SPATIOTEMPORAL MODELS FOR EXPLORING VARIABILITY IN SCALLOP CONDITION ACROSS THE BAY OF FUNDY

by

Jiaying Liu

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
Dec 2023

Mindfulness.

Contents

List of Tables

List of Figures

Abstract

Sea scallops (*Placopecten magellanicus*) comprise the fifth largest fishery in Canada, the vast majority of which occurs in the Maritimes. To ensure the long-term sustainability of the scallop fishery, fisheries scientists provide essential information to DFO, including annual scallop biomass, enabling the dynamic adjustment of fishing policies to maintain a healthy scallop population. Measuring scallop meat weights is more challenging and time-consuming compared to measuring their shell heights. As a result, a Length-Weight Relationship (LWR) is commonly used to estimate scallop meat weights based on their shell heights. However, both meat weight and shell height exhibit temporal and spatial variability. The original LWR lacks the ability to comprehensively account for both aspects of variability, resulting in predictions that lack spatiotemporal accuracy. Consequently, we have developed the Joint Weight Height Model (JWHM) to enhance the foundational LWR by effectively addressing the intricacies of spatial and spatiotemporal variations in both meat weight and shell height. The JWHM is formulated within the Spatiotemporal Model (STM) framework to capture these variations through a Matérn Gaussian Markov Random Field (GMRF). This model accommodates the potential influence of environmental variables including depth, temperature, salinity, and stress, which can impact the both scallop meat weight and shell height.

Our goal is to propose a JWHM to improve available estimates of scallop meat weights in the Bay of Fundy. The resulting JWHM uncovers intriguing patterns related to scallop conditions and significantly improves current predictions of scallop meat weight in the Bay of Fundy.

List of Abbreviations Used

BNAM	Bedford Institute of Oceanography North Atlantic Model	14, 15, 17
CV	Cross Validation	40–42, 62
DEM	Digital Elevation Model	17
DFO	Department of Fisheries and Oceans Canada	viii, 2, 3, 8, 11, 14, 17, 55, 62, 63
GLM	Generalized Linear Model	29, 38, 39
GLMM	Generalized Linear Mixed Model	29, 30, 43
GMRF	Gaussian Markov Random Field	viii, 28, 31–33, 36, 43, 61
GRF	Gaussian Random Field	30–32

HMC	Height Model Component	45–49, 51, 56, 60–62
JWHM	Joint Weight Height Model	viii, 44, 45, 55, 56, 58–63
LWR	Length-Weight Relationship	viii, 3, 4, 19, 26, 29, 45, 61
MSPE	Mean Square Prediction Error	41–43, 51, 62
MWSH	Meat Weight Shell Height	8–11, 13, 14, 16, 17, 19, 21–24, 26, 49, 55

PCC	Pearson Correlation Coefficient	vi, 24, 25
RQR	Randomized Quantile Residual	28, 38, 39, 43, 47, 60, 61
SCV	Stratified Cross Validation	28, 42, 43, 49, 51, 60, 62
SH	Shell Height	12–14, 16, 17, 22, 23, 26, 49
SHF	Shell Height Frequency	11, 12, 26
SM	Spatial Model	49, 51, 62
SPA	Scallop Production Areas	3, 7, 8, 63
SPDE	Stochastic Partial Differential Equation	31–33, 35

STM	Spatiotemporal Model	viii, 5, 28, 30, 31, 35, 37, 43, 45, 49, 51, 55, 62, 63
STM-D	Spatiotemporal Model incorporating Depth	52, 54, 60, 62
TAC	Total Allowable Catch	7
TMB	Template Model Builder	35
UTM	Universal Transverse Mercator	33
WMC	Weight Model Component	45–49, 51, 55, 56, 58–62

Acknowledgements

I would like to express my heartfelt gratitude to Dr. Joanna Mills Flemming and Dr. Orla Murphy, my dedicated supervisors, whose unwavering support and encouragement inspired me to approach my research project with creativity and determination. Their guidance has been indispensable, and I am truly grateful for the invaluable insights they provided. I am grateful for the time and effort they dedicated to engaging in thoughtful meetings with me, contributing to the refinement of my ideas and the overall quality of my thesis.

I extend my sincere thanks to Jessica Sameoto from Fisheries and Oceans Canada for generously sharing her knowledge and patiently addressing my inquiries. I also want to acknowledge the exceptional contributions of Dr. David Keith and Dr. Mike Dowd, whose insightful feedback significantly enhanced the quality of this document.

I am indebted to three individuals who played important roles in my research journey. Firstly, I extend my heartfelt thanks to Yihao Yin for generously providing his expertise in programming methodologies, which enriches my understanding in this crucial area. Secondly, my sincere appreciation goes to Balagopal Pillai for his invaluable technical assistance. His technical assistance was instrumental in successfully completing complex programming projects within tight deadlines. Lastly, I would like to extend my sincere appreciation to Janice MacDonald Eddington for her patient and meticulous assistance. Her insightful discussions and valuable writing suggestions greatly enriched and professionalized my thesis.

Gratitude is also extended to my friends and family for their unwavering emotional support. Their encouragement proved to be a beacon of strength during challenging moments, motivating me to persist when faced with perplexing obstacles.

I would like to express my sincere thanks to the Ocean Frontier Institute (BEcoME Project) for providing the funding that has been instrumental in supporting my pursuit of a master's degree.

Finally, I acknowledge and commend my own unwavering dedication and resilience throughout this academic journey. Overcoming obstacles and earning this degree required immense effort, and I am proud of the determination I exhibited.

Chapter 1

Introduction

Sea scallops (*Placopecten magellanicus*) play an important role for the economy in fisheries in Atlantic Canada as they comprise the fifth largest fishery in Canada, the vast majority of which occurs in the Maritimes (Fisheries and Oceans Canada, 2021).

Figure 1.1: Sea scallops (*P. magellanicus*).

Photo credit: Jessica Sameoto, 2023

Indeed, overfishing stands as a significant contributor to the decline in fish stocks and the destruction of marine habitats (Fisheries and Oceans Canada, 2009). This practice can lead to unhealthy growth patterns in fish populations and, in severe cases, result in increased mortality rates. Currently, the earth's marine ecosystems are

Figure 1.2: Sea scallop fishing vessel from the Maritimes.

Photo credit: Jessica Sameoto, 2023

subject to excessive fishing pressure, sometimes surpassing their capacity to maintain healthy fish populations. This distressing trend has led numerous species perilously close to extinction. The consequences of overfishing extend beyond ecological implications, as they also inflict substantial economic losses upon the fishery industries (ECO, 2018). For instance, during the 1990s, Newfoundland and Labrador experienced a catastrophic collapse of the Atlantic Canadian cod fishery, an event widely recognized as a result of overfishing (Myers et al., 1997). This devastating collapse had profound repercussions on the Atlantic regional economy (Higgins, 2009).

To ensure the long-term sustainability of the scallop fishery, DFO has implemented regulations under the Atlantic Fishery Regulation (Canada, 2018). These regulations are designed to protect overexploitation of scallops and minimize the impact on the seabed caused by excessive trawling (Branch, 2022).

Fisheries science entails a rigorous scientific process of analysing available data to provide decision makers with the necessary information (e.g., maximum sustainable yield) to make reasoned choices (Canada, 2021). Fisheries scientists use mathematical and statistical models called stock assessment models to analyze and understand the

impact of fishing and environmental factors on fish stocks (NOAA Fisheries, 2023).

Length-Weight Relationship (LWR) plays a crucial role in fisheries science as they offer valuable information about the growth, general well-being, and fitness of fish species in marine habitats (Jisr et al., 2018). By analyzing LWRs, experts can better understand the growth patterns, condition, and overall status of fish populations, enabling more informed decision-making in sustainable fisheries management.

DFO has conducted an annual tow survey to monitor the Atlantic sea scallop biomass in the Bay of Fundy since 1981. The survey design follows a stratified random approach, with strata defined to align with historical areas of fishing effort (Fisheries and Oceans Canada, 2017). Higher fishing effort is assumed to correlate with better scallop habitat. In this design, tows are randomly distributed within each survey stratum within the Scallop Production Areas (SPA) during each sampling event (Fisheries and Oceans Canada, 2017). For all tows, the shell heights are measured and recorded in 5 mm size bins. However, weighing scallops is more challenging and time-consuming than simply measuring their heights. DFO aims to sample approximately half of all tow locations to measure both the meat (adductor muscle) and the exact shell height (Yin et al., 2022). However, various factors including weather conditions and vessel limitations, may result in fewer than half of the tows being sampled on occasion. Given that only a subsample of scallops caught in the field are actually weighed, having knowledge of the LWRs becomes crucial for estimating scallop abundance. LWRs provide a valuable tool to estimate the weight of scallops based on their length, allowing researchers to approximate the weight of the entire population without weighing each individual. To be more specific, by measuring the lengths of a representative subset of scallops and applying established LWRs, it becomes possible to estimate the weights of the entire sample accurately. This approach not only saves time and resources but also enables researchers to obtain essential weight data for

population assessments, growth analyses, and fisheries management decisions. By understanding and utilizing LWRs, scientists can derive important insights into the health, condition, and size structure of scallop populations, even when direct weighing of all individuals is not feasible.

Environmental factors play an important role in the LWRs of sea scallops. To be more specific, scallop growth can be influenced by environmental factors, including depth and salinity, which can be related to the quality of their living habitats (Silina, 2023).

Depth can affect factors such as light availability and substrate composition, which in turn can impact scallop growth and survival (Côté et al., 1993). Bottom salinity levels also play a crucial role, as wild scallops may have specific osmotic gradient requirements to maintain physiological balance with their surroundings (Urbina & Glover, 2015). Furthermore, bottom temperature can be an influential factor for most fish species, including sea scallops. Generally, increasing temperatures can lead to higher growth rates, up to a certain threshold that varies for each species. Once this critical limit is reached, growth rates may decline abruptly (Lindmark et al., 2022). Bottom stress is of interest as well, as it is believed to exert a mechanistic influence on benthic communities for the majority of sea species (Jackson-Bué et al., 2022).

Scallops are filter-feeding bivalves, and their habitat preferences are closely tied to environmental factors that influence the availability of suitable food sources (Brand, 2006). For instance, depth and temperature can also influence the phytoplankton growth and phytoplankton forms the base of the marine food web (Palomares-Garcia et al., 2006). Enhancing phytoplankton growth can potentially improve the sufficiency of food source for scallops (Kong et al., 2022).

In the field of modern statistics, Spatiotemporal Model (STM) has emerged as regarded framework for modeling data that exhibits correlations in both time and space. When it comes to understanding the trends of scallop meat weight over time, using a fixed shell height to predict the meat weight has proven valuable (Yin et al., 2022). However, in order to accurately reflect the scallop population in different areas in the Bay of Fundy, it is crucial to improve meat weight predictions by incorporating the spatial variability in shell height. Our primary objective is to construct a joint STM that will account for the variability in both scallop meat weight and shell height within the study area of the Bay of Fundy. By using this model, some interesting patterns can be discovered about the condition of scallops and can improve current estimates of scallop meat weight for this area. Additionally, we investigate whether the inclusion of environmental variables can improve the estimates from the joint STM.

This chapter serves as an introduction to the research, we have now introduced our scientific of interest. Chapter 2 describes the available data and provides data visualizations of the variables of interest. Chapter 3 provides the background of statistical methodology. It also discusses the procedures used to compare the resulting models. Moving forward, Chapter 4 presents model fitting and prediction results within the study area. It also provides insights into how well the joint model aligns with the sampled data and its ability to make accurate predictions across the whole study area. Lastly, Chapter 5 discusses the interpretation of the results and future directions of this work.

Chapter 2

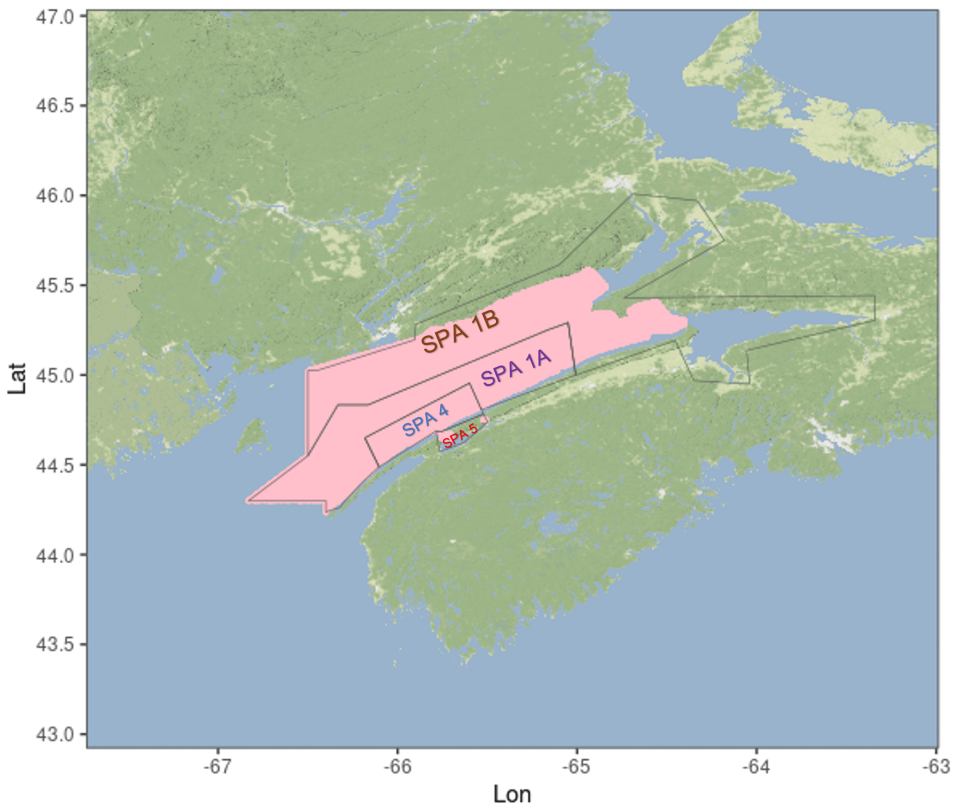
Data Analysis

This chapter describes the available data and provides visualizations of all variables of interest. Section 2.1 provides a detailed description of the study area and the datasets under consideration, offering a comprehensive overview of the foundational aspects of the research. In Section 2.2, the focus shifts to a visual exploration of the data through the use of graphs, aiming to uncover discernible patterns and relationships among the variables of interest.

2.1 Data Description

Since 1981, comprehensive annual surveys of sea scallop (*P. magellanicus*) in the Bay of Fundy and surrounding areas have been consistently carried out (Glass, 2017). These surveys have been conducted aboard vessels operated by both the Canadian Coast Guard and commercial fishing entities. The primary objective of these surveys is to evaluate the biomass within the Scallop Production Areas (SPA) shown in Figure 2.1. These invaluable survey data are a critical asset for DFO Resource Management. These survey data inform scientific recommendations for establishing Total Allowable Catch (TAC) limits in each specific area (Glass, 2017).

Figure 2.1: Fishing management areas in the Bay of Fundy.

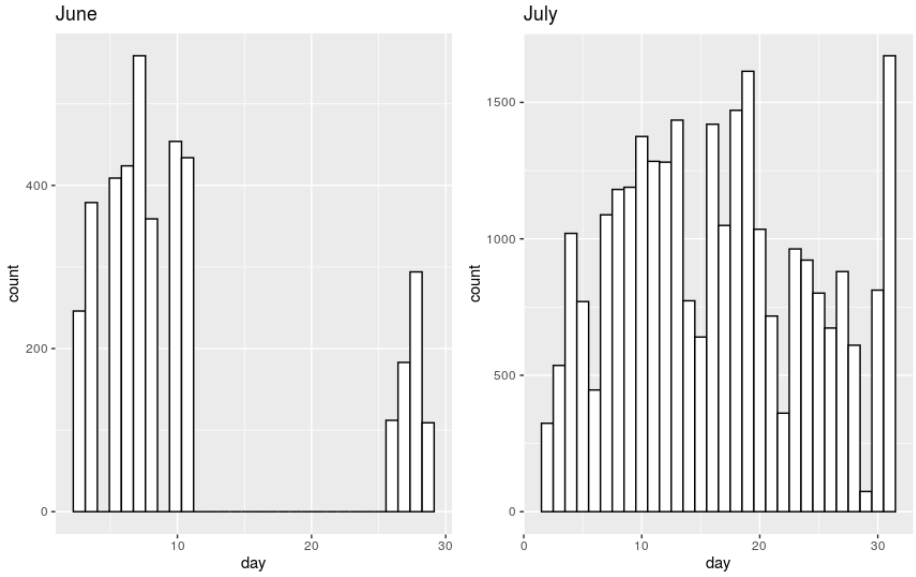


Notes. The study area designated for modeling and predictive analysis is depicted in pink, while the black lines demarcate the boundaries of the scallop fishing management areas encompassing SPA 1A, SPA 1B, SPA 4, and SPA 5.

To sustainably manage scallop populations, it is essential to understand LWRs. In this relationship, length pertains to scallop shell height, while weight corresponds to scallop meat (adductor muscle) weight. DFO conducts an annual Meat Weight Shell Height (MWSH) survey of the Bay of Fundy to monitor changes in scallop LWRs.

There are 33,333 observations in the MWSH survey dataset from DFO Maritimes Region Inshore SPAs spanning 2012 to 2019. There are nine cruises in the dataset: BF2012, BF2013, BF2014, BF2015, BF2016, BF2017, BF2018, BF2019, and RF2012. Of the nine, the last cruise, RF2012, was excluded from the analysis. This exclusion was based on the fact that RF2012 serves as a comparative survey which is not used for investigation. There are no observations from June 12 to June 26, and only the years 2012 and 2015 have recorded observations during the month of June. Due to this discontinuity (Figure 2.2), it was decided to remove the June data from the MWSH dataset. In the end, we have data for the first eight cruises, all of which are from the month of July spanning 8 years. Each year there were around 3,400 to 5,000 scallops captured in July in the Bay of Fundy.

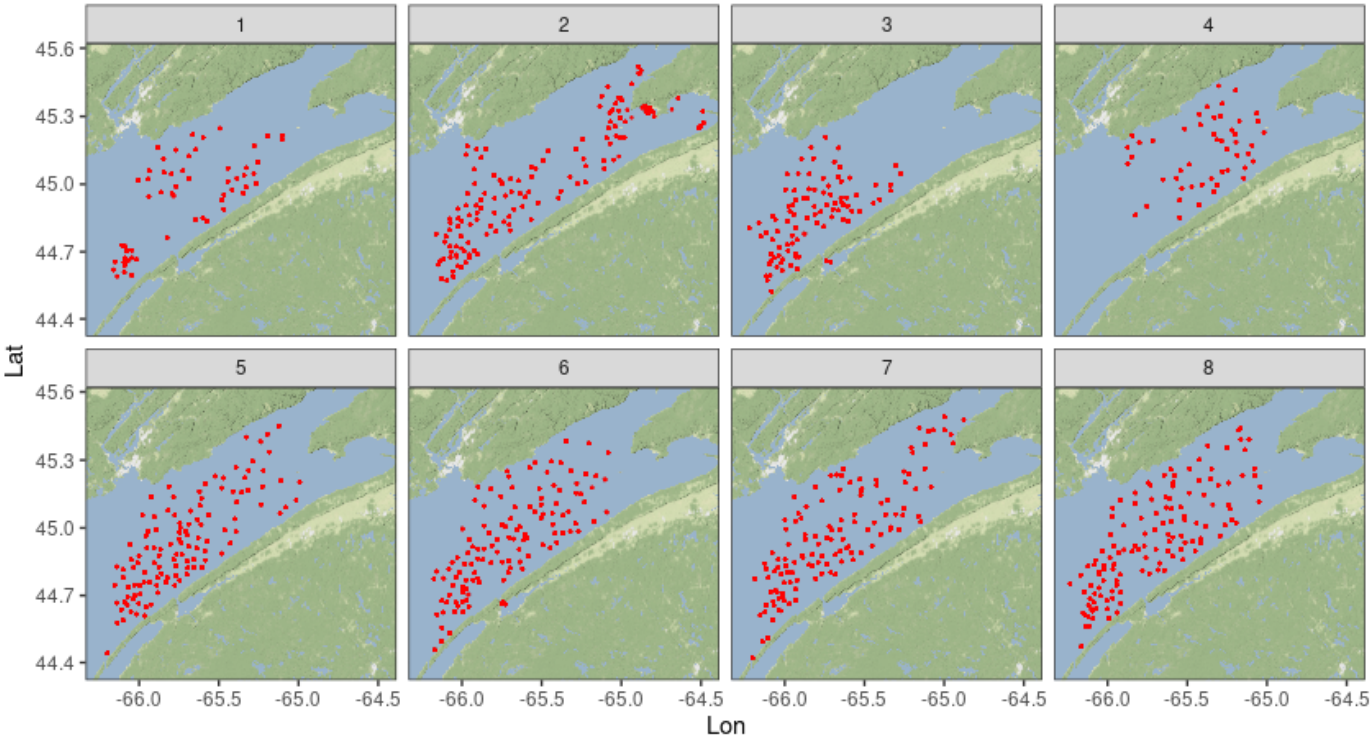
Figure 2.2: 2012-2019 MWSH survey data by month.



Notes. x axis: the day of the month. y axis: the count of scallops.

After cleaning the MWSH survey dataset, there are 28,415 observations located in 826 unique locations across 8 years (Figure 2.3).

Figure 2.3: MWSH survey tow locations in the Bay of Fundy across 2012-2019.



Notes. Red points represent the middle longitude and middle latitude values of each individual tow location.

Table 2.1 presents information for five randomly selected scallop observations from the cleaned MWSH dataset, including year, latitude, longitude, shell height (mm), and meat weight (g). For example, in the first row of the table, this scallop was caught in 2012, measuring 71 mm in height and 6.4 g in weight, where the location of tow 2012_186 was centered at longitude 44.5591° and latitude -66.1296°. The scallop shell heights are carefully measured, rounded to the nearest millimeter, while the corresponding meats undergo a thorough process where excess water is extracted, and they are weighed with precision to the nearest 0.1 g (Yin et al., 2022).

Table 2.1: Five random observations from the MWSH survey data.

year	lon	lat	ID_TOW	weight	height
2012	-65.32970	44.98778	2012_186	6.4	71
2018	-65.12256	45.26342	2018_129	7.6	84
2019	-65.04251	45.18196	2019_69	22.3	150
2016	-65.77654	45.05731	2016_173	12.3	105
2013	-65.12456	45.38990	2013_86	10.3	89

Notes. lon and lat represent the middle longitude and latitude of each tow location. ID_TOW is a categorical variable that identifies the unique tow locations.

It is important to note that in the MWSH data, the selection of tows per year has been carried out through a random-stratified approach from all sampled tows conducted during the study period (Yin et al., 2022). Therefore, the tow locations from the MWSH dataset represent a subset of all tow locations in the Bay of Fundy. In contrast, the Shell Height Frequency (SHF) data provided by DFO is collected at all tow locations throughout the entirety of the Bay of Fundy during the study period. The including of the SHF survey data can give us a better understanding of the variation in shell heights across the Bay of Fundy, thereby enriching our insights into the scallop LWRs in this region.

In the SHF dataset, there are 40 bins (Bin 1, Bin 2,..., Bin 40) for each tow location. These bins serve to categorize scallops according to their shell heights, ranging from 0 to 199.99 mm. Scallops at each of the tow locations are sorted into bins based on their shell heights, with each bin measuring precisely 5 mm in size (Yin et al., 2022). As an illustration, Bin 1 contains scallops with shell heights ranging from 0 to 4.99 mm, while Bin 40 corresponds to scallops with shell heights between 195 and 199.99 mm. To ensure comparability of SHF data between tows, the data are standardized by adjusting for variations in tow size. Specifically, the standardization process involves rescaling the scallop number to correspond to a consistent tow length of 800 m and a tow width of 5.334 m. As an example, consider a tow,

which originally measures 700 m in length and 5 m in width and contains 1 scallop. After standardization to the specified dimensions, the standardized value for this tow would be approximately 1.2 (i.e., $\frac{800\,m}{700\,m} \times \frac{5.334\,m}{5\,m}$).

To model shell height we have to convert the height count data to height measurements. Table 2.2 presents information for five randomly selected scallop observations from the cleaned SHF dataset, including information such as year, latitude, longitude, and shell height interval. For instance, there are 1.7 scallops at tow location 2018_13 and their shell heights are between 100 and 104.99 mm (see the highlight row in Table 2.2). In this context, 1.7 scallops can be understood as representing the presence of one scallop, accompanied by a 0.7 probability of another scallop being present. To determine the count of scallops along with their respective shell height values, we employed the *rbinom (Binomial random variate generator)* R function, which yielded two scallops. To obtain the precise shell heights, we made the assumption that within each shell height interval, all possible shell height outcomes have an equal probability. This assumption implies a uniform distribution of shell heights within each bin. Therefore, we utilized the *runif (pseudo-random number generator)* R function to assign shell height values of 103.7028 and 101.9114 mm to these simulated scallops. By applying the *rbinom* and *runif* functions within the context of 40 bins, the cumulative outcome yields 39 scallops within the 2018_13 location. When this process is extended to encompass all tow locations, it grants us the capability to calculate reasonable scallop quantities and shell heights for each unique tow location. The scallop Shell Height (SH) dataset consists of shell heights resulting from this procedure.

Table 2.2: Five random observations from the SHF data.

Year	ID_TOW	lon	lat	[0, 4.99]	[5, 9.99]	[10, 14.99]	...	[100, 104.99]	...	[195, 199.99]
2017	2017_308	-64.95479	45.25436	0	0	0	...		...	0
2016	2016_272	-65.08351	45.33167	0	0	0	...		...	0
2013	2013_355	-66.08446	44.60441	0	0	0	...	0	...	0
2018	2018_13	-66.07152	44.67362	0	0	0	...	1.7	...	0
2012	2012_128	-65.63013	44.86039	0	0	0	...		...	0

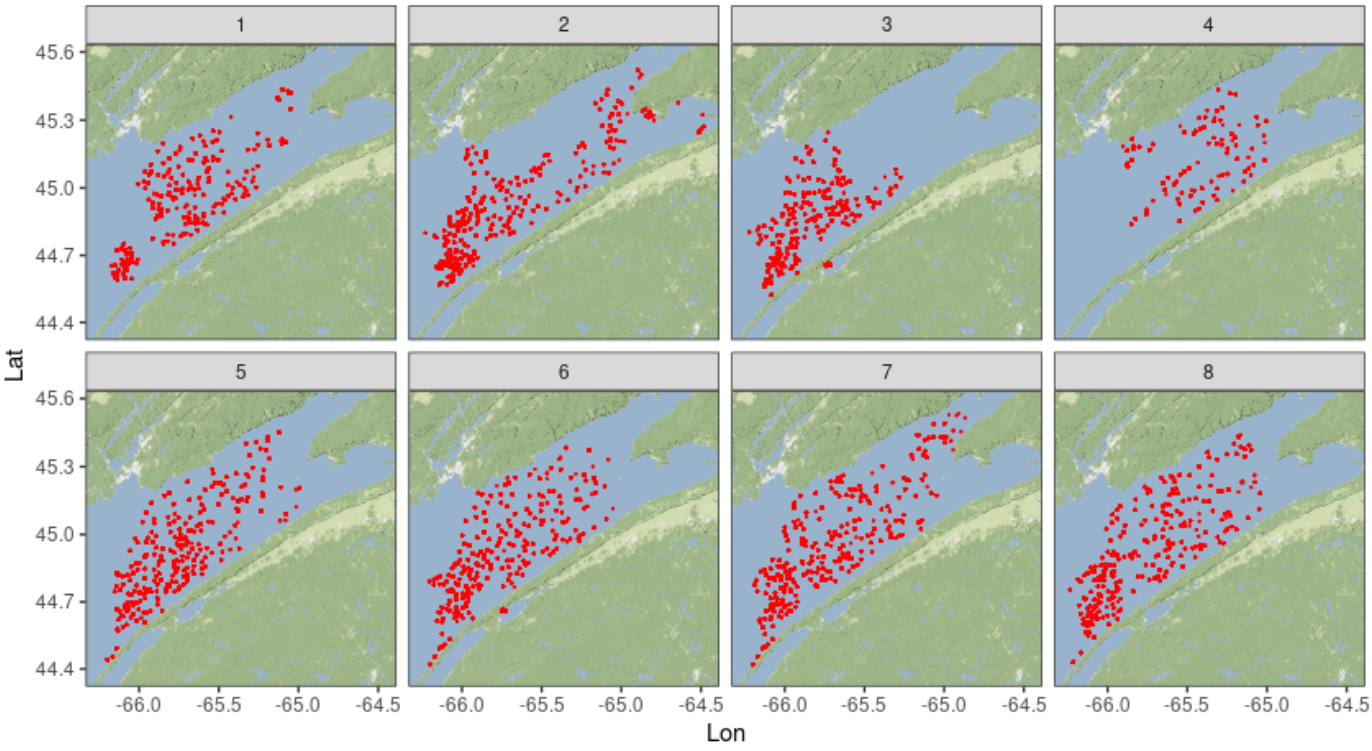
Similar to the MWSH dataset, across the entire SH dataset, scallops are grouped by their respective tow locations. Each location in the SH dataset contains scallops from some bins in order. As an illustration, Table 2.3 provides an overview of the recording structure based on the first five observations at tow location 2018_13 within the SH dataset. At tow location 2018_13, the first two scallops are assigned to Bin 21 (encompasses scallop shell heights ranging from 100 to 104.99 mm) with respective shell heights 103.7028 and 101.9114 mm. Following that, the subsequent two scallops are attributed to Bin 22 (covers shell heights in the range of 105 to 109.99 mm) with respective shell heights 109.0785 and 105.5080 mm.

Table 2.3: The first five observations at 2018_13 from the SH dataset.

year	lon	lat	ID_TOW	height
2018	-66.07152	44.67362	2018_13	103.7028
2018	-66.07152	44.67362	2018_13	101.9114
2018	-66.07152	44.67362	2018_13	109.0785
2018	-66.07152	44.67362	2018_13	105.5080
2018	-66.07152	44.67362	2018_13	113.5786

The SH dataset comprises 50,880 observations, distributed over 2,791 unique locations across 8 years (Figure 2.4). Notably, the tow locations in the MWSH dataset constitute a subset, comprising approximately one third of the tow locations within the SH dataset.

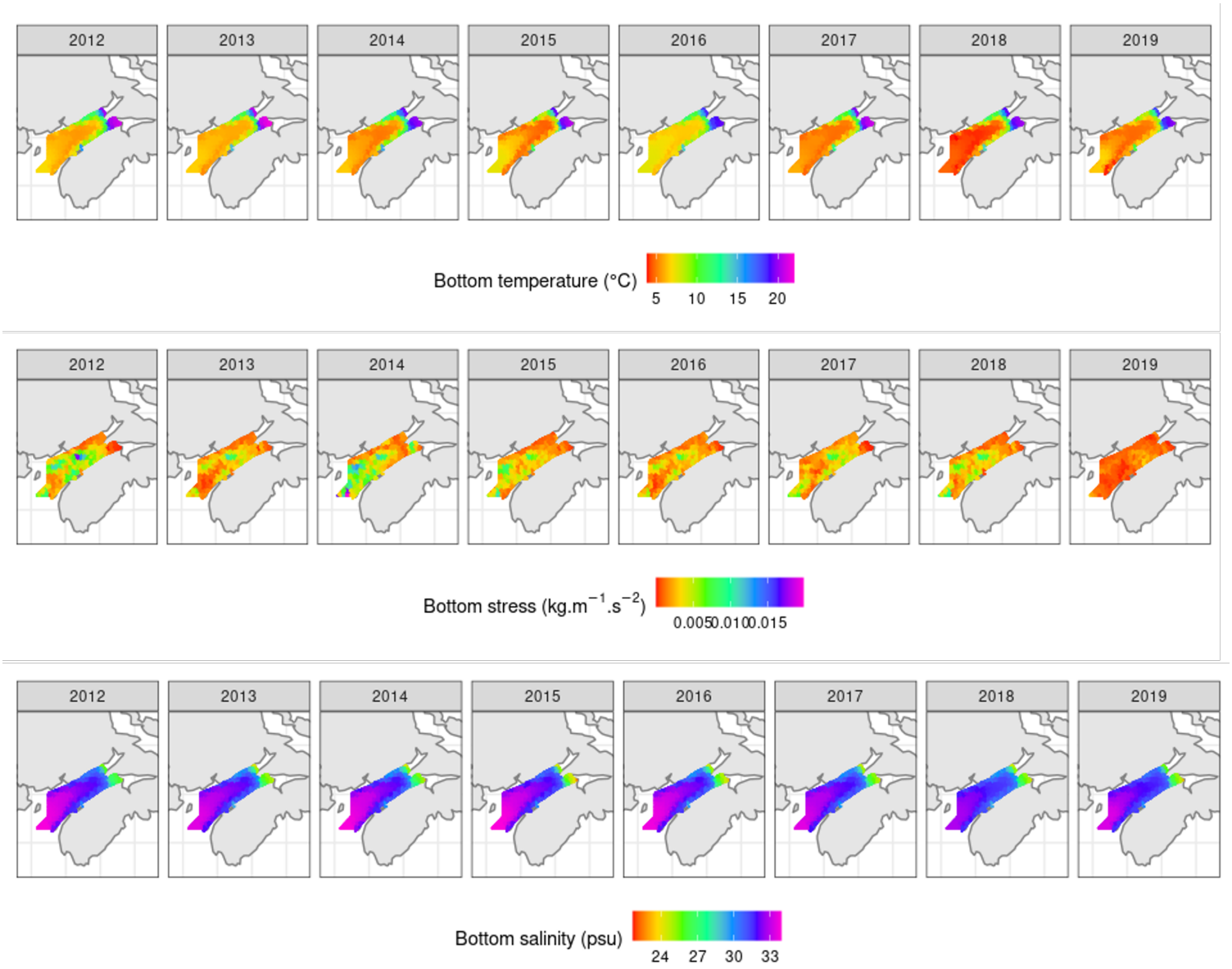
Figure 2.4: SH observations of the sampled sea scallop tow locations in the Bay of Fundy across 2012-2019.



The Bedford Institute of Oceanography North Atlantic Model (BNAM) has been developed with the purpose of assisting various DFO monitoring programs (Wang et al., 2018). It accomplishes this by offering a comprehensive dataset that includes hindcast simulations and future climate projections of a range of different variables (Wang et al., 2018). The relevant BNAM outputs are bottom temperature, stress, and salinity. These data are represented as monthly averaged raster values from 2012

to 2019, each with a spatial resolution of 7 km (Figure 2.5). Additionally, it is important to note that these BNAM data are spatiotemporally indexed, meaning that the value of each data point is associated with both a specific year and tow location. BNAM is designed with a focus on the North Atlantic region, making it well-suited for broad-scale applications. It has proven to be suitable in the Scotian Shelf and the outer Gulf of Maine. However, the current version of the BNAM model is of coarse resolution for coastal areas and does not include tides (Wang et al., 2018). Since the Bay of Fundy is strongly dominated by tidal circulation, the variable bottom shear stress should be interpreted with caution.

Figure 2.5: BNAM temperature, stress, and salinity.

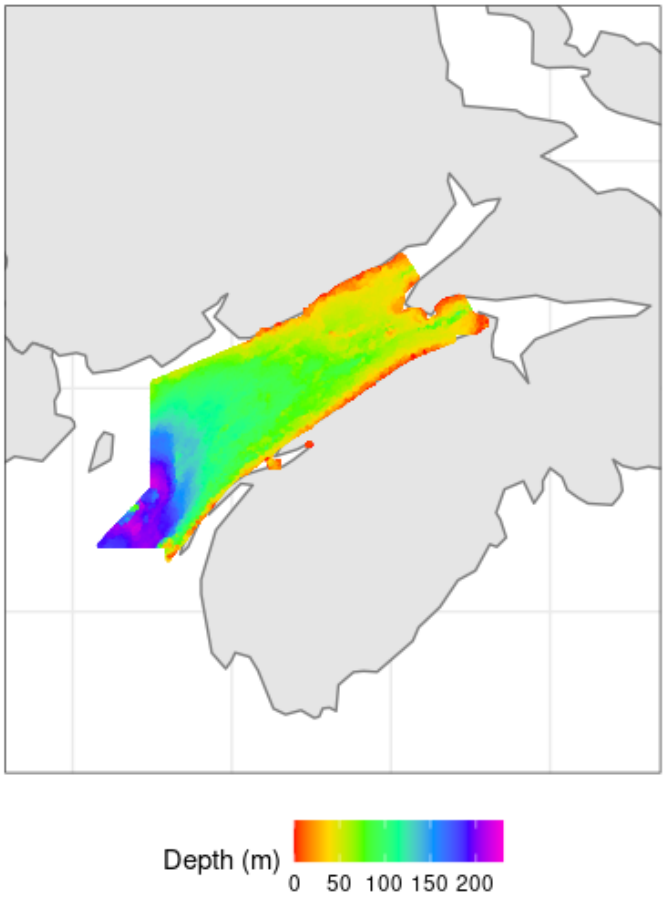


In Figure 2.5, it is evident that the spatial patterns temperature and stress variations vary across different years. However, when it comes to salinity, the temporal variations are less conspicuous; in the majority of locations in the region of interest, salinity remains relatively stable across years. To integrate these pertinent environmental variables into our MWSH and SH datasets, we employ the *extract* R function that spatiotemporally extracts the bottom temperature, stress, and salinity values

from the BNAM dataset.

The Digital Elevation Model (DEM) is another valuable tool for DFO (Davies et al., 2019), facilitating spatial analysis by providing high-resolution depth data standardized to a mean water level with a 50 m resolution (Davies et al., 2019; Glass, 2017), see Figure 2.6. A region with a depth value below 50 meters is defined as a shallow water region. SPA 5 (see Figure 2.1) is situated within a shallow water region. Given that depth data is collected spatially, we can effectively utilize the *extract* R function to extract the depth values from the DEM dataset and integrate them into our MWSH and SH datasets, based on the respective tow locations.

Figure 2.6: DEM depth.



The final set of environmental covariates is listed in Table 2.4.

Table 2.4: List of environmental variables

Environmental Variable	Data resource	Resolution
Depth (m)	DEM	50 m
Bottom temperature(°C)	BNAM	7 km
Bottom stress ($\mathrm{kg.\,m^{-1}.\,s^{-2}}$)	BNAM	7 km
Bottom salinity (psu)	BNAM	7 km

2.2 Variable Visualization

In biology, the "Cube Law" refers to a principle related to the scaling of biological organisms, particularly the relationship between their size and certain physiological parameters (Gayon, 2000). This principle is also applied in the context of the LWR in fisheries science. The contemporary form of the LWR was established by Keys (1928), linking length, L, and weight, W, in the field of fisheries science, which is expressed as

$$W = a\,L^{b} \tag{2.1}$$

The nonlinear LWR can be observed in the left plot of Figure 2.7.

Estimating coefficients (a and b) in a non-linear regression (Equation 2.1) is more challenging than in a linear regression. To transform this relationship into a linear form, the logarithmic equivalent relationship, denoted as Equation 2.2, is commonly employed in fishery analysis. After performing a log transformation on both the height and weight variables in the MWSH dataset, the relationship between these two variables becomes approximately linear (the right plot of Figure 2.7) and the scales for weight and height become closer.

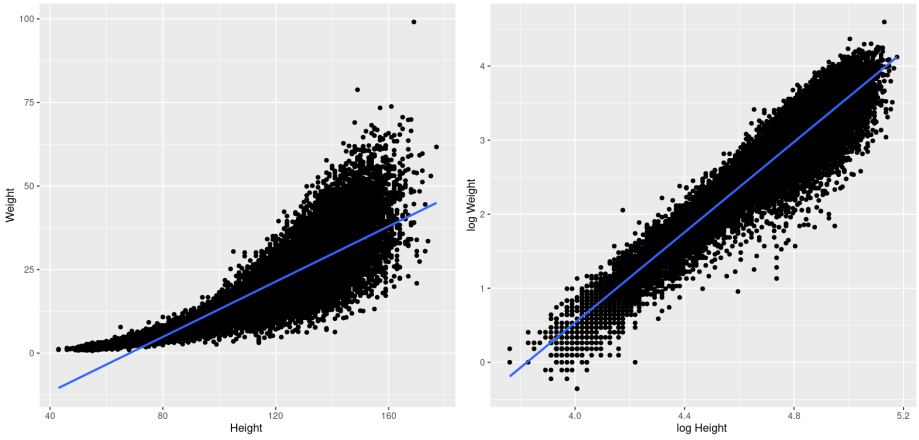
$$log(W) = log(a) + b\,log(L) \tag{2.2}$$

Where $log(a)$ represents the intercept, and b represents the slope, also called growth rate (Jisr et al., 2018).

While, allometric growth in this context refers to the disproportionate increase

or decrease in weight in relation to length as the fish grows (Gayon, 2000). In Equation 2.2, if b is less than 3, it suggests negatively allometric growth, indicating that the fish, on average, become lighter as they grow longer. If b is greater than 3, it indicates positively allometric growth, suggesting that the fish, on average, become heavier as they grow longer (Mazumder et al., 2016). If b being equal to 3, it represents isometric growth due to the geometric properties of three-dimensional objects (Ricker, 1959).

Figure 2.7: The LWRs for scallop meat weights and shell heights in the MWSH dataset.



Notes. The original LWR (left), the log LWR (right).

Figure 2.8 presents the relationship between environmental variables and weight in the MWSH dataset and Figure 2.9 shows the relationship between environmental variables and height in the SH dataset. These relationships are not apparent, as they lack a distinct ascending or descending pattern, unlike the relationship between weight and height in the MWSH dataset. Given that the MWSH dataset employs log transformation for both weight and height, it is reasonable to consider applying a similar log transformation to the environmental variables within the same dataset. However, even after applying log transformations for environmental variables and weight, as depicted in Figure 2.10, the relationships between weight and environmental variables still appear somewhat ambiguous. Nevertheless, it is worth noting that the log transformation has the beneficial effect of narrowing the ranges for environmental variables, bringing them closer in scale to that of the log height in the MWSH dataset.

Figure 2.8: Weights versus environmental variables in the MWSH dataset.

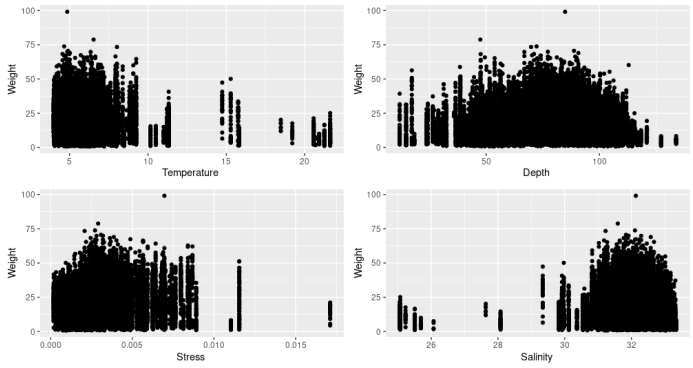


Figure 2.9: Heights versus environmental variables in the SH dataset.

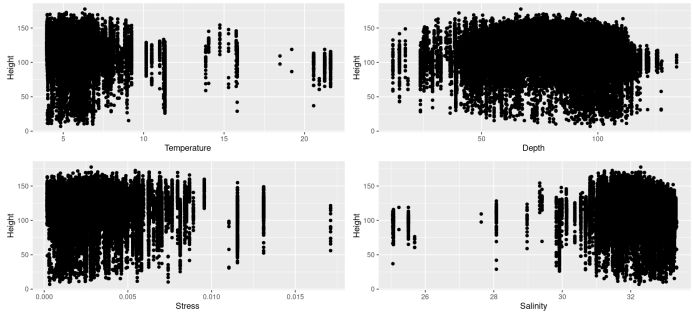


Figure 2.10: Weights versus environmental variables after log transformations in the MWSH dataset.

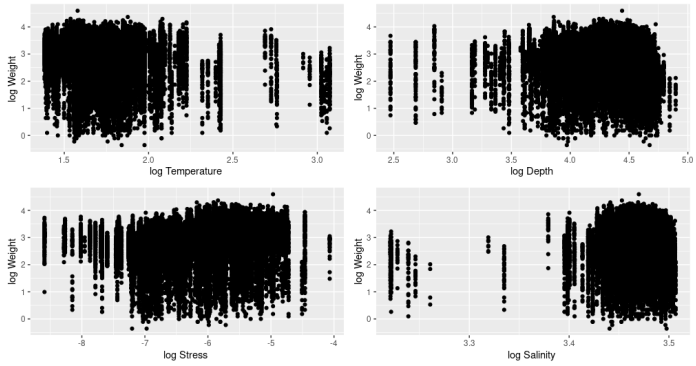
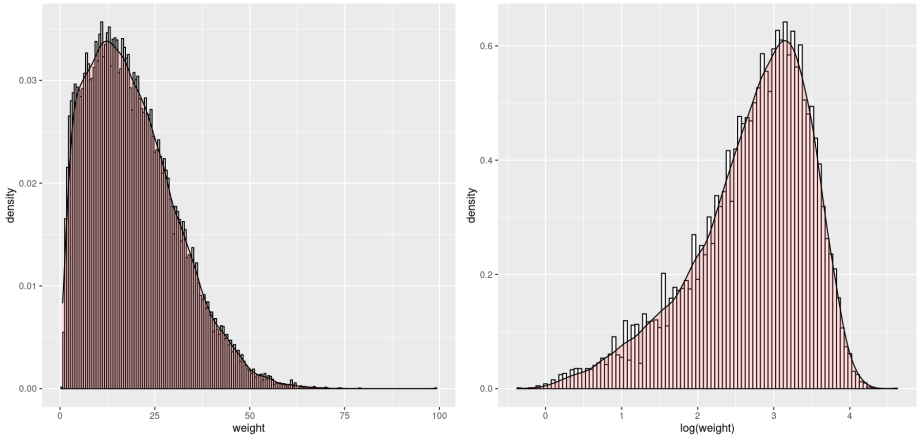


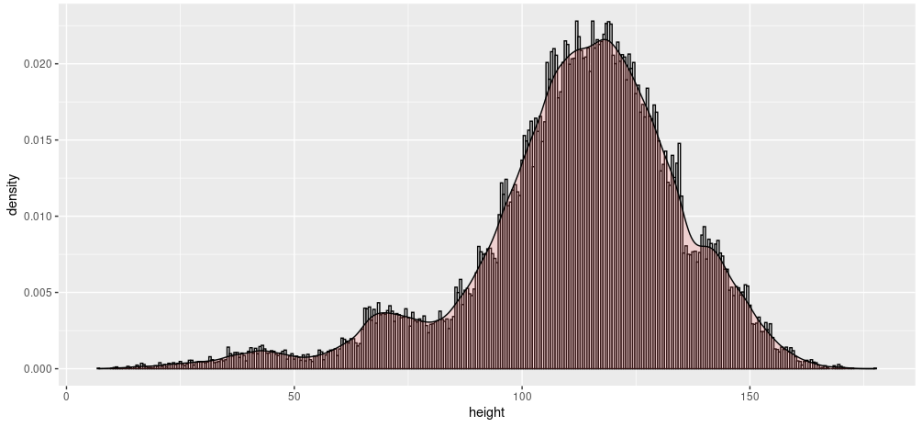
Figure 2.11 illustrates the empirical histograms of scallop meat weights in the MWSH dataset, both before and after undergoing a log transformation. Prior to this transformation, the weight histogram is heavily right-skewed, while after the log transformation, it becomes more symmetric with a notable heavy tail. This adjustment through a log transformation suggests a potential mitigation of skewness for the original weight distribution. Figure 2.12 shows the empirical histogram of scallop shell heights in the SH dataset. The histogram exhibits a heavy-tailed distribution of scallop shell heights.

Figure 2.11: Histograms of weight and log weight from the MWSH dataset.



Notes. The original weight histogram (left), the weight histogram after a log transformation (right).

Figure 2.12: Histogram of height from the SH dataset.



In Figures 2.13 and 2.14, we can discern the spatial distributions of mean meat weights and mean shell heights derived from the MWSH dataset and SH dataset, respectively. Over the span of eight years, SPA 1A, SPA 4 and SPA 5 have heavier and/or larger scallops on average compared to SPA 1B.

Figure 2.13: Mean meat weights calculated from the MWSH dataset.

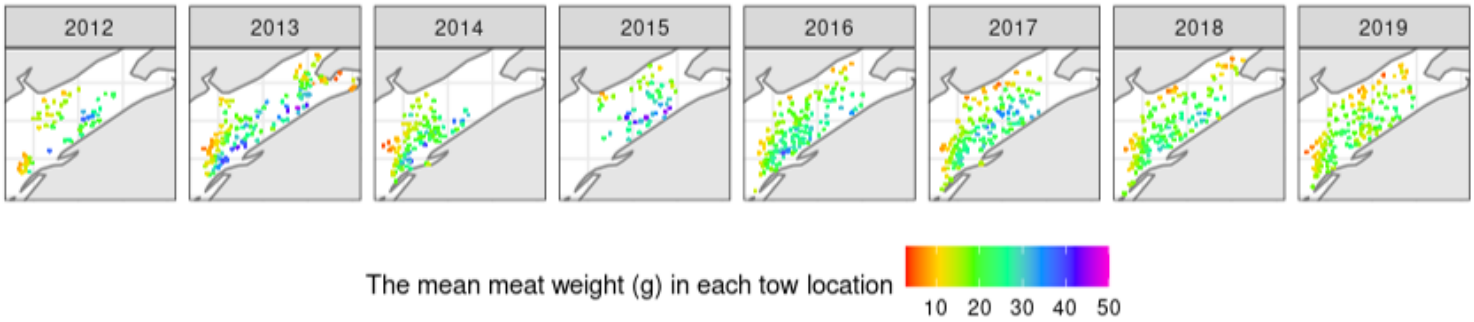
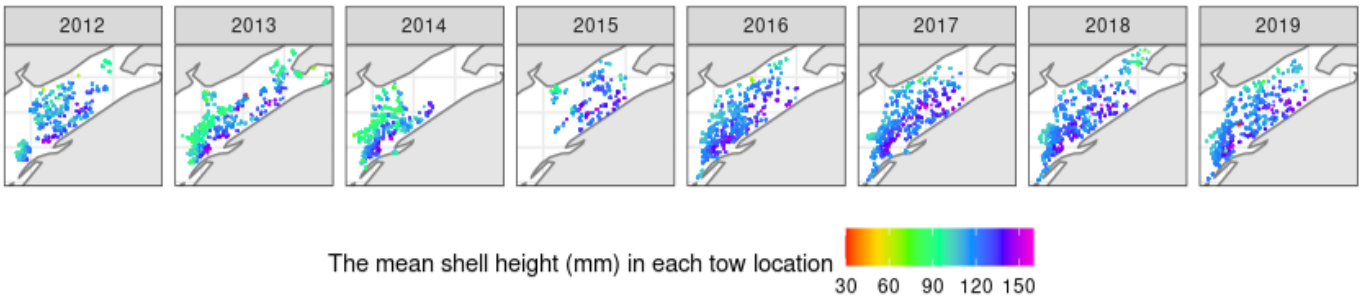
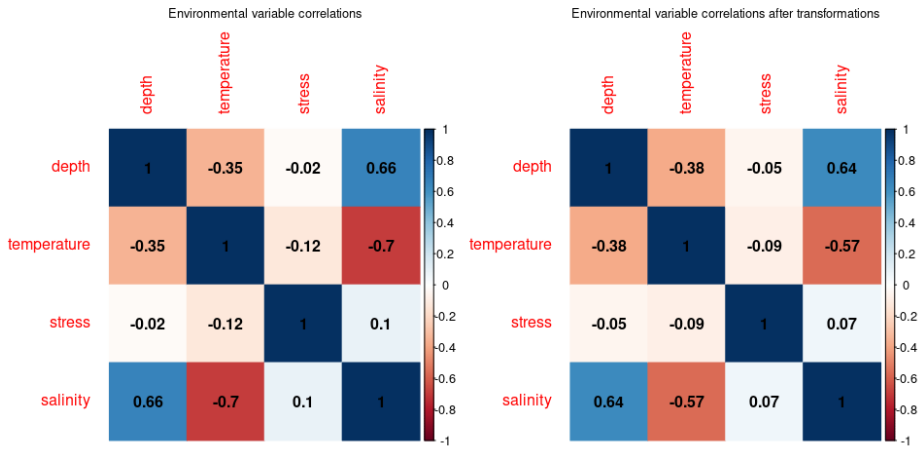


Figure 2.14: Mean shell heights calculated from the SH dataset.



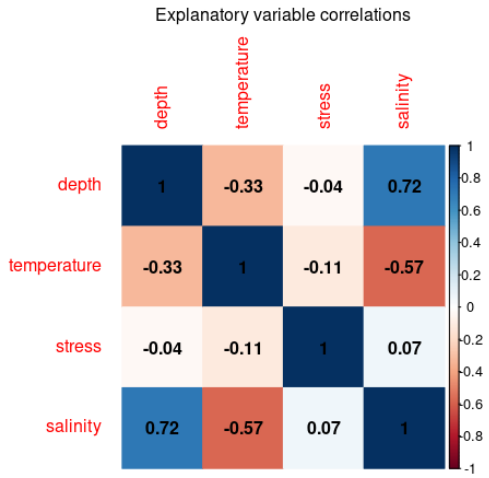
We compute the Pearson Correlation Coefficient (PCC) for each pair of environmental variables before and after transformations and draw heat maps (Figures 2.15 and 2.16). PCC, ρ, ranges between -1 and 1. If $|\rho| \geq 0.7$, in this case, we say there is a strong linear correlation between x_1 and x_2 (Ratner, 2009). In Figure 2.15, it is evident that there is a strong correlation (ρ = -0.7) between salinity and temperature before the log transformation. However, it is noteworthy that this strong correlation in the MWSH dataset is eliminated following the log transformation. Meanwhile, in Figure 2.16, a strong correlation ($\rho = 0.72$) emerges between salinity and depth.

Figure 2.15: The PCCs between environmental variables in the MWSH dataset.



PCCs between original variables (left) and PCCs between log transformed variables (right).

Figure 2.16: The PCCs between environmental variables in the SH dataset.



2.3 Summary

In this chapter, we introduced our study area and datasets. Our designated study area is situated within the Bay of Fundy, encompassing four SPAs. The data for our analysis are sourced from DFO and comprise both the MWSH and SHF survey datasets from Maritimes Region Inshore SPAs spanning the years 2012 to 2019. The MWSH dataset contains the count of scallops captured in each tow with their corresponding shell heights and meat weights. However, in the SHF dataset, only the scallop counts per height bin are provided for each tow location. In the first section of this chapter, we described the procedure for converting count data into approximate shell heights to form the SH dataset.

Next, we provided a comprehensive review of the relationship between scallop shell height and meat weight, as well as the correlations among environmental variables (depth, temperature, stress, and salinity) within the dataset. The adoption of the logarithmic LWR is common practice in fisheries science due to its ability to establish a closely linear relationship between shell height and meat weight. Our histograms revealed that both the log-transformed weight from the MWSH dataset and height from the SH dataset are heavy-tailed distributions. It is worth noting that in the MWSH dataset, a strong correlation exists between salinity and temperature. While in SH dataset, a strong correlation exists between salinity and depth. However, through the application of log transformation to all environmental variables in the MWSH dataset, the previously observed strong correlations appear to reduce to moderate levels.

Overall, this chapter provided a comprehensive description of all the datasets utilized throughout this thesis and presented crucial data visualization outputs. In Chapter 3, we describe methodologies for appropriately analysing such spatiotemporal

data. Moving forward to Chapter 4, we embark on the modeling phase, leveraging the insights gained from these datasets and visualizations to inform our analysis.

Chapter 3

Statistical Methodologies

This chapter describes the statistical methodologies we use for analysing the LWRs of sea scallops in the Bay of Fundy. Section 3.1 provides an introduction to STMs with an emphasis on the Gaussian Markov Random Field (GMRF) as a random effect correlation structure. A comprehensive explanation of the STM fitting and prediction processes using R project is also provided. Section 3.2 is the procedures of model fitting. Section 3.2.1 explains Randomized Quantile Residual (RQR) and how it is used to validate model assumptions and Section 3.2 unveils concepts of k-Fold Stratified Cross Validation (SCV) as a measure of model predictive performance.

3.1 The Spatiotemporal Model Framework

Ecological data are recognized for their spatial and temporal variability. In recent years, there has been a shift towards capturing such spatiotemporal patterns by incorporating random effects into statistical modeling frameworks, which frequently leads to the development of mixed-effects models (Thorson & Minto, 2015). Generalized Linear Mixed Model (GLMM) has now become a cornerstone in ecological research, facilitating a deeper understanding of complex ecological processes and enhancing the quality of scientific investigations in the field (Bolker et al., 2009). GLMM not only inherits the advantages of Generalized Linear Model (GLM), such as handling non-normal data, but also expands its ability to include both fixed and random effects.

The generic GLMM formulation (Yin et al., 2022) is defined as

$$\begin{aligned} g(\mathbb{E}(y)) &= X\alpha + Z\beta, \\ y &\sim D, \\ \beta &\sim \Theta \end{aligned} \tag{3.1}$$

Here, y is the response variable, and g is a link function relating the expected response to explanatory variables. X and Z are design matrices containing data on the explanatory variables, and α and β are the coefficients of the fixed and random effects, respectively. β is assumed to follow a generic distribution Θ. D denotes the response distribution for y and can be any exponential family distribution.

Incorporating geographical variations in the living conditions of sea scallops is essential for accurately estimating LWRs (Thorson & Haltuch, 2019). Furthermore, fisheries data are often collected through time. Those changes in space and time are

known as spatiotemporal variation (Thorson et al., 2015).

When dealing with data collected across both space and time, basic GLMMs may lack the ability to capture the spatiotemporal variation. To solve this problem, STMs have emerged as a powerful tool (Mailman School of Public Health, 2016). STMs incorporate spatial and spatiotemporal variability. They provide a comprehensive framework for analyzing how the variable of interests under study evolve and interact across different space and time. In essence, STM can be seen as a specific type of GLMM with a correlation structure that is based on random effects in both space and time. By utilizing STMs, researchers can gain deeper insights into the intricate interplay between environmental conditions, sea scallop populations, and other relevant factors, leading to more accurate and robust assessments of sea scallop LWRs.

Correlating the random effects in STMs using the Gaussian Random Field (GRF) is a widely adopted and proven method that significantly enhances prediction capabilities (Thorson & Haltuch, 2019). A random field $\{Z(s) : s \in S\}$ is a GRF, if for all choices of points within the region of interest S, $Z(s)$ has a multivariate Gaussian distribution, which can be characterized by its mean $\mu(s)$ and its covariance matrix $\Sigma(s)$.

In many fisheries science studies, the observations collected from the field represent only a sample of the entire spatial domain. In such cases, making the most efficient use of the geographical information from the sampled observations becomes critical. Indeed, Tobler's first law of geography states "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 240), which emphasizes the fundamental principle of spatial correlation. This law is a cornerstone concept in geography and spatial analysis and has implications in

various fields, including fisheries science. The covariance function of a GRF must be specified. The Matérn correlation functions are commonly used in geostatistics. For a random field $Z(s)$, the Matérn covariance function can be expressed as

$$\Sigma(s_i, s_j) = Cov(Z(s_i), Z(s_j)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||s_i - s_j||)^\nu K_\nu(\kappa||s_i - s_j||) \tag{3.2}$$

where, K_ν is the modified Bessel function of the second kind, σ^2 is the marginal variance parameter, $\kappa > 0$ is the scale parameter, and $\nu > 0$ is the smoothness parameter.

Generally speaking, ν is chosen before estimating other parameters in an STM, and $\rho = \frac{\sqrt{8\nu}}{\kappa}$ is the distance between two locations s_i and s_j, where the spatial autocorrelation diminishes to 0.1.

GRFs are continuous and relatively convenient for spatial and spatiotemporal modelling, but they can be computationally expensive. The "big N problem" often arises when we have a large number of sample locations, because it requires an $N \times N$ covariance matrix factorization (Jona Lasinio et al., 2012). To tackle this problem, the Stochastic Partial Differential Equation (SPDE) approach proves to be a valuable solution (Lindgren et al., 2011). This method allows us to transform the continuously indexed GRF, $Z(s)$, into a discretely indexed GMRF by means of finite basis functions defined on a Delaunay triangulation of the region, commonly referred to as a mesh (Krainski et al., 2021).

A GMRF is a Gaussian vector in N dimensions characterized by a mean, μ, and a precision matrix (inverse of the covariance matrix), Q. In a GMRF, conditional independencies and dependencies among vertices are determined by the absence or presence of edges connecting them (Sidén & Lindsten, 2020). Vertices that lack an

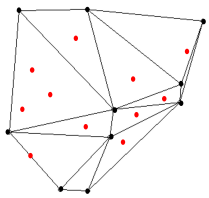
edge between them are identified as non-neighbors, and as a consequence, they are defined as conditionally independent of each other. While, vertices that are connected directly by an edge are identified as neighbors, and as a consequence, they are defined as conditionally dependent of each other. This inherent characteristic gives rise to a sparse precision matrix Q, offering significant computational advantages compared to dealing with a dense covariance matrix (Sidén & Lindsten, 2020).

Figure 3.1 illustrates a mesh structure, given a random sample of 11 observations. Equation 3.3 reflects the relationship between GRFs and GMRFs; that is, $Z(s)$ is a weighted average of the GMRF in the nodes of the triangles containing the location of the observation.

$$Z(s) = \sum_{i=1}^{M} \phi_i(s) Z_i, \tag{3.3}$$

Where, $\phi_i(.)$ represents a piece-wise polynomial basis function for each triangle, Z_i is zero-mean Gaussian distributed, and M is the number of nodes in the mesh.

Figure 3.1: An example of a Delaunay triangulation.



Notes. The red points represent the sampled observations and the black points are the nodes of triangles.

The INLA R package (Rue et al., 2023) is one of the software packages used widely in applying the SPDE approach, which combines analytical approximations and numerical integration to approximate posterior distributions. In this package,

the *inla.mesh.create* function is used to create a mesh for a group of the observations. Fisheries data are typically collected with geographic information, including longitude and latitude, which falls under the unprojected system, referencing observation locations on the episode earth in decimal degrees (DD) or degree, minutes, and seconds (DMS). In this unprojected system, calculating the spatial distance between two observations is impractical. Consequently, before creating a mesh for spatial observations, it is essential to project their geographical locations on the Universal Transverse Mercator (UTM) coordinate system (Moraga, 2019). This transformation can be automatically accomplished using the PBSmapping R package (Schnute et al., 2022).

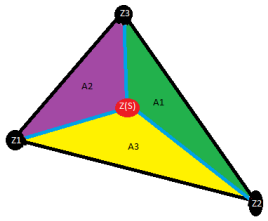
The SPDE method defines a mesh that creates an artificial set of neighbours over the study area by calculating the spatial autocorrelation among observations to ensure that non-neighboring components remain conditionally independent (Belmont, 2022). This property greatly enhances computational efficiency. By leveraging GMRFs, we can significantly expedite computations, as we only need to factorize the sparse precision matrix Q, which closely approximates the inverse of the Matérn covariance matrix Σ (Lindgren et al., 2011).

The use of the associated projection matrix is a powerful tool for interpolating any location in the field of interest. Once the mesh has been established, the projection matrix A can be computed by a translation of spatial locations on the mesh into corresponding vectors in the model (Moraga, 2019). In A, the number of rows matches the total sampled locations, N, and the number of columns precisely corresponds to the number of vertices, M, in the Delaunay triangulation.

A simple illustration of the interpolation process in a random field is shown in Figure 3.2. In this Delaunay triangulation, the random field only contains one

triangle and s represents any observation in the triangle region S.

Figure 3.2: An example of a 3-vertex mesh.



Notes. The red dot indicates the specific location we are interested in estimating, while the black points represent Delaunay nodes (Z_1,Z_2, and Z_3). The blue lines divide the Delaunay triangulation into three areas, referred to as A_1, A_2, and A_3.

Applying Equation 3.3,

$$Z(s) = \frac{A_1}{A_1+A_2+A_3} Z_1 + \frac{A_2}{A_1+A_2+A_3} Z_2 + \frac{A_3}{A_1+A_2+A_3} Z_3.$$

Written in matrix form,

$$Z(s) = A \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix}$$

where A is the projection matrix defined by

$$A = \begin{bmatrix} \frac{A_1}{A_1+A_2+A_3} & \frac{A_2}{A_1+A_2+A_3} & \frac{A_3}{A_1+A_2+A_3} \end{bmatrix}$$

Rather than manually creating a projection matrix for interpolation, the sdmTMB R package (Anderson et al., 2023) provides an automated solution for constructing STMs and performing predictions within the region of interest using the SPDE method. Template Model Builder (TMB) (Kristensen et al., 2023) is used to speed up the optimization process by using the Laplace approximation and automatic differentiation.

Using the same parameter definitions as Equation 3.1, the general STM form based on the SPDE approach can be defined as

$$
\begin{aligned}
g(\mathbb{E}(y)) &= X\alpha + Z\beta, \\
y &\sim D, \\
\beta &\sim GMRF(0, Q)
\end{aligned}
\tag{3.4}
$$

where, $GMRF(0, Q)$ represents the GMRF process, and Q is the precision matrix, which is the inverse of the Matérn covariance Σ .

3.2 Assessing Model

3.2.1 Randomized Quantile Residuals

In fisheries science, it is common to assume that certain continuous response variables, such as size and weight, follow a normal distribution (Guy & Brown, 2007). To assess the validity of this normality assumption, a simple graphical tool - the quantile-quantile plot, often referred to as the QQ plot (Ford, 2015), is used to visually inspect whether such an assumption is sensible.

When non-normal patterns emerge within observations, such as heavily skewed data, it becomes untenable to assume a normal distribution. Furthermore, in the linear modelling context, if the connection between predictor variables and the response variable is non-linear, the normality assumption is not suitable either (Kumar, 2022). For instance, the logarithmic equivalent LWRs mentioned in Chapter 2.

The STM framework offers a solution to address these complexities and accommodates exponential family distributions for the response and relationships including non-linear ones between the response and predictors. The exponential family distributions is defined as

$$p(y, \theta) = h(y) \exp(\eta^T(\theta) T(y) - B(\theta)) \tag{3.5}$$

where $p(y, \theta)$ is the probability density function of variable y with parameter θ. $h(y)$ and $T(y)$ are functions that can only contain y, while $\eta^T(\theta)$ and $B(\theta)$ are functions that can only contain θ.

The response residuals from a STM are often not normally distributed, thus, a

straightforward application of a QQ plot yields limited insights, rendering it less advantageous for assessing the distributional characteristics of these residuals. Pearson residuals and deviance residuals are commonly used when dealing with GLMs. However, assessing these residuals is challenging, as they deviate from a typical normal distribution, both marginally and conditionally. Residual QQ plots exhibit several parallel curves to the x-axis, further complicating their interpretation, as there is no standard reference distribution available for comparison (Feng et al., 2020).

Dunn and Smyth (1996) defined RQRs and demonstrated their applications with GLMs. Due to randomization, the quantile residuals are able to maintain continuity (Dunn & Smyth, 1996). This property allows them to be especially well-matched with exponential family distributions. Moreover, RQRs can be computed faster and interpreted more easily than other alternative forms of residuals, which makes them a favorable choice for statistical analyses in GLMs (Dunn & Smyth, 1996). For continuous response variables, a general form of the RQR can be simplified to

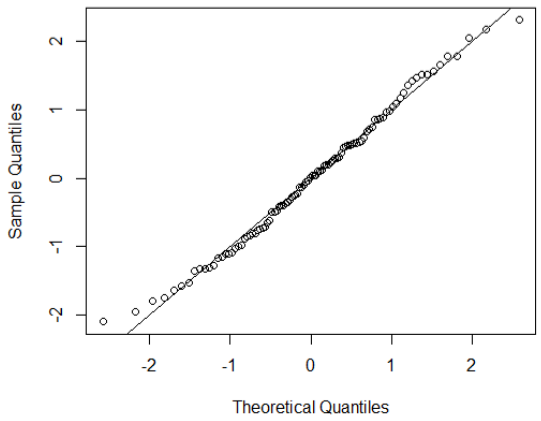
$$r_i^q = \Phi^{-1}(F(R_i \leq r_i)) \tag{3.6}$$

where $R_i = h_i(y_i, \hat{u_i})$, is the crude residual, defined by Cox and Snell (1968), in correspondence with a certain relationship function h_i, that relates the actual response value y_i to the estimated mean $\hat{u_i}$. F is the cumulative distribution function of the response defined by the assumed model (and estimated parameters) and Φ is the cumulative distribution function of the standard normal. Apart from sampling variability in mean and variance, the r_i^q converges to a standard normal distribution if all model parameters are consistently estimated (Dunn & Smyth, 1996).

The *qqnorm* R function is used to create RQR plots (Dunn & Smyth, 2018) for

GLMs. Similar to the conventional QQ plot for normal distributions, the RQR plot (Figure 3.3) retains the standard normal distribution's theoretical quantiles along the x-axis. Conversely, the y-axis portrays the RQRs as defined in Equation 3.6. In most instances, when the model's distributional assumptions hold approximately true, the RQR plot tends to an approximately straight line. This alignment signifies the concordance between the model's assumptions and the empirical data, validating the model's appropriateness.

Figure 3.3: An example of a RQR plot.



Notes. The line goes through the first and third quartiles.

3.2.2 k-Fold Stratified Cross Validations

Cross Validation (CV) is a statistical technique employed to evaluate the predictive performance of a proposed model (Great Learning Team, 2020). This method entails a random division of the dataset into distinct training and test subsets. The model is fitted using the training data, then its performance is evaluated on the test data. A model might achieve an excellent fit with the training data by increasing its complexity, which involves adding more parameters, but it could exhibit poor performance when evaluated with the test data. Such models are often referred to as overfited models, as they exhibit low bias in the training data but high variance in the test data. Conversely, a model may adopt an overly simplistic approach, failing to capture important features within the training dataset, but performing reasonably well in the test data. These models are termed underfitting models, characterized by high bias in the training data but low variance in the test data (Huilgol, 2020).

An optimal model strikes a balance between overfitting and underfitting, effectively managing the trade-off between bias and variance (Hali, 2022). In the case of such optimal models, both training and test errors are typically minimized (Singh, 2018). By evaluating the training error of the model, we can obtain a precise assessment of the bias incurred during the training process (Rai, 2020). Conversely, measuring the testing error of the model provides us with a precise understanding of its predictive accuracy and generalization ability (Shah, 2020).

Balancing variance and bias remains a challenge even when utilizing CV. The variability in data splitting can highly affect the model evaluations obtained during CV, given that each data point can be selected either for training or testing.

To overcome this limitation, a k-fold CV algorithm can be employed. By implementing the k-fold CV algorithm, every data point is utilized as both training and

test data. As a consequence, the overall testing results become independent of the specific data partition. In this algorithm, the data observations from a given dataset are randomly split into k folds: each fold contains approximately the same number of data observations. In each iteration, $k-1$ folds of the data observations are chosen to be the training set, the remaining fold becomes the test set. There are k separate iterations of testing (Pandian, 2022). The predictive performance is evaluated by averaging the squared prediction errors across all data observations in the given dataset after the k iterations of testing, which is known as the Mean Square Prediction Error (MSPE) (Arnholt, 2021). A general form of MSPE is

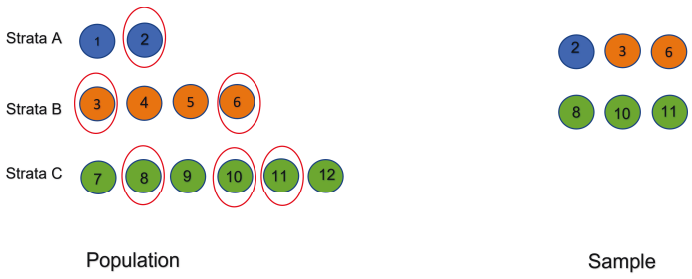
$$MSPE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y_i})^2 \tag{3.7}$$

where n is the total number of the data observations in the given dataset, y_i is the observed value of each data observation, and $\hat{y_i}$ is the estimated value of each data observation.

In k-fold CV, individual observations are randomly assigned to different folds. However, if we want to use k-fold CV to examine a model's spatial predictions then we have to account for the spatial structure across the sampling domain, otherwise it can cause prediction bias (Ploton et al., 2020). To ensure that observations from every location and year are included in each fold proportionally to their occurrence in the total population, one of the widely employed techniques is stratified sampling (Yin et al., 2022). This involves partitioning the entire dataset into distinct strata, defined by the combined attributes of location and year for each observation. The ID_TOW variable, as discussed in Chapter 2, defines our distinct strata. Consequently, during the sampling process, a random selection of sample observations can be achieved across all strata, where the samples are selected in the same proportion

as they appear in the population, ensuring a balanced spatial and temporal representation across all folds (Muralidhar, 2023). Figure 3.4 illustrates an application of stratified sampling. In this example, if the population of interest has two observations from strata A, four observations from strata B, and six observations from strata C, then we randomly choose $\frac{2}{(2+4+6)}$ of the sample from strata A, $\frac{4}{(2+4+6)}$ of the sample from strata B, and $\frac{6}{(2+4+6)}$ of the sample from strata C. If the required sample size is six, then the random selection process will involve one observation from strata A, two observations from strata B, and three observations from strata C. As a result, observation 2, 3, 6, 8, 10, 11, and 12 are randomly chosen from their respective strata.

Figure 3.4: An example of stratified sampling.



Notes. The 12 numbered observations comprise the total population, which is divided into three strata (A, B, and C). Red circles represent the random selection process.

CV through stratified sampling is commonly referred to as SCV. The *createFolds* R function (Kuhn, 2023) can be utilized to generate k folds for a given dataset using stratified sampling. Each fold is uniquely identified by numbers (1, 2, ..., k) for clarity. In order to assign each data observation to its respective fold, a foldID factor can be created in the dataset. This foldID factor then functions as a categorical variable, indicating the fold number to which each data observation belongs. The foldID factor is subsequently used in the *sdmTMB_cv* R function (Anderson et al., 2023) to compute MSPEs through k-fold CV for the models of interest.

3.3 Summary

In this Chapter, we focused on the the statistical methodologies we use for analysing these survey data. We first introduced STMs and demonstrated their connection to basic GLMMs through the GMRF. In contrast to basic GLMMs, STMs are capable of capturing intricate spatiotemporal variations present in data. By utilizing STMs, we can gain a deeper understanding of the relationships between scallop meat weight and shell height, enabling us to attain more precise predictions of scallop meat weight. Additionally, we introduced crucial R packages for modeling STMs. These packages not only streamline the process of fitting our models but also facilitate faster and more efficient predictions, enhancing the overall ease and effectiveness of our modeling efforts. Next, we introduced a valuable tool for validating model distribution assumptions, known as the RQR plot. Typically, when the model's underlying distributional assumptions hold approximately true, the RQR plot exhibits a tendency to form an approximately straight line. Lastly, when comparing the predictive performance in both spatial and temporal dimensions across different models, we employed k-Fold SCV with MSPE, which allows us to compare models without introducing bias either spatially or temporally.

Overall, this chapter summarizes the theoretical foundation that underlies the methodology we apply in Chapter 4.

Chapter 4

Modeling and Results

This chapter is focused on the predictions, derived from our Joint Weight Height Model (JWHM). Section 4.1 describes the model fitting and selection process. Section 4.2 presents and discusses the results of scallop meat weight predictions obtained from the optimal model.

4.1 Model Fitting

In Chapter 2, we observed spatial variability in shell height and meat weight, as illustrated in Figures 2.13 and 2.14, respectively. To extend the original LWR (captured in Equation 2.2), in order to account for both spatial and spatiotemporal variability in shell height and meat weight, thus delivering more precise predictions for meat weight across the Bay of Fundy, we propose the JWHM as

$$p_{W,H}(w\,,h|X,\Theta) = p_{W|H}(w|h,X,\Theta)p_{H}(h|X,\Theta) \tag{4.1}$$

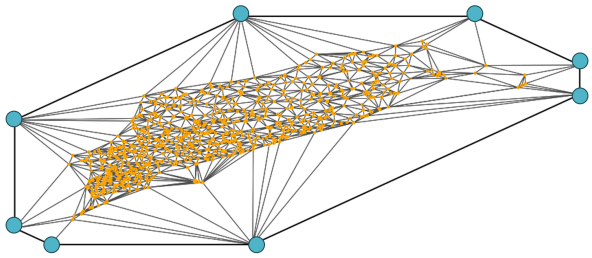
where $p(.)$ is the density function, w is meat weight, and h is shell height. X contains all the fixed effects (time and environmental variables) and Θ contains all the random effects (spatial and spatiotemporal random effects).

The JWHM, $p_{W,H}(w\,,h|X,\Theta)$, is a product of the Weight Model Component (WMC), $p_{W|H}(w|h,X,\Theta)$, and the Height Model Component (HMC), $p_{H}(h|X,\Theta)$. Both WMC and HMC are spatiotemporal models and can be fitted seperately by using the function *sdmTMB* (Anderson et al., 2023) described in Chapter 3. The WMC is designed to specifically capture the variability in meat weight. It serves the purpose of predicting meat weight given a fixed shell height value across both space and time. The HMC is employed to effectively capture the variability in shell height across both space and time.

In Chapter 3, we elucidated that the process of fitting an STM using the SPDE approach necessitates the creation of a triangulation. We also provided a comprehensive guide on how to employ the INLA package (Rue et al., 2023) to generate such a mesh. It is worth noting that this mesh offers various customizable features, allowing us to control triangle sizes and the number of nodes. In our case, we can opt to utilize

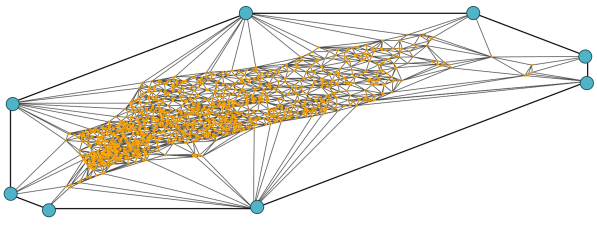
the default mesh, as recommended by Yin et al. (2022). In this default mesh, the nodes are positioned at tow locations and boundary extension points (created automatically by the *inla.mesh.create* function). We created a mesh for the WMC and the HMC, respectively shown as Figure 4.1 and Figure 4.2. Ultimately, the WMC mesh was configured with 834 nodes, while the HMC featured an expanded mesh with 2799 nodes.

Figure 4.1: The Delaunay triangulation used for the GMRF presence of the WMC.



Notes. Orange dots represent the 826 tow locations and blue dots represent the 8 boundary extension points from the MWSH dataset used in the WMC fitting.

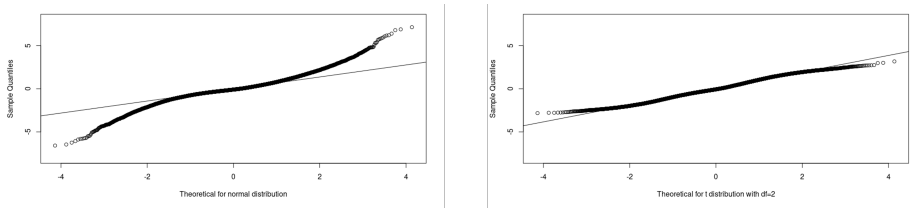
Figure 4.2: The Delaunay triangulation used for the GMRF presence of the HMC.



Notes. Orange dots represent the 2791 tow locations and blue dots represent the 8 boundary extension points from the SH dataset used in the HMC fitting.

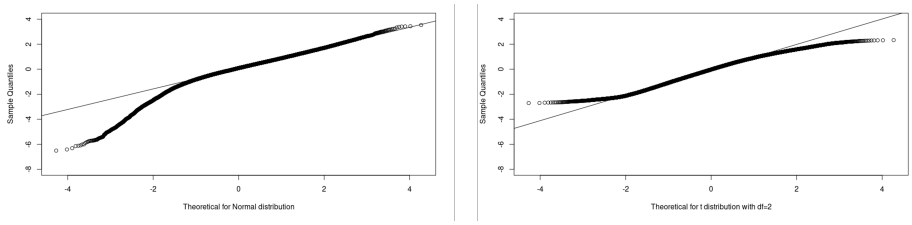
From analyzing the empirical histograms representing log weight and height, we can initially consider a normal distribution for both model components. Notably, both empirical histograms exhibit the potential for heavy tails. In addressing heavy-tailed distributions, the Student t distribution with 2 degrees of freedom, sdt_2, emerges as a suitable choice. Therefore, it is reasonable for us to also assume that both meat weight and shell height adhere to a sdt_2 distribution in both model components. In order to check the distributional assumptions for both model components, we can utilize the RQR plot introduced in Chapter 3. Figures 4.3 and 4.4 depict the distribution comparison between the normal distribution and sdt_2 for the WMC and the HMC respectively. It is evident that sdt_2 outperforms the normal distribution in both model components, as evidenced by the RQR plots, which tend to closely resemble a straight line. This observation underscores the superior performance of sdt_2 in capturing the distribution characteristics of the data.

Figure 4.3: The RQR plots for the WMC.



Notes. The RQR plot for the WMC with a normal distribution (left) and the RQR plot for the WMC with a sdt_2 (right).

Figure 4.4: The RQR plots for the HMC.



Notes. The RQR plot for the HMC with a normal distribution (left) and the RQR plot for the HMC with a sdt_2 (right).

Furthermore, in Chapter 2, we thoroughly explored the benefits of employing log transformations for not only shell height but also for the environmental variables in the weight height relationship. Consequently, in the framework of the WMC, all continuous explanatory variables will be represented in their log scales.

The specific WMC is shown as Equation 4.2 and the specific HMC is shown as Equation 4.3.

$$
\begin{aligned}
p_{w|h}(w_i|h_i, X_i, \Theta_i) : log(\mathbb{E}(w_i)) &= \alpha + \alpha_0\,(t_i) + \beta_0(s_i) + \beta_0(s_i, t_i) + \alpha_1\, log(h_i) + \alpha_2\, log(d_i) \\
&+\alpha_3\, log(c_i) + \alpha_4\, log(e_i) + \alpha_5\, log(f_i) \\
w_i &\sim\ sdt_2, \\
i &= 1, 2..., 28415, \\
\beta_0(s_i) &\sim GMRF(0, Q_1(s)), \\
\beta_0(s_i, t_i) &\overset{\text{i.i.d}}{\sim} GMRF(0, Q_1(s, t))
\end{aligned}
\tag{4.2}
$$

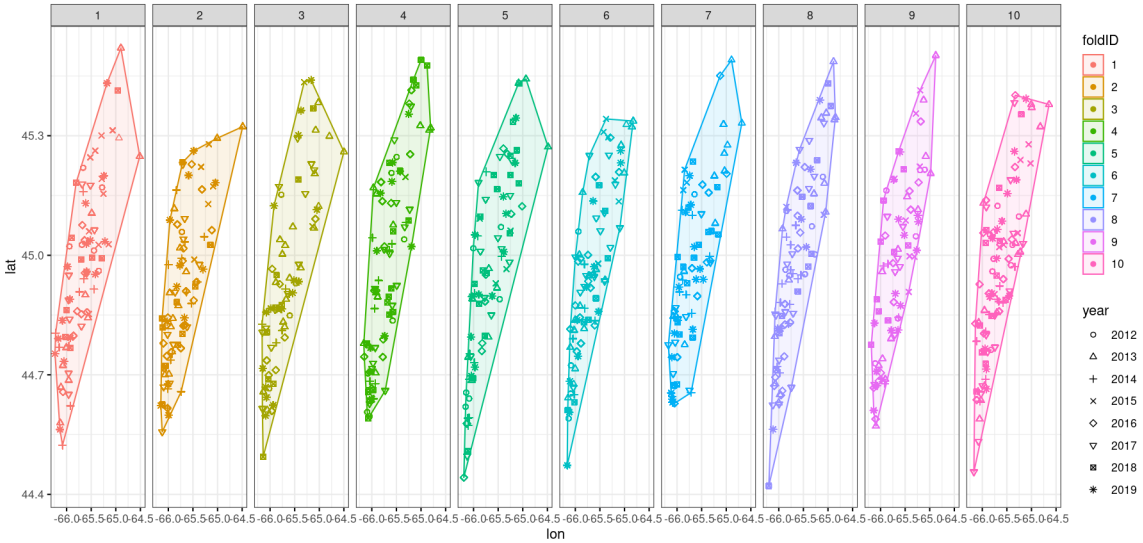
$$
\begin{aligned}
p_{h}(h_j|X_j, \Theta_j) : \mathbb{E}(h_j) &= \gamma + \gamma_0(t_j) + \zeta_0(s_j) + \zeta_0(s_j, t_j) + \gamma_1\, d_j + \gamma_2\, c_j + \gamma_3\, e_j + \gamma_4\, f_j \\
h_j &\sim\ sdt_2, \\
j &= 1, 2..., 50880, \\
\zeta_0(s_j) &\sim GMRF(0, Q_2(s)), \\
\zeta_0(s_j, t_j) &\overset{\text{i.i.d}}{\sim} GMRF(0, Q_2(s, t))
\end{aligned}
\tag{4.3}
$$

where, t is the year (2012-2019), s is the sample tow location, d is depth, c is bottom temperature, e is bottom stress, and f is bottom salinity. $\alpha_0(t_i)$ and $\gamma_0(t_j)$ are the

fixed temporal effects. $\beta_0(s_i)$, $\zeta_0(s_j)$ and $\beta_0(s_i, t_i)$, $\zeta_0(s_j, t_j)$ are the spatial and time-independent spatiotemporal random effects. $Q_1(s)$, $Q_2(s)$ and $Q_1(s,t)$, $Q_2(s,t)$ are the spatial and spatiotemporal precision matrices.

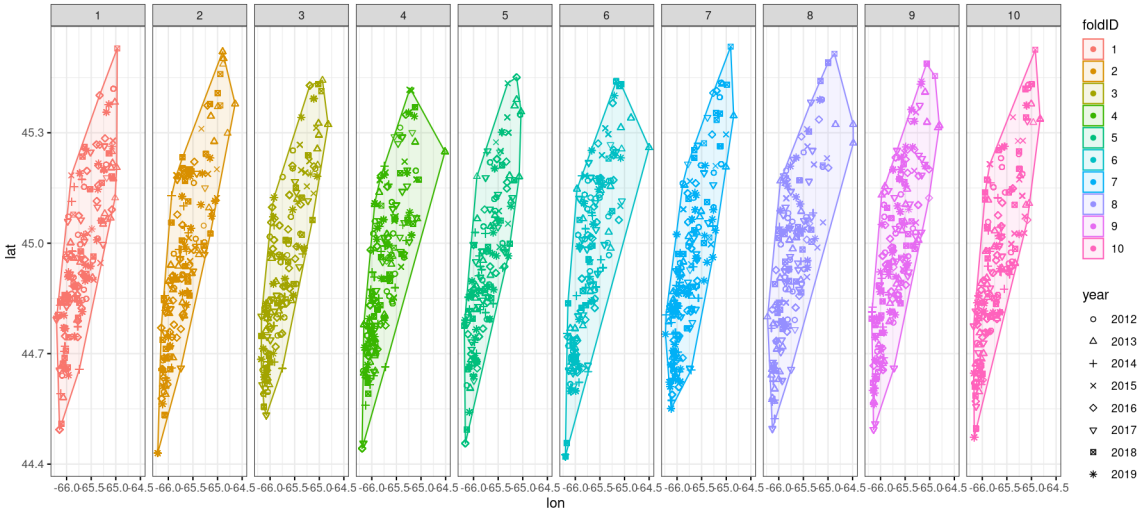
Initially, both the WMC and HMC were assumed to incorporate both spatial and spatiotemporal random effects. Indeed, if a model incorporates only spatial random effects, it is referred to as a Spatial Model (SM). On the other hand, if the model includes both spatial and spatiotemporal random effects, it is termed an STM. To ascertain the significance of these two random effects in both model components, we can employ the 10-fold SCV method. Stratified sampling necessitates dividing the entire survey dataset into distinct strata based on both tow location and year, which is identified as ID_TOW. Within each fold, a random selection of the sampled scallops can be performed, encompassing all ID_TOW categories. Figures 4.5 and 4.6 display the spatial distribution of sample observation locations and the corresponding years within the 10 folds for both the MWSH and SH datasets.

Figure 4.5: Spatial distributions of observations from the MWSH dataset across 10 folds by stratified sampling



Notes. Folds are ordered by numbers (foldID) and distinguished by different colors for clarity. The x-axis represents the longitude of the sample observations, while the y-axis corresponds to their latitude. The shapes of the observation points indicate the year.

Figure 4.6: Spatial distributions of observations from the SH dataset across 10 folds by stratified sampling



These visualizations provide a comprehensive view of the sampling patterns and sample observation distribution across different folds. It is evident that there

are observations selected from each year and nearly every tow location in each fold, which shows representative and unbiased sampling. Therefore, 10-fold SCV will aid us in making an informed decision regarding the choice between an SM or STM by comparing their MSPEs for both WMC and HMC without bias from either spatial or temporal dimensions.

The results reveal that for the WMC , the STM yields an MSPE of 14.7520, while the SM produces an MSPE of 17.2248. In the case of the HMC, the STM achieves an MSPE of 403.6835, and the SM results in an MSPE of 426.8411. There is a reduction of 14.4% and 5.4% in MSPE for the WMC and the HMC, respectively, when transitioning from an SM to an STM The deduction rates for MSPE are greater than 1% in both model components, underscoring that STM models demonstrate superior predictive performance compared to SM models for both components. Also, by examining spatial graphs illustrating the random effects for both model components (as depicted in Figures B.1 and B.2), we discern noteworthy distinctions in these random effects, encompassing spatial as well as spatiotemporal dimensions. Positive spatial random effects predominantly manifest in SPA 1A, 4, and 5, whereas negative spatial random effects tend to be more pronounced in SPA 1B for both the model components. In other words, locations in SPA 1A, 4, and 5 have positive spatial effects and tend to have higher meat weight and shell height predictions on average across the 8 years, while locations in SPA 1B have negative spatial effects and tend to have lower meat weight and shell height predictions on average across the 8 years. Additionally, spatiotemporal random effects exhibit variations across different years for both the WMC and HMC. These variations manifest as distinct differences in the spatial distribution of colors across different years.

We now employ the STM framework to determine the significant environmental variables. Our environmental variables are considered fixed effects, and therefore, we

can use a backward variable selection method to identify the significant environmental ones (with a p-value less than 0.05). The procedure starts with a model containing depth, temperature, salinity, and stress. At each step, we remove the least significant variable from the current model and re-evaluate the model. We continue removing variables until all remaining variables in the current model are deemed significant. Ultimately, this process reveals that depth is the only significant environmental variable for both model components. One potential explanation for this phenomenon is that depth exhibits a significantly higher level of resolution when compared to other environmental variables. Additionally, depth serves as a critical factor influencing ecological dynamics in marine environments. For instance, it affects factors like light availability, which is pivotal for the growth of phytoplankton (Brand, 2006). Phytoplankton, being the foundation of the marine food web, holds particular importance for the growth of scallops (Kong et al., 2022). Another reason might be that the spatial and spatiotemporal effects have already accounted for the variability in the environment, potentially rendering the other environmental variables non-significant in a linear fashion with respect to the two model components.

Furthermore, the spatiotemporal model Spatiotemporal Model incorporating Depth (STM-D)exhibits slightly smaller MSPEs for both model components (402.9253 for the HMC and 14.7231 for the WMC) when compared to the STM. As a result, there is a reduction of 2.0% and 1.9% in MSPE for the WMC and the HMC, respectively, when transitioning from an STM to an STM-D. The deduction rates for MSPE are greater than 1% in both model components, underscoring that the STM-D framework performs better than the STM framework in both components. Consequently, STM-D is selected for the JWHM, serving as the optimal model for predicting both shell height and meat weight.

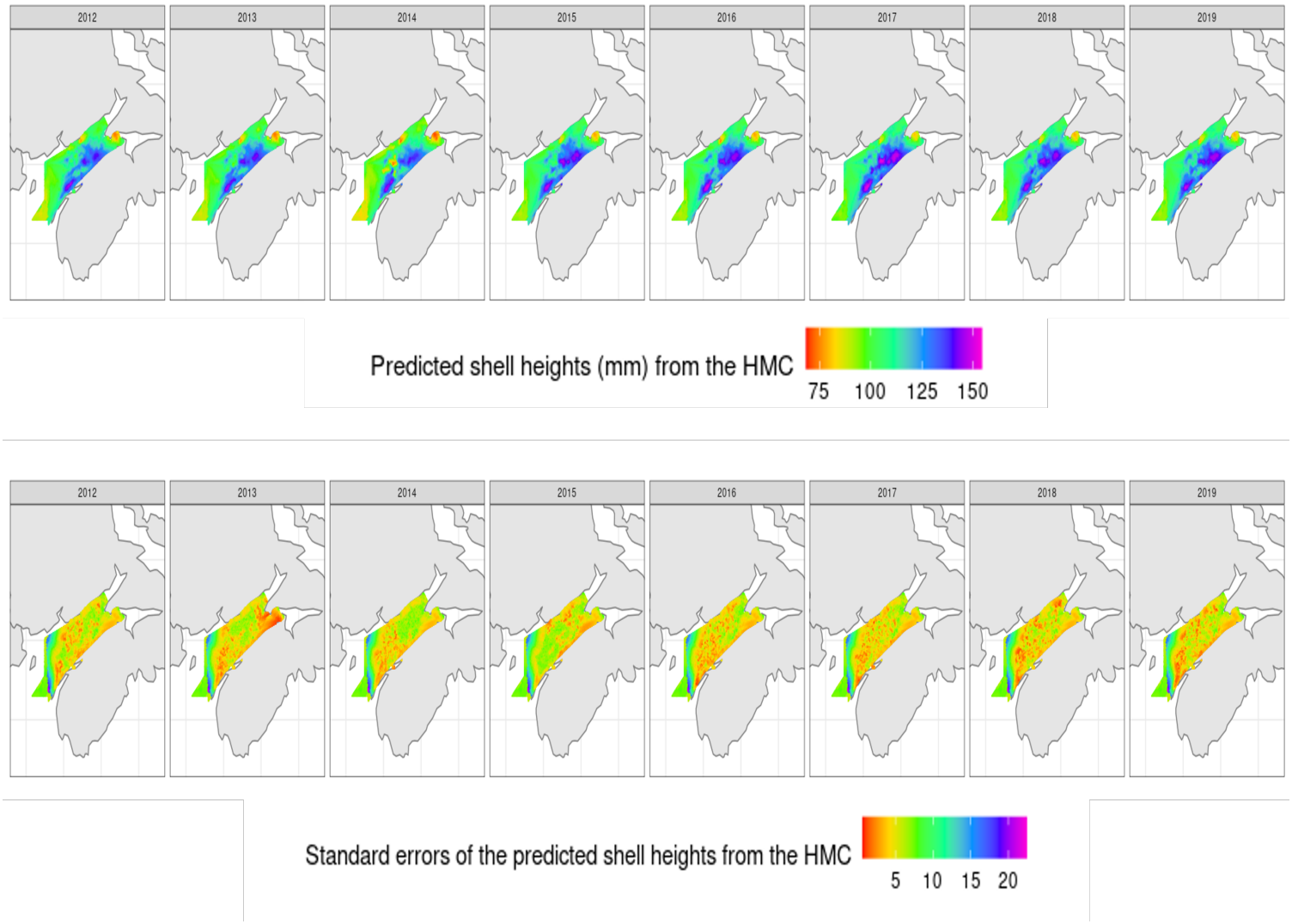
The STM-D parameter estimations for both model components are Table A.1

and Table A.2. From Table A.2, $\alpha_1 = 2.88$ in Equation 4.2 is less than 3, it means negatively allometric growth, indicating that the scallops in the study area, on average, become lighter as they grow larger.

4.2 Model Predictions

First we employed the STM-D to predict shell height across the study area in the Bay of Fundy (see Figure 4.7, refer to Figure 2.1 for the study area). Over the span of 8 years, the predictions indicate that larger scallops with shell heights greater than 130 mm (visualized as color purple), are more likely to be found near the Nova Scotia coast of the Bay of Fundy, mainly encompassing SPA 1A, SPA 4, and SPA 5, than the New Brunswick coast, primarily covering SPA 1B. In the HMC, the cumulative impact of spatial and spatiotemporal effects (see Figure B.1) reveals positive effects in SPA 1A, 4, and 5, while negative effects are observed in SPA 1B. These varying influences from random effects potentially contribute to reduced shell height predictions in SPA 1B and elevated predictions in terms of shell height for SPA 1A, 4, and 5. In the majority of the study area locations, the predicted shell height is approximately 110 mm (visualized as color blueish-green) , with a standard error of about 7.

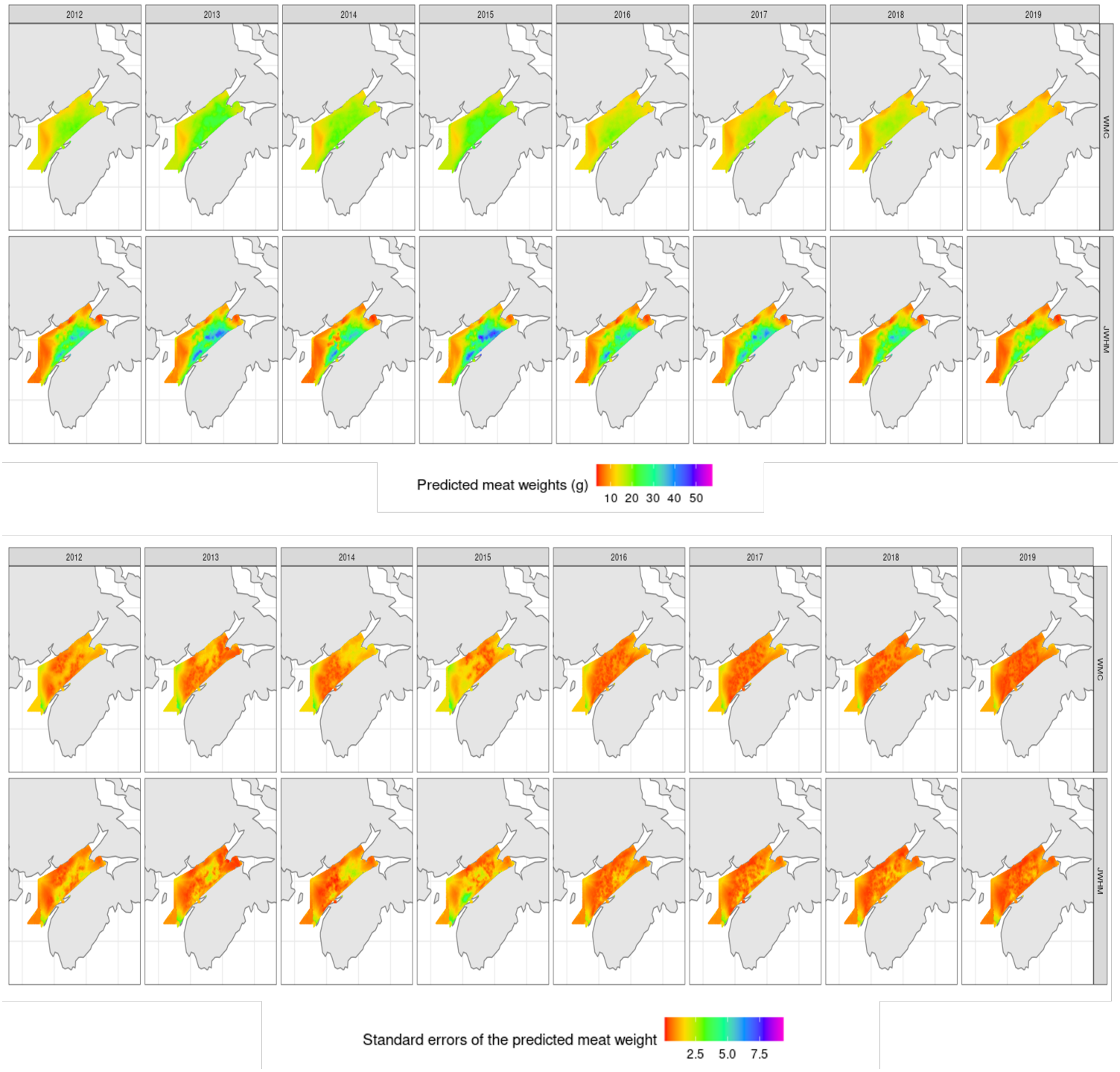
Figure 4.7: Shell height predictions and standard errors from the HMC.



It is worth noting that the WMC closely resembles the STM framework established by Yin et al. (2022) and has been actively utilized by DFO. However, the WMC does not have the ability to capture the shell height variability. To underscore the significance of incorporating shell height variability in the prediction of meat weight, we conducted a comparison between using the WMC with a fixed height and the JWHM. In the meat weight prediction utilizing the WMC, a fixed shell height of 114 mm (the average shell height from the MWSH dataset) was applied across the entire study area. Conversely, for the meat weight prediction using the JWHM, the shell

height values predicted by the HMC were employed consistently across the study area. Figure 4.8 displays the meat weight predictions and their associated standard errors obtained from both the WMC and JWHM, respectively.

Figure 4.8: Comparison of meat weight predictions and standard errors between the WMC and JWHM.



Notes. WMC: predicted meat weights (g) with the fixed shell height (114 mm). JWHM: predicted meat weights (g) with the predicted shell heights (mm).

In the central study area, the differences in spatial predictions of meat weight are hardly noticeable when utilizing the WMC with a fixed height. Moreover, it is worth noticing that the meat weight prediction in this central study area seems to exhibit a declining trend with year. There is no obvious decreasing patterns in the outer study area and the WMC yields an average meat weight prediction around 25 g (visualized as color green) from 2012 to 2015 and around 15 g (visualized as color yellowish-green) from 2016 to 2019 in the central study area and around 10 g (visualized as color orange) in the outer study area across 8 years.

In contrast, the JWHM reveals noticeable differences in meat weight predictions within the central study area and more substantial differences between the central area and the outer study area. In the central study area, the predicted meat weights remain consistently around 25 g each year in the majority of locations. Interestingly, the predictions indicate the presence of heavier scallops, with meat weights exceeding 45 g (visualized as color purple) near the Nova Scotia coast (SPA 1A, 4 and 5). One possible reason for predicting these heavier scallops in SPA 1A, 4 and 5 is the predicted shell heights from the HMC are also larger on average in these areas (see Figure 4.7). There is no obvious declining trend of the predictions with year, but in 2018 and 2019, the predictions indicate the scarcity of these heavier scallops near the Nova Scotia coast. In the WMC, the spatiotemporal effects exhibit distinct influences, particularly notable in 2019 (see Figure B.2). To be more specific, there are discernible negative spatiotemporal effects in SPA 1A, 4, and 5. This specific pattern may contribute to lower predictions of meat weight on average in these areas during 2019 in comparison to other years. The JWHM predicts lighter scallops around 5 g (visualized as color reddish-orange) occur in the outer area compared to the predictions made by the WMC.

When comparing the prediction standard errors, there is little difference between

the WMC and JWHM results. In the majority of the study area locations in the Bay of Fundy, both the WMC and JWHM exhibit standard errors of less than 2. However, it is important to highlight that the JWHM does indeed show slightly larger standard errors in some locations compared to the WMC. This variation can be attributed to the fact that the JWHM incorporates shell height predictions, which the WMC does not account for. Shell height tend to be larger in these locations are more likely to exhibit a larger meat weight variability in nature.

Based on the observations from Figures 4.7 and 4.8, it becomes apparent that over the course of 8 years, heavier and larger scallops are more likely to predicted in SPA 1A, SPA 4, and SPA 5 compared to SPA 1B. This finding matches up the empirical result in Chapter 2 (see Figures 2.13 and 2.13).

Hence, it demonstrated that the JWHM provides more accurate meat weight predictions by incorporating shell height variability, making it a superior model compared to the WMC with a fixed shell height across the study area.

4.3 Summary

In this chapter, we applied our statistical methodology and obtained modelling results. To enhance the original LWR model's capability to encompass variability in both shell height and meat weight over space and time, we proposed the JWHM, which is a product of the WMC and the HMC. Initially, we elucidated the process of model fitting, which encompasses distribution validation, as well as the selection of random effects and environmental variables for both model components. By examining the RQR plots, it became evident that the sdt_2 exhibited superior performance when contrasted with the normal distribution for both model components. We used 10-fold SCV to select random effects and backward selection to choose environmental variables. Our results indicated that the STM-D yielded the most favorable predictive performance for both model components.

Subsequently, we employed the STM-D to forecast shell height and meat weight across the study area in the Bay of Fundy. To underscore the significance of accounting for shell height variability, we conducted a comparison between the meat weight predictions from the WMC with a fixed shell height and the JWHM with predicted shell heights. Our analysis revealed that the JWHM appeared to provide more accurate meat weight predictions, predicting heavier scallops in SPA 1A, 4, and 5, compared to SPA 1B.

Overall, this chapter details the journey of obtaining the JWHM and presents the resulting predictions. In Chapter 5, we draw the entire thesis to a close, providing a comprehensive conclusion and delving into potential avenues for future research.

Chapter 5

Conclusions

We encounter a more complex and time-intensive challenge when it comes to measuring the weight of scallops, in contrast to assessing their heights. Hence, the LWR emerges as a valuable approach for estimating scallop meat weights, relying on their shell heights. Given the anticipated strong correlation in the growth of scallop individuals when they are in proximity to each other (Carsen et al., 1996), it is imperative to consider the geographical correlations in environmental conditions that affect both the heights and weights of sea scallops. Hence, we used the widely adopted Matérn GMRF for incorporating these important geographical correlations into a basic LWR.

We developed the JWHM to enhance the foundational LWR by accommodating the complexities of spatial and spatiotemporal variations in meat weight and shell height. The JWHM is the result of combining the WMC and the HMC, which were fitted by the *sdmTMB* function (Anderson et al., 2023). When it comes to model distributional assumption checking, our approach involved employing RQR plots to demonstrate that sdt_2 outperforms the normal distribution for both model components.

In both model components, there are two random effects and four environmental variables. However, overly complicated model structures can lead to potential

overfitting of data. Hence, we decided to choose significant random effects and environmental variables to avoid the unintentional selection of models that ultimately yield less accurate predictions. Therefore, in our study, we implemented 10-fold SCV to underscore that opting for the STM framework incorporating both spatial and spatiotemporal efffects for both model components enhances their predictive performance compared to utilizing the SM framework incorporating just spatial effects. During the stratified sampling process, we showed that observations were chosen from each year and from nearly every tow location within each fold. This strategic approach was implemented to mitigate potential biases from either the spatial or temporal dimension during the CV procedure. Subsequently, we applied the STM framework to both model components to refine our selection of environmental variables. The backward selection method revealed that depth was a significant variable for both model components, supported by a p-value of 0.05. Additionally, we computed MSPEs for the STM-D in both model components. It served as a further confirmation to validate that the STM-D framework is the optimal choice for both model components. By independently optimizing the WMC and the HMC, we determined that STM-D is the most suitable model framework for the JWHM.

We emphasized the importance of considering shell height variability by conducting a comparison between meat weight predictions from the WMC with a fixed shell height, which has been actively utilized by DFO, and the JWHM with predicted shell heights. Our analysis unveiled that while the WMC with a fixed shell height is valuable for discerning trends in scallop meat weight over time, the JWHM offers more precise meat weight predictions. The JWHM predicts heavier scallops on average in SPA 1A, 4, and 5, in contrast to SPA 1B. To be more specific, in the central study area, the predicted meat weights remain consistently around 25 g each year in the majority of locations. Interestingly, the predictions indicate the presence of

heavier scallops, with meat weights exceeding 45 g near the Nova Scotia coast (SPA 1A, 4 and 5). There is no obvious declining trend of the predictions with year, but in 2018 and 2019, the predictions indicate the scarcity of these heavier scallops near the Nova Scotia coast. Also, lighter scallops around 5 g tend to occur in the outer study area in all 8 years.

Fisheries scientists are now delving into the intricacies of scallop LWRs, where the constantly shifting environmental conditions that scallops inhabit pose a challenge in accurately modeling LWRs (Yin et al., 2022). Therefore, while the current JWHM demonstrates remarkable efficacy in predicting within-year meat weight, its applicability to precasting or forecasting remains uncertain.

Sea scallops (*P. magellanicus*) hold considerable significance for both human well-being and Canada's economy. In pursuit of long-term sustainability for the scallop fishery, fishery stock assessments provide DFO management with essential data, including annual scallop biomass, to dynamically adjust fishing policies and maintain a healthy scallop population. This research study shows a significant advancement in enhancing meat weight predictions by using the JWHM, building upon the Yin et al. (2022) STM framework by incorporating both meat weight and shell height variability. In a comprehensive meat weight prediction comparison between the JWHM and the current STM model employed by DFO, significant enhancements stemming from the incorporation of both meat weight and shell height variability were validated. Our JWHM offers more accurate predictions of meat weight, therefore, the meat weight prediction results obtained from the JWHM can provide DFO with a more intuitive understanding of varying scallop biomass levels within different SPAs. This enhanced insight can significantly improve DFO's ability to dynamically monitor scallop populations and make well-informed decisions across different regions of the Bay of Fundy.

Bibliography

Anderson, S. C., Ward, E. J., Barnett, L. A. K., English, P. A., James T., T., Joe, W., & Julia, I. (2023, January). sdmTMB: Spatial and Spatiotemporal SPDE-Based GLMMs with 'TMB'. https://cran.r-project.org/web/packages/sdmTMB/index.html

Arnholt, A. (2021). Cross-validation hand out. https://stat-ata-asu.github.io/STT3851ClassRepo/Rmarkdown/Cross-ValidationInClassHO.html

Belmont, J. (2022). RPubs - Building a mesh with INLA. https://rpubs.com/jafet089/886687

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Branch, L. S. (2022, September). Consolidated federal laws of canada, atlantic fishery regulations, 1985. https://laws-lois.justice.gc.ca/eng/regulations/sor-86-21/page-5.html#h-892127

Brand, A. R. (2006, January). Chapter 12 Scallop ecology: Distributions and behaviour. In S. E. Shumway & G. J. Parsons (Eds.), *Developments in Aquaculture and Fisheries Science* (pp. 651–744, Vol. 35). Elsevier. https://doi.org/10.1016/S0167-9309(06)80039-6

Canada, F. a. O. (2018). Offshore scallop. https://www.dfo-mpo.gc.ca/fisheries-peches/ifmp-gmp/scallop-petoncle/2018/index-eng.html

Canada, F. a. O. (2021). Fisheries science: Overview. https://www.dfo-mpo.gc.ca/science/species-especes/fisheries-halieutiques/about-sur/index-eng.html

Carsen, A. E., Hatcher, B. G., & Scheibling, R. E. (1996). Effect of flow velocity and body size on swimming trajectories of sea scallops, Placopecten magellanicus (Gmelin): A comparison of laboratory and field measurements. *Journal of Experimental Marine Biology and Ecology*, *203*(2), 223–243. https://doi.org/10.1016/0022-0981(96)02578-6

Côté, J., Himmelman, J. H., Claereboudt, M., & Bonardelli, J. C. (1993). Influence of density and depth on the growth of juvenile sea scallops (Placopecten magellanicus) in suspended culture. *Canadian Journal of Fisheries and Aquatic Sciences*, *50*(9), 1857–1869. https://doi.org/10.1139/f93-208

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *30*(2), 248–275. https://www.jstor.org/stable/2984505

Davies, S. C., Gregr, E. J., Lessard, J., Bartier, P., & Peter, W. (2019). Coastal digital elevation models integrating ocean bathymetry and land topography for marine ecological analyses in Pacific Canadian waters. https://publications.gc.ca/site/archivee-archived.html?url=https://publications.gc.ca/collections/collection_2019/mpo-dfo/Fs97-6-3321-eng.pdf

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244. https://doi.org/10.2307/1390802

Dunn, P. K., & Smyth, G. K. (2018). Chapter 8: Generalized linear models: Diagnostics. In P. K. Dunn & G. K. Smyth (Eds.), *Generalized Linear Models With Examples in R* (pp. 297–331). Springer. https://doi.org/10.1007/978-1-4419-0118-7_8

ECO. (2018, February). Overfishing is a huge problem. Here's what you need to know. https://www.ecomagazine.com/news/policy/overfishing-is-a-huge-problem-here-s-what-you-need-to-know

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, *20*(1), 175. https://doi.org/10.1186/s12874-020-01055-2

Fisheries and Oceans Canada. (2009). Global consequences of overfishing - international fisheries. https://www.dfo-mpo.gc.ca/international/isu-global-eng.htm

Fisheries and Oceans Canada. (2017). Inshore scallop. https://www.dfo-mpo.gc.ca/fisheries-peches/ifmp-gmp/scallop-petoncle/scallop-petoncle2015-sec4-5-6-7-eng.html

Fisheries and Oceans Canada. (2021, October). Research Document - 2013/004. https://www.dfo-mpo.gc.ca/csas-sccs/Publications/ResDocs-DocRech/2013/2013_004-eng.html

Ford, C. (2015). Understanding qq plots. https://data.library.virginia.edu/understanding-q-q-plots/

Gayon, J. (2000). History of the concept of allometry. *American Zoologist*, *40*(5), 748–758. https://doi.org/10.1093/icb/40.5.748

Glass, A. (2017). Maritimes region inshore scallop assessment survey: Detailed technical description. https://publications.gc.ca/site/archivee-archived.html?url=https://publications.gc.ca/collections/collection_2018/mpo-dfo/Fs97-6-3231-eng.pdf

Great Learning Team. (2020). What is cross validation in machine learning? Types of cross validation. https://www.mygreatlearning.com/blog/cross-validation/

Guy, C. S., & Brown, M. L. (2007). Analysis and interpretation of freshwater fisheries data. In U.S. Geological Survey Montana Cooperative Fishery Research Unit Montana State University & Department of Wildlife and Fisheries Sciences South Dakota State University (Eds.), *Analysis and Interpretation of Freshwater Fisheries Data*. American Fisheries Society. https://doi.org/10.47886/9781888569773.ch1

Hali, B. (2022). Understanding bias-variance trade-off in 3 minutes. https://www.kdnuggets.com/understanding-bias-variance-trade-off-in-3-minutes.html

Higgins. (2009). Cod moratorium in newfoundland and labrador. https://www.heritage.nf.ca/articles/economy/moratorium.php

Huilgol, P. (2020). Bias and variance in machine learning - a fantastic guide for beginners! https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/

Jackson-Bué, T., Williams, G. J., Whitton, T. A., Roberts, M. J., Goward Brown, A., Amir, H., King, J., Powell, B., Rowlands, S. J., Llewelyn Jones, G., & Davies, A. J. (2022). Seabed morphology and bed shear stress predict temperate reef habitats in a high energy marine

region. *Estuarine, Coastal and Shelf Science*, *274*, 107934. https://doi.org/10.1016/j.ecss.2022.107934

Jisr, N., Younes, G., Sukhn, C., & El-Dakdouki, M. H. (2018). Length-weight relationships and relative condition factor of fish inhabiting the marine area of the Eastern Mediterranean city, Tripoli-Lebanon. *The Egyptian Journal of Aquatic Research*, *44*(4), 299–305. https://doi.org/10.1016/j.ejar.2018.11.004

Jona Lasinio, G., Mastrantonio, G., & Pollice, A. (2012). Discussing the "big n problem". *Statistical Methods & Applications*, *22*(1), 97–112. https://doi.org/10.1007/s10260-012-0207-2

Keys, A. B. (1928). The weight-length relation in fishes. *Proceedings of the National Academy of Sciences*, *14*(12), 922–925. https://doi.org/10.1073/pnas.14.12.922

Kong, N., Liu, Z., Yu, Z., Fu, Q., Li, H., Zhang, Y., Fang, X., Zhang, F., Liu, C., Wang, L., & Song, L. (2022). Dynamics of phytoplankton community in scallop farming waters of the bohai sea and north yellow sea in china. *BMC Ecology and Evolution*, *22*(1), 48. https://doi.org/10.1186/s12862-022-02002-z

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2021). Advanced spatial modeling with stochastic partial differential equations using r and inla. https://becarioprecario.bitbucket.io/spde-gitbook/index.html

Kristensen, K., Bell, B., Skaug, H., Magnusson, A., Berg, C., Nielsen, A., Maechler, M., Michelot, T., Brooks, M., Forrence, A., Albertsen, C. M., & Monnahan, C. (2023, July). TMB: Template Model Builder: A General Random Effect Tool Inspired by 'ADMB'. https://cran.r-project.org/web/packages/TMB/index.html

Kuhn, M. (2023). *The caret Package*. https://topepo.github.io/caret/

Kumar, A. (2022). Generalized linear models explained with examples. https://vitalflux.com/generalized-linear-models-explained-with-examples/

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. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lindmark, M., Audzijonyte, A., Blanchard, J. L., & Gårdmark, A. (2022). Temperature impacts on fish physiology and resource abundance lead to faster growth but smaller fish sizes and yields under warming. *Global Change Biology*, *28*(21), 6239–6253. https://doi.org/10.1111/gcb.16341

Mailman School of Public Health. (2016, August). Spatiotemporal analysis. https://www.publichealth.columbia.edu/research/population-health-methods/spatiotemporal-analysis

Mazumder, S. K., Das, S. K., Bakar, Y., & Ghaffar, M. A. (2016). Effects of temperature and diet on length-weight relationship and condition factor of the juvenile Malabar blood snapper (Lutjanus malabaricus Bloch & Schneider, 1801). *Journal of Zhejiang University. Science. B*, *17*(8), 580–590. https://doi.org/10.1631/jzus.B1500251

Moraga, P. (2019). Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny. https://www.paulamoraga.com/book-geospatial/sec-arealdataexamplespatial.html

Muralidhar, K. S. V. (2023). What is stratified cross-validation in machine learning? https://towardsdatascience.com/what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e

Myers, R. A., Hutchings, J. A., & Barrowman, N. J. (1997). Why do fish stocks collapse? The example of cod in atlantic canada. *Ecological Applications*, *7*(1), 91–106. https://doi.org/10.1890/1051-0761(1997)007[0091:WDFSCT]2.0.CO;2

NOAA Fisheries. (2023, January). Stock assessment model descriptions — noaa fisheries. https://www.fisheries.noaa.gov/insight/stock-assessment-model-descriptions

Palomares-Garcia, R., Bustillos-Guzman, J. J., & Lopez-Cortes, D. (2006). Pigment-specific rates of phytoplankton growth and microzooplankton grazing in a subtropical lagoon. *Journal of Plankton Research*, *28*(12), 1217–1232. https://doi.org/10.1093/plankt/fbl051

Pandian, S. (2022). K-fold cross validation technique and its essentials. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Rai, K. (2020). Assessing the performance (Types and sources of error) in machine learning. https://medium.com/analytics-vidhya/assessing-the-performance-types-and-sources-of-error-in-machine-learning-e5d28b71da6b

Ratner, B. (2009). The correlation coefficient: Its values range between +1/1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, *17*(2), 139–142. https://doi.org/10.1057/jt.2009.5

Ricker, W. (1959, September). *Handbook of computations for biological statistics of fish populations.* (Vol. 34). https://waves-vagues.dfo-mpo.gc.ca/library-bibliotheque/10161.pdf

Rue, H., Lindgren, F., Niekerk, J. v., Krainski, E., & Fattah, E. A. (2023). R-inla project - documentation. https://www.r-inla.org/documentation

Schnute, J. T., Boers, N., Haigh, R., Couture-Beil, A., Chabot, D., Grandin, C., Johnson, A., Wessel, P., Antonio, F., Lewin-Koh, N. J., & Bivand, R. (2022, September). PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools. https://cran.r-project.org/web/packages/PBSmapping/index.html

Shah, T. (2020). About train, validation and test sets in machine learning. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Sidén, P., & Lindsten, F. (2020). Deep Gaussian Markov random fields. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 8916–8926.

Silina, A. V. (2023). Effects of temperature, salinity, and food availability on shell growth rates of the Yesso scallop. *PeerJ*, *11*, e14886. https://doi.org/10.7717/peerj.14886

Singh, S. (2018). Understanding the Bias-Variance Tradeoff. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Thorson, J. T., & Haltuch, M. A. (2019). Spatiotemporal analysis of compositional data: Increased precision and improved workflow using model-based inputs to stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(3), 401–414. https://doi.org/10.1139/cjfas-2018-0015

Thorson, J. T., & Minto, C. (2015). Mixed effects: A unifying framework for statistical modelling in fisheries biology. *ICES Journal of Marine Science*, *72*(5), 1245–1256. https://doi.org/10.1093/icesjms/fsu213

Thorson, J. T., Shelton, A. O., Ward, E. J., & Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *Ices Journal of Marine Science*, *72*(5), 1297–1310. https://doi.org/10.1093/ICESJMS/FSU243

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Urbina, M. A., & Glover, C. N. (2015). Effect of salinity on osmoregulation, metabolism and nitrogen excretion in the amphidromous fish, inanga (Galaxias maculatus). *Journal of Experimental Marine Biology and Ecology*, *473*, 7–15. https://doi.org/10.1016/j.jembe.2015.07.014

Wang, Z., Lu, Y., Greenan, B., & DeTracey, B. (2018). BNAM: An eddy-resolving North Atlantic Ocean model to support ocean monitoring. https://open.canada.ca/data/en/dataset/c44a8574-9f7d-45b7-afda-27802353a04c

Yin, Y., Sameoto, J. A., Keith, D. M., & Flemming, J. M. (2022). Improving estimation of length–weight relationships using spatiotemporal models. *Canadian Journal of Fisheries and Aquatic Sciences*, *79*(11), 1896–1910. https://doi.org/10.1139/cjfas-2021-0317

Appendices

Appendix A

Tables

Table A.1: HMC parameter estimation.

Fixed effects	Estimate	Std. Error
Intercept	116.9656	6.1867
2013	0.5949	1.4208
2014	-3.3501	1.4853
2015	3.6160	1.6061
2016	3.6376	1.3581
2017	6.7444	1.3301
2018	4.7099	1.3311
2019	4.9021	1.3222
depth	-0.1444	0.03273
Random effects	Estimate	Std. Error
Dispersion ϕ	2.3725	0.005019
Spatial precision τ_s	-1.7926	0.07651
Spatial scale κ_s	-2.3916	0.1662
Spatiotemporal precision τ_{st}	-2.2871	0.07685
Spatiotemporal scale κ_{st}	-0.8745	0.08056

Notes. All random effect parameters are log-transformed for the optimization process and hence, estimates reported are on the log scale. Precision parameter τ^2 is the inverse of the marginal variance σ^2.

Table A.2: WMC parameter estimation.

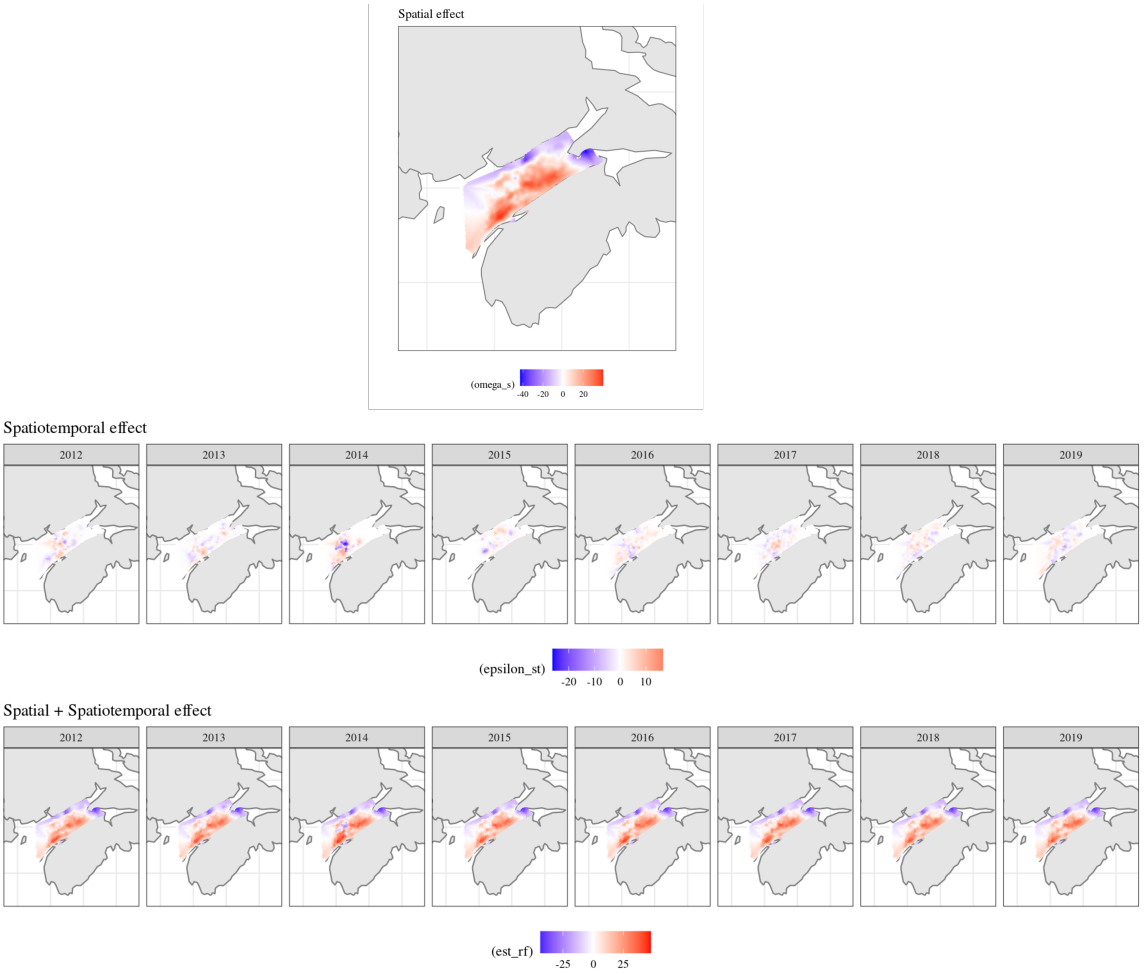
Fixed effects	Estimate	Std. Error
Intercept	-10.6301	0.1395
2013	0.1552	0.03732
2014	0.0734	0.03993
2015	0.1575	0.04085
2016	-0.05211	0.03688
2017	-0.07551	0.03689
2018	-0.1484	0.03723
2019	-0.2704	0.03712
log(height)	2.8785	0.006406
log(depth)	-0.05900	0.01702
Random effects	Estimate	Std. Error
Dispersion ϕ	0.6386	0.006923
Spatial precision τ_s	3.2004	0.1156
Spatial scale κ_s	-2.8910	0.2471
Spatiotemporal precision τ_{st}	3.2259	0.09881
Spatiotemporal scale κ_{st}	-2.0436	0.1197

Notes. All random effect parameters are log-transformed for the optimization process and hence, estimates reported are on the log scale. Precision parameter τ^2 is the inverse of the marginal variance σ^2.

Appendix B

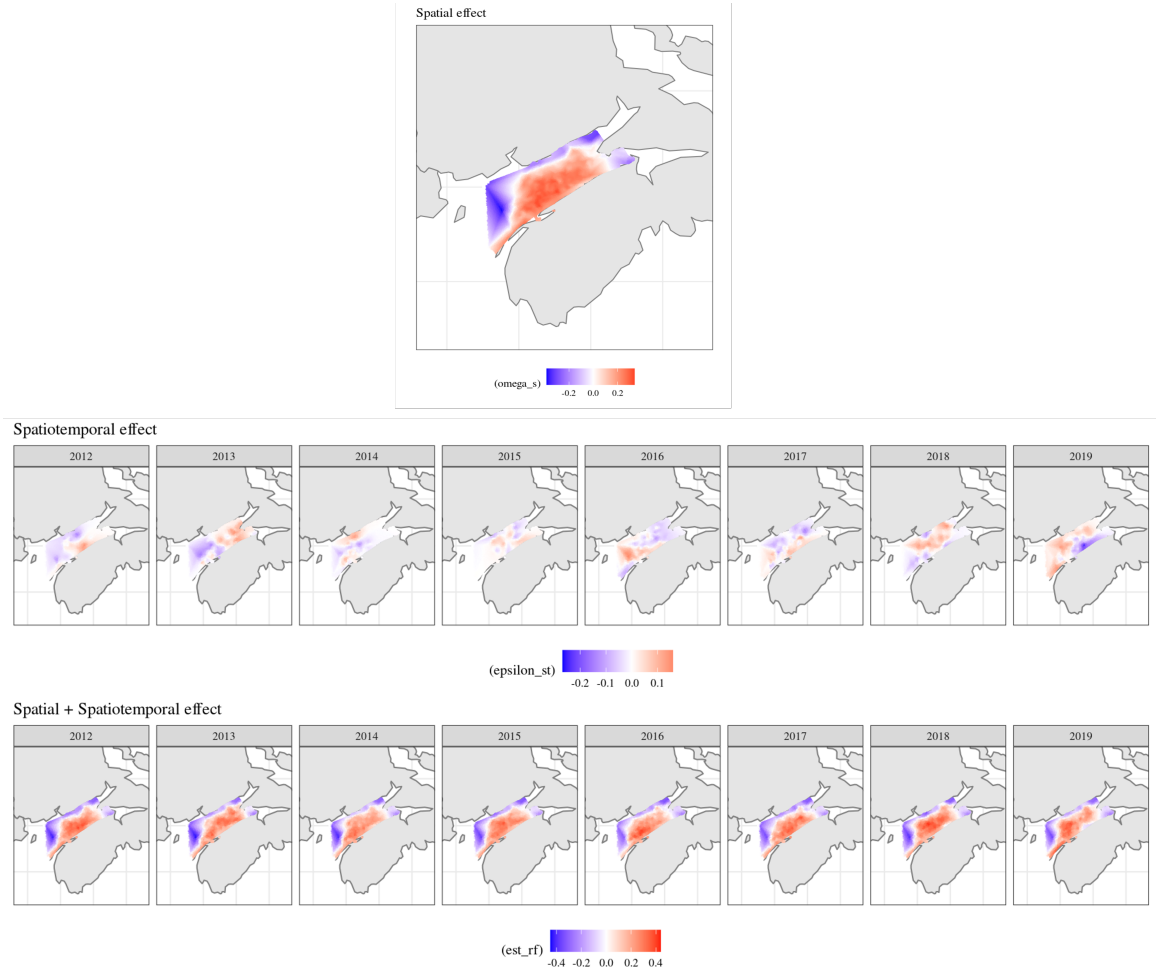
Graphs

Figure B.1: Visualization of the spatial, and spatiotemporal effects with their sum for the HMC.



Notes. omega_s represents $\zeta_0(s_j)$ and epsilon_st represents $\zeta_0(s_j, t_j)$, and est_rf is the sum of both omega_s and epsilon_st

Figure B.2: Visualization of the spatial, and spatiotemporal effects with their sum for the WMC.



Notes. omega_s represents $\beta_0(s_i)$ and epsilon_st represents $\beta_0(s_i, t_i)$, and est_rf is the sum of both omega_s and epsilon_st.

Appendix C

R coding

```
library(ggplot2)
library(dplyr)
library(tidyverse)
require(maptools)
library(mapdata)
library(maps)
library(mapproj)
library(gridExtra)
library(grid)
library(caret)
library(TMB)
library(INLA)
library(raster)
library(sf)
library(sdmTMB)
library(rgdal)
library(ggpubr)
library(ggmap)

# HMC

# clean the SHF data
heightdata=read.csv("bof.shf.unlined.gear.2012to2019.csv")
heightdata=na.omit(heightdata) # remove RF
heightdata$year=NA
heightdata$ID_TOW=NA

for(i in 1:length(heightdata[,1]))
{
  if(grepl((heightdata$CRUISE[i]), "BF2012")==T   )
  {
    heightdata$year[i]=2012
  }
```

```
  if(grepl((heightdata$CRUISE[i]), "BF2013")==T  )

  {
    heightdata$year[i]=2013
  }
  if(grepl((heightdata$CRUISE[i]), "BF2014")==T   )
  {
    heightdata$year[i]=2014
  }
  if(grepl((heightdata$CRUISE[i]), "BF2015")==T   )
  {
    heightdata$year[i]=2015

  }
  if(grepl((heightdata$CRUISE[i]), "BF2016")==T   )
  {
    heightdata$year[i]=2016
  }
  if(grepl((heightdata$CRUISE[i]), "BF2017")==T   )
  {
    heightdata$year[i]=2017
  }
  if(grepl((heightdata$CRUISE[i]), "BF2018")==T   )
  {
    heightdata$year[i]=2018
  }
  if(grepl((heightdata$CRUISE[i]), "BF2019")==T  )
  {
    heightdata$year[i]=2019
  }
}

heightdata$year = as.factor(heightdata$year)
heightdata$ID_TOW =as.factor(paste(heightdata$year,heightdata$TOW_NO,sep = '_'))


# create the SH  dataset
set.seed(1)
test=c()
binocoun=matrix(0,length(heightdata[,1]),40)

for(i in 1: length(heightdata[,1]))
```

```
{
  for(k in 1:40)
  {
    binocoun[i,k]=floor(heightdata[i,k+5])
    + rbinom(1,1,((heightdata[i,k+5])%% 1))
    new_value =runif( binocoun[i,k],min=(k-1)*5,max=((k-1)*5)+4.99)
    test=c(test,new_value)
  }
}


for(i in 1: length(heightdata[,1]))

{
  heightdata$bcoun[i]=sum(binocoun[i,seq(from=1,to=40, by=1)])
}


fullset <- as.data.frame(lapply(heightdata, rep, heightdata$bcoun ))
fullset$gheight=test
Julyfullset=fullset[fullset$month==7,]

draster=raster("BathyCHS_GEBCO_SEAM_mixedData_BOF_ExtentClip_100m_LatLong.asc")
r2012=raster("BtmTemp_Jul_2012.asc")
r2013=raster("BtmTemp_Jul_2013.asc")
r2014=raster("BtmTemp_Jul_2014.asc")
r2015=raster("BtmTemp_Jul_2015.asc")
r2016=raster("BtmTemp_Jul_2016.asc")
r2017=raster("BtmTemp_Jul_2017.asc")
r2018=raster("BtmTemp_Jul_2018.asc")
r2019=raster("BtmTemp_Jul_2019.asc")

s2012=raster("BtmStress_Jul_2012.asc")
s2013=raster("BtmStress_Jul_2013.asc")
s2014=raster("BtmStress_Jul_2014.asc")
s2015=raster("BtmStress_Jul_2015.asc")
s2016=raster("BtmStress_Jul_2016.asc")
s2017=raster("BtmStress_Jul_2017.asc")
s2018=raster("BtmStress_Jul_2018.asc")
s2019=raster("BtmStress_Jul_2019.asc")

a2012=raster("BtmSalinity_Jul_2012.asc")
a2013=raster("BtmSalinity_Jul_2013.asc")
a2014=raster("BtmSalinity_Jul_2014.asc")
```

```
a2015=raster("BtmSalinity_Jul_2015.asc")
a2016=raster("BtmSalinity_Jul_2016.asc")
a2017=raster("BtmSalinity_Jul_2017.asc")
a2018=raster("BtmSalinity_Jul_2018.asc")
a2019=raster("BtmSalinity_Jul_2019.asc")


# assign environmental values  for every observation
Julyfullset$temperature=NA
Julyfullset$salinity=NA
Julyfullset$stress=NA

# spatially extract the depth values
Julyfullset$depth=raster::extract(draster, y = cbind(Julyfullset$mid.lon,
                                                     Julyfullset$mid.lat))

# spatiotemporally extract the other environmental variable values
h1=Julyfullset[Julyfullset$year==2012,]
h2=Julyfullset[Julyfullset$year==2013,]
h3=Julyfullset[Julyfullset$year==2014,]
h4=Julyfullset[Julyfullset$year==2015,]
h5=Julyfullset[Julyfullset$year==2016,]
h6=Julyfullset[Julyfullset$year==2017,]
h7=Julyfullset[Julyfullset$year==2018,]
h8=Julyfullset[Julyfullset$year==2019,]

h1$temperature=raster::extract(r2012, y = cbind(h1$mid.lon , h1$mid.lat))
h2$temperature=raster::extract(r2013, y = cbind(h2$mid.lon , h2$mid.lat))
h3$temperature=raster::extract(r2014, y = cbind(h3$mid.lon , h3$mid.lat))
h4$temperature=raster::extract(r2015, y = cbind(h4$mid.lon , h4$mid.lat))
h5$temperature=raster::extract(r2016, y = cbind(h5$mid.lon , h5$mid.lat))
h6$temperature=raster::extract(r2017, y = cbind(h6$mid.lon , h6$mid.lat))
h7$temperature=raster::extract(r2018, y = cbind(h7$mid.lon , h7$mid.lat))
h8$temperature=raster::extract(r2019, y = cbind(h8$mid.lon , h8$mid.lat))

h1$stress=raster::extract(s2012, y = cbind(h1$mid.lon,h1$mid.lat))
h2$stress=raster::extract(s2013, y = cbind(h2$mid.lon, h2$mid.lat))
h3$stress=raster::extract(s2014, y = cbind(h3$mid.lon, h3$mid.lat))
h4$stress=raster::extract(s2015, y = cbind(h4$mid.lon, h4$mid.lat))
h5$stress=raster::extract(s2016, y = cbind(h5$mid.lon, h5$mid.lat))
h6$stress=raster::extract(s2017, y = cbind(h6$mid.lon, h6$mid.lat))
h7$stress=raster::extract(s2018, y = cbind(h7$mid.lon, h7$mid.lat))
h8$stress=raster::extract(s2019, y = cbind(h8$mid.lon, h8$mid.lat))
```

```
h1$salinity=raster::extract(a2012, y = cbind(h1$mid.lon, h1$mid.lat))
h2$salinity=raster::extract(a2013, y = cbind(h2$mid.lon, h2$mid.lat))
h3$salinity=raster::extract(a2014, y = cbind(h3$mid.lon, h3$mid.lat))
h4$salinity=raster::extract(a2015, y = cbind(h4$mid.lon, h4$mid.lat))
h5$salinity=raster::extract(a2016, y = cbind(h5$mid.lon, h5$mid.lat))
h6$salinity=raster::extract(a2017, y = cbind(h6$mid.lon, h6$mid.lat))
h7$salinity=raster::extract(a2018, y = cbind(h7$mid.lon, h7$mid.lat))
h8$salinity=raster::extract(a2019, y = cbind(h8$mid.lon, h8$mid.lat))

Julyfullset$temperature=c(h1$temperature,h2$temperature,h3$temperature,
                          h4$temperature,h5$temperature,h6$temperature,
                          h7$temperature,h8$temperature)
Julyfullset$stress=c(h1$stress,h2$stress,h3$stress,h4$stress,
                     h5$stress,h6$stress,h7$stress,h8$stress)
Julyfullset$salinity=c(h1$salinity,h2$salinity,h3$salinity,h4$salinity,
                       h5$salinity,h6$salinity,h7$salinity,h8$salinity)


# SH dataset
myheight= Julyfullset%>%filter(
  !is.na(temperature),
  !is.na(depth),
  !is.na(stress),
  !is.na(salinity)
) %>% transmute(
  height = gheight,
  year = as.factor(year),
  lon= mid.lon, lat= mid.lat ,
  TOW_NO=TOW_NO,
  ID_TOW = as.factor(paste(year,TOW_NO,sep = '_')),
  depth = -depth ,
  temperature = temperature,
  salinity=salinity,
  stress=stress)  # make sure no NA, all values should be positive

write.csv(myheight,"myheight.csv")
myheight=read.csv("myheight.csv")


# create a UTM dataset for the SH dataset
hall_set <- myheight %>%
  mutate(
    depth = depth,
    temperature = temperature,
```

```
    salinity=salinity,
    stress=stress) %>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
  `attr<-`("zone", "20") %>%
  PBSmapping::convUL()
# always check if the year is a factor


# create GPS map for the SH dataset
bbox <- c(left = min(myheight$lon)-1, bottom = min(myheight$lat)-1,
          right = max(myheight$lon)+1, top = max((myheight$lat)+1))
latitude =    myheight$lat
longitude = myheight$lon
year=as.factor(myheight$year)
site_df = as.data.frame(cbind(latitude,longitude,year))
site_map = ggmap(get_stamenmap(bbox, maptype = "terrain-background"))+
  geom_point(data = site_df, aes(x = longitude, y = latitude),
             size = 0.1, color = "red")+
  geom_point(data = site_df, aes(x = longitude, y = latitude),
             pch= 21, size = 0.1, color = "red")+
  facet_wrap(year ~ ., ncol = 4)+
  theme_bw() +
  labs(x = "Lon", y = "Lat")


# create base map used for ggplot
base_map <- ggplot() +
  borders("world", colour="gray50", fill = "gray90",
          xlim = c(-66.24-1 ,-64.49+1), ylim = c(44.42-1,45.52+1   )) +
  coord_map(xlim = c(-66.24-1 ,-64.49+1), ylim = c(44.42-1,45.52 +1  )) +
  theme_bw() +
  scale_color_continuous(low = "white", high = "red") +
  scale_size_continuous(guide = FALSE) +
  theme(axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "bottom")


# graphs for predictor Variables vs the response
h3<-ggplot(myheight, aes(x =( temperature) , y =  height    )) +
  geom_point() +labs(x = "Temperature ", y = "Height" )+
  geom_smooth(method='lm', formula= y~x)
```

```
h4<-ggplot(myheight, aes(x = depth , y =   height  )) +
  geom_point() +labs(x = "Depth ", y = "Height")+
  geom_smooth(method='lm', formula= y~x)

h5<-ggplot(myheight, aes(x = stress , y =   height  )) +
  geom_point() +labs(x = "Stress ", y = "Height")+
  geom_smooth(method='lm', formula= y~x)

h6<-ggplot(myheight, aes(x = salinity , y =   height  )) +
  geom_point() +labs(x = "Salinity ", y = "Height")+
  geom_smooth(method='lm', formula= y~x)

grid.arrange(h3,h4,h5,h6, nrow = 2, ncol=2)


# graph for explanatory variable correlation in the SH dataset
corr_matrix <- myheight[,c(7,8,10,9)] %>%
  cor(method="pearson", use="pairwise.complete.obs")
mtext("Explanatory variable correlations", at=2.5, line=-0.5, cex=1)


# graph for mean heights in each tow location
myheight=myheight%>%
  group_by(ID_TOW) %>%
  mutate(mean_height=mean(height,na.rm=T))

base_map +
  geom_point(data = myheight, aes(x=lon, y=lat, color = mean_height),
             shape = 20,size=0.05) +
  scale_color_gradientn(colours = rainbow(7), limits=c(30,160),
                        oob=scales::squish)  +
  labs(colour = "The mean shell height (mm) in each tow location")+
  facet_grid(~year)


# graph for the height distribution
ggplot(myheight, aes(x=height)) +
  geom_histogram(aes(y=..density..),
                 binwidth=.5,
                 color="black",  fill="white") +
  geom_density(alpha=.2,color="black", fill="#FF6666")
```

```
#SPDE mesh of the HMC
mesh1 = inla.mesh.create(hall_set[,c("X","Y")], refine = F, extend = F)
mesh <- make_mesh(hall_set, xy_cols = c("X", "Y"),mesh=mesh1) # SPDE
plot(mesh1, family = "serif", cex.main = 2, main = "")
points(cbind(hall_set$X, hall_set$Y), col = "orange", cex = 0.4)


# Normal vs t_2
fit_sdm<- sdmTMB(
  height ~ year,
  family = gaussian(link = "identity"), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)

saveRDS(fit_sdm,"fit_sdm.rds")

fit_sdmt <- sdmTMB(
  weight ~ year,
  family = student(link = "identity",df=2), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)

saveRDS(fit_sdmt,"fit_sdmt.rds")


# RQR plot for the HMC
rq_res1 <- residuals(fit_sdmt)
rq_res1 <- rq_res1[is.finite(rq_res1)]
qqnorm(rq_res1,xlab="Theoretical  df=2")
qqline(rq_res1)

rq_res2 <- residuals(fit_sdmt)
rq_res2 <- rq_res2[is.finite(rq_res2)]
qqnorm(rq_res2)
qqline(rq_res2)


# stratified sampling for the SH dataset
set.seed(111)
list_tow <- unique(hall_set$ID_TOW)
folds_tow <- caret::createFolds(list_tow, k = 10, list = T, returnTrain = F)
folds <- lapply(folds_tow, function(x) which(hall_set$ID_TOW %in% list_tow[x]))
```

```
hall_set$foldID=NA
hall_set$obs=c(1:length(hall_set[,1]))
hall_set$foldID[which(hall_set$obs %in% folds$Fold01)]=1
hall_set$foldID[which(hall_set$obs %in% folds$Fold02)]=2
hall_set$foldID[which(hall_set$obs %in% folds$Fold03)]=3
hall_set$foldID[which(hall_set$obs %in% folds$Fold04)]=4
hall_set$foldID[which(hall_set$obs %in% folds$Fold05)]=5
hall_set$foldID[which(hall_set$obs %in% folds$Fold06)]=6
hall_set$foldID[which(hall_set$obs %in% folds$Fold07)]=7
hall_set$foldID[which(hall_set$obs %in% folds$Fold08)]=8
hall_set$foldID[which(hall_set$obs %in% folds$Fold09)]=9
hall_set$foldID[which(hall_set$obs %in% folds$Fold10)]=10
hall_set$foldID=as.factor(hall_set$foldID)


# spatial distributions of observations from the SH dataset across 10 folds
heightcluter=ggscatter(
  hall_set, x = "lon", y = "lat",
  color = "foldID",  ellipse = TRUE, ellipse.type = "convex", shape="year",
  size = 1.5,  legend = "right", ggtheme = theme_bw(),
  xlab = paste0("lon" ),
  ylab = paste0("lat" )
) +facet_grid(~foldID)


# backward variable selection for the HMC
fit_sdmtDTSS<- sdmTMB(
  height ~ year+depth+temperature+stress+salinity,
  family = student(link = "identity", df = 2), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(fit_sdmtDTSS$sd_report, select = "fixed", p.value = TRUE)

fit_sdmtDTSa<- sdmTMB(
  height ~ year+depth+temperature+salinity,
  family = student(link = "identity", df = 2), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(fit_sdmtDTSa$sd_report, select = "fixed", p.value = TRUE)

fit_sdmtDT<- sdmTMB(
```

```
  height ~ year+depth+temperature,
  family = student(link = "identity", df = 2), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(fit_sdmtDT$sd_report, select = "fixed", p.value = TRUE)

fit_sdmtD<- sdmTMB(
  height ~ year+depth,
  family = student(link = "identity", df = 2), data = hall_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(fit_sdmtD$sd_report, select = "fixed", p.value = TRUE)

saveRDS(fit_sdmtD,"fit_sdmtD.rds")


# CV for potential models of HMC
hspm_cv <- sdmTMB_cv(
  height ~ year,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "off",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(hstm_cv$data,"hspm_cvdata")
h0=read.csv("hspm_cvdata.csv")

hstm_cv <- sdmTMB_cv(
  height ~ year,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
```

```
)

write.csv(hstm_cv$data,"hstm_cvdata")
h1=read.csv("hstm_cvdata.csv")

hstmD_cv <- sdmTMB_cv(
  height ~ year+depth,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(hstmD_cv$data,"hstmD_cvdata")
h2=read.csv("hstmD_cvdata.csv")

hstmT_cv <- sdmTMB_cv(
  height ~ year+temperature,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(hstmT_cv$data,"hstmT_cvdata")
h3=read.csv("hstmT_cvdata.csv")

hstmT_cv <- sdmTMB_cv(
  height ~ year+temperature,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
```

```
  use_initial_fit = FALSE
)

write.csv(hstmT_cv$data,"hstmT_cvdata")
h4=read.csv("hstmT_cvdata.csv")

hstmDT_cv <- sdmTMB_cv(
  height ~ year+temperature+depth,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(hstmDT_cv$data,"hstmT_cvdata")
h5=read.csv("hstmDT_cvdata.csv")

hstmDTSS_cv <- sdmTMB_cv(
  height ~ year+temperature+depth+stress+salinity,
  family = student(link = "identity",df=2), data = hall_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(hstmDTSS_cv$data,"hstmTSS_cvdata")
h6=read.csv("hstmDTSS_cvdata.csv")


# create a table for all potential model MSPEs of the HMC
h0=read.csv("hspm_cvdata.csv")
h1=read.csv("hstm_cvdata.csv")
h2=read.csv("hstmD_cvdata.csv")
h3=read.csv("hstmT_cvdata.csv")
h4=read.csv("hstmDT_cvdata.csv")
h5=read.csv("hstmDTSS_cvdata.csv")
```

```
hres <- data.frame(
  "lon"=h1$lon,
  "lat"=h1$lat,
  "Year"=as.factor(h1$year),
  "STM"=h1$ height-h1$cv_predicted,
  "STM-D"=h2$ height-h2$cv_predicted,
  "STM-T"=h3$ height-h3$cv_predicted,
  "STM-DT"=h4$ height-h4$cv_predicted,
  "STM-DTSS"=h5$ height-h5$cv_predicted
) %>%
  tidyr::gather(model,resid,-lon,-lat,-Year) %>%
  mutate(model = factor(model, ordered = T))

hres.sp <- hres %>%
  group_by(Year, lon, lat, model) %>%
  summarise(m.resid = mean(resid), sd.resid = sd(resid),
            m.abs.resid = mean(abs(resid)),m.sq.resid = mean((resid)^2)) %>%
  ungroup()
bind_rows(
  res %>%
    group_by(Year, model) %>%
    summarise(indiv.resid.mean = paste0(format(round(mean((resid^2)),4),
                                               nsmall=4, scientific=F))) %>%
    spread(model, indiv.resid.mean),
  hres %>%
    group_by(model) %>%
    summarise(indiv.resid.mean = paste0(format(round(mean((resid^2)),4),
                                               nsmall=4, scientific=F))) %>%
    spread(model, indiv.resid.mean) %>% mutate(Year = "2012-2019")
) %>%
  xtable::xtable() %>%
  print(include.rownames=F)



# create shape files for the study area and the Bay of Fundy
studyarea =readOGR(dsn = path.expand("BoF_Strata_extent4Joy.shp"),
                   layer = "BoF_Strata_extent4Joy")
shape1 <- readOGR(dsn = path.expand("SPA1A_polygon_NAD83.shp"),
                  layer = "SPA1A_polygon_NAD83")
shape2 <- readOGR(dsn = path.expand("SPA1B_polygon_NAD83.shp"),
                  layer = "SPA1B_polygon_NAD83")
shape3 <- readOGR(dsn = path.expand("SPA4_polygon_NAD83.shp"),
```

```
                  layer = "SPA4_polygon_NAD83")
shape4 <- readOGR(dsn = path.expand("SPA5_polygon_NAD83.shp"),
                  layer = "SPA5_polygon_NAD83")
studyarea <- union(studyarea,shape4)
writeSpatialShape(studyarea, "studyarea.ship")
studyarea =st_read("studyarea.shp")
subs_union1 <- union(shape1,shape2)
subs_union2 <- union(shape3,shape4)
BayofFundy=union(subs_union1,subs_union2)
writeSpatialShape(BayofFundy, "Bay of Fundy.ship")


# generate point locations for the shape file
sf_use_s2(FALSE)
set.seed(100)
times = 8
N=100000
nc_point <- st_sample(x = studyarea, size = N)
nc_point=as.matrix(nc_point)
nc_point <- do.call(rbind, st_geometry(nc_point)) %>%
  as_tibble() %>% setNames(c("lon","lat"))

points=as.data.frame(nc_point)%>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
  `attr<-`("zone", "20") %>%
  PBSmapping::convUL()
points=as.matrix(points)


# assign environmental values to all generated locations
mg=as.data.frame(nc_point)
mg$temperature=NA
mg$salinity=NA
mg$stress=NA

mg1=mg # for all environmental data in 2012
mg2=mg # for all environmental data in 2013
mg3=mg # for all environmental data in 2014
mg4=mg # for all environmental data in 2015
mg5=mg # for all environmental data in 2016
mg6=mg # for all environmental data in 2017
mg7=mg # for all environmental data in 2018
mg8=mg # for all environmental data in 2019
```

```
mg1$temperature=raster::extract(r2012, y = cbind(mg1$lon, mg1$lat))
mg2$temperature=raster::extract(r2013, y = cbind(mg2$lon, mg2$lat))
mg3$temperature=raster::extract(r2014, y = cbind(mg3$lon, mg3$lat))
mg4$temperature=raster::extract(r2015, y = cbind(mg4$lon, mg4$lat))
mg5$temperature=raster::extract(r2016, y = cbind(mg5$lon, mg5$lat))
mg6$temperature=raster::extract(r2017, y = cbind(mg6$lon, mg6$lat))
mg7$temperature=raster::extract(r2018, y = cbind(mg7$lon, mg7$lat))
mg8$temperature=raster::extract(r2019, y = cbind(mg8$lon, mg8$lat))

mg1$stress=raster::extract(s2012, y = cbind(mg1$lon, mg1$lat))
mg2$stress=raster::extract(s2013, y = cbind(mg2$lon, mg2$lat))
mg3$stress=raster::extract(s2014, y = cbind(mg3$lon, mg3$lat))
mg4$stress=raster::extract(s2015, y = cbind(mg4$lon, mg4$lat))
mg5$stress=raster::extract(s2016, y = cbind(mg5$lon, mg5$lat))
mg6$stress=raster::extract(s2017, y = cbind(mg6$lon, mg6$lat))
mg7$stress=raster::extract(s2018, y = cbind(mg7$lon, mg7$lat))
mg8$stress=raster::extract(s2019, y = cbind(mg8$lon, mg8$lat))

mg1$salinity=raster::extract(a2012, y = cbind(mg1$lon, mg1$lat))
mg2$salinity=raster::extract(a2013, y = cbind(mg2$lon, mg2$lat))
mg3$salinity=raster::extract(a2014, y = cbind(mg3$lon, mg3$lat))
mg4$salinity=raster::extract(a2015, y = cbind(mg4$lon, mg4$lat))
mg5$salinity=raster::extract(a2016, y = cbind(mg5$lon, mg5$lat))
mg6$salinity=raster::extract(a2017, y = cbind(mg6$lon, mg6$lat))
mg7$salinity=raster::extract(a2018, y = cbind(mg7$lon, mg7$lat))
mg8$salinity=raster::extract(a2019, y = cbind(mg8$lon, mg8$lat))


# expand mg to store environmental data across all years
rows= c(1:nrow(mg))
mg=mg[rep(rows, times),]
mg$year=c(rep(2012,N),rep(2013,N),rep(2014,N),rep(2015,N),
          rep(2016,N),rep(2017,N),rep(2018,N),rep(2019,N))
mg$temperature=c(mg1$temperature,mg2$temperature,mg3$temperature,
                 mg4$temperature,mg5$temperature,mg6$temperature,
                 mg7$temperature,mg8$temperature)
mg$salinity=c(mg1$salinity,mg2$salinity,mg3$salinity,mg4$salinity,
              mg5$salinity,mg6$salinity,mg7$salinity,mg8$salinity)
mg$stress=c(mg1$stress,mg2$stress,mg3$stress,mg4$stress,
            mg5$stress,mg6$stress,mg7$stress,mg8$stress)
mg$depth=raster::extract(draster, y = cbind(mg$lon, mg$lat))

mg= mg%>%filter(
```

```
  !is.na(depth),
  !is.na(temperature),
  !is.na(stress),
  !is.na(salinity)
) %>% transmute(
  year = as.factor(year),
  lon= lon, lat= lat ,
  depth = -depth ,
  temperature = temperature,
  stress=stress,
  salinity=salinity)
mg=mg[mg$depth>0,]


# make a study set with year and environmental information for all locations
hstudy_set <- mg %>%
  mutate(
    depth = as.numeric(((depth))),
    temperature = as.numeric(((temperature))),
    stress = as.numeric(((stress))),
    salinity = as.numeric(((salinity)))) %>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
  `attr<-`("zone", "20") %>%
  PBSmapping::convUL()

# graph for the SPAs
BayofFundy =st_read("BayofFundy.shp")
bbox1 <- c(left = min(myheight$lon)-1.5, bottom = min(myheight$lat)-1.5,
           right = max(myheight$lon)+1.5, top = max((myheight$lat)+1.5))
BD2=BayofFundy%>% dplyr::select(SP_ID, geometry)
ggmap(get_stamenmap(bbox1, maptype = "terrain-background")) +
  coord_sf(crs = hst_crs(3857)) +
  geom_point(data = hstudy_set, aes(x = lon, y = lat),
             size = 0.5, color = "pink")+
  geom_sf(data =  BD2, fill = NA,inherit.aes = FALSE)+
  theme_bw() +
  labs(x = "Lon", y = "Lat")


# graphs for environmental variable
base_map +
  geom_point(data = study_set, aes(x=lon, y=lat, color =  depth),shape = 20,
             size=0.5) +
```

```
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  labs(colour = "Depth (m)")

base_map +
  geom_point(data = study_set, aes(x=lon, y=lat, color = temperature),
             shape = 20,size=0.1) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  facet_grid(~year) + labs(colour = "Bottom temperature (°C)")

base_map +
  geom_point(data = study_set, aes(x=lon, y=lat, color = stress),
             shape = 20,size=0.1) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  facet_grid(~year) +
  labs(colour = expression("Bottom stress"* " (kg."*"m"^-1*".s"^-2*")"))

base_map +
  geom_point(data = study_set, aes(x=lon, y=lat, color =  salinity),
             shape = 20,size=0.1) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)  +
  facet_grid(~year) + labs(colour = "Bottom salinity (psu)")


# use STM-D for HMC to predict shell height in the study area
Studyheightt2STMD<- predict(fit_sdmtD,hstudy_set,
                            type = "response",re_form = NULL)
write.csv(Studyheightt2STMD,"Studyheightt2STMD.csv")
Studyheightt2STMD=read.csv("Studyheightt2STMD.csv")


# graphs for random effects in the HMC
base_map +facet_wrap(year ~ ., ncol = 8) +
  geom_point(aes(x =  lon , y = lat ,colour =(epsilon_st )),size = 0.1,
             data =Studyheightt2STMD , alpha =.5)  + scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white")+
  ggtitle("Spatiotemporal effect")

base_map  +
  geom_point(aes(x =  lon , y = lat ,colour =(omega_s )),size = 0.1,
             data =Studyheightt2STMD, alpha =.5)  + scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white") +
  ggtitle("Spatial effect")
```

```
base_map  + facet_wrap(year ~ ., ncol = 8) +
  geom_point(aes(x =  lon , y = lat ,colour =( est_rf  )),size = 0.1,
             data =Studyheightt2STMD, alpha =.5)  + scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white")+
  ggtitle("Spatial + Spatiotemporal effect")


# make a dataset for shell height prediction
pred.interpolation <- data.frame(
  "lon"=Studyheightt2STMD$lon,
  "lat"=Studyheightt2STMD$lat,
  "Year"=Studyheightt2STMD$year,
  "STMD"=Studyheightt2STMD$est
)
pred.interpolation=na.omit(pred.interpolation)


# graph for shell height prediction
base_map +
  geom_point(data = pred.interpolation, aes(x=lon, y=lat, color =STMD),
             shape=20,  size =0.05, alpha=0.5) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  facet_grid(~Year)+
  labs(colour = "Predicted shell heights (mm) from the HMC")


# graph for shell height sd errors
p <-predict(fit_sdmtD,type = "response",re_form = NULL,
            newdata=hstudy_set, nsim = 500)
predictor_dat=hstudy_set
predictor_dat$se<- apply(p, 1, sd)

base_map +
  geom_point(data = predictor_dat, aes(x=lon, y=lat, color =se),
             shape = 20,size=0.5) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  facet_grid(~year) +
  labs(colour = "Standard errors of the predicted shell heights from the HMC")
```

```
##############################################################################

# WMC and JWHM

# clean the MWSH dataset
scallops=read.csv("bof.mwsh.JuneJuly.2012to2019.clean.csv")
scallops=scallops%>% filter(month ==7)
scallops = scallops[!scallops$CRUISE == "RF2012",]
mydata= scallops%>% transmute(weight = WET_MEAT_WGT,
                              height = HEIGHT,
                              year = as.factor(year),
                              TOW_NO =TOW_NO ,
                              lon= mid.lon, lat= mid.lat
)
mydata$depth=raster::extract(draster, y = cbind(mydata$lon , mydata$lat))

obs1=mydata%>% filter(year ==2012)
obs2=mydata%>% filter(year ==2013)
obs3=mydata%>% filter(year ==2014)
obs4=mydata%>% filter(year ==2015)
obs5=mydata%>% filter(year ==2016)
obs6=mydata%>% filter(year ==2017)
obs7=mydata%>% filter(year ==2018)
obs8=mydata%>% filter(year ==2019)

obs1$salinity=raster::extract(a2012, y = cbind(obs1$lon, obs1$lat))
obs2$salinity=raster::extract(a2013, y = cbind(obs2$lon, obs2$lat))
obs3$salinity=raster::extract(a2014, y = cbind(obs3$lon, obs3$lat))
obs4$salinity=raster::extract(a2015, y = cbind(obs4$lon, obs4$lat))
obs5$salinity=raster::extract(a2016, y = cbind(obs5$lon, obs5$lat))
obs6$salinity=raster::extract(a2017, y = cbind(obs6$lon, obs6$lat))
obs7$salinity=raster::extract(a2018, y = cbind(obs7$lon, obs7$lat))
obs8$salinity=raster::extract(a2019, y = cbind(obs8$lon, obs8$lat))
mydata$salinity=c(obs1$salinity,obs2$salinity,obs3$salinity,obs4$salinity,
                  obs5$salinity,obs6$salinity,obs7$salinity,obs8$salinity)

obs1$temperature=raster::extract(r2012, y = cbind(obs1$lon, obs1$lat))
obs2$temperature=raster::extract(r2013, y = cbind(obs2$lon, obs2$lat))
obs3$temperature=raster::extract(r2014, y = cbind(obs3$lon, obs3$lat))
obs4$temperature=raster::extract(r2015, y = cbind(obs4$lon, obs4$lat))
obs5$temperature=raster::extract(r2016, y = cbind(obs5$lon, obs5$lat))
obs6$temperature=raster::extract(r2017, y = cbind(obs6$lon, obs6$lat))
obs7$temperature=raster::extract(r2018, y = cbind(obs7$lon, obs7$lat))
obs8$temperature=raster::extract(r2019, y = cbind(obs8$lon, obs8$lat))
```

```
mydata$temperature=c(obs1$temperature,obs2$temperature,obs3$temperature,
                     obs4$temperature,obs5$temperature,obs6$temperature,
                     obs7$temperature,
                     obs8$temperature)

obs1$stress=raster::extract(s2012, y = cbind(obs1$lon, obs1$lat))
obs2$stress=raster::extract(s2013, y = cbind(obs2$lon, obs2$lat))
obs3$stress=raster::extract(s2014, y = cbind(obs3$lon, obs3$lat))
obs4$stress=raster::extract(s2015, y = cbind(obs4$lon, obs4$lat))
obs5$stress=raster::extract(s2016, y = cbind(obs5$lon, obs5$lat))
obs6$stress=raster::extract(s2017, y = cbind(obs6$lon, obs6$lat))
obs7$stress=raster::extract(s2018, y = cbind(obs7$lon, obs7$lat))
obs8$stress=raster::extract(s2019, y = cbind(obs8$lon, obs8$lat))
mydata$stress=c(obs1$stress,obs2$stress,obs3$stress,obs4$stress,
                obs5$stress,obs6$stress,obs7$stress,obs8$stress)

mydata= mydata%>%filter(
  !is.na(depth),
  !is.na(temperature),
  !is.na(stress),
  !is.na(salinity)
) %>% transmute(weight = weight,
                height = height,
                depth=-depth,
                year = as.factor(year),
                lon= lon, lat= lat ,
                ID_TOW = as.factor(paste(year,TOW_NO,sep = '_')),
                depth = depth ,
                temperature = temperature,
                stress=stress,
                salinity=salinity)
# make sure no NA, all values should be positive


# make a UTM dataset for the MWSH dataset
all_set <- mydata %>%
  mutate(height = as.numeric((log(height))),
         depth = as.numeric((log(depth))),
         temperature = as.numeric((log(temperature))),
         stress = as.numeric((log(stress))),
         salinity = as.numeric((log(salinity)))) %>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
  `attr<-`("zone", "20") %>%
```

```
  PBSmapping::convUL()


# graph for height vs weight
h1<-ggplot(mydata, aes(x =height  , y =   weight)) +
  geom_point() +labs(x = "Height ", y = "Weight") +
  geom_smooth(method = "lm")

h2<-ggplot(mydata, aes(x =log(height)  , y =  log(weight))) +
  geom_point() +labs(x = "log Height ", y = "log Weight") +
  geom_smooth(method = "lm")

grid.arrange(h1,h2, nrow = , ncol=2)

# graph for environmental variables vs weight
h3<-ggplot(mydata, aes(x =( temperature) , y =  weight)) +
  geom_point() +labs(x = "Temperature ", y = "Weight" )

h4<-ggplot(mydata, aes(x = depth , y =   weight)) +
  geom_point() +labs(x = "Depth ", y = "Weight")


h5<-ggplot(mydata, aes(x = stress , y =   weight)) +
  geom_point() +labs(x = "Stress ", y = "Weight")


h6<-ggplot(mydata, aes(x = salinity , y =   weight)) +
  geom_point() +labs(x = "Salinity ", y = "Weight")

grid.arrange(h3,h4,h5,h6, nrow = 2, ncol=2)


# graph for log environmental variables vs log weight
h3<-ggplot(mydata, aes(x =log( temperature) , y =  log(weight))) +
  geom_point() +labs(x = "log Temperature ", y = "log Weight ")

h4<-ggplot(mydata, aes(x = log(depth) , y =  log(weight))) +
  geom_point() +labs(x = "log Depth ", y = "log Weight")


h5<-ggplot(mydata, aes(x = log(stress) , y =   log(weight))) +
  geom_point() +labs(x = "log Stress ", y = "log Weight")
```

```
h6<-ggplot(mydata, aes(x = log(salinity) , y =  log(weight))) +
  geom_point() +labs(x = "log Salinity ", y = "log Weight")

grid.arrange(h3,h4,h5,h6, nrow = 2, ncol=2)


# graph for environmental variable correlation in the MWSH dataset
par(mfrow=c(1,2))
corr_matrix <- mydata[,c(7,8,9,10)] %>%
  cor(method="pearson", use="pairwise.complete.obs")

corrplot(corr_matrix ,method="color", addCoef.col = "black",
         mar=c(0,0,5,0), tl.offset = 1)
mtext("Environmental variable correlations", at=2.5, line=-0.5, cex=0.8)

corr_matrix <- all_set[,c(7,8,9,10)] %>%
  cor(method="pearson", use="pairwise.complete.obs")

corrplot(corr_matrix ,method="color", addCoef.col = "black",
         mar=c(0,0,5,0), tl.offset = 1)
mtext("Environmental variable correlations after transformations",
      at=2.5, line=-0.5, cex=0.8)


# graph for mean weights in each tow location
mydata=mydata%>%
  group_by( ID_TOW  ) %>%
  mutate(mean_weight=mean(weight,na.rm=T))

base_map +
  geom_point(data = mydata, aes(x=lon, y=lat, color = mean_weight),
             shape = 20,size=0.05) +
  scale_color_gradientn(colours = rainbow(7), limits=c(3,50),
                        oob=scales::squish)  +
  labs(colour = "The mean meat weight (g) in each tow location")+
  facet_grid(~year)


# graph for weight distributions
w1=ggplot(mydata, aes(x=weight)) +
  geom_histogram(aes(y=..density..),
                 binwidth=.5,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666")
```

```
w2=ggplot(mydata, aes(x=log(weight))) +
  geom_histogram(aes(y=..density..),
                 binwidth=.05,
                 color="black",  fill="white") +
  geom_density(alpha=.2,color="black", fill="#FF6666")
grid.arrange(w1,w2, nrow = 1, ncol=2)


# SPDE mesh of the WMC
mesh1 = inla.mesh.create(all_set[,c("X","Y")], refine = F, extend = F)
plot(mesh1, family = "serif", cex.main = 2, main = "")
points(cbind(all_set$X, all_set$Y), col = "orange", cex = 0.4)
mesh <- make_mesh(all_set, xy_cols = c("X", "Y"),mesh=mesh1)


# Normal vs t_2
weight_sdm <- sdmTMB(
  weight ~ year+height,
  family = gaussian(link = "log"), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)

saveRDS(weight_sdm,"weight_sdm.rds")

weight_sdmt <- sdmTMB(
  weight ~ year+height,
  family = student(link = "log",df=2), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)

saveRDS(weight_sdmt,"weight_sdmt.rds")


#WMC RQR plot
rq_res1 <- residuals(weight_sdmt)
rq_res1 <- rq_res1[is.finite(rq_res1)]
qqnorm(rq_res1,xlab="Theoretical  df=2")
qqline(rq_res1)

rq_res2 <- residuals(weight_sdm)
rq_res2 <- rq_res2[is.finite(rq_res2)]
```

```
qqnorm(rq_res2)
qqline(rq_res2)


# stratified sampling for the MWSH dataset
set.seed(111)
list_tow <- unique(all_set$ID_TOW)
folds_tow <- caret::createFolds(list_tow, k = 10, list = T, returnTrain = F)
folds <- lapply(folds_tow, function(x) which(all_set$ID_TOW %in% list_tow[x]))

all_set$foldID=NA
all_set$obs=c(1:length(all_set[,1]))
all_set$foldID[which(all_set$obs %in% folds$Fold01)]=1
all_set$foldID[which(all_set$obs %in% folds$Fold02)]=2
all_set$foldID[which(all_set$obs %in% folds$Fold03)]=3
all_set$foldID[which(all_set$obs %in% folds$Fold04)]=4
all_set$foldID[which(all_set$obs %in% folds$Fold05)]=5
all_set$foldID[which(all_set$obs %in% folds$Fold06)]=6
all_set$foldID[which(all_set$obs %in% folds$Fold07)]=7
all_set$foldID[which(all_set$obs %in% folds$Fold08)]=8
all_set$foldID[which(all_set$obs %in% folds$Fold09)]=9
all_set$foldID[which(all_set$obs %in% folds$Fold10)]=10
all_set$foldID=as.factor(all_set$foldID)


# spatial distributions of observations from the MWSH dataset across 10 folds
weightcluter=ggscatter(
  all_set, x = "lon", y = "lat",
  color = "foldID",  ellipse = TRUE, ellipse.type = "convex", shape="year",
  size = 1.5,  legend = "right", ggtheme = theme_bw(),
  xlab = paste0("lon" ),
  ylab = paste0("lat" )
) +facet_grid(~foldID)


# backward variable selection for the WMC
weight_sdmtDTSS<- sdmTMB(
  weight ~ year+ height+depth+temperature+stress+salinity,

  family = student(link = "log", df = 2), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(weight_sdmtDTSS$sd_report, select = "fixed", p.value = TRUE)
```

```
weight_sdmtDTSt<- sdmTMB(
  weight ~ year+ height+depth+temperature+stress,

  family = student(link = "log", df = 2), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(weight_sdmtDTSt$sd_report, select = "fixed", p.value = TRUE)

weight_sdmtDT<- sdmTMB(
  weight ~ year+ height+depth+temperature,
  family = student(link = "log", df = 2), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(weight_sdmtDT$sd_report, select = "fixed", p.value = TRUE)

weight_sdmtD<- sdmTMB(
  weight ~ year+ height+depth,
  family = student(link = "log", df = 2), data = all_set, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid",
  share_range=FALSE
)
summary(weight_sdmtD$sd_report, select = "fixed", p.value = TRUE)
saveRDS(weight_sdmtD,"weight_sdmtD.rds")


# CV for potential models of WMC
spm_cv <- sdmTMB_cv(
  weight ~ year+height,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "off",
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(spm_cv$data,"spm_cvdata.csv")
w0=read.csv("spm_cvdata.csv")
r0=mean((w0$weight-w0$cv_predicted)^2)
```

```
stm_cv <- sdmTMB_cv(
  weight ~ year+height,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(stm_cv$data,"stm_cvdata")
w1=read.csv("stm_cvdata.csv")
r1=mean((w1$weight-w1$cv_predicted)^2)

stmD_cv <- sdmTMB_cv(
  weight ~ year+height+depth,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)
write.csv(stmD_cv$data,"stmD_cvdata")
w2=read.csv("stmD_cvdata.csv")

stmT_cv <- sdmTMB_cv(
  weight ~ year+height+temperature,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(stmT_cv$data,"stmT_cvdata")
w3=read.csv("stmT_cvdata.csv")
```

```
stmDT_cv <- sdmTMB_cv(
  weight ~ year+height+depth+temperature,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(stmDT_cv$data,"stmDT_cvdata")
w4=read.csv("stmDT_cvdata.csv")

stmDTSS_cv <- sdmTMB_cv(
  weight ~ year+height+depth+temperature+salinity+stress,
  family = student(link = "log",df=2), data = all_set,
  time = "year",  spatial = "on", spatiotemporal = "iid",
  share_range=FALSE,
  mesh = mesh,
  fold_ids = "foldID",
  k_folds = 10,
  parallel = TRUE,
  use_initial_fit = FALSE
)

write.csv(stmDTSS_cv$data,"stmDTSS_cvdata")
w5=read.csv("stmDTSS_cvdata.csv")


# MSPE for all potential model of the WMC
w0=read.csv("spm_cvdata.csv")
w1=read.csv("stm_cvdata.csv")
w2=read.csv("stmD_cvdata.csv")
w3=read.csv("stmT_cvdata.csv")
w4=read.csv("stmDT_cvdata.csv")
w5=read.csv("stmDTSS_cvdata.csv")

res <- data.frame(
  "lon"=w1$lon,
  "lat"=w1$lat,
  "Year"=as.factor(w1$year),
```

```
  "SM"=w0$ weight-w0$cv_predicted,
  "STM"=w1$ weight-w1$cv_predicted,
  "STM-D"=w2$ weight-w2$cv_predicted,
  "STM-T"=w3$ weight-w3$cv_predicted,
  "STM-DT"=w4$ weight-w4$cv_predicted,
  "STM-DTSS"=w5$ weight-w5$cv_predicted
) %>%
  tidyr::gather(model,resid,-lon,-lat,-Year) %>%
  mutate(model = factor(model, ordered = T))

# WMC MSPE table
res.sq <- res %>%
  dplyr::group_by(Year, lon, lat, model) %>%
  dplyr::summarise(m.resid = mean(resid), m.abs.resid = mean(abs(resid)),
                   m.sq.resid = mean((resid)^2)) %>%
  ungroup()

bind_rows(
  res %>%
    dplyr::group_by(Year, model) %>%
    dplyr::summarise(indiv.resid.mean = paste0(format(round(mean((resid^2)),4),
                                                      nsmall=4, scientific=F))) %>%
    spread(model, indiv.resid.mean),
  res %>%
    group_by(model) %>%
    dplyr::summarise(indiv.resid.mean = paste0(format(round(mean((resid^2)),4),
                                                      nsmall=4, scientific=F))) %>%
    spread(model, indiv.resid.mean) %>% mutate(Year = "2012-2019")
) %>%
  xtable::xtable() %>%
  print(include.rownames=F)



# make a study set with a fixed shell height
mg$height=c(rep((mean(mydata$height) )))
study_set <- mg %>%
  mutate(height = as.numeric((log(height))),
         depth = as.numeric((log(depth))),
         temperature = as.numeric((log(temperature))),
         stress = as.numeric((log(stress))),
         salinity = as.numeric((log(salinity)))) %>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
```

```
  `attr<-`("zone", "20") %>%
  PBSmapping::convUL()


# use STM-D for both the WMC and the JWHM
weight_sdmtD=readRDS("weight_sdmtD.rds")


# meat weight predictions by the WMC using the fixed height
Studyweightt2STMD<- predict(weight_sdmtD,study_set,type = "response",
                            re_form = NULL)
write.csv(Studyweightt2STMD,"Studyweightt2STMD.csv")


# graphs for random effects in the WMC
base_map +facet_wrap(year ~ ., ncol = 8)+
  geom_point(aes(x =  lon , y = lat ,colour =(epsilon_st)),
             size = 0.1, data =Studyweightt2STMD , alpha =.5)+
  scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white")+
  ggtitle("Spatiotemporal effect")

base_map  +
  geom_point(aes(x =  lon , y = lat ,colour =(omega_s)),
             size = 0.1, data =Studyweightt2STMD , alpha =.5)+
  scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white")+
  ggtitle("Spatial effect")

base_map  + facet_wrap(year ~ ., ncol = 8)+
  geom_point(aes(x =  lon , y = lat ,colour =(est_rf)),
             size = 0.1, data =Studyweightt2STMD , alpha =.5)+
  scale_color_continuous()+
  theme(text=element_text(size=11, family="serif"))+
  scale_color_gradient2(low = "blue", high = "red", mid = "white")+
  ggtitle("Spatial + Spatiotemporal effect")



#make a study set with predicted shell heights from the HMC
Studyheightt2STMD=read.csv("Studyheightt2STMD.csv")
colnames(Studyheightt2STMD)[colnames(Studyheightt2STMD) == 'est'] <- 'height'
```

```
Studyheightt2STMD=Studyheightt2STMD[,c(2:11)]
Studyheightt2STMD$year=as.factor(Studyheightt2STMD$year)
Studyheightt2STMD=Studyheightt2STMD%>%
  mutate(height = as.numeric((log(height))),
         depth = as.numeric((log(depth))),
         temperature = as.numeric((log(temperature))),
         stress = as.numeric((log(stress))),
         salinity = as.numeric((log(salinity)))) %>%
  mutate(X=lon, Y=lat) %>%
  `attr<-`("projection", "LL") %>%
  `attr<-`("zone", "20") %>%
  PBSmapping::convUL()


# meat weight prediction by the JWHM using  predicted shell heights from the HMC
Studywh2STMD<- predict(weight_sdmtD,Studyheightt2STMD,
                       type = "response",re_form = NULL)
write.csv(Studywh2STMD,"Studywh2STMD.csv")


# make a set for meat weight prediction comparison between the WMC and the JWHM
pred.interpolation.comp <- data.frame(
  "lon"=pred.interpolation$lon,
  "lat"= pred.interpolation$lat,
  "Year"=pred.interpolation$Year,
  "WMC"=pred.interpolation$weightSTMD ,
  "JWHM"=whpred.interpolation$STMD
) %>%
  gather(model,pmw,-lon,-lat,-Year) %>%
  mutate(model = factor(model,
                        levels = c("WMC", "JWHM")))


# graph for meat weight prediction comparison between the WMC and the JWHM
base_map +
  geom_point(data = pred.interpolation.comp, aes(x=lon, y=lat, color = pmw),
             shape=20,  size =0.05, alpha=0.5) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish)+
  facet_grid(model~Year) +
  labs(colour = "Predicted meat weights (g)")


# sd errors for the WMC and the JWHM
pw <-predict(weight_sdmtD,type = "response",re_form = NULL,
```

```
             newdata = study_set, nsim = 500)
wpredictor_dat=study_set
wpredictor_dat$se<- apply(pw, 1, sd)
colnames(wpredictor_dat)[11] <- "WMC"

pwh <-predict(weight_sdmtD,type = "response",re_form = NULL,
              newdata = Studyheightt2STMD, nsim = 500)
whpredictor_dat=Studyheightt2STMD
whpredictor_dat$se<- apply(pwh, 1, sd)
colnames(whpredictor_dat)[11] <- "JWHM"


# make a set for sd error comparison between the WMC and the JWHM
sedata <- data.frame(
  "lon"=whpredictor_dat$lon,
  "lat"=whpredictor_dat$lat,
  "Year"=whpredictor_dat$year,
  "WMC"=wpredictor_dat$WMC ,
  "JWHM"=whpredictor_dat$JWHM
) %>%
  gather(model,se,-lon,-lat,-Year) %>%
  mutate(model = factor(model,
                        levels = c("WMC", "JWHM")))


# graph for sd comparison between the WMC and the JWHM
base_map +
  geom_point(data = sedata, aes(x=lon, y=lat, color = se),
             shape=20,  size =0.05, alpha=0.5) +
  scale_color_gradientn(colours = rainbow(7),oob=scales::squish) +
  facet_grid(model~Year) +
  labs(colour = "Standard errors of the predicted meat weight")
```