Fall 2016

Computationally Efficient Specifications of Spatial Point Process Models and Spatio-Temporal Gaussian Models: Combining Remote Sensing Drivers with Geospatial Disease Case Data to Enhance Geographic Epidemiology

Beth Louise Ziniti
University of New Hampshire, Durham

COMPUTATIONALLY EFFICIENT SPECIFICATIONS OF SPATIAL POINT PROCESS MODELS AND SPATIO-TEMPORAL GAUSSIAN MODELS: COMBINING REMOTE SENSING DRIVERS WITH GEOSPATIAL DISEASE CASE DATA TO ENHANCE GEOGRAPHIC EPIDEMIOLOGY

BY

BETH L. ZINITI

B.A., McGill University, Quebec Canada, 2005

M.S., University of New Hampshire, 2012

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Statistics

September, 2016

This dissertation has been examined and approved in partial fulfillment of the requirements for the degree of Ph.D. in Statistics by:

Dissertation Director, Dr. Ernst Linder,
Professor of Mathematics & Statistics

Dr. Linyuan Li,
Professor of Mathematics & Statistics

Dr. Haiying Wang,
Assistant Professor of Mathematics & Statistics

Dr. Mark Lyon,
Associate Professor of Mathematics & Statistics

Dr. Nathan Torbick,
Director, Human & Environment Interactions
Applied Geosolutions, LLC

On August 9 2016

Original approval signatures are on file with the University of New Hampshire Graduate School.

DEDICATION

To my parents, Barbara and Bill

ACKNOWLEDGEMENTS

This dissertation could not have been written without the support and guidance from many individuals, and I would like to express my deepest appreciation to some of these people in particular.

To begin, I would like to give special thanks to my advisor, Dr. Ernst Linder, for introducing me to Spatial Statistics and the EAR model, and for all his time and guidance that he has provided me in the research and writing of this dissertation. I was glad to revisit the topic of Markov Chains, which I first saw during my undergraduate studies and most enjoyed.

I would also like to thank the members of my committee Drs. Linyuan Li, Haiying Wang, Mark Lyon and Nathan Torbick for taking the time to discuss relevant research topics with me and to provide me useful suggestions and valuable comments.

I would like to give particular thanks to Dr. Nathan Torbick for introducing me to such an important and complex scientific application, as well as for providing all the water quality data and much guidance and suggestions in my modeling. Also I am thankful to all the employees of Applied Geosolutions for providing such a friendly and welcoming atmosphere, and especially to Megan Corbiere for all her work in processing the data.

I am thankful to all the people in the Dartmouth group in particular, Dr. Elijah Stommel, Dr. Angeline Andrew, Dr. Xun Shi, Dr. Tracie Caller and Ms. Patricia Henegan for providing the case data, for their interest in this work, and for all their helpful suggestions and comments.

The work in this dissertation was funded from the National Science Foundation grant GSS (BCS-1433756) and is part of a larger collaborative project that includes researchers from the University of New Hampshire, Applied Geosolutions LLC, Dartmouth College and the Dartmouth Hitchcock Medical Center. I am honored to have been a part of this project.

I am very grateful for the generous financial support I received during my doctorial studies and that allowed me to complete this dissertation. I would like to thank the UNH Mathematics and Statistics Department for providing me funding in the form of teaching assistantships as well as for funding to attend conferences. I would like to thank the UNH Graduate School for also providing funding to attend conferences. I am thankful for some additional funding I received as a consultant in the UNH Mathematics and Statistics department's statistical consulting service, in particular for a project requested by Senior Vice Provost for Research Jan Nesbit. I am thankful for the LandScan background population data paid for by the UNH Library.

I would like to thank Professor Edward Hinson and Professor Rita Hibschweiler for their guidance during my time at UNH, and to thank Jan Jankowski, Ellen O'Keefe, and April Flore in the math office for their various forms of support. I am thankful for the writing and editing assistance I received at the UNH writing center, in particular from Lauren Short and Molly Tetreault. I am grateful to my fellow graduate students; Meng Zhao, Michael Willliams, Shuang Wu, and many more for their many helpful and interesting conversations and for joining me on many hiking, running and x-country skiing adventures.

Finally, I am most grateful for my supportive and understanding family, especially thank you to my husband, Charlie Ziniti who has helped me through some difficult days, and to my father, Bill Roberts, for all his support, for his patience and for his persistence. I have finally found that something technical that I truly enjoy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

COMPUTATIONALLY EFFICIENT SPECIFICATIONS OF SPATIAL POINT PROCESS MODELS AND SPATIO-TEMPORAL GAUSSIAN MODELS: COMBINING REMOTE SENSING DRIVERS WITH GEOSPATIAL DISEASE CASE DATA TO ENHANCE GEOGRAPHIC EPIDEMIOLOGY

by

Beth L. Ziniti

University of New Hampshire, September, 2016

In this dissertation, the flexibility of Bayesian hierarchical models specified using a latent Gaussian Markov Random Field (GMRF) are evaluated for use in analyzing large complex spatial and spatio-temporal data with the goal of contributing to an interdisciplinary effort of developing an eco-epidemiological model that quantifies the relationship between remotely sensed water quality and the incidence of ALS (Amyotrophic Lateral Sclerosis or Lou Gehrig’s Disease) over large areas such as Northern New England (NNE).

In particular, a Log-Gaussian Cox Process (LGCP) specified by the logarithm of a GMRF on a regular lattice is shown to allow for simultaneous estimation of the spatial distribution of ALS risk and its relationship to remotely sensed water quality metrics. This approach improves

on previous analyses of the dataset considered by explicitly accounting for the spatial uncertainty in determining locations of ALS “hotspots” needed in the estimation of the hotspots’ relationship to the water quality of lakes in NNE.

Finally, since warming lake temperatures have been associated with more frequent cyanobacteria blooms (blue-green algae), which is a possible risk factor of ALS, a spatially varying coefficient model specified with an Extended Autoregression (EAR) latent process is used in an analysis of remotely sensed surface water temperatures of Lake Champlain. New interpretations of the EAR model are suggested and issues relating to its parameter’s identifiability are investigated.

I INTRODUCTION

The field of eco-epidemiology seeks to identify causes of spatial and temporal patterns in disease incidence and mortality at multiple scales and in relationship to the Earth's changing ecologies (Susser, 2004). It is an emerging discipline but has roots at least as far back as 1854 when Dr. John Snow used maps to link London England's cholera outbreak to the contamination of the water supply in certain neighborhoods. Its questions lie at the intersection of diverse fields including epidemiology, geography, the environmental sciences, computer science, mathematics and statistics and require a collaborative effort in the collection and analysis of complex spatio-temporal datasets.

Spatio-temporal data contains measurements that are geographically and temporally indexed. These data generally follow the principle stated by the geographer Waldo Tobler, "Everything is related to everything else, but near things are more related than distant things." In geographic space, the "near things" are measurements taken at a small Euclidean distance apart while in the time domain, the "near things" are measurements taken at a small time interval apart. In statistics, this concept is known as positive autocorrelation and is commonly modeled using a Gaussian Process. The Gaussian Process is characterized by a mean function and a covariance function. The mean function is generally interpreted to describe large-scale trends, while the covariance function describes the small-scale dependencies or autocorrelations. In the analysis of count or binary data common in epidemiology and ecology, the Gaussian Process is often used in a hierarchical model as a latent process that drives the discrete outcome.

The advancement of remote sensing technologies, such as satellite imaging that include information from global positioning systems and timestamps, has greatly facilitated the collection of large environmental spatio-temporal datasets. In the area of public health, spatio-temporal datasets have also become more available in recent years. The spatial and temporal scales of remotely sense datasets often vary by technology, while the scales of public health data vary for reasons of ethics and confidentiality. As the world populations become more connected and interdependent, and the need for research in eco-epidemiology grows, it becomes increasingly important to have methods to analyze spatio-temporal data that are flexible in their ability to combine data sources collected at various temporal and spatial scales, that are robust in the presence of abnormalities unrelated to the questions under study and computationally efficient for big data. Furthermore, for the purpose of decision/policy making, it is not enough that an analysis produces an answer, but it must be able to account for the uncertainty in its answer that comes from various sources such as measurement error, model specification and parameter estimation.

In statistics, a probability framework is used to make inference on the processes from which data are sampled and uncertainty is quantified by variance. This framework is particularly useful in the study of public health since although medical science can explain many of the biological mechanisms by which disease occurs, not all persons in contact with the suspected causes will become diseased. Thus, the statistical approach quantifies a person’s disease risk as their probability of contracting the disease and tests exposures to identify and quantify the ones that significantly modify a person’s risk (Waller & Gotway, 2004).

One disease that is not very well understood by medical science is the progressive neurodegenerative disease, Amyotrophic Lateral Sclerosis (ALS) also known as Lou Gehrig's disease. Except in very few cases, ALS is fatal and has an expected survival time of approximately 3 years after onset (Noonan, White, Thurman, & Wong, 2005). ALS is also a rare disease with annual incidence rates varying spatially within the United States ranging from 1 to 1.8 in 100,000 people, although studies have indicated that these rates have been increasing over time (Noonan et al., 2005). Incidence rates have been shown to be age and sex related with the highest incidence occurring for ages between 55-75 years and in males, but the change in incidence over time shows a diminished effect of sex (Noonan et al., 2005). The change in incidence however could be attributed to several factors including the aging population, improved diagnosis, and the creation of registries for more complete case ascertainment (Caller, Chipman, Field, & Stommel, 2013). Only between 5-10% of cases can be attributed to a genetic cause, while the rest known as sporadic ALS (sALS) are assumed to be the result of genetic susceptibility combined with environmental exposures (Noonan et al., 2005; Caller et al., 2013; Torbick, Hession, Stommel, & Caller, 2014). One exposure of particular interest is the presence of the neurotoxin, beta-methylamino-L-alanine (BMAA) produced by cyanobacteria (blue-green algae), which has been linked to the high incidence of ALS and Parkinson's disease occurring in Guam during the early 1950's (Caller et al., 2009; Caller et al., 2013; Torbick et al., 2014; Banack et al., 2015).

ALS in Northern New England

In Northern New England (NNE), which comprised the US states of Maine, New Hampshire and Vermont, several clusters of ALS, regions with higher ALS cases than would be expected, have been identified based on case event data for 1997 to 2009 that were aggregated to the census block group level (Caller et al., 2013). Since ALS is a non-contagious disease, the presence of clusters in this region is likely due to environmental heterogeneity. In particular, at least one cluster occurs adjacent to a lake with frequent harmful cyanobacteria blooms and in which BMAA has been found in fish and filtered aerosol samples (Caller et al., 2009; Banack et al., 2015). Torbick et al. (2014) reanalyzed the case event data for this region aggregating at the census track level and also found similar clusters to those identified by Caller et al. (2013). The fact that each study arrives at a similar conclusion having used different aggregation scales for cluster determination gives strong evidence in favor of the existence of an environmental risk factor for ALS. Furthermore, Torbick et al. (2014) shows that cases close to lakes with poorer water quality had an increased chance of belonging to an ALS cluster.

However, since there is no one coordinated medical system for the entire NNE region and no mandatory national ALS registry at the time of data collection, incomplete case ascertainment could be an issue for the ALS case dataset used in Caller et al. (2013) and Torbick et al. (2014). It is suspected that case counts in this dataset underestimate the risk of ALS in NNE during 1997-2009, in particular for all of Maine, parts of southern New Hampshire that are close to Boston, Massachusetts and southwestern Vermont that are close to Albany, New York. Since the case ascertainment appears to vary spatially, the identified clustering may be an artifact of this ascertainment. In an Irish study where a nearly complete ascertainment was possible due to a national ALS registry, no high-risk ALS clusters were found (Rooney et al., 2014; Rooney et al., 2015). Furthermore, the relationship between ALS risk and the water quality of lakes in the region is only based on the census track clusters, the larger of the two scales, and does not explicitly account for the spatial uncertainty of the clusters in the logistic regression used to estimate the effect of poor water quality exposure.

An alternate approach would restrict the study region to areas where nearly complete case ascertainment was possible and would use Bayesian Inference of a Log-Gaussian Cox Process (LGCP) (Møller, Syversveen, & Waagepetersen, 1998) to simultaneously estimate the spatial distribution of ALS risk and its relationship to the water quality of the region's lakes. Furthermore, an aggregation scale chosen to approximate a continuous LGCP, along with the Markov Property assumption for spatial autocorrelation would allow for a computationally efficient estimation and yet sufficient approximation (Waagepetersen 2004; Lindgren, Rue, & Lindström, 2011; Diggle, Moraga, Rowlingson, & Taylor, 2013; Simpson, Illian, Lindgren, Sørbye, & Rue, 2016).

Study Goals

In this dissertation, the ALS case dataset discussed in Caller et al. (2013) and Torbick et al. (2014) is reanalyzed using Bayesian inference of a LGCP specified by the logarithm of a Gaussian Markov Random field (GMRF) on a regular lattice (grid). The modeled component of the LGCP intensity, which represents relative risk due to environmental heterogeneity, will include spatial random effects used to model small-scale autocorrelations and fixed effects of the region's lake water quality used to model spatial trends. Two forms of GMRF spatial random effects will be compared, the Besag-York-Mollié model (Besag, York, & Mollié, 1991) which is a convolution of the intrinsic CAR and independent random effects, and the Leroux model (Lee, 2011) which will be shown to have random effects equivalent to those of the Czado CAR (Czado & Prokopenko, 2008) and equivalent to the EAR model (Yuan, 2011) with smoothing parameter,$\theta = 1$. Although these two models are widely used in the statistical literature on disease mapping, there is little investigation of what effect varying spatial scales will have on the estimated parameters. In this dissertation, two scales will be compared. Furthermore, it has been suggested that the use of these random effects models to account for spatial autocorrelations, may bias fixed effect estimations due to collinearities of the spatial random effects with the fixed effects (Reich, Hodges, and Zadnik, 2006; Hughes and Haran, 2013; Hughes, 2015). However, this issue has only been demonstrated with the BYM model and not the Leroux model. Since the Leroux model is shown to be connected to the EAR model developed by Yuan (Yuan, 2011), this dissertation will investigate the effect of this issue with the EAR model.

Fixed effects will be derived from the lake water quality data used by Torbick et al. (2014) which are measurements of satellite derived lake average secchi depth (SD in meters), total nitrogen (TN in micrograms per liter (ug/L)), and chlorophyll-a (Chl-a in ug/L) of lakes sized 6 hectares or more except for Lake Champlain, as well as satellite derived measures of phycocyanin (PC in ug/L) aggregated on various lake scales for all lakes sized 6 hectares or more including Lake Champlain. Lake aggregation scales include original Landsat measurement scale of 30-meter resolution pixels covering each lake, lake average, lake maximum, lake area weighted watershed average at the HUC12 and HUC10 scales, and the lake area weighted watershed maximum at the HUC12 and HUC10 scales. Lake Champlain includes a couple watersheds and so an entire lake average is not used for this lake, but instead local averages within the lake are considered based on these watersheds. In order to account for the spatial misalignment between the lake scales smaller than the watershed size and the residential locations (i.e. houses are not built in the water), the small lake scales are matched to the regular lattice summarizing case intensity using a few different deterministic approaches including a fixed distance water area-weighted average and an inverse distance water area-weighted average. Distances are calculated to the centroids of the cells in the lattice used to summarize case intensity. Since deterministic approaches of matching lake water quality metrics to the summarized case intensity grid may underestimate the uncertainty in the effect of the exposure (Waller & Gotway, 2004, pp. 405-409), an exploratory analysis of spatial variability of the PC variable at the HUC12 scale will also be considered.

To account for the incomplete cases ascertainment, different area boundaries will be considered. The population at risk for the different area boundaries will be based on Census 2000 population estimates. Although this does not account for changes that may occur in the population at risk over the 13-year period from 1997 to 2009, it represents the best available data. Furthermore, to avoid the spatial misalignment that would result from using Census block or Census track aggregation scales of the population at risk with the regular lattice chosen to summarized case intensity, the background population will be estimated from two modeled population products of the 2000 Census, SEDAC’s 1km resolution population product for the 2000 US census (Seirup & Yetman, 2006) and ORNL’s LandScan 2000 1km resolution population product (LandScan 2000). The LGCP model parameters calculated using these two different backgrounds will be compared to check the sensitivity of these parameter estimates to small differences in the backgrounds.

Finally, there exists some temporal misalignment between the water quality data and the ALS disease onset of case data. The ALS disease onset for cases in this study occurred before or during the time period from 1997 to October of 2009 (Caller et al., 2013; Torbick et al., 2014). While, the lake average TN, SD and Chl-a were derived from Landsat images taken in late August and early September (days of year 242, 244, 248) for the years 2009 and 2010 (Torbick et al., 2014), and the PC metrics were derived from Landsat images taken in summers of 2014 and 2015. The use of these water quality metrics may produce biased exposure estimates particularly if the spatial distribution of the water quality in this region was quite different at the time of onset of the ALS cases. Little is known about the spatial distribution of the water quality in this region over time. There is recent evidence that lakes in this region have warmed and that the spatial distribution of the warming is heterogeneous (Smeltzer, Shambaugh, & Stangel, 2012; Torbick, Ziniti, Wu, & Linder, 2016). In order to address this concern, a general water quality metric, Trophic Status Index (TSI) is also considered as a fixed effect in the LGCP, aggregated over different years. In particular, the TSI metric is derived from Landsat images taken during the months of July, August and September between 2000 and 2005, and is averaged for four time periods, 2000 to 2001, 2002 to 2003, 2004 to 2005 and 2000 to 2005. Furthermore, a local analysis of the spatial distribution of lake skin temperature trends within Lake Champlain is considered. This investigation of temperature may be useful because there is evidence that warming temperatures have been associated with more frequent algal blooms (Paerl and Huisman, 2008, 2009).

The dissertation's chapters are organized as follows. Chapter II details terminology commonly used in the Epidemiological literature and details the expected counts estimation. Chapter III provides details about the ALS dataset and the metrics used for water quality. Chapter IV begins with an introduction to Bayesian Hierarchical models and estimation using Markov Chain Monte Carlo (MCMC) and Integrated Nested Laplace Approximation (INLA). Next is a brief introduction to Gaussian Process models and Spatial Point Process Models. This section ends with details about Gaussian Markov Random Fields. EAR Models from Hupper (2005) and Yuan (2011) are reviewed and the issue of collinearity between fixed effects and random effects is discussed. Chapter V outlines the use of computational grids for estimating ALS risk and also in the spatio-temporal modeling of Lake Champlain surface water temperatures. Chapter VI applies the BYM and Leroux models estimated via INLA and MCMC to various scales and boundaries of the NNE region to investigate spatial clustering of cases. Results are compared with clusters found in Caller et al. (2013) and Torbick et al. (2014). In Chapter VII, BYM model fits are compared with non-spatial Bayesian Poisson regression model fits estimated using INLA in order to estimate exposure effects of lake water quality on ALS risk. Various scales of the exposures variables (PC, TSI) are considered and ranked by significance. Chapter VIII discusses two case studies of exposure modeling. The first applies the EAR model (Yuan, 2011) in a spatially-varying coefficient model similar to the one also discussed in Hupper (2005) to satellite derived lake skin temperatures of Lake Champlain between 1984 and 2011. The second case study applies the EAR model with smoothing parameter equal to 1 to the spatial distribution of HUC12 aggregated maximum lake phycocyanin (in ug/L) derived from Landsat images during the summers of 2014 and 2015. In Chapter IX, conclusions are discussed and Chapter X ends with suggestions for future research.

II BASICS OF EPIDEMIOLOGY

2.1 TERMINOLOGY

In Epidemiology, interest lies in comparing risk between individuals or groups of individuals with different exposures. Exposure is a general term. It can refer to demographic factors such as age and sex, activities such as smoking and swimming in polluted lakes or location information such as town of residence.

When the spatial distribution of a particular disease is of interest, an epidemiologist will need to collect public health data in one of two forms, case event data or case count data. Case event data is a collection of geographically referenced points usually representing the residence of a disease case, while case count data is a collection of areal units such as census regions, administrative areas or zip codes with a count of the total disease cases having a residence location within the given areal unit. The two forms of data are related since case count data are spatially aggregated case event data. From a confidentiality point of view, spatially aggregated data is preferred since a map of exact case locations would still allow patient identification even when other identifying information such as patient name or social security number are not included in a dataset (Lawson, 2012). This is particularly true for rare diseases in rural areas. On the other hand, a relationship found between an exposure and the case counts on an areal unit that indicates an increased risk in the presence of the exposure may not exist or be the opposite (decreased risk with the exposure) when the exposure is analyzed with respect to the individual cases. This is known as the ecological fallacy, which is another version of Simpson’s paradox in the analysis of contingency tables or a special case of the modifiable areal unit problem (MAUP) discussed in the geography literature and the change of support problem in the geostatistics literature (Waller & Gotway, 2004, pp. 29-31). Thus when the research question pertains to individual level risk and/or effects at small spatial scales such as effects of pollution sources on surrounding communities, case event data is preferred because the unit of study is the individual and not an areal unit (Lawson, 2012).

From a modeling viewpoint, the distinction between the two types of spatial public health data, case event and case count, may be important dependent on other assumptions about the disease in question (Lawson, 2012; Waller & Gotway, 2004). In general, case count data is modeled using the binomial distribution or when the case counts are divided by expected counts, i.e. transformed to standardized incidence ratios (SIR), they can be modeled using the Normal approximation, while case event data are best modeled using the Poisson distribution in a spatial point process. However, when the disease is rare, both case event data and case count data are modeled with a Poisson distribution (Waller & Gotway, 2004; Li, Brown, Gesink, & Rue, 2012). The justification for the use of a Poisson distribution for case count data when the disease is rare comes from the relationship between the Binomial and Poisson distributions. As the probability of disease decreases and the study population increases, such that their product approaches a fixed number, the binomial probability of x disease cases in a population of n, approaches the Poisson probability of x disease cases in a fixed region. Furthermore, when the disease is rare a Normal approximation for the standardized incidence ratios (SIR) can be poor since case counts are small or zero unless the aggregation scale is quite large in space or time (Li et al., 2012).

A risk ratio is a common method used to compare risk estimates between groups having different exposures, and it addresses relative risk, the multiplicative impact of exposure. When estimating risk, the epidemiology literature is very precise about the quantities calculated. In particular, the definition of risk includes a specified time period that is required in determining the probability a person contracts the disease in question (Waller & Gotway, 2004, p. 9). Further distinctions are made with respect to this time frame for both the numerator and denominator of risk ratios. In the numerator, when only people who contracted the disease within the specified time period are counted this is called incidence and is the preferred quantity of interest when the study question is about what exposures are related with disease onset. Prevalence is a count of the all the people having the disease within the specified time period who may of either contracted it during the time period or before. Prevalence will be related to exposures that effect both onset and duration. In some cases, it may be impossible to know the exact time a disease was contracted and only a prevalence count can be calculated, but for diseases that have a short duration, the prevalence will be close to the incidence (Waller & Gotway, 2004, pp. 8-9).

The distinction between a rate and a proportion is defined by the quantity in the denominator of a risk ratio. In particular, a rate has a denominator that is the sum of the observation time for each person in the population at risk, while a proportion has a denominator that is the count of the population at risk. The population at risk fluctuates over time; people leave the study region; people die of unrelated causes and when people contract the disease in question, they no longer are part of the population at risk. For example, if 5 people make up a population at risk for a disease under study for 5 years, but 1 person contracts the disease after 1 year, and 2 people die in car crash at 2 years, the denominator for the proportion will be 5 people but the denominator for the rate will be 1+2*2+(5-3)*5 = 15 person years. When studying large populations at risk in an observational setting, it may be impossible to determine how long each person remains in the population at risk. However, when the disease is rare and/or the study period is short, proportions provide good approximations of rates (Waller & Gotway, 2004, pp. 9-10).

Standardized ratios are used to adjust raw ratios by removing the effect of known exposures not of interest for the study. For example, it is known that locations with higher population densities have more disease cases. Furthermore, some demographic factors such as age are known to correlate to disease occurrence. Disease occurrence is generally higher among older individuals than younger ones. Since a study population's density and demographic composition will vary spatially, it is important to calculate standardized ratios when one is interested in comparing risks between individuals at different locations in order to make the ratios from the different locations comparable (Waller & Gotway, 2004, p. 11). The standardized incidence ratio (SIR) also known as the standardized mortality ratio (SMR) compares the number of disease cases observed in a study population to the number that would be expected based on demographic (age, sex, a known factor not of interest) specific rates from a standard population (Waller & Gotway, 2004, p. 15). A number of different standard populations are possible. For example, when comparing risks between locations A and B, one could define the standard population to be the population at location A, the population at location B, the sum of the populations at locations A and B, or the population at a location C. In general, one chooses a standard population that makes the most sense for the particular study question, but issues of cost and accuracy of available information about a certain population will also be deciding factors. When the study question is related to determining if spatial clusters of disease cases exist, a common approach is to use the combined population from all sub-regions within the study region as the standard population.

2.2 ALS Expected counts

The denominator of the standardized incidence ratio (SIR) is called the standardized expected counts, which represent the number of the cases expected in the study population if the study population contracts a disease at the same rate as the standard population (Waller & Gotway, 2004, pp. 14-15). The question of interest for risk of ALS in Northern New England (NNE) is if the spatial distribution of ALS has clusters (or hot spots). Thus the standard population for the calculation of expected counts is defined to be the collection of all sub-regions under study. Since the ALS dataset being used in this analysis is suspected to underestimate the risk for the entire NNE region, different boundary areas for the sub-regions under study are considered. These include the boundary area of all NNE, the boundary area of just the states of Vermont and New Hampshire (VTNH) and the boundary area covered by all Vermont counties except for Bennington and all New Hampshire counties except for Cheshire, Hillsborough, Rockingham, and Strafford (VTNH_rmcty). These boundary areas are pictured in Figure II-1.

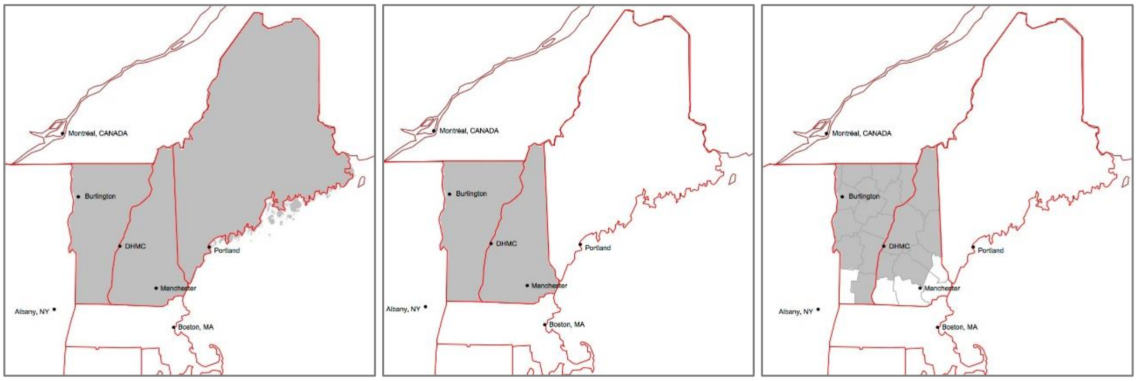


Figure II-1. Boundary areas of the study region are in gray; NNE (right); VTNH (middle); VTNH_rmcty (left).

Since the onset of ALS is age and sex related (Noonan, White, Thurman, & Wong, 2005), expected ALS cases in the study population will depend on the age/sex specific rates in the standard population. Noonan et al. (2005) defines 12 age/sex classes given as M & 0-44, M & 45-54, M & 55-64, M & 65-74, M & 75-84 and M & 85+, F & 0-44, F & 45-54, F & 55-64, F & 65-74, F & 75-84 and F & 85+, where M stands for male and F stands for female. Thus three sets of age/sex specific rates, one for each of the boundary areas, are calculated as follows:

$$r_i = \frac{\sum_{all\ x} O_i(x)}{\sum_{all\ x} n_i(x)},$$

where $i = 1, 2, \dots, 12$ is the identifier for one of the 12 age/sex classes defined by Noonan et al. (2005), x represents one of the sub-regions within a boundary area, $O_i(x)$ is the number of observed ALS cases in sub-region x having age/sex class i, and $n_i(x)$ is the total population at risk in sub-region x having age/sex class i. Following the calculation of the age/sex specific rates, the standardized expected counts for each sub-region x are calculated as:

$$E(x) = \sum_{all\ i} n_i(x) * r_i \,. \qquad [\mathbf{2.2.1}]$$

Note that the above standard population rates differ slightly from the rates used in both Caller et al. (2013) and Torbick et al. (2014) who use the sex specific direct-age adjusted rates that Noonan et al. (2005) calculates for motor neuron disease mortality over the entire United States for the time period 1994-1998. The method described in this analysis more explicitly corrects for the effect of age since the sex specific rates of Noonan et al. (2005) are direct-age adjusted to the age distribution for the entire United States that may different from the age distribution of Northern New England.

A caveat is that the age of diagnosis is not available for the dataset used in this analysis, and the observed age class counts are based on the case's age at year 2000 determined from the case's date of birth. Furthermore, several cases are missing date of birth or sex values. Counts of observed cases by age and sex class with the missing values also separated are shown in Table II-1.

Table II-1. Observed ALS case counts by age/sex class and boundary area

	sex	age	NNE	VTNH	VTNH_rmcty
1	F	0 - 44	28	20	15
2	F	45 - 54	60	44	29
3	F	55 - 64	63	49	34
4	F	65 - 74	60	43	26
5	F	75 - 84	39	22	16
6	F	85+	*	*	*
7	F	unknown	61	31	23
8	M	0 - 44	63	42	22
9	M	45 - 54	96	65	36
10	M	55 - 64	79	54	34
11	M	65 - 74	88	62	51
12	M	75 - 84	41	31	23
13	M	85+	*	*	*
14	M	unknown	68	43	32
15	unknown	unknown	*	*	*
16		Totals	762	516	347

*Values less than 10 are suppressed for privacy reasons.

In order to estimate the age/sex specific rates, cases with missing data were not excluded but added to the known data in such a way as to keep the proportions of the known data fixed. For example, consider the case of adding female cases of unknown age to those female cases with known age. Let $N(y)$ represent the number of y, and $P(y|A)$ represent the probability of y given A. The following expression describes the procedure for calculating total female and age class counts from both the known ages and unknown ages:

$$N(age_i \,\&\, F) + P(age_i|F) * N(unknown\ age \,\&\, F)$$

$$where \quad P(age_i|F) = \frac{N(age_i \,\&\, F)}{\sum_{all\ i} N(age_i \,\&\, F)}.$$

The modified case counts by age/sex and boundary area used to estimate age/sex specific rates are given in Table II-2.

Table II-2. Modified observed ALS case counts by age/sex class and boundary area

Age/sex class	NNE	VTNH	VTNH_rmcty
F & 0-44	34.71	23.37	17.80
F & 45-54	74.38	51.41	34.42
F & 55-64	78.10	57.26	40.36
F & 65-74	74.38	50.24	30.86
F & 75-84	48.35	25.71	18.99
F & 85+	*	*	*
M & 0-44	74.68	49.00	26.17
M & 45-54	113.80	75.83	42.82
M & 55-64	93.65	63.00	40.44
M & 65-74	104.32	72.33	60.66
M & 75-84	48.60	36.17	27.36
M & 85+	*	*	*
total	762	516	347

*Values less than 10 are suppressed for privacy reasons.

The sub-regions x of the study area from $[\mathbf{2.2.1}]$ are defined in Caller et al. (2013) as census block groups and the $n_i(x)$ values are the 2000 Census block group population counts by sex. In Torbick et al. (2014) sub-regions x are defined as census tracts and the $n_i(x)$ values are the 2000 Census tract population counts by sex. In this study, the sub-regions are 1km by 1km square grid cells, for which there are 221232 cells covering all of Maine, New Hampshire and Vermont and the $n_i(x)$ values are derived from two gridded population products at a 1km resolution representing the region’s population by the 12-age/sex classes at the year 2000. One of the gridded population products is provided by the Socioeconomic Data and Applications Center (SEDAC) (Seirup & Yetman, 2006) and the other is a product of OakRidge National Laboratory (ORNL) called LandScan (LandScan 2000).

SEDAC's 1km population product is lightly modeled meaning Census 2000 block counts were disaggregated using one data layer, while ORNL's LandScan product uses more data layers. ORNL's LandScan 2000 population product is also based off of population values of the 1990 Census projected to the year 2000. Given the difference in input variables used to model the populations between the two organizations, there is also a slightly different meaning to the population locations for each. SEDAC's population estimates were modeled using nighttime lights and so represent nighttime population locations while LandScan represents daytime populations (*What Are the Differences Between GPW, GRUMP and Landscan?* 2015). Neither of the SEDAC or LandScan 2000 1km population products are available by the age/sex classes defined in Noonan et al. (2005). Thus a simple approach that assumes demographic characteristics are constant at the census block level was used to apply the age/sex class totals from the 2000 Census at the census block level. Although this assumption may not be true for large census blocks, this represents the best-known available data.

First, age/sex class proportions were calculated for each 2000 Census block within the NNE region. Then the shape file of Census 2000 block polygons was rasterized to a 20-meter resolution lattice with the same extent and origin as the 1km SEDAC lattice and all lattice cells with centroid in the same polygon received the age/sex count from that polygon. The 20-meter resolution was chosen since this was the smallest width or length dimension of any of the census block polygons. Rasterized census block age/sex proportions made up a 12-layer raster brick, where each layer contained the proportion of each census block population belonging to a specific age/sex class. This 20-meter resolution raster brick was then aggregated to a 1km resolution by averaging the proportions in each layer separately. For a 1km cell falling completely inside a census block, the 1km cell would have the same value as all the 20-m cells. For 1km cells intersecting several census blocks, the proportion represents an area-weighted average of proportions from different census blocks, and finally if regions with no population were included in the 1km cells, these regions were excluded in the average calculation. After age/sex proportions were aggregated to the 1km resolution, the 12 layers were summed in order to verify each cell had a value of either 1 (at least one of the age/sex classes is present) or 0 (no population). Each of the age/sex class proportion layers was then multiplied by the population counts for each SEDAC and LandScan 2000 separately. The LandScan 2000 global lattice at 1km resolution cropped to the NNE region had a slightly different origin than SEDAC’s NNE 1km population raster. Thus, population counts from the LandScan 2000 were resampled to match the SEDAC lattice.

Table II-3. Comparison of total population counts by age/sex class for all of NNE

	Age/sex class	Census blocks	SEDAC	LandScan
1	F & 0-44	971469	826144.9408	837851.9386
2	F & 45-54	236326	252879.8987	258295.7725
3	F & 55-64	146905	177108.8892	178379.3266
4	F & 65-74	115673	135658.4292	139148.6193
5	F & 75-84	85539	93696.24544	97578.34566
6	F & 85+	37291	32589.75896	33682.87883
7	M & 0-44	979005	845637.6616	853779.2156
8	M & 45-54	234088	252396.0614	259875.2447
9	M & 55-64	142861	170199.82	173161.5455
10	M & 65-74	99533	124757.9363	126240.4717
11	M & 75-84	56594	67465.08834	68286.85924
12	M & 85+	14252	14499.35824	14502.82532
13	Total Female	1593203	1518078.162	1544936.881
14	Total Male	1526333	1474955.926	1495846.162
15	Total (after)	------------	2993034.088	3040783.044
16	Total (before)	3119536	3119535.991	3145258

Total population counts by age/sex class for the entire NNE region are summarized in the Table II-3. Rows 15 and 16 compare the total population estimate by SEDAC and LandScan 2000 before proportions were applied and afterward. Before proportion were applied the SEDAC total population for NNE matches the Census 2000 count, while the LandScan 2000 count is slightly larger. Recall, LandScan 2000 counts were developed before the 2000 US census was complete and is thus based on projections from the 1990 US census. After the proportions were applied, both the total populations for SEDAC and LandScan 2000 decrease. This is likely due to spatial uncertainty in locations without populations. Age/sex class differences between the Census blocks totals and calculated values for SEDAC and LandScan 2000 show that calculated values are less than Census values for both sexes younger than 45 and greater than Census values for both sexes older than 44 with the exception of females older than 84. It is important to note that all these estimates, Census 2000, SEDAC and LandScan 2000 contain some unquantifiable uncertainty about the true population in this region for this time. However, by using two different estimates of expected counts, the effect of estimates can be partially quantified.

III DESCRIPTION OF ALS CASE AND EXPOSURE DATA

3.1 NNE ALS CASE DATA

The ALS case datasets discussed in Caller et al. (2013) and Torbick et al. (2014), included case data collected from medical records at Dartmouth Hitchcock Medical Center (DHMC) and the Muscular Dystrophy Association (MDA) of Northern New England. Both studies report a thorough data collection quality control process that only included cases having year of diagnosed (or in some cases year of MDA registration) between January 1997 and October 2009 and who had a primary address in Northern New England (NNE). The collected information about each case included age at diagnosis, side of symptom onset, year of diagnosis, year of death, family history, dwelling address at time of diagnosis, and in some instances historical dwelling addresses (Caller et al., 2013; Torbick et al., 2014). Note that given the uncertainty with some of the times of diagnosis, this data is more likely a representation of ALS prevalence, however since life expectancy for ALS patients is short, the prevalence will be close to the incidence.

In the current analysis, only the variables date of birth, sex, and longitude/latitude coordinates for an approximate dwelling address at time of diagnosis were made available from the ALS case dataset used in Caller et al. (2013) and Torbick et al. (2014) for reasons of confidentiality. The dataset originally included 772 cases and also included two other columns, one with notes indicating additional information about the longitude/latitude coordinates and the other with a color-coding of red, yellow, or none. Location information was missing (missing coordinates and location notes) for 8 of the 772 cases. Thus these 8 cases were excluded, and a total of 764 cases were used in the analyses that follow.

There are some additional uncertainties present in the 764 cases used in the following analyses that could have biased some results. In particular, location notes were only recorded for 22 of the remaining 764 cases of which 9 gave a town name and the other 11 indicated an error message. Since the cases with a town name note were missing longitude and latitude coordinates needed for summarizing the data on the computational grid, these cases were assigned a set of coordinates using the geocode function in the R package ggmap, which makes use of Google Maps (Kahle & Wickham, 2013). Unique town names were available for 7 of the 9 cases and these were assigned to town centroids, while the other 2 were assigned to a town hall location and a public library location within the town respectively. This procedure does add spatial uncertainty for distances less than the town aggregation level. Furthermore, 4 of the 764 cases, only contained 2 unique date of birth, sex and location combinations. Since, these are likely duplicate cases, in the analysis of ALS within the VTNH boundary, 2 of these 4 cases were removed, giving a total of 762 cases considered. Despite the uncertainties in the ALS case data used for the following analyses, the author believes the data are sufficient for the goals outlined at the beginning of this dissertation.

3.2 Exposure Metrics for water Quality

There are several hypothesized routes by which people may be exposed to BMAA, the neurotoxin produced by cyanobacteria (blue-green algae) that is suspected to be a risk factor for ALS. These include drinking water, water sport activities such as swimming and boating, aerosolization of cyanobacteria blooms, and dietary exposure from eating seafood (Caller et al., 2013). In Northern New England, there are over 6,000 water bodies that could pose as a possible exposure source. Given the costs, both in time and money, of traditional water quality assessment needed for these over 6,000 water bodies, the use of satellite remote sensing technology is a viable alternative. The choice of a particular satellite sensor depends on the goals of the application since the spatial and temporal resolutions as well as the spectral sensitivity vary between different technologies. This application requires a fine spatial resolution with sufficient spectral sensitivity to derive metrics related to cyanobacteria for the entire NNE region. Torbick et al. (2014) show that the Landsat Thematic Mapper is a useful tool for deriving cyanobacteria related water quality metrics at 30 meter resolution pixels covering all lakes sized 6 hectares or more in NNE, through the use of band ratio regression techniques calibrated with in-situ lake sampled measurements.

3.2.1 Torbick et al. (2014) Water Quality Metrics

The water quality metrics used by Torbick et al. (2014) to explain census tract based ALS cluster membership include secchi depth (SD in meters), total nitrogen (TN in micrograms per liter (ug/L)), and chlorophyll-a (Chl-a in ug/L), and were derived from Landsat images taken in late August and early September (days of year 242, 244, 248) for the years 2009 and 2010. Lake averages of these metrics were calculated for all lakes sized 6 hectares or more, which included a total of 4453 lakes for which 3298 were found in Maine, 925 in New Hampshire, and 230 in Vermont. The average for Lake Champlain however was not included since it is located at the very western boundary of NNE and is likely not well represented by one average value due to its size (Torbick et al., 2014).

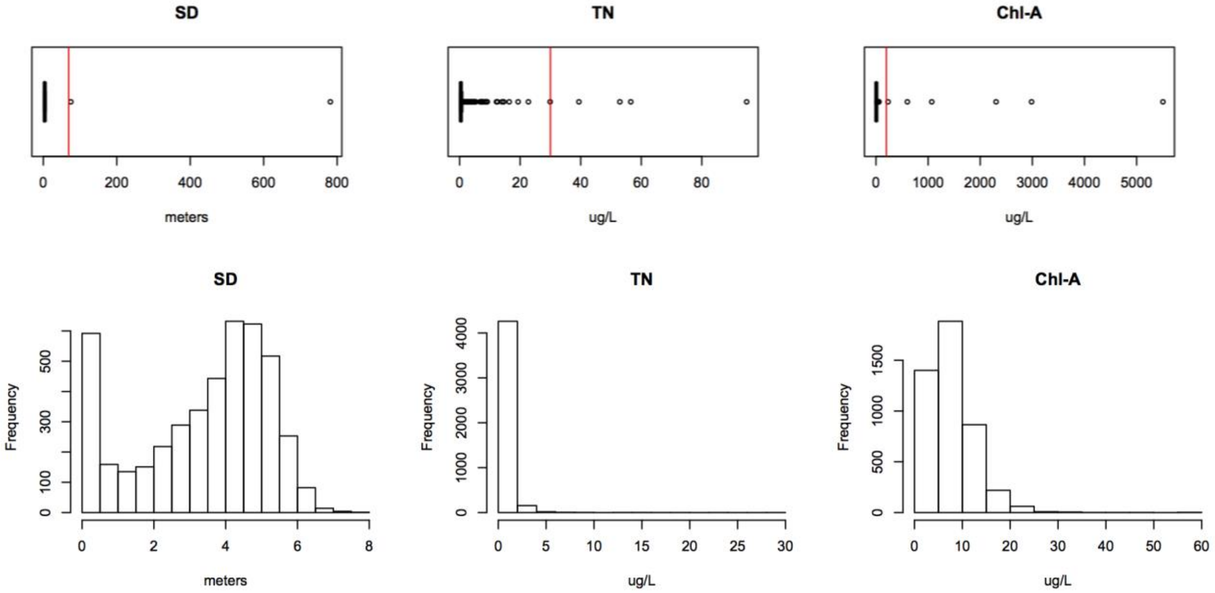


Figure III-1. Distribution of lake average SD, TN and Chl-a satellite derived metrics for all 4453 lakes in NNE greater than 6 hectares. Top: Metric boxplots of all lakes; Bottom: Metric histogram of non-outliers values only.

The box plots in Figure III-1 show that some of the satellite derived SD, TN and Chl-a metrics have values outside the range of observed in-situ measurement values of these metrics (values beyond red line). In particular, values of SD greater than 69 meters, TN greater than 30 ug/L and Chl-a greater than 200 ug/L were marked as outliers as these values are highly improbable for lakes in NNE. In general, both TN and Chl-a values have a positive skew (bottom of Figure III-1), and in general are positively correlated (Figure III-2): lakes with low average Chl-a values also have low average TN values.

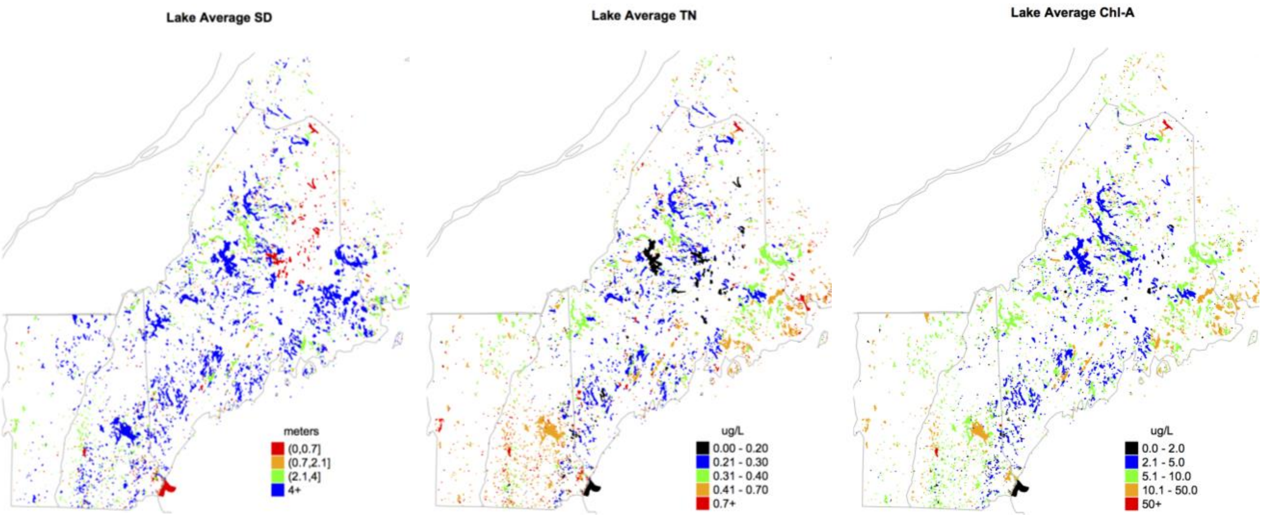


Figure III-2. Late Summer 2009/2010 snapshot values of lake average secchi depth, total nitrogen, and chlorophyll-a for lakes sized 6 hectares for more across Northern New England

Torbick et al. (2014) reports that these metrics show most water bodies in NNE over 6 hectares are considered healthy waters for this late summer 2009/2010 time snapshot, while only about 4% would be considered nutrient-rich lakes in which frequent and intense algal blooms would be likely to occur. Maps in Figure III-2 show a spatial distribution of these water quality metrics with the warmer colors (red/orange) indicating poorer water quality. In particular, in the right most map, lakes colored orange have Chl-a values that the World Health Organization (WHO) would classify as a moderate acute health risk and lakes colored red have Chl-a values that classify as a high acute health risk according to their guidelines for recreational waters (*Guidelines and Recommendations* 2016).

3.2.2 Phycocyanin (PC) Metric

Another satellite-derived metric used in this study that is related to the presence of cyanobacteria is phycocyanin (PC in ug/L). Phycocyanin is generally positively correlated with Chl-a and TN and negatively correlated with SD. This metric however is thought to be more closely related than Chl-a, TN and SD to the amount of BMAA to which the population is exposed since phycocyanin is a pigment of the cyanobacteria that produce the BMAA. Although, there is evidence for a disparity between the presence of cyanobacteria and the toxins it produces, which is not well understood as there is little mechanistic understanding related to why and when cyanobacteria produce toxins (Loftin et al., 2016).

The PC metric used in this study represents a late summer 2014/2015 snapshot of the amount of PC (ug/L) in all 30 meter water pixel covering all lakes sized 6 hectares or more (Figure III-3).

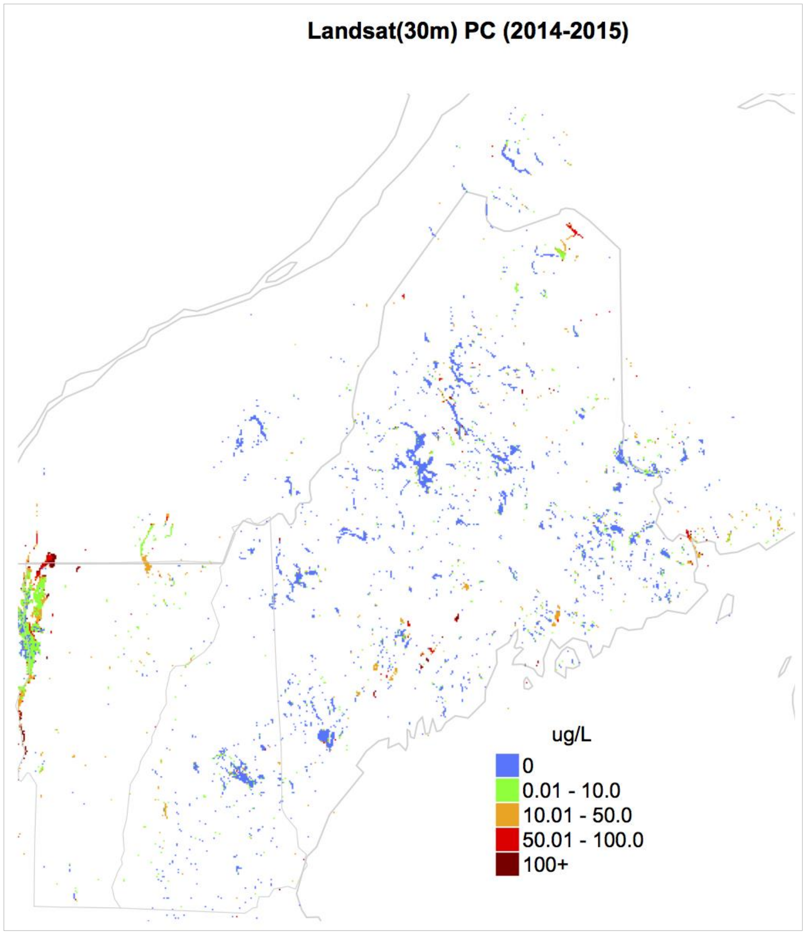


Figure III-3. Late Summer 2014/2015 Snapshot Representation of the amount of phycocyanin for each 30m water pixel for all lakes sized 6 hectares or more in Northern New England

Lake averages of these pixel PC values range from about 0 to 1200 ug/L with the majority of lake averages below 10 ug/L (Figure III-4).

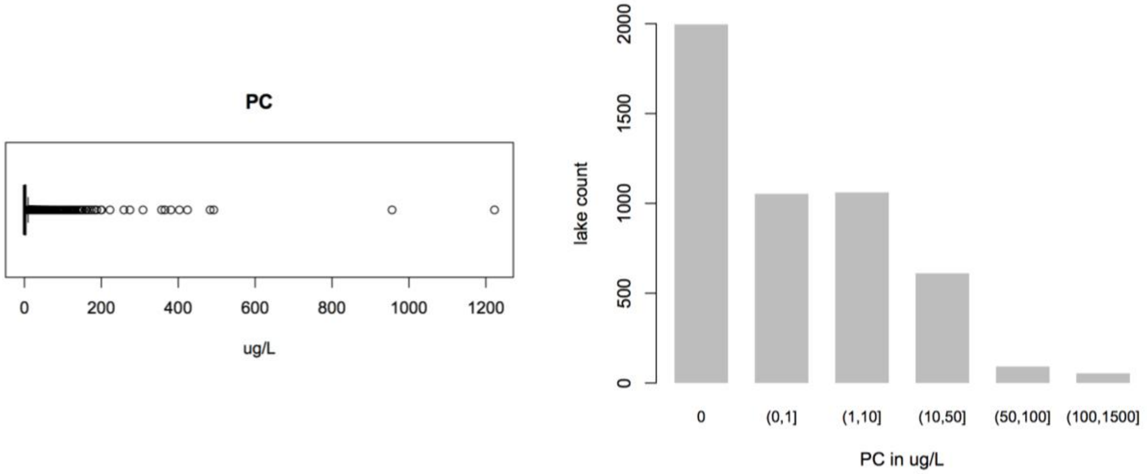
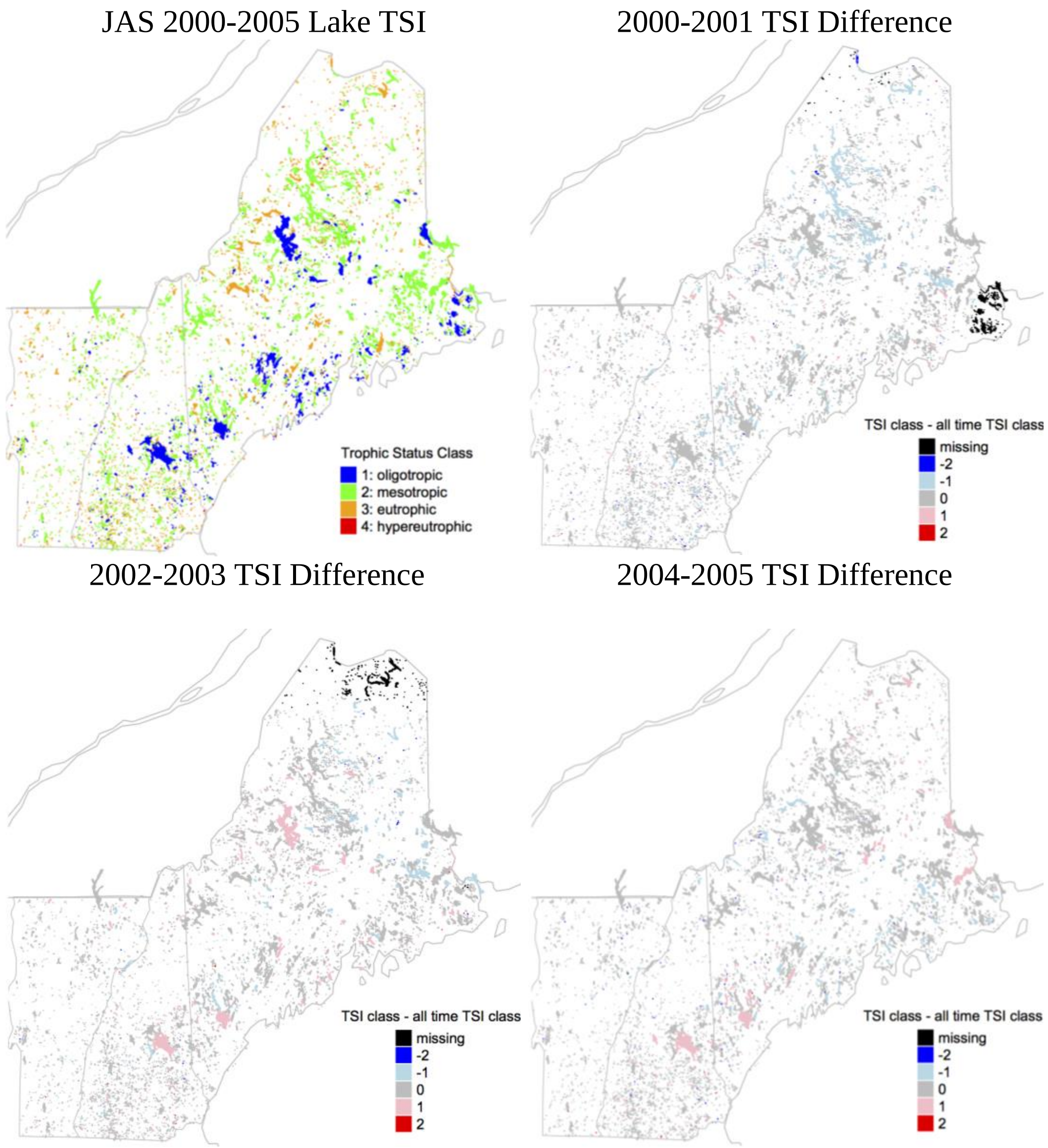


Figure III-4. Lake average phycocyanin (ug/L) for 4867 lakes sized 6 hectares or more across Northern New England based on Landsat derived metric at 30 meter resolution pixels.

3.2.3 Tropic Status Index (TSI) Metric

Trophic Status Index (TSI) is an ordinal metric used by ecologists to classify water bodies based on their varying amounts of nutrients and is an indicator of water health (Carlson 1977). Generally, an entire water body is given one classification, however for large water bodies with spatially varying amounts of nutrients, sections of the water body may be classified differently. Index values are 1: oligotrophic, 2: mesotrophic: 3: eutrophic and 4: hypereutrophic, where eutrophic and hypereutrophic water bodies contain the most nutrients. These lakes are most likely to experience frequent and intense algae blooms, while oligotrophic lakes are clear and good sources of drinking water. A number of metrics may be used to classify the trophic status of a water body including amounts of total nitrogen, total phosphorus, abundance of algae, total chlorophyll-a, and secchi depth.

This study considers a TSI metric based on secchi depth derived from the spectral bands of Landsat Thematic Mapper calibrated using in-situ secchi depth measurements from lakes across the Continental United States that were included in the National Lake Assessment 2007. This TSI metric gives a classification for each 30-meter water pixel covering all lakes in Northern New England sized 6 hectares or more for the particular day and year of a Landsat overpass. A lake classification for a particular day and year is based on the mode (most common) classification of all 30-meter water pixels within a given lake. Due to its size, a lake classification was not given to Lake Champlain. Lake classifications based on Landsat images taken during the months of July, August and September (JAS) for the years between 2000 and 2005, were then used to calculate JAS classifications for the time periods, 2000 to 2001, 2002 to 2003, 2004 to 2005 and 2000 to 2005. JAS classifications were determined by using the median classification of all 50% cloud free Landsat overpass dates within the specified time period for each lake. The JAS 2000-2005 TSI lake classifications for lakes (n=5,809) across NNE are shown in Figure III-5. The three other time period lake classifications are compared to the JAS 2000-2005, where blue colors indicate a clearer classification than the one for 2000-2005 and red colors indicate a murkier classification than the one for 2000-2005.



JAS 2000-2005 Lake TSI

2000-2001 TSI Difference

2002-2003 TSI Difference

2004-2005 TSI Difference

Figure III-5. JAS 2000-2005 TSI lake classifications compared with the early period 2000-2001 classification, the middle period 2002-2003 classification and the late period 2004-2005 classification.

Landsat overpass dates vary by location in NNE because Landsat has a square footprint of about 180km and requires 12 frames to create wall-to-wall maps of TSI across all NNE. Furthermore, approximately 10 to 11 Landsat overpass dates are possible in a given year. However, the number of images available for each lake will vary by location because adjacent Landsat orbit paths overlap resulting in additional observations of some lakes and cloud cover also varying by day and location will result in less observations for other lakes. Overpass frequency and day of year variability by location for each of these time periods is shown in Figure III-6 and Figure III-7 respectively.

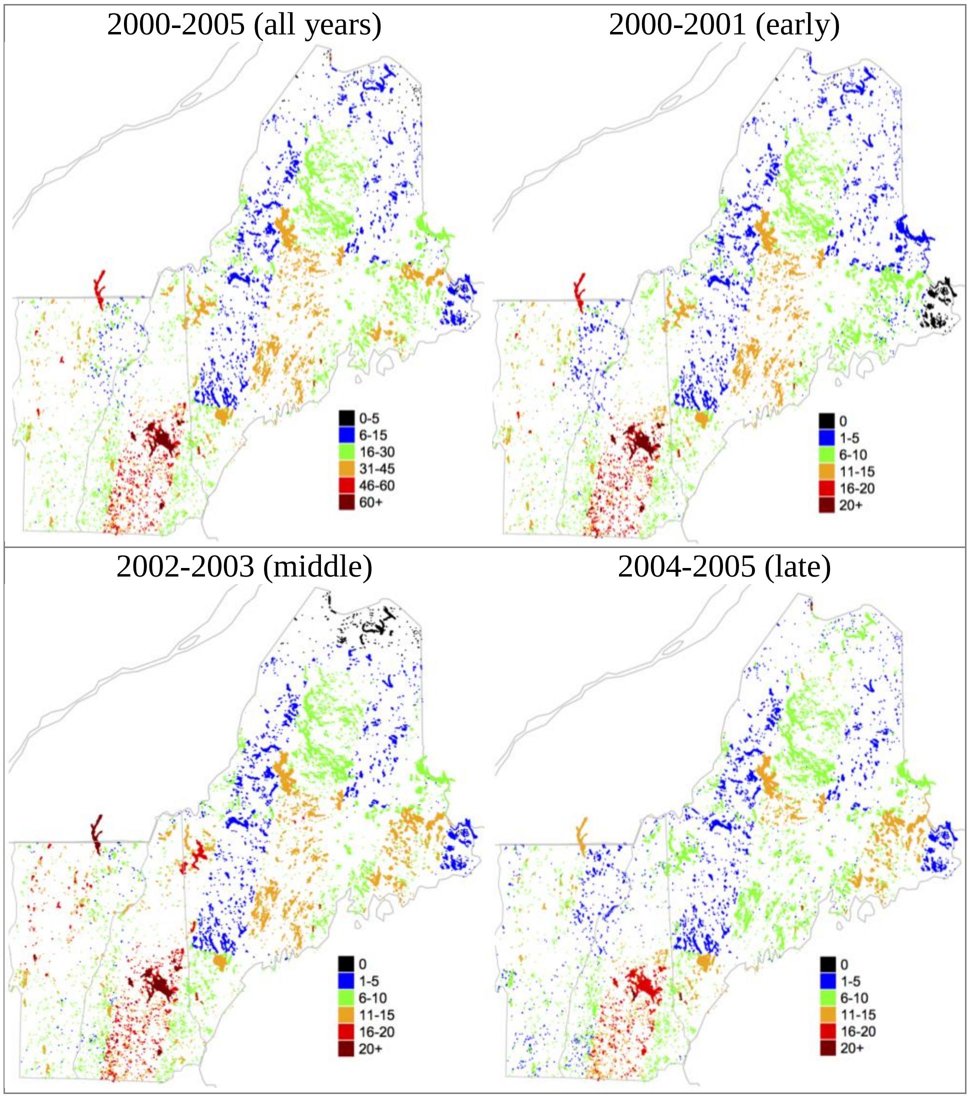


Figure III-6. Number of 50% cloud free Landsat overpasses by lake for each time period

The variability in cloud-free overpass dates and frequency may explain some of the variability in TSI values between time periods, but more research is need to fully characterized the effects of the irregular sampling pattern on the TSI spatial variability across NNE. However, almost half the lakes (n=2813) have the same classification for all time periods and only about 6% of lakes (n=330) vary by 2 or 3 classes.

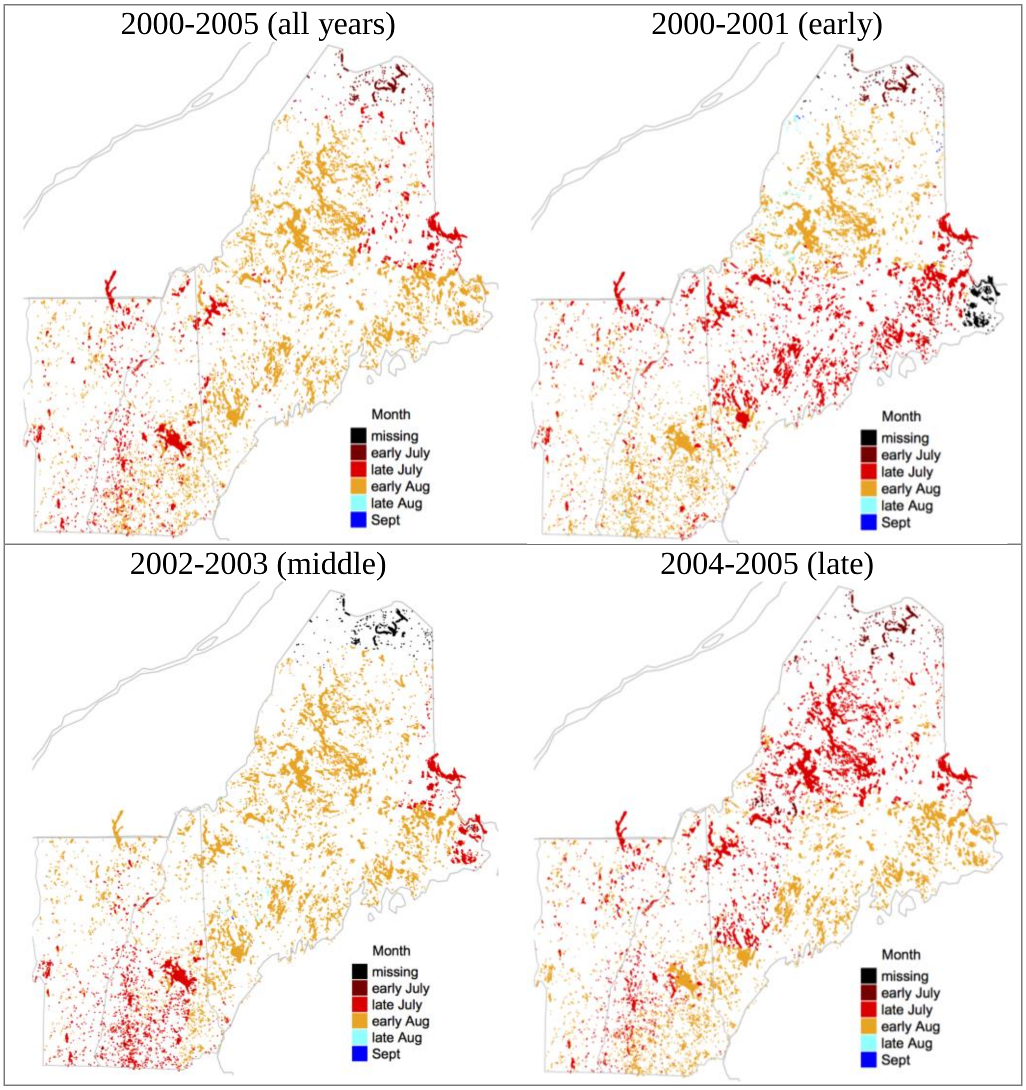


Figure III-7. Average time of month for all 50% cloud free Landsat overpass dates by lake for each time period.

3.2.4 Lake Champlain Surface Water Temperatures

Lake water temperature is a central driver regulating lake ecology. Recent studies suggest temperatures are increasing in many inland lakes (Schneider & Hook, 2010; Smeltzer, Shambaugh, & Stangel, 2012; Torbick, Ziniti, Wu, & Linder, 2016). These temperature increases are suspected to increase the frequency, magnitude, and duration of cyanobacterial harmful algal blooms. Understanding the spatio-temporal dynamics of lake temperature change may provide valuable insight in how other ecological processes vary within the lake, including the development and duration of algal blooms.

Lake Champlain is the largest lake in the NNE region for which there exists several valuable long-term water temperature records and thus is an ideal candidate for a case study of the spatio-temporal dynamics of lake temperature change. In a previous study, temperature data from two water monitoring programs, a survey by early limnologists E.B Henson and M. Potash from the University of Vermont from 1964 to 1974 and the Long-Term Water Quality and Biological Monitoring Program on Lake Champlain supported by the Lake Champlain Basin Program from 1992 to 2009, were analyzed showing that August mean surface water temperatures of Lake Champlain rose by 1.6-3.8 degrees Celsius between 1964 and 2009 (Smeltzer et al., 2012). Furthermore, this increase was statistically significant for 8 out of the 10 sampling sites considered. It is likely that the temperature trends at nearby sites exhibit positive spatial autocorrelation and that trends could be calculated with less uncertainty, if this autocorrelation is included in the calculation. The limited number of sites in this previous study however do not provide sufficient information to make this modeling useful. On the other hand, Landsat Thermal Imagery can be used instead since the spectral bands provide satellite-derived measurements of water temperature for a 31-year period from 1984 to 2014 in a 30 by 30 meter uniform gridded format covering the entire lake.

For this case study, surface water temperatures derived from Landsat images calibrated to in-situ water measurements were collected for the months of July, August, and September (JAS) for the investigation of the summertime spatio-temporal temperature trends within Lake Champlain. Only Landsat images of path row tile 014029, which covers most of Lake Champlain except for the very narrow southern section were included and the JAS metric was chosen to represent temperatures in which algae most commonly bloom. Two Landsat satellites were in operation during the 31-year period, each having a 16-day interval between images. Landsat LT5 captured images during the years 1984 to 2012 and Landsat LE7 began collecting in 1999 with an 8-day collection offset to Landsat LT5 (Figure III-8).

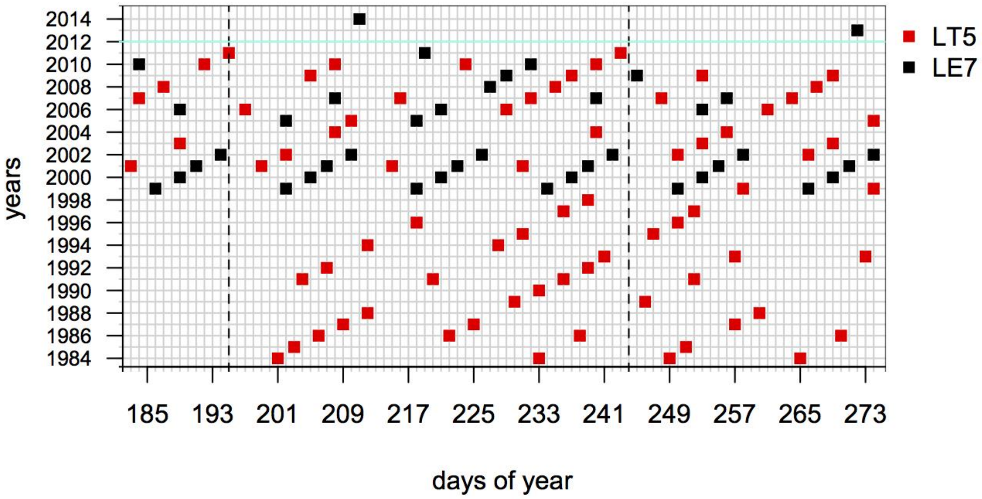


Figure III-8. Sampling Pattern of Landsat LT5 and LE7 for mostly cloud free images of path row tile 014029 over Lake Champlain

Although the two satellites cumulatively captured a total of 270 images of tile 014029 during this period, only 114 of these images could be used to calculate water temperatures due to poor image quality and/or extensive cloud cover. Also day of year sampling patterns vary from year to year. In particular no samples before day of year 200 are available for years before 1999, and the sampling frequency is very sparse or nonexistent for years after 2011. Furthermore, average lake temperatures exhibit a systematic quadratic variability by day of year for the July to September collection months (Figure III-9).

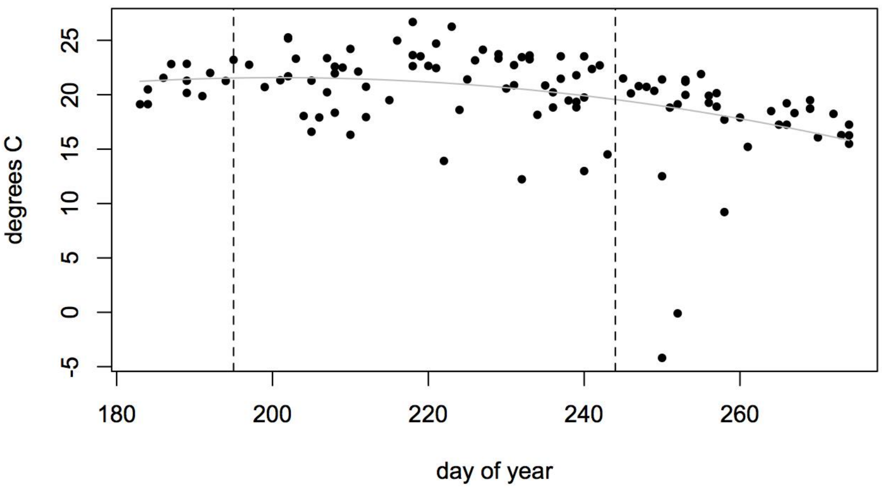


Figure III-9. Average lake satellite-derived surface water temperatures of Lake Champlain from 1984-2014 by day of year

Thus, to control for trend bias due to the sparse and irregular sampling pattern and due to the day of year variability, only images taken between day of year 195 and 244 and for years before 2012 were used for spatio-temporal trend estimates. This gave a total of 63 time points over the 27-year period from 1984 to 2011.

Satellite derived temperatures also have the potential for more noise than in-situ measured temperatures due to varying atmospheric conditions. Figure III-10 shows that there are a number of outlier values, particularly lower outliers, in the satellite-derived temperatures of Lake Champlain compared with in-situ buoy measured temperatures.

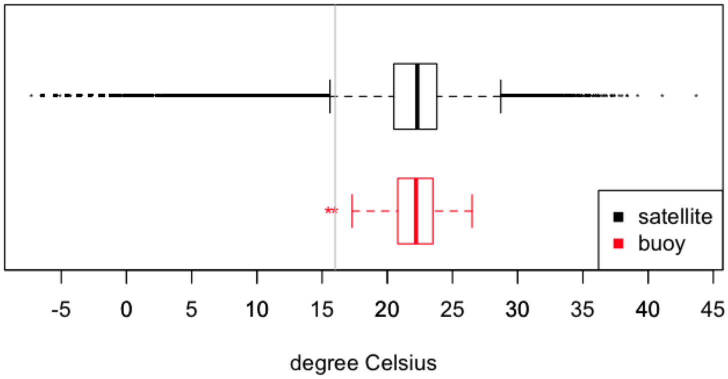


Figure III-10. Satellite-derived surface water temperatures of Lake Champlain from 1984-2011 for days of year 195 to 244 vs. buoy in-situ water temperatures from 1992-2013 for days of year 195 to 244 and depths below 4 meters. (Gray line at 16 degrees Celsius.)

Cloud masking algorithms are used to remove effects from clouds, but when there are conditions of light haze, these algorithms can be ineffective, resulting in low values. For example, Figure III-11 shows the satellite-derived temperatures for year 1989 on day of year 230 next to a Landsat image of the section of lake near St. Albans Bay for the same day. Satellite derived temperatures values are between 10 and 16 degrees Celsius for the same location the image shows hazy cloud cover.

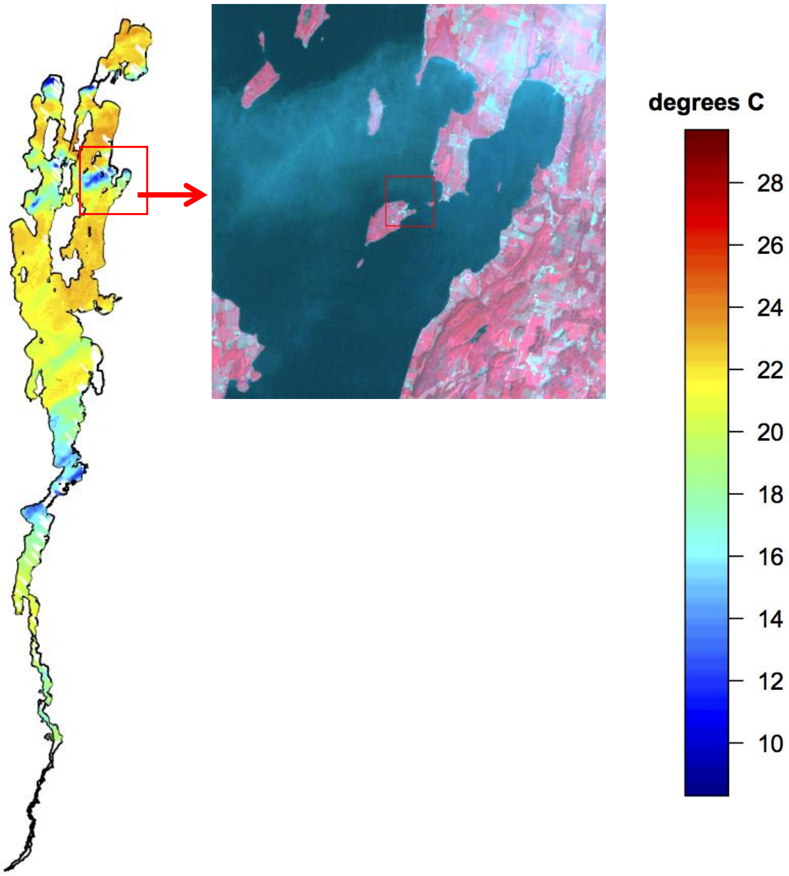


Figure III-11. Satellite-derived surface water temperatures of Lake Champlain and Landsat image near St. Albans Bay for Year 1989 and Day of Year 230

In order to remove the effect of temperatures estimated when conditions of light haze were present, a low temperature threshold was chosen to be 16˚C, and temperatures less than 16 were marked as missing.

IV STATISTICAL MODELS

4.1 Bayesian Inference for Hierarchical Models

In Bayesian statistics all quantities of interest are considered random and are modeled as a probability distribution, which, when conditioned on data lead to a posterior distribution by applying Bayes' theorem (Gelman, Carlin, Stern, & Rubin, 2004).

Bayes' Theorem (Simple model)

Let Y represent the data,
θ the parameter or quantity of interest,
and $P(..)$ the probability function or density, then:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{\int P(Y|\theta)P(\theta)d\theta} \propto P(Y|\theta)P(\theta)$$

$P(\theta|Y)$ is the posterior distribution.
$P(Y|\theta)$ is the data likelihood.
$P(\theta)$ is the prior distribution of the quantities of interest.

Through a sequence of hierarchical specification that applies repeated conditioning, a more complex yet realistic likelihood can be defined. For example, uncertainty in the data can be specified conditionally on a process, which, in turn, can be specified conditionally on parameters for which uncertainty is captured in prior distributions (Cressie & Wikle, 2011). This framework of repeated conditioning is known as a Bayesian hierarchical model.

Bayes' Theorem (Hierarchical Models)

Let Y represent the data,
Z a latent process,
θ the parameters,
and P(..) the probability function or density, then:

$$P(\theta, Z|Y) \propto P(Y|Z,\theta)P(Z|\theta)P(\theta)$$

where $P(Y|Z,\theta)$ is the probability distribution of the data layer,
$P(Z|\theta)$ is the probability distribution of the process layer
and $P(\theta)$ is the probability distribution of the prior layer.

In the analysis of spatial public health data in the form of case event data, the most appropriate model is a spatial point process (Simpson, Illian, Lindgren, Sørbye, & Rue, 2016). An Inhomogeneous Poisson Process (IPP) is a spatial point process that is useful for modeling clustering of cases having a non-contagious disease, where the clustering would result from environmental heterogeneity instead of case interactions because the IPP assumes independent counts in non-overlapping areas have a Poisson distribution conditional on a spatially varying intensity, $\boldsymbol{\Lambda}$, (mean number of points per square area) (Diggle 2013). In particular, the spatially varying intensity of an IPP used for disease event data is assumed to factor into a deterministic component,$\boldsymbol{E}$, representing the background population of people at risk and a modeled component representing relative risk due to environmental heterogeneity (Lawson 2012). When the modeled component is also stochastic, the IPP is known as a Cox process (Diggle 2013). A Log-Gaussian Cox Process (LGCP) is a particularly flexible yet tractable Cox Process that uses the logarithm of a Gaussian Process (GP) to model the relative risk component of the intensity function (Møller, Syversveen, & Waagepetersen, 1998). A Bayesian hierarchical model for the LGCP is specified with the data layer having the disease counts, $\boldsymbol{O}$, follow a Poisson distribution, a process layer specified by the logarithm of a Gaussian Process (GP) and the prior layer specified by non-informative or weakly informative distributions for the parameters of the Gaussian Process.

Bayesian Hierarchical Log Gaussian Cox Process

Data layer	$\boldsymbol{O} \sim Poisson(\boldsymbol{\Lambda})$
Process layer	$\log(\boldsymbol{\Lambda}) = \boldsymbol{X\beta} + \boldsymbol{\omega} + \log(\boldsymbol{E})$ $\boldsymbol{\omega} \sim GP[0, \sigma^2 \boldsymbol{R}(\boldsymbol{\phi})]$
Prior layer	$\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}$ are given non-informative distributions

A similar hierarchical model can be defined for spatial environmental data. When the data represent a continuous process such as temperature, the data layer is specified with the data following a Normal distribution, which depends on a mean parameter and a variance parameter. The process layer is specified with the mean parameter from the data layer as following a Gaussian Process and then the prior layer includes non-informative or weakly informative distributions for the parameters of the Gaussian Process.

Bayesian Hierarchical Gaussian Process

Data layer	$\boldsymbol{Y} \sim Normal\left(\boldsymbol{\mu}, \sigma_y^2\right)$
Process layer	$\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{\omega}$ $\boldsymbol{\omega} \sim GP[0, \sigma_\omega^2 \boldsymbol{R}(\boldsymbol{\phi})]$
Prior layer	$\sigma_y^2, \boldsymbol{\beta}, \sigma_\omega^2, \boldsymbol{\phi}$ are given non-informative distributions

For high-dimensional parameter models, such as the hierarchical model, the normalizing constant of the posterior distribution, which is the denominator $\int P(\boldsymbol{Y}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}$, is difficult to compute and/or may not have an analytical solution. Modern Bayesian data analysis thus relies on Markov Chain Monte Carlo (MCMC) simulation or numerical approximations such as

Integrated Nested Laplace Approximation (Rue, Martino, & Chopin, 2009) to obtain samples from the posterior distribution.

MCMC is an iterative sampling method that draws samples from approximate distributions, called transition distributions, which through the use of a Markov Chain are improved with each iteration and eventually converge to a target distribution which is the posterior distribution of the hierarchical model (Gelman et al., 2004, pp. 285-287). The transition distribution at any one time, t, depends only on the previous time, $t - 1$. This is the defining characteristic of the Markov property. When the transition distribution is defined in such a way that the Markov chain is ergodic, there exists a steady state distribution. Furthermore, the transition distribution can be defined so that this steady state distribution is the posterior distribution. For analysis, the beginning samples of the chain are discarded as warmup or burn-in samples since these samples are not drawn from the converged distribution (Gelman et al., 2004, pp. 294-295). In practice, it is not known how many iterations are needed before the chain converges to a steady state. Thus it is common to choose a large number of samples ahead of time and to discard a fixed number for warmup. Convergence is then verified by various methods such as running multiple chains simultaneously with different starting values and then comparing their post warmup samples using the Gelman-Rubin statistic, which calculates the ratio of the between chain variability to the within chain variability. Values of the Gelman-Rubin statistic that are close to 1 indicate the chains have converged (Gelman & Rubin, 1992).

One general class of methods for constructing the transition distribution of a Markov chain that leads to an arbitrary posterior distribution is known as Metropolis-Hastings (Hastings 1970). In this approach, after choosing starting values,$\boldsymbol{\theta}^0$, proposal values for the next iteration,

$\boldsymbol{\theta}^*$, are drawn from a jumping distribution $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{\boldsymbol{t-1}})$ and then the next values, $\boldsymbol{\theta}^t$, $t = 1,2,\ldots$ are determined by an acceptance rule:

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* & with\, probability\ \ \min(1,\mathrm{r}) \\ \boldsymbol{\theta}^{t-1} & otherwise \end{cases}$$

where

$$r = \frac{p(\boldsymbol{\theta}^*|y)/J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}{p(\boldsymbol{\theta}^{t-1}|y)/J_t(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}$$

If the jumping distribution is symmetric, as is the case for the Normal and Uniform distributions, then $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}) = J_t(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)$ and r simplifies to

$$r = \frac{p(\boldsymbol{\theta}^*|y)}{p(\boldsymbol{\theta}^{t-1}|y)}$$

This was an assumption used in the original Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). Note that $p(\theta^t|y)$ is the posterior distribution, which is always known up to a normalizing constant and the ratio of $\frac{p(\boldsymbol{\theta}^*|y)}{p\left(\boldsymbol{\theta}^{t-1}|y\right)}$ does not depend of this normalizing constant. In cases where the full conditional posterior distribution $P(\theta_i|\theta_{-i},Y)$ of the parameter θ_i is known, where θ_{-i} denotes all θ except for θ_i, the parameters can be sampled directly. This individual direct sampling is known as Gibbs sampling (Gelman et al., 2004, pp. 287 -288). If the unknown parameters can be circularly sequenced so that for each parameter in the sequence the full conditionals are known, then the Gibbs sampler lends itself for a sequential framework of individual parameter updating which has become the core of MCMC. If for any parameter in the sequence, the full conditional is not known, a Gibbs sampler step is replaced by the more general Metropolis (Hastings) step. This is generally referred to as "Metropolis within Gibbs" MCMC.

The use of a jumping distribution in the Metropolis Hasting algorithm and the Gibbs Sampler results in a random walk Markov Chain. When attempting to sample from a hierarchical distribution, the behavior of a random walk Markov Chain can be quite undesirable as it may not efficiently explore the posterior distribution and samples may retain significant autocorrelations. There are a number of remedies suggested in the literature such as running the chain much longer than would be necessary to reach a desired Monte Carlo error threshold, and then to thin the chain, throwing out every nth sample (Gelman et al., 2004, pp. 291-308).

An alternative to a random walk Metropolis, is to use Hybrid or Hamiltonian Monte Charlo (HMC), which instead of using a jumping distribution, uses Hamiltonian dynamics to get a proposal value. In order to use Hamiltonian dynamics, extra “momentum” variables are introduced and the leapfrog method is generally used to approximate the system (Neal 2011). Since the leapfrog method is time reversible, after the proposal values, $\boldsymbol{\theta}^*$, are obtained, the next values of the Markov chain, $\boldsymbol{\theta}^t$, $t = 1,2, \ldots$ can be determined as in the original Metropolis algorithm (Metropolis et al., 1953) by the acceptance rule:

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* \;\; with\, probability \min(1, r) \\ \boldsymbol{\theta}^{t-1} \;\; with\, probability\; 1 - \min(1, r) \end{cases}$$

where

$$r = \frac{p(\boldsymbol{\theta}^*|y)}{p\left(\boldsymbol{\theta}^{t-1}|y\right)} \; .$$

MCMC designed using HMC can converge to the steady state distribution with less iterations and posterior samples will have much less autocorrelations than MCMC designed using random walk Metropolis, provided careful tuning of the step size and number of steps for the leapfrog algorithm is implemented (Neal 2011). However, Hoffman and Gelman (2014) have developed an implementation of HMC called the No-U-Turn Sampler (NUTS) that eliminates the need to preset the number of steps and adaptively selects the step size making it useful for ready-to-use software available to general practitioners. The No-U-Turn Sampler is the default sampler available in the recently developed Stan software (Stan Development Team 2016).

For hierarchical models having a data layer defined by a probability distribution of the exponential family and a process layer defined by a Gaussian Markov Random Field, a numerical approximation known as Integrated Nested Laplace Approximation (INLA) can be used for very fast estimation of the posterior marginal distributions, which in the cases of non-Gaussian data probability distributions have no closed form solution (Rue et al., 2009). The Log-Gaussian Cox Process is such an example. INLA estimates the marginal posterior distributions

$$P(z_i|\boldsymbol{y}) = \int P(z_i|\,\boldsymbol{\theta},\boldsymbol{y})P(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta}$$

and

$$P\left(\theta_j\middle|\boldsymbol{y}\right) = \int P(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta_{-j}}$$

by obtaining a Laplace approximation for $P(\boldsymbol{\theta}|\boldsymbol{y})$, a Laplace approximation for $P(z_i|\,\boldsymbol{\theta},\boldsymbol{y})$ and then using these approximations in a numerical integration (Rue et al., 2009). This is computational fast because $\boldsymbol{\theta}$ has few dimensions and the Gaussian Markov Random Field’s precision matrix is sparse. A Laplace approximation uses the Gaussian density to approximate an integral (Gelman et al., 2004, pp. 341-342). There is debate about whether in practice INLA or MCMC is better for Bayesian posterior model estimation (Taylor & Diggle, 2014). However, opponents of INLA still agree INLA can be particularly useful for model selection when many candidate models are available for consideration (Diggle, Moraga, Rowlingson, & Taylor, 2013).

4.2 Introduction to Gaussian Processes

A Gaussian Process, $GP(\theta)$, is a real-valued function having domain $\theta \in \mathbb{R}^d$, where in the analysis of spatio-temporal data $d = 3$, such that for any $\{\theta_n |\, n \in \mathbb{N}\}$, $\boldsymbol{Y} = [GP(\theta_1) \quad \dots \quad GP(\theta_n)]$ has a n-dimensional multivariate Normal distribution with probability density function

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}_{nxn}|^{-\frac{1}{2}} \, exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}$$

where Σ is required to be positive definite (Rasmussen & Williams, 2006).

The Gaussian Process is characterized by a mean function $E[GP(\theta)]$ and covariance function $Cov[GP(\theta_1, \theta_2)]$. The mean function is generally interpreted in the spatio-temporal context to describe large-scale space-time trends, while the covariance function describes small-scale dependencies or autocorrelations, although trends and autocorrelations are not necessarily mutually exclusive properties. For example, with the right set of basis functions $g_n(\boldsymbol{X})$ that span the spatio-temporal domain, the variability of the observed process $\boldsymbol{Y}$ can be explained by the mean function $E[GP(\theta)] = \sum_n g_n(\boldsymbol{X})$ and a simple covariance function

$$Cov\left[GP\left(\theta_i, \theta_j\right)\right] = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases},$$

which implies the space-time locations θ are independent conditional on the basis functions. In this case, the probability density function simplifies to a product of n univariate Normal densities, and this is the widely used multiple linear regression model. On the other hand, the same observed process $\boldsymbol{Y}$ can be explained by a constant mean function $E[GP(\theta)] = \mu$ and a complex covariance function

$$Cov[GP(\theta_i, \theta_j)] = \begin{cases} \rho_{ij}\sigma_i\sigma_j & i \neq j \\ \sigma_i^2 & i = j \end{cases}, where\ \rho_{ij} = \rho_{ji}.$$

This second method implies the interpretation that the process is generated chiefly by the dependence structure, and this is a form of non-parametric regression (Rasmussen & Williams, 2006, pp. 2).

In practice, simplifying assumptions are needed for the mean function and/or the covariance function in order for estimation from observed data to be possible, and the choice between modeling the mean or the covariance is often based on the scientific theory about how the data are likely generated. In the spatio-temporal context where data commonly exhibit positive spatial and/or temporal correlations, it is common to model some or all of the variation in the covariance function, and even when much of the variation can be explain by predictor variables in the mean function, prediction can be improved by not assuming an independent covariance function.

The requirement of positive definiteness for the covariance function limits the types of covariance functions one can use. In general, the choice of covariance function is motivated by the form of the observed data. When modeling spatial data observed at irregular point locations known as geostatistical data, it is common to calculate an empirical variogram, which compares the value of the measured variable to pairwise distances between locations at which it was measured. The variogram is then used to choose a particular covariance function from a list of functions known to satisfy the positive definiteness requirement. The most widely used class of geostatistical covariance functions is the Matérn class (Stein 1999), which under the assumptions of stationarity and isotropy is defined as:

$$\text{Cov}(h) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^{\nu} K_{\nu}(\kappa h), h \geq 0$$

where $h = \left\|\theta_i - \theta_j\right\|$, $\nu > 0$ is a smoothing parameter, $\kappa > 0$ is a range parameter and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind. The exponential covariance function is a special case of the Matérn when $\nu = \frac{1}{2}$, and the Gaussian or squared exponential covariance function is a special case as $\nu \rightarrow \infty$. Stationarity is the assumption that dependency is a function of distance alone, and isotropy is the assumption that dependency is not a function of direction. These are common assumptions when other variables are used in the mean function to describe large-scale spatial trends. However, for examples of non-stationary spatial covariance functions see Sampson and Guttorp (1992), Higdon (1998), Nychka, Wikle and Royle (2002), and Paciorek (2006).

When spatial data are observed at regular locations on a grid or as aggregated spatial units such as administrative districts known as areal data, Gaussian Markov Random Fields (GMRF) are commonly used to model the dependency structure (Rue & Held, 2005).

Definition (Rue & Held, 2005, pp. 21-22):
A GMRF is a random vector $\boldsymbol{x}$ having a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and positive definite precision matrix (inverse of the covariance matrix) $\boldsymbol{Q}$, where $\boldsymbol{x}$ is indexed by a set $\mathcal{V} = \{1, \ldots, n\}$ containing the vertices of an undirected labeled graph G, and $Q_{ij} \neq 0$ when $\{i, j\} \in \mathcal{E}$ for $i \neq j$. $\mathcal{E}$ is the set of two element subsets of $\mathcal{V}$ called the set of edges in the labeled graph G.

Thus, the dependence structure of a GMRF is determined by the edges $\mathcal{E}$ in the graph G. In particular, by the defining the neighbors of the vertex i as:

$$ne(i) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$$

the conditional distribution of $x_i \in \boldsymbol{x}$ given $\boldsymbol{x}_{-i}$, which denotes all $x \in \boldsymbol{x} \backslash x_i$ is as follows:

$$P(x_i | \boldsymbol{x}_{-i}) = P\left(x_i \middle| \boldsymbol{x}_{ne(i)}\right).$$

In general the edges are defined so that the dimension of $\boldsymbol{x}_{ne(i)}$ is much smaller than the dimension of $\boldsymbol{x}_{-i}$, and the conditional distributions give a convenient and computationally simplified form of specifying the GMRF for Bayesian estimation using a Gibbs sampler MCMC.

The conditional specification of the GMRF was pioneered by Besag (1974) and is commonly referred to by the name conditional autoregression or CAR model. One of the most commonly used CAR in hierarchical models is the intrinsic conditional autoregression (ICAR) model (Besag & Kooperberg, 1995). It has been used extensively in disease mapping applications where the data follow a Poisson distribution. One reason for the use of these models in disease mapping is that standardized incidence ratios (SIR) may be unstable when estimating risks of rare diseases and/or diseases is small populations and spatial models enable the pooling of many SIR's thus overcoming the small populations (or small area) problem (Waller & Gotway, 2004, pp. 97). In particular the conditional distributions of the ICAR model are specified as:

$$x_i | \boldsymbol{x}_{-i} \sim Normal\left(\bar{x}_{ne(i)}, \frac{\sigma^2}{d_i}\right)$$

where $\bar{x}_{ne(i)}$ is the mean of the neighbor vector $\boldsymbol{x}_{ne(i)}$ and d_i is the number of neighbors or the dimension of $\boldsymbol{x}_{ne(i)}$. Thus the ICAR smooths an unstable SIR with an average of the neighboring regions' SIRs. The joint distribution implied by the conditional distribution of the ICAR is however improper because the covariance matrix is only semi-positive definite (singular matrix), but it can still be used in a hierarchical model because when combined with the data likelihood, it leads to a proper posterior. In particular, the ICAR's distribution is proper for a $n-1$ dimensional multivariate normal distribution, and so the interpretation is that the ICAR models the residuals. Another model commonly used in disease mapping applications known as the Besag-York- Mollie (BYM) is a convolution of the ICAR with independent random effects, where the ICAR models spatial clustering and the independent random effects model region wide heterogeneity (Besag, York, & Mollié, 1991).

A proper conditional autoregression (CAR) model can be specified when modeling areal data in a non-hierarchical model. A common specification for a proper CAR generalizes the ICAR model by introducing a parameter ρ, which is often restricted to $[0,1)$. Thus ω is a proper CAR if

$$\omega \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{Q}^{-1})$$

$$\boldsymbol{Q} = \boldsymbol{D} - \rho \boldsymbol{A} \qquad [4.2.1]$$

where $N(\quad)$ denotes the n-dimensional multivariate Normal distribution, $\boldsymbol{A}$ is the n by n adjacency matrix defined by the graph G, and $\boldsymbol{D}$ is the diagonal n by n matrix of row sums of $\boldsymbol{A}$ (Besag & Kooperberg, 1995). The restriction of ρ to $[0,1)$ is generally used because positive spatial correlation is more common in real-world applications. Recall positive spatial correlation is when data at close locations are more alike. Alternatively $\rho < 0$ implies data are negatively correlated, meaning dissimilar values are close together, which is rarely seen in practice. However, the restriction of ρ to $[0,1)$ is not necessary for Q to be nonsingular. In fact, it is sufficient that $\rho \in \left(\frac{1}{\min \boldsymbol{\lambda}}, \frac{1}{\max \boldsymbol{\lambda}}\right)$ where $\boldsymbol{\lambda}$ is the vector of eigenvalues for $\boldsymbol{D}^{-1}\boldsymbol{A}$ (Banerjee, Carlin, & Gelfand, 2004, p. 80).

4.3 Computational Efficiency Issues

In the analysis of spatial public health data, the distinction between the form of Gaussian process used, Matérn covariance function for case event data versus Gaussian Markov Random Field for case count data, comes in whether the intensity function is believed to be continuous or piecewise constant. This distinction is blurred however since a continuous function can be approximated by a piecewise constant function. For example, Li et al. (2012) note that the Besag-York-Mollié (BYM) model is a special case of the LGCP where the log-intensity is modeled as a piecewise constant function on the areal units. Li et al. (2012) further note that if the areal units are meaningful to the disease in question the BYM model would be a sufficient parameterization of the LGCP. However, when the areal units are unrelated to the disease in question, as is often the case, the assumption of the constant intensity within the arbitrary aggregation units may be poor. Furthermore, since spatial structure for the BYM model is dependent on a graph G defined by the boundaries around the arbitrary regions, the overall estimated spatial properties of the process may be poor (Wall 2004). It is for these reasons, Wall (2004) recommends the use of a spatial structure that is a function of distance, which would be the case with a LGCP having a Gaussian Process with Matérn covariance function.

However, when fitting a LGCP with a distance based covariance function over a large spatial extent, model estimation suffers “the Big-N problem” (Banerjee, Gelfand, Finley, & Sang, 2008), which is the computational bottleneck created from the covariance matrix inversion calculation that takes $O(n^3)$ flops, when n, representing the number of spatial locations, is large.

Diggle et al. (2013) recommends the use of a computational grid (regular lattice) in combination with spectral decomposition of the covariance function to speed estimation. Diggle et al. (2013) demonstrates the use of this computational grid to aggregate case event data and to change the aggregation of case count data using the data augmentation step proposed by Li et al. (2012) in a number of applications ranging from tree density estimation to a spatio-temporal model for gastro-enteric disease (food poisoning) risk prediction. Waagepetersen (2004) justifies the use of the regular lattice and shows that the approximate posterior expectations of the LGCP converge to exact posterior expectations when lattice cell sizes tend to zero. However after further study of the discretization resolution effects on the posterior expectations, Waagepetersen (2004) cautions that in practice a given discretization should be interpreted with care and researchers should test a few different grid resolutions. To determine an appropriate resolution in practice, Diggle et al. (2013) recommends the use of minimum contrast estimates to obtain a preliminary estimate of the spatial variability of the disease process and then to choose a resolution that is smaller than the disease process spatial variability but still large enough to be computationally feasible.

When the spatial extent is very large and the spatial variability very small, one may not be able to reach a sufficient compromise in the choice of a grid resolution. Decreasing the number of locations in the grid is not however the only method to deal with “the Big-N problem”. In fact the computational bottleneck of the distance based LGCP is more importantly related to the structure of covariance matrix of the GP. As long as the covariance matrix is sparse (many zero’s in off diagonal elements), parameter estimation can still be computational feasible even for a large number of locations.

One advantage of GMRF models over general geostatistical models is that the likelihood calculation does not require the inversion of the covariance matrix since the precision matrix (inverse of the covariance matrix) is modeled directly and because of the Markov Property, which informally states that a value at one location only depends on its neighbors and not all other locations, the precision matrix of the GMRF is sparse. Furthermore, Lindgren et al. (2011) show that the stochastic partial differential equation having a GP with the Matérn covariance function as an exact solution, has a finite element method weak solution which is equivalent to a GMRF. Since the Matérn covariance function is the most popular covariance function used for modeling GPs this is a very useful result. Thus, by using a regular lattice to aggregate disease cases in combination with a specification of the GMRF as given in Lindgren et al. (2011), one can have a faster estimation than if a distance based covariance function is used, particularly since INLA is a estimation method that can be used with these models (Rue et al., 2009), and the result will not give the undesirable properties demonstrated by Wall (2004).

4.4 Gaussian Markov Random Fields for Smooth Processes

Pettitt, Weir and Hart (2002) make use of the computational advantage in using GMRF models to define a proper CAR for irregularly spaced spatial data, the PWH CAR. The data $\boldsymbol{z} = (z_1, z_2, \dots, z_n)'$ observed at locations $\{s_1, s_2, \dots, s_n\}$ is said to be distributed PWH CAR if

$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{Q}^{-1})$$

$$\boldsymbol{Q} = \boldsymbol{I} + |\phi|\boldsymbol{D} - \phi\boldsymbol{A} = \begin{cases} \boldsymbol{I} - \phi(\boldsymbol{A} - \boldsymbol{D}), & \phi > 0, \\ \boldsymbol{I}, & \phi = 0, \\ \boldsymbol{I} - \phi(\boldsymbol{A} + \boldsymbol{D}), & \phi < 0, \end{cases}$$

where $N(\quad)$ denotes the n-dimensional multivariate Normal distribution, $\boldsymbol{I}$ is the n by n identity matrix, $\boldsymbol{D}$ is the diagonal n by n matrix of row sums of $\boldsymbol{A}$, and $\boldsymbol{A} = [\gamma_{ij}]$ is an n by n matrix. The elements of $\boldsymbol{A}$ are defined by

$$\gamma_{ij} = \begin{cases} \gamma\big(d_{ij}\big), & i \neq j, \\ 0, & i = j, \end{cases} \qquad [4.4.1]$$

where d_{ij} denotes the Euclidean distance between locations s_i and s_j and $\gamma\colon [0,\infty) \rightarrow [0,\infty)$ is continuous and non-increasing on $[0,\delta)$ and zero on $[\delta,\infty)$, where $\delta > 0$ (2002). Since $d_{ij} = d_{ji}$ and $\gamma(d_{ij}) \in \mathbb{R}$, the matrices $\boldsymbol{A}-\boldsymbol{D}$ and $\boldsymbol{A}+\boldsymbol{D}$ have an eigenvalue decomposition. Let the eigenvalue decomposition of $\boldsymbol{A}-\boldsymbol{D}$ be

$$\boldsymbol{A}-\boldsymbol{D} = \boldsymbol{E}\boldsymbol{\Lambda}\mathbf{E}'$$

and then

$$\boldsymbol{I}-\phi(\boldsymbol{A}-\boldsymbol{D}) = \boldsymbol{I}-\phi\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' = \boldsymbol{E}(\boldsymbol{I}-\phi\boldsymbol{\Lambda})\boldsymbol{E}'$$

and thus the PWH CAR not only avoids the covariance matrix inverse calculation by directly specifying the precision matrix, but also simplifies the calculation of the covariance matrix's determinant, which is also needed for the likelihood calculation. This is because the determinant of Q is the product of its eigenvalues, $\boldsymbol{\zeta}$, which can be calculated as

$$\zeta_i = \begin{cases} 1-\phi\lambda_i, & \phi > 0, \\ 1, & \phi = 0, \\ 1-\phi v_i, & \phi < 0, \end{cases}$$

where $\boldsymbol{\lambda}$ are the eigenvalues of $\boldsymbol{A}-\boldsymbol{D}$ and $\boldsymbol{v}$ are the eigenvalues of $\boldsymbol{A}-\boldsymbol{D}$.

The PWH parameterized CAR can be further adjusted for computational efficiency through the use of a pre-whitening transformation as defined in Linder (2001), where the PWH CAR is used to model average March 1990 air temperatures measured at irregularly spaced stations located in the Beaufort region of the pan-arctic and then to spatially interpolate these temperatures to a fine grid over the entire region. When data are assumed to exhibit positive spatial autocorrelation, the precision matrix of PWH CAR simplifies to:

$$\boldsymbol{Q} = \boldsymbol{I}-\phi(\boldsymbol{A}-\boldsymbol{D})$$

where $\phi \in [0, \infty)$, which can be alternatively represented by its eigenvalue decomposition.

$$\boldsymbol{Q} = \boldsymbol{E}(\boldsymbol{I} - \phi\boldsymbol{\Lambda})\boldsymbol{E}'$$

The eigenvectors, columns of $\boldsymbol{E},$ contain the spatial structure of the data, i.e. represent the data's patterns of spatial clustering at different scales (Paciorek 2013; Hughes & Haran, 2013), and are not dependent on the parameter ϕ. Thus applying a linear transformation of the data $\boldsymbol{z} = (z_1, z_2, \dots, z_n)'$ defined by $\boldsymbol{E}'$ removes the spatial structure

$$\boldsymbol{E}'\boldsymbol{z} \sim N(\boldsymbol{E}'\boldsymbol{\mu}, \sigma^2\boldsymbol{E}'\boldsymbol{Q}^{-1}\boldsymbol{E}\,)$$

$$\boldsymbol{E}'\boldsymbol{Q}\boldsymbol{E} = \boldsymbol{E}'\boldsymbol{E}(\boldsymbol{I} - \phi\boldsymbol{\Lambda})\boldsymbol{E}'\boldsymbol{E} = (\boldsymbol{I} - \phi\boldsymbol{\Lambda})$$

$$\boldsymbol{E}'\boldsymbol{z} \sim N(\boldsymbol{E}'\boldsymbol{\mu}, \sigma^2(\boldsymbol{I} - \phi\boldsymbol{\Lambda})^{-1}\,)$$

and allows for fast estimation of the model parameters (Linder 2001).

Although this pre-whitening transformation significantly speeds the estimation of the PWH model parameters, the initial calculation of the eigenvalue decomposition for $\boldsymbol{Q}$ would require significant memory resources and calculation time for a very large number of locations. However, when the neighbor structure of the data can be mapped to a torus, Q will be circulant and the Fast Fourier Transform can be used to remove the spatial structure from the data (Rue & Held, 2005; Hupper 2005; Yuan 2011). Techniques for mapping a neighborhood structure from an irregular region to a torus known as circulant embedding are beyond the scope of this dissertation but the interested reader can find more details in (Wood & Chan, 1994), (Stroud, Stein, & Lysen, 2016), and (Guinness, Fuentes, Hesterberg, & Polizzotto, 2014).

Despite the computational advantages of the PWH CAR that can be used for both geostatistical and areal data, it has been criticized for being too simplistic for use in many geostatistical data applications and for being restrictive (Lindgren et al., 2011; Paciorek 2013; Czado & Prokopenko, 2008). In particular, Paciorek (2013) shows that distance based CARs such as the PWH CAR poorly represent very smooth spatial processes in comparison to the second-order CAR defined by Rue and Held (2005, pp. 114-116) despite the inclusion of more neighbors at farther distances. Furthermore, Czado and Prokopenko (2008) argue that the PWH CAR is restrictive because as $\phi \rightarrow \infty$, the conditional variance of $z_i|z_{-i}$ approaches 0. Alternatively, when the precision matrix of the PWH CAR is modified to:

$$\boldsymbol{Q} = \boldsymbol{I} + |\psi|(\boldsymbol{D} - \boldsymbol{I}) - \psi\boldsymbol{A}$$

where $\psi = \frac{\phi}{1+\phi} \in (-1,1)$, the conditional variance of $z_i|z_{-i}$ approaches $\frac{\sigma^2}{d_i}$ as $\phi \rightarrow \infty$ (2008). Note that is the same conditional variance implied by the ICAR.

Czado and Prokopenko (2008) use the modified PWH CAR in a hierarchical model with data layer defined by a binary distribution. However, a similar model known as the Leroux CAR (Lee, 2011) has been used when the data layer is defined by the Poisson distribution in the disease mapping literature. In this case, $\psi \in [0,1)$ and the precision matrix becomes:

$$\boldsymbol{Q} = (1 - \psi)\boldsymbol{I} + \psi(\boldsymbol{D} - \boldsymbol{A})$$

Lee (2011) argues that the Leroux CAR is better than the BYM since it only uses one set of random effects but still allows for a combination of spatial dependence implied by the ICAR and spatial independence.

Both the inability to model smooth processes and the restrictive conditional variance of the PWH CAR can be remedied in one parsimonious two-parameter model proposed by Hupper (2005) and improved by Yuan (2011), called the Extended Autoregression (EAR). The data $\boldsymbol{z} = (z_1, z_2, \dots, z_n)'$ observed at locations $\{s_1, s_2, \dots, s_n\}$ are said to have an EAR distribution if

$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{Q}^{-1}) \qquad [4.4.2]$$

$$\boldsymbol{Q} = [\boldsymbol{I} - \psi(\boldsymbol{A} - \boldsymbol{D} + \boldsymbol{I})]^{\theta}$$

where $\boldsymbol{A}$, $\boldsymbol{D}$, and $\boldsymbol{I}$ are defined as in the PWH CAR, and ψ is defined as in the modified PWH CAR (Yuan, 2011).

Additionally, the modified PWH CAR (Czado and Prokopenko, 2008) is a special case of the EAR when $\theta = 1$, thus the EAR retains the computational advantages available in the PWH CAR, such as the computational efficient form for the determinant of the precision matrix and the pre-whitening transformation defined in Linder (2001). Furthermore, since the modified PWH CAR (2008) is equivalent to the Leroux CAR (Lee, 2011), the Leroux CAR is also a special case of the EAR when $\theta = 1$. Thus, the EAR's precision matrix can also be thought of as a linear combination of spatial independence and spatial dependence as defined by the ICAR.

Hupper (2005) showed the EAR model when defined for data that are observed on a regular lattice where $\delta = 1$ and $\gamma(d_{ij}) = 1$ is useful for modeling smooth processes. In particular, Hupper used the EAR to model spatially varying temporal trend components in a hierarchical model for an application of gridded river runoff data from the pan-Artic region. The principal goal of her analysis was to determine locations within the larger region that had significant changes in river runoff over time, while also accounting for spatial correlations between nearby locations that may exhibit extreme smoothness. When compared with a by pixel simple linear regression approach and a hierarchical model with intercept and trend components specified with a PWH CAR prior, the hierarchical model with EAR specified intercept and trend components estimated more regions with a significantly positive or negative change.

Yuan (2011) further investigates the properties of the EAR when defined for data that are observed on a regular lattice where $\delta = 1$ and $\gamma(d_{ij}) = 1$. Yuan (2011, p. 93) showed that the EAR model is equivalent to the higher-order CAR given by Rue and Held (2005, pp. 114-116) when defined on a torus. Yuan (2011, pp. 95-96) further showed that there is an identifiability issue between the EAR smoothing parameter θ and the EAR interaction parameter ψ, which is similar to the identifiability issue found in geostatistical models using a Matérn class covariance between the range κ and smoothing parameter ν (Zhang, 2004). In order to deal with this identifiability issue, Yuan (2011) proposes a new EAR called the intrinsic EAR (IEAR) model, similar to ICAR (Besag & Kooperberg, 1995), which is an EAR model with $\psi = 1$. The precision matrix for the IEAR is then as follows:

$$\boldsymbol{Q} = [\boldsymbol{D} - \boldsymbol{A}]^{\theta}$$

where $\boldsymbol{A}$, and $\boldsymbol{D}$, are defined as in the EAR (Yuan, 2011). Like the ICAR, the IEAR is an improper distribution, but proper in $n - 1$ dimensions, and can only be used as a latent process in a hierarchical model.

Yuan (2011, pp. 96) notes that in the case of space-time data such as in the application considered by Hupper (2005) the repeated spatial measurements at different times would be sufficient to resolve the identifiability issue between the smoothing and interaction parameters of the EAR. However, when the EAR model is used as a latent process for spatial random effects in addition to other regression parameters, the spatial random effects are collinear with the regression parameters, and inference on the regression parameters may be unreliable. This follows from an argument given by Reich et al. (2006) who showed that when ICAR spatial random effects are used in additional to regression variables, the spatial random effects will be collinear with the regression variables. Consider the case of the intensity function in the process layer of a LGCP used to model disease case counts:

$$\log(\boldsymbol{\Lambda}) = \boldsymbol{X\beta} + \boldsymbol{\omega} + \log(\boldsymbol{E})$$

Furthermore recall that the precision matrix in the EAR model can be written as the eigenvalue decomposition:

$$\boldsymbol{Q} = \boldsymbol{E}(\boldsymbol{I} - \psi\boldsymbol{\Lambda})^{\theta}\boldsymbol{E}'$$

where

$$\boldsymbol{E}\Lambda\mathrm{E}' = \boldsymbol{A} - \boldsymbol{D} + \boldsymbol{I}$$

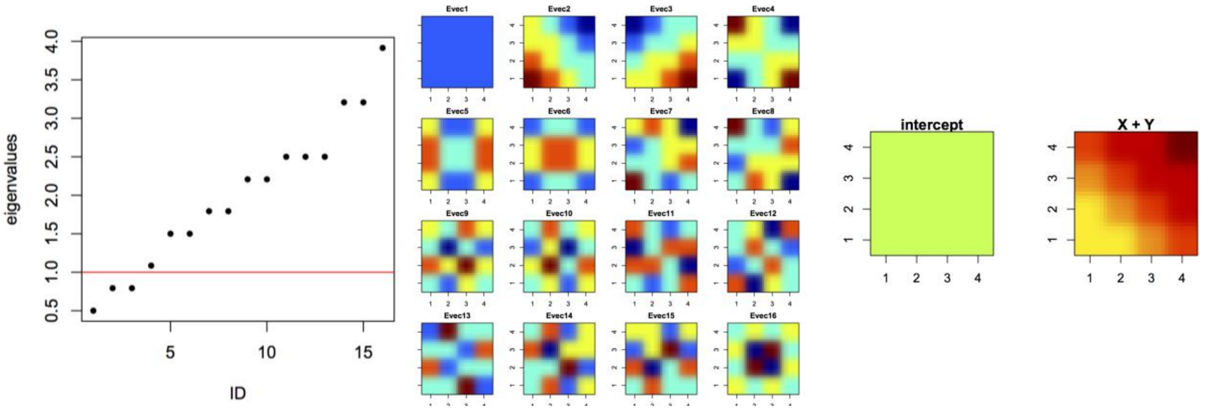
The process layer can then be written as

$$\log(\boldsymbol{\Lambda}) = \boldsymbol{X\beta} + \boldsymbol{E}(\boldsymbol{E}'\boldsymbol{\omega}) + \log(\boldsymbol{E})$$

where $\boldsymbol{E}$ can be thought of as the design matrix for the transformed random effects $\boldsymbol{E}'\boldsymbol{\omega}$. Then since the columns of design matrix $\boldsymbol{E}$ are n orthogonal eigenvectors that span the space of $\mathbb{R}^n$, the columns of $\boldsymbol{X}$ are linear combinations of the columns of $\boldsymbol{E}$, i.e. $\boldsymbol{X}$ and $\boldsymbol{E}$ are collinear.

The issue of collinearity is a problem (as for non-spatial regression) when the goal of the study is scientific explanation because it causes variance inflation that may make important predictors appear insignificant. The ICAR and EAR model are not the only models with this disadvantage, as Hughes (2015) remarks that the argument used by Reich et al. (2006) could be used to show that any spatial random effects, ones with a proper CAR distribution or even a GP with distance based covariance function, are collinear with regression variables. Reich et al. (2006) note that the variance inflation will be particularly problematic in the Bayesian framework context if there is little smoothing from the prior distribution of the variance parameter σ^2 given in [4.4.2] and when columns of $\boldsymbol{X}$ are nearly collinear with the columns of $\boldsymbol{E}$ associated with the smallest eigenvalues $(\boldsymbol{I} - \phi\boldsymbol{\Lambda})^{\theta}$.

As an example, consider the relationship between an EAR process defined on the 4 by 4 regular square lattice with $\delta = 1$ and $\gamma\left(d_{ij}\right) = 1$ and the common set of spatial regression covariates including an intercept and the positive diagonal gradient. The eigenvalues of $\boldsymbol{Q} = \boldsymbol{E}(\boldsymbol{I} - \psi\boldsymbol{\Lambda})^{\theta}\boldsymbol{E}'$ when $\theta = 1$ and $\psi = 0.5$ range from 0.5 to about 4 (Figure IV-1) with three eigenvalues less than 1. The corresponding eigenvectors for these three smallest eigenvalues show the same spatial patterns as the spatial regression covariates. Thus variance inflation would be a problem here if the spatial regression covariates were of particular scientific interest.



Eigenvalues of Q Eigenvectors of Q Spatial Regression Covariates

Figure IV-1. Eigenvalues and Eigenvectors of EAR's precision matrix Q compared with Spatial Regression Covariates

As an aside, this example also helps to illustrate why the EAR process can model smooth processes. The smallest eigenvalues of Q (the precision) are the largest values of the variance-covariance matrix, since these are inversely related. In particular, as the smoothing parameter $\theta \to \infty$, the eigenvalues of Q less than 1 will approach 0, which means in the variance they will get very large, while the eigenvalues of Q greater than 1 will be diminished in the variance. Furthermore, the eigenvectors associated with the smallest eigenvalues of Q represent large-scale spatial variability, so as their corresponding eigenvalues become larger these will dominate the spatial correlation structure. This can be seen in the simulated mean 0 EAR process defined on the 10 by 10 regular square lattice with $\delta = 1$ and $\gamma(d_{ij}) = 1$ (Figure IV-2).

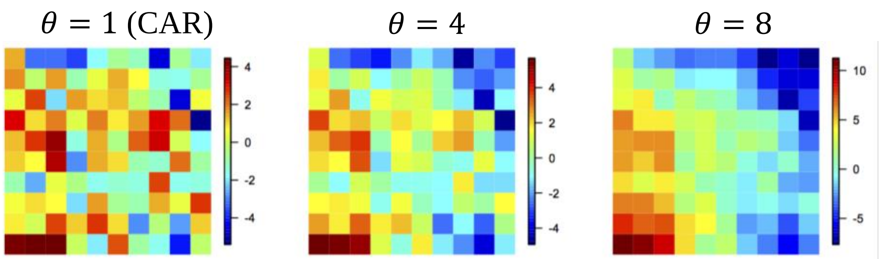

$\theta = 1$ (CAR)
$\theta = 4$
$\theta = 8$

Figure IV-2. Simulated EAR process with mean 0 and $\psi = 0.5$ *for different values of* θ

Given the collinearity between regression covariates and spatial random effects, one may simply prefer to use the regression covariates without the random effects when a scientific explanation of the regression parameters is of interest, but then regression parameters will be biased due to the spatial correlation in the data. In the case of the ICAR random effects, Reich et al. (2006) offers a solution that transforms the ICAR random effects in such a way that the random effects will smooth orthogonal to the regression covariates. Hughes and Haran (2013) improve on this transformation of the ICAR random effects to account for the structure of the underlying graph G and further reduce the dimensions of the random effects to improve computational efficiency using MCMC. Hughes (2015) notes that the use of the augmented linear predictor (transformation) in the case of the ICAR random effects cannot easily be used in an efficiently computationally way when a proper CAR (Besag & Kooperberg, 1995) is used. This is because in the proper CAR as defined by [4.2.1] the eigenvectors and not just the eigenvalues of the precision matrix depend on a parameter ρ. However, in the case of the PWH CAR (Pettitt et al., 2002), the modified PWH CAR (Czado & Prokopenko, 2008), and the EAR model (Hupper 2005; Yuan 2011), the eigenvectors do not depend on any of the precision matrix parameters, and thus the methods described in Reich et al. (2006) and Hughes and Haran (2013) could likely be extended to these models quite easily.

V SUMMARIZING DATA TO COMPUTATIONAL GRIDS

5.1 COMPUTATIONAL GRIDS FOR ESTIMATING ALS RISK

Following Diggle et al. (2013) a regular lattice (computational grid) is used to summarize ALS case counts across the NNE region. It is likely this aggregation scale will provide a better approximation of a believed continuous ALS risk function than the aggregations scales used by Caller et al. (2013), the census block group, and by Torbick et al. (2014), the census track, since the census unit aggregations are large in rural areas where there may be within census unit varying environmental heterogeneity due to the varying locations of lakes. A caveat in using the regular lattice computational grid is that properties of the background population used in the deterministic component of the spatially-varying intensity function are generally only available at census aggregation units (Diggle et al., 2013) that are spatially misaligned with the chosen computational grid. Diggle et al. (2013) notes that this can be somewhat avoided by using a modeled population product such as the Socioeconomic Data and Applications Center (SEDAC) available on 1km and 5km grids (Seirup & Yetman, 2006).

In this study, two gridded population products at a 1km resolution representing the region's population at the year 2000, one provided by the Socioeconomic Data and Applications Center (SEDAC) (2006) and the other a product of OakRidge National Laboratory (ORNL) called LandScan (LandScan 2000) are used to calculate the background population component of the LGCP's intensity function and are compared for model sensitivity due to small differences in the backgrounds. However, since a 1km resolution would require a 221,232 by 221,232 adjacency matrix (the neighbor structure matrix) to be saved in memory, which was not possible on the 2.4 GHz Intel Core i5 with 16GB of memory used for the analysis, two coarser resolutions, one at a 4km resolution and another at a 8km resolution, are chosen, which attempt to balance computational efficiency with disease process spatial variability and confidentially concerns.

The choice of the 4km resolution was based on crude preliminary values of Exponential covariance function parameters estimated (scale and variance) via minimum contrast following the recommendation of Diggle et al. (2013). Minimum contrast estimates parameters by minimizing the squared discrepancy between the assumed parametric form of the point process's second-order characteristics and a nonparametric estimate (Taylor, Davies, Rowlingson, & Diggle, 2015). Minimum contrast estimation was carried out using the function *minimum.contrast()* in the R package lgcp (Taylor et al., 2015). This function requires coordinate locations of case event data specified for a given border region. Longitude/latitude coordinates of the 764 cases across all NNE were re-projected so that distances between points would be calculated in meters and U.S. Census Tiger shapefiles of the states of Vermont, New Hampshire and Maine were combined to define the border region (U.S. Census Bureau, 2013). Furthermore, the function allows for a deterministic background intensity and for a transformation function to be applied to function in the contrast, and requires a choice of non-parametric estimator based on either the pair correlation function g or Ripley's K function (K) (Taylor et al., 2015). Estimates of the scale parameter (in km) are given in the last column of Table V-1 based on the two different non-parametric estimators, a deterministic background derived from a smoothed version of the ALS case locations, and a logarithm transformation. The deterministic background of smoothed ALS case locations used the R function *density.ppp()* from the R package spatstat, in which Diggle’s edge correction can be used (Table V-1 column diggle) and a standard deviation of the isotropic Gaussian kernel density can be specified (Table V-1 column Dens sigma) (Baddeley et al., 2015).

Table V-1. Minimum contrast estimates of the scale and variance parameters for the exponential covariance

	Non-parametric estimator	Dens sigma	diggle	scale	variance	Squared_discrepancy	scaleKM
1	*g*	NA	F	5756.645	3.2236	8589.15	5.7566
2	*K*	NA	F	4393.622	3.9716	2239.08	4.3936
3	*g*	NA	T	5460.953	3.3737	10298.32	5.4610
4	*K*	NA	T	4084.108	4.2457	2655.43	4.0841
5	*g*	10000	T	906.600	2.8840	16885.42	0.9066
6	*K*	10000	T	33.792	13.4065	28706.61	0.0338
7	*g*	5000	T	4048.825	0.0000	33755.66	4.0488
8	*K*	5000	T	4048.825	0.0000	94667.43	4.0488
9	*g*	30000	T	2563.726	3.9791	11518.09	2.5637
10	*K*	30000	T	1543.303	5.7505	7795.49	1.5433

Values of the scale parameter in this investigation mostly ranged from about 1km to 5km, and 4km resolutions had the lowest squared discrepancy.

The 8km resolution was chosen since it is double that of the 4km resolution and because it may better represent some of the spatial uncertainty in the case locations, particularly the 9 cases that have only a town name location. These towns ranged in area from 6.2km^2 to 157.7km^2 with a median square area of 66.3km^2.

Following the definition of the 4 and 8km resolution computational grids, the three boundary regions, NNE, VTNH and VTNH_rmcty (Figure II-1), where used to crop the grids in order to investigate the effect of low case ascertainment assumed for all of Maine, parts of southern New Hampshire and southwestern Vermont. Water quality metrics were matched deterministically to the computational grids using several scales.

5.1.1 NNE 8km Computational Grid

In a first analysis, the boundary area is set to the entire NNE region and the relationship between ALS risk and a location's closest lake average water quality metrics is explored. In this analysis, the two suspected duplicate cases are not removed. Thus there are a total of 764 cases that are first rasterized to counts on the 1km SEDAC lattice, and then aggregated (counts summed) to the 8km resolution computational grid.

Expected counts for this analysis were based on the SEDAC background population estimates only. The general procedure for calculating the expected counts was as outlined in Chapter II section 2.2. However, Age/sex specific rates included two more cases in total than was shown in Table II-2, and the original Census 2000 block age/sex totals were used as denominators instead of the SEDAC estimate totals. These counts are summarized in Table V-2. Furthermore, in the calculation of SEDAC estimated age/sex specific population counts by 1km lattice cells, census block polygons were directly rasterized to the 1km lattice. It is expected this procedure resulted in more uncertainty in densely populated areas than the method described in Chapter II section 2.2 that first rasterized polygons to a 20m lattice.

Table V-2. Age/sex class case and population counts

	Age/sex classes	ALS cases	Census 2000
1	**F & 0-44**	34.71	971469
2	**F & 45-54**	74.38	236326
3	**F & 55-64**	78.09	146905
4	**F & 65-74**	74.38	115673
5	**F & 75-84**	48.34	85539
6	**F & 85+**	*	37291
7	**M & 0-44**	74.62	979005
8	**M & 45-54**	114.90	234088
9	**M & 55-64**	93.57	142861
10	**M & 65-74**	105.42	99533
11	**M & 75-84**	48.56	56594
12	**M & 85+**	*	14252
Total:		764	3119536

*Values less than 10 are suppressed for privacy reasons.

In order to account for the spatial misalignment between the lake polygons used for summarizing average Chl-a, SD and TN, and the 8km resolution computational grid, a distance and direction (azimuth: degrees of angle with respect to North) are calculated from case to nearest lake point using the R package geosphere (Hijmans, 2015). Then the lake average SD, TN and Chl-a at the nearest lake, and distance and direction to closest lake point are averaged for all cases within the same 8km grid cell. Lake Champlain was excluded from the possible closest lake points considered lake average values of Chl-a, SD and TN were not calculated for this lake. When more than one case occur in the same 8km grid cell, there is the potential that the relationship between ALS risk and water quality at this 8km resolution is opposite from their relationship at the individual case scale. Table V-3 summarizes the frequencies of lattice cells by case count.

Table V-3. Lattice cell frequency by ALS case count

ALS case count	Lattice cell count
0	3205
1	312
2	64
3	34
4	18
5	7
6 – 7	7
8 - 13	7
Total	3654

For cells containing 0 cases, the centroid is used to calculate distance and direction to closest lake point, and SD, TN and Chl-a values for these cells are the lake average for the lake closest to the cell centroid. The average distance to nearest lake for each cell is then log-transformed and the direction for each cell is transformed to a sine and cosine component. An illustration of the effect of averaging distances and directions is shown in Figure V-1.

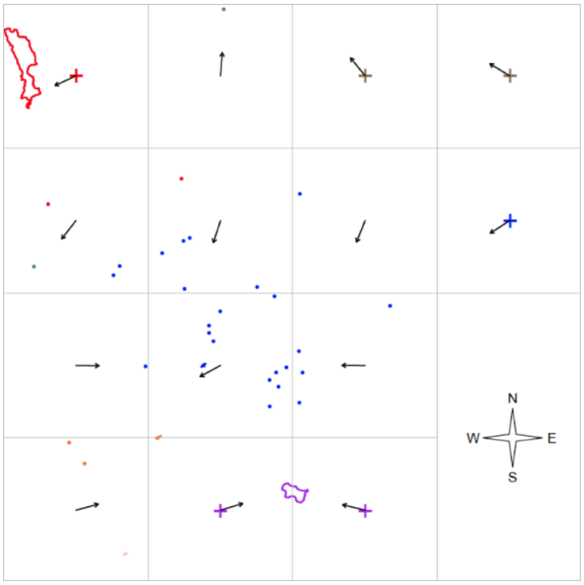


Figure V-1. Illustration of direction aggregation effect (colors red, brown, green, blue, orange, pink purple represent locations and their nearest lake)

The illustration includes 5 visible lake polygons colored red, pink, orange, blue and purple, as well as 2 lake polygons colored green and brown that fall outside the image extent. Locations closest to a given lake appear in the same color as the lake polygon. Case locations are represented by dots and centroid locations are represented by crosses and by the origin of the small black arrows. Crosses appear to indicate cell centroids when the cell has a 0 case count. The small black arrows point in the direction of the closest lake based on the aggregation scheme, average of all case directions within cell to nearest lake or direction of cell centroid to nearest lake. In several cases, the averaged direction agrees with the direction of the individual locations. For example, in the third column of the third row, the closest lake (blue) of all the individual locations are to the west of the blue lake. However, in some cases such the first column of the second row, individual locations are closest to different lakes in different directions. Thus the interpretation of the southwest-pointing arrow does not necessarily agree with the directions of the closest lakes at the individual level. In regions with many lakes close in distance, the effect of the averaged direction to closest lake will likely not be a good approximation of individual level locations.

When the closest lake to a given grid cell centroid with no cases had a least one metric, Chl-a, SD and TN deemed as an outlier, the grid cell was removed from further analysis. This procedure removed 13 cells giving a final count of 3641 lattice locations covering most of NNE, which is shown in Figure V-2.

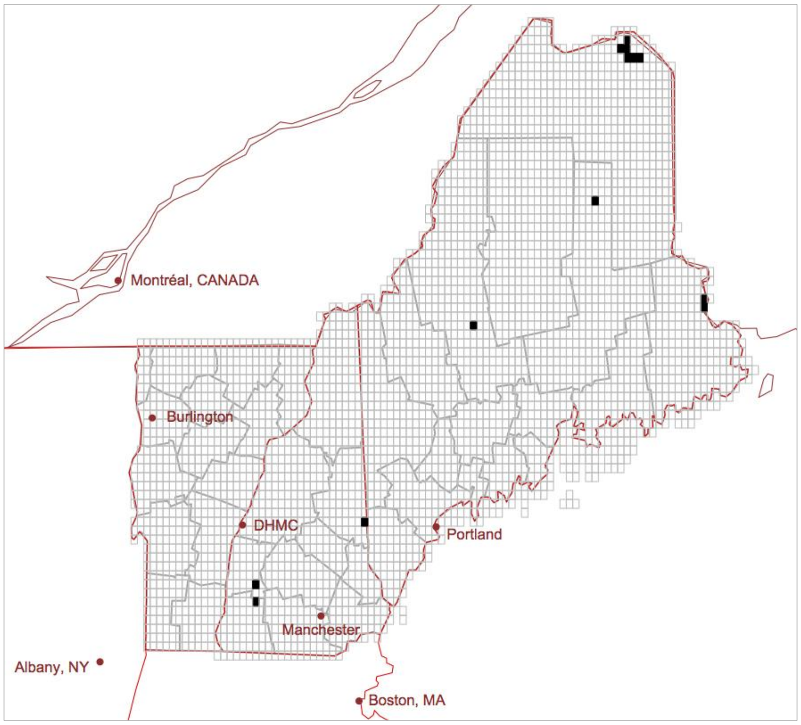


Figure V-2. NNE region overlaid with 8km resolution lattice (black cells are 13 excluded cells)

5.1.2 VTNH and VTNH_rmcty 4 and 8km Computational Grids

For the analysis of the relationship between ALS risk and water quality within the boundary areas of VTNH and VTNH_rmcty, both the 4km and 8km computational grids are compared as well as two sets of expected counts, one based on the SEDAC background population estimates and the other on the Landscan 2000 background population estimates. The general procedure for calculating the expected counts was as outlined in Chapter II section 2.2. Unlike in the 8km NNE analysis, age/sex specific rate denominators were based on the SEDAC and Landscan population backgrounds for the specific boundary areas. Also, unlike in the 8km NNE analysis, census block polygons were rasterized first to the 20m lattice and then aggregated to the 1km lattice for the calculation of age/sex specific population counts by 1km lattice cells. Furthermore, the two suspected duplicate cases are removed. Thus there are a total of 762 cases that are first rasterized to counts on the 1km SEDAC lattice, and then aggregated (counts summed) to a 4km and 8km lattice, along with the expected counts.

Water quality metrics for the VTNH and VTNH_rmcty regions include various spatial aggregation scales of the PC metric and Lake TSI metric at 4 temporal resolution scales. Recall that PC metric is available at a 30m pixel resolution for surface water in lakes sized 6 hectares or more, while the Lake TSI metric is a value available at the lake scale for all lakes sized 6 hectares or more except that Lake Champlain was excluded. The spatial aggregation scales used to summarize water quality exposure based on the PC metric and the 4 temporal Lake TSI metrics are described in Table V-4.

Table V-4. Aggregation Scales used to summarize PC in ug/L & Lake TSI

<table>
<tr><th>Water Quality Metric</th><th>Aggregation scale name</th><th colspan="2">Description</th></tr>
<tr><td rowspan="8"><b>PC</b></td><td>huc10_max_pc</td><td colspan="2">Maximum of lake averages within the Hydrological Unit Code 12 (HUC10) boundary assigned to 4(8)km lattice cells if cell centroid also falls within HUC10 boundary</td></tr>
<tr><td>huc10_mean_pc</td><td colspan="2">Mean of lake averages within the HUC10 boundary assigned to 4(8)km lattice cells if cell centroid also falls within HUC10 boundary</td></tr>
<tr><td>huc12_max_pc</td><td colspan="2">Maximum of lake averages within the Hydrological Unit Code 12 (HUC12) boundary assigned to 4(8)km lattice cells if cell centroid also falls within HUC12 boundary</td></tr>
<tr><td>huc12_mean_pc</td><td colspan="2">Mean of lake averages within the HUC12 boundary assigned to 4(8)km lattice cells if cell centroid also falls within HUC12 boundary</td></tr>
<tr><td>pc4kmAVG</td><td colspan="2">Mean PC of all 30m lake pixels within a 4km radius of 4(8) km lattice cell, when no lake intersects this radius a value of 0 is assigned.</td></tr>
<tr><td>pc8kmAVG</td><td colspan="2">Mean PC of all 30m lake pixels within an 8km radius of 4(8) km lattice cell, when no lake intersects this radius a value of 0 is assigned.</td></tr>
<tr><td>pc10kmAVG</td><td colspan="2">Mean PC of all 30m lake pixels within a 10km radius of 4(8) km lattice cell, when no lake intersects this radius a value of 0 is assigned.</td></tr>
<tr><td>pc_idw6</td><td colspan="2">Inverse distance weighted mean PC of 30m lake pixel centroids to each 4(8)km lattice cell centroid, where 30m cell centroids greater than a distance of 50km were not included. Weights were $\frac{1}{distance^6}$.</td></tr>
<tr><td rowspan="12"><b>TSI</b></td><td>lakeTSI_2000_4kmAVG</td><td>2000-2001 time period</td><td rowspan="4">Area-weighted mean of lake TSI metric for each lake intersecting a 4km radius from the lattice cell centroid, when no lake intersects a lattice cell a value of 0 is assigned.</td></tr>
<tr><td>lakeTSI_2002_4kmAVG</td><td>2002-2003 time period</td></tr>
<tr><td>lakeTSI_2004_4kmAVG</td><td>2004-2005 time period</td></tr>
<tr><td>lakeTSI_alltime_4kmAVG</td><td>2000-2005 time period</td></tr>
<tr><td>lakeTSI_2000_8kmAVG</td><td>2000-2001 time period</td><td rowspan="4">Area-weighted mean of lake TSI metric for each lake intersecting an 8km radius from the lattice cell centroid, when no lake intersects a lattice cell a value of 0 is assigned.</td></tr>
<tr><td>lakeTSI_2002_8kmAVG</td><td>2002-2003 time period</td></tr>
<tr><td>lakeTSI_2004_8kmAVG</td><td>2004-2005 time period</td></tr>
<tr><td>lakeTSI_alltime_8kmAVG</td><td>2000-2005 time period</td></tr>
<tr><td>tsi2000_idw6</td><td>2000-2001 time period</td><td rowspan="4">Inverse distance weighted mean Lake TSI metric of 30m lake pixel centroids to each 4(8)km lattice cell centroid, where 30m cell centroids greater than a distance of 50km were not included. Weights were $\frac{1}{distance^6}$.</td></tr>
<tr><td>tsi2002_idw6</td><td>2002-2003 time period</td></tr>
<tr><td>tsi2004_idw6</td><td>2004-2005 time period</td></tr>
<tr><td>tsialltime_idw6</td><td>2000-2005 time period</td></tr>
</table>

5.2 Computational grids for Lake Champlain Surface Water Temperatures

A spatial varying coefficient model similar to the one used by Hupper (2005) is considered in the analysis of surface water temperatures of Lake Champlain, which were not used explicitly in the models relating ALS disease risk to the area's water quality but provided as a case study of the spatio-temporal dynamics of how the water quality may be varying during the time period (1997-2009) under study of ALS risk. This application has a similar goal as the one discussed in Hupper (2005) in that there is interest in knowing if temperatures have significantly increased and if these increases are spatially varying within the lake. There is also interest in knowing if the entire study period (1984-2011) average temperatures significantly varies within the lake, since these variations might correlate well with cyanobacteria blooms and could be used as a cyanobacteria exposure proxy metric.

This application introduces new challenges not present in the application considered by Hupper (2005). In particular, the original resolution of the data is too large (over 100,000 locations) for the calculation of the precision matrix's eigenvalue decomposition when using a 2.4 GHz Intel Core i5 with 16GB of memory and a Fast Fourier Transform cannot be used in its place because the long narrow shape of Lake Champlain is not conducive to circulant embedding. Thus, following the approach Diggle et al. (2013) uses in the estimation of the LGCP to deal with computational challenges, a computational grid will be defined for the Lake Champlain temperature data. Since the estimation would likely depend on the resolution of the computational grid, two resolutions are compared, one at a 1080-meter resolution and another at a 2160-meter resolution. . Furthermore, missing data due to cloud cover could bias trend estimates. To account for this missing data, missing pixels are coded as 0 which is much smaller than expected summer Celsius water temperatures and then a regression variable will be introduced that gives the percent of missing pixels within a given computational grid cell that is at a much larger user defined resolution.

In order to choose the resolutions of the computational grid, a similar logic as Diggle et al. (2013) suggests for LGCP is applied. First a preliminary estimate of the spatial variability of the trend process and the average temperature process are estimated and then a computational grid resolution is selected that is smaller than the estimated variabilities yet large enough to be computationally efficient. In order to get a preliminary estimate of the spatial variability, by pixel least square regressions were estimated defined by:

$$\boldsymbol{y}_i = \alpha_i + \beta_i \boldsymbol{t}_i$$

where $\boldsymbol{y}_i$ is a vector of temperatures and $\boldsymbol{t}_i$ is the corresponding vector of time points (in years) when the temperatures were measured for a given pixel i. Then variograms were calculated separately for the intercept parameters α_i and the slope parameters β_i for each pixel. Recall the variogram compares the pairwise distances between points (in this case pixel centroids) to the values of the least squares estimated intercept and slope parameters. Plots of the semivariance for each least squares estimated parameter are show in *Figure V-3*. After about 15km the intercept parameter values reach a sill and after about 5km the slope parameters reach a sill, which means a grid no larger than 5km can be used. Both the 1080m resolution grid and the 2160m resolution grid are less than 5km in resolution.

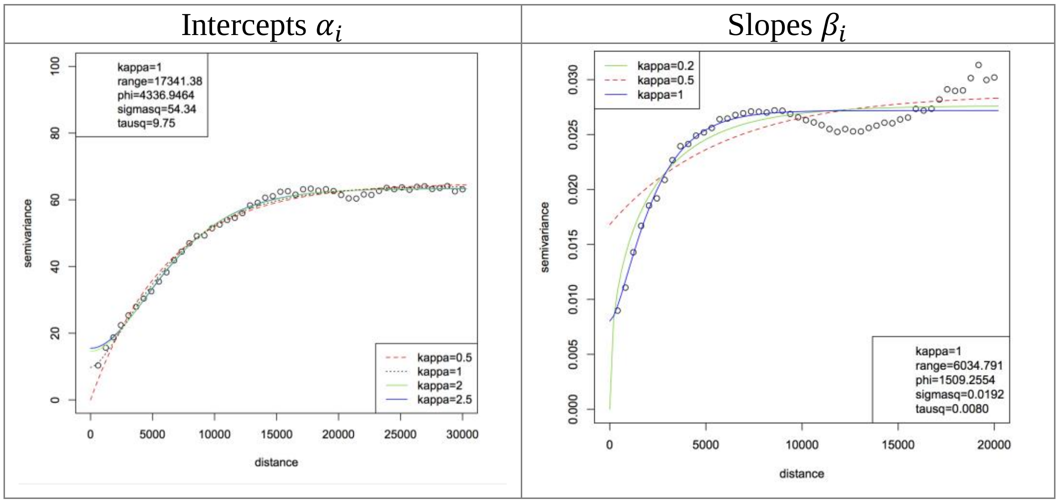


Figure V-3. Semivariance plots for each least squares estimated parameter

The 1080m resolution grid includes 1217 cell locations covering Lake Champlain and the 2160m resolution grid includes 352 cell locations covering Lake Champlain. These grids are show in Table V-4.

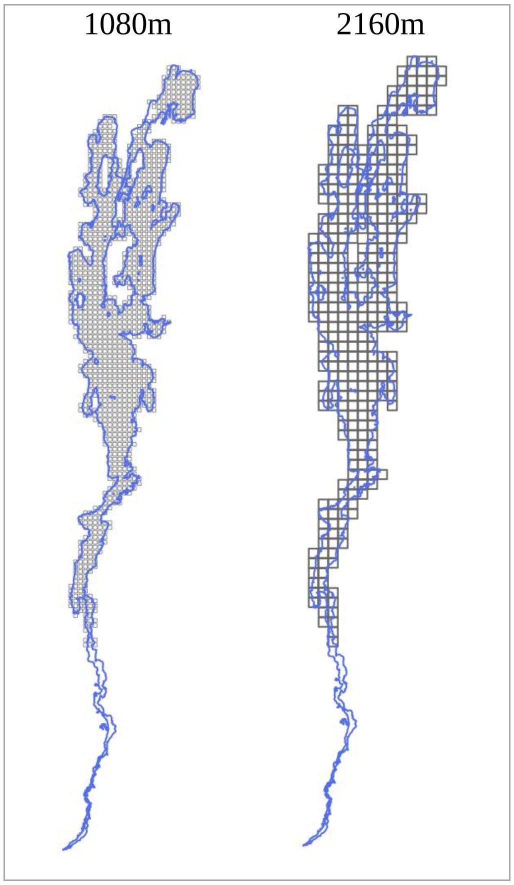


Figure V-4. Computational Grids used to model remotely sensed surface water temperatures of Lake Champlain

VI SPATIAL DISTRIBUTION OF ALS

6.1 GOALS & ORGANIZATION

An analysis of the spatial distribution of ALS risk was done with the goal of assessing the sensitivity of this distribution to the fixed components of the modeling framework as well as to provide a holistic comparison with the cluster analyses done by Caller et al. (2013) and Torbick et al. (2014) which were based on Local Moran's I statistics (Waller & Gotway, 2004, pp. 236-240). The components assessed for sensitivity included the boundary regions, NNE, VTNH, and VTNH_rmcty, the background populations, SEDAC versus Landscan 2000, the choice of aggregation scale resolution, 4km versus 8km and the choice of spatial random effects, BYM versus Leroux. Because of the large number of combinations possible from varying all fixed modeling components (3*2*2*2=24), a subset of 10 combinations was chosen with computational time in mind (shown in Figure VI-1).

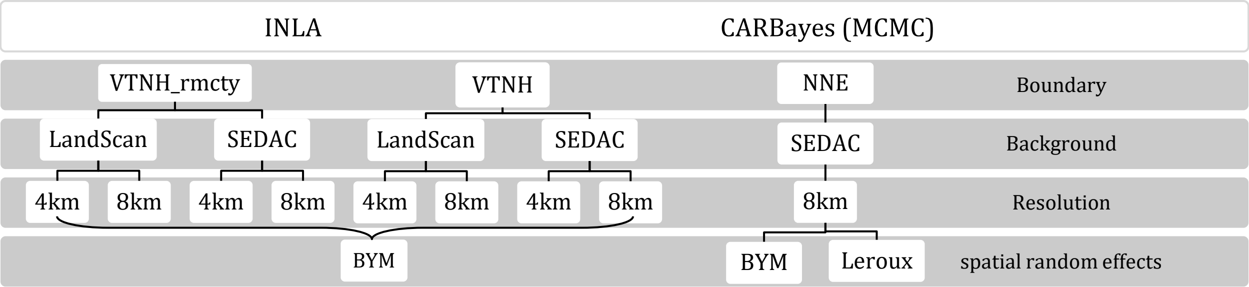


Figure VI-1. Fixed modeling component combinations used in estimation of ALS spatial distribution

An exploratory approach is taken for this sensitivity analysis, where posterior probabilities of predicted SIR are plotted and visually compared. Furthermore, the deviance information criterion (DIC) for each model is calculated and compared (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002).

6.2 Model

A Bayesian hierarchical model is used to estimate the spatial distribution of the ALS risk. The data layer of the hierarchical model is defined by case counts, $\boldsymbol{O}$, in each lattice cell, which has a 4km or 8km resolution. These counts are modeled using a Poisson distribution having a spatially varying intensity, $\boldsymbol{\Lambda}$.

$$\boldsymbol{O} \sim Poisson(\boldsymbol{\Lambda})$$

The process layer models the intensity as the logarithm of a GMFR using either BYM random effects (Besag, York, & Mollié, 1991) or Leroux random effects (Lee 2013) and includes a fixed offset of logarithmically transformed expected counts, $\boldsymbol{E}$, based on one of the two background populations, SEDAC or LandScan 2000. Since some locations are estimated to have 0 expected counts due to very sparse population, all expected counts are augmented by a small number, $1e^{-9}$, in order to still include these locations in the analysis. The model is given as:

$$\log(\boldsymbol{\Lambda}) = \beta + \boldsymbol{\omega} + \log(\boldsymbol{E} + 1e^{-9})$$

where β represents the logarithm of the average risk for the entire region and $\boldsymbol{\omega}$ are the random effects defined as:

$$\omega_{BYM} = \boldsymbol{\omega} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$$

$$\boldsymbol{\omega} \sim N(0, \tau^2 [D - A]^-)$$

or as:

$$\omega_{LEROUX} = \boldsymbol{\omega}$$

$$\boldsymbol{\omega} \sim N(0, \tau^2[\rho(D - A) + (1 - \rho)I]^{-1})$$

where $N(\quad)$ denotes the multivariate Normal distribution, $[\quad]^{-}$ denotes a generalized inverse, $\boldsymbol{A}$ is the n by n adjacency matrix defined by the graph of first order neighbors (N,S, E, W) on the regular lattice having either 4km or 8km resolution and boundary of NNE, VTNH or VTNH_rmcty, $\boldsymbol{D}$ is the n by n matrix of row sums of $\boldsymbol{A}$, $\boldsymbol{I}$ is the n by n identity matrix, and n is the number of lattice cells covering the region, which depends on both the resolution and boundary used. The values for n are summarized in Table VI-1.

Table VI-1. Number of lattice cells by boundary region and lattice resolution

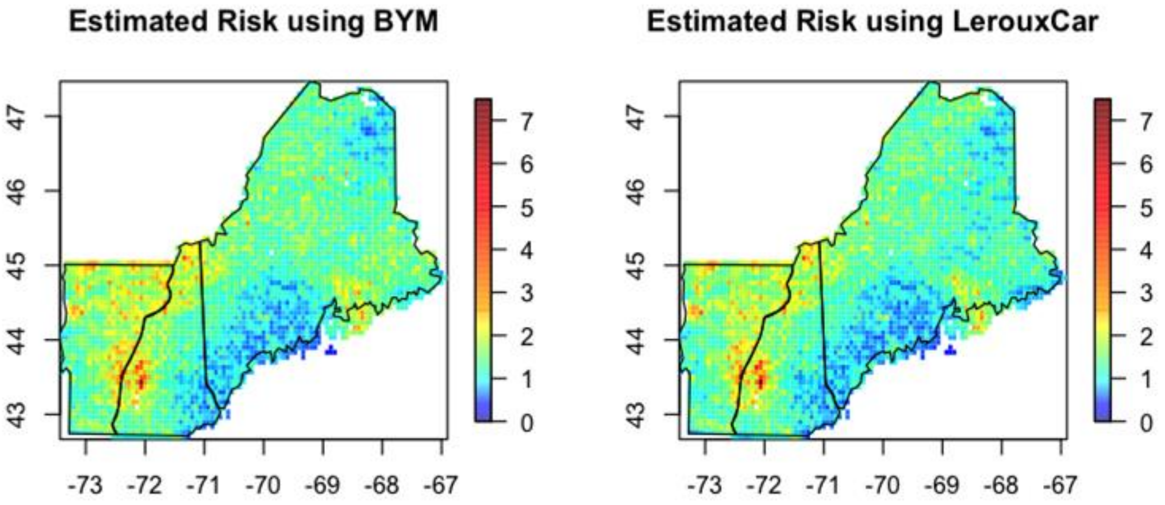
Boundary	Lattice scale	n
VTNH	4km	5108
VTNH	8km	1318
VTNH_rmcty	4km	4231
VTNH_rmcty	8km	1099
NNE	8km	3641

The spatial distribution of ALS for all NNE using BYM and Leroux random effect models were fit using the R package CARBayes (Lee, 2013), which uses a Gibbs sampler in combination with Metropolis Hasting step to run Markov Chain Monte Carlo (MCMC) in order to obtain samples from the posterior distribution. Package default prior distributions were selected for each model. For the variance parameters τ^2 and σ^2, the default prior was Unif (0, 1000). Chains contained 500 warm-up iterations followed by 10,000 iterations that were then thinned by 10 to decreases effects of autocorrelations due to poor mixing. This gave a total of 1000 posterior samples.

The other 8 model combinations that estimated ALS risk within Vermont and New Hampshire used only the BYM random effects and were estimated using the Integrated Nested Laplace Approximation (INLA) implemented in the R package INLA (Rue et al., 2009).

6.3 ALS Spatial distribution in NNE

The mean posterior relative risk estimates for the NNE boundary are plotted in Figure VI-2. In general, there is little difference between the BYM estimates and the Leroux estimates in terms of the spatial distribution of the mean posterior relative risk estimates. Areas of high risk (orange/red) occur between the Vermont and New Hampshire border, along the Northern Maine coastline, and in the northwest section of Vermont. These areas correspond to the locations of identified clusters at the census block group scale and the census track scale (Caller et al., 2013; Torbick et al., 2014).


Estimated Risk using BYM

Estimated Risk using LerouxCar

Figure VI-2. Posterior relative risk mean estimates

There is also large area of estimated low risk (blue) along the coastlines of New Hampshire and southern Maine. It is suspected this is due to an underestimate of case counts, as people in this region are likely to seek medical treatment in the Boston MA area hospitals.

6.4 ALS Spatial distribution in VTNH vs VTNH_rmcty

The spatial distribution of ALS for the boundary regions of VTNH and VTNH_rmcty are shown in Figure VI-3 and Figure VI-4 respectively. Unlike Figure VI-2, which shows the mean posterior risk, these plots show the probability that the posterior risk is above a threshold value of 1.5. The interpretation of a relative risk of 1.5 is that there are 50% more predicted observed cases than would be expected. In further discussions, relative risk above 1.5 is defined as high risk. The probabilities of locations being high risk are defined by Li et al (2012) as exceedance probabilities. Exceedance probabilities incorporate the estimated relative risk as well as its uncertainty and probabilities greater than 0.8 or 0.95 are considered notable (Li, Brown, Gesink, & Rue, 2012).

The region along the border of Vermont and New Hampshire has the highest probabilities of high risk for both the plots with boundary VTNH (Figure VI-3) and boundary VTNH_rmcty (Figure VI-4), although only for the boundary VTNH would the exceedance probabilities be notable. This was also the region of highest relative risk shown for the boundary NNE in Figure VI-2. In the VTNH plots, locations in northwestern Vermont also are shown to have higher probabilities of high risk and these correspond to locations in Figure VI-2, which have high relative risk estimates (2 to 5) based on the NNE boundary. The VTNH boundary plots show an area of very low probability of high risk in southern New Hampshire, which includes the counties of Strafford, Rockingham, and parts of Merrimack and Belknap counties. In the VTNH_rmcty boundary plots where Strafford and Rockingham counties are among the removed counties, the parts of Merrimack and Belknap counties with low exceedance probabilities in the VTNH boundary plots, still remain locations with low exceedance probabilities. Although, areas of low exceedance probability cannot necessarily be interpreted as low risk areas, the NNE boundary plots show this same region to have a predicted low relative risk (less than 1). Since it was suspected that locations closer to Boston have lower case ascertainment, and that there are locations with lower risk/lower probabilities of high risk in each of the boundary area plots that are close to Boston, it appears even the smallest boundary area does not achieve the goal of only including regions not effected by missing cases. By comparing the probabilities in Figure VI-3 and Figure VI-4, the regions of lowest probability near Boston could be better contained by a boundary defined by the major central highways (Route 3 between Nashua and Manchester and I93) and the northeastern borders of Belknap and Stafford counties.

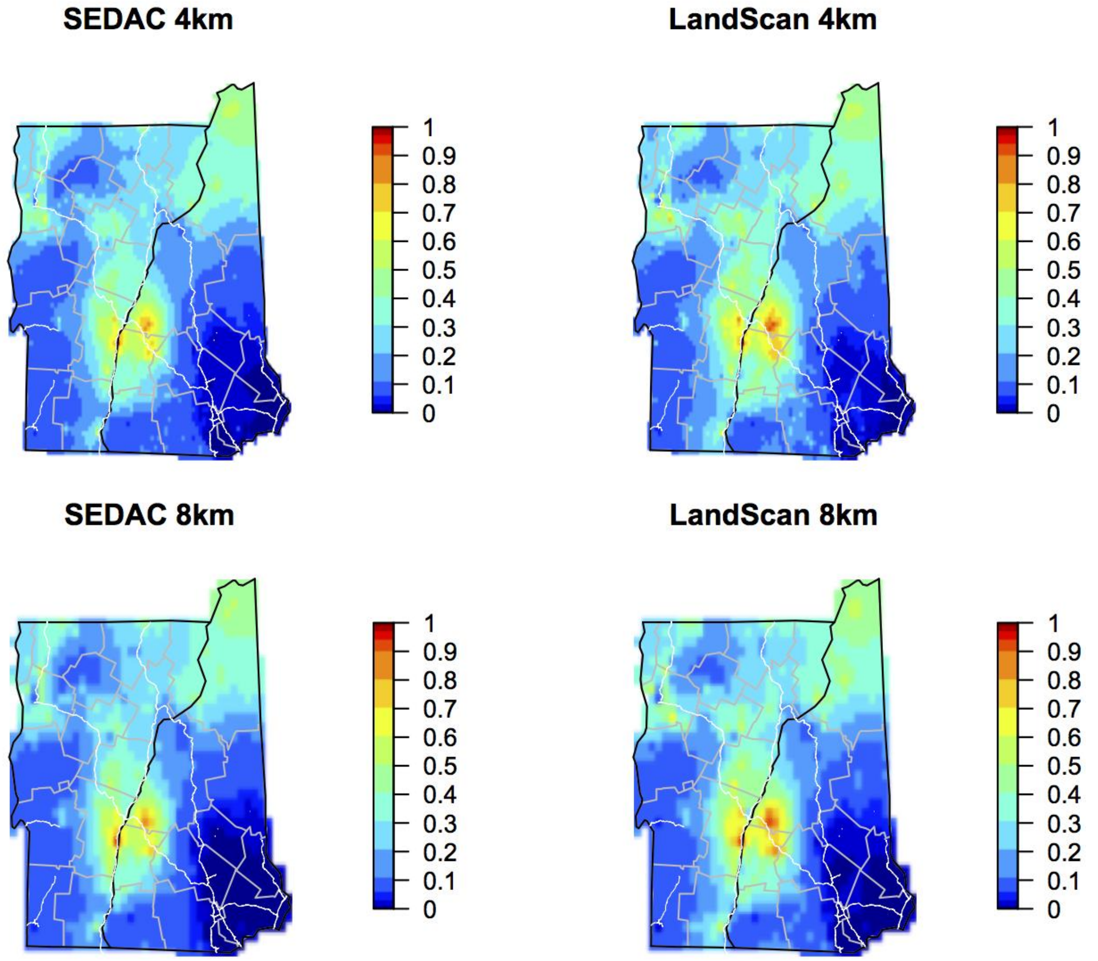


Figure VI-3. Probability Relative Risk is greater than 1.5 for VTNH by resolution scale and background population

While the overall spatial pattern of high risk and low risk appears the same between the three boundary regions of NNE, VTNH, and VTNH_rmcty, the significance of these patterns do appear to vary according to the choice in boundary. When comparing the VTNH boundary versus the VTNH_rmcty boundary, the VTNH boundary has more areas with higher exceedance probabilities and the highest exceedance probabilities are notable. In particular, in northwestern Vermont, locations have lower probabilities of high risk (0.1 to 0.3) when the VTNH_rmcty boundary is used compared to when the VTNH boundary is used (0.3 to 0.7).

The overall spatial pattern between the plots with different background populations, LandScan 2000 versus SEDAC, is also very similar. Although, areas with high probabilities of risk greater than 1.5, appear slightly higher when the LandScan background population is used. In particular, locations in southeastern Vermont and southwestern New Hampshire have higher probability areas that extend to more areas when LandScan is used versus when SEDAC is used. When comparing the 4km resolution and the 8km resolution, 8km probabilities appear smoother, but the overall spatial pattern is quite similar.

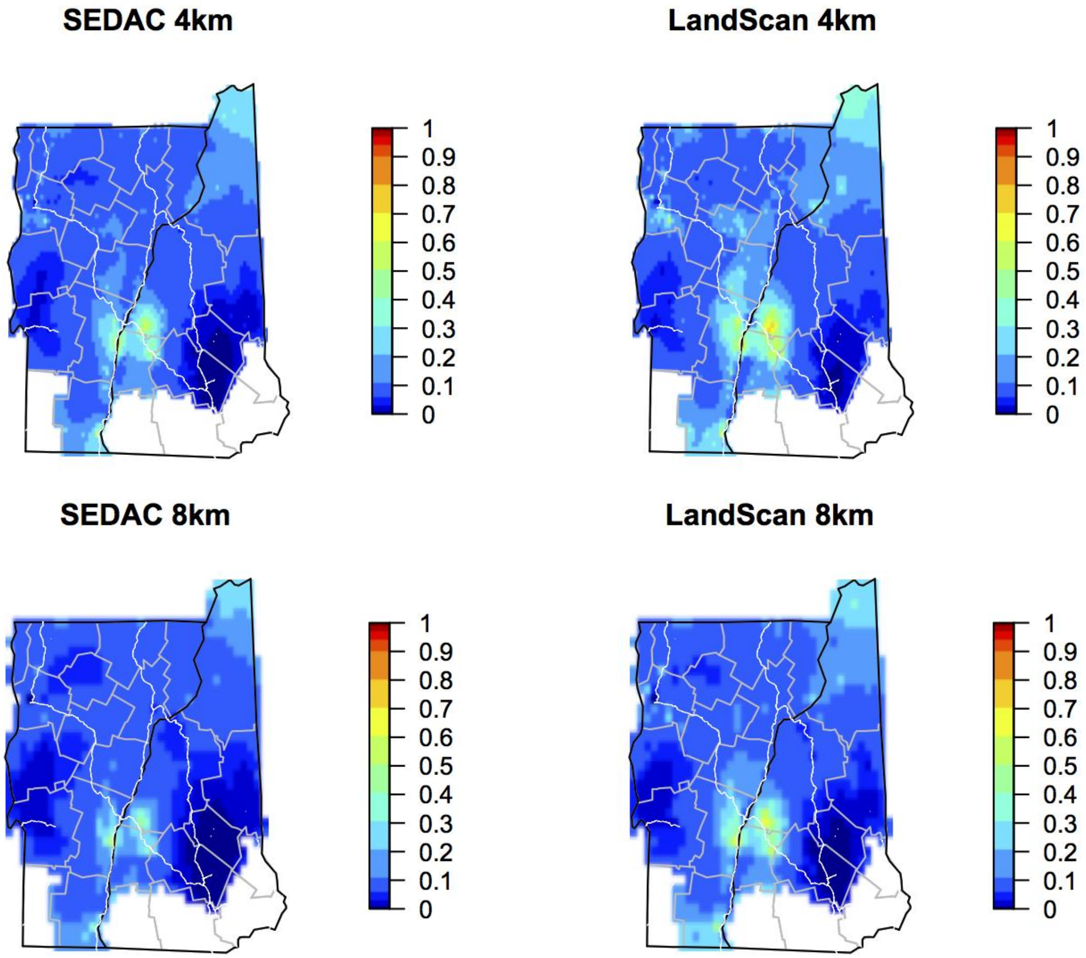


Figure VI-4. Probability Relative Risk is greater than 1.5 for VTNH_rmcty using SEDAC by resolution scale; (left) 4km and (right) 8km

In terms of model fit as quantified by the DIC, the choice of boundary area and the choice of grid size resolution used for the computational grid have the most impact (Figure VI-5). Better fit (lower DIC) values are calculated for models with VTNH_rmcty as the boundary area and a computational grid having an 8km resolution. One reason that the 8km grid may fit better is that some of the case locations contained spatial uncertainty that may be better represented by the 8km resolution scale. In particular, the 8 cases with only a town name that were placed at the centroid locations of the towns were found in towns with median size of 66.3km^2. There is a minimal difference between DIC values for models using Landscan 2000 as the background population for expected counts versus those using SEDAC as the background population.

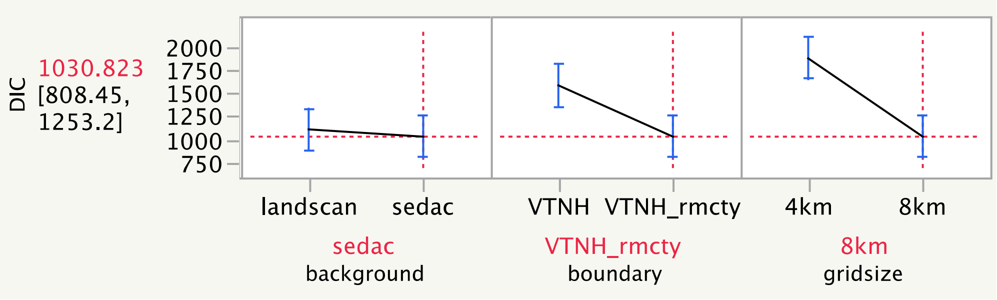


Figure VI-5. Comparison of DIC values between the 8 models estimating the spatial distribution of ALS using different population backgrounds, different boundaries and different computational grid size resolutions.

VII ESTIMATING EXPOSURE EFFECTS

In order to estimate the effect of the water quality metrics on the risk of ALS, regression variables are added at the process layer of the models described in Chapter VI. Recall the process layer in the ALS risk models was defined as:

$$\log(\boldsymbol{\Lambda}) = \beta + \boldsymbol{\omega} + \log(\boldsymbol{E} + 1e^{-9})$$

where $\boldsymbol{\omega}$ are the random effects defined as:

$$\omega_{BYM} = \boldsymbol{g} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$$

$$\boldsymbol{g} \sim N(0, \tau^2 [D - A]^-)$$

or as:

$$\omega_{LEROUX} = \boldsymbol{g}$$

$$\boldsymbol{g} \sim N(0, \tau^2 [\rho(D - A) + (1 - \rho)I]^{-1})$$

Thus in this model, β is replace by $\boldsymbol{X\beta}$ where $\boldsymbol{X}$ is the n by p matrix of regressors and $\boldsymbol{\beta}$ is the vector of regression parameters of length, p. Different water metrics are used in $\boldsymbol{X\beta}$ for the boundary region NNE than for the boundary regions VTNH and VTNH_rmcty.

7.1 RELATIONSHIP OF WATER QUALITY TO ALS RISK OVER NNE

7.1.1 Model

The analysis for the boundary region NNE is provided to allow holistic comparison of the Bayesian hierarchical approach used in this dissertation to the methods discussed in Torbick et al. (2014). Furthermore, regression parameter estimates are compared between a model using BYM random effects to a model using Leroux random effects, which are equivalent to EAR random effects when $\theta = 1$. The grid size for these models is 8km and the background population is based on SEDAC.

The metrics used for the NNE boundary region are similar to the ones used by Torbick et al. (2014), which include lake average Chl-a, SD and TN. Unlike in Torbick et al. (2014) where several area weighed distance averages of these metrics from the case locations, the model used here matches the grid locations to the values of these metrics at the nearest lake. For this reason, the regression additionally includes the variables logarithm of the distance to nearest lake point (smallest Euclidean distance between grid centroid and vertices of lake polygons), and the sine and cosine of the direction to nearest lake point. Thus the regression is:

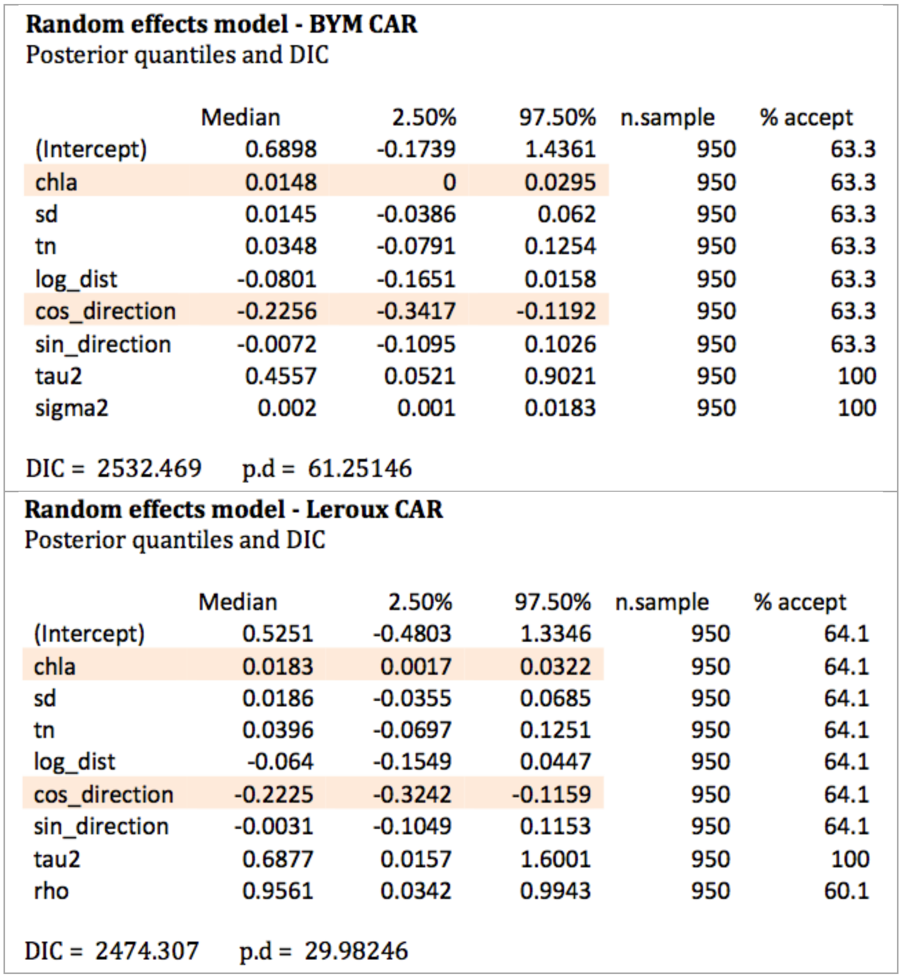
$$\boldsymbol{X\beta} = \beta_0 + \beta_1 * ChlA + \beta_2 * SD + \beta_3 * TN + \beta_4 * logdist + \beta_5 * \mathrm{cosdirect} + \mathrm{\beta}_6 * sindirect$$

The R package CARBayes (Lee, 2013), which implements MCMC, is used to fit these two models. For the regression coefficients β_i the default priors are the non-informative Normal distributions with mean 0 and variance 1000, and for the variance parameters τ^2 and σ^2, the default priors are the Uniform distribution, Unif (0, 1000).

7.1.2 Results

Both Leroux and BYM random effects lead to parameters estimates that agree with Torbick et al. (2014)'s conclusion that poorer water quality is associated with higher ALS risk. Although, the strength of the association depends on the random effects used, which may also be related to the collinearity between the regression variables and the random effects. The median as well as 95% credibility interval limits for the parameter posterior samples of each model are summarized in Table VII-1.

Table VII-1. Summary of parameter posterior samples for BYM and Leroux models fit using CARBayes package in R with 1 chain having 1000 samples



Random effects model - BYM CAR
Posterior quantiles and DIC

	Median	2.50%	97.50%	n.sample	% accept
(Intercept)	0.6898	-0.1739	1.4361	950	63.3
chla	0.0148	0	0.0295	950	63.3
sd	0.0145	-0.0386	0.062	950	63.3
tn	0.0348	-0.0791	0.1254	950	63.3
log_dist	-0.0801	-0.1651	0.0158	950	63.3
cos_direction	-0.2256	-0.3417	-0.1192	950	63.3
sin_direction	-0.0072	-0.1095	0.1026	950	63.3
tau2	0.4557	0.0521	0.9021	950	100
sigma2	0.002	0.001	0.0183	950	100

DIC = 2532.469 p.d = 61.25146

Random effects model - Leroux CAR
Posterior quantiles and DIC

	Median	2.50%	97.50%	n.sample	% accept
(Intercept)	0.5251	-0.4803	1.3346	950	64.1
chla	0.0183	0.0017	0.0322	950	64.1
sd	0.0186	-0.0355	0.0685	950	64.1
tn	0.0396	-0.0697	0.1251	950	64.1
log_dist	-0.064	-0.1549	0.0447	950	64.1
cos_direction	-0.2225	-0.3242	-0.1159	950	64.1
sin_direction	-0.0031	-0.1049	0.1153	950	64.1
tau2	0.6877	0.0157	1.6001	950	100
rho	0.9561	0.0342	0.9943	950	60.1

DIC = 2474.307 p.d = 29.98246

Both models estimate a positive effect of average Chl-a on ALS risk, however the BYM's estimate is smaller and its lower 95% credibility interval bound is 0. The odds ratios of these estimates when the average Chl-a is 1 ug/L and 100 ug/L are summarized in Table VII-2. The estimated odds ratio where the nearest lake has on average 100 ug/L of Chl-a for the BYM model means there is on average 4 times the number of cases as would be expected based on the background population and for the Leroux model means there is on average 6 times the number of cases as would be expected based on the background population. However, there is considerable uncertainty in these estimates, as these range from a 0% or 18% increase to 19 or 25 times more cases as would be expected.

Table VII-2. Odds Ratios (OR) of average Chl-a effect on ALS risk based on BYM and Leroux models

	Average Chl-A	Median OR	Lower 95% OR	Upper 95% OR
BYM	1 ug/L	1.0149	1	1.0299
	100 ug/L	4.3929	1	19.1059
Leroux	1 ug/L	1.0185	1.0017	1.0327
	100 ug/L	6.2339	1.1853	25.0281

The other metrics used for water quality, SD and TN do not have a significant effect based on these models. This differs from the results reported by Torbick et al. (2014) where TN and SD were more significant than Chl-a at predicting clustering membership. However, it could be that the TN and SD at the nearest lake point underestimate the exposure when many lakes are in close proximity, while the approach used by Torbick et al. (2014) averaged these water quality metrics for all lakes within a fixed distance from each case. Furthermore, this difference could be attributed to a number of other factors including, the different background population, census track versus 8km square aggregation scale used for summarizing ALS risk, and/or the two-stage model versus Bayesian hierarchical model.

As for the geographic variables considered in this analysis, both models suggest a significant negative effect of the cosine component of direction to nearest lake point on ALS risk (95% credibility interval bounds are both negative). The interpretation of this effect, as illustrated in Figure VII-1, is that 8km cells to the east of nearest lake points have higher average case counts than those on the west. This could provide some additional evidence for the theory that cyanotoxins can become aerosolized and carried in the wind, since the principal wind direction is from west to east (Stommel, Field, & Caller, 2013; Murby & Haney, 2015). However, it could also be that more of the population in general lives on the eastern side of lakes, and/or this relationship is only true at the 8km scale and not the individual scale. Note that Lake Champlain was not included in this analysis. However, its inclusion may weaken the effect as it is found at the western boundary of the ALS case study region and only has grid cells to its east.

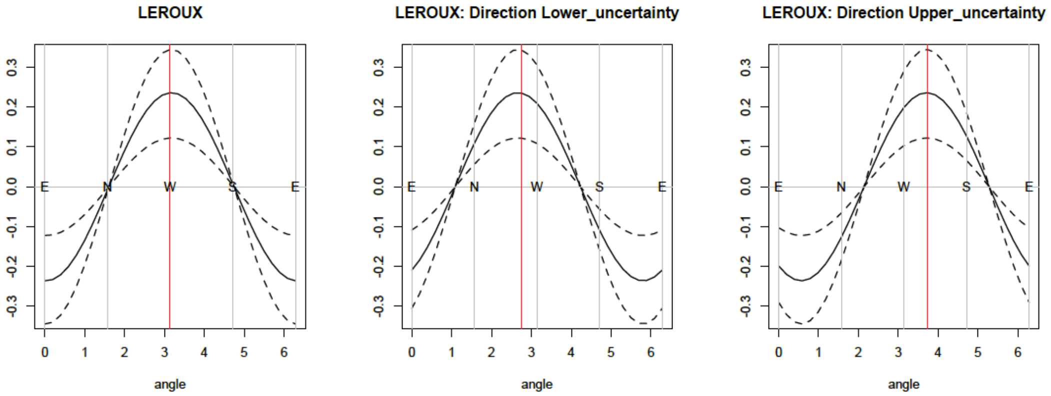


Figure VII-1. Illustration of direction Effect

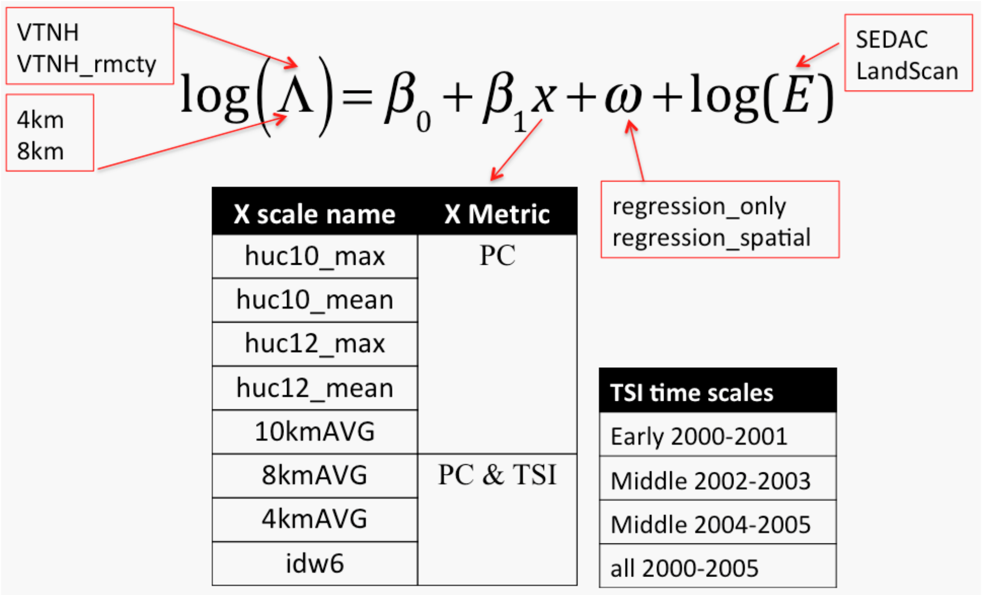
7.2 Relationship of water quality to ALS Risk over VTNH and VTNH_rmcty

7.2.1 Model

In the analysis of the relationship of water quality to ALS risk over VTNH, the two water quality metrics PC and Lake TSI are used at various spatial scales in separate models with the form of the regression added at the process level of the LGCP defined as:

$$\boldsymbol{X\beta} = \beta_0 + \beta_1 X$$

and the parameter estimates of these water quality metrics β_1 are compared between several versions of the model with varied fixed modeling components. In particular, 16 versions of a model with a specific water quality metric are considered where the boundary is chosen to be either VTNH or VTNH_rmcty, the computational grid resolution is either 8km or 4km, the background population expected counts are based on SEDAC or LandScan 2000 and BYM random effects are used (regression_spatial) or are not used (regression_only). Since the PC metric is considered at 8 different spatial aggregation scales and the lake average TSI metric is considered at 3 different spatial aggregation scales and 4 different time aggregation scales, a total of 320 models are fit. Since many candidate models are considered INLA is used for parameter estimation instead of MCMC due to its fast estimation. Figure VII-2 provides a summary of the differences between each model.


$$\log(\Lambda) = \beta_0 + \beta_1 x + \omega + \log(E)$$

X scale name	X Metric
huc10_max	PC
huc10_mean	
huc12_max	
huc12_mean	
10kmAVG	
8kmAVG	PC & TSI
4kmAVG	
idw6	

TSI time scales
Early 2000-2001
Middle 2002-2003
Middle 2004-2005
all 2000-2005

Figure VII-2. Illustration showing the different fixed components of the model that are varied

The goal of this analysis is two-fold. First we seek to identify if any of the spatial aggregation scales of the PC metric or the spatial temporal scales of the lake TSI metric have a significant effect despite differences between the models, and second we seek to explain the impact of the model differences on the model's fit as measured by the DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002).

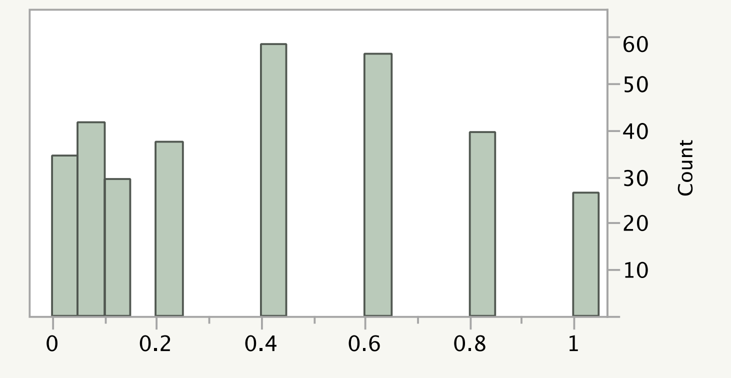
7.2.2 Results

In order to determine significance, percentiles of the regression coefficient's marginal posterior distribution in each of the models are used to calculate p-values. For example, if the smallest positive percentile is the 5th percentile, then the p-value is less than 2*0.05 =0.1. All spatial aggregation scales of the PC metric and all but two space-time aggregation scales of the Lake TSI metric have at least one of the 16 models with a p-value less than 0.05 (Table VII-3), which is often considered as a statistically significant result. The two variables with little evidence of a statistically significant relationship to ALS risk include the area weighted 8km average of Lake TSI for the middle time period 2002-2003 and area weighted 8km average of Lake TSI for the entire time period 2000-2005.

Table VII-3: Summary of P-values for each variables 16 models

Variable	Scale	Min	Q25	Q75	Max	Number of p-values under 0.05
PC	PC_idw6	0.01	0.01	0.0875	0.2	12
PC	PC8kmAVG	0.01	0.02	0.35	1	8
PC	PC_huc10_mean	0.01	0.0625	0.6	0.8	4
PC	PC_huc10_max	0.01	0.0625	0.6	1	4
PC	PC_huc12_mean	0.01	0.1	0.75	1	2
PC	PC10kmAVG	0.05	0.05	0.4	0.8	5
PC	PC4kmAVG	0.05	0.1	0.4	1	2
PC	PC_huc12_max_pc	0.05	0.125	0.6	1	3
TSI	TSI_2004_idw6	0.01	0.01	0.175	0.6	9
TSI	TSI_2002_idw6	0.01	0.05	0.55	0.8	5
TSI	TSI_2004_4kmAVG	0.01	0.0875	0.6	1	4
TSI	TSI_2002_4kmAVG	0.01	0.125	0.6	1	3
TSI	TSI_alltime_idw6	0.01	0.05	0.6	1	5
TSI	TSI_2000_idw6	0.01	0.2	0.75	1	3
TSI	TSI_alltime_4kmAVG	0.01	0.4	0.6	1	3
TSI	TSI_2000_8kmAVG	0.05	0.2	0.8	1	1
TSI	TSI_2000_4kmAVG	0.05	0.4	0.8	1	1
TSI	TSI_2004_8kmAVG	0.05	0.4	0.8	1	1
TSI	TSI_2002_8kmAVG	0.2	0.6	0.8	1	0
TSI	TSI_alltime_8kmAVG	0.4	0.6	1	1	0

Since multiple versions of the same model are fit, it is possible to observe significant results by chance. In particular, for each separate variable, one would expect about 1 of the 16 models to have a p-value less than 0.05 by chance. However, several of the variables have many more than 1 significant p-value, such as the inverse distance weighted average of PC, the 8km area weighted average of PC, and the inverse distance weighted average of Lake TSI for the later time period 2004-2005.When all models are considered together, one would expect 16 of the 320 models to have a p-value less than 0.05 by chance. However, the distribution of the p-values for all models shows that 75 of the 320 have a p-value less than 0.05, which is much larger than would be expected to occur by chance (Figure VII-3).



Approximate P-value	count	proportion
0.01	34	0.106
0.05	41	0.128
0.1	29	0.091
0.2	37	0.116
0.4	58	0.181
0.6	56	0.175
0.8	39	0.122
1	26	0.081
Total	320	1

Figure VII-3. Distribution of p-values for all 320 models

Thus even when there are differences in the background population used for expected counts, differences in computational grid resolutions, differences in boundary areas and despite the difficulty of inference on fixed parameters when random effects are used, this analysis still shows that depending on the aggregation scale, the water quality metrics used here have a statistically significant relationship to the risk of ALS in Vermont and New Hampshire.

Based on the distribution of p-values for each aggregation scale of PC and Lake TSI shown in Figure VII-4, the inverse distance weighted PC average is the most related water quality metric and aggregation scale to the spatial distribution of ALS risk.

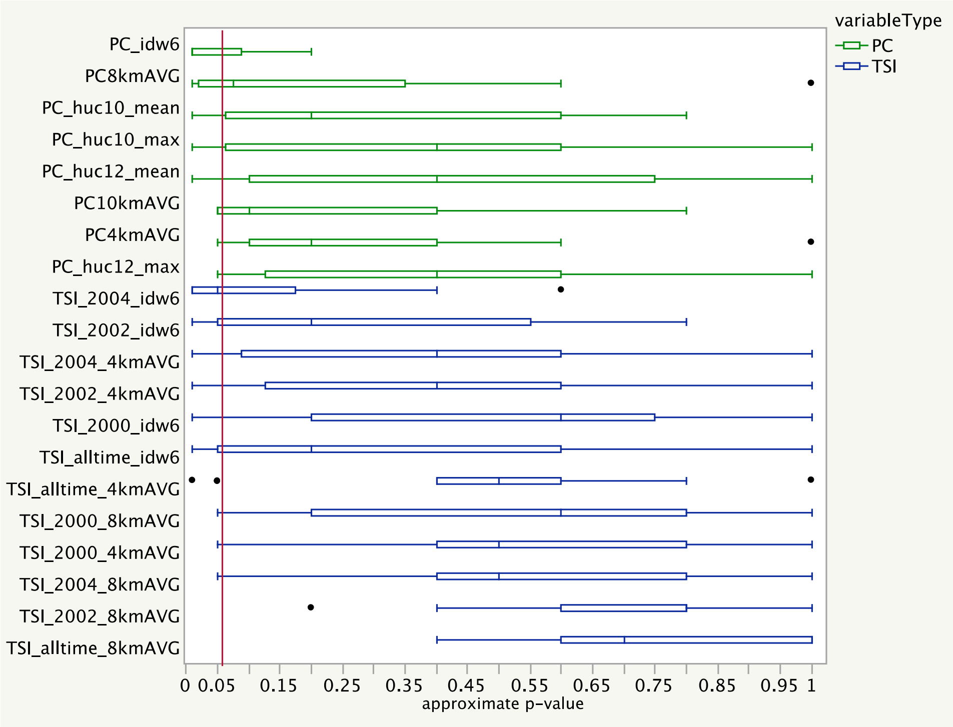


Figure VII-4. Boxplots (distributions) of 16 p-values for each of the 20 variables

The estimated effect of the IDW6 PC metric is positive (column ‘mean’ of Table VII-4), which means that higher amounts of PC are associated with higher prevalence counts of ALS in Vermont and New Hampshire. This is consistent with the results of Torbick et al. (2014), which showed poorer water quality is associated with ALS hotspot membership and provides further evidence in support of the hypothesis that neurotoxins produced by cyanobacteria are a risk factor for ALS.

Table VII-4. Marginal Posterior summaries of the coefficients for the 16 models that included PC IDW6

Random Effect	background	boundary	grid size	mean	**95% CI**		DIC	p-value
none	sedac	VTNH	4km	0.0052	0.0020	0.0082	2563.4	0.01
none	landscan	VTNH	4km	0.0040	0.0007	0.0071	2698.8	0.05
none	sedac	VTNH_rmcty	4km	0.0065	0.0030	0.0099	1801.2	0.01
none	landscan	VTNH_rmcty	4km	0.0052	0.0016	0.0085	1917.1	0.01
none	sedac	VTNH	8km	0.0061	0.0024	0.0095	1587.0	0.01
none	landscan	VTNH	8km	0.0053	0.0016	0.0087	1653.9	0.01
none	sedac	VTNH_rmcty	8km	0.0062	0.0022	0.0098	1145.0	0.01
none	landscan	VTNH_rmcty	8km	0.0056	0.0016	0.0092	1195.2	0.01
BYM	sedac	VTNH	4km	0.0039	-0.0004	0.0079	2477.2	0.1
BYM	landscan	VTNH	4km	0.0033	-0.0012	0.0076	2591.0	0.2
BYM	sedac	VTNH_rmcty	4km	0.0071	0.0026	0.0114	1771.4	0.01
BYM	landscan	VTNH_rmcty	4km	0.0066	0.0018	0.0113	1873.0	0.01
BYM	sedac	VTNH	8km	0.0046	-0.0001	0.0090	1512.8	0.1
BYM	landscan	VTNH	8km	0.0045	-0.0003	0.0090	1564.20	0.1
BYM	sedac	VTNH_rmcty	8km	0.0060	0.0011	0.0106	1124.24	0.05
BYM	landscan	VTNH_rmcty	8km	0.0059	0.0009	0.0107	1166.63	0.05

More specifically, the effect of PC at the IDW6 aggregation scale for the model with the smallest DIC, can be specified in terms of an odds ratio for differing amounts of PC (Table VII-5). When there is 10 ug/L of IDW6 PC, the odds ratio is 1.06, meaning there is 6% increase in average ALS risk and when there is 100 ug/L of IDW6 PC the odds ratio is 1.83, which means there is an 83% increase in average ALS risk.

Table VII-5. Odds ratio of effect of PC IDW 6 when model includes BYM random effects, SEDAC expected counts, the boundary is VTNH_rmcty, and the grid resolution is 8km

ug/L of PC	OR	95% Confidence	
1	1.01	1.00	1.01
10	1.06	1.01	1.11
100	1.83	1.12	2.88

The Lake TSI metric aggregated for years 2004-2005 and using the spatial aggregation IDW6 is the next water quality metric that comes up with mostly significant p-values (Table VII-6).

Table VII-6. Marginal Posterior summaries of the coefficients for the 16 models that included TSI_2004_IDW6

Random Effect	background	boundary	grid size	mean	95% CI		DIC	p-value
none	sedac	VTNH	4km	-0.22	-0.38	-0.07	2565.0	0.01
none	landscan	VTNH	4km	-0.31	-0.47	-0.15	2689.8	0.01
none	sedac	VTNH_rmcty	4km	-0.08	-0.27	0.11	1812.3	0.6
none	landscan	VTNH_rmcty	4km	-0.18	-0.37	0.01	1920.9	0.1
none	sedac	VTNH	8km	-0.23	-0.39	-0.07	1588.8	0.01
none	landscan	VTNH	8km	-0.30	-0.46	-0.14	1647.9	0.01
none	sedac	VTNH_rmcty	8km	-0.18	-0.37	0.01	1150.2	0.1
none	landscan	VTNH_rmcty	8km	-0.28	-0.47	-0.08	1194.3	0.01
BYM	sedac	VTNH	4km	-0.15	-0.33	0.03	2480.0	0.2
BYM	landscan	VTNH	4km	-0.24	-0.43	-0.05	2590.9	0.05
BYM	sedac	VTNH_rmcty	4km	-0.12	-0.35	0.10	1779.2	0.4
BYM	landscan	VTNH_rmcty	4km	-0.24	-0.47	-0.01	1878.2	0.05
BYM	sedac	VTNH	8km	-0.14	-0.33	0.05	1515.7	0.2
BYM	landscan	VTNH	8km	-0.21	-0.40	-0.02	1565.6	0.05
BYM	sedac	VTNH_rmcty	8km	-0.22	-0.44	0.01	1125.2	0.1
BYM	landscan	VTNH_rmcty	8km	-0.31	-0.55	-0.08	1165.9	0.01

The Lake TSI estimated effect is negative, which means lower values of Lake TSI are associated with higher ALS risk. Recall lower values of TSI indicate better lake water quality. This result contradicts the previous findings (Torbick et al., 2014). However other temporal aggregation scales of the lake TSI do not show the same significance, and it could be that this particular time aggregation does not represent typical summer lake water quality. Recall, that differences between the time aggregations may be related to the varying cloud-free satellite overpass dates. Furthermore, Lake TSI is a general water quality metric, which in this analysis is treated as a continuous variable, but is actually an ordered categorical variable. Figure VII-5 shows Lake TSI for the 2004-2005 time period is mostly a contrast between a classification of 2 and 3. Thus more research is needed to determine the usefulness of this metric.

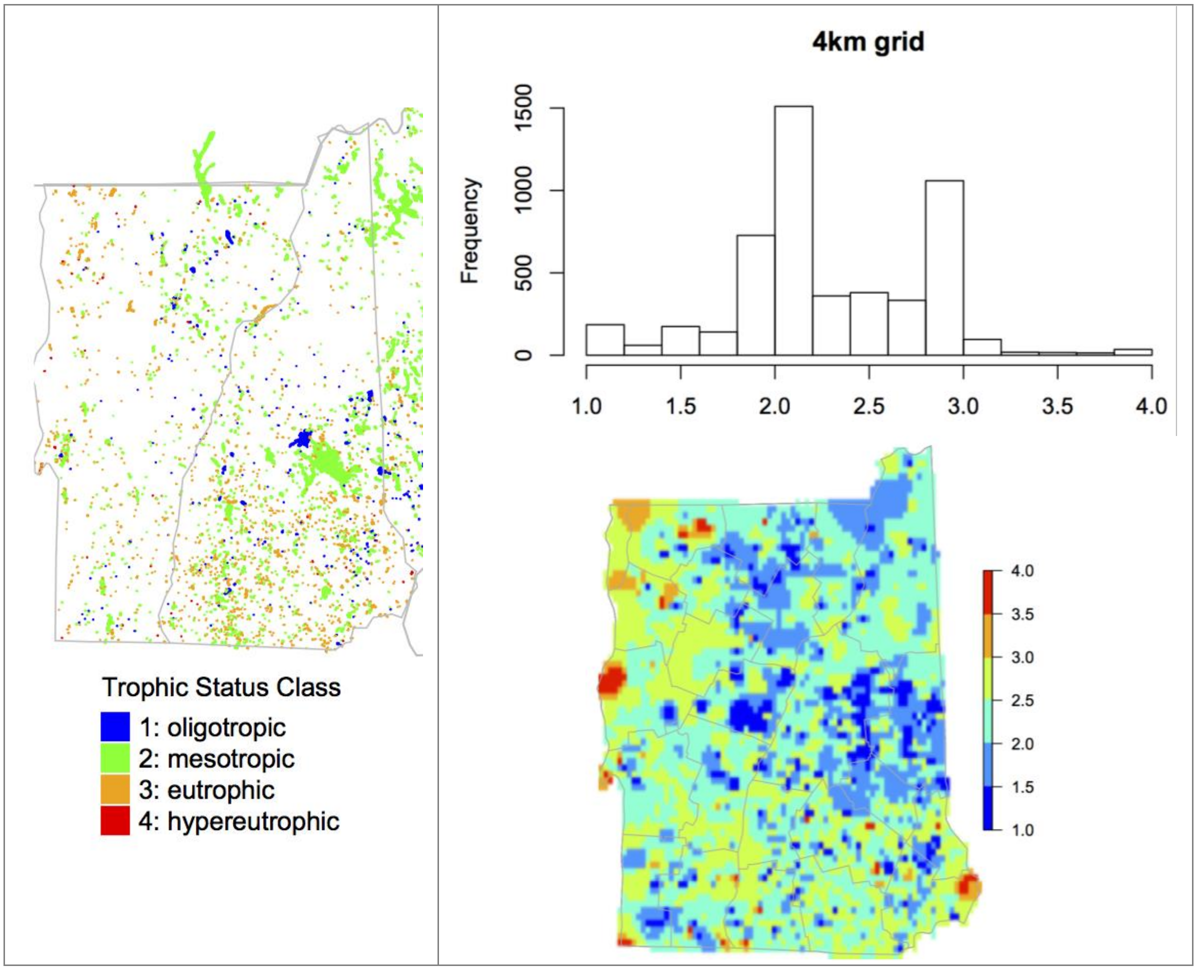


Figure VII-5. Distribution of lake TSI 2004-2005 temporal aggregation; left is original lake aggregation scale; upper right is the histogram of TSI 2004-2005 values when aggregated to match 4km grid using IDW6 scale; lower right is the spatial distribution of the histogram values.

In order to explain the impact of the model differences on the model's fit as measured by the DIC, DIC values for all 320 models are compared by the different model components, background, boundary, grid size use of TSI or PC, and use of random effects (regression_only or regression_spatial). Figure VII-6 shows that the choice of grid size and boundary have the largest effect on the model's DIC, with small DIC (better fit) occurring for models with boundary VTNH_rmcty and grid size of 8km. When the VTNH_rmcty boundary is chosen and the grid size is 8km, there is little difference between models with the different background populations, models that used random effects and those that did not, and the variable choice.

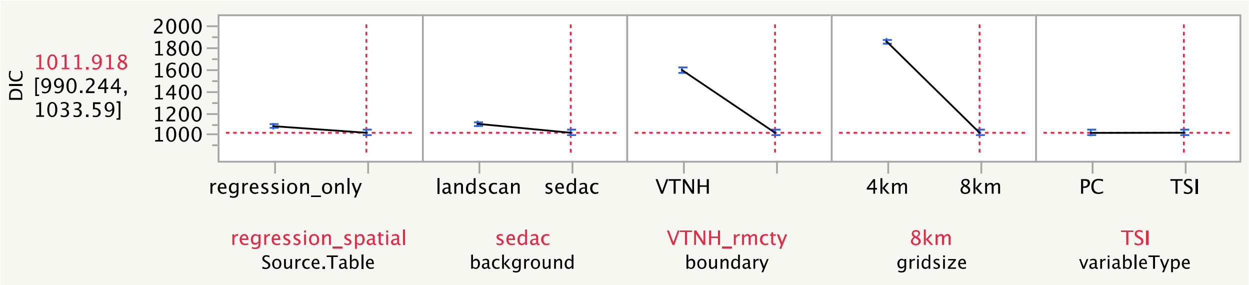


Figure VII-6. Comparison of DIC values between the 320 models estimating the effect of water metric regression variables on the spatial distribution of ALS using different population backgrounds, different boundaries, different computational grid size resolutions, different water metrics and using BYM or no spatial random effects.

VIII EXPOSURE MODELING

8.1 SPATIAL DISTRIBUTION OF HUC12 MAX PC IN VT AND NH

An analysis of the spatial distribution of the PC metric itself was done with the goal of determining if certain locations had higher than regional average PC and with the goal of exploring the spatial uncertainty of one of the spatial aggregation scales of PC. This analysis was only done for the MaxPC_huc12 variable before values were matched to the regular lattice used to summarize ALS disease risk. Figure VIII-1 shows a plot of the HUC12 boundaries intersecting with the state boundaries of Vermont and New Hampshire colored by the observed maximum lake average PC in ug/L. Log-transformed MaxPC_huc12 values were modeled using a Gaussian Process in a Bayesian hierarchical model. The data layer models the log-transformed MaxPC_huc12, $\boldsymbol{y}$, as Normally distributed and in the process layer the mean of this Normal distribution is modeled using the EAR model with smoothing parameter $\theta = 1$, which is equivalent to Czado's CAR and the Leroux CAR. The model is specified by:

$$\text{data layer:} \quad \log(\boldsymbol{y}) \sim N(\boldsymbol{\beta} + \boldsymbol{\omega}, \sigma^2)$$

$$process\ layer\text{:} \quad \boldsymbol{\omega} \sim N(\boldsymbol{0}, \boldsymbol{Q^{-1}})$$

$$\boldsymbol{Q} = [\boldsymbol{I} - \psi(\boldsymbol{A} - \boldsymbol{D} + \boldsymbol{I})]^1$$

where $N()$ denotes the multivariate Normal distribution, $\boldsymbol{A}$ is the 495 by 495 adjacency matrix defined by the graph of first order neighbors (those sharing a border) on the irregular lattice defined by the HUC12 boundaries, $\boldsymbol{D}$ is the diagonal 495 by 495 matrix of row sums of $\boldsymbol{A}$, and $\boldsymbol{I}$ is the 495 by 495 identity matrix.

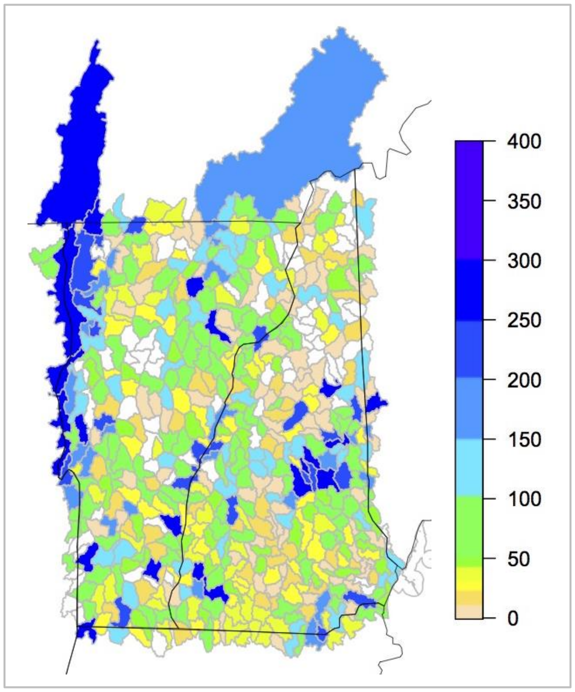


Figure VIII-1. HUC12 boundaries colored by Max lake average PC in ug/L (White areas do not contain any lakes sized a least 6 hectares.)

The VTNH regional average MaxPC_huc12, $\boldsymbol{\beta}$, is given a non-informative Normal prior with mean 4 and standard deviation 100. The variance parameter σ is given a half-cauchy prior, $cauchy(0,5)$, following advice of Gelman et al. (2006). The ψ parameter is transformed as $\phi = \frac{\psi}{1-\psi}$ and given the prior, $lognormal(0, 1.6)$.

The model is then fit using the Hamiltonian Markov Monte Carlo sampler available in the Stan software. Four chains were run for 5000 iterations after 5000 iterations of warm-up to give a total of 20,000 posterior samples. Model convergence was accessed using the Gelman-Rubin statistic (Rhat) (Gelman & Rubin, 1992).

Model predicted MaxPC_huc12 values at original data scale in micrograms per liter (ug/L) are shown in Figure VIII-2, along with 95% credibility interval bounds. HUC12 boundaries including the southern section of Lake Champlain have the highest modeled MaxPC_huc12 values between 100-150 ug/L. However there is a large amount uncertainty in these estimates. In particular the HUC12 area including the northern section of Lake Champlain (Missiquoi Bay) has a slightly lower estimated value (between 50-100 ug/L) than the southern region, but both regions have similar uncertainty bounds, lower of less than 50 ug/L and upper between 300 - 400 ug/L.

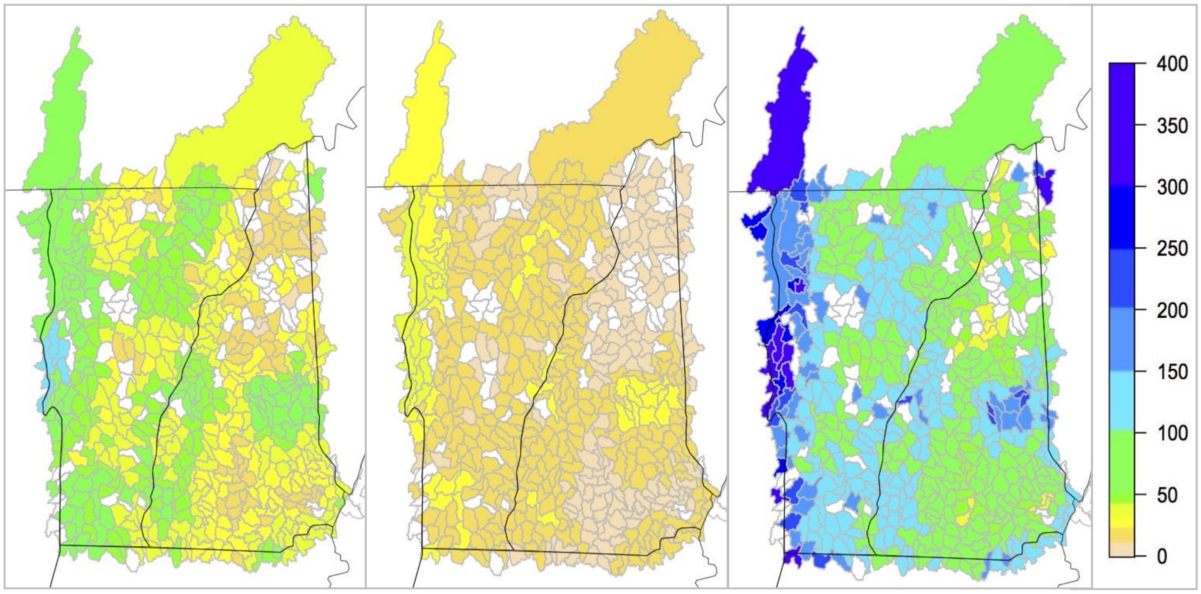


Figure VIII-2. Model predicted MaxPC_huc12 (left); Lower 95% credibility bound (middle); Upper 95% credibility bound (right)

The distribution of the spatial random effect $\boldsymbol{\omega}$ can be used to assess for regions with higher than average MaxPC_huc12 values. Figure VIII-3 shows the probabilities (percent of posterior samples of $\boldsymbol{\omega}$) that any particular HUC12 boundary area's spatial random effect was positive. The HUC12 boundary areas with the highest probabilities include parts of Lake Champlain.

Other higher probability regions include the area around Lake Winnipesaukee as well as along the western border between Vermont and New Hampshire.

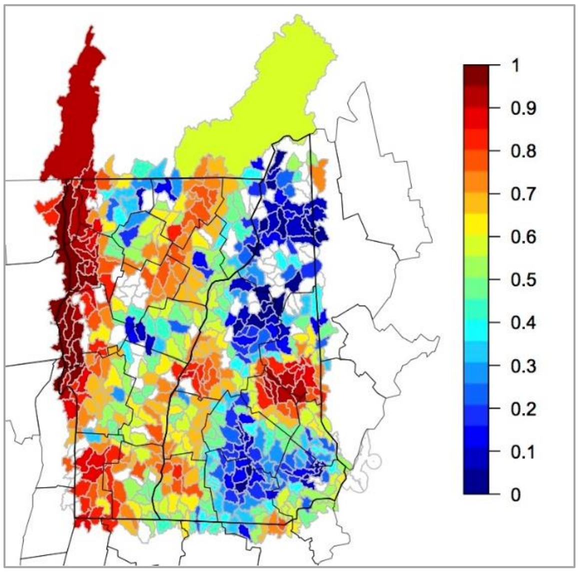


Figure VIII-3. Proportion of posterior spatial random effect samples greater than 0.

8.2 SPATIO-TEMPORAL SURFACE TEMPERATURES OF LAKE CHAMPLAIN

Since much remains unknown about spatial distribution of water quality over time, a case study of water temperature within Lake Champlain is investigated with the goal of determining if surface water temperatures vary within the lake, if the surface water temperatures have warmed since 1984 and if the amount of warming varies spatially within the lake. Warmer water temperatures are of interest because of their association with the frequency and extent of algae blooms (Paerl and Huisman, 2008, 2009).

8.2.1 Model

A Bayesian spatially varying coefficient model is used to analyze trends in the satellite-derived surface water temperatures. The surface temperatures in tenths of degree Celsius,$y(s, t)$, which are indexed by s, a location in Lake Champlain, and by t, a time point between mid-July to late August (day of year 195 to 244) for years 1984 to 2011, form the data layer of the hierarchical model and are considered to be a realization of a continuous temperature process $\mu(s,t)$ measured with Gaussian noise ε. Thus

$$y(s,t) = \mu(s,t) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

where σ_ε^2 is commonly referred to as a nugget in geostatistics. The spatial locations,$\{s_i, i = 1, 2, \ldots, k\}$, form a regular lattice covering Lake Champlain where the temperature at a given point s_i represents the average water temperature for the area of the lake covered by the lattice cell. Two lattice resolution are used, one at a 2160m resolution and the other at a 1080m resolution, $k = 352$ for the 2160m case and $k = 1217$ for the 1080m case. The time points $\{t_i, i = 1, 2, \ldots, 63\}$ are centered continuous year values. For example for a measurement taken during 1999 on day of year 219 is $t_i = 1999.6 - mean(years)$.

We model the continuous temperature process $\mu(s,t)$ as

$$\mu(s,t) = \beta_1 * pcloud(s,t) + \omega_0(s) + \omega_1(s) * t$$

where β_1 is the mean difference in temperature due to missing data and $\{\omega_i(s), i = 0, 1\}$ are spatial random effects. In particular, $\omega_0(s)$ are the spatially varying intercepts and $\omega_1(s)$ are the spatially varying linear trend parameters. The spatial random effects $\{\omega_i(s), i = 0, 1\}$, are modeled with an IEAR prior. Recall the IEAR is an improper prior defined by:

$$\omega_i(s) \sim \mathcal{N}(\mathbf{0}, \tau_i^2 \boldsymbol{Q}^-)\,, \;\; for\ i = 1, 2$$

$$\boldsymbol{Q} = (\boldsymbol{D} - \boldsymbol{A})^{\boldsymbol{\theta_i}}$$

where $\boldsymbol{A}$ is a k by k matrix defined as in [4.4.1] with $\delta = 1$ and $\gamma(d_{ij}) = 1$, $\boldsymbol{D}$ is the k by k diagonal matrix of the row sums of $\boldsymbol{A}$, and $\theta_i \in [0, \infty)$ is the smoothness of the spatial process with larger values indicating a longer range of spatial correlation. In particular recall that larger values of θ increase the eigenvalues of the covariance matrix $\boldsymbol{Q}^{-1}$greater than 1, which are associated with the eigenvectors describing large-scale spatial variability. For the 1080m lattice, the eigenvalues are plotted in order smallest to largest in Figure VIII-4.

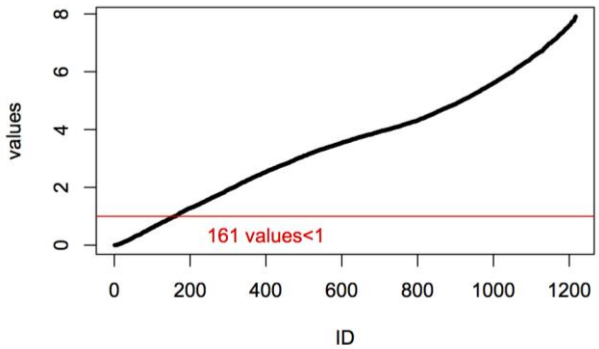


Figure VIII-4. Eigenvalues by ID for the matrix D-A when the 1080m lattice is used.

Notice that eigenvalues 1 to 161 are less than 1, so inversely are greater than 1, and these are associated with the large-scale eigenvectors for which a few examples are shown in the top of Figure VIII-5. Conversely, eigenvalues 162 to 1217 will all be less than 1 in the inverse and these are associated with the small-scale eigenvectors for which a few examples are show in the bottom of Figure VIII-5.

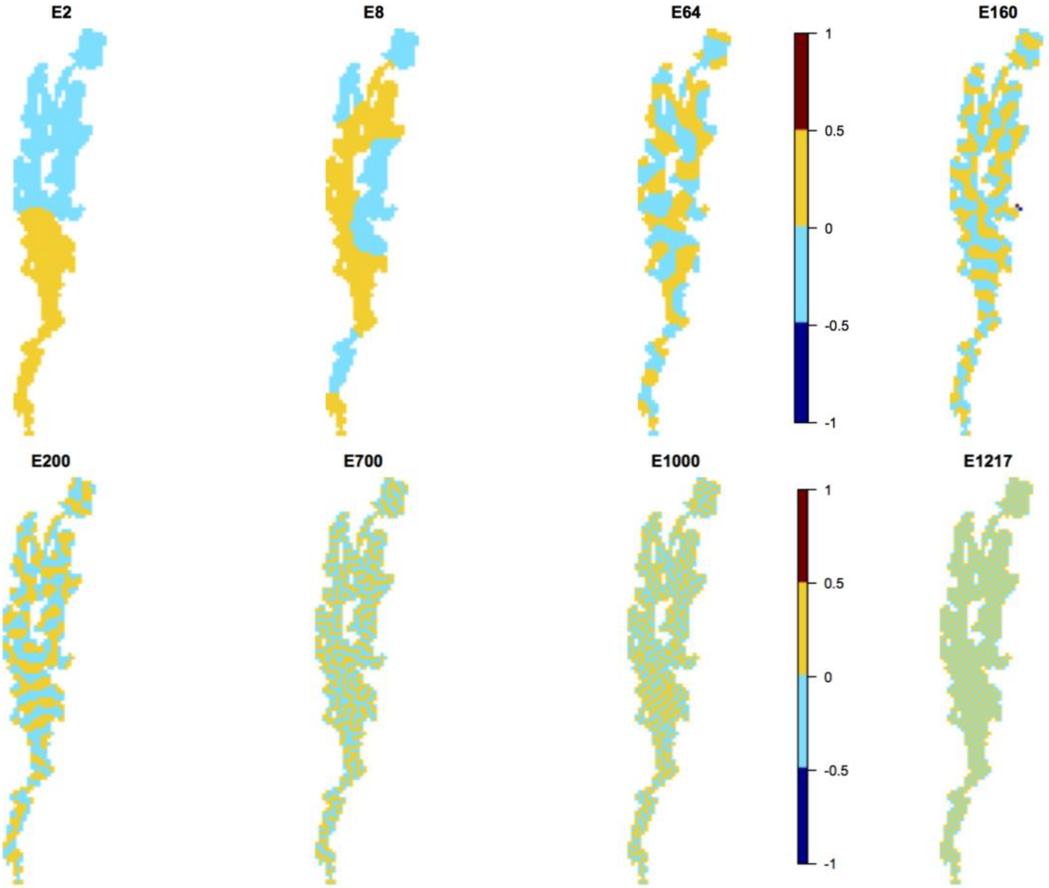


Figure VIII-5.Example Eigenvectors of the Matrix D – A for the 1080m computational lattice. (Top: Large-scale spatial variability; Bottom: Small-scale spatial variability).

To improve computational efficiency of the model parameter estimation the spatial structure was removed following a similar pre-whitening procedure as outlined by Yuan (2011). First an eigenvalue decomposition of the matrix $(\boldsymbol{A}-\boldsymbol{D}+\boldsymbol{I})$ is calculated:

$$\boldsymbol{A}-\boldsymbol{D}+\boldsymbol{I}=\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^{\boldsymbol{T}}$$

where $\boldsymbol{E}$ is the orthonormal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. Since the first column of $\boldsymbol{E}$ is associated with the 0 eigenvalue (IEAR is only semi-positive definite), it is removed to define the matrix used for pre-whitening

$$\widetilde{\boldsymbol{E}}=\boldsymbol{E}[\boldsymbol{1}:\boldsymbol{k}\,,\boldsymbol{2}:\boldsymbol{k}].$$

Then the data and missing pixel covariate values are pre-whitened. Pre-whitened data have the distribution:

$$\widetilde{\boldsymbol{E}}^T y(s,t) \sim N\left(\beta_1 \cdot \left(\widetilde{\boldsymbol{E}}^T pcloud\right) + \widetilde{\boldsymbol{E}}^T \omega_0(s) + \widetilde{\boldsymbol{E}}^T \omega_1(s) \cdot t, \sigma_\varepsilon^2\right)$$

and the transformed spatially varying parameters have the distribution:

$$\widetilde{\boldsymbol{E}}^T \omega_i(s) \sim \mathcal{N}\left(\boldsymbol{0}, \tau_i^2 \left[\boldsymbol{I} - \widetilde{\boldsymbol{\Lambda}}\right]^{\boldsymbol{\theta}_i}\right), \quad for\ i = 1, 2$$

where $\boldsymbol{I} - \widetilde{\boldsymbol{\Lambda}}$ is a diagonal matrix of size $k-1$ by $k-1$, with eigenvalues λ_i

, $i = 2, \dots, k$. After model estimation, estimates of the spatially varying parameters are obtained from the back transformation:

$$\widetilde{\boldsymbol{E}}\left[\widetilde{\boldsymbol{E}}^T \omega_i(s)\right] = \omega_i(s)$$

8.2.2 Results & Discussion

Parameter estimation of the hierarchical model is carried out using the Hamiltonian Monte Carlo No-U-turn MCMC algorithm available in Stan in order to obtain samples from the posterior distribution (Stan Development Team 2016). The prior distributions used for the variance parameters σ_ε^2, $\{\tau_i^2, i = 1, 2\}$, and the smoothing parameters $\{\theta_i, i = 1,2\}$ along with the other details of the model are given in Appendix B. After four separate chains are run with 10000 iterations and a warm-up of 5000 iterations for each chain is discarded, model convergence is accessed using the Gelman-Rubin statistic, Rhat (Gelman and Rubin, 1992).

Posterior sample summaries for all parameters except the random effects are shown in Table VIII-1 for the 2160m resolution computational grid and in Table VIII-2 for the 1080m resolution computational grid. All Rhat values are 1 indicating the models have converged. Mean posterior parameter estimates are similar for the 2160m resolution and for the 1080m resolution.

However, the effective samples sizes for the process layer prior parameters are smaller for the 1080m resolution than for the 2160m resolution. This indicates the sampler had some poor mixing and did not efficiently explore the posterior parameter distribution space.

Table VIII-1. Summary of posterior samples for model parameters (2160m grid)

Level	Parameter	Mean	95% Credibility		Effective sample size	Rhat
Data	β_1	-216.74	-217.2	-216.28	20,000	1
	σ_y	9.27	9.18	9.35	20,000	1
Random intercept $\omega_0(s)$	θ_0	1.54	1.32	1.77	2,050	1
	τ_0	3.32	2.93	3.77	2,738	1
Random trend $\omega_1(s)$	θ_1	1.42	1.07	1.82	3,806	1
	τ_1	0.08	0.06	0.11	20,000	1

Furthermore, parameters related to the random intercepts had the smallest effective sample sizes for both resolutions, and this could be related to collinearity issues between the pcloud regression parameter and the random intercepts.

Table VIII-2. Summary of posterior samples for model parameters (1080m grid)

Level	Parameter	Mean	95% Credibility		Effective sample size	Rhat
Data	β_1	-216.39	-216.65	-216.14	20,000	1
	σ_y	10.03	9.98	10.09	20,000	1
Random intercept $\omega_0(s)$	θ_0	1.53	1.39	1.66	700	1
	τ_0	4.13	3.85	4.44	1,014	1
Random trend $\omega_1(s)$	θ_1	1.68	1.46	1.94	2,288	1
	τ_1	0.06	0.05	0.08	8,868	1

The model estimated late summer average surface water temperatures for 1984-2011 are shown in Figure VIII-6. Including the uncertainty estimates, these temperatures range from 21 to 25°C. These are based on a least squares estimated entire lake average of 22.5°C and the model estimated random intercepts.

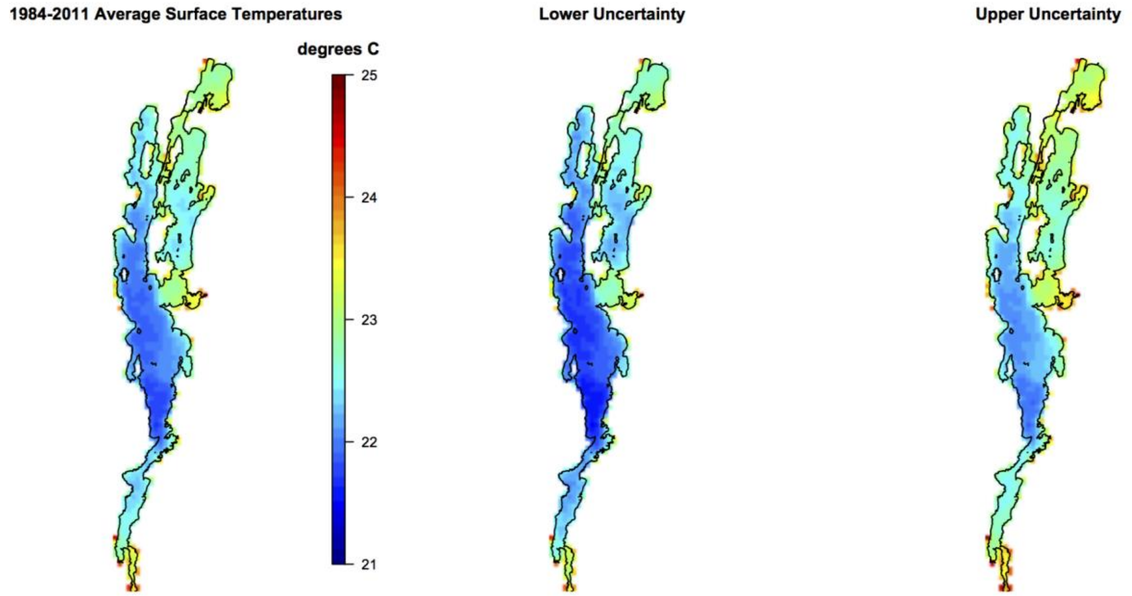


Figure VIII-6. Posterior Mean of Average Surface water temperatures (right) and 95% credibility interval (center and left) based on the 1080m resolution grid.

.

Based on these model estimates, surface water temperatures do vary within the lake. In particular, locations (red) where surface water temperatures are greater than lake average are shown in Figure VIII-7, and these include Missiquoi Bay, St. Albans Bay, and Mallets Bay, which are all known to have cyanobacteria blooms (Torbick & Corbiere, 2015). These locations correspond to the random intercepts having a marginal posterior distribution with a 2.5 percentile that is greater than 0. The general spatial pattern is the same for the 2160m and the 1080m resolutions.

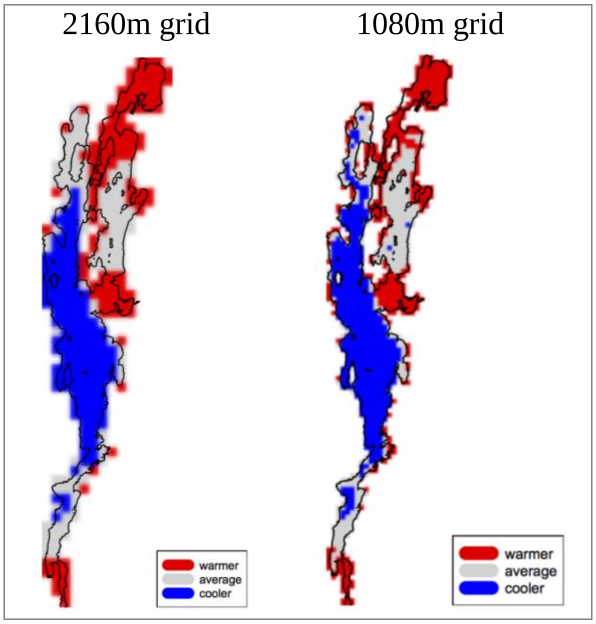


Figure VIII-7. Locations of significantly warmer and significantly cooler than lake average surface water temperatures.

It is possible that the in-situ average water temperatures are not warmer in these areas, but that the satellite derived temperatures are measured as warmer due to the presence of cyanobacteria blooms. In this case, temperature may be a good proxy metric from cyanobacteria bloom. For example, the satellite derived surface temperatures for day of year 227 and year 2008 compared with the Landsat image are shown in Figure VIII-8. Temperatures of approximately 30°C are measured for locations that are shown to have algae blooms in the Landsat image.

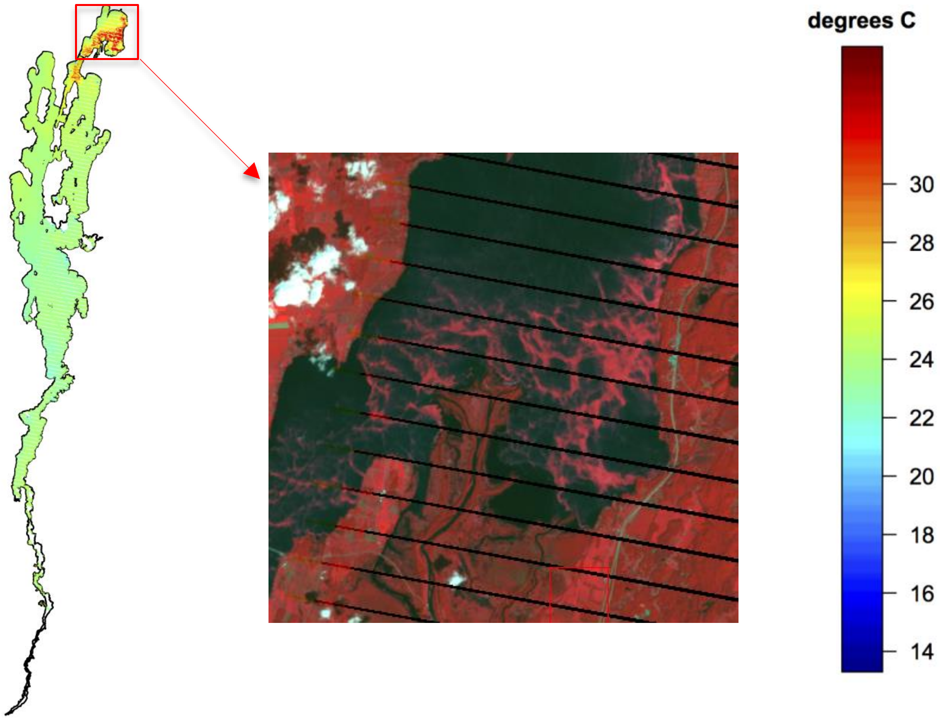


Figure VIII-8. Satellite-derived surface water temperatures of Lake Champlain and Landsat image near Missiquoi Bay for Year 2008 and Day of Year 227

The spatial random slope parameters are estimated to be between -0.5 and 0.6 °C change per decade (Figure VIII-9). This range includes both the uncertainty bounds and the mean posterior values. Significantly increasing trends are found in the northern sections of the lake including Missiquoi Bay, but also near Shelburne Bay and South Lake.

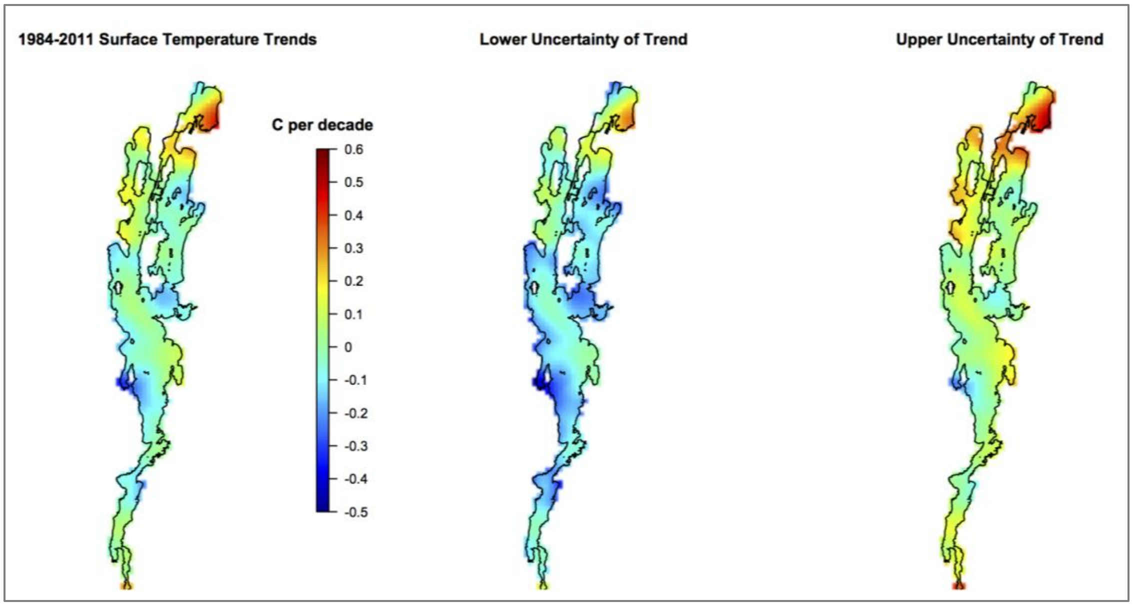

1984-2011 Surface Temperature Trends
Lower Uncertainty of Trend
Upper Uncertainty of Trend
C per decade
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5

Figure VIII-9. Posterior Mean of Temperature Trends (right) and 95% credibility interval (center and left) based on the 1080m resolution grid

The significance of these trends depends on the chosen resolution, where more areas have higher significance for the 1080m resolution than the 2160m resolution (Figure VIII-10). However, since more locations are tested in the 1080m resolution, this could be a result of the multiple testing problem, where there is a greater chance to find significant results by chance.

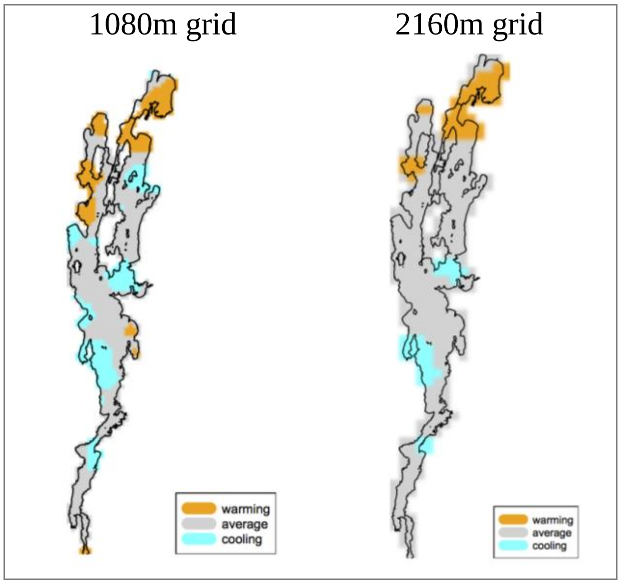


Figure VIII-10. Locations with significantly warmer or cooler temperatures based on a p-value of 0.05 and confidence interval of 95%

The spatial pattern of higher probabilities of increasing trends is similar for both resolutions (Figure VIII-11).

The model also does estimate a few regions of significantly decreasing trends (Figure VIII-10), which contradicts previous studies of temperature trends in Lake Champlain (Smeltzer et al., 2012). However, these decreasing trends could be an artifact of the EAR model, since the eigenvectors represent high and low regions at multiple scales. For example, when there is an area of increase, the model also requires an area of decrease. For this reason, it would make more sense to add an overall lake trend and then the spatial random slopes would represents regions higher or lower than lake average. However, in the current approach this average trend is not identifiable, likely because of remaining collinearity issues between the fixed and random effects.

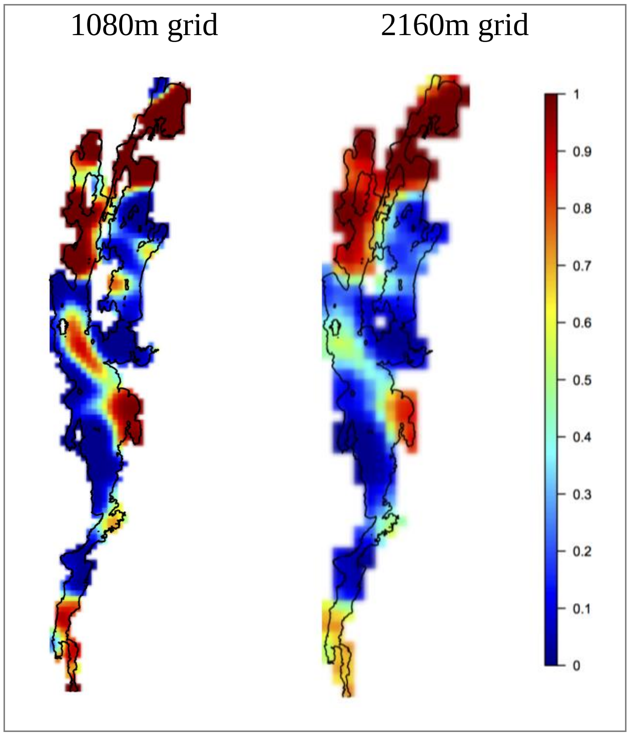


Figure VIII-11. Probability of increasing trends

IX CONCLUSIONS

In this dissertation, the ALS case dataset discussed in Caller et al. (2013) and Torbick et al. (2014) is reanalyzed using Bayesian inference of a LGCP, which includes spatial random effects specified by the logarithm of a GMRF on a regular lattice. The LGCP identifies locations of high risk within Northern New England determined using exceedance probabilities that correspond to the locations of ALS clusters or “hotspots” identified in Caller et al. (2013) and in Torbick et al. (2014). The issue of artifact clusters due to incomplete case ascertainment is investigated by considering three boundary regions, where the smaller regions attempt to remove locations where low case ascertainment is suspected. The significance of the high-risk areas decreases when these lower case ascertainment regions are removed, suggesting clustering may be an artifact of the case ascertainment. However, the smallest boundary region still includes areas with very low probability of high risk that are adjacent to the areas removed because of low case ascertainment. The estimation of clusters also depends on other fixed aspects of the model. In particular, the Irish study, which had nearly complete case ascertainment and did not find “hotspots”, modeled the cases aggregated by Ireland’s electoral divisions using a BYM model (Rooney et al., 2014; Rooney et al., 2015). However if the aggregation unit is not meaningful to the disease risk process, the BYM model may give poor or unexpected results (Wall 2004; Li et al., 2012). Therefore, in this dissertation, the effects of resolution choice of the regular lattice, the choice of background population, and the choice of spatial random effects (BYM versus Leroux) are investigated in terms of model fit by comparing the DIC values of different models. It is shown that the resolution choice and boundary area choice have the largest impact on model fit, where the larger of the two resolutions and the smallest boundary area considered give the smallest DIC values. It is suspected that the larger resolution provides a better model fit because of the spatial uncertainty in some case locations that have only a town name.

For the estimation of the relationship between ALS risk and satellite derived metrics of the region’s lake water quality, the modeled component of the intensity for the LGCP is defined by the water quality metrics as regression parameters combined with spatial random effects. This specification improves on Torbick et al. (2014)’s analysis by explicitly accounting for the spatial uncertainty of ALS “hotspots”. A further improvement in this dissertation is the inclusion of the PC metric, which is more closely related to the presence of cyanobacteria than the Chl-a metric, SD metric and the TN metric. In this specification, the Chl-a metric and the PC metric are found to be positively associated to ALS risk, which agrees with the conclusion in Torbick et al. (2014).

However, the significance of these associations depends on several fixed components of the modeling framework. In particular, the use of random effects to account for spatial autocorrelations can create variance inflation in the fixed effects, meaning significant fixed effects may appear insignificant (Reich et al., 2006; Hughes & Haran, 2013; Hughes 2015). In order to deal with this difficulty as well as investigate the resolution choice of the regular lattice, the choice of background population, and the choice of spatial aggregation for the fixed effects on the significance of the fixed effects, Bayesian inference using INLA is carried out for 320 different models that vary each of these components. For the random effects, half of the models included the BYM random effects and the other half did not include any spatial random effect. Despite model differences, it was found that the PC metric matched to the regular lattice using an inverse distance weighted average between the centroid of the lattice cells and the centroids of the 30m pixels of PC covering the region’s lakes, has a significantly positive relationship with ALS risk in most model variations.

This dissertation does not account for the temporal uncertainty in the PC metric, which is temporally misaligned to the case incidence times. Instead another more general lake quality metric, TSI, is considered, which is more temporally aligned to the times of case incidence. Several space-time aggregation scales are considered for the TSI metric. However, the TSI metric is significant for fewer models than the PC variable and when it is found statistically significant, it has a negative association with ALS risk, which contradicts the PC analysis and results from Torbick et al. (2014). However, there are several reasons that can account for this difference. In particular, only one of the time aggregation scales gives significant results and differences in time aggregations scales may be related to varying satellite overpass dates.

Given the temporal misalignment difficulty, a space-time model of surface water temperatures is developed for Lake Champlain as a case study in order to provide ideas for future research, since warmer lake temperatures are associated with increased cyanobacteria blooms (Paerl and Huisman, 2008, 2009). This model makes use of the IEAR developed by Yuan (2011) and is shown to account for long range spatial autocorrelations by increasing the eigenvalues of the variance matrix that are associated with the eigenvectors explaining large-scale spatial patterns. The model estimates lake temperatures and lake temperature trends that vary spatially within the lake, where some of the trends are increasing while others are decreasing. It is suspected the decreasing trends are an artifact of the model since the large-scale spatial pattern contrasts regions of low and high values. Thus when an increasing trend is estimated, the model forces a decreasing trend. This could be corrected by adding an overall lake trend fixed effect, but this results in identifiability issues.

X FUTURE WORK

There are a number of areas of possible future work that could improve upon the analyses discussed in this dissertation, both in terms of the scientific questions related to water quality and ALS as well as in terms of statistical innovation.

For one, the relationship between water quality and ALS risk could be investigated in different regions. Of particular interest with respect to the NNE region would be to expand the boundaries to include areas of Quebec and upstate New York that are near the northern parts of Lake Champlain. There seems to be little published studies about ALS in Quebec, except for one in the Saguenay region, which investigated ALS incidence between the years of 1985 and 2009. Interestingly, this study shows that the incidence in this region is significantly higher for the last 5-year period of the study 2005-2009, and further notes that this region's population is expected to be at high risk for genetic disorders related to the "founder's effect", but this study only found one case with a family history of ALS (Lareau-Trudel et al., 2013). Since the Saguenay region has a large lake, Lac St. Jean, a case study of the water quality space-time dynamics of this lake might provide some useful insights into the relationship between cyanobacteria exposure time and ALS onset, and thus improving upon the issues in this dissertation created from the temporal misalignment between the case incidence times and the water quality metrics.

There are also a few alternatives to deal with the incomplete case ascertainment of the ALS dataset analyzed in this dissertation as well as in Caller et al. (2013) and Torbick et al. (2014). For one, an alternative subset of the region VTNH could be defined using major highways to delineate the section to remove. Another alternative is to use the preferential sampling method describe in Diggle et al. (2010) and Simpson et al. (2016), which defines a function that takes values between 0 and 1 giving the percent of non-missing cases. This function could be based on expert opinion or perhaps estimated based on the mortality rates of ALS for this region at the county level available from the Wonder CDC website.

In terms of statistical innovation, this dissertation only compares differences between parameters in the LGCP based on two different resolutions of a regular lattice. However, Simpson et al. (2016) argues that using the triangulation method described in Lindgren et al. (2011) to construct the approximation of the GP and a second grid defined by connecting the centroids of the triangulation's edges to estimate the intensity of the point process allows for a more flexible and second-order accurate approximation to a continuous intensity LGCP, where the regular lattice leads to an approximation that is only first-order accurate.

Furthermore, this dissertation only considers deterministic functions at various scales to deal with the spatial misalignment between the water quality measurement locations and the computational grid locations used to summarize the case intensity. An alternative approach would spatially (and even better spatially and temporally) model the water quality metrics simultaneously with the ALS intensity using a version of the multivariate CAR (MCAR) such as the one used in Terres et al. (2015). Instead of a MCAR, a two-step model could be set up, where several (1000 or 10000) simulations from a spatial-temporal model of water quality could then be used in the LGCP. This method would also give a more realistic measure of the uncertainty in the exposure effect (Waller & Gotway, 2004, pp. 400-409).

Other statistical innovations could be developed specifically for the EAR model. For example, the methods described in Reich et al. (2006) and Hughes and Haran (2013) could likely be extended to the EAR and IEAR models since the eigenvectors of these models do not depend on any of the precision matrix parameters and this may resolve the collinearity issues that result when these models are used with fixed effects. Also, since the Leroux model is the same as the EAR model when $\theta = 1$ and two recent papers have develop methods for estimating the Leroux model in INLA, these method could possibly be extended to estimating the EAR model in INLA (Lee & Mitchell, 2013; Ugarte et al., 2014).

REFERENCES

Banack, S. A., Caller, T., Henegan, P., Haney, J., Murby, A., Metcalf, J. S., Stommel, E. (2015). Detection of cyanotoxins, β-n-methylamino-l-alanine and microcystins, from a lake surrounded by cases of amyotrophic lateral sclerosis. *Toxins*, *7*(2), 322-36. doi:10.3390/toxins7020322

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & hall/CRC monographs on statistics & applied probability. CRC Press.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825-848.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192-236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1-20.

Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*(4), 733-746.

Caller, T. A., Chipman, J. W., Field, N. C., & Stommel, E. W. (2013). Spatial analysis of amyotrophic lateral sclerosis in northern new england, USA, 1997-2009. *Muscle & Nerve*, *48*(2), 235-41. doi:10.1002/mus.23761

Caller, T. A., Doolin, J. W., Haney, J. F., Murby, A. J., West, K. G., Farrar, H. E., Stommel, E. W. (2009). A cluster of amyotrophic lateral sclerosis in new hampshire: A possible role for toxic cyanobacteria blooms. *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases*, *10 Suppl 2*, 101-8. doi:10.3109/17482960903278485

Carlson, R. E. (1977). A trophic state index for lakes. *Limnology and Oceanography*, *22*(2), 361-369.

Cressie, N., & Wikle, C. (2011). *Wiley Series in Probability and Statistics: Statistics for spatio-temporal data*. Hoboken, New Jersey: John Wiley & Sons.

Czado, C., & Prokopenko, S. (2008). Modelling transport mode decisions using hierarchical logistic regression models with spatial and cluster effects. *Statistical Modelling*, *8*(4), 315-345. doi:10.1177/1471082x0800800401

Diggle, P. J., Menezes, R., & Su, T. -L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 191-232.

Diggle, P. J., Moraga, P., Rowlingson, B., & Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, *28*(4), 542-563. doi:10.1214/13-sts441

Diggle, P. J. (2013). *Monographs on Statistics and Applied Probability: Statistical analysis of spatial and spatio-temporal point patterns* (third ed.). Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

Gelman, A., Carlin, J. B., Stern, H., & Rubin, D. (2004). *Texts in Statistical Science: Bayesian data analysis* (second ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis, 1*(3), 515-534.

Guidelines and Recommendations. (2016). Retrieved from https://www.epa.gov/nutrient-policy-data/guidelines-and-recommendations

Guinness, J., Fuentes, M., Hesterberg, D., & Polizzotto, M. (2014). Multivariate spatial modeling of conditional dependence in microscale soil elemental composition data. *Spatial Statistics, 9*, 93-108. doi:10.1016/j.spasta.2014.03.009

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika, 57*(1), 97-109.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics, 5*(2), 173-190.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research, 15*(1), 1593-1623.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(1), 139-159.

Hughes, J. (2015). CopCAR: A flexible regression model for areal data. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America, 24*(3), 733-755. doi:10.1080/10618600.2014.948178

Hupper, V. P. (2005). Contributions to modeling and computer efficient estimation for gaussian space-time processes. *PhD dissertation*. Department of Mathematics and Statistics: University of New Hampshire.

Kahle, D., and Wickham, H., (2013). ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

LandScan. (2000). *High resolution global population data set* . Copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government has certain rights in this Data Set.

Lareau-Trudel, Fortin, Gauthier, M., Lavoie, S., Morissette, & Mathieu, J. (2013). Epidemiological surveillance of amyotrophic lateral sclerosis in saguenay region. *The Canadian Journal of Neurological Sciences*, *40*(05), 705-709.

Lawson, A. B. (2012). Bayesian point event modeling in spatial and environmental epidemiology. *Statistical Methods in Medical Research*, *21*(5), 509-29. doi:10.1177/0962280212446328

Lee, D. (2011). A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, *2*(2), 79-89. doi:10.1016/j.sste.2011.03.001

Lee, D. (2013). CARBayes: An R package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, *55*(13), 1-24.

Lee, D., & Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(4), 593-608.

Li, Y., Brown, P., Gesink, D. C., & Rue, H. (2012). Log gaussian cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, *21*(5), 479-507. doi:10.1177/0962280212446326

Linder, E. (2001). Computer-efficient spatial estimation and interpolation based on conditional gaussian autoregressive models. Proceedings: Joint Statistical Meetings. 2001. American Statistical Association. Alexandria, VA.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423-498.

Loftin, K. A., Graham, J. L., Hilborn, E. D., Lehmann, S. C., Meyer, M. T., Dietze, J. E., & Griffith, C. B. (2016). Cyanotoxins in inland lakes of the united states: Occurrence and potential recreational health risks in the EPA national lakes assessment 2007. *Harmful Algae*, *56*, 77-90.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451-482.

Murby, A. L., & Haney, J. F. (2015). Field and laboratory methods to monitor lake aerosols for cyanobacteria and microcystins. *Aerobiologia.* doi:10.1007/s10453-015-9409-z

Neal, R. M. (2011). MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

Noonan, C. W., White, M. C., Thurman, D., & Wong, L. -Y. (2005). Temporal and geographic variation in united states motor neuron disease mortality, 1969–1998. *Neurology*, *64*(7), 1215-1221. doi:10.1212/01.WNL.0000156518.22559.7F

Nychka, D., Wikle, C., & Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Stat Modelling, 2*(4), 315-331. doi:10.1191/1471082x02st037oa

Paciorek, C. J., & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics, 17*(5), 483-506. doi:10.1002/env.785

Paciorek, C. J. (2013). Spatial models for point and areal data using markov random fields on a fine grid. *Electron. J. Statist., 7*(0), 946-972. doi:10.1214/13-ejs791

Paerl HW, Huisman J. 2008. Blooms like it hot. Science 320:57-58.

Paerl HW, Huisman J. 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environmental Microbiology Reports 1(1): 27–37.

Pettitt, A. N., Weir, I. S., & Hart, A. G. (2002). A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing, 12*(4), 353-367.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning.* Cambridge, Massachusetts: MIT Press.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics, 62*(4), 1197-206. doi:10.1111/j.1541-0420.2006.00617.x

Rooney, J., Heverin, M., Vajda, A., Crampsie, A., Tobin, K., Byrne, S., Hardiman, O. (2014). An exploratory spatial analysis of ALS incidence in ireland over 17.5 years (1995-july 2013). *PloS One, 9*(5), e96556. doi:10.1371/journal.pone.0096556

Rooney, J., Vajda, A., Heverin, M., Elamin, M., Crampsie, A., McLaughlin, R., Hardiman, O. (2015). Spatial cluster analysis of population amyotrophic lateral sclerosis risk in ireland. *Neurology, 84*(15), 1537-44. doi:10.1212/WNL.0000000000001477

Rue, H., & Held, H. (2005). *Monographs on Statistics & Applied Probability: Gaussian markov random fields: Theory and applications.* Chapman & Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 71*(2), 319-392.

Sampson, P. D., & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association, 87*(417), 108-119.

Schneider, P., & Hook, S. J. (2010). Space observations of inland water bodies show rapid surface warming since 1985. *Geophys. Res. Lett., 37*(22). doi:10.1029/2010gl045059

Seirup, & Yetman. (2006). *U.S. Census grids (summary file 1), 2000*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved from http://dx.doi.org/10.7927/H4B85623

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: Computationally efficient inference for log-gaussian cox processes. *Biometrika, 103*(1), 49-70. doi:10.1093/biomet/asv064

Smeltzer, E., Shambaugh, A. d., & Stangel, P. (2012). Environmental change in lake champlain revealed by long-term monitoring. *Journal of Great Lakes Research, 38,* 6-18. doi:10.1016/j.jglr.2012.01.002

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

Stan Development Team. (2016). *Stan modeling language users guide and reference manual, version 2.11.0*. http://mc-stan.org.

Stein, L. (1999). *Springer Series in Statistics: Interpolation of spatial data: Some theory for kriging.* New York, NY: Springer-Verlag.

Stommel, E. W., Field, N. C., & Caller, T. A. (2013). Aerosolization of cyanobacteria as a risk factor for amyotrophic lateral sclerosis. *Medical Hypotheses, 80*(2), 142-5. doi:10.1016/j.mehy.2012.11.012

Stroud, J. R., Stein, M. L., & Lysen, S. (2016). Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics,* (just-accepted).

Susser, E. (2004). Eco-epidemiology: Thinking outside the black box. *Epidemiology, 15*(5), 519-520.

Taylor, B. M., & Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes. *Journal of Statistical Computation and Simulation, 84*(10), 2266-2284.

Taylor, B., Davies, T., Rowlingson, B., & Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in R. *Journal of Statistical Software, 63,* 1-48.

Terres, M. A., Fuentes, M., Hesterberg, D., & Polizzotto, M. (2015). Bayesian spectral modeling of microscale spatial distributions in a multivariate soil matrix. *ArXiv Preprint ArXiv:1505.07798.*

Torbick, N., Hession, S., Stommel, E., & Caller, T. (2014). Mapping amyotrophic lateral sclerosis lake risk factors across northern new england. *International Journal of Health Geographics, 13,* 1. doi:10.1186/1476-072X-13-1

Torbick, N., & Corbiere, M. (2015). A multiscale mapping assessment of lake champlain cyanobacterial harmful algal blooms. *International Journal of Environmental Research and Public Health, 12*(9), 11560-78. doi:10.3390/ijerph120911560

Torbick, N. (2015). *Lake surface water temperatures in unit tenths of degree Celsius at 90m resolution pixels covering Lake Champlain (Landsat path row tile 014029)* [114 rasterfile datasets].

Torbick, N. (2016). *Phycocyanin in ug/L 2014-2015 snapshot at 30m resolution pixels across all lakes in Northern New England sized 6 hectares or more* [rasterfile dataset].

Torbick, N. (2016). *Lake Trophic Status Index for 5,812 lakes across Northern New England sized 6 hectares or more for Landsat overpass dates during July-Sept from 2000-2005* [mulitple csv datasets].

Torbick, N., Ziniti, B., Wu, S., and Linder, E. (2016). *Mapping spatiotemporal lake skin temperature trends in the northeast USA.* Manuscript submitted for publication.

Ugarte, M. D., Adin, A., Goicoa, T., & Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate bayesian inference. *Statistical Methods in Medical Research*, *23*(6), 507-30. doi:10.1177/0962280214527528

Waagepetersen, R. (2004). Convergence of posteriors for discretized log gaussian cox processes. *Statistics & Probability Letters*, *66*(3), 229-235.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*(2), 311-324. doi:10.1016/s0378-3758(03)00111-3

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data.* Hoboken, N.J.: John Wiley & Sons.

What Are the Differences Between GPW, GRUMP and Landscan? (2015). Retrieved from https://sedac.uservoice.com/knowledgebase/articles/130821-what-are-the-differences-between-gpw-grump-and-la

Wood, A. T. A., & Chan, G. (1994). Simulation of stationary gaussian processes in [0, 1] d . *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *3*(4), 409-432. doi:10.1080/10618600.1994.10474655

Yuan, C. (2011). Models and methods for computationally efficient analysis of large spatial and spatio-temporal data. *PhD dissertation*. Department of Mathematics and Statistics: University of New Hampshire.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, *99*(465), 250-261. doi:10.1198/016214504000000241

APPENDIX A: R CODE FOR CHAMPLAIN ANALYSIS

```
### Lake Champlain surface water temperatures
### day of year 2000 to 239
### resolution 2160m (24 times original 90m)

### data preparpation steps

library(rgdal);library(raster)

## original data
champ_rast=brick("Landsat_data/rasterChamp.tif")
time = read.csv("Landsat_data/champlain_JAS_84-14_wtemp_bands.csv")


# get location of just water
percent_water=brick("cloud_percent.tif")
percent_water[percent_water==0]<-NA
percent_water[percent_water>0]<- 1
waterarea=percent_water

champ_rast=waterarea*champ_rast
names(champ_rast)=paste(time$Year,time$DOY,sep="_")

## subset data for day of year between 195 (~july 14) to 244(~aug 31) (about 50days)
d195to244=which(time$DOY<195 | time$DOY>244)
champ_rast_195to244=dropLayer(champ_rast,d195to244)
time_195to244=time[-d195to244,c(3,4,5)]


## remove values less than 16C  (5% of data for DOY 200 to 239)
cloud=champ_rast_195to244

cloud[cloud<160]<- 0   ## cloud is counter intuitive (1 means not cloud)
cloud[cloud>=160]<- 1

champ_rast_195to244[champ_rast_195to244<160]<- 0  ## threshold
## resolution decreased
champ2160=aggregate(champ_rast_195to244,fact=24,mean)
champ1080=aggregate(champ_rast_195to244,fact=12,mean)


#get number of cells averaged
numcell2160=aggregate(waterarea,fact=24,sum)
numcell1080=aggregate(waterarea,fact=12,sum)

#get percent of average that is below 16
num_not_cloud2160=aggregate(cloud,fact=24,sum)
num_not_cloud1080=aggregate(cloud,fact=12,sum)

percent_cloud2160=1-(num_not_cloud2160/numcell2160)
names(percent_cloud2160)=names(champ2160)
percent_cloud1080=1-(num_not_cloud1080/numcell1080)
names(percent_cloud1080)=names(champ1080)

#####
ynames=names(champ1080)

champ1080=champ1080*cell1080rm
percent_cloud1080=percent_cloud1080*cell1080rm
names(champ1080)=ynames
names(percent_cloud1080)=ynames

champ2160=champ2160*cell2160rm
percent_cloud2160=percent_cloud2160*cell2160rm
names(champ2160)=ynames
names(percent_cloud2160)=ynames

## save as data

cloud_per100_2160=as.data.frame(rasterToPoints(percent_cloud2160))
champ2160_pts=as.data.frame(rasterToPoints(champ2160))

#ncells=as.data.frame(rasterToPoints(numcell))
cloud_per100_1080=as.data.frame(rasterToPoints(percent_cloud1080))
champ1080_pts=as.data.frame(rasterToPoints(champ1080))

write.csv(cloud_per100_2160,file="cloud_per100_2160.csv")
write.csv(champ2160_pts,file="champ2160_pts.csv")

write.csv(cloud_per100_1080,file="cloud_per100_1080.csv")
write.csv(champ1080_pts,file="champ1080_pts.csv")



### Fitting EAR MODEL in STAN

### champlain 2160m resolution 1984 - 2011 analysis with clouds as 0

champ2160=read.csv("champ2160_pts.csv")
names(champ2160)[1]="cellid"
champ2160=champ2160[,-67]  ### removing year 2014

clouds=read.csv("cloud_per100_2160.csv")
names(clouds)[1]="cellid"
clouds2160=clouds[,-67]

### Getting Neighborhood matrix
locs=champ2160[,2:3]
distance=dist(locs)
distance[which(distance>2160)]=0
distance[which(distance==2160)]=1
A=as.matrix(distance)
d=(apply(A,1,sum))
summary(as.factor(d))
### getting Eigenvalue decomposition of A-D+I
svd=eigen(A-diag(d-1))
F_s=svd$vectors
lam=svd$values


## for fitting IEAR (could considered removing a few more to reduce dimension of
 "random effects")
lam_star=lam[-1]
F_s_star=F_s[,-1]

### prewhitening data and covariates #ignores DOY variability
temp=as.matrix(champ2160[,-c(1:3)])
cloudp=as.matrix(clouds2160[,-c(1:3)])
year=as.numeric(substr(names(champ2160[,-c(1:3)]),2,5)) +
 as.numeric(substr(names(champ2160[,-c(1:3)]),7,9))/365.25

##############
## rough OLS of all data pooled ignoring spatial variability
## use intercept and cp estimes for normal mean priors for fixed effects
olsdat=data.frame(t=stack(champ2160[,-c(1:3)])[,1],cp=stack(clouds2160[,-
 c(1:3)])[,1],yr=sort(rep(year-mean(year),352)))
dim(olsdat)
names(olsdat)
olsfit1=lm(t~cp+yr,data=olsdat)
summary(olsfit1)
####################

temp=temp-summary(olsfit1)$coef[1,1]  ## center temp

temp_w=t(F_s_star)%*%temp
ones_w=t(F_s_star)%*%matrix(1,nrow=352,ncol=63)
percent_missing_w=t(F_s_star)%*%cloudp
years_w=t(F_s_star)%*%matrix(year-mean(year),nrow=352,ncol=63,byrow=TRUE)
years=matrix(year-mean(year),nrow=351,ncol=63,byrow=TRUE)

### preparing data for STAN
N=dim(A)[1]-1   ## number of spatial random effects
n=N*63
temp_w=stack(as.data.frame(temp_w))[,1]
ones_w=stack(as.data.frame(ones_w))[,1]
percent_missing_w=stack(as.data.frame(percent_missing_w))[,1]
years_w=stack(as.data.frame(years_w))[,1]
years=stack(as.data.frame(years))[,1]
cell=rep(1:N,63)
data=c("N","n","temp_w","percent_missing_w","years","cell","F_s_star","lam_star")

### running model in STAN
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

ear_champ2160=stan(file="stan_model_July18_iear_cloudpercent_Champ_model.stan",data=da
 ta,chains=4,iter=10000,control=list(adapt_delta=0.99,max_treedepth=20),refresh=500)


### champlain 1080m resolution 1984 - 2011 analysis with clouds as 0 (any value less
 16C)
champ1080=read.csv("champ1080_pts.csv")
names(champ1080)[1]="cellid"
champ1080=champ1080[,-67]  ### removing year 2014

clouds=read.csv("cloud_per100_1080.csv")
names(clouds)[1]="cellid"
clouds1080=clouds[,-67]

### Getting Neighborhood matrix
locs=champ1080[,2:3]
distance=dist(locs)
distance[which(distance>1080)]=0
distance[which(distance==1080)]=1
A=as.matrix(distance)
d=(apply(A,1,sum))
summary(as.factor(d))


### getting Eigenvalue decomposition of A-D+I
svd=eigen(A-diag(d-1))
F_s=svd$vectors
lam=svd$values

## for fitting IEAR (could considered removing a few more to reduce dimension of
 "random effects")
lam_star=lam[-1]
F_s_star=F_s[,-1]

### prewhitening data and covariates #ignores DOY variability
temp=as.matrix(champ1080[,-c(1:3)])
cloudp=as.matrix(clouds1080[,-c(1:3)])
year=as.numeric(substr(names(champ1080[,-c(1:3)]),2,5)) +
 as.numeric(substr(names(champ1080[,-c(1:3)]),7,9))/365.25

##############
## rough OLS of all data pooled ignoring spatial variability
## use intercept and cp estimes for normal mean priors for fixed effects
olsdat=data.frame(t=stack(champ1080[,-c(1:3)])[,1],cp=stack(clouds1080[,-
 c(1:3)])[,1],yr=sort(rep(year-mean(year),1217)))
dim(olsdat)
names(olsdat)
olsfit1=lm(t~cp+yr,data=olsdat)
summary(olsfit1)
####################

temp=temp-summary(olsfit1)$coef[1,1]  ## center temp

temp_w=t(F_s_star)%*%temp
percent_missing_w=t(F_s_star)%*%cloudp
years=matrix(year-mean(year),nrow=1216,ncol=63,byrow=TRUE)


### preparing data for STAN
N=dim(A)[1]-1   ## number of spatial random effects
n=N*63
temp_w=stack(as.data.frame(temp_w))[,1]
percent_missing_w=stack(as.data.frame(percent_missing_w))[,1]
years=stack(as.data.frame(years))[,1]
cell=rep(1:N,63)


data=c("N","n","temp_w","percent_missing_w","years","cell","F_s_star","lam_star")


### running model in STAN
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

ear_champ1080=stan(file="stan_model_July18_iear_cloudpercent_Champ_model1080.stan",dat
a=data,chains=4,iter=10000,control=list(adapt_delta=0.99,max_treedepth=20),refresh=500
)
```

APPENDIX B: STAN MODEL CODE FOR CHAMPLAIN ANALYSIS

```
functions {
  vector transform_znorm(vector znorm,vector lam,real theta,real sigma, int N){
  vector[N] eigen;
  for(j in 1:N) eigen[j] <- (1 - lam[j])^(-theta/2);
  return sigma * eigen .* znorm;
 }
}

data {
   int<lower=1> N;                      // number of spatial random effects  (spatial locations - 1)
   int<lower=1> n;                      // N*(number of time points)
   vector[n] temp_w;                    // Temp in CX10 after pre-whitening
   vector[n] percent_missing_w;         // prewhitened
   vector[n] years;                     // just centered
   int<lower=1> cell[n];                // grid id number
   matrix[N+1,N] F_s_star;              // eigenvectors of A-D+I (except for first eigenvector)
   vector[N] lam_star;                  // eigenvalues of A-D+I (except for lam=1)
}

parameters {
   real beta_1;                 // effect of missing values (missing values set to 0)
   real<lower=1e-5> sigma_y;
   real<lower=1e-5> sigma_z_0;
   real<lower=1e-5> sigma_z_1;
   real<lower=0> theta_0;
   real<lower=0> theta_1;
   vector[N] z_norm_0;                          // standard normal
   vector[N] z_norm_1;                          // standard normal
}


model {
vector[n] mu;
vector[N] z_star_0;
vector[N] z_star_1;
beta_1 ~ normal(-226,100);             // for 2160m used normal(-227,100);
sigma_y ~ lognormal(2.5,0.75);         // for 2160m used lognormal(1,1.3)
sigma_z_0 ~ lognormal(0,1);            // for 2160m used  lognormal(0.5,0.5)
sigma_z_1 ~ lognormal(-1,1.5);         // for 2160m used  lognormal(0.5,0.5)
theta_0 ~ lognormal(1,0.5);
theta_1 ~ lognormal(1,0.5);
z_norm_0 ~ normal(0,1);
z_norm_1 ~ normal(0,1);
z_star_0 <- transform_znorm(z_norm_0,lam_star,theta_0,sigma_z_0,N);
z_star_1 <- transform_znorm(z_norm_1,lam_star,theta_1,sigma_z_1,N);

mu <- beta_1*percent_missing_w + z_star_0[cell] + z_star_1[cell].*years;
temp_w ~ normal(mu, sigma_y);

}

generated quantities{
vector[N+1] omega_0;
vector[N+1] omega_1;
omega_0 <- F_s_star*transform_znorm(z_norm_0,lam_star,theta_0,sigma_z_0,N);
omega_1 <- F_s_star*transform_znorm(z_norm_1,lam_star,theta_1,sigma_z_1,N);
}
```