$ for SWCV or no windows for LOOCV).

2. In each window, construct a model: select 10 primary covariates using a cluster analysis and depending on the model, apply LASSO using the 10 covariates in each window.

3. Calculate $LS$.

4. Compare scores against climatology ($\hat{F}_1$ and $\hat{F}_2$) using $\bar{H}$.

All calculations are done using $\alpha = 0.05$, and thus 95% confidence interval are presented in all cases. All the analysis were done using R (R Core Team [2013]), and the code needed to implement the verification method with or without the variable selection procedure can be found in https://github.com/malfaro18/Verification_SSTCF.
4.4 Results

4.4.1 Selected Predictors

Models using only the clustering procedure and both cluster and LASSO are used to select predictors for each of the forecasts $F$. It is important to highlight that the models might select different predictors in each window since the idea is to construct and test the model with different subsets of data (in each window). For that reason, a good summary of the most important predictors is a count of how many times each variable was selected. The results are presented as the most frequently selected variables using only the clustering, and also using both the clustering and LASSO.

Table 4.3 presents the five most selected predictors using the LASSO with SWCV method in each of the $F$ for the Atlantic Basin. February EPO, MDRSLP, MDRV200, and TSA are in all models of the March Outlook. May Outlook also includes February MDRSLP and incorporates March TSA, April NGST, and February TNA. ENSO JAS Outlooks include NINO12 from August in $F_{3B}$, $F_{3N}$ and $F_{3L}$ as one of the most commonly selected predictors. Figure B.1, shows a tile plot for each response, with the percentage of times a predictor is selected using the cluster analysis and the SWCV method. In each plot, the twenty variables that were selected the highest number of times are presented for each response.

Figure B.2 show the percentage of times a predictor is selected using clustering with SWCV method. May Outlook Models ($F_{2B}$, $F_{2N}$, $F_{2L}$) tend to select different variables than the March
Table 4.3: Atlantic Basin tropical storms: five most selected variables using LASSO, per $F$. NINO variables are presented in bold letters.

<table>
<thead>
<tr>
<th>$F$</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Fourth</th>
<th>Fifth</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{1B}$</td>
<td>EPO02 (0.80)</td>
<td>MDRSLP02 (0.63)</td>
<td>MDRV20002 (0.54)</td>
<td>MDRSLP01 (0.49)</td>
<td>TSA02 (0.46)</td>
</tr>
<tr>
<td>$F_{1N}$</td>
<td>MDRSLP02 (0.80)</td>
<td>EPO02 (0.69)</td>
<td>AMM01 (0.57)</td>
<td>MDRV20002 (0.54)</td>
<td>TSA02 (0.49)</td>
</tr>
<tr>
<td>$F_{1L}$</td>
<td>EPO02 (0.66)</td>
<td>MDRSLP02 (0.60)</td>
<td>LHF.WIN3 (0.57)</td>
<td>MDRV20002 (0.49)</td>
<td>TSA02 (0.43)</td>
</tr>
<tr>
<td>$F_{2B}$</td>
<td>TSA03 (0.74)</td>
<td>TNA02 (0.54)</td>
<td>NINO1202 (0.54)</td>
<td>MDRSLP02 (0.46)</td>
<td>NINO1208 (0.40)</td>
</tr>
<tr>
<td>$F_{2N}$</td>
<td>TSA03 (0.71)</td>
<td>NGST04 (0.57)</td>
<td>TNA02 (0.51)</td>
<td>MDRSLP02 (0.43)</td>
<td>NINO1208 (0.34)</td>
</tr>
<tr>
<td>$F_{2L}$</td>
<td>TSA03 (0.69)</td>
<td>TNA02 (0.57)</td>
<td>NINO1208 (0.51)</td>
<td>MDRSLP02 (0.43)</td>
<td>NAO03 (0.40)</td>
</tr>
<tr>
<td>$F_{3B}$</td>
<td>EPO02 (0.60)</td>
<td>MDRV20002 (0.60)</td>
<td>MDRSLP02 (0.57)</td>
<td>NINO1208 (0.51)</td>
<td>TSA02 (0.46)</td>
</tr>
<tr>
<td>$F_{3N}$</td>
<td>MDRSLP02 (0.66)</td>
<td>EPO02 (0.63)</td>
<td>MDRV20002 (0.57)</td>
<td>NINO1208 (0.51)</td>
<td>AMM01 (0.51)</td>
</tr>
<tr>
<td>$F_{3L}$</td>
<td>EPO02 (0.66)</td>
<td>LHF.WIN3(0.6)</td>
<td>MDRSLP02 (0.54)</td>
<td>NINO1208 (0.51)</td>
<td>MDRV20002 (0.46)</td>
</tr>
</tbody>
</table>

Outlook and ENSO JAS models, in all responses. The variables highlighted with two black vertical lines, represent the ENSO-related indices. Only the models for Atlantic tropical storms, hurricanes and Gulf of Mexico major hurricanes select an ENSO-related index as a covariate, and within these cases, these covariates get selected only in the ENSO JAS models (in 40% or more of the windows). This result is important in terms of how variables like EPO from January or February are selected more often and thus could contribute more with the model predictive skill. Furthermore, most of the ENSO-related indices are selected in the ENSO JAS model, which suggests that the important correlation between ENSO and number of storms is during the season, and not before (January - May), or that at least the other ENSO variables correlate with the counts in a similar way than the ENSO JAS indices.

Some predictors are selected consistently for some responses, across all $F$. For tropical cyclones in the Atlantic Basin (TS, HU, and MH), February EPO, MDRSLP and MDRV200 are selected throughout the models. The Caribbean region presents March TSA, February TNA and MDRSLP in each model within the three possible CAR responses (TS, HU, MH). Finally, February EPO and MDRSLP, along with August NINO12 are consistently selected in the models describing the Gulf of Mexico tropical cyclones.
4.4.2 Verification Results

The objective of the verification procedure is to compare how well each of the forecasts predicted the observed number of tropical storms, hurricanes and major hurricanes. The results are presented as a comparison between the skill score $\bar{H}$ of all the nine $F$s. Two climatologies ($\hat{F}_1$ and $\hat{F}_2$) are compared. Then, the best option is applied to a total of nine responses (TS, HU, MH for the three regions of the Atlantic Basin). There are three goals: the first is to compare the models generated using LASSO and those generated using only cluster for variable selection, the second one is to investigate whether or not to use forecasted ENSO data (March vs March with ENSO) and if a forecast released at the beginning of the season is better than one released in April using forecasted ENSO data (May vs March ENSO). This will evaluate how necessary the NINO forecast indices are in order to predict the number of storms well. Lastly, the third goal is to compare the best model against climatology and see if there is a significant improvement in the forecast.

With respect to the climatology scores, even though one can argue that the empirical version is more robust to misspecification of the distribution and that it sets a higher standard to compare model scores, it presents problems when the observed values are unique. Also, $\hat{F}_1$ is a fair comparison given that all the models are constructed assuming a Poisson distribution. Thus, to make sure the Poisson distribution is a realistic choice, a goodness of fit test is performed for all models, using the residual deviance. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. Therefore, if the residual difference is very small, the goodness of fit test will not be significant, indicating that the model fits the data. The results are presented in Table 4.4, where it is clear that the distribution is a good fit for all models. For this reason, $\hat{F}_1$ is chosen to be the reference, but it is important to highlight that the use of the empirical distribution is recommended in cases where several distribution choices are being compared.

When comparing two cross-validation options, it is important to highlight the need for
Table 4.4: Median p-values for deviance goodness-of-fit test per response. Null hypothesis is 
the Poisson model is a good fit for the data.

<table>
<thead>
<tr>
<th></th>
<th>$F_{1B}$</th>
<th>$F_{1N}$</th>
<th>$F_{1L}$</th>
<th>$F_{2B}$</th>
<th>$F_{2N}$</th>
<th>$F_{2L}$</th>
<th>$F_{3B}$</th>
<th>$F_{3N}$</th>
<th>$F_{3L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATL TS</td>
<td>0.67</td>
<td>0.75</td>
<td>0.65</td>
<td>0.86</td>
<td>0.94</td>
<td>0.89</td>
<td>0.88</td>
<td>0.92</td>
<td>0.86</td>
</tr>
<tr>
<td>ATL HU</td>
<td>0.77</td>
<td>0.86</td>
<td>0.84</td>
<td>0.87</td>
<td>0.90</td>
<td>0.85</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>ATL MH</td>
<td>0.66</td>
<td>0.67</td>
<td>0.74</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>CAR TS</td>
<td>0.24</td>
<td>0.22</td>
<td>0.39</td>
<td>0.60</td>
<td>0.63</td>
<td>0.55</td>
<td>0.61</td>
<td>0.59</td>
<td>0.58</td>
</tr>
<tr>
<td>CAR HU</td>
<td>0.52</td>
<td>0.57</td>
<td>0.55</td>
<td>0.72</td>
<td>0.74</td>
<td>0.78</td>
<td>0.73</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>CAR MH</td>
<td>0.53</td>
<td>0.52</td>
<td>0.48</td>
<td>0.55</td>
<td>0.59</td>
<td>0.54</td>
<td>0.58</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>GOM TS</td>
<td>0.84</td>
<td>0.82</td>
<td>0.83</td>
<td>0.90</td>
<td>0.82</td>
<td>0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>GOM HU</td>
<td>0.58</td>
<td>0.63</td>
<td>0.62</td>
<td>0.72</td>
<td>0.69</td>
<td>0.74</td>
<td>0.69</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>GOM MH</td>
<td>0.65</td>
<td>0.63</td>
<td>0.76</td>
<td>0.67</td>
<td>0.62</td>
<td>0.75</td>
<td>0.77</td>
<td>0.77</td>
<td>0.84</td>
</tr>
</tbody>
</table>

having a realistic validation method, that respects the temporal order of the observations. As 
mentioned before, evaluating forecasts using LOOCV or any hindcast method that uses future 
data to evaluate past forecast is not realistic, and thus not recommended in this case. Thus, 
since SWCV does not use future data to evaluate past forecasts and gives slightly different 
results than using LOOCV (see Figure 4.5), it is chosen as the best cross-validation method, 
given that is considered a more realistic scenario in terms of the forecasting procedure.

In section 4.3.3, the procedure to construct and validate the models was outlined. The 
variable selection procedure is explained in the second step, where a hierarchical clustering 
algorithm is used to select 10 primary covariates before applying LASSO. Before drawing con- 
clusions using the LASSO results, a comparison between the two sets of models with and 
without LASSO was done, to follow the principle of using the simplest best model. Results 
are presented in Figure 4.6 where there are some cases where the LASSO has superior skill 
compared to the cluster model ($\bar{H} < 0$). Given that there is at least one case where LASSO is 
significantly better in terms of skill, and no cases where $\bar{H} > 0$, the LASSO are used to draw 
conclusions in this case.

In order to identify those models that have a superior skill compared to climatology, another 
comparison was done using the average over $i$ from $H_i = LS(F_{clim}, y_i) - LS(F_{LASSO}, y_i)$. In 
this case, mean values of $H$ that are positive and significantly different from zero, denote models
Figure 4.5: In the left column, the difference between climatology and MSE for Atlantic Basin tropical storm models $F$ are presented for each model, and for SWCV and LOOCV. The percentage of change between MSE for model $F$ and climatology is presented on top of each observation. Positive values mean that model $F$ has lower MSE than climatology. Right column present the results for $H_i = LS(F_{clim}, y_i) - LS(F_{LASSO}, y_i)$ for each Atlantic Basin tropical storm model and SWCV and LOOCV. Intervals not overlapping zero represent significant difference between LASSO and cluster models. Positive values indicate superior skill of the cluster model, compared to LASSO.
with better skill compared to climatology. Figure 4.6 shows how there is not one model that performs significantly better than climatology. Thus, it is not meaningful to compare between models in each response. The conclusions in this case would be very different if LOOCV and percentage of improvement of MSE would have been used instead (Figure 4.5). As an example, if we calculate the mean percentage of improvement $PE = (MSE_{\text{clim}} - MSE_F)/MSE_{\text{clim}}$ for each $F$ in the Atlantic Basin tropical storms, we would have obtained that $F_{2B}$ had a 22.3% of improvement of skill compared to climatology. If the conclusions were drawn from here, most of the models for Atlantic Basin storms would be considered more skillful than climatology.

It is worth highlighting that in case of finding a set of models with superior skill for one response, the difference between scores of those models should be used as statistic, and its respective SE should be calculated using $\text{var}(H)$ formula, correcting for multiple testing with the significance level of $\alpha/c$ where $c$ is the number of tests using the same full model.

### 4.4.3 Use of method to verify past Outlooks

The data for 1990-2015 Outlooks was collected from the Colorado State University group (Klotzbach [2017]) since they are the only group with more than 10 years of forecasts. The data was collected taking into account those Outlooks released in April of each year (1996-2015) and in June for 1990-1995. The observed data and the climatology (average of the previous 40 years) are collected and the score $H$ is calculated with the formula $H_i = LS(F_{\text{clim}}, y_i) - LS(F_{\text{model}}, y_i)$ for year $i$, assuming a Poisson model, to then calculate its mean and SE using bootstrap with $B = 10000$. The results for the TS outlooks ($CI = (0.0252, 0.8464)$) are positive and $H$ is significantly different from zero, which means that the CSU model for TS is superior to climatology. Nevertheless, the other two models for HU ($CI = (-0.4326, 0.3885)$) and MH ($CI = (-0.3208, 0.4963)$) are not significantly different from zero and thus, they are not significantly better in terms of forecast skill compared to climatology.
Figure 4.6: First column is the mean $H_i = LS(F_{LASSO}, y_i) - LS(F_{cluster}, y_i)$ by response. Intervals not overlapping zero (in red) represent significant difference between LASSO and cluster models. Positive values indicate superior skill of the cluster model, compared to LASSO. Second column is the mean $H_i = LS(F_{clim}, y_i) - LS(F_{LASSO}, y_i)$ by response. Intervals not overlapping zero (in red) represent significant difference between climatology and LASSO. Positive values indicate superior skill of LASSO, compared to climatology.
4.5 Discussion

A reliable verification procedure is essential due to the economic and societal implications of releasing a poor forecast. A probabilistic verification method was presented, which incorporates a variable selection procedure into the method, using sliding window cross-validation (SWCV) to resample the data and logarithmic score (LS) to measure the prediction skill of the statistical seasonal hurricane forecasts. It is shown that LS for climatology gives similar results when using SWCV compared to LS using the traditional LOOCV, but how when using the combination of LOOCV and MSE as a verification method gives very different results as with the proposed method. In here it is argued that SWCV is a more realistic way to verify forecasts, and that the SE correction for autocorrelation between windows can be neglected in this case. SWCV and LS take into account the probabilistic nature of the forecast and provides a more realistic scenario, given that it does not use future data to evaluate past forecasts.

An important result is that using cluster analysis to preselect the variables in each model performs almost as well as using LASSO in terms of their logarithmic score. The verification method can be used to evaluate how necessary it is to use forecasted values and to evaluate any other set of covariates, or modeling decisions. The article shows how to implement the procedure and how to select possible variables from a set of covariates in the seasonal hurricane forecast context. This procedure – and the code available in the specified repository, provides a way to quantify and characterize the uncertainty in seasonal hurricane forecasts, to be used with any other set of covariates.
Bibliography


Appendices
Appendix A

Tables and Figures for Chapter 2

Figure A.1: MSE per location for each response: (a) Tropical Storms (TS), (b) Hurricanes (HU) and (c) Major Hurricanes (MH).
Table A.1: Proportion of locations with estimated coefficients statistically different from zero (credible intervals do not include zero). From those coefficients different from zero (+) indicates that more than 60% of the locations have a positive coefficient for that specific trimester, response and score. In a similar way, (-) indicates that more than 60% of the locations have a negative coefficient for that specific trimester, response and score. Bold values are significant factors. Red values are those displayed in Figures 2.4, 2.5, A.3 and A.4.

<table>
<thead>
<tr>
<th>Resp, Level</th>
<th>Score</th>
<th>Sea Surface Temperature</th>
<th>Latent Heat Flux</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>w = 1</td>
<td>w = 2</td>
</tr>
<tr>
<td>$Y_1 = TS, p$</td>
<td>$r = 1$</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td><strong>0.10</strong></td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td><strong>0.45</strong></td>
<td><strong>0.22</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td><strong>0.48</strong></td>
<td>0.01</td>
</tr>
<tr>
<td>$Y_1 = TS, \lambda$</td>
<td>$r = 1$</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td><strong>0.77</strong></td>
<td><strong>0.11</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td><strong>0.81</strong></td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>$Y_2 = HU, p$</td>
<td>$r = 1$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$Y_2 = HU, \lambda$</td>
<td>$r = 1$</td>
<td><strong>0.25</strong></td>
<td>+ <strong>0.23</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td><strong>0.35</strong></td>
<td>+ 0.08</td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td>0.07</td>
<td>- <strong>0.13</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td><strong>0.33</strong></td>
<td>- <strong>0.19</strong></td>
</tr>
<tr>
<td>$Y_3 = MH, p$</td>
<td>$r = 1$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td>0.06</td>
<td>- 0.03</td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$Y_3 = MH, \lambda$</td>
<td>$r = 1$</td>
<td><strong>0.17</strong></td>
<td>+ 0.04</td>
</tr>
<tr>
<td></td>
<td>$r = 2$</td>
<td>0.03</td>
<td>+ <strong>0.10</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 3$</td>
<td><strong>0.17</strong></td>
<td>+ <strong>0.10</strong></td>
</tr>
<tr>
<td></td>
<td>$r = 4$</td>
<td>0.07</td>
<td>- 0.07</td>
</tr>
</tbody>
</table>
Figure A.2: MSE per year for each response: (a) Tropical Storms (TS), Hurricanes (HU) and Major Hurricanes (MH).
Figure A.3: Summer: SST or LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.
Figure A.4: Fall: LHF Features (first left column) and their respective statistically significant coefficients. Warm colors (red, orange) are positive, and cold colors (purple, green) are negative. EOFs range from -0.2 to 0.2 and coefficients from -1 to 1.

Figure A.5: Example of simulation settings realizations for covariate 1.
Figure A.6: First four empirical orthogonal functions for sea surface temperature anomalies in the Atlantic Basin, per season.
Figure A.7: First four empirical orthogonal functions for latent heat flux anomalies in the Atlantic Basin, per season.
Figure A.8: Time series for each $\xi_{\ell,r}(t,w)$. 2 scores for SST are presented in the first row and 2 scores for LHF in the second row.
Figure A.9: Statistically significant factors per season ($w$), EOF ($r$) and covariate ($\ell$). First column represents SST coefficients, left column LHF coefficients. First row are coefficients that associate covariates with the occurrence of TS, second row with the occurrence of HU and third row are coefficients that associate covariates with the occurrence of MH.
Figure A.10: Statistically significant factors per season (w), EOF (r) and covariate (ℓ). First column represents SST coefficients, left column LHF coefficients. First row are coefficients that associate covariates with the intensity parameter of TS, second row with the intensity parameter of HU and third row with the intensity parameter of MH.
Appendix B

Definition of Climate Indices and Extra Figures for Chapter 4

- AMM (Atlantic Meridional Mode): The result of a maximum covariance analysis of SSTs and the zonal and meridional winds over the region 21S-32N, 74W-15E. This index is obtained from http://sahale.aos.wisc.edu:4080/MModes/Data/AMM.txt.

- AMO (Atlantic Multi-decadal Oscillation): An index based on North Atlantic SSTs. This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/amon.us.data

- TNA (Tropical Northern Atlantic): Anomaly of the average of the monthly SST from 5.5N to 23.5N and 15W to 57.5W. This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/tna.data

- TSA (Tropical Southern Atlantic): Anomaly of the average of the monthly SST from 0-20S and 10E-30W. This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/tna.data

- DM (Atlantic Dipole Mode): The difference between TNA and TSA SSTs.

- WHWP (Western Hemisphere Warm Pool): Monthly anomaly of the ocean surface area warmer than 28.5°C in the Atlantic and eastern North Pacific. This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/whwp.data

- NINO12: SST index in the Nino1+2 region. This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/nina1.data. To represent ENSO impacts, NINO12 values for the July/August/September average during the hurricane season were used in building the model. And the forecast values for NINO12 obtained from the National Center for Environmental Protection (NCEP) coupled forecast system (CFS) model were used for forecasts of the upcoming hurricane season from 1 June 31 November 2016.

- SOI (Southern Oscillation Index): This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/soi.data.

- NAO (Northern Atlantic Oscillation): This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/nao.data.

- EPO (East Pacific/North Pacific Oscillation): This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/epo.data.

- PNA (Pacific North American Index): This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/pna.data.

- AO (Arctic Oscillation): This index is obtained from http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/monthly.ao.index.b50.current.ascii.table.

- QBO (Quasi-Biennial Oscillation): This index is obtained from http://www.esrl.noaa.gov/psd/data/correlation/qbo.data.

- MDRSLP, MDRSST, MDRVWS: Monthly averaged sea-level pressure (MDRSLP), SST (MDRSST), and vertical wind shear (MDRVWS, wind shear between 200 and 850 mb), over the main development region (MDR, 10-20N, 80-20W). Data derived from the NCEP/NCAR Reanalysis I dataset by http://www.esrl.noaa.gov/psd/cgi-bin/data/timeseries/timeseries.pl?ntype=1&var=OLR&level=2000&lat1=20&lat2=10&lon1=-80&lon2=-20&iseas=0&mon1=0&mon2=0&iarea=0&typeout=1&Submit=Create+Timeseries

Figure B.1: Percentage of times a Variable X is selected using LASSO models, per response and for SWCV. In total, there are 35 windows (repetitions of the model). Nino variables are highlighted with black lines.
Figure B.2: Percentage of times a Variable X is selected using cluster models, per response and for SWCV. In total, there are 35 windows (repetitions of the model). Nino variables are highlighted with black lines.