2016

Space-time modelling of emerging infectious diseases: Assessing leptospirosis risk in Sri Lanka

Cameron C F Plouffe
Wilfrid Laurier University, plou7570@mylaurier.ca

Space-time modelling of emerging infectious diseases: Assessing leptospirosis risk in Sri Lanka

by

Cameron C.F. Plouffe

BSc (Hons) in Geography and Geomatics, Wilfrid Laurier University, 2009

THESIS/DISSERTATION

Submitted to the Department/Faculty of Geography and Environmental Studies

In partial fulfilment of the requirements for

Master of Science in Geography

Wilfrid Laurier University

Abstract

In this research, models were developed to analyze leptospirosis incidence in Sri Lanka and its relation to rainfall. Before any leptospirosis risk models were developed, rainfall data were evaluated from an agro-ecological monitoring network for producing maps of total monthly rainfall in Sri Lanka. Four spatial interpolation techniques were compared: inverse distance weighting, thin-plate splines, ordinary kriging, and Bayesian kriging. Error metrics were used to validate interpolations against independent data. Satellite data were used to assess the spatial pattern of rainfall. Results indicated that Bayesian kriging and splines performed best in low and high rainfall, respectively. Rainfall maps generated from the agro-ecological network were found to have accuracies consistent with previous studies in Sri Lanka. These rainfall data were then used as the primary predictor in a family of time series leptospirosis forecasting models at varying spatial scales across Sri Lanka. Several modelling scenarios were evaluated using proper scoring rules and numerous other metrics to assess model fit and calibration. A negative binomial integer-valued autoregressive conditional heteroscedasticity (INGARCH) model that included current and previous rainfall covariates, as well as regression on previous cases of leptospirosis at a local and seasonal time scale was selected as the best performing model. It was found that rainfall did not have a significant correlation with leptospirosis incidence in Sri Lanka, but the family of INGARCH models developed was able to forecast leptospirosis incidence and effectively provide early warning for leptospirosis outbreaks at the district level across Sri Lanka.

Acknowledgments

I would first like to thank my advisor Dr. Colin Robertson, and committee member Dr. Susan Elliott, for having a seemingly limitless amount of patience during my Master's thesis. I would also like to thank the Department of Meteorology in Sri Lanka for providing all rainfall data used to perform interpolations involved in this project, and the Ministry of Health in Sri Lanka for providing leptospirosis data that were used to model disease risk. I am grateful to the Social Sciences and Humanities Research Council (SSHRC) for providing necessary funding for this research. Lastly, I would like to thank Dr. Sam Daniel for helping to facilitate relations with numerous different organizations within Sri Lanka that were essential to this research.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1. Research context

An emerging infectious disease (EID) can be thought of as an infectious disease that has recently appeared in existing populations, or that has had its incidence rapidly increase in the recent past (Morse, 1995). Most EIDs are caused by pathogens that are already present in the environment, but changes in the underlying environmental conditions and in the human-environment relationship (e.g., land use change, immigration of human populations to previously uncultivated areas), can lead to the emergence or resurgence of such diseases (Mayer, 2000; Morse, 1995). Social, ecological, and geographical changes can all play an important role in the emergence or resurgence of infectious diseases, and given the increasing worldwide attention in recent decades that has been given to EIDs (e.g., AIDS, SARS) and the ways that they can affect society, it is important to continue to develop a better understanding of these drivers of emergence so that future EID outbreaks can be prevented (Mayer, 2000). The first step to prevention of the emergence or resurgence of infectious diseases is developing effective global disease surveillance systems (Morse, 1995). By developing surveillance systems for EIDs in key areas around the world, early warning of emerging infections or outbreaks can be had which can help to prevent and minimize future outbreaks before they become more global issues (Morse, 1995). Given the changing geography of the late 20th and early 21st century and increased mobility of human populations around the world, it is also important to analyse geographical aspects of emergence of infectious diseases (Haggett, 1994).

Drivers of the emergence of infectious diseases can be difficult to account for, as they are often a collection of many different social, ecological, economic, and environmental factors

(Mayer, 2000; Morse, 1995). Many of these factors are anthropogenic, as humans are perhaps the most important agents of ecological and environmental change, but natural changes in climate and weather can have just as pronounced an effect, and have typically been associated with the emergence of infectious diseases (Ashford et al., 2000; Mayer, 2000; Morse, 1995; Robertson et al., 2012; Vinetz et al., 2005). For example, with signs of increasing climate change such as rising global temperatures and varying trends in precipitation, it is thought that in the future, more drastic global environmental changes will occur (Jayawardene et al., 2005a; Pachauri et al., 2007). These natural environmental changes are likely to impact the emergence of infectious diseases and the risks they pose to human populations (e.g., increase in geographic range within which disease vectors can survive). To better understand the relationship between environmental change and EID incidence in human populations, work was done to correlate the two. Increased environmental variability influences the incidence of EIDs, and thus was an important aspect to consider when trying to understand the dynamics of emergence (Lau et al., 2010; Morse, 1995). By developing environmentally-driven forecasting models for disease risk and outbreak, progress was made in understanding the dynamics of EIDs.

Leptospirosis is a waterborne zoonotic disease of worldwide importance, as its incidence is continually increasing in developed and developing countries around the world (Vijayachari et al., 2008; WHO, 1999). Incidence of human infection tends to be higher in tropical areas and temperate regions (Bharti et al., 2003). Symptoms of leptospirosis are variable, and can include fever, headache, myalgia, nausea, and abdominal pain (Ashford et al., 2000). Severe and potentially fatal forms of leptospirosis can cause more adverse symptoms, and in recent decades, endemic and epidemic severe pulmonary haemorrhage has increasingly been identified as a symptom of leptospiral infection (Bharti et al., 2003; Levett, 2001). Leptospirosis incidence is

often underestimated due to lack of public awareness of the disease, and its symptoms being similar to other more well-known diseases (e.g., malaria) (Bharti et al., 2003). Human infection is caused by exposure to water that has been contaminated by the infected urine of carrier mammals (e.g., rodents, dogs) (Bharti et al., 2003). The occupation of an individual can often play a role in contracting leptospirosis – specifically, occupations which put an individual in contact with animal reservoirs or occupations that involve increased contact with potentially contaminated water (e.g., farming and agricultural work) will put one at greater risk (Levett, 2001).

Leptospirosis is known to have an association with environmental variables, and thus it was a suitable case for developing forecasting models for EID incidence (Ashford et al., 2000; Sarkar et al., 2012, 2002; Vinetz et al., 2005). Several studies have assessed the effects that different environmental variables (e.g., temperature, rainfall) have on leptospirosis transmission and incidence (Ashford et al., 2000; Chadsuthi et al., 2012; Pappachan et al., 2004). While there is an observed relationship between environmental factors (e.g., changes in precipitation and temperature dictated by seasonality) and leptospirosis, this research aimed to develop a more acute understanding of the dynamics of this relationship and as a result, contributions were made to the fields of epidemiology and EID modelling, and more specifically, to leptospirosis research. This research considered an outbreak of leptospirosis in Sri Lanka as a case study. Modelling methodologies were analyzed for determining correlation between environmental factors and leptospirosis incidence. Through the use of geographic information systems (GIS) and spatial analysis, relationships were investigated in a spatial context.

2. Research questions and objectives

Developing an understanding of the relationship between EID incidence and environmental factors is required to forecast where new EIDs will emerge and spread. While many studies have speculated certain patterns pertaining to correlation between precipitation and EID incidence, most agree that research must continue to be performed – specifically regarding leptospirosis – to further clarify the correlation (Ashford et al., 2000; Chadsuthi et al., 2012; Lau et al., 2010; Pappachan et al., 2004; Sarkar et al., 2012, 2002; Zhang et al., 2008). I developed leptospirosis risk models (i.e., models that forecasted leptospirosis incidence) and assessed them in a spatial context by taking into account information and data specific to the underlying landscape to elucidate mechanisms of transmission. The primary goal of this research was to improve and further understanding of EID outbreak, and specifically, leptospirosis outbreak, by examining how environmental drivers affect the spatial and temporal distribution of the disease. This research was conducted by answering two primary research questions.

2.1 **Can spatial interpolation techniques be employed to effectively predict precipitation across Sri Lanka?**

When constructing models that rely heavily on climate variables, using large-scale climate data sets can help models yield desirable and realistic results, as incorporating climate data specific to the study area into models allows for the relationship between what is being modelled and the underlying climate to be more accurately defined. Data are increasingly becoming available due to improvements in different measurement technologies such as remote sensing, but most climate and rainfall data are still collected by networks of permanent and irregularly dispersed weather stations. To assess leptospirosis incidence and its correlation to precipitation, I predicted rainfall across all of Sri Lanka by generating continuous surfaces of

rainfall. Several spatial interpolation techniques were assessed in an effort to best approximate rainfall in Sri Lanka. Rainfall was then included as a primary predictor for modelling leptospirosis incidence in Sri Lanka.

Spatial interpolation methods including inverse-distance weighting, thin-plate smoothing splines, ordinary kriging, and Bayesian kriging were evaluated by applying them to a rainfall data set received from the Department of Meteorology of Sri Lanka. Daily precipitation data were obtained from a network of weather stations distributed across Sri Lanka from 2005 to 2011, but there were considerable gaps between weather stations in the data set. Spatial interpolation was performed on this data set to first, test whether rainfall data values could be accurately predicted, and second, compare the interpolation methods against each other. From this, conclusions were drawn as to which method was the most effective at predicting rainfall in Sri Lanka. Once accurate rainfall data were predicted, they were used to model leptospirosis incidence in Sri Lanka.

2.2 Does precipitation data provide a reliable early-warning signal for leptospirosis outbreaks in Sri Lanka?

To investigate the relationship between precipitation and outbreak events of leptospirosis, a family of local independent time-series models was explored to analyze and clarify the effect that precipitation has on leptospirosis incidence in Sri Lanka. Using these models, predictions were made concerning areas and districts of likely leptospirosis outbreak. The predictions were then evaluated to assess whether the models constructed could be used to provide a reliable early warning of leptospirosis outbreak in Sri Lanka, and if rainfall and wetness of the physical environment were significant predictors for forecasting leptospirosis risk.

3. Contributions

This research helped to further understanding of the effect that environmental drivers have on the distribution of emerging infectious diseases. Methods that have previously not been applied to assessing leptospirosis in Sri Lanka were used in an effort to understand the spatial and temporal dynamics of disease outbreak. Contributions were made to leptospirosis research, and more generally, the spatial and temporal analysis of zoonotic diseases (i.e., diseases that can be transmitted between animals and humans). Methodological improvements in the field of epidemiology were made through the use of families of local independent time-series models, which have not previously been applied to evaluating leptospirosis risk.

There has been an abundance of research done in the past on factors which influence the outbreak of zoonotic diseases. While anthropogenic factors can be important drivers of disease transmission, environmental factors can carry just as much weight, as most of the anthropogenic factors are affected by these environmental factors (Morse, 1995). By conducting this research and developing empirically-driven models that incorporate environmental variables to forecast leptospirosis incidence in Sri Lanka, understanding was expanded by more acutely exploring the links between rainfall and leptospirosis incidence.

The country of Sri Lanka can benefit greatly from the findings of this research. The forecasting models for leptospirosis risk in Sri Lanka can be used to alert the Sri Lanka Ministry of Health of possible upcoming leptospirosis outbreaks. With this information, early-warning protocols can be developed and implemented to help reduce the extent of future outbreaks, and prepare the population if a potential leptospirosis outbreak is expected given the underlying environmental conditions. Examples of possible methods of outbreak prevention include

improving sanitation measures, administering leptospirosis vaccines, and administering equipment to individuals in regions of heightened leptospirosis risk to minimize human contact with contaminated water (e.g., waterproof boots and gloves) (Bharti et al., 2003). Understanding drivers of leptospirosis outbreak is relevant and necessary in a developing country like Sri Lanka, as this type of information is not generally available, and potential for large-scale outbreak is high due to the tropical environmental conditions present.

Chapter 2: Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka

An edited version of this paper was published in the journal Environmental Modelling & Software:

Plouffe, C.C.F., Robertson, C., Chandrapala, L., 2015. Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka. *Environ. Model. Softw.* 67, 57–71. doi:10.1016/j.envsoft.2015.01.011

1. Introduction

Ecological forecast models that rely on climate data are increasingly used in a variety of contexts. For example, detailed climate data are necessary when modelling outbreak patterns of emerging infectious diseases (Briët et al., 2008; Robertson et al., 2012). While data are increasingly becoming available due to the advent of smaller and cheaper environmental sensors, most climate data – specifically, precipitation data – are still collected by a network of geographically dispersed weather stations. This leads to data that contain considerable gaps in coverage of areas where stations are more isolated. However, additional data sources such as citizen sensors (Goodchild, 2007), unofficial and/or semi-official networks of rain gauges (Wickramaarachchi et al., 2013), and satellite-derived data products (Kummerow et al., 1998) may be used to augment estimates from ground-based stations. To leverage these auxiliary sources of data, new approaches are required to integrate data from multiple sources, and to help evaluate the best performing models (Bennett et al., 2013). In this paper, I investigate the integration of additional sources of data for comparison of rainfall interpolation methods in Sri Lanka. Firstly, I aim to evaluate an unofficial network of rain gauges across Sri Lanka by comparing different interpolations against official meteorological station recordings. Secondly,

as part of the unofficial rainfall station network validation, I examine spatial patterns in predicted rainfall in relation to satellite-derived estimates of rainfall.

Spatial interpolation techniques are widely used to estimate seamless spatial coverage of rainfall over large areas, yet there is little consensus on the optimal interpolator for rainfall, especially where spatial rainfall pattern is highly variable (Dirks et al., 1998; Price et al., 2000; Vicente Serrano et al., 2003). Table 2.1 displays a summary of several different studies evaluating interpolation methods applied for rainfall prediction in different settings. Previous studies have come to different conclusions regarding the most effective techniques for spatial interpolation of rainfall data, and more generally, measuring performance for any given environmental model is intrinsically case-dependent (Bennett et al., 2013). Robson (2014) suggests that opting for the simplest model possible is desirable unless it has been found to be inadequate when compared to more complex models. The literature reveals that accuracy of precipitation interpolation varies greatly by region and temporal scale (Table 2.1). Interpolation errors are related to measurement error, the density of the station network, topography, and the type of rainfall (Abtew et al., 1993). Tropical and monsoonal environments in particular have proven difficult to characterize with seamless spatial coverages of rainfall (Jayawardene et al., 2005b)(Jayawardene et al., 2005). The amount of rainfall in the tropics is often highly variable in intensity and seasonality (Malhi and Wright, 2004), and interpolating rainfall for these areas can be quite difficult, as weather often dramatically changes over space and time.

The primary differences between the statistical methods used to interpolate rainfall are how they are conceptually formulated and mathematically constructed (Burrough and McDonnell, 1998). Some approaches to spatial interpolation are more effective at predicting certain types of spatial processes, and thus context-specific applications of interpolation methods

are common. Comparative studies have been conducted to determine which method of spatial interpolation is best suited for different contexts, but as of yet, no decisive conclusions have been made (Zimmerman et al., 1999). It is important to continue research in this direction to gain a better understanding of proper applications of these interpolation techniques.

The tropical country of Sri Lanka is used here as a case study for this exploration of rainfall interpolators. It was hypothesized that one of the geostatistical methods would yield the most accurate results. A review of relevant literature found that kriging is the most effective interpolation method for precipitation data (Jeffrey et al., 2001; Vicente-Serrano et al., 2003; Zimmerman et al., 1999).

1.1 **Objectives**

The objectives of this research were three-fold. Firstly, I aimed to determine the most effective spatial interpolation methods for rainfall data for application to countrywide environmental modelling in Sri Lanka. Specifically, I required a methodology for estimating seamless spatial coverage of monthly precipitation. While the focus here is Sri Lanka, I aim to add to the literature on interpolation comparisons, with specific emphasis on tropical areas that exhibit large variability in rainfall throughout the year. To investigate this, four different spatial interpolation methods were evaluated: inverse distance weighting (IDW), thin plate smoothing splines, ordinary kriging, and Bayesian kriging. These methods were chosen on the basis that many studies in the past have employed these techniques in rainfall interpolation (Daly et al., 1994; Dirks et al., 1998; Jeffrey et al., 2001; Oke et al., 2009; Vicente Serrano et al., 2003). The results of these comparisons will be used as input for a spatial-temporal model used for surveillance and forecasting of waterborne infectious disease risk in Sri Lanka. The second

objective was evaluate the suitability of a novel source of data, community managed weather stations which form an agro-ecological monitoring network, as a source of data for mapping rainfall over the whole country. Currently in Sri Lanka, these stations are not used for modelling rainfall at the country scale. Additionally, given that the infectious disease being investigated (i.e., Leptospirosis) tends to be of greater risk to human populations in agricultural areas, accuracy of interpolated rainfall values in these areas was an underlying research goal. Finally, I aimed to investigate the use of a novel map comparison method, structural similarity (SSIM) index, to evaluate the spatial structure of interpolated rainfall maps derived from station readings.

To meet these objectives, analysis occurred in three distinct stages. Firstly, I interpolated rainfall maps for each of the four methods. Secondly, I used official meteorological station ground truth data to compare errors of the interpolation methods at different times of year. The validation data were independent from the data used for model development, providing an objective assessment of map accuracy. Finally, I employed the SSIM index to compare spatial rainfall patterns obtained from satellite imagery over corresponding locations and times. By evaluating the quality of the rainfall maps using multiple criteria (i.e., across interpolators, relative to independent data, and compared to satellite imagery), I hypothesized that I could determine the strengths and weaknesses of each interpolation method when dealing with the high climatic variability present in tropical climates.

2. Material and methods

2.1 **Interpolation**

Spatial interpolation methods can be grouped into four categories: global methods (trend surfaces and regression models), local methods (Thiessen polygons, IDW, and splines), geostatistical methods (kriging), and mixed methods, which involve a combination of all of the previously listed methods (Vicente-Serrano et al., 2003). While these different types of methods present viable options for interpolation, only compared local and geostatistical methods were compared. These were chosen based on a number of factors, including computational complexity, ease of implementation in an operational forecasting system in Sri Lanka, the capabilities of the software being utilized, the size of the data sets being studied, and a thorough review of recent literature. Given these criteria, IDW, splines, ordinary kriging, and Bayesian kriging were chosen as the appropriate spatial interpolation techniques.

Interpolation techniques predict the variable of interest at a specific location by taking known values from the surrounding region into account, and using them to estimate the value at a location where it is unknown. Normally, prediction at a location using interpolation can be expressed generally by the following formula:

$$\hat{z}(s_i) = \sum_{i=1}^{n} f(s_i) + \epsilon(s_i) \quad (1)$$

Where $\hat{z}(s_i)$ is the estimated value at location s_i, f is a function specific to the particular interpolation technique that takes in known values to make a prediction at location s_i, and $\epsilon(s_i)$ are the random errors associated with that particular location. Based on the type of interpolation

being performed, the values predicted at location s_i will differ. A brief description of the more commonly used interpolation methods that were evaluated is outlined in Table 2.2.

Bayesian kriging differs from ordinary kriging in that prior distributions are put on parameters of the semivariogram, and estimation yields a posterior distribution for each of the semiovariogram parameters (range, sill, and nugget). Prior distributions allow inclusion of expert knowledge and uncertainty into the estimation procedure and model outputs. Yet typical implementations of Bayesian kriging do not include informative prior distributions (Berger et al., 2001). I modelled rainfall as a realization of a Gaussian random field, as is common in environmental applications of Bayesian kriging, such that

$$z_i|S \sim N(\beta(z_i) + S(z_i)), \tau^2) \tag{2}$$

where rainfall z_i is a linear combination of spatial trend β, a Gaussian process S (Diggle and Ribeiro, 2002) and the nugget variance τ^2. The Gaussian process is

$$S(z_i) \sim N(0, \sigma^2 R(h; \phi, \mathrm{K}) \tag{3}$$

where R is specified as a Matérn covariance function for spatial lag h, correlation parameter ϕ, smoothness parameter K. The full parameter vector for the Bayesian kriging model was therefore $[\beta, \tau^2, \sigma^2, \phi, K]$. Samples from the posterior distributions were obtained from simulation (Diggle and Ribeiro, 2002).

Since full posterior distributions are available for inference, parameter uncertainty was incorporated into the spatial predictions of rainfall. As the initial rainfall values used to perform ordinary and Bayesian kriging were not Gaussian, a Box-Cox transformation was performed on the data set to satisfy model assumptions. The interpolated data were later back-transformed for

analysis and interpretation. The R package geoR (Diggle and Ribiero, 2007) was used to generate spatial predictions from the fitted model.

2.1.1. Model parameters

Parameters for each interpolation model were set based on visual inspection and qualitative analysis of multiple test interpolations in high and low rainfall scenarios. Particular attention was given to not overfitting the models to the test data. The purpose of this research was to determine a suitable interpolation model for predicting monthly rainfall in Sri Lanka year round, and considering the nature of the data being interpolated, it would be an oversight to fit the models too closely to only the months used for analysis (i.e., May and November).

IDW yielded the most accurate predictions by taking into account n = 12 nearest neighbours and using an inverse distance power of $n = 1$. Jayawardene et al. (2005) found an inverse distance power of $n = 1$ attained the highest correlation coefficient and the lowest RMSE of all tested values of n for IDW interpolation of rainfall in Sri Lanka. Thin-plate smoothing spline parameters include the order of the polynomial expansion, and the smoothing parameter. The smoothing parameter was chosen by generalized cross validation and polynomial was second order based on experimentation and literature review (Alvarez et al. 2014). A spherical semivariogram was used to fit the model for ordinary kriging, which had a range of 15 km, and a cutoff of 100 km. A uniform prior distribution was used for parameters of the semivariogram for Bayesian kriging, and a Matérn covariance function was used to define the spatial correlation structure.

2.2 **Study area**

Sri Lanka is situated in the Indian Ocean, off the southeastern tip of the Indian subcontinent. The climate is tropical, and weather is characterized by two seasonal monsoons. The northeast monsoon typically lasts from December to February, while the southwest monsoon lasts from April until September. The southwest area of Sri Lanka receives significant rainfall particularly during the southwest monsoon season, while the northern and eastern regions of the country become predominantly dry during this time. There are also two inter-monsoonal seasons, which last from March to April, and October to November. During these inter-monsoonal seasons, Sri Lanka can experience considerable amounts of convectional rainfall.

2.3 **Data**

2.3.1. Model development data

Rainfall data were collected from the Department of Meteorology of Sri Lanka, and include daily rainfall measurements (millimeters) from a network of ~370 small-scale community-managed weather monitoring stations, many of which were located in agricultural areas (Figure 2.1). The spatial distribution of the station network varies considerably with population, climate, and land use. Daily rainfall measurements were aggregated into total monthly rainfall. Multiple subsets of these data were extracted for the months of May and November for the years of 2006 through 2010. These months were chosen as they coincided with periods of peak rainfall for Sri Lanka (Zubair, 2002). The quality of the data collected at stations was largely unknown, as many stations were located in remote areas, and maintenance of each station was situation-dependent.

2.3.2. Model validation data

Rainfall validation data were obtained from 20 to 22 (dependent on year) official meteorological stations managed by the Department of Meteorology of Sri Lanka (Figure 2.1). The data set was aggregated into monthly rainfall for May and November of 2006 through 2010. The official meteorological station data (n=~20) were used to validate the interpolations generated using the larger (n=~370, varying by year and month) community-managed weather monitoring data set.

2.3.3. Spatial structure evaluation data

While weather station data are accurate estimates of local rainfall, 'ground truth' of spatial rainfall patterns across large areas is difficult to assess from point observations. To this end, satellite-derived rainfall maps were obtained to represent the general spatial pattern of rainfall, which were then compared to interpolated surfaces. Remotely sensed hourly rainfall estimates for May and November of 2006 through 2010 were acquired from the Tropical Rainfall Measuring Mission (TRMM), a collaborative mission between the National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA) designed to monitor and study tropical rainfall. The GSMaP_MVK rainfall product for global hourly precipitation was used, which employs a Kalman filter to estimate surface rainfall rates at a 0.1° latitude x 0.1° longitude resolution by incorporating data from LEO microwave and GEO infrared radiometers (Ushio et al., 2009).

2.4 **Accuracy assessment**

2.4.1. Evaluation metrics

To evaluate which interpolation methods generated the most accurate rainfall predictions, a number of accuracy assessment metrics were employed. Comparisons were made between the rainfall values obtained from the 20 official meteorological stations, and each of the

interpolations' predictions at those same locations. Mean absolute error (MAE), median percent error (MdPE), and standardized root-mean-square error (SRMSE) were used to evaluate interpolations for May and November of 2006 to 2010.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$\text{MdPE} = \text{median}\left(\frac{|y_i - \hat{y}_i|}{\hat{y}_i}\right) \times 100 \tag{5}$$

$$\text{SRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{6}$$

The SRMSE was calculated by taking the RMSE of interpolations' predicted values compared to observed values, and standardizing by mean rainfall for that particular month and year. It provides a dimensionless measure which is beneficial for comparing values between data sets with different means and has been used in similar analyses (Chemel et al., 2011). Statistical error (SE), which simply involves subtracting the known values from the predicted values, was used to evaluate interpolated values from 2007 and 2010.

$$\text{SE} = \hat{y}_i - y_i \tag{7}$$

MAE, MdPE, and SRMSE were used when evaluating the results averaged over the 20 station locations to check for yearly trends, while SE was used to evaluate rainfall at the individual station level to investigate regional trends. Observed vs. predicted error plots were also created for all interpolation methods for the years of 2007 and 2010.

2.4.2. Justification of metrics

All evaluation metrics were chosen to evaluate model performance, and account for inadequacies of other individual metrics. In a paper by Bennett et al. (2013) regarding characterization of model performance, it was stressed that metrics should be chosen to compliment the weaknesses of the other metrics being employed. MAE was chosen as a measure of overall accuracy for the interpolation models. While RMSE is a more common metric to measure accuracy, it gives greater weight to extreme outliers present in the results. Given the uncertain quality of the data used to fit the models, giving less weight to extreme outliers helped determine which models performed consistently well and were not affected as adversely by stations with erroneous data. MdPE was used in conjunction with MAE as an alternative measure of overall accuracy. MdPE was used to further reduce the impact of extreme outlier errors that could considerably skew MAE depending on the magnitude of the outlier. Using a percentage as opposed to a unit of measure (e.g., mm) also gave perspective on the magnitude of the error for the given month when paired with the MAE.

SE was used to account for station-level bias for 2007 and 2010 – years with the lowest and highest mean rainfall, respectively. Stations were geographically plotted from north to south to monitor spatial trends in the interpolations. SE indicated whether errors at the station level were consistently positive or negative, which helped reveal possible issues with the underlying data used to fit the models. SRMSE allowed for easier cross-seasonal comparison of model fit. The SRMSE metric was standardized by mean monthly rainfall for each month being analysed, which provided a metric independent of the magnitude of rainfall for any given month of study. SRMSE was especially useful when comparing results from years with substantially different amounts of rainfall.

2.5 **Spatial structure evaluation**

Using accuracy assessment statistics is valuable when looking to measure the accuracy of results at specific locations, but these metrics fail to account for global spatial trends that are present within the data. In a recent review of model performance evaluation by Bennett (2013), the notion of evaluating 'data pattern' was highlighted as an important aspect of model performance. Typically evaluating pattern preservation is done a-spatially, such as the correlation coefficient or more recently for temporal data methods to estimate curve similarity have been proposed (e.g., Ehret and Zehe, 2011). I extend these ideas to the spatial domain to compare the spatial similarity of rainfall maps using the SSIM index (see Robertson et al., 2014 for details). Briefly, by comparing interpolated rainfall maps as a whole to other rainfall maps of the same study region, it can be determined whether spatial patterns of rainfall are being accurately predicted. Map comparisons were performed between the generated interpolations and TRMM remotely sensed rainfall estimates by employing the structural similarity index (SSIM), a quality assessment methodology originally intended to assess the quality of image compression algorithms (Wang et. al, 2004). Hagen-Zanker (2006) later suggested this method for assessing the structure of continuous maps. SSIM was employed to extend this notion to comparing the differences between multiple interpolation methods' outputs and remotely sensed rainfall estimates. By comparing differences between the spatial patterns and structure of the interpolations and the remotely sensed rainfall estimates, an assessment of interpolation quality at the pattern-level could be made. SSIM takes three components into account for map comparison: luminance, contrast, and structure, concerning local differences in mean, variance, and correlation, respectively (Wang et al. 2004). For this research, only structure will be assessed:

$$\sigma_a = \left(\sum_{i=1}^{n} w_i (a_i - \mu_a)^2 \right)^{\frac{1}{2}} \quad (8)$$

$$\sigma_{ab} = \sum_{i=1}^{n} w_i (a_i - \mu_a)(b_i - \mu_b) \quad (9)$$

In these two equations, *a* and *b* represent raster maps, the index *i* iterates through *n* cells in a set region, and w_i are the spatial weights that control the smoothness of the local region effect.

$$S(a,b) = \frac{\sigma_{ab} + c_3}{\sigma_a \sigma_b + c_3} \quad (10)$$

The formula above represents one of the three components that the SSIM is comprised of; structure (*S*). In the formula, the constant c_3 is used for stability in situations where the mean or variability is close to zero which would be the case with large homogeneous patches. The component *S* ranges from -1, to 1, indicating a negative or positive correlation coefficient between cells in each window. Interpreting maps of *S* values allows for analyzing local spatial patterns. Where structure is high, the spatial pattern in values will be similar; even if the magnitudes of the pixels are dramatically different. The SSIM and specifically the structure component provide a novel methodology for evaluating continuous maps. It should be noted that the original TRMM rasters were recorded at a 0.1° latitude x 0.1° longitude resolution. These rasters were converted to a 5 km x 5 km resolution using bilinear interpolation such that comparisons could be made between the rainfall interpolations and the TRMM satellite based rainfall estimates.

2.6 **Software**

Several different types of software were used for data management and processing during this project. The programming language Python was used for data preprocessing – specifically, parsing the unformatted data sets. Python was also used to download remotely sensed hourly

rainfall estimates acquired from TRMM for the months being studied. The statistical programming language R (version 2.14.2) was used for a variety of tasks, including processing the official meteorological station data, aggregating the data from daily into monthly rainfall, accumulating the TRMM remotely-sensed hourly rainfall estimates into monthly rainfall, and generating the IDW, spline, ordinary kriging, and Bayesian kriging interpolations. The R packages gstat (Pebesma, 2004), fields (Nychka et al., 2012), and geoR (Diggle and Ribiero, 2007) were used to perform all interpolations. Figure 2.2 depicts a flow chart of the entire workflow taken throughout this research.

3. Results

Mean yearly rainfall of both official meteorological stations and community-managed weather stations for May and November of 2006 to 2010 is presented in Figure 2.3. Both meteorological stations and community-managed weather stations exhibited similar trends of mean rainfall by year. While the magnitude of rainfall tended to be slightly greater for the community-managed stations, peak rainfall years of 2006 and 2010 were present for both station networks. Generally, the yearly patterns for the month of May paralleled those for November, with the major difference being the larger magnitude of rainfall always present in November (as would be expected given the typical seasonal distribution of rainfall in Sri Lanka). An unusually low mean rainfall was found for the year of 2007, specifically when looking at November, which will be considered as a dry year in the quantitative analysis. The years of 2007 and 2010 will be analysed specifically, as they represent years with the minimum and maximum mean rainfall, respectively.

3.1 **Accuracy assessment**

Table 2.3 displays MAEs, MdPEs, and SRMSEs for May and November from 2006 to 2010 for all interpolations methods utilized in this study. Overall, moderate MAEs were found for all interpolation methods, while MdPEs were found to be much more varied, specifically for the month of May. The SRMSEs were perhaps more indicative of the actual predictive error associated with each interpolation technique, as they were standardized for each month and year. These SRMSEs were found to be relatively moderate in range. For the interpolations concerning the month of November, the MAEs were noted to be considerably larger (occasionally double the size) than those concerning the month of May, but this was directly related to the magnitude of rainfall being experienced in November, as the MdPEs were found to be lower for November than May. The SRMSEs were consistent in magnitude regardless of month. The differences found between the MAEs and MdPEs were thought to be related to how much the rainfall measurements varied given the month being assessed, and extreme station readings dramatically affecting the MAEs. While the majority of Sri Lanka received significant rainfall during the peak rainfall month of November, the north of Sri Lanka became dry during May.

MAEs for the month of May ranged anywhere from 20.81 mm to 87.92 mm given the method of interpolation, while MdPEs ranged between 17.58% and 69.55%, and SRMSEs ranged from 0.285 to 0.578. Overall, May experienced much more varied rainfall (relative to the norm for that month) than November, and this can be accounted for when evaluating the error metrics. When looking solely at the yearly MdPEs, none of the interpolation methods performed particularly well, with Bayesian kriging performing the best of all methods, with an average MdPE of 33.32% over the five year study period, and thin plate smoothing splines performing the worst, with an average MdPE of 41.93% over the same 5 year span. The SRMSEs depicted a

different trend for May, as while Bayesian kriging performed the best for two of the five years (2007 and 2008), it proceeded to perform the worst for all other years studied. The two years that Bayesian kriging attained the lowest SRMSEs coincide with the two years that attained the lowest MAEs, and the two years that had the lowest mean yearly rainfall. No method consistently outperformed all other methods, as results were largely dependent on the mean rainfall for that year. In years with low rainfall, Bayesian kriging performed well, while in years with high rainfall, the local interpolation methods (IDW and splines) attained the lowest errors.

The interpolations produced for the month of November generally had larger MAEs than those attained from May, but they also demonstrated much lower MdPEs. Interestingly, the SRMSEs were of similar magnitudes of those found for May, with only one year (2010) where November SRMSEs were much lower than their corresponding SRMSEs for May (approximately half the size of those found in May). Of all the methods tested, thin plate smoothing splines and IDW performed the best, with an average MdPE over the 5 year study period of 14.50% and 14.11%, and an average SRMSE of 0.307 and 0.314, respectively. Interpolations from November of 2010 (Table 2.3) had the lowest MdPEs and SRMSEs for any month of any year in the study period.

3.2 **2007 and 2010 accuracy assessment**

The years of 2007 and 2010 exhibited the lowest and highest mean rainfall, respectively. To assess which interpolation methods produced the best results given the underlying conditions, these years were selected for in-depth analysis, as they demonstrate some of the most extreme conditions that occurred during the study period.

The year of 2007 had the lowest mean rainfall when taking into account both months being studied. For both May and November of 2007, Bayesian kriging produced the lowest errors when evaluated by all three error metrics (MAE, MdPE, SRMSE). It is unlikely that this is a coincidence, considering all three error metrics were in concordance. For May of 2007, Bayesian kriging attained an SRMSE of 0.360, which was lower than all three other interpolation methods tested by at least 0.044. When evaluating the plots of observed vs. predicted rainfall for 2007 (Figure 2.4), certain trends can be identified. May of 2007 tended to have slightly lower predicted rainfall than observed rainfall for all four methods. All error plots for 2007 tended to be heteroscedastic, with a general trend of increasing variance as rainfall values increased. Of all the error plots analysed for either of the years being focused on, the plots from November of 2007 seemed to depict the least linear trend, as the data were relatively isotropic for some interpolation methods. Specifically, stations where between 200 and 300 mm of monthly rainfall was observed had weak correlation to the predicted values of all interpolation methods.

Plots of statistical error (SE) by station (Figure 2.5) revealed interesting regional trends in the interpolations. Each plot depicts how much each interpolation method differs from the official meteorological station value at each station, sorted from south (station 1) to north (station 20). For May of 2007, generally, SEs in the southernmost stations were much greater than those found in the north. All SEs were within +/- 110 mm of the actual observed rainfall data at each meteorological station. Most of the sizeable SEs in the south of Sri Lanka were overestimations by the interpolation techniques, with only one station (station 4) demonstrating a large underestimation of rainfall. Ordinary kriging tended to predict the most extreme values when compared to other interpolation methods. Overall, all interpolation methods followed similar

trends with regard to positive and negative predictions, with the main difference between methods being the magnitude of the errors found.

November of 2007 also had its highest SEs present in the southernmost stations of the Sri Lanka. As stations locations moved north, the SEs associated with the interpolations at those locations decreased. SEs were +/-180 mm from the actual observed value, which were slightly larger than those found for May. The SEs can likely be attributed to greater magnitudes of convectional rainfall present during the second inter-monsoonal period, and the start of the northeast monsoon. In November, the stations located in the north of Sri Lanka experienced SEs larger than those exhibited in May, as the stations received little to no rainfall in May, resulting in very low SEs. The largest SE present was at station 1, where all interpolation methods underestimated rainfall at that location. Again, all interpolation methods followed similar trends of positive and negative SE, with only a few minor exceptions.

2010 had the highest mean rainfall for both months of all the years that were studied. Thin plate smoothing splines attained the lowest errors for both months for the majority of the error metrics. The only metric which did not have splines receiving the lowest error was the MdPE for May of 2010. The SRMSEs for both May and November of 2010 were considerably lower for splines than any other method. For May, the SRME for splines was 0.318 – 0.056 lower than any of the other interpolation methods, while for November, the SRMSE for splines was 0.142 (the lowest SRMSE for all years and months being assessed). The observed vs. predicted rainfall plots for 2010 (Figure 2.4) showed more concordance between the predicted and observed values than those of 2007. In particular, the error plots for November of 2010 depict quite homoscedastic trends for all interpolation methods except Bayesian kriging, which contained one station location that was greatly overestimated. Anisotropy was also observed in the November error

plots, and to a lesser extent, for May. Overall, there was stronger correlation between observed and predicted values for 2010 than present in 2007.

The plots of SE by station (Figure 2.5) are interesting to evaluate for 2010. It should be noted that there were two more official meteorological stations recording rainfall available in 2010 that were used to evaluate the SEs. Compared to 2007, the magnitudes of the SEs were much greater, but this was largely due to the overall higher rainfall associated with 2010. For May of 2010, the same trend present throughout all of the SE by station plots was apparent; SEs tended to be much larger in the south, while as station locations moved north, the SEs decreased. The SEs covered a range of +/- 500 mm of the actual observed values at each station location, with Bayesian kriging often predicting the most extreme values where there was error among the interpolation methods. Most SEs were underestimates of the rainfall observed, with one notable exception being station 8, which exhibited large overestimation of rainfall by all of the interpolation methods. All interpolation methods again followed similar trends with regard to positive and negative predictions.

Plots of SE by station for November of 2010 were quite different than any of the other SE plots, as the SEs did not decrease as the station locations moved north. Instead, all SEs were reasonably consistent from south to north, with one particular area of Sri Lanka exhibiting much larger errors than the surrounding regions. Station 8 had very large SEs that were both positive and negative depending on the interpolation method. Bayesian kriging – which overestimated rainfall by 537.89 mm, the highest prediction error from all SE plots for both years – and IDW both had positive SEs, while splines and ordinary kriging had negative SEs. It should be noted that station 8 had the most extreme SEs for both months studied in 2010.

Spatial maps generated from each interpolation method (Figure 2.6) help visualize some of the spatial patterns of rainfall in Sri Lanka for 2007 and 2010. Immediately, it can be noticed that the vast majority of Sri Lanka receives little rainfall in May.

While May is considered a peak rainfall month in Sri Lanka due to the combination of convectional rainfall from the first inter-monsoonal season and the onset of the southwest monsoon, only the southwest corner of the country experiences significant rainfall, which demonstrates the stark contrast between May and November's rainfall distribution (Zubair, 2002). For May of 2007, Bayesian kriging (which attained the best results using the three error metrics) appeared to account for more micro-scale changes in the amount of rainfall than any of the other interpolation methods. Thin plate smoothing splines seemed to have the opposite effect, and only picked up large-scale variations in rainfall, which was readily apparent when evaluating the spatial maps from November of 2007. Splines again generated spatial rainfall patterns dictated by large-scale variations of rainfall across the entire country, whereas Bayesian kriging and IDW seemed to interpolate to a much finer scale. In particular, the Bayesian kriging interpolation for November of 2007 did not look reasonable, as there was far more small-scale variation than would be expected during a monsoonal rainfall event.

The spatial outputs from 2010 followed similar patterns to those for 2007, but the magnitude of rainfall experienced for this year was much greater. Again, Bayesian kriging interpolated the most small-scale variation in rainfall, which in turn led to the best overall results for May. For November, Bayesian kriging depicted irregular spatial patterns of rainfall. Splines tended to produce interpolated spatial rainfall patterns that closer approximated what would be expected for continuous rainfall events, and this is in agreement with the three error metrics employed, which found splines to be the best predictor of rainfall for November of 2010.

3.3 **Map comparison**

Evaluating the SSIM between rainfall interpolations and TRMM remotely sensed rainfall estimates for Sri Lanka produced very low SSIM values (means of ~0.4 and ~0.3 for May and November, respectively), which represented low correlation in rainfall values associated between the two rasters. Low correlation was likely a product of the differences in the amount of rainfall predicted by the interpolation methods versus the amount of rainfall observed in the TRMM rasters. TRMM remotely sensed rainfall estimates have been found to consistently underestimate the amount of rainfall when compared to ground-based gauge measurements (Wang and Wolff, 2010). For this reason, the structure component (*S*) of the SSIM was primarily focused on, as it does not take rainfall magnitudes into account, and focuses solely on spatial patterns in the data. Assessing the *S* component of the interpolation methods compared to the TRMM rasters revealed that the spatial patterns for the month of May were considerably more similar than spatial patterns for November.

In May, Bayesian kriging attained the highest mean *S* value over the 5 year study period of 0.85, while thin plate smoothing splines attained the lowest mean *S* value of 0.77. The *S* values for all interpolation methods were high, demonstrating correlation between the spatial patterns present in the TRMM rasters. Southern Sri Lanka accounts for the majority of the dissimilarity between the interpolations and the TRMM rasters. The only year that did not demonstrate high *S* (greater than 0.75 for all interpolation methods) was 2010, where Bayesian kriging attained the highest *S* value of 0.69. November did not demonstrate nearly as high *S* values as May.

All the interpolations' mean *S* values over the 5 year study period for November were below 0.70 – a considerable amount lower than May – with thin plate smoothing splines

attaining the highest of 0.67, and ordinary kriging attaining the lowest of 0.61. Spline interpolations' *S* value for 2008 and 2009 (0.73 and 0.81) were at least 0.15 higher than the next closest *S* value.

Figure 2.7 depicts maps of the structure component of the SSIM for November of 2009 which was chosen for analysis because MAEs, MdPEs, and SRMSEs for all interpolation methods were similar, so the structure component was the only variable truly being assessed. The spline interpolation exhibited much higher *S* values all around Sri Lanka, while IDW and Bayesian kriging interpolations showed much lower and occasionally negative *S* values, depending on the area. Major differences in structure for IDW were focused in the central areas of the country, while for Bayesian kriging, low *S* values were present from the southwest corner of Sri Lanka up through the northeast. November of 2007 had very low *S* values for all of the interpolation methods, which brought down the mean *S* values for November considerably and should be taken into account when comparing November to May.

4. Discussion

Some of the error associated with the results can be accounted for when looking at Sri Lanka's regional rainfall trends. As a general trend, most of Sri Lanka receives continuous rainfall during the second inter-monsoonal season in November. The southwest monsoon – which experiences peak rainfall in May – exhibits very different spatial patterns of rainfall, as much of the country is dry, receiving little rainfall. Given that the north of Sri Lanka is much less densely sampled than the south (Figure 2.1), the rainfall discrepancies during the southwest monsoon can be problematic, as large regions of the country's rainfall are being predicted given little input data (i.e., trace rainfall and an extremely sparse network of station). Predicting

rainfall for areas located in the boundary zones between the south, which is experiencing substantial rainfall in May, and the north, which is dry, can be especially difficult, and thought of as a key source of error. Finding the boundary zones between the monsoonal south and the dry north is a research area that could significantly improve rainfall maps for Sri Lanka, perhaps through the use of automated satellite-derived rainfall maps, or densification of the existing network in these transition areas.

It can be speculated that Bayesian kriging, which was generally found to perform the best in low rainfall conditions, was able to account for the very local variations in rainfall between the north and the south for the southwest monsoon. Bayesian kriging produced low values from all three error metrics when the year was considered to have low rainfall. However, while Bayesian kriging had the lowest SRMSEs in May in years where low rainfall was observed (2007, 2008), it also had the highest SRMSEs for 2006, 2009, and 2010 in both May and November. These years can be thought to represent years of medium to high rainfall. Bayesian kriging was more effective when interpolating rainfall for years with very low rainfall, as when rainfall levels increase, the errors associated with Bayesian kriging increased at a rapid rate. The increased errors of Bayesian kriging in high rainfall conditions could be a result of oversensitivity to erroneous data present in the dataset used to fit the models (i.e., the agricultural weather station network), and prior distributions not being properly tuned is another possible source of error. Model parameterization can be difficult when dealing with Bayesian inference, specifically when trying to determine general parameters that could be used to fit models given different underlying conditions (e.g., months with high or low rainfall totals). Temporally-dynamic prior-parameterization might improve this aspect of the modelling. It should be noted that for many applications of large-area rainfall maps such as diseases risk forecasting, incorrectly estimating

rainfall at extremely low magnitudes (i.e. <10 mm) would not have nearly as dramatic an effect on the model as incorrectly estimating rainfall at higher magnitudes (i.e. >200 mm).

When evaluating interpolation methods based on overall performance regardless of month or year, thin plate smoothing splines seemed to perform the best. Looking at SRMSE, splines acquired the lowest SRMSEs for four of the ten months studied, the highest percentage of any of the interpolation methods assessed. May of 2007 and 2008 were the only years that splines attained the highest SRMSEs. These years were both noted for particularly low rainfall, and given that May in general experiences lower rainfall totals than November, it may be that splines predict less accurately with low yearly rainfall. As rainfall patterns in May revealed sharp spatial discontinuities, splines sometimes performed worse than other methods due to the smoothing nature of the interpolator (Figure 2.6), while the more continuous nature of rainfall in November was more suited to local methods. The geostatistical methods tested did not consistently predict rainfall with greater accuracy, which could be due to the fact that much of northern Sri Lanka was sparsely sampled, and spatial non-stationarities were stronger than the modelled correlation structure.

The quality of the rainfall data obtained from the community-managed weather stations that were utilized to produce the interpolations in this research were uncertain, and gave incentive to test whether results produced using these stations would be on par with the official meteorological stations. While errors varied from year to year, the results were largely accurate so long as the areas around the official meteorological station being evaluated were sufficiently sampled by community-managed weather stations. Where this was particularly evident was official meteorological station 8 in November of 2010 (Figure 2.5). Station 8 showed all interpolation methods predicting rainfall poorly and inconsistently at that location. Figure 2.8

depicts the locations of community-managed weather stations overlaid on Voronoi polygons created from the locations of official meteorological stations. The highlighted area in the figure represents the location of station 8, and how the community-managed weather stations were situated around it. None of the community-managed weather stations were located close to official meteorological station 8, thus resulting in poor prediction. The reason for the lack of community-managed weather stations was that this area represents Colombo, the capital city and urban center of Sri Lanka. Most of the community-managed weather stations used for the interpolations were located in agricultural areas on farms, and therefore there was not adequate sampling in predominantly urban regions, such as the areas surrounding meteorological station 8. Inadequate sampling in urban regions is a noted shortcoming of using the network of community-managed weather stations for interpolations, but being that there are few urban areas in Sri Lanka, most predictions were not dramatically affected by this. Future interpolations using rainfall data from both the official meteorological stations and the community-managed weather stations would likely produce the most accurate results in all areas of Sri Lanka and alleviate this issue.

4.1 **Map comparison**

In May, the *S* values for all interpolation methods were high, which implied correlation between the spatial patterns present in the TRMM rasters and the interpolations. The correlation is thought largely to be due to the areas in northern Sri Lanka that received little to no rainfall during the southwest monsoon. Areas in northern Sri Lanka demonstrated very high *S* values, as the spatial patterns were close to identical since there was very little rainfall. Finding the structure component of these areas would likely results in *S* values very close to 1, representing nearly identical spatial patterns. One year where May interpolations did not attain high *S* values

when compared to the TRMM raster was 2010. Looking at the interpolations compared to the TRMM raster for this year, it is evident that the spatial patterns between the two are much different. Although the north of Sri Lanka is still experiencing little to no rainfall in all rainfall maps, the interpolations depict heavy rainfall only in the southwest of Sri Lanka, while the TRMM raster shows rainfall only in the southeast. 2010 was the year that Sri Lanka experienced the most rainfall, so this heavy rainfall in the southwest was expected, but the difference in spatial pattern between the interpolations and TRMM rasters was unexpected. The TRMM raster depicts a very different trend than is present in all the interpolations, and is clearly the cause of the low *S* values present for this year.

For November, it was noted before that the spline interpolations' *S* value for 2008 and 2009 (0.73 and 0.81) were at least 0.15 higher than the next closest *S* value, which is thought to be due primarily to the nature of how splines are generated. Splines ensure a smooth fit of rainfall across the study region, which is congruent with the spatial patterns of the TRMM rasters for these years. All of the other interpolation methods demonstrated more local variations in rainfall, and when assessing spatial pattern similarity, this caused their *S* values to be considerably lower. Since Sri Lanka experienced rainfall across the entire country in November, the smoothing nature of splines is effective at predicting this continuous spatial pattern of rainfall.

SSIM analysis between interpolations and the TRMM remotely-sensed rainfall estimates proved to be a useful tool for assessing the spatial patterns of rainfall in Sri Lanka. While using SSIM analysis on its own does not provide meaningful feedback with regard to the accuracy of prediction from interpolation, using it in conjunction with other error metrics helped clarify the deficiencies of certain interpolation methods with respect to others. It can be concluded that

employing this type of spatial pattern analysis is beneficial for assessing global patterns in data, and is recommended for future studies focused on predicting spatial phenomena.

It is difficult to contextualize the empirical findings of this study, as there have been few scientific studies conducted on interpolating rainfall in Sri Lanka. This is especially true when taking the entire country of Sri Lanka into account – most studies have only aimed at predicting rainfall in the dry zones of the country (Jayawardene et al., 2005; Punyawardena and Kulasiri, 1999). Jayawardene et al. (2005) examined interpolating daily rainfall totals in the dry zone of Sri Lanka between 1970 and 1999. Ordinary kriging and IDW were employed, and were both found to predict rainfall at a similar level of quality. Since the data being interpolated were daily rainfall, it is difficult to compare to the monthly totals being predicted in this study. Jayawardene et al. (2005) calculated mean absolute percentage errors for IDW between 34.2% and 48.8%; depending on the month being studied, these were in line with the MdPEs calculated from this study. The data used by Jayawardene et al. were thought to be derived from the same official meteorological station network used for validation in this research, which based on the magnitude of MdPE would suggest that the community-managed station network used to predict rainfall was as viable a data source for building interpolation models as the official meteorological station network. Scatterplots depicting observed rainfall versus predicted rainfall for both IDW and ordinary kriging exhibited a closer linear relationship than most of the plots from this study, but it must be considered that these results were only attained for the dry zone of Sri Lanka, where the maximum observed rainfall values were less than 50 mm. Lower rainfall magnitudes would obviously affect the degree of error associated with each rainfall value. While it is encouraging that ordinary kriging and IDW performed similarly well in both studies,

drawing any more definitive conclusions about the quantitative findings is difficult due to the differences in scale and study region between the two studies.

Punyawardena and Kulasiri (1998) researched interpolation of weekly rainfall across the dry zone of Sri Lanka using an exponential model compared to local interpolation methods (i.e., IDW, local mean). The study found the exponential model to perform marginally better than the local methods in the time period of two dry seasons. Punyawardena and Kulasiri interestingly found that no method of spatial interpolation was effective within the dry zone when the distance between station locations was greater than 100 km. The most northern regions of Sri Lanka were the only areas where the agricultural weather station network used in this study had any gaps of coverage greater than 100 km. Figure 2.5 which depicts SE of rainfall by meteorological validation station shows that station locations found in the northern half of Sri Lanka were predicted relatively well by all interpolation methods. These low SEs provide more evidence that the network of community managed weather stations used in this study is a viable data source for fitting interpolation models in Sri Lanka.

Upon reviewing all quantitative analysis performed on the interpolations, it is difficult to determine one particular method that consistently produced more accurate results that the others in all scenarios (e.g., high and low rainfall years). What can be seen is that accuracy of rainfall prediction in Sri Lanka is very much dependent on season due to the dramatically different distributions of rainfall. Each of the study months could practically be treated as entirely different study periods, as where one method performed well for May, it may have performed poorly for November. When low yearly rainfall was observed, Bayesian kriging was often able to attain the most accurate results based on the three error metrics employed. However, as yearly mean rainfall increased, Bayesian kriging's prediction accuracy quickly deteriorated. When

moderate to high yearly mean rainfall was present, thin plate smoothing splines produced the most accurate results. It also was found to most closely approximate the spatial structure of rainfall events in Sri Lanka, specifically during the second inter-monsoonal season. In future research of rainfall in Sri Lanka, it may be beneficial to determine a threshold value for yearly mean rainfall, under which it can be assumed that Bayesian kriging would produce accurate results. If the yearly mean rainfall exceeds the threshold value, thin plate smoothing splines could be employed to interpolate rainfall.

5. Conclusions

This study assessed four different interpolation techniques and their ability to accurately predict monthly rainfall in Sri Lanka. IDW, splines, ordinary kriging, and Bayesian kriging were selected as appropriate methods for interpolating rainfall for May and November of 2006 through 2010. Community managed weather stations were used to interpolate rainfall and evaluated using official meteorological station readings to validate the quality of the results. TRMM remotely sensed rainfall estimates were also used to compare to the interpolations and assess the spatial pattern of rainfall in Sri Lanka. Certain methods performed better dependent on the month being interpolated, and the spatial pattern of rainfall. Splines tended to perform well in situations where all of Sri Lanka experienced high rainfall, whereas Bayesian kriging performed well when the north of Sri Lanka experienced dry conditions, and the south experienced rainfall. Comparing the spatial structure of the interpolation to the remotely sensed images demonstrated that most methods approximated the spatial distribution of rainfall at a similarly high level for May, largely due to much of Sri Lanka experiencing uniformly low rainfall other than the southwest portion of the country. November showed thin plate smoothing splines closely approximating the spatial pattern of the TRMM rasters, which was thought to be

due to the smooth spatial pattern that spline interpolations produce. The community managed weather stations attained similar performance.

Interpolating a climate variable such as rainfall is very much dependent on the setting, and thus more methods must be tested to determine the best means of prediction. Involving all other months in this research would help elucidate which methods perform well on a consistent basis, and will be taken into account in future research. Similarly, interpolation models are, like all models, context-specific representations of more complex processes. The data and models analyzed in this paper will eventually be used to forecast the spatial distribution of waterborne infectious disease risk in Sri Lanka. With accurate precipitation maps, models can be constructed that will identify correlation between rainfall and disease incidence and reveal when and where rainfall-driven outbreaks are likely. While the models presented here have demonstrated the utility of including both community-managed data and satellite imagery in the rainfall mapping methodology, I suspect the insights from this analysis will be applicable to a wide array of environmental modelling contexts.

5.1.1. Abbreviations

IDW – Inverse Distance Weighting
MAE – Mean Absolute Error
MdPE – Median Percent Error
SE – Statistical Error
SRMSE – Standardized Root Mean Square Error
SSIM – Structural Similarity Index
S – Structure component of SSIM
TRMM – Tropical Rainfall Measuring Mission

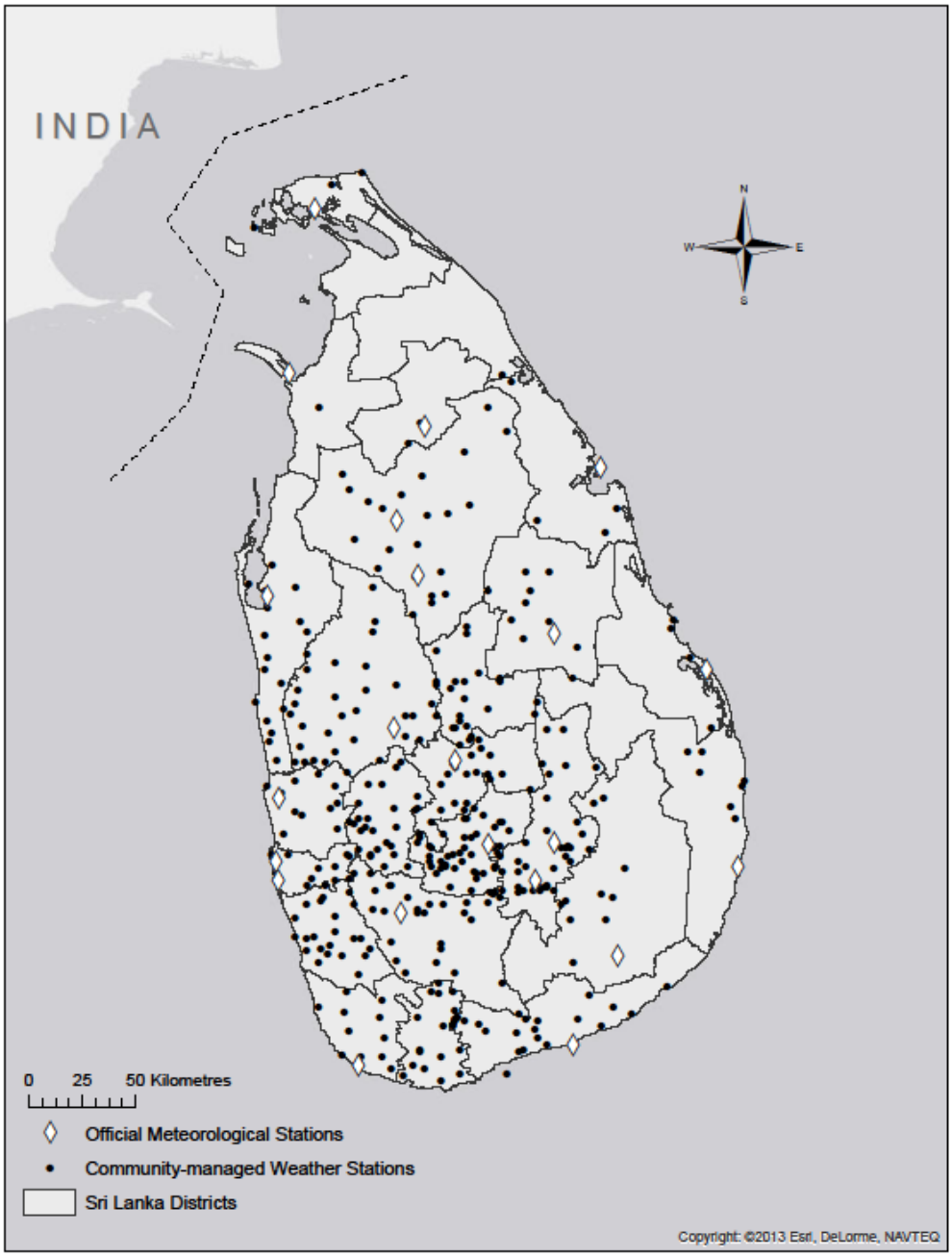


Figure 2.1. Locations of official meteorological stations and community-managed weather stations in Sri Lanka.

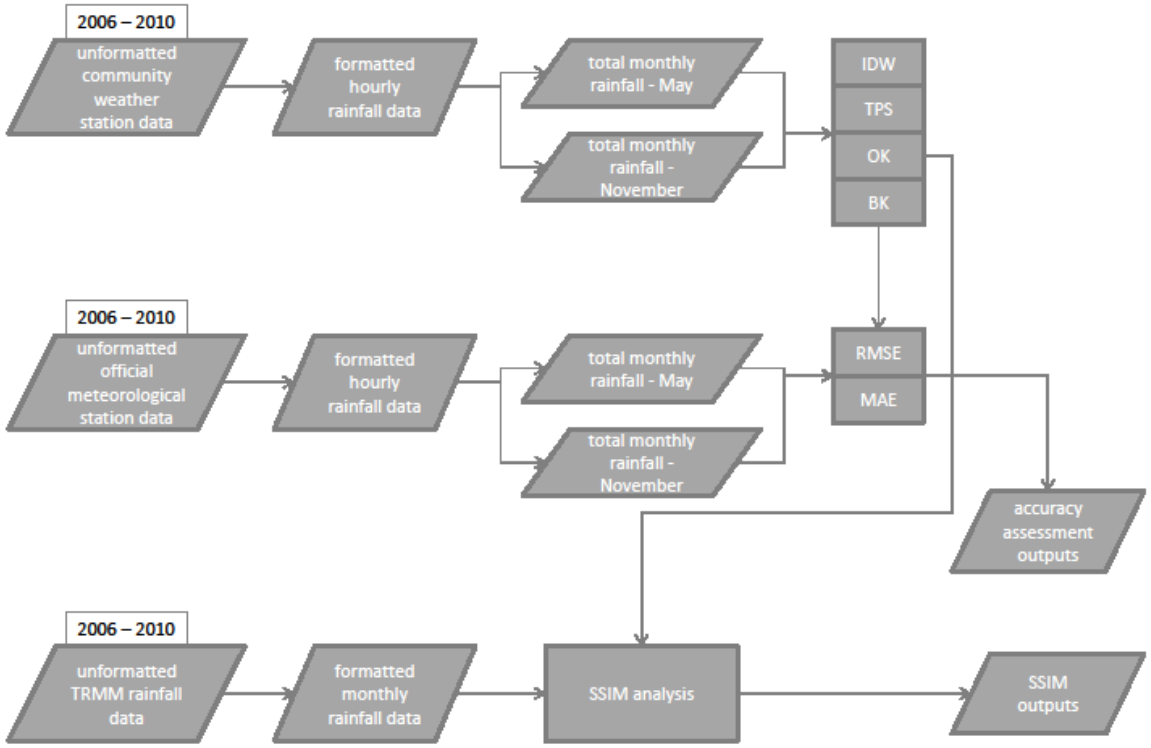


Figure 2.2. Flow chart illustrating steps taken to carry out research.

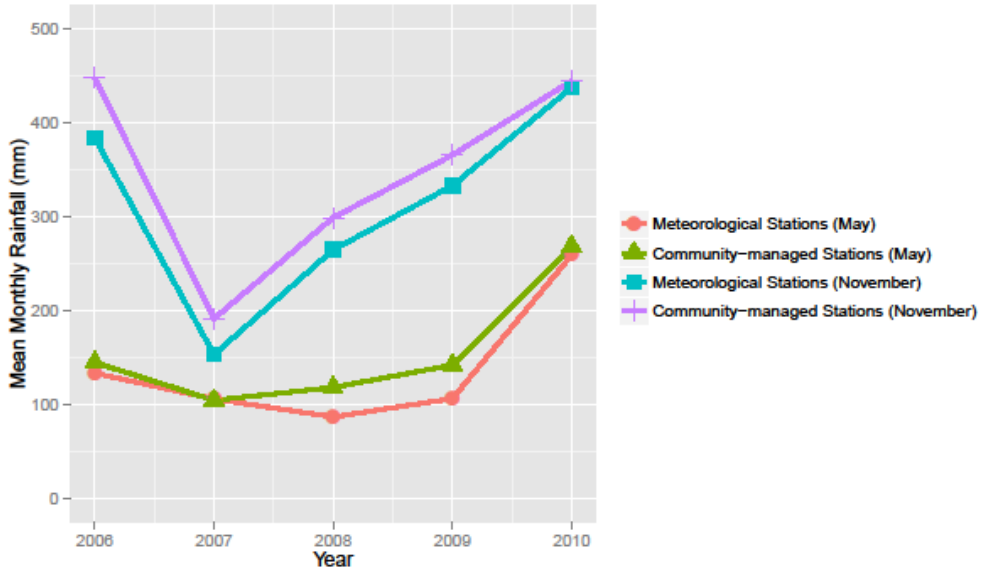
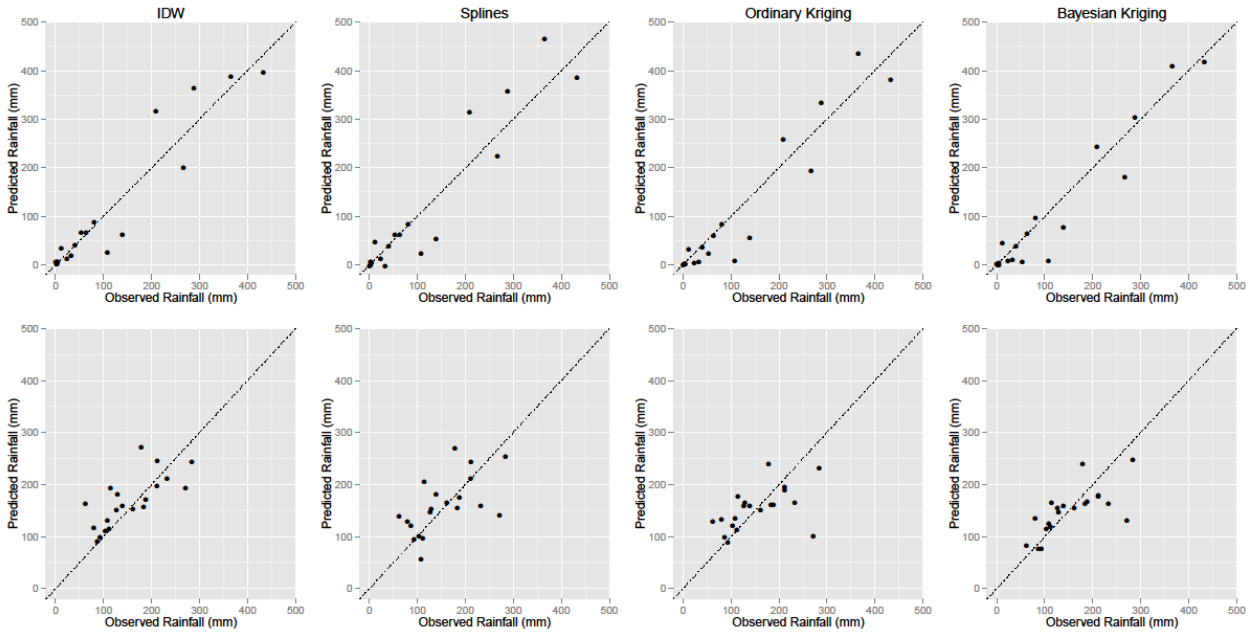


Figure 2.3. Mean monthly rainfall by year of official meteorological stations and community-managed community-managed weather stations for May and November of 2006 – 2010.

2007 (May on top, November on bottom)



2010 (May on top, November on bottom)

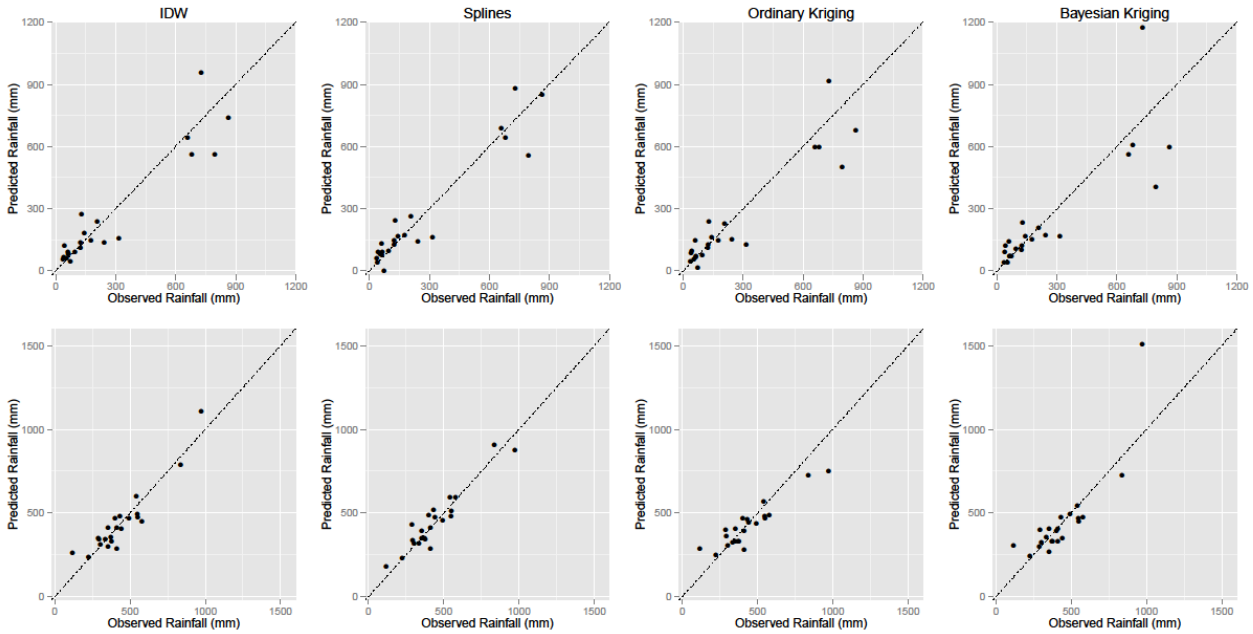


Figure 2.4. Scatterplots of Observed vs. Predicted values for all interpolation methods of May and November 2007 and 2010.

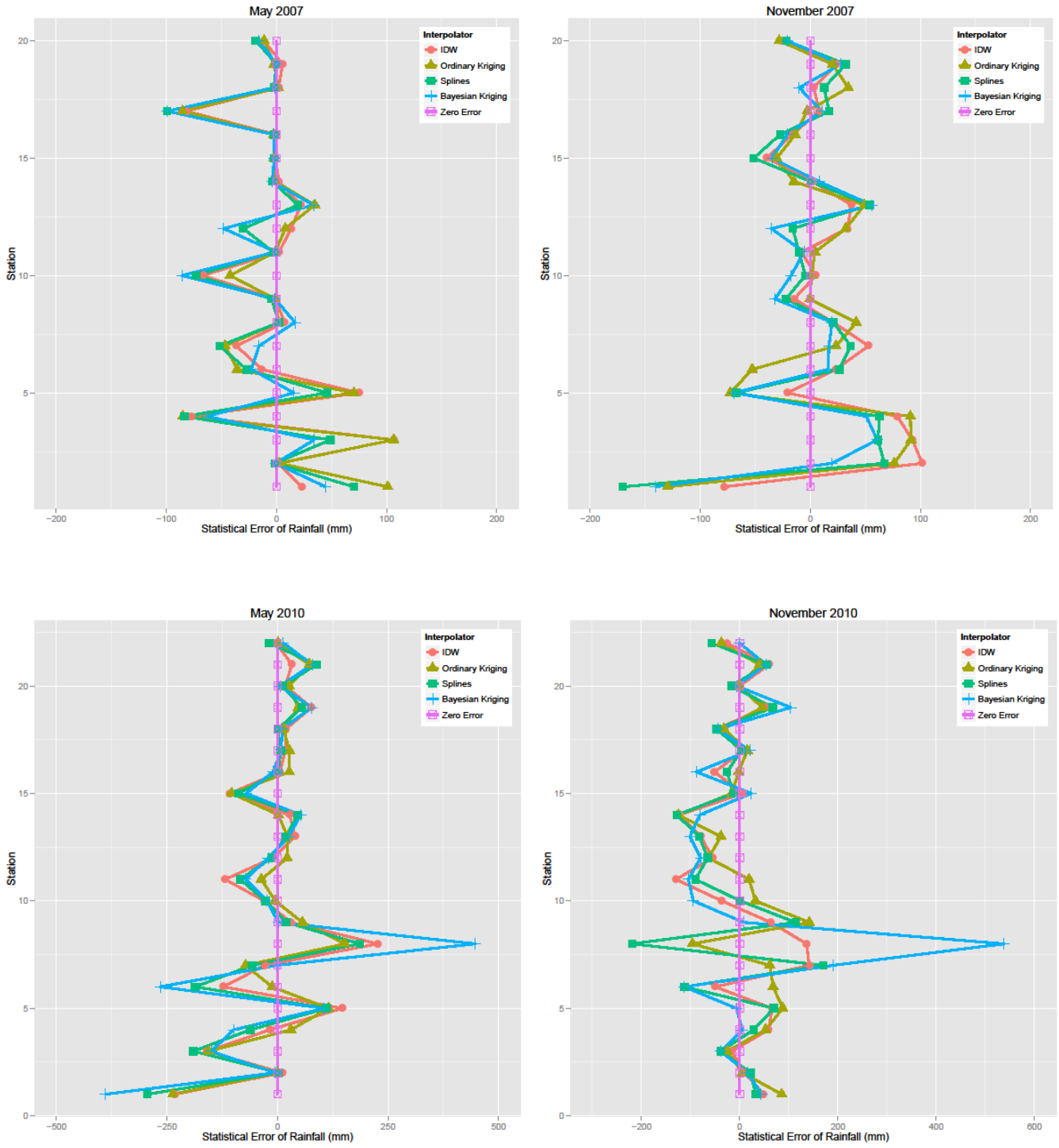
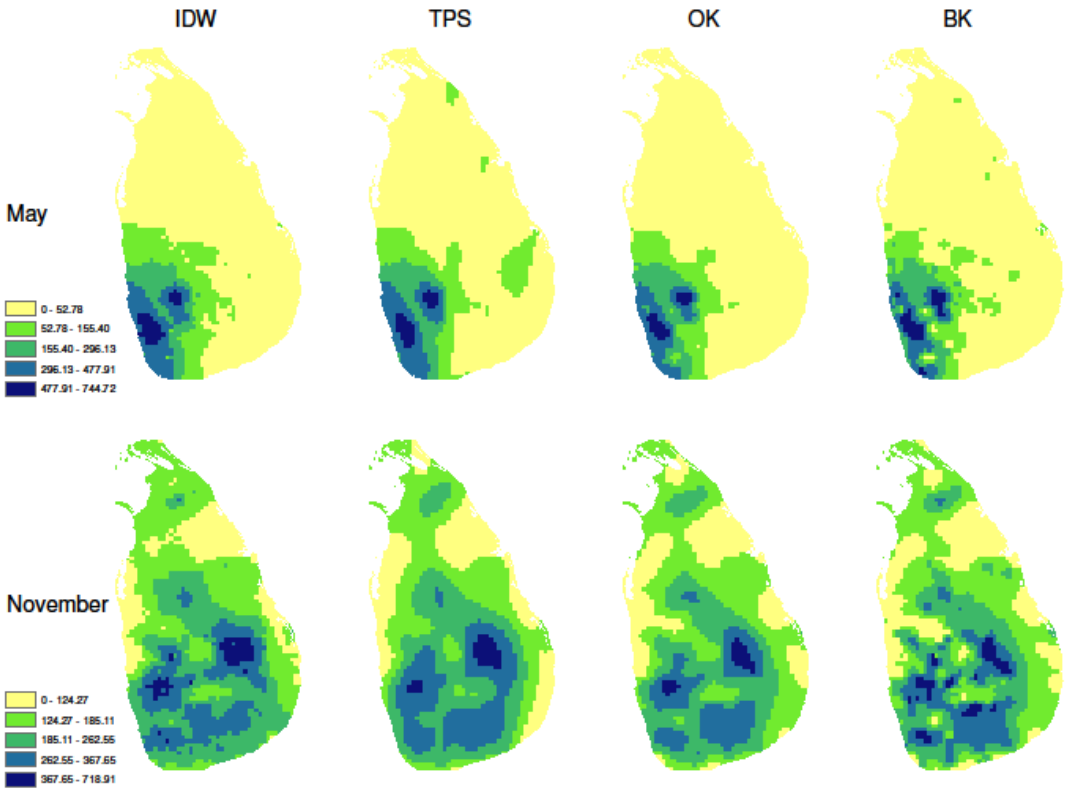


Figure 2.5. Statistical error of rainfall in between interpolation methods and official meteorological station rainfall measurements delineated by meteorological station location sorted from south (1) to north (20 - 22).

2007


IDW
TPS
OK
BK
May
0 - 52.78
52.78 - 155.40
155.40 - 296.13
296.13 - 477.91
477.91 - 744.72
November
0 - 124.27
124.27 - 185.11
185.11 - 262.55
262.55 - 367.65
367.65 - 718.91

2010

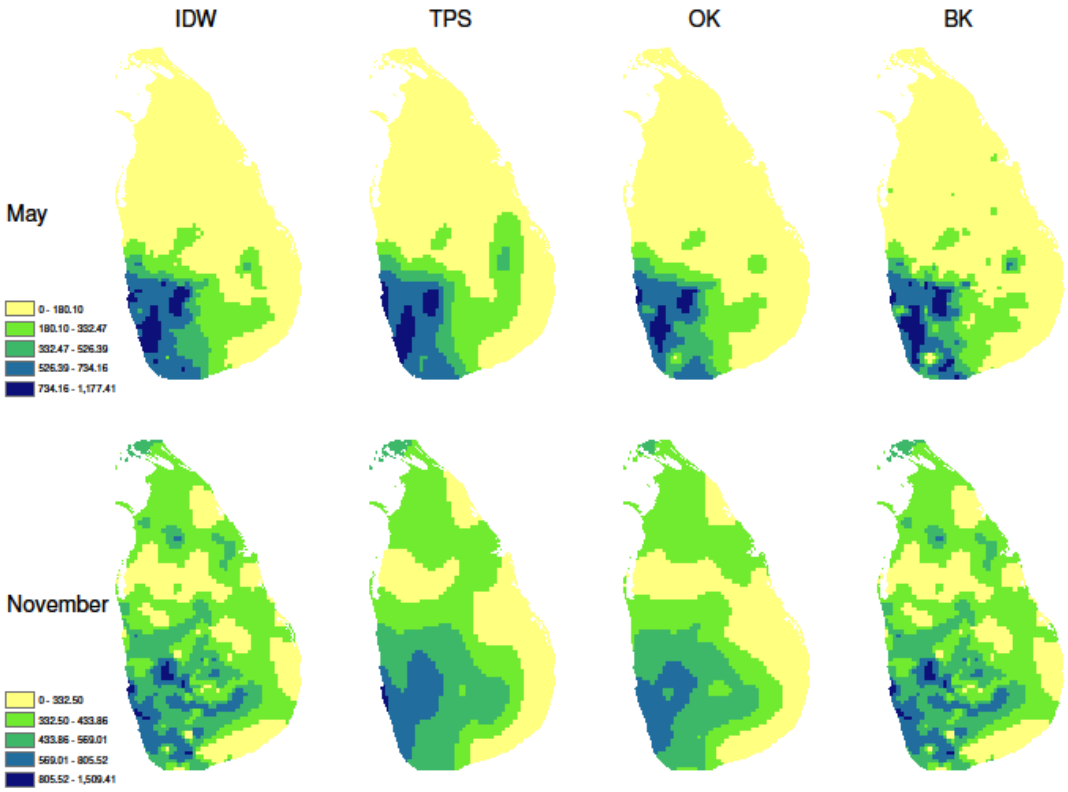


Figure 2.6. Spatial outputs of all interpolation methods for 2007 and 2010. The legend denotes the amount of predicted rainfall (mm).

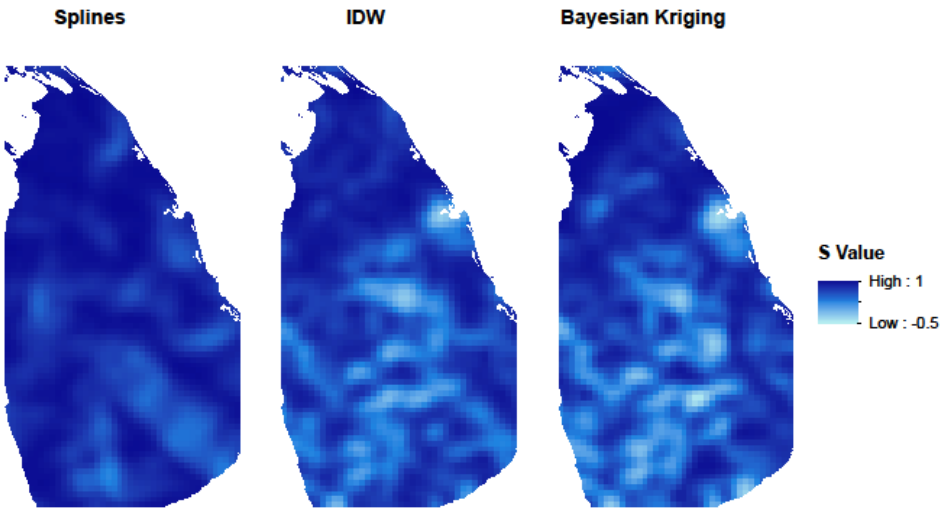


Figure 2.7. Maps of structure component for November 2009. Mean S value over study area: Splines, S = 0.81; IDW, S = 0.66; Bayesian kriging, S = 0.63.

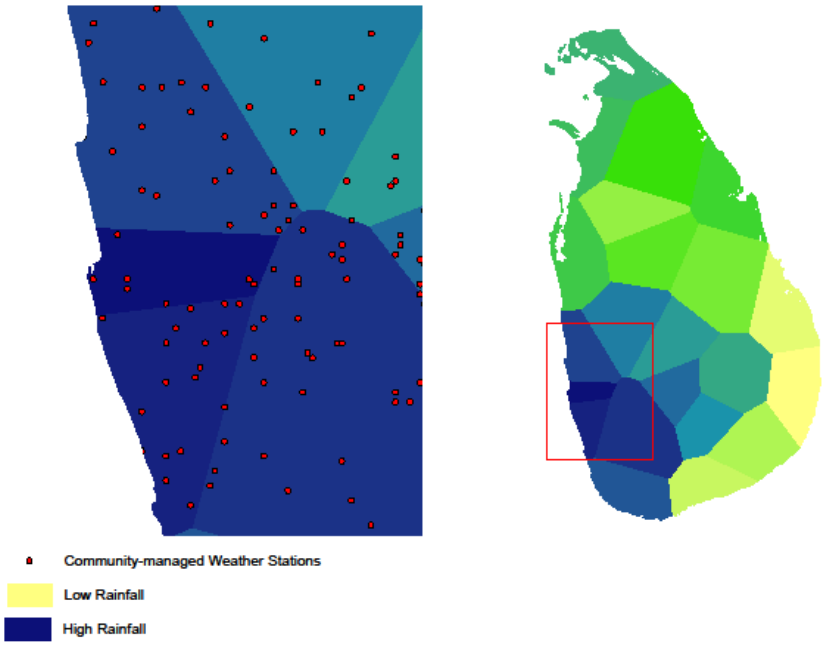


Figure 2.8. Locations of community-managed weather stations for November 2010 with Voronoi polygons based on official meteorological station locations. Highlighted area focuses on official meteorological station 8.

Table 2.1. Summary of studies of rainfall interpolation. Studies are sorted by author, process, the interpolation methods being employed, and the overall findings of each study.

Study	Process	Interpolation Methods Employed	Findings
Hutchinson (1995)	Interpolated annual rainfall for a region of south eastern Australia	-Thin plate smoothing splines	-Splines required no prior estimation of the spatial auto covariance structure, which could prove beneficial when the data set being used could contain errors distributed across entire spatial network of observation stations
Dirks et al. (1998)	Interpolated rainfall obtained for Norfolk Island, of the coast of Australia	-Areal-mean -IDW -Kriging -Thiessen polygons	-All methods found to perform at a similar level -Thiessen polygons produced most unrealistic results, due to discrete rainfall boundaries -IDW deemed most appropriate method, due to accurate interpolations produced, and low performance requirements
Price et al. (2000)	Interpolated monthly mean climate data for study sites in British Columbia/Alberta, and Ontario/Quebec, Canada.	-ANUSPLIN, a software based on thin plate smoothing splines -GIDS, a regression-based model	-Extreme outliers that exceeded 100% difference between observed and predicted values present for precipitation interpolations -ANUSPLIN produced slightly more accurate predictions, as it could more easily account for changes in elevation -Regions with sparse data occasionally exhibited negative precipitation values being predicted
Vicente-Serrano et al. (2003)	Analysed validity of precipitation and temperature maps of the Ebro Valley in northeast Spain	-Empirical regression models -IDW -Kriging methods -Thiessen polygons -Thin plate smoothing splines -Trend surfaces	-Regression modelling and kriging methods produced the highest correlation between observed and predicted rainfall -Trend surfaces and Thiessen polygons were determined to produce the least accurate predictions according to validation statistics
Oke et al. (2009)	Investigated prediction of rainfall across Australia	-Cokriging -Ordinary kriging -Simple kriging with a locally varying mean	-Prediction errors from all three methods found to be similar (negative errors implying underestimation of gauge rainfall present for all methods) even with the inclusion of satellite-based TRMM rainfall estimates for some methods -Satellite rainfall data potentially improved spatial prediction in areas that were not adequate

			sampled -Errors were consistently negative in coastal regions, while errors in higher inland areas tended to be positive
Newlands et al. (2010)	Evaluated three interpolation methods for precipitation and temperature across much of the Canadian landmass	-ANUSPLIN, a software based on thin plate smoothing splines -HYBRID inverse-distance/natural-neighbour model -DAYMET Weighted-truncated Gaussian filter	-All models predicted reasonably well, with ANUSPLIN producing the most accurate daily mean precipitation values at a 10 km scale -High error variance for precipitation was exhibited in summer along the coasts and in winter in the Prairies region -Authors recommend employing a Bayesian/mixed models methodology for future climate prediction in Canada

Table 2.2. Brief descriptions and formulas of several commonly used interpolation methods employed in this research.

Interpolation Method	**Formula**	**Description**
Inverse distance weighting	$\hat{z}(s_i) = \sum_{i \neq j}^{n} \frac{1}{d_{ij}} z(s_i)$ $\hat{z}(s_i)$ is the estimated value at location s_i, d_{ij} is the distance between s_i (unknown value) and s_j (known value), n is the number of known value locations within the set radius, and $z(s_i)$ is the sampled value within the set radius	-IDW is based on the assumption that a climatic value at an unsampled site is the distance weighted average of climatic values from all sampled sites surrounding it within a given radius. -As distance increases between the sampled location and the location being interpolated, the weight associated with the sampled value decreases.
Spline interpolation	$\hat{z}(s_i) = \sum_{i=1}^{n} f(s_i) + \epsilon(s_i)$ $\hat{z}(s_i)$ is the estimated value at location s_i, f is a radial basis function, and $\epsilon(s_i)$ are the random errors associated with that location	-Spline interpolation generalizes IDW by expanding the local function to a group of continuous functions adapted to local variations in the sampled data. -A radial basis function is created for all of the sampled data within the radius of each estimation location. -A bivariate spline function was used to model the spline surfaces in this study, where there was a spatially varying dependence on rainfall.
Ordinary kriging	$\hat{z}(s_i) = \sum_{i=1}^{n} w_{ij}(s_j)$ $\hat{z}(s_i)$ is the estimated value at location s_i, and a graph of semivariance is used to select a model to make predictions at unsampled locations by deriving the optimal set of weights w_{ij} to use in a linear combination of neighbouring values	-In ordinary kriging, the continuous variable used to generate the surface consists of a spatially-correlated random component. -The spatial variance of the climate variable being interpolated is used in a function that is determined using a semivariogram model which estimates semivariance as a function of spatial distance. -In this research, spherical semivariogram models were used for rainfall, which is a commonly available model in most geostatistical software packages.

Table 2.3. Mean absolute errors (MAE), median percent errors (MdPE), and standardized root-mean-square errors (RMSE) between interpolations and official meteorological station rainfall measurements (mm) for May and November of 2006 – 2010. RMSEs were standardized by mean official meteorological station rainfall of all stations given year and month.

MAEs and MdPEs (in brackets).

May

	IDW	Splines	Ordinary Kriging	Bayesian Kriging
2006	30.24 (23.29%)	29.23 (35.11%)	30.94 (38.42%)	35.65 (29.94%)
2007	27.49 (45.23%)	32.17 (46.74%)	29.50 (59.59%)	25.38 (39.67%)
2008	23.70 (53.73%)	26.60 (69.55%)	21.70 (27.55%)	20.81 (22.48%)
2009	27.91 (39.18%)	29.45 (33.95%)	24.87 (44.52%)	31.93 (56.92%)
2010	67.27 (28.69%)	56.20 (24.32%)	71.42 (20.87%)	87.92 (17.58%)

November

	IDW	Splines	Ordinary Kriging	Bayesian Kriging
2006	74.78 (9.73%)	86.04 (13.97%)	79.24 (14.2%)	96.08 (26.26%)
2007	34.30 (14.54%)	40.52 (16.92%)	39.14 (17.3%)	33.74 (14.96%)
2008	83.81 (20.45%)	69.28 (14.68%)	88.26 (22.51%)	75.05 (19.41%)
2009	76.49 (16.15%)	76.59 (16.69%)	76.10 (18.44%)	82.87 (16.57%)
2010	57.44 (11.65%)	49.61 (8.31%)	66.35 (12.18%)	80.10 (12.76%)

Standardized RMSEs.

May

	IDW	Splines	Ordinary Kriging	Bayesian Kriging
2006	0.316	0.285	0.288	0.360
2007	0.406	0.460	0.404	0.360
2008	0.423	0.501	0.418	0.385
2009	0.355	0.422	0.362	0.481
2010	0.374	0.318	0.405	0.578

November

	IDW	Splines	Ordinary Kriging	Bayesian Kriging
2006	0.283	0.317	0.262	0.323
2007	0.296	0.347	0.349	0.295
2008	0.513	0.427	0.601	0.589
2009	0.317	0.301	0.293	0.366
2010	0.162	0.142	0.196	0.312

Chapter 3: Forecasting leptospirosis risk in Sri Lanka using interpolated rainfall

1. Introduction

The drivers behind the emergence of an infectious disease are often difficult to identify and account for, as they are usually an assemblage of several varied factors, including ecological changes, human demographics and behaviour, travel and movement of people and goods through space and time, and failings in pre-existing public health measures (Morse, 1995). Trying to account for all of these social, environmental, and economic factors can become an increasingly complex problem, and in many situations, may not be feasible. As an alternative, it can be a more viable approach to look to more tangible underlying conditions, such as the physical environment, that are known to have a meaningful relationship with these complex drivers of emergence.

When assessing emerging infectious disease (EID) risk in a spatial context, environmental factors can often play a major role when trying to predict areas of future outbreak (Briët et al., 2008; Robertson et al., 2012). In developing countries where data can be sparse or are often not available, if climate data can be obtained, they can provide a means for developing disease risk forecasting models, and can act as a proxy to more complex drivers of transmission. Leptospirosis is an EID whose incidence is increasing in developed and developing countries around the world (Vijayachari et al., 2008). This increase in incidence – specifically in developing countries where outbreak events can have major health repercussions and administering treatment may be difficult due to financial cost and physical distribution limitations – provides incentive to develop leptospirosis risk models to establish early warning protocols to limit future outbreak events. If effective early warning can be successfully

implemented, it can help prevent future outbreaks from occurring when environmental conditions are present that indicate high probability of leptospirosis transmission, thus limiting potential costs associated with trying to mitigate disease transmission after an outbreak event is already underway. In this paper, I employed a modelling approach for forecasting leptospirosis risk in the country of Sri Lanka. A variety of different modelling techniques were considered to identify the best rainfall variables predictive of suspected leptospirosis cases across several districts of Sri Lanka. If models can be developed that effectively project leptospirosis risk based on local meteorological data, they will be suggested for use in early warning systems in districts of Sri Lanka where major leptospirosis outbreaks have occurred in the past.

1.1 **Leptospirosis**

Leptospirosis is a globally significant EID, as it is thought to be the most widespread zoonotic disease in the world (Levett, 2001; Sarkar et al., 2012; WHO, 1999). In the recent past, leptospirosis incidence has increased in developed and developing countries around the world (Vijayachari et al., 2008). For example, in Sri Lanka, a general trend of increasing leptospirosis incidence has been observed since 2006 with a country-wide outbreak in 2008 (Table 3.2). Human infection is caused by exposure to the pathogenic Leptospira species. This pathogen is usually spread to humans through contact with water contaminated by urine of infected animals (Bharti et al., 2003). Leptospirosis is often misdiagnosed, as it has variable symptoms that mimic many other infectious diseases (Lau et al., 2010). Fast recognition of leptospirosis is important because early treatment is crucial if morbidity and mortality are to be limited. Direct human-to-human transfer of leptospirosis is noted to be very rare, and will not be considered as a realistic means of infection in this research (Levett, 2001). The incidence of this disease is much higher in tropical climates than in temperate zones, as the Leptospira species is able to survive

much longer in warm, humid environments (Bharti et al., 2003; Levett, 2001). This also can be attributed to the fact that most warm climate countries are developing countries as well, where exposure to animal hosts is increased due to the greater role of agriculture in national economies (Bellack et al., 2006; Madsen and Shine, 1999; Mendelsohn and Dinar, 1999). Usually, leptospirosis is contracted through cuts or skin abrasions and subsequent immersion in contaminated water. Occupation often plays a role in leptospirosis risk, and occupations which involve interaction with animal reservoirs put one at greater risk (Levett, 2001). The most important vectors of leptospirosis are often small mammals, of which the most significant in Sri Lanka are rodents. These rodents may transfer the infection to other domestic farm animals, dogs, and humans. The extent to which leptospirosis is transmitted relies on many variables, such as climate, population density, and the degree to which there is contact between hosts and sources of infection.

1.2 Environmental Risk Factors for Leptospirosis

Precipitation is thought to have a pronounced effect on the incidence of many rodent-borne diseases. Several different authors have demonstrated a link between fluctuations in rodent reservoir populations and oscillations in new human cases of disease (Heyman et al., 2001; Mills and Childs, 1998; Olsson et al., 2003; Rose et al., 2003). This link is a mechanism of changes in densities of rodent populations due to ecological factors, such as abundant food supply, causing corresponding changes in frequency of contact between humans and infected rodents (Heyman et al., 2001). Also, a number of studies have linked large amounts of rainfall with an increased number of human cases of rodent-borne disease (e.g., Davis and Calvet, 2005; Enscore et al., 2002). It has been hypothesized from this that high precipitation can lead to increased rodent populations, which consequently results in higher rodent-borne disease

incidence. While monitoring fluctuations and movements of rodent populations may be the most effective way to model leptospirosis incidence, it is very difficult to do, and thus using rainfall as a proxy for rodent populations is a potential tool for early warning systems for rodent-borne disease.

The prevalence of leptospirosis is significantly higher in warm, humid regions (Levett, 2001). In some tropical climates, rodent-borne disease incidence is tied to seasonality, which is characterized by large monsoonal events, for example, the northeast monsoon season in Sri Lanka (Robertson et al., 2011). The substantial rainfall in monsoon season is speculated to increase food sources for rats, which in turn improves conditions for rat reproduction (Madsen and Shine, 1999). This results in a spike in rat populations following the monsoonal rainfall, which allows for increased contact between humans and rats, and thus, increased disease risk. Another means in which rainfall can affect the incidence of disease which occurs over a smaller time scale, is concerning flooding events. In massive rainfall-induced flood events, rats can be displaced from their normal burrows into areas where there is potential for more human exposure (e.g. households, urban environments) (Madsen and Shine, 1999). This is perhaps the most common association found between rainfall and incidence of rodent-borne disease (Madsen and Shine, 1999). Lastly, agricultural activity is regularly dictated by seasonal variation in rainfall, which can increase exposure risk for agricultural workers. In Sri Lanka, agriculture makes up a large portion of the workforce. Such occupational exposures are thought to be the leading cause of infection of leptospirosis (Sri Lanka Epidemiology Unit, 2008).

1.3 Exposure Risk and Relation to Environmental Variables

To properly assess how environmental variables such as precipitation can affect the distribution and risk of EIDs, efforts must be made to assess whether there is significant relationship between these factors and increased EID risk to the human population. Leptospirosis has traditionally been associated with occupational exposures, with high incidence groups being farmers, miners, construction workers, and sewer workers. Studies have found that working in outdoor environments with exposure to sewage, floodwater, or mud, can lead to higher leptospirosis risk (Sarkar et al., 2002). Conditions of heavy rain and flooding will increase the amount of exposure to these listed outdoor factors, and thus will increase exposure risk. Activities such as gathering wood, grinding grain, and husking corn which would be performed on a daily basis for certain occupations, have also shown significant correlation with leptospirosis infection (Ashford et al., 2000). Reasoning behind this correlation can be drawn to heightened exposure to infected animals, and contaminated surfaces or mud. When precipitation levels are high, infected hosts such as rats, are often forced out of their regular burrows to areas of higher human contact (Madsen and Shine, 1999). This leads to an abundance of contaminated surfaces and mud, as the amount of rainfall causes the infected urine of animals to be dispersed throughout the environment. It is important to note that direct exposure to rats is not thought to be a significant agent for transmission, suggesting that the primary mechanism is through exposure to environments which are contaminated by the urine of rodent reservoirs (Sarkar et al., 2002).

Ashton et al. (2000) used multivariate logistic models to evaluate preventive measures against leptospiral infection. When assessing transmission of leptospirosis in developing countries where low socio-economic status dictates the availability of basic amenities, it was

found that having an indoor water source may be the most effective preventive measure against leptospiral infection, implying that ingestion of contaminated water can be a significant risk factor. In heightened rainfall events, large amounts of runoff from the groundwater tables can seep into outdoor water reservoirs and wells. If the surrounding environment was previously contaminated, all those obtaining and ingesting water from the reservoirs will be at increased risk of infection. Also, high temperatures can lead to need for higher water consumption, and thus can be considered in conjunction with this risk factor as a means of increased infection. Many of the environmental conditions Ashton et al. (2000) assessed are present in Sri Lanka, specifically after large seasonal monsoonal events, so they are of particular interest to consider when constructing models for the region. While leptospirosis risk may be indirectly affected by precipitation, it is clear that precipitation is important to consider when assessing risk factors of the disease.

Given recent trends such as climate change, increased extreme weather events such as floods, population growth, and urbanization, many have speculated leptospirosis incidence will continue to increase (Lau et al., 2010). Enhanced surveillance techniques must be used to understand how these environmental factors affect the transmission dynamics of leptospirosis. Space-time surveillance can be a potentially useful tool for estimating current and future disease burden as a result of environmental change (Robertson et al. 2010).

1.4 **Objectives**

There were two primary objectives of this research. Firstly, I looked to identify a significant relationship between rainfall and leptospirosis by evaluating important lag times between leptospirosis cases and weekly rainfall. Secondly, I aimed to use these identified lags to

develop a surveillance model for early warning of leptospirosis in Sri Lanka. In 2008, there was a country-wide outbreak of leptospirosis across much of Sri Lanka. By employing leptospirosis case count data from 2006 through 2010 to fit the models, I hoped to develop a framework that could in time be used by the Ministry of Health (MOH) of Sri Lanka to mitigate future leptospirosis outbreaks. By incorporating rainfall as the primary model covariate within the modelling framework, I assessed if it could be used to accurately predict leptospirosis outbreak events in Sri Lanka.

Practical considerations must be made when evaluating realistic methods for predicting disease outbreak in developing countries. Factors such as computational complexity and ease of use are of prime importance if the proposed methods are expected to be employed in any nation-wide surveillance system where financial and computational resources may be limited. It should also be noted that literature suggests that model parsimony is desirable unless the model has been found to be inadequate when compared to more complex models (Robson, 2014). The modelling approaches considered in this research were selected to take these factors into account so that once models were developed, they could be implemented and maintained by the Sri Lanka MOH workers with variable experience with probabilistic modelling and spatial data processing.

To accomplish the objectives set out, analysis was performed in several stages. Firstly, I compared rainfall and leptospirosis notified case counts at a weekly regional scale to identify optimal time lags for previous rainfall events for forecasting notable leptospirosis outbreaks. A variety of modelling scenarios were then evaluated that incorporated these varying lags of weekly rainfall as covariates to forecast leptospirosis risk. Lastly, I assessed model fit quality and prediction accuracy for three districts of Sri Lanka where there is a history of known leptospirosis outbreaks using several metrics to determine if the models produced were of high

enough quality to be used for nation-wide early warning in Sri Lanka. Through the use of geographic information systems (GIS), I evaluated model fit and forecast quality over time and space, which is of critical importance when building and operating spatially explicit disease risk models (Robertson, 2015).

2. Material and methods

2.1 **Modelling EID risk**

It is important to consider previous approaches to EID surveillance to be able to make an informed decision on the best methods to employ given the study area and research objectives. Model-based approaches to disease surveillance have been shown to be successful and yield accurate results in a variety of different research settings (Ashford et al., 2000; Chien and Yu, 2014; Guisan and Zimmermann, 2000; Held et al., 2006; Hii et al., 2012; Kleinman et al., 2004; Robertson et al., 2011; Tassinari et al., 2008).

Traditionally, disease surveillance has been carried out using standard hypothesis-testing statistical methods, where an outbreak is detected as a significant departure from the null hypothesis (Waller, 2003). In a study by Kleinman et al. (2004), a Generalized Linear Mixed Models (GLMM) approach was proposed to provide predictions of the expected number of cases in absence of an outbreak, and then compared to the observed number of cases. Though the objective of these two methods appear similar, hypothesis-testing statistics generate a "yes/no" (detection/no detection) binary answer to the research question, while the GLMM approach places much more emphasis on describing the pattern found in the data (Waller, 2004). In the field of disease surveillance, issues arise with traditional methods, as often the nature of surveillance is continuous through time with no discrete endpoints, and conducted for several

areas/outcomes simultaneously. While modelling does not necessarily negate these problems, it shifts focus to characterizing and understanding patterns in the data, and has the potential to accommodate expected values that vary over time. Thus, modelling could be better suited for exploring observed trends (Waller, 2004).

In Kleinman et al. (2004), GLMMs were used for detecting cases of acute lower respiratory infection as a method of early recognition of possible bioterrorism events. Logistic regression was used to predict the probability of a person being a case on a given day by using various predictors to describe the day of observation. It is important to take the spatial distribution of disease risk into account when modelling, as there could be areas where certain populations may be more likely to become infected than others. Kleinman et al. (2004) introduced the following formulae to account for changes in risk based on location, time and the individual:

$$E(y_{ijt}|b_i) = p_{ijt} \quad (1)$$

$$\text{logit}(p_{it}) = X_{ijt}\beta + b_i \quad (2)$$

where y_{ijt} denotes if person j in area i is a case on day t, p_{ijt} is the probability that he/she is a case, X_{ijt} is a set of covariates measured for person j and/or area i and/or day t, β is a vector of fixed effects, and b_i is a random effect for area i. While this model did consider different probabilities of disease risk for each region, it was not truly spatial, as it weighed the random effect for an area based on the population for that region, as opposed to the neighbouring areas (Kleinman et al., 2004).

GLMM techniques have been used in the past to assess risk factors for leptospirosis (Tassinari et al., 2008). In a study by Tassinari et al. (2008), a GLMM approach was taken to evaluating risk factors for comparison between cluster and non-cluster cases of leptospirosis in

Rio de Janeiro, Brazil. A cluster case was defined as a leptospirosis case that belonged to a cluster, which was found using spatial scan statistics (Tassinari et al., 2008). The study incorporated two spatial scales: individual level, and 32 Voronoi polygons that were situated around meteorological measurement stations. Cases were related to mean daily rainfall (measured at the nearest meteorological station) that had occurred anywhere between 3-20 days preceding the onset of symptoms. What was found was that the summer season – associated with high rainfall and flooding – had correlation to leptospirosis case clusters (Tassinari et al., 2008). When comparing a cluster case to a non-cluster case, it was found that a threshold value of greater than 4 mm mean daily rainfall had significant association with leptospirosis cluster events (Tassinari et al., 2008).

Another type of modelling that can be considered for disease surveillance is ecological niche modelling. The ecological niche of a species of disease can be modelled by evaluating relationships between observations of disease occurrence and predictor variables of the abiotic conditions present in that area (Guisan and Zimmermann, 2000). With the use of GIS, ecological niche models can produce spatially explicit predictions of the probability of EID spread to unsampled locations (Meentemeyer et al., 2008). Ecological niche models do not take into account the current distribution of the disease, but produce predictions based on underlying environmental conditions suitable for growth (Meentemeyer et al., 2008). Using this type of model can be very effective for early detection of disease outbreak, as the model is not dependent on the current disease distribution, but the environmental variables associated with disease occurrence.

When evaluating ecological niche models as a viable method for disease surveillance, it must be considered that they are prone to false positives, as they do not take the pathogen being

studied into account in their predictions. This can result in the flagging of many areas that are 'potentially' suitable for heightened disease risk, but in actuality, there may be no observed cases in these areas. In certain research contexts, these false positives may not be a major issue, but when working in a disease surveillance setting, false positives are not ideal, specifically when considering the implications of early warning and the associated financial costs. When looking to incorporate previous case counts to improve prediction accuracy, time-series based modelling methods offer an effective solution, as they account for temporal autocorrelation and seasonal variations in the data.

Time series regression models incorporate current and past observations of predictor variables ordered by time to predict the response variable (e.g., leptospirosis case counts). An important component of time series models is the use of the previous values of the response variable to predict future values (i.e., autoregressive dynamics). Incorporating an autoregressive component in a model is very useful when assessing any process where there is a strong serial dependence on previous values of the response, for example, infectious disease outbreak. Time series models take the general form:

$$y_t = X_t\beta + e_t \tag{3}$$

where y_t is the observed response at time *t*, X_t is a time-varying covariate vector, β represents the contributions of individual predictors, and e_t are the errors associated with *t*.

While incorporating spatial dynamics into a model can often yield accurate predictions, implementing a family of time series models for regions of a given study area can also produce meaningful, accurate results. Hii et al. (2012) implemented Poisson a multivariate regression model that incorporated mean temperature and cumulative rainfall as covariates to predict dengue fever incidence in Singapore. The forecasting model was developed to provide timely

early warning of dengue in Singapore. Weekly dengue cases from 2000 through 2011 were used in conjunction with daily mean rainfall to predict weekly cases of dengue for 2011 and 2012. Several time lags between dengue and weather variables were considered to determine the optimal period for dengue forecasting. A lag of 16 weeks was found to predict dengue most accurately. The model developed was able to distinguish between outbreak and non-outbreak events with 96% confidence from 2004-2010, and 98% confidence in 2011, and was able to predict an known outbreak in 2001 accurately with less than a 3% chance of false alarm (Hii et al., 2012). Perhaps the most intriguing aspect of this research was that Hii et al. (2012) were able to forecast dengue accurately with relatively simple models that only incorporated rainfall and temperature. This finding is of particular importance, as the models in this research are planned to be used for early warning of disease outbreak in the developing country of Sri Lanka. Data and computational resources for Sri Lanka are limited, and thus developing simple, low computational cost models could aid in national surveillance.

Due to the strong serial dependence that must be considered when assessing EID risk, and considering the financial and computational limitations that were discussed in the objectives, I chose to implement a family of time series multivariate regression models at multiple spatial scales: the MOH area level, and the district level (Figure 3.1).

2.1.1. Model construction

Given that one of the primary goals of this research was to detect an early warning signal for leptospirosis outbreaks, regions where notable outbreaks had occurred in the past were selected for detailed analysis, and regions where no known outbreaks had been reported were not evaluated (Figure 3.1). Three districts of Sri Lanka – Colombo, Kalutara, and Matale – were selected to represent regions of high and medium leptospirosis risk and were assessed in detail.

These districts were composed of 13, 10, and 12 MOH areas respectively. The models constructed for the MOH area level and the district level forecasted weekly leptospirosis cases and were fit by regressing on multiple independent variables that included retrospective leptospirosis cases, weekly total rainfall at varying time lags, lagged weekly cumulative rainfall, and seasonal factors. Models were analysed by employing a variety of performance metrics to determine if leptospirosis could be accurately forecasted using rainfall as the primary covariate.

We developed eight different multivariate regression models for each MOH area and district of interest in Sri Lanka. All models were integer-valued autoregressive conditional heteroskedasticity (INGARCH) models (Ferland et al., 2006; Heinen, 2003). INGARCH models used were of the general form

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}\left(Y_{t-i_k}\right) + \sum_{\iota-1}^{q} \alpha_\iota g(\lambda_{t-j_\iota}) + \eta X_t \tag{4}$$

where g is a link function, $\tilde{g}$ is a transformative function, Y_t denotes a count time series, X_t is a time-varying covariate vector, λ_t is the conditional mean, and η is a parameter vector corresponding to the effects of the covariates (Liboschik et al., 2015). The main advantages of using this type of model is that they are flexible, parsimonious, and generally easy to estimate using maximum-likelihood based methods (Heinen, 2003). Four INGARCH models with a negative binomial distribution as well as four INGARCH models with a Poisson distribution with varying permutations of covariates were then compared for each MOH area in a given district of study, and for each district. Table 3.1 provides an outline of all of the different models that were fit, and their respective covariates.

2.1.2. Serial correlation of leptospirosis cases and weekly rainfall

One notable characteristic that must be taken into account when modelling EID risk is the serial dependence of current cases on past cases (Hii et al., 2012). With regard to leptospirosis, this research hypothesized that there may also be dependence on prior rainfall. Probable lag times were estimated for serial correlation of leptospirosis cases and cross-correlation of rainfall by analysing data using autocorrelation functions (ACF) and cross-correlation functions (CCF), and reviewing qualitative findings on leptospirosis transmission. Seasonality of the dependent variable was captured by regressing on λ_{t-52} – the unobserved conditional mean of leptospirosis cases from 52 weeks (i.e., one year) before. CCF analysis of weekly rainfall and weekly leptospirosis case counts indicated the strongest correlation between leptospirosis and the current week's rainfall, and with rainfall at a lag of 23 weeks (i.e., correlation between rainfall at *t-23* and leptospirosis cases at *t*). To account for the overall wetness of the environment at time *t*, a moving windowed sum of cumulative rainfall from weeks *t-12* through *t* (i.e., the past three months) was incorporated as a model covariate. Using a three-month window would adequately account for overall wetness of the environment in the case of flooding after a major rainfall event.

2.1.3. Model evaluation criteria

To assess model performance, model fitted values were compared with actual observed values using the standardized root-mean-square error (SRMSE). The SRMSE can be defined as

$$\text{SRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{5}$$

where y_i is an observed value, and $\hat{y}_i$ is a predicted or fitted value. Standardizing the value by the mean of the observations is useful for comparing values from different data sets (e.g., different districts of Sri Lanka) and it provides a meaningful dimensionless measure which has

been used in model evaluation in a variety of different contexts including forecasting disease incidence (Chemel et al., 2011; Hii et al., 2012; Plouffe et al., 2015). This metric was found to be especially useful for evaluating prediction accuracy of the models at both the MOH area level and the district level, as by standardized the errors found, all models were able to be compared and contrasted without needing to consider the magnitude of the values being predicted (i.e., leptospirosis case counts).

For select MOH areas and districts of interest, the probability integral transform (PIT) was used to assess calibration of the respective model's predictive distribution. A predictive distribution can be thought to be correctly calibrated if events with a probability q occur a proportion q of the time on average (Gneiting et al., 2005; Jones and Spiegelhalter, 2012). The PIT will assess that this is true by checking that if you supply a random model variable into the model's respective cumulative distribution function (CDF), it will output a uniform distribution (Jones and Spiegelhalter, 2012). Gneiting et al. (2007) stress that uniformity of the PIT is a necessary but not sufficient indicator that a forecasting model is ideal. Jones and Spiegelhalter (2012) and Gneiting et al. (2007) suggest employing proper scoring rules to assess models' predictive distributions and model sharpness. A 'sharp' or well calibrated model should have high statistical consistency between its predictive distribution and its observations (Christou and Fokianos, 2015). Proper scoring rules provide numerical values (i.e., scores) that measure the predictive performance of the model, and are usually employed when looking to compare competing forecasting models (Christou and Fokianos, 2015).

We calculated both the mean logarithmic score and the ranked probability score (RPS) for all models being evaluated. The mean logarithmic score is defined as

$$-\log\big(f(y_i)\big) \tag{6}$$

where f is a predictive density function specific to the model, and y_i is an observed value. This is a commonly used scoring rule that has been employed in a variety of different modelling contexts, for example, weather forecasting (Bröcker and Smith, 2008). One notable issue when employing the mean logarithmic score is that it has been found to be highly sensitive to individual extreme cases (Gneiting and Raftery, 2007). Given the possibility of extreme individual case counts in the case of an outbreak event, other complimentary scoring rules can be used to help assess model calibration. The ranked probability score (RPS) has been recommended as a more robust scoring alternative (Christou and Fokianos, 2015; Gneiting and Raftery, 2007). The RPS is defined as

$$RPS(P_t, y_t) = \sum_{x=0}^{\infty} (P_t(x) - 1(y_t \leq x))^2 \tag{7}$$

where P_t is the forecast probability CDF for the time t, and y_t is an observed value. The average RPS is calculated over all modelling units to determine the model's mean RPS value. The mean RPS reduces to $|\widehat{y_t} - y_t|$ (Gneiting and Raftery, 2007), and as such can be considered a generalization of the mean absolute difference for probabilistic forecasting models (Jones and Spiegelhalter, 2012). This allows for easy evaluation and interpretation of the mean RPS. Both of the scoring rules used to assess model calibration are 'proper', in that the calculated score is minimized if one's beliefs are reported honestly (Jones and Spiegelhalter, 2012). Lastly, the Akaike information criterion (AIC) was evaluated for all models in the selected districts. The AIC measures the relative quality of a model for a given set of source data, and includes a penalty term for higher model complexity that favours model parsimony. In practice, I attempted to minimize the values of all the above model evaluation metrics to produce the best calibrated models for the selected areas of study.

To assess the values predicted by the leptospirosis risk models in a surveillance context, a Cumulative Sum (CUSUM) analysis was performed between the predicted and observed values for the best performing model based on the criteria above. The objective of CUSUM analysis is to detect a change (e.g., an outbreak) given an underlying process (Robertson et al., 2010). For a given region, a moving sum of deviations for each time period is calculated as follows:

$$S_t = \max(0,\ S_{t-1} + y_t - k) \tag{8}$$

where S_t is the cumulative sum alarm statistic, y_t is the case count at time *t*, and *k* represents the slack term which allows one to adjust the sensitivity of the CUSUM analysis. Observed counts that exceed *k* are then accumulated, and an alarm is triggered if S_t is greater than a set threshold parameter *h* (Robertson et al., 2010). CUSUM analysis can be used as a concrete decision support tool to use a model to flag time periods of possible outbreak, and signal an alarm for early warning of leptospirosis.

2.2 **Study area**

Sri Lanka is a country that is found off of the southeastern coast of the Indian subcontinent. Sri Lanka's climate is tropical, and annual seasonal variations in weather are characterized by the northeast monsoon and the southwest monsoon. The northeast monsoon generally begins in December, and last until the following February, whereas the southwest monsoon begins in April, and lasts until September. The most populous area of Sri Lanka is located in the southwest (e.g., Colombo), which experiences heavy rainfall particularly during the southwest monsoon. The less populous northern and eastern areas of Sri Lanka become predominantly dry during the southwest monsoon, and generally have not experienced leptospirosis outbreak events of the same magnitude as the areas located in the southwest. Two separate inter-monsoonal rainfall seasons – during which Sri Lanka can experience relatively

large amounts of convectional rainfall – last from March until April and from October until November. Leptospirosis is endemic across Sri Lanka, but there were notable leptospirosis outbreaks in the years of 2008 and 2009. Table 3.2 outlines leptospirosis case counts by year for all of Sri Lanka, and case counts at the district level by year.

2.3 **Data**

2.3.1. Leptospirosis data

Reported and confirmed leptospirosis weekly case counts for the years of 2006 through 2010 were obtained from the Epidemiology Unit of the MOH in Sri Lanka. These data were aggregated by MOH administrative areas and by district (Figure 3.1). Weekly counts were separated into two categories: reported leptospirosis cases (i.e., suspected cases), and confirmed leptospirosis cases (i.e., clinically tested cases). A reported leptospirosis case was recorded if an individual visited a clinic and exhibited symptoms associated with leptospirosis, whereas a confirmed leptospirosis case was recorded when an individual's blood and urine were serologically tested and a leptospiral infection was confirmed. In this research, I opted to use the reported case counts in the models developed. When looking to detect an early warning signal in a leptospirosis surveillance context, minimizing the amount of time between when a person contracts leptospirosis and when that case is first recorded is of importance, and thus the reported cases were selected to be used in the models over the confirmed cases. A drawback of using reported cases is that there is a higher degree of uncertainty introduced into the models, as they were recorded based only on clinical suspicion, which is variable. To be recorded as a confirmed case, subjects were required to get tested at a local clinic, which in some of the more rural areas of Sri Lanka, were not easily accessible. It is believed that this may have led to a slight

underestimation of actual leptospirosis cases, and thus, for all the reasons stated, the reported case counts were preferred for fitting the models.

2.3.2. Rainfall data

Rainfall data were obtained from the Department of Meteorology of Sri Lanka, and consisted of daily rainfall measurements recorded in millimetres for the years of 2006 through 2010. These data were collected from two separate meteorological station networks: a network of small-scale community-managed weather monitoring stations, and a network of official meteorological stations maintained by the Department of Meteorology in Sri Lanka.

The network of small-scale community-managed stations was composed of ~370 weather monitoring stations (varying by year), many of which were located in agricultural areas. The spatial distribution of these stations varied considerably based on factors such as population, climate, and land use. The quality of measurements taken from these stations could not be verified, as many stations were located in remote areas of Sri Lanka where station maintenance was situational. The network of official meteorological stations was composed of 20 to 22 meteorological stations (varying by year), where measurements were verified as accurate by the Department of Meteorology of Sri Lanka. These stations were irregularly distributed across the country, with the majority of stations being situated in the more populous southwest region of Sri Lanka.

Data obtained from both station networks were combined into a master rainfall data set for use in this study. This data set was then aggregated by week and used to interpolate weekly rainfall for the years of 2005 through 2010. In previous research (Chapter 2), the quality of the network of community-managed weather stations was evaluated for use in a modelling context

by comparing interpolated rainfall surfaces to remotely-sensed imagery of Sri Lanka (Plouffe et al., 2015). The network of community-managed stations was found to be as viable a data source as the network of official meteorological stations for building interpolation models so long as the network of community-managed stations had adequate spatial coverage across the area being interpolated (Plouffe et al., 2015). Many of the gaps in coverage from the network of community-managed stations were accounted for by incorporating data from the network of official meteorological stations, which was the reasoning behind combining both rainfall data sets for interpolating rainfall in this research. Several different interpolation methods were also evaluated for use in context of predicting rainfall in Sri Lanka. Findings indicated that when magnitude rainfall was low, Bayesian kriging performed best, whereas in high rainfall conditions, thin-plate smoothing splines produced the most accurate rainfall predictions (Plouffe et al., 2015). Since rainfall magnitudes at the temporal scale (i.e.., weekly) investigated were considered low with respect to the monthly rainfall totals examined in the previous study, Bayesian kriging was used to produce the most accurate rainfall predictions.

Once weekly interpolations were produced, a spatial mean of rainfall for each MOH area was extracted from the interpolated surfaces for each week included in the study period (i.e., 2006 to 2010), and included in the leptospirosis risk models. It should be noted that due to the civil war between the Sri Lankan Army and the Tamil Tigers (LTTE) during the period of study (e.g., the year of 2008), there were many missing rainfall measurements in the north of the country due to lack of maintenance at community-managed weather stations. These null measurement values were not considered in the interpolations, and may have had an effect on the quality of the produced interpolations, however Bayesian kriging can leverage prior distributions and nearby values to 'fill-in' data gaps. These effects were assumed to be minimal, as most of

the missing rainfall values were in the north where leptospirosis outbreak was much less pronounced than in the southwest.

2.4 **Software**

Several different types of software were used in this research for data management, data processing, and modelling. The programming language Python was used for initial extraction and parsing of rainfall and leptospirosis data from large text-based files. The statistical programming language R (version 3.1.1) was used for all other analyses, including data processing and aggregation, interpolating rainfall, and modelling disease risk. The R packages gstat (Pebesma, 2004) and geoR (Ribeiro Jr and Diggle, 2001) were used to produce interpolations, and the package tscount (Liboschik et al., 2015) was used for model construction and scoring.

3. Results

Total leptospirosis case counts by district from 2006 to 2010 are presented in Figure 3.2. The districts located in the southwest region of Sri Lanka experienced much higher case counts than districts in the north and the east. Weekly leptospirosis case counts and weekly rainfall for each district being assessed in this study (i.e., Colombo, Kalutara, and Matale) from 2006 to 2010 are presented in Figure 3.3. Generally, no common trends for leptospirosis cases by district can be observed, whereas periods of heightened rainfall can be seen to be more congruent between the districts being studied. Certain spikes in number of leptospirosis cases (e.g., the second half of the year of 2009) can be observed in multiple districts at the same time, which can be thought to signify an epidemic period of leptospirosis outbreak. Visual inspection of Figure 3.3 for meaningful lags between the amount of rainfall and the number of leptospirosis cases

does not yield any easily discernable pattern, but for the year of 2008, it can be seen that rainfall from the southwest monsoon midway through the year may have corresponded with increased leptospirosis incidence in the latter half of the year. This would be consistent with the correlation observed between rainfall at *t-23* and leptospirosis cases at *t* during the CCF analysis. Having perspective on these global trends can help explain some of the behaviour exhibited by the models that were evaluated in this study.

3.1 **Model selection**

3.1.1. MOH area level

To keep analysis concise, all modelling scenarios being analysed in this study will be referred to by the letter assigned to them in Table 3.1. ACF analysis and visual inspection of leptospirosis case counts were used to determine that regressing on the previous four weeks of observations would be suitable to account for serial dependence. To compare modelling scenarios and select the most appropriate model for further analysis, ranks of the various model evaluation metrics were assessed. If any of the values being ranked were sufficiently similar (i.e., were equivalent up to three decimal places), they were assigned the same rank. Ranks were attributed in ascending order with a rank of 1 indicating a model attained the best result from all modelling scenarios. In all model assessment tables for MOH areas, MOH areas 101 through 113 correspond to Colombo, 301 through 310 correspond to Kalutara, and 501 through 512 correspond to Matale. MOH area 512 exhibited very low leptospirosis case counts (i.e., under 20 leptospirosis cases for the entire period of study) and thus was not included in this analysis, as the primary goal was to effectively model areas that experienced an outbreak of leptospirosis and detect epidemic status. First, the ranks of SRMSEs between observed and predicted case counts will be assessed.

Table 3.3 presents SRMSE ranks for all fitted models for each MOH area in each district of study. A trend that is immediately apparent is that all models fit with the same covariates for both negative binomial and Poisson distributions attained the same rank. In general, models D and H (i.e., models incorporating rainfall, rainfall lagged by 23 weeks, summed cumulative rainfall, and leptospirosis cases lagged by a year as covariates) attained the lowest SRMSEs relative to other models for their respective MOH area. Models D and H received the highest rank in 15 of the 34 different MOH areas, and the second highest rank in 11 of the MOH areas. By assessing the total summed rank at the bottom of the Table 3.3, it can be observed that models D and H achieved a lower SRMSE value than other competing models the majority of the time. Models A and E (i.e., models incorporating only rainfall and rainfall lagged by 23 weeks) consistently received the highest SRMSE, and in 20 of the 34 MOH areas, were ranked as the worst performing models when being assessed using SRMSE.

Comparing the ranked RPSs (Table 3.4) followed some of the same trends as the ranked SRMSEs, but there was a much more pronounced difference when comparing between models that were fit using negative binomial regression versus Poisson regression. Overall, model D attained the lowest RPS in the majority of the MOH areas, and was the best calibrated model in 19 of the 34 MOH areas. When comparing the RPSs attained by models A, B, C, and D, which were fit using negative binomial regression, to models E, F, G, and H, which were fit using Poisson regression, a very distinct trend is apparent. In the majority of the MOH areas, the negative binomial models exhibited lower RPSs than any of the Poisson models. This trend is especially evident when evaluating the summed total RPS ranks at the bottom of Table 3.4, where the most poorly calibrated negative binomial model, model A, had the same summed total RPS as the as the best calibrated Poisson model, model H. This indicates that the leptospirosis

count data being used to fit the models were likely overdispersed, as negative binomial models are known to account for overdispersion better than Poisson models. Overdispersion in the count data seems like a likely possible occurrence, given the highly variable nature of leptospirosis counts, and that the mean case counts at the MOH area level was likely very close to zero provided that there were very few if any cases of leptospirosis when an outbreak event was not being observed (i.e., most weeks of the study period). Overall, the worst performing model, model E, was ranked last in RPS in 15 of the 34 MOH areas.

The ranked mean logarithmic scores followed an almost identical pattern to that of the ranked RPSs. The four negative binomial regression models consistently produced lower ranked mean logarithmic scores than the Poisson regression models, with models C and D generally ranked first or second. Model D yielded the lowest mean logarithmic score in 19 of the 34 MOH areas, while model C yielded the lowest mean logarithmic scores in 13 of the MOH areas.

Lastly, the ranked AIC was evaluated for each MOH area to best select the appropriate model. Many of the same trends that were observed for the ranked RPSs and ranked log mean scores were also present in the ranked AIC. All of the negative binomial regression models generally attained lower AIC values than the Poisson regression models, with very few exceptions. A notable difference is that model C actually yielded the lowest AIC values in 21 of the 34 MOH areas, where as model D only ranked first in 8 MOH areas. AIC is a useful indicator of goodness of fit, but it also penalizes for incorporating more covariates in a given model, which is likely the reason for model C's comparatively low AIC value to with respect to model D. I opted to calculate the mean AIC value yielded from all MOH areas to elucidate if model C's higher AIC rank was a realization of a notably lower AIC score. Mean values for all model assessment metrics can be found in Table 3.5. What can be deduced by inspecting Table

3.5 is that the mean AIC value for model C (512.22) was only marginally lower than model D (513.46), and this difference is orders of magnitude smaller than the observed mean AIC values thus making it negligible.

As a last means of model selection, several PIT histograms were produced as spot checks to assess calibration of the respective model's predictive distribution. Figure 3.4 depicts three PIT histograms comparing models C, D, and H for MOH area 102. Model D and C both exhibit uniformity across the histogram indicating that the models were properly calibrated, although model D does exhibit slightly stronger uniformity than model C. Model H exhibits a strong U-shaped histogram, which indicates that the predictive distribution (i.e., in this case, the Poisson distribution) is underdispersed (Czado et al., 2009; Dawid, 1984). This confirmed the supposition that the leptospirosis case count data were likely overdispersed at the MOH area level, and that a Poisson regression model would not be able to adequately account for that.

3.1.2. District level

The same trends that were observed at the MOH area level were also observed at the district level. The negative binomial models consistently outperformed the Poisson models when evaluating model calibration using either of the proper scoring rules and when evaluating the AIC. Model D generally yielded the lowest values, but with a few exceptions, e.g., the AIC produced for model B for the Kalutara district was slightly lower than the AIC for model D (1435.699 and 1437.007, respectively). The SRMSEs at the district level were found to be less consistent, which is thought to be due to the fact that only one model was being evaluated for each district. When finding the mean SRMSE over all MOH areas, I was able to assess how well each model predicted over a larger sample size.

One notable difference between the MOH area level and the district level models was the PIT histograms. At the district level, model D still exhibited the most uniform PIT histograms, but they were not of the same level of uniformity as many of the models that were spot checked at the MOH area level. Figure 3.5 presents model D's PIT histograms for each of the three districts being studied. Model D for Colombo produced an inverse-U shaped histogram, which may signify that the data were underdispersed, while both Kalutara's and Matale's histograms were relatively uniform. Unfortunately, the PIT histograms for the Poisson models demonstrated that the models were not able to account for this less dispersed data, and still produced strong U-shaped histograms.

By taking all model evaluation metrics into account at both spatial scales, it was deduced from this model selection analysis that model D (i.e., a negative binomial model that incorporated rainfall at *t*, rainfall at *t-23*, leptospirosis at *t-52*, the sum of cumulative rainfall, a regression on the previous four observations to account for serial dependence, and a regression on unobserved conditional mean at *t-52*) was the best fitting and best calibrated model at the MOH and district level, and that it would be able to perform the most consistently across both spatial scales being assessed. As such, it was employed in all further modelling and analysis.

3.2 **Model assessment**

Further analysis was carried out to assess how well the selected model was calibrated and how accurately it was fitted in different scenarios. Figure 3.6 depict SRMSEs between fitted and observed leptospirosis case count values mapped for each MOH area (labelled by MOH ID) that was assessed. MOH area 512 was not included in this analysis for reasons noted earlier, and thus it is coloured grey on the map to reflect that. Models constructed for MOH areas 105 and 106

both produced some of the highest SRMSEs of any of the MOH areas. Table 3.6 presents the exact values for all model assessment metrics used in this study. Interestingly, MOH area 106 had the highest SRMSE of any of the MOH areas, but it also had the lowest value for AIC, RPS, and the mean logarithmic score, which demonstrates that even if a model is thought to be properly calibrated, it can still perform quite poorly when assessing accuracy of prediction. The opposite can also be true, and interestingly, the MOH area with the highest value for both of the proper scoring rules and the AIC also was one of the lowest SRMSEs. To gain a better understanding of these MOH areas, a time series plot of fitted and observed values from each model (Figure 3.7) was investigated. By inspecting predicted and observed leptospirosis cases over time, it can be seen that the total number of cases in MOH area 106 was very low. Given that SRMSE is standardized to minimize the effect of magnitude on the observed error, the SRMSE is very high, as the model is underpredicting leptospirosis cases during periods of relatively high leptospirosis incidence. This MOH area is a good example of a situation where the magnitude of the response variable being predicted must be taken into account. While the SRMSE is high for this particular MOH area, it is not of importance to this study, as predicted values in MOH areas where there are very few leptospirosis cases does not need to be considered when looking to develop an early warning system for outbreak events.

The model for MOH area 106 can easily be contrasted with the model for MOH area 103 (Figure 3.7). MOH area 103 exhibited the highest leptospirosis case counts of any MOH area in Colombo by a large margin (456 cases more than the next highest MOH area), with 1106 cases observed between 2006 and 2010. Upon inspection, it can be seen that the observed versus fitted case counts for this MOH area during periods of outbreak (e.g., early and late 2008, late 2009), were quite different, with the model considerably underpredicting the magnitude of the number

of cases. This underprediction of case counts is thought to be the primary reason why the highest values in both of the proper scoring rules and AIC across all MOH areas in this study were observed. Other than the two noted outbreak events, the model fit well to the observed values. The SRMSE calculated for MOH area 103 was one of the lowest observed, which gives strengths to the idea that while the model may not be calibrated as well for this area, it was still fit relatively well if the extreme magnitudes of the case counts were not taken into account. When evaluating the plot, this is evident, as the fitted values closely approximated the more global trends of the observed values – just not to the same magnitude.

We also evaluated each of the three models that were fit at the district level. Figure 3.8 depicts fitted and observed leptospirosis case counts over the study period. Many of the same trends that were present in the MOH area models were also present in the district level models – while model fit approximated the overall pattern of the observed values, each model tended to underpredict as case counts approached more extreme magnitudes. This tendency was much less pronounced in the districts of Kalutara and Matale, but it was noticeable in Colombo during a time of major outbreak (e.g., late 2009).

Overall, models fit at the MOH and district level varied in quality of fit and prediction depending on several factors, such as the number of leptospirosis case counts, and how fast the onset of an outbreak event was. The effect of rainfall on outbreak events tended to be minimal in both the family of MOH area models and the district models, with the standard error of the rainfall-related covariates often being larger than the estimates themselves. For example, the model fit for the district of Colombo had estimates of $1.92e^{-8}$, $2.34e^{-5}$, and $2.92e^{-7}$ for rainfall, rainfall at *t-23*, and the windowed cumulative sum of rainfall, respectively. The standard errors associated with these estimates were $4.17e^{-4}$, $4.03e^{-4}$, and $1.03e^{-6}$, respectively, all of which are

larger than the estimates they are associated with. This finding indicates that there was no observable or meaningful relationship between leptospirosis incidence and rainfall in Sri Lanka during the study period.

3.2.1. Prediction accuracy case study

To determine if the models developed in this study were of high enough quality to be employed as a means of early warning for leptospirosis in Sri Lanka, the district level model for Colombo was refit for 2006 and 2007,and was then used to forecast weekly leptospirosis cases from 2008 to 2010 using 1-step-ahead and 2-step-ahead prediction. Colombo's model was selected as it represented an area where a major leptospirosis outbreak had occurred. Refitting the model for the years of 2006 and 2007 was performed so that the effectiveness of the model to detect outbreak events could be assessed. In more ideal conditions, the models already produced could be used to forecast a more recent outbreak event, but due to the unavailability of more recent data, this approach was thought to be the next best alternative. The precision of the forecasted values were analysed by comparing the predicted leptospirosis case counts to the observed. Predicted and observed leptospirosis case counts for Colombo from 2008 to 2010 can be seen in Figure 3.9. Similar trends to those present in the fit models are noticeable when comparing the predicted case counts to the observed. In outbreak events, the model tended to underpredict the number of cases, with a noticeable lag in prediction. Otherwise, the model did approximate the trends in observed values, with no notable periods of time where the model consistently overpredicted the number of leptospirosis cases.

One of the primary objectives of this research was to determine if an early warning signal could be detected for leptospirosis before the onset of an epidemic state (i.e., an outbreak). This was assessed this by performing a CUSUM analysis to determine if periods of outbreak could be

effectively predicted by the Colombo model, and identified as periods of outbreak to signal an alarm. The CUSUM analysis was performed using predicted values from the Colombo model and previously observed values from 2008 to 2010. The *k* (as defined in the Methods sections) for the CUSUM analysis was set to 5, to allow for leniency in sensitivity of the analysis up to plus or minus 5 case counts per unit of time. Setting *k* to 5 ensured that the CUSUM analysis was not oversensitive to slight under- or overpredictions of leptospirosis cases by the Colombo model when looking to identify periods of leptospirosis outbreak. Figure 3.10 depicts the results from the CUSUM analysis, indicating points in time to signal an alarm of a potential outbreak event. What can be seen is that the Colombo model was reasonably effective in determining states of outbreak, and signaling an alarm relatively early at the onset of an outbreak event. While the magnitudes of outbreak events were not adequately predicted by the model, it was able to effectively detect periods of outbreak, which is of prime importance when determining if a model can be used to detect an early warning signal of leptospirosis outbreak.

4. Discussion

The goals of this research were to 1) elucidate the relationship between rainfall and leptospirosis incidence, and 2) assess whether models could be developed to provide early warning for leptospirosis in Sri Lanka. What was found was that rainfall was not a significant predictor of leptospirosis risk, and that trying to use rainfall as a proxy for other factors that influence leptospirosis risk may not be an adequate way to explain variations in leptospirosis incidence within the model. Instead, identifying specific mechanisms that are believed to have a more direct effect on leptospirosis risk may be a more suitable approach. Figure 3.11 presents a theoretical leptospirosis risk model that incorporates many different mechanisms of leptospiral transmission that were not explicitly considered in the models developed in this research.

Accounting for environmental factors such as presence of animal reservoirs, and type of land cover may be attainable if appropriate data sets can be obtained. Also, consideration of different mechanisms of outbreak, e.g., cultivation of previously uncultivated land with large preexisting rodent reservoir populations, could help lead to more accurate model predictions.

In spite of rainfall not being a significant predictor of leptospirosis incidence, simple models that could be used in an early warning context in Sri Lanka were still able to be developed. It was important in this research to limit the complexity of the models that were developed and to fit models as best as possible given the relatively limited number of covariates available. While it is believed that the models were calibrated as best as possible given the data available, issues still arose when attempting to characterize trends across many different regions (i.e., MOH areas across Sri Lanka). This study aimed to find the best overall model that could be fit for many different regions, each of which had its own unique environmental conditions. This led to certain regions' models performing much differently than others.

While one of the research questions of this study was based on rainfall as a predictor for leptospirosis outbreak, there is not an abundance of prior academic evidence proving or denying this as a meaningful relationship. Pappachan et al. (2004) studied leptospirosis risk in the Indian state of Kerala for the year of 2002, and found that there was a strong relationship between heavy rainfall events and the onset of leptospirosis cases at a 7 to 10 day lag. While this is an interesting finding, the amount of leptospirosis cases was quite low when compared to the study area that was evaluated in this paper, and the period of study was also much shorter. The models developed for this research were able to predict leptospirosis incidence accurately when dealing with a low number of leptospirosis cases.

Chadsuthi et al. (2012) developed an Autoregressive Integrated Moving Average (ARIMA) to study seasonal trends of climate factors and their effects on leptospirosis incidence in Thailand. They noted the strong seasonality of reported leptospirosis cases that was present year round. When inspecting Figure 3.3 from this research, there is a weak seasonal pattern of leptospirosis outbreak at the district level that shows both major outbreak events occurring in the latter half of the year, but there is no strong seasonal trend in total weekly rainfall, especially when looking for a meaningful lag at which leptospirosis incidence increased. Instead, serial dependence on previous leptospirosis cases tended to play a much more pronounced role than correlation with previous rainfall, especially when experiencing outbreak conditions. This can be seen when assessing any of the fitted or predicted case counts compared to the observed, and was reflected when evaluating the proper scoring rules. For example, calculating the Pearson correlation coefficient between RPS and leptospirosis cases for the district level model of Colombo yielded a value of 0.99, whereas the correlation between RPS and rainfall at *t-23* for that same model yielded a value of 0.11.

The CUSUM analysis in this study indicated that by using predicted values from the leptospirosis risk models, it was possible to detect an early warning signal for outbreak events, even if the number of leptospirosis cases during outbreak were underpredicted. It is a notable finding that the models produced were able to accurately assess periods of leptospirosis outbreak, and employing these models in an early warning context would aid with signaling alarms when there is a high probability of a future outbreak occurring in the near future for a given district or MOH area. This is an important finding, as one of the primary objectives of this research was to be able to develop models suitable for early warning of leptospirosis in Sri Lanka.

While data for other covariates to include in the leptospirosis risk models (e.g., if a location is within an agricultural area, the distance of a location from rivers and animal reservoirs) are not easily obtainable for Sri Lanka, it is believed that the inclusions of other meaningful predictors into the models would yield better results, and allow for more effective prediction of leptospirosis cases with a shorter lag time, and with higher sensitivity to the number of cases. Even considering this, the models that were produced were able to detect early warning signals of leptospirosis outbreak effectively, and will be recommended as a starting point for a nation-wide leptospirosis surveillance system in Sri Lanka.

To use the models that were constructed in this research and present the results in a meaningful way, development of a graphic user interface (GUI) for Sri Lanka MOH workers that would allow for dynamic calibration of models without any programming would be a useful extension unto the research outlined in this paper. The R package Shiny, which provides a framework for developing Web applications that are powered by R in the backend, would be a suitable medium for developing such a GUI, and will be considered for future research projects (Chang et al., 2015).

5. Conclusions

In this study, it was assessed whether a modelling approach could be taken to forecasting leptospirosis incidence in Sri Lanka. Firstly, meaningful lags between rainfall and leptospirosis were identified by performing correlation analyses. A notable dependence between rainfall events and leptospirosis cases was found at a lag of 23 weeks.. Next, several INGARCH time series regression models were evaluated and compared at two different spatial scales: the MOH area level, and the district level. Numerous model calibration and predictive quality metrics

were employed to select the most performant model from a set of modelling scenarios. The best calibrated model was a negative binomial INGARCH model with covariates for rainfall at *t*, rainfall at *t-23*, leptospirosis at *t-52*, and a windowed cumulative sum of rainfall from the past 12 weeks. What was found was that models at the MOH area level and district level were able to approximate trends in leptospirosis outbreak quite well, and also predict periods of high and low leptospirosis risk with a reasonable degree of certainty. District level models were validated by a CUSUM analysis of the predicted case counts versus previously observed leptospirosis cases, and demonstrated that the models developed in this research could be beneficial if employed in an early warning context in Sri Lanka. Interestingly, it was found that the suspected relationship between rainfall and leptospirosis incidence in Sri Lanka was insignificant, and in future modelling efforts, more data concerning numerous other covariates will be obtained to help capture significant relationships that were not explained by the models in this research. While the models presented in this study were able to adequately provide early warning for leptospirosis in Sri Lanka, it is hoped that the methods demonstrated and insights gained during this study can be used in the future to help provide early warning for leptospirosis and other waterborne EIDs.

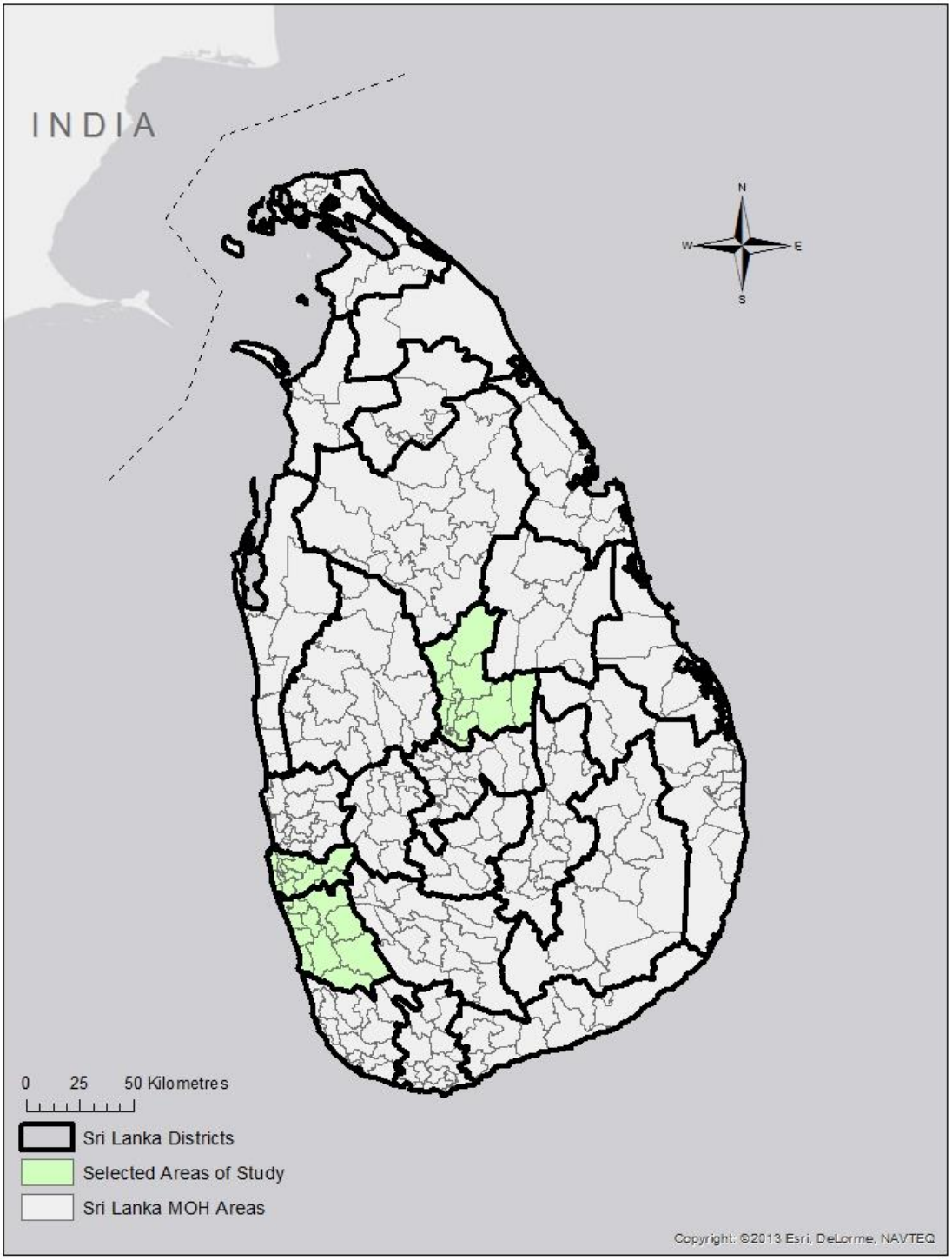


Figure 3.1. Map of Sri Lanka with districts and MOH areas. Areas that were selected for extensive analysis given known leptospirosis outbreak events during the study period are highlighted in green.

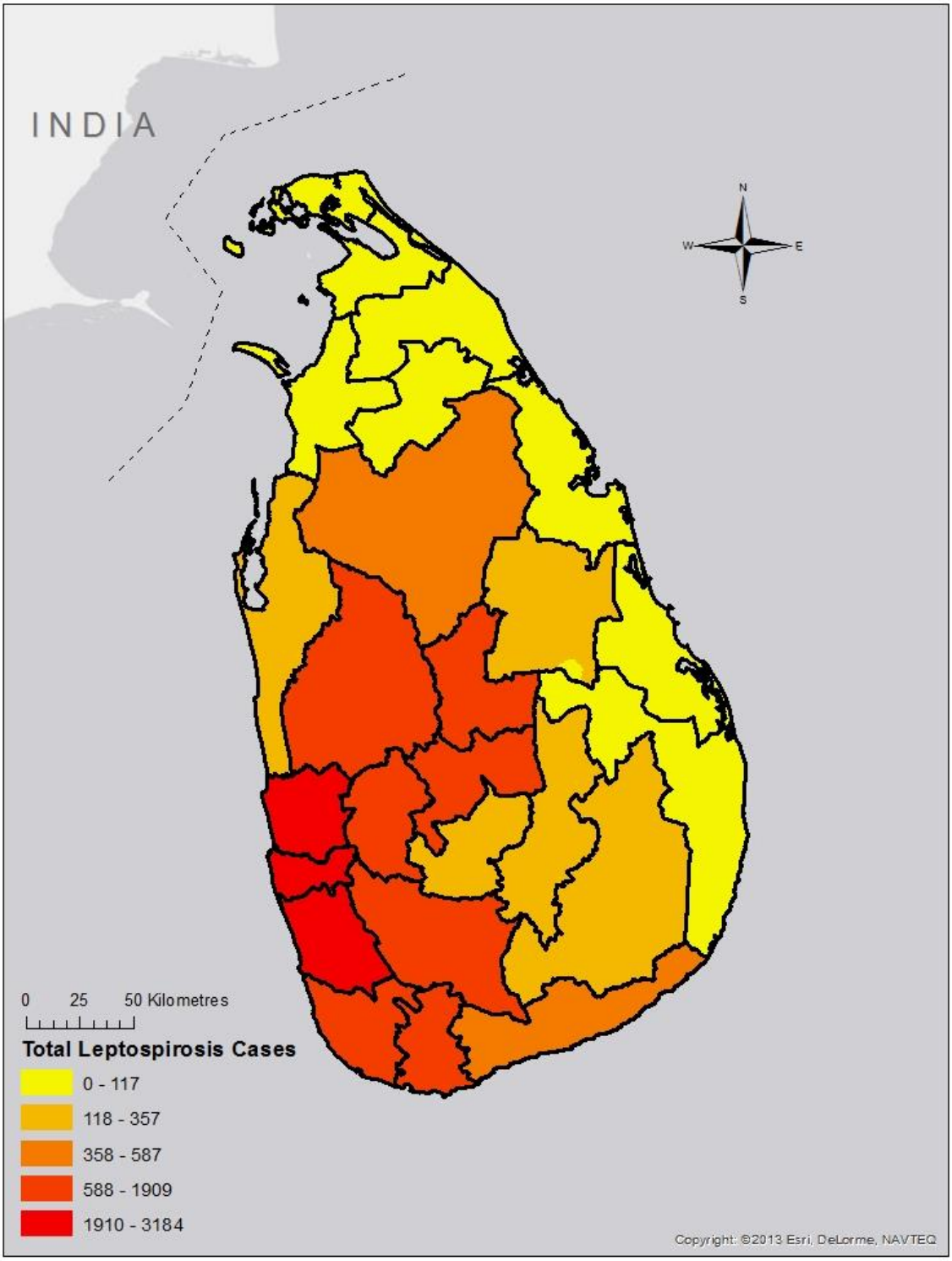


Figure 3.2. Map of total leptospirosis case counts from 2006 to 2010 by district.

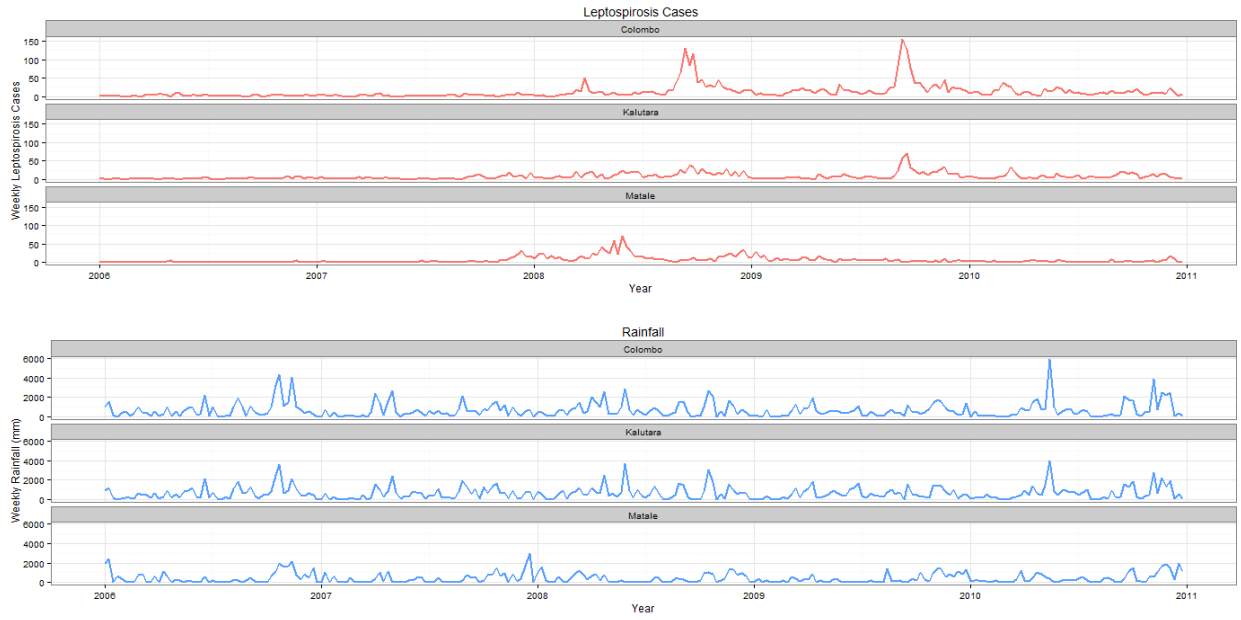


Figure 3.3. Weekly leptospirosis case counts and rainfall for each district of study from 2006 to 2010.

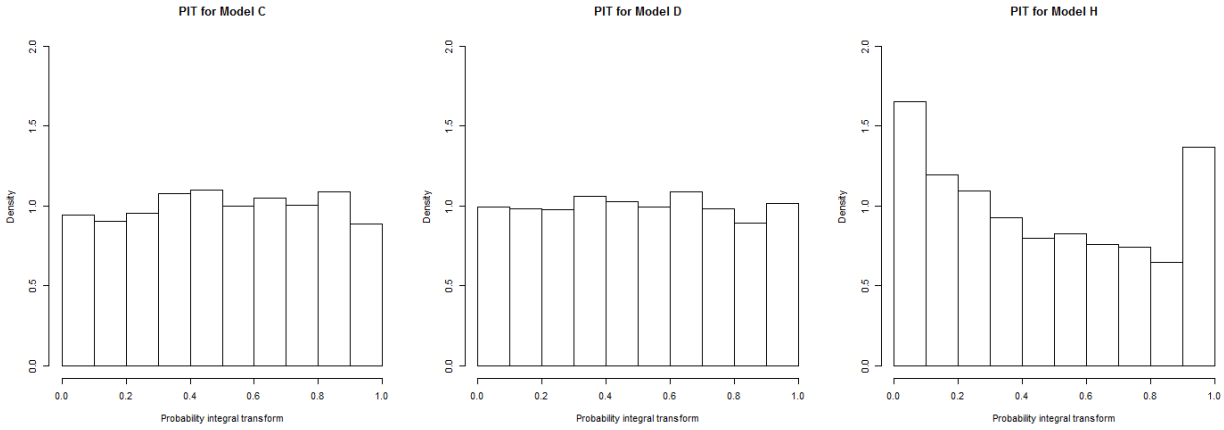


Figure 3.4. PIT histograms comparing models C, D, and H for MOH area 102.

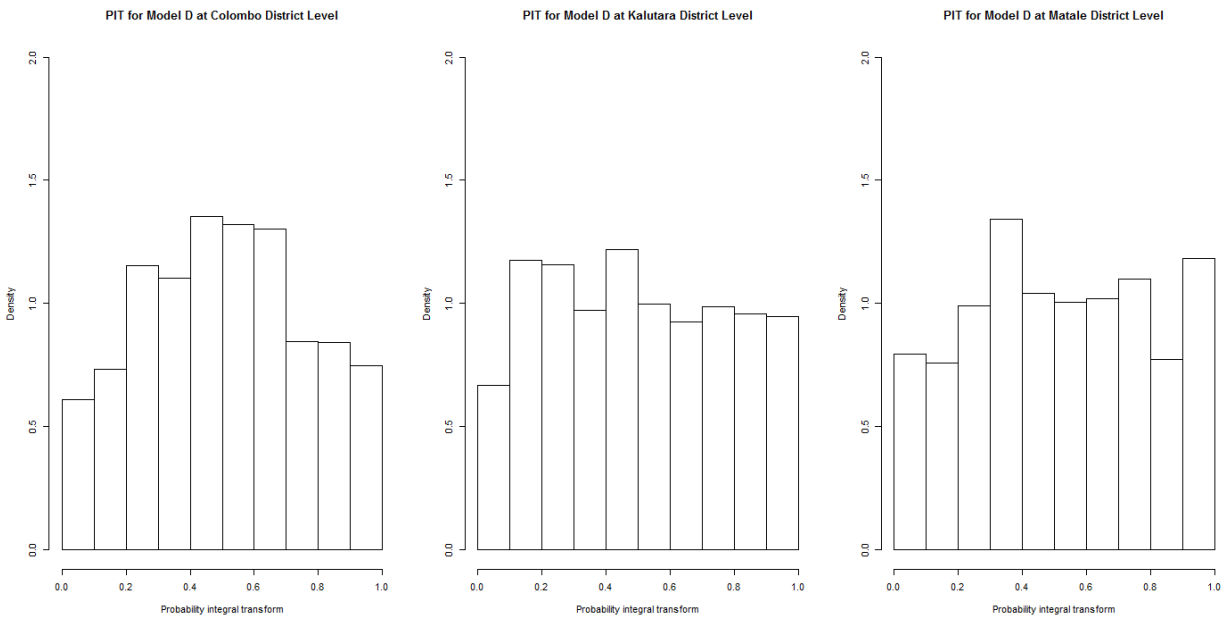


Figure 3.5. PIT histograms for model D in Colombo, Kalutara, and Matale.

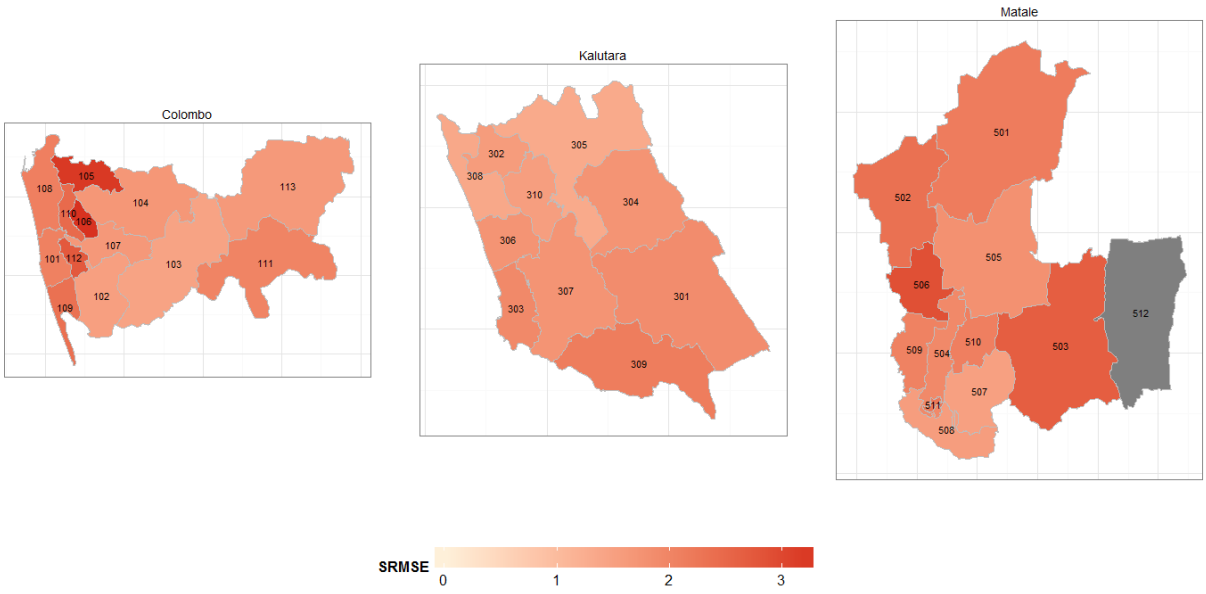


Figure 3.6. SRMSEs between fitted and observed leptospirosis case count values mapped for each MOH area (labelled by MOH ID).

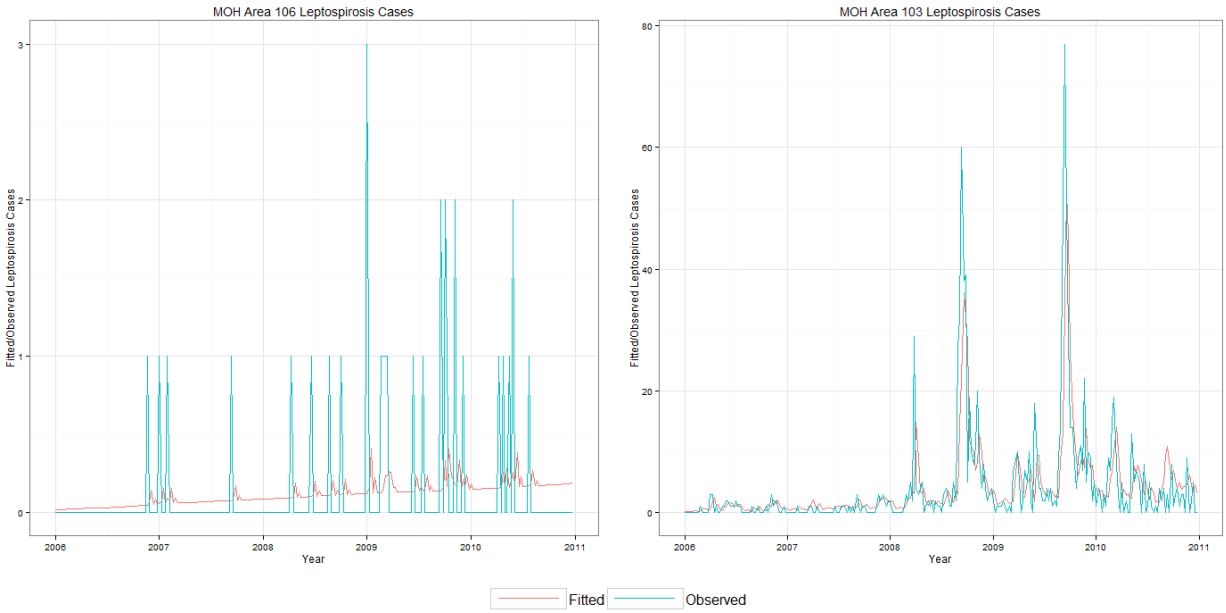


Figure 3.7. Fitted and Observed leptospirosis cases from 2006 to 2010 for MOH area 106 and MOH area 103.

Figure 3.8. Fitted and observed leptospirosis cases from 2006 to 2010 for Colombo, Kalutara, and Matale district level models.

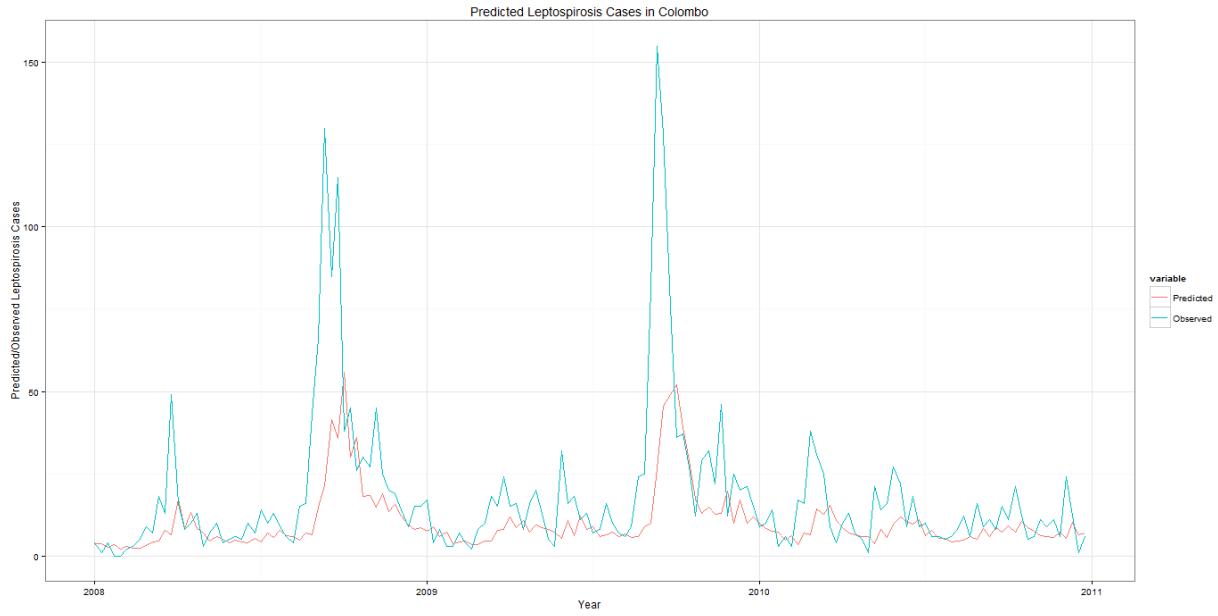


Figure 3.9. Predicted and observed leptospirosis cases in Colombo from 2008 to 2010.

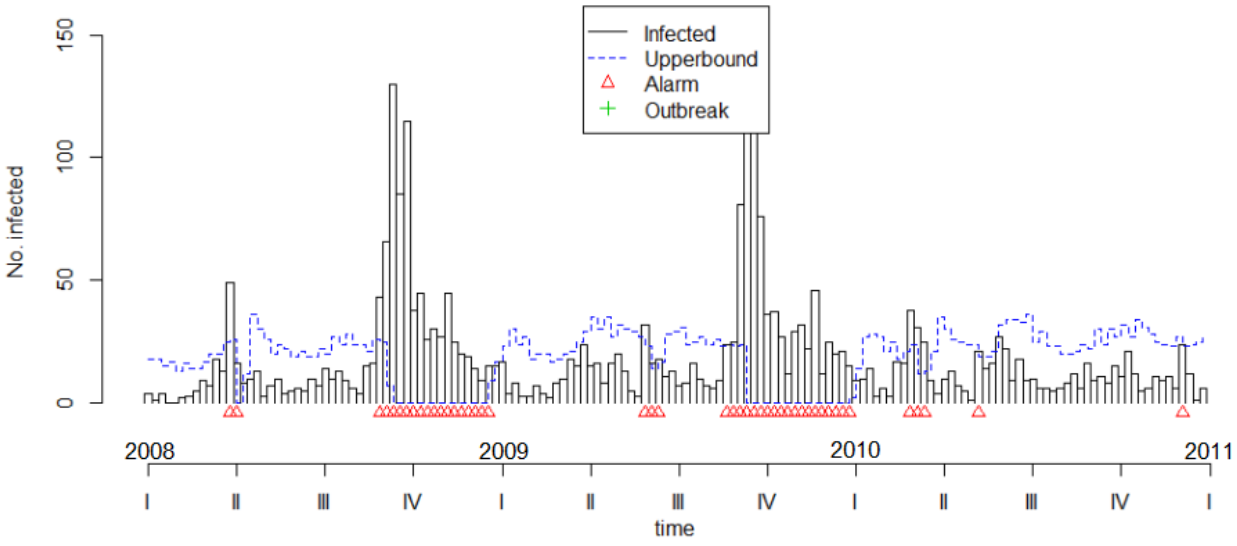


Figure 3.10. CUSUM analysis of Colombo model predicted values from 2008 to 2010. Red triangles indicate periods of time where alarm of an outbreak should be triggered.

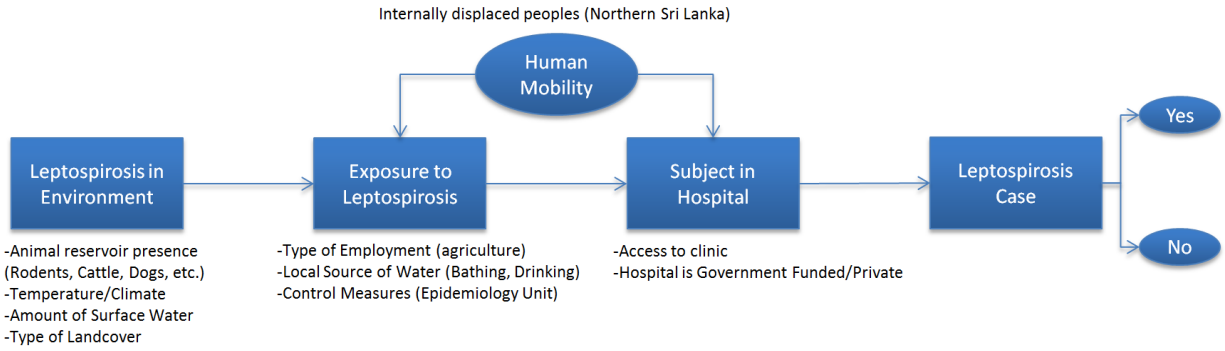


Figure 3.11. Theoretical leptospirosis risk model.

Table 3.1. Outline of all different models that were fit for each MOH area, and their respective covariates. *t* denotes the current week.

Model ID	Model Distribution	Covariates
A	Negative binomial	• Rainfall at *t* • Rainfall at *t-23*
B	Negative binomial	• Rainfall at *t* • Rainfall at *t-23* • Leptospirosis at *t-52*
C	Negative binomial	• Rainfall at *t* • Rainfall at *t-23* • Sum of cumulative rainfall
D	Negative binomial	• Rainfall at *t*, • Rainfall at *t-23* • Leptospirosis at *t-52* • Sum of cumulative rainfall
E	Poisson	• Rainfall at *t* • Rainfall at *t-23*
F	Poisson	• Rainfall at *t* • Rainfall at *t-23* • Leptospirosis at *t-52*
G	Poisson	• Rainfall at *t*, • Rainfall at *t-23* • Leptospirosis at *t-52* • Sum of cumulative rainfall
H	Poisson	• Rainfall at *t*, • Rainfall at *t-23* • Leptospirosis at *t-52* • Sum of cumulative rainfall Sum of cumulative rainfall

Table 3.2. A) Leptospirosis case counts by year for all of Sri Lanka, and B) leptospirosis case counts at the district level for each year of study.

A)

Year	Leptospirosis cases
2006	1582
2007	2198
2008	7419
2009	4980
2010	4554

B)

District	2006 cases	2007 cases	2008 cases	2009 cases	2010 cases
Ampara	17	9	27	16	36
Anuradhapura	47	41	270	102	127
Badulla	39	49	74	103	92
Batticaloa	6	0	12	16	14
Colombo	143	163	1073	1195	610
Galle	78	170	447	262	192
Gampaha	211	311	830	499	597
Hambantota	53	57	142	111	116
Jaffna	3	0	2	1	1
Kalmunai	1	1	4	7	3
Kalutara	139	221	692	596	440
Kandy	102	151	537	242	195
Kegalle	290	247	594	347	431
Killinochchi	0	0	2	0	3
Kurunegala	75	87	694	194	397
Mannar	1	2	0	0	15
Matale	32	178	855	342	141
Matara	175	289	501	251	393
Moneragala	31	56	104	18	50
Mulattivu	0	0	0	0	0
Nuwara Eliya	12	14	76	48	36
Polonnaruwa	22	22	112	81	101
Puttalam	21	31	69	99	82
Rathnapura	79	84	262	419	434
Trincomalee	3	12	34	23	45
Vavuniya	2	3	6	8	3

Table 3.3. SRMSE ranks for all fitted models for each MOH area of study. MOHID denotes the MOH area which the rank can be attributed to, while each column labeled by letter denotes the model being evaluated (see Table 3.1). Red text indicates lowest total SRMSE.

MOHID	A	B	C	D	E	F	G	H
101	4	3	2	1	4	3	2	1
102	4	2	3	1	4	2	3	1
103	4	2	3	1	4	2	3	1
104	3	4	2	1	3	4	2	1
105	2	3	1	4	2	3	1	4
106	4	3	1	2	4	3	1	2
107	4	1	3	2	4	1	3	2
108	2	3	1	4	2	3	1	4
109	4	3	2	1	4	3	2	1
110	3	4	2	1	3	4	2	1
111	3	4	1	2	3	4	1	2
112	4	2	1	3	4	2	1	3
113	4	3	2	1	4	3	2	1
301	4	3	1	2	4	3	1	2
302	4	2	3	1	4	2	3	1
303	4	3	2	1	4	3	2	1
304	2	4	3	1	2	4	3	1
305	4	2	3	1	4	2	3	1
306	4	3	1	2	4	3	1	2
307	4	2	3	1	4	2	3	1
308	4	3	1	2	4	3	1	2
309	4	2	3	1	4	2	3	1
310	4	1	3	2	4	1	3	2
501	1	2	3	4	1	2	3	4
502	3	1	4	2	3	1	4	2
503	3	1	4	2	3	1	4	2
504	4	2	3	1	4	2	3	1
505	4	3	1	2	4	3	1	2
506	2	1	4	3	2	1	4	3
507	2	4	1	3	2	4	1	3
508	3	4	2	1	3	4	2	1
509	4	3	1	2	4	3	1	2
510	3	2	1	4	3	2	1	4
511	2	1	4	3	2	1	4	3
Total	114	86	75	65	114	86	75	65

Table 3.4. RPS ranks for all fitted models for each MOH area of study. MOHID denotes the MOH area which the rank can be attributed to, while each column labeled by letter denotes the model being evaluated (see Table 3.1). Red text indicates lowest total RPS.

MOHID	A	B	C	D	E	F	G	H
101	4	3	2	1	4	3	2	1
102	4	2	3	1	4	2	3	1
103	4	2	3	1	4	2	3	1
104	3	4	2	1	3	4	2	1
105	2	3	1	4	2	3	1	4
106	4	3	1	2	4	3	1	2
107	4	1	3	2	4	1	3	2
108	2	3	1	4	2	3	1	4
109	4	3	2	1	4	3	2	1
110	3	4	2	1	3	4	2	1
111	3	4	1	2	3	4	1	2
112	4	2	1	3	4	2	1	3
113	4	3	2	1	4	3	2	1
301	4	3	1	2	4	3	1	2
302	4	2	3	1	4	2	3	1
303	4	3	2	1	4	3	2	1
304	2	4	3	1	2	4	3	1
305	4	2	3	1	4	2	3	1
306	4	3	1	2	4	3	1	2
307	4	2	3	1	4	2	3	1
308	4	3	1	2	4	3	1	2
309	4	2	3	1	4	2	3	1
310	4	1	3	2	4	1	3	2
501	1	2	3	4	1	2	3	4
502	3	1	4	2	3	1	4	2
503	3	1	4	2	3	1	4	2
504	4	2	3	1	4	2	3	1
505	4	3	1	2	4	3	1	2
506	2	1	4	3	2	1	4	3
507	2	4	1	3	2	4	1	3
508	3	4	2	1	3	4	2	1
509	4	3	1	2	4	3	1	2
510	3	2	1	4	3	2	1	4
511	2	1	4	3	2	1	4	3
Total	114	86	75	65	114	86	75	65

Table 3.5. Mean model assessment metric values by all MOH areas in districts of study. Red indicates the minimum value for that particular metric.

Model Assessment Metric	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
AIC	520.02	518.61	512.22	513.46	578.87	575.4	570.33	569.76
SRMSE	2.0275	2.0151	2.016	2.007	2.0275	2.0151	2.016	2.007
Mean log score	0.9654	0.9589	0.9466	0.9451	1.0824	1.0719	1.0622	1.0572
RPS	0.49689	0.4908	0.48527	0.48321	0.51191	0.50846	0.50555	0.50341

Table 3.6. Table of Model Evaluation metrics for each MOH area in each district of study. Red indicates the highest (or worst scoring) values, while blue indicates the lowest (or best scoring) values for the given metric.

MOH ID	SRMSE	AIC	RPS	Mean Logarithmic Score
101	2.075	425.388	0.298	0.776
102	1.514	824.723	0.948	1.544
103	1.444	1083.983	2.119	2.042
104	1.669	615.708	0.589	1.142
105	3.212	286.876	0.189	0.509
106	3.287	207.270	0.107	0.356
107	1.644	653.337	0.633	1.214
108	2.098	449.424	0.324	0.822
109	2.353	329.474	0.211	0.591
110	2.470	288.600	0.164	0.513
111	2.015	681.852	0.928	1.269
112	2.741	263.249	0.142	0.464
113	1.621	717.257	0.680	1.337
301	1.862	456.257	0.328	0.835
302	1.578	604.095	0.509	1.119
303	1.922	524.698	0.421	0.967
304	1.722	446.754	0.330	0.817
305	1.317	780.818	0.824	1.459
306	1.724	653.986	0.582	1.215
307	1.730	667.486	0.639	1.241
308	1.360	797.433	0.797	1.491
309	2.139	329.557	0.212	0.591
310	1.528	576.306	0.474	1.066
501	2.156	492.879	0.488	0.906
502	2.350	422.975	0.356	0.771
503	2.663	279.332	0.170	0.495
504	1.927	450.674	0.368	0.824
505	1.773	335.781	0.199	0.603
506	2.898	396.107	0.332	0.719
507	1.501	572.995	0.516	1.060
508	1.552	616.879	0.559	1.144
509	2.057	509.095	0.499	0.937
510	2.096	375.925	0.264	0.681
511	2.239	340.338	0.229	0.612

Chapter 4: Conclusions

1. Discussion and Conclusions

The goal of this research was to advance the understanding of the role that specific environmental drivers can play in the emergence of infectious diseases. This goal was considered and addressed in a spatial context, and as a case study, rainfall and its effect on leptospirosis incidence was assessed in Sri Lanka. To accomplish this goal, I set out two primary research objectives: 1) determine if spatial interpolation techniques could be employed to predict rainfall effectively across the country of Sri Lanka, and 2) determine if precipitation data could provide a reliable early-warning signal for leptospirosis outbreaks in Sri Lanka. I was able to compare a variety of spatial interpolation methods successfully, and determine the most effective methods for predicting rainfall in a tropical setting based on specific underlying environmental conditions. I then used insights from this research to interpolate weekly rainfall data across Sri Lanka for use as a primary predictor to model leptospirosis incidence at numerous spatial scales, and evaluate whether models could be employed to forecast leptospirosis incidence. Forecasted results were found to be of high enough quality to allow the models to be used as a starting point by the Sri Lanka Ministry of Health (MOH) to provide timely early warning for leptospirosis outbreak events in Sri Lanka.

Chapter 2 demonstrated that by using a network of small-scale community-managed stations, accurate country-wide rainfall predictions could be made across the country of Sri Lanka. By comparing inverse distance weighting, thin-plate smoothing splines, ordinary kriging, and Bayesian kriging, it was determined that in a tropical setting with extremely variable climatic conditions over space and time such as Sri Lanka, interpolation method selection should be based on evaluating specific conditions present in the environment at that point in time. For

example, it was found that in high rainfall conditions, thin-plate smoothing splines were able to predict the magnitude of rainfall most accurately, as well as approximate the global spatial pattern associated with large monsoonal and convectional rainfall events (i.e., continuous over space). It was also found that Bayesian kriging was a more effective method when considering relatively low rainfall conditions, as it was able to approximate the more discrete nature of minor rainfall events, and predict rainfall with higher accuracy at lower magnitudes. Several complimentary error metrics were employed to evaluate the interpolated results against ground truth data. The Structural Similarity Index was also used to assess overall spatial patterns and similarity between the interpolations and remote-sensed imagery (Robertson et al., 2014). I believe that this research will help others looking to predict rainfall in a tropical setting by providing novel ways to evaluate and select appropriate interpolation methods. Additionally, this research added to previous modelling literature on situation-specific (e.g., in Sri Lanka) selection of interpolation techniques, as measuring performance for any environmental model is intrinsically case-dependent (Bennett et al., 2013).

In Chapter 3, time-series regression models at two different spatial scales were used to predict leptospirosis risk in Sri Lanka, and determine if rainfall was a meaningful predictor for future leptospirosis incidence. Based on research performed in Chapter 2, Bayesian kriging was selected as the most appropriate interpolation method to generate rainfall surfaces at a weekly time scale. Rainfall values for each MOH area and district across Sri Lanka were extracted from the Bayesian kriging interpolations. Using various forms of correlation analysis, meaningful lags between rainfall and leptospirosis incidence were identified. These lags were then used to help select several different permutations of covariates and predictive distributions to include in the leptospirosis risk models. After careful evaluation of all models produced across varying spatial

scales using a multitude of model calibration metrics, a negative binomial integer-valued autoregressive conditional heteroscedasticity (INGARCH) model that included current and previous rainfall covariates, as well as regression on previous cases of leptospirosis at a local and seasonal time scale was selected for modelling leptospirosis incidence. Regions of interest based on a known leptospirosis outbreak events in 2008 and 2009 were used to model leptospirosis risk in Sri Lanka. It was found that while there was no significant correlation between previous or current rainfall events and leptospirosis incidence in Sri Lanka, there was a strong serial-dependence on previous leptospirosis cases that allowed for accurate prediction of future leptospirosis risk and outbreak events. A CUSUM analysis of forecasted leptospirosis cases for the Colombo district of Sri Lanka indicated that the model was able to provide early warning for major leptospirosis outbreaks in Sri Lanka.

2. Research limitations

Through this research, I was able to complete the objectives set out for this thesis successfully, but there were several limitations that affected the quality of results and implementations. Limitations in both the amount and quality of the data used in these studies must be taken into account when assessing any of the results that were produced. In developing countries such as Sri Lanka, data are often sparse or completely missing in certain regions of the study area, and when performing analysis in a spatial context, irregularly distributed data can lead to biased results for a given area. For example, in the interpolation assessment in Chapter 2, it was found that the most populous district of Sri Lanka, Colombo, experienced very large prediction errors due to the spatial network of community-managed stations not providing adequate coverage for particular regions. Other than incorporating a more complete data set in the analysis (which was not available at the time this research was conducted), the research was

limited by the quality and distribution of the data sets used, and thus certain results attained could be more indicative of data quality than the relationships that were observed and assessed. Given that this research occurred in two separate stages – one which was dependent on results from the on the other – it is important to mention that any errors encountered due to data quality may have potentially compounded when those data were employed in further research.

This research was also limited by the associated computational and financial costs that would be required if automation of this workflow was implemented (e.g., by the Sri Lanka MOH and Department of Meteorology to develop an early warning system for leptospirosis outbreak). While powerful processors have become more affordable over the last decade, mathematically complex algorithms, such as those used to generate Bayesian kriging interpolations, still require a relatively large amount of computational power to be completed in a timely fashion. I considered these costs as criteria for selecting methods in all of the research conducted, but minimum computational requirements for implementing a workflow where results are needed in for real-time disease surveillance may still be an issue. A concerted effort was made to select methods that would be able to generate accurate and usable results using the simplest methods available without incurring a cost on the quality of predicted outputs.

3. Research Contributions

The research conducted in this thesis made contributions to elucidating the relationships between environmental factors and EID incidence. Specifically, by employing novel modelling methods to assess the EID leptospirosis and evaluating the effect that precipitation has on its incidence in Sri Lanka, I was able to determine that serial dependence on previous leptospirosis cases plays a more important role in the dynamics of leptospirosis transmission. Little research

has been done assessing drivers of leptospirosis risk in Sri Lanka, and with a known outbreak in 2008 where greater than 7000 people were affected, it is important determine which environmental variables should and should not be considered as drivers of outbreak when attempting to prevent future outbreaks.

Another contribution of this research was in the field of spatial interpolation, and specifically, assessment of rainfall in tropical settings which observe highly variable climates. I was able to determine the strengths and weaknesses of various interpolation methods when dealing with variable weather patterns in a relatively small study area, and from that, make suggestions as to the best method to employ provided certain underlying conditions. I was also able to characterize global monsoonal and convectional rainfall patterns, and evaluate their spatial structures using SSIM – a novel image comparison algorithm – in conjunction with more standard error metrics. This approach of using spatial structure evaluation in conjunction with standard empirical error metrics provided a means to better understand how well interpolation methods were approximating actual rainfall magnitudes and distribution, and could be used in a spatial modelling context when evaluating congruency in pattern between ground truth data and predicted results.

Through completing this research, a master data set was constructed by aggregating several different data sets together (e.g., leptospirosis case count data, rainfall data, Sri Lanka district level and MOH area level spatial data). This new data set is a vast improvement over any previous data sets available for assessing leptospirosis in Sri Lanka, as it has been standardized both spatially and temporally. MOH area boundaries were subject to change for each year of the study period and required a considerable amount of manipulation to standardize to a set of common boundaries for the entire study period. This new master data set will be of use to the

Ministry of Health in Sri Lanka, as never before has a data set containing leptospirosis data and rainfall data been available for analysis. Moreover, as a script was written to assemble this data set, other social, ecological, economic, or environmental variables can potentially be appended to the data set with relatively little programming.

Lastly, by developing an effective INGARCH time series modelling framework for Sri Lanka, I was able to provide a means of early warning for leptospirosis outbreak in Sri Lanka. I plan to continue refining the leptospirosis risk models and have them assessed for implementation at the nation-wide scale, and to work with the Ministry of Health in Sri Lanka to improve leptospirosis surveillance efforts, and attempt to develop early warning protocols for leptospirosis risk.

References:

Abtew, W., Obeysekera, J. Shih, G., 1993. Spatial Analysis for Monthly Rainfall in South Florida. JAWRA Journal of the American Water Resources Association, 29(2), 179–188. doi: 10.1111/j.1752-1688.1993.tb03199.x

Alvarez, O., Guo, Q., Klinger, R.C., Li, W., Doherty, P., 2014. Comparison of elevation and remote sensing derived products as auxiliary data for climate surface interpolation. International Journal of Climatology, 34(7), 2258–2268.

Ashford, D.A., Kaiser, R.M., Spiegel, R.A., Perkins, B.A., Weyant, R.S., Bragg, S.L., Plikaytis, B., Jarquin, C., Reyes, J.D.L., Amador, J.J., 2000. Asymptomatic infection and risk factors for leptospirosis in Nicaragua. Am. J. Trop. Med. Hyg. 63, 249–254.

Bellack, N.R., Koehoorn, M.W., Macnab, Y.C., Morshed, M.G., 2006. A conceptual model of water's role as a reservoir in Helicobacter pylori transmission: a review of the evidence. Epidemiol. Infect. 134, 439–449. doi:10.1017/S0950268806006005

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., others, 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Berger, J.O., De Oliveira, V., Sansó, B., 2001. Objective Bayesian Analysis of Spatially Correlated Data. Journal of the American Statistical Association, 96(456), 1361–1374.

Bharti, A.R., Nally, J.E., Ricaldi, J.N., Matthias, M.A., Diaz, M.M., Lovett, M.A., Levett, P.N., Gilman, R.H., Willig, M.R., Gotuzzo, E., others, 2003. Leptospirosis: a zoonotic disease of global importance. Lancet Infect. Dis. 3, 757–771.

Briët, O.J., Vounatsou, P., Gunawardena, D.M., Galappaththy, G.N., Amerasinghe, P.H., 2008. Temporal correlation between malaria and rainfall in Sri Lanka. Malar. J. 7, 77.

Bröcker, J., Smith, L.A., 2008. From ensemble forecasts to predictive distribution functions. Tellus A 60. doi:10.3402/tellusa.v60i4.15387

Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S., Triampo, W., 2012. Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. Asian Pac. J. Trop. Med. 5, 539–546. doi:10.1016/S1995-7645(12)60095-9

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2015. Shiny: web application framework for R. R Package Version 011 1.

Chemel, C., Sokhi, R.S., Dore, A.J., Sutton, P., Vincent, K.J., Griffiths, S.J., Hayman, G.D., Wright, R.D., Baggaley, M., Hallsworth, S., others, 2011. Predictions of UK regulated power station contributions to regional air pollution and deposition: a model comparison exercise. J. Air Waste Manag. Assoc. 61, 1236–1245.

Chien, L.-C., Yu, H.-L., 2014. Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. Environ. Int. 73, 46–56.

Christou, V., Fokianos, K., 2015. On count time series prediction. J. Stat. Comput. Simul. 85, 357–373.

Cressie, N., 1993. Statistics for Spatial Data (Wiley Series in Probability and Statistics) (Revised Edition.). Wiley-Interscience.

Czado, C., Gneiting, T., Held, L., 2009. Predictive model assessment for count data. Biometrics 65, 1254–1261.

Daly, C., Neilson, R. P., Phillips, D. L., 1994. A statistical-topographical model for mapping climatological precipitation over mountainous terrain. J Clim Appl Meterol, 33, 140-158.

Davis, S., Calvet, E., 2005. Fluctuating rodent populations and risk to humans from rodent-borne zoonoses. Vector-Borne Zoonotic Dis. 5, 305–314.

Dawid, A.P., 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. J. R. Stat. Soc. Ser. Gen. 278–292.

Diggle, P.J. Ribeiro, P.J., 2002. Bayesian Inference in Gaussian Model-based Geostatistics. Geographical and Environmental Modelling, 6(2), 129–146.

Dirks, K.N., Hay, J.E., Stow, C.D., Harris, D., 1998. High-resolution studies of rainfall on Norfolk Island: Part II: Interpolation of rainfall data. J. Hydrol. 208, 187–193.

Ehret, U., Zehe, E., 2011. Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrol. Earth Syst. Sci., 15, 877-896. doi:10.5194/hess-15-877-2011.

Enscore, R.E., Biggerstaff, B.J., Brown, T.L., Fulgham, R.E., Reynolds, P.J., Engelthaler, D.M., Levy, C.E., Parmenter, R.R., Montenieri, J.A., Cheek, J.E., others, 2002. Modeling relationships between climate and the frequency of human plague cases in the southwestern United States, 1960-1997. Am. J. Trop. Med. Hyg. 66, 186–196.

Ferland, R., Latour, A., Oraichi, D., 2006. Integer-Valued GARCH Process. J. Time Ser. Anal. 27, 923–942.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Hagen-Zanker, A., 2006. Comparing continuous valued raster data: A cross disciplinary literature scan. Netherlands Environmental Assessment Agency. Research Institute for Knowledge Systems bv. 59p.

Haggett, P., 1994. Geographical aspects of the emergence of infectious diseases. Geogr. Ann. Ser. B Hum. Geogr. 91–104.

Heinen, A., 2003. Modelling time series count data: an autoregressive conditional Poisson model. Available SSRN 1117187.

Held, L., Hofmann, M., Höhle, M., Schmid, V., 2006. A two-component model for counts of infectious diseases. Biostatistics 7, 422–437.

Heyman, P., Vervoort, T., Escutenaire, S., Degrave, E., Konings, J., Vandenvelde, C., Verhagen, R., 2001. Incidence of hantavirus infections in Belgium. Virus Res. 77, 71–80.

Hii, Y.L., Zhu, H., Ng, N., Ng, L.C., Rocklöv, J., 2012. Forecast of dengue incidence using temperature and rainfall.

Hutchinson, M. F., 1995. Interpolating mean rainfall using thin plate smoothing splines. International Journal of Geographical Information Systems, 9(4), 385–403.

Jayawardene, H., Sonnadara, D.U.J., Jayewardene, D.R., 2005. Trends of rainfall in Sri Lanka over the last century. Sri Lankan J. Phys. 6, 7–17.

Jayawardene, H., Sonnadara, D. U. J., Jayewardene, D. R., 2005. Spatial interpolation of weekly rainfall depth in the dry zone of Sri Lanka. Climate Research, 29(3), 223.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., Beswick, A. R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling Software, 16(4), 309–330.

Jones, H.E., Spiegelhalter, D.J., 2012. Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models. J. R. Stat. Soc. Ser. A Stat. Soc. 175, 729–747.

Kleinman, K., Lazarus, R., Platt, R., 2004. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am. J. Epidemiol. 159, 217–224.

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., Simpson, J., 1998. The tropical rainfall measuring mission (TRMM) sensor package. J. Atmospheric Ocean. Technol. 15, 809–817.

Lau, C.L., Smythe, L.D., Craig, S.B., Weinstein, P., 2010. Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? Trans. R. Soc. Trop. Med. Hyg. 104, 631–638.

Levett, P.N., 2001. Leptospirosis. Clin. Microbiol. Rev. 14, 296–326.

Liboschik, T., Fokianos, K., Fried, R., 2015. tscount: An R package for analysis of count time series following generalized linear models.

Madsen, T., Shine, R., 1999. Rainfall and rats: Climatically-driven dynamics of a tropical rodent population. Aust. J. Ecol. 24, 80–89.

Malhi, Y., Wright, J., 2004. Spatial patterns and recent trends in the climate of tropical rainforest regions. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1443), 311–329.

Mayer, J.D., 2000. Geography, ecology and emerging infectious diseases. Soc. Sci. Med. 50, 937–952.

Meentemeyer, R.K., Anacker, B.L., Mark, W., Rizzo, D.M., 2008. Early detection of emerging forest disease using dispersal estimation and ecological niche modeling. Ecol. Appl. 18, 377–390.

Mendelsohn, R., Dinar, A., 1999. Climate change, agriculture, and developing countries: does adaptation matter? World Bank Res. Obs. 14, 277–293.

Mills, J.N., Childs, J.E., 1998. Ecologic studies of rodent reservoirs: their relevance for human health. Emerg. Infect. Dis. 4, 529.

Morse, S.S., 1995. Factors in the emergence of infectious diseases. Emerg. Infect. Dis. 1, 7.

Nychka, D., Furrer, R., Sain, S., 2012. fields: Tools for spatial data.

Olsson, G.E., Dalerum, F., Hörnfeldt, B., Elgh, F., Palo, T.R., Juto, P., Ahlm, C., 2003. Human hantavirus infections, Sweden. Emerg. Infect. Dis. 9, 1395.

O'Sullivan, D., Unwin, D. J., 2003. Geographic information analysis. John Wiley & Sons Inc.

Pachauri, R.K., Reisinger, A., others, 2007. Contribution of working groups I, II and III to the fourth assessment report of the intergovernmental panel on climate change. IPCC Geneva Switz. 104.

Pappachan, M.J., Sheela, M., Aravindan, K.P., 2004. Relation of rainfall pattern and epidemic leptospirosis in the Indian state of Kerala. J. Epidemiol. Community Health 58, 1054–1054. doi:10.1136/jech.2003.018556

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

Plouffe, C.C.F., Robertson, C., Chandrapala, L., 2015. Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka. Environ. Model. Softw. 67, 57–71. doi:10.1016/j.envsoft.2015.01.011

Price, D.T., McKenney, D.W., Nalder, I.A., Hutchinson, M.F., Kesteven, J.L., 2000. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agric. For. Meteorol. 101, 81–94.

Punyawardena, B.V.R., Kulasiri, D., 1999. Spatial interpolation of rainfall in the dry zone of Sri Lanka. J Nat Sci Counc Sri Lanka, 26(3), 247–262.

Robertson, C., Nelson, T. A., Stephen, C., 2012. Spatial epidemiology of suspected clinical leptospirosis in Sri Lanka. Epidemiol Infect, 140(4), 731–43. doi: 10.1017/S0950268811001014

Ribeiro Jr, P.J., Diggle, P.J., 2001. geoR: A package for geostatistical analysis. R News 1, 14–18.

Robertson, C., 2015. Towards a geocomputational landscape epidemiology: surveillance, modelling, and interventions. GeoJournal 1–18. doi:10.1007/s10708-015-9688-5

Robertson, C., Long, J.A., Nathoo, F.S., Nelson, T.A., Plouffe, C.C., 2014. Assessing quality of spatial models using the structural similarity index and posterior predictive checks. Geogr. Anal. 46, 53–74.

Robertson, C., Nelson, T.A., MacNab, Y.C., Lawson, A.B., 2010. Review of methods for space–time disease surveillance. Spat. Spatio-Temporal Epidemiol. 1, 105–116.

Robertson, C., Nelson, T.A., Stephen, C., 2012. Spatial epidemiology of suspected clinical leptospirosis in Sri Lanka. Epidemiol. Infect. 140, 731–743.

Robertson, C., Sawford, K., Gunawardana, W.S., Nelson, T.A., Nathoo, F., Stephen, C., 2011. A hidden markov model for analysis of frontline veterinary data for emerging zoonotic disease surveillance.

Robson, B.J., 2014. State of the art in modelling of phosphorus in aquatic systems: review, criticisms and commentary. Environ. Model. Softw. 61, 339–359.

Rose, A.M., Vapalahti, O., Lyytikäinen, O., Nuorti, P., 2003. Patterns of Puumala virus infection in Finland. Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 8, 9–13.

Sanner, M.F., others, 1999. Python: a programming language for software integration and development. J Mol Graph Model 17, 57–61.

Sarkar, J., Chopra, A., Katageri, B., Raj, H., Goel, A., 2012. Leptospirosis: a re-emerging infection. Asian Pac. J. Trop. Med. 5, 500–502.

Sarkar, U., Nascimento, S.F., Barbosa, R., Martins, R., Nuevo, H., Kalofonos, I., Kalafanos, I., Grunstein, I., Flannery, B., Dias, J., others, 2002. Population-based case-control investigation of risk factors for leptospirosis during an urban epidemic. Am. J. Trop. Med. Hyg. 66, 605–610.

Sri Lanka Epidemiology Unit, 2008. An interim analysis of leptospirosis outbreak in Sri Lanka - 2008. Colombo Epidemiol. Unit Minist. Health Care Nutr. 1–8.

Tassinari, W.S., Pellegrini, D.C., Sá, C.B., Reis, R.B., Ko, A.I., Carvalho, M.S., 2008. Detection and modelling of case clusters for urban leptospirosis. Trop. Med. Int. Health 13, 503–512.

Vicente Serrano, S.M., Sánchez, S., Cuadrat, J.M., others, 2003. Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. Clim. Res. 24, 161–180.

Vijayachari, P., Sugunan, A.P., Shriram, A.N., 2008. Leptospirosis: an emerging global public health problem. J. Biosci. 33, 557–569.

Vinetz, J.M., Wilcox, B.A., Aguirre, A., Gollin, L.X., Katz, A.R., Fujioka, R.S., Maly, K., Horwitz, P., Chang, H., 2005. Beyond disciplinary boundaries: leptospirosis as a model of incorporating transdisciplinary approaches to understand infectious disease emergence. EcoHealth 2, 291–306.

Waller, L.A., 2004. Invited commentary: surveilling surveillance—some statistical comments. Am. J. Epidemiol. 159, 225–227.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13, 600–11.

Wang, J., Wolff, D. B., 2010. Evaluation of TRMM ground-validation radar-rain errors using rain gauge measurements. Journal of Applied Meteorology and Climatology, 49(2), 310–324.

WHO, 1999. Leptospirosis worldwide, 1999. Wkly. Epidemiol. Rec. Health Sect. Secr. Leag. Nations 74, 237–242.

Wickramaarachchi, T.N., Ishidaira, H., Wijayaratna, T.M.N., 2013. Applicability of global public domain data versus local detailed data for distributed hydrological modelling: a study form Gin river basin Sri Lanka. Appl. Water Sci. 3, 545–557.

Zhang, Y., Bi, P., Hiller, J.E., 2008. Climate change and the transmission of vector-borne diseases: a review. Asia. Pac. J. Public Health 20, 64–76.

Zubair, L., 2002. El Nino–southern oscillation influences on rice production in Sri Lanka. Int. J. Climatol. 22, 249–260.