

COMPARATIVE STATISTICAL ANALYSIS OF SEA SURFACE TEMPERATURE TIME
SERIES MODELING IN THE GULF OF MEXICO

A Thesis

by

MADISON RIBA

BS, Texas A&M University-Corpus Christi, 2022

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

Texas A&M University-Corpus Christi
Corpus Christi, Texas

December 2023

© Madison Riba
All Rights Reserved
December 2023

COMPARATIVE STATISTICAL ANALYSIS OF SEA SURFACE TEMPERATURE TIME
SERIES MODELING IN THE GULF OF MEXICO

A Thesis

by

MADISON RIBA

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

Lei Jin, PhD
Chair

Xinping Hu, PhD
Committee Member

José Guardiola, PhD
Committee Member

Alexey Sadovski, PhD
Graduate Faculty Representative

December 2023

ABSTRACT

This work provides an analysis of sea surface temperature (SST) data in the Gulf of Mexico (GoM), and compares modeling methods to assess their performance through statistical simulation. The GoM is a region of global economic and environmental importance, including the notable Flower Garden Banks National Marine Sanctuary's coral habitats. However, the area has largely not been focused on in terms of advanced SST modeling. The objective of this work is to provide insight into SST variability in the GoM by modeling time series data, concentrating on seasonal patterns and the overall linear warming trend. In the study, varying methods compared include a dynamic seasonal mean model, a dynamic sinusoidal regression model, a cyclic spline model, and a seasonal mean regression model with sandwich estimator. At the 95% nominal confidence level, methods ignoring the autocorrelation of error terms maintain empirical coverage between 60% and 80% to capture the linear yearly warming trends in the simulation. The recommended modeling approach, a dynamic sinusoidal regression model, provides such coverage of approximately 91.6%. It also offers a balance between accuracy, simplicity, and continuity, especially for frequently collected data. Analysis estimates the East Flower Garden Bank has an annual warming trend of approximately $0.028 \pm 0.016^{\circ}\text{C}$. This comparative analysis of modeling techniques contributes to a better understanding of SST dynamics in this critical marine ecosystem.

ACKNOWLEDGEMENTS

I would first like to thank my committee for their time and assistance throughout this project. I want to express my deepest gratitude my thesis advisor, Dr. Lei Jin. This project would not have been possible without his guidance, and I am thankful to have had such a caring and dedicated advisor. I am thankful for Dr. Xinping Hu for offering his unique scientific perspective, lab access, and kindness, and I am grateful Dr. José Guardiola for his advice and compassion. I send my thanks to the faculty of the Department of Mathematics and Statistics as well as its chair, Dr. Sadovski, for their continued support throughout my education.

I appreciate my peers and STAGES program colleagues for their support inside and outside of the classroom. I am eternally grateful to my family, friends, and my partner for always putting up with me. I could not have done this without them.

This material is based upon work supported by the National Science Foundation under Grant No. DGE-2152131. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

The Surface Ocean CO₂ Atlas (SOCAT) is an international effort, endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS), and the Integrated Marine Biosphere Research (IMBeR) program, to deliver a uniformly quality-controlled surface ocean CO₂ database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT.

TABLE OF CONTENTS

	Page
ABSTRACT	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	vii
LIST OF TABLES	viii
1. INTRODUCTION	1
2. DATA	4
2.1 Data manipulation	5
3. METHODS	11
3.1 Statistical models	12
3.1.1 Dynamic seasonal mean model	12
3.1.2 Cyclic spline model	14
3.1.2.1 Dynamic sinusoidal regression model	15
3.1.2.2 Regression model with sandwich method	17
3.1.2.3 Thin spline spatial model	18
4. RESULTS	20
4.1 Simulation	20
4.1.1 Data analysis	24
5. CONCLUSION	31
5.1 Summary	31
5.2 Discussion	31
REFERENCES	33

LIST OF FIGURES

	Page
2.1 Cyclic splines of SST and CO ₂ of SOCAT observations.	6
2.2 Maps of SOCAT SST observation locations.	9
2.3 Histograms of grouped SOCAT data.	10
2.4 Histograms of daily-aggregated data by month.	10
3.1 Monthly SST from January 16, 1989, to December 16, 2020, at the East Flower Garden Bank.	11
3.2 Monthly-aggregated SST observations in the three FGB locations.	16
3.3 Maps of observational data used to create the thin plate spline model.	19
4.1 Fitted cyclic spline models to observed data.	26
4.2 ACF plot of residuals from cyclic spline model of centered observed data.	26
4.3 Contour plots of thin plate spline fit for SST in the Western GoM.	30

LIST OF TABLES

	Page
2.1 List of variables of original SOCAT data.	8
2.2 SOCAT data variables after manipulation for analysis.	9
4.1 Estimates of monthly coefficients from the linear model of EFGB observed SST data.	21
4.2 Simulation results: MAE, variance, and the CI capture proportion of various models.	28
4.3 Estimated linear yearly trend and associated 95% CIs of EFGB observed data.	29
4.4 Estimates of coefficients from the dynamic monthly means model of EFGB observed data.	29
4.5 Estimates of coefficients from the dynamic sinusoidal regression model of EFGB observed data.	29

1. INTRODUCTION

Sea surface temperature (SST) is a useful marker for understanding marine health and coastal conditions (O'carroll et al., 2019). SST is an important covariate for many different oceanic conditions, such as CO₂, which can be used to assess oceanic acidification. In addition, changes in SST can affect climate conditions including storm intensification. Warmer oceanic temperatures also lead to the acceleration of thermal expansion.

The Gulf of Mexico (GoM) is the world's largest gulf, encompassing approximately 600,000 square miles. Of great economic, ecological, and biological importance, the GoM is a major source of oil and petroleum deposits and is home to many species of marine life such as sea turtles and corals. Although SST modeling has been performed in many different areas, the GoM is typically not analyzed despite its importance.

The interest in comprehensive comparative analysis derives from the myriad modeling methods currently used throughout the world's oceans. This work attempts to compare multiple methods of SST modeling in the GoM to determine accuracy.

SST has been modeled across the world in different ways. In the Chinese Bohai Sea, there is evidence of seasonal duration changes over time (Yuan et al., 2023). Seasons were determined by the warmest and coolest percentiles of the inter-annual temperature range rather than by date. Spring and fall were shown to be contributing factors for the magnitude of the lengthening and shortening trend, most notably the wind speeds and cloud cover throughout the seasonal transition periods.

In Australia, many analyses have been performed for SST. In Foster et al. (2014), remotely-sensed data analysis was conducted using a seasonally decomposed linear model. Error terms are random and normally distributed. Outliers were determined using a spline interpolation. Most waters surrounding Australia warmed between 0 and 0.5°C per decade, with deeper waters warming faster than coastal waters. Currents seem to play a large role in warming patterns.

SST analysis has also been performed in North America. In San Francisco, many data sets were simulated and modeled for different scenarios (Bashevkin et al., 2022). To generate the

models, each component received individual treatment. Residuals used an AR(1) model, spatial components used a thin plate spline, the temperature slope was modeled using a tensor product, seasonality used a cyclic spline, and time-of-day was corrected using a thin plate smooth function. Time periods were varied for different analyses, which resulted in more accurate predictions for each interval. Certain areas tended to warm quicker than others, specifically the northernmost channels. Generally, winter has been warming the most seasonally.

In the Caribbean, regional averages of satellite data were aggregated into a time series for analysis of SST and precipitation (Glenn et al., 2015). A correlation was found between ENSO (El Niño-Southern Oscillation) and warmer SST. Late rainfall season is warming the fastest at 0.0209°C per year, with no significant correlation with dry seasons. The Atlantic Warm Pool is growing over time in the late rainfall season. Hydrological cycles are changing, which are associated with an increase in SST and precipitation.

In the GoM, Empirical Orthogonal Function analysis was conducted on SOCAT data from 1996-2017 (Kealoha et al., 2020). The GoM was broken up into different subsections: three coastal and six open ocean areas. CO_2 and SST were found to frequently be correlated, in addition to production and salinity. The article explains some biological drivers of variation in these relationships, with the Mississippi Delta being a large catalyst for spatial changes as well as seasonality. Coastal environments tended to fluctuate in conditions much more than open ocean environments. Regarding CO_2 , different environments have different levels – certain areas are CO_2 sources while others are sinks. The relationship between CO_2 is typically negative between SST and salinity but can be positively correlated with salinity in some areas. Data collection was inconsistent over time. The ocean is becoming acidified (coastal quicker than the ocean), with the main factors most likely being human activity and an increase in SST.

Globally, the relationship between SST and Chlorophyll-A levels has been investigated (Dunstan et al., 2018). Splines were used with varying knot points to determine deviating behavior. The change in covariance between SST And Chl-a is associated with external factors such as wind and precipitation. Generally, SST follows a Gaussian distribution, while Chl-a follows a Gamma

distribution. It has also been found through linear analysis of global SST averages that more warming occurs in the warmer months than in the cooler months (Gil-Alana et al., 2019).

Different methods have been developed for SST analysis. The most common types of analysis used for SST include autoregressive–moving-average (ARMA) and spline models. B-splines have been used to predict SST in the East and South China Seas (Liu et al., 2021). In this study, a model that combined components of B-splines, attention mechanisms, and long short-term memory (LSTM) networks was the most accurate and useful for interpolation.

The motivation for this work is to apply multiple existing models to a different context and area. SST has been modeled in many ways in previous work, and this work aims to synthesize and compare those model types. This work focuses on modeling specifically in the GoM. Much of the previous work focuses on areas in the Pacific Ocean, so performing analysis in a different body of water will further investigate the inter-oceanic variation of SST behavior. Additionally, the novel introduction of statistical inference and simulation will provide more insight into the precision of the models created. Because the GoM is an area of national and global importance, improving SST modeling in the area can help aid researchers whose work depends on these models. Re-contextualizing and further developing extant models will bring further understanding of SST in the region.

2. DATA

In this work, multiple data sets will be used for analysis. Each has its own benefits and disadvantages.

One data set of this work originates from the Surface Ocean CO₂ Atlas (SOCAT) (Bakker et al., 2023). This database is an aggregation of oceanic data collected around the world. An advantage of the use of this database is that despite data coming from multiple sources, all data are quality-controlled and collected in situ rather than using interpolations generated by remote sensing. In Foster et al. (2014), satellite observations tended to be cooler than drifter observations. In utilizing the SOCAT, this disparity is eliminated.

Data were accessed online through the NOAA Live Access Server. Location was used to filter data only in the GoM: the boundaries set were (18, 32)[°]N and (80, 100)[°]W. The time span of the data set ranges from 1992 to 2022. Conditions including time, location, SST, CO₂, bathymetry, and pressure were recorded. In total, the data set contains 960,993 observations throughout time and space. To reduce outliers, observations with temperature recordings < 10°C were excluded from the final data set. The resulting final, filtered data set contains 960,948 observations from January 22, 1996, to December 8, 2022.

There is much variation in both the observational locations and observation frequency between years. Histograms of the filtered SOCAT data in all locations grouped by different criteria are displayed in Figure 2.3. Temporally, the data are unequally spaced. The distribution of observations by time of day is approximately uniform, with 01:00-05:00 UTC slightly less frequent. Regarding the distribution of observations throughout the year, the most frequently observed months include September, July, and August, and the least frequently observed months include December, January, and February. The majority of observations were recorded in the 21st century, with the most observations occurring in 2009, 2013, and 2017. Observations collected in 2020 are very limited.

A map of SST observations throughout the GoM contained in the SOCAT in select years are seen in Figure 2.2. The most common collection areas include the western coastline of Florida due to a common docking location as well as coastal Louisiana. In later years, gulf-wide data collection

occurred more often. The western Florida coast has been consistently observed throughout the years, and the Gulf shoreline has been observed more often in recent years.

In addition to the SOCAT database, monthly observational data collected at the Flower Garden Banks National Marine Sanctuary (FGB) were also accessed (Johnston et al., 2022). This data set allows for the analysis of consistent observations through time at specific locations. The time frame of the data spans from January 1989 to June 2021, resulting in 378 total observations. This is considerably less dense than the SOCAT data. Across three locations, the Stetson Bank (STET), the East Flower Garden Bank (EFGB), and the West Flower Garden Bank (WFGB), SST observations were taken at 00:00 CST on the 16th day of each month regardless of daylight savings time. This resulted in an equidistant data set of observations.

2.1 Data manipulation

Tables 2.1 and 2.2 depict the difference between the imported SOCAT data set and the working data set used for analysis. As seen in Table 2.1, there are 27 variables in the original SOCAT data. The variables of interest for this work include SST, CO₂, time (date and time, year, month, and hour), and location. These variables were selected from the original data set, and a date variable was created. This resulted in the creation of the working data set as seen in Table 2.2.

With the large volume of observations in the SOCAT data set, using data manipulation to aggregate data in different ways proved useful for summarizing. Data were stored by averaging observations grouped by different variables into respective data sets: group by date, month, year, and time of day were the primary methods.

The cyclic splines modeling intra-annual variability of SST and CO₂ from the SOCAT data set are shown in Figure 2.1. For SST, temperatures are warmest on average in early-mid August and coolest in mid-February. CO₂ behaves similarly, with the greatest value in August and the least value in early February. However, the CO₂ data are less smooth in pattern - there are local maxima in early January, mid-May, early August, and late September. This can be due to the varying activity of organisms throughout a given year whose productivity contributes to fluctuations in CO₂ levels. Within a day, both SST and CO₂ peak in the evening, with SST being greatest around 21:00 and

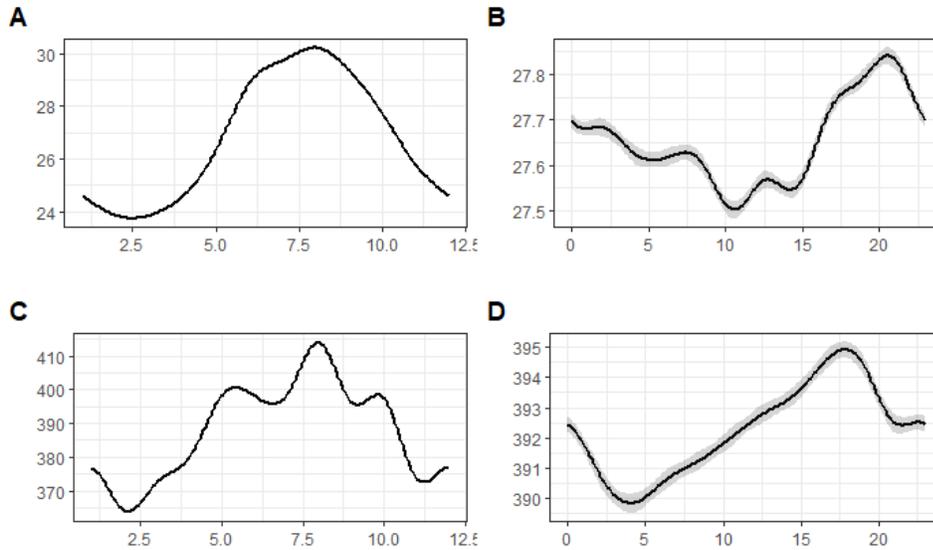


Figure 2.1
 Cyclic splines of SST and CO₂ of SOCAT observations. A: SST aggregated by month; B: SST aggregated by hour; C:CO₂ aggregated by month; D:CO₂ aggregated by hour.

CO₂ around 18:00. The lowest CO₂ levels occur in the early morning around 04:00, and SST has four local minima located approximately at 00:30, 05:00, 10:30 and 14:30. The lack of smoothness could be caused by varying atmospheric factors such as wind, which leads to mixing and changes in SST.

It is obvious from the graphs that these trends are highly nonlinear. Therefore, nonlinear modeling types will be needed to estimate smaller-scale effects of the data.

Because certain models are spatial in nature, sub-setting the data in a way that contains observations throughout the study area was useful in the model-building process. The subset used for this analysis was determined as the collection of observations since 2010, aggregated daily. Histograms of observations throughout different years from 2012 until 2022 are shown in Figure 2.4. Overall, the most common points of the year in which data are collected are in the spring and late summer. However, the temporal distribution of these observations varies greatly from year to year. Due to the irregularity of observations throughout time, the aggregation process allows for the data set to be more complete. Additionally, due to the strong autocorrelation of these data, creating a new, more regularly-spaced aggregated data set controls this effect between observations. It is

important to note that instances with fewer observations (such as December observations aggregated monthly) can account less for linear long-term trends than more observed instances (such as September observations) due to less prevalent data throughout the years. If winter expeditions only occurred from 2000-2005, controlling for yearly trends outside of that time frame is not possible. Otherwise, if summer expeditions were completed each year, the yearly change in averages would be accounted for in the averaging process for aggregation.

The inclusion of both SOCAT and FGB data allows for the comparison of modeling methods for data sets of different natures. SOCAT data include a spatial component, are much more observation-dense, and are irregularly spaced. FGB data are regularly spaced, simpler, and contain significantly fewer observations. Models that are spatial, such as thin plate splines, can be created using the SOCAT database, and models that favor regularity can be utilized with the FGB data.

	Observation 1	Observation 2
DATETIME	20-MAR-2011 00:17	20-MAR-2011 06:17
LATITUDE	30	30
LONGITUDE	-88	-88
SOCAT_DOI	10.1594/PANGAEA.853509	10.1594/PANGAEA.853509
QC_FLAG	C	C
SOCAT_VERSION	3.0N	3.0N
DEPTH	-1e+34	-1e+34
YEAR	2011	2011
MONTH	3	3
DAY	20	20
HOUR	0	6
MINUTE	17	17
SECOND	0	0
SAL	27.724	27.080
TEMP	20.507	20.951
TEMPERATURE_EQUI	20.507	20.951
PRESSURE_ATM	-1e+34	-1e+34
PRESSURE_EQUI	1023.1	1024.3
WOA_SSS	32.9525	32.9525
PRESSURE_NCEP_SLP	1024.1	1024.4
ETOPO2	24.99972	24.99972
DIST_TO_LAND	55.65976	55.65976
GVCO2	395.769	395.769
FCO2_RECOMMENDED	341.1743	346.6720
FCO2_SOURCE	2	2
WOCE_CO2_WATER	2	2
DELTA_TEMP	0	0

Table 2.1

List of variables of original SOCAT data.

	Observation 1	Observation 2
lon	-88	-88
datetime	2011-03-20 00:17:00	2011-03-20 06:17:00
date	2011-03-20	2011-03-20
year	2011	2011
month	3	3
hour	0	6
code	316420110319	316420110319
temp	20.507	20.951
co2	341.1743	346.6720

Table 2.2
SOCAT data variables after manipulation for analysis.

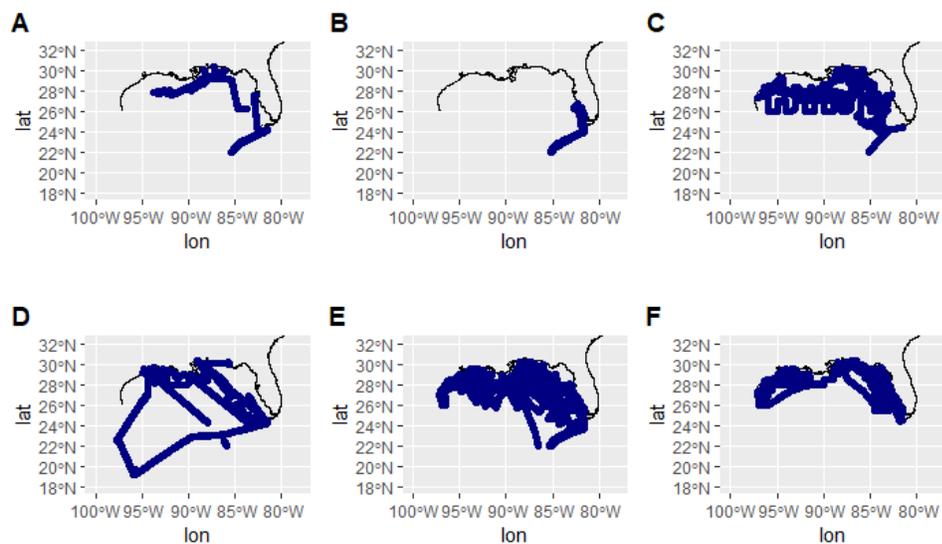


Figure 2.2
Maps of SOCAT SST observation locations. A: 2003, B: 2005, C:2008, D:2013, E:2018, F:2022.

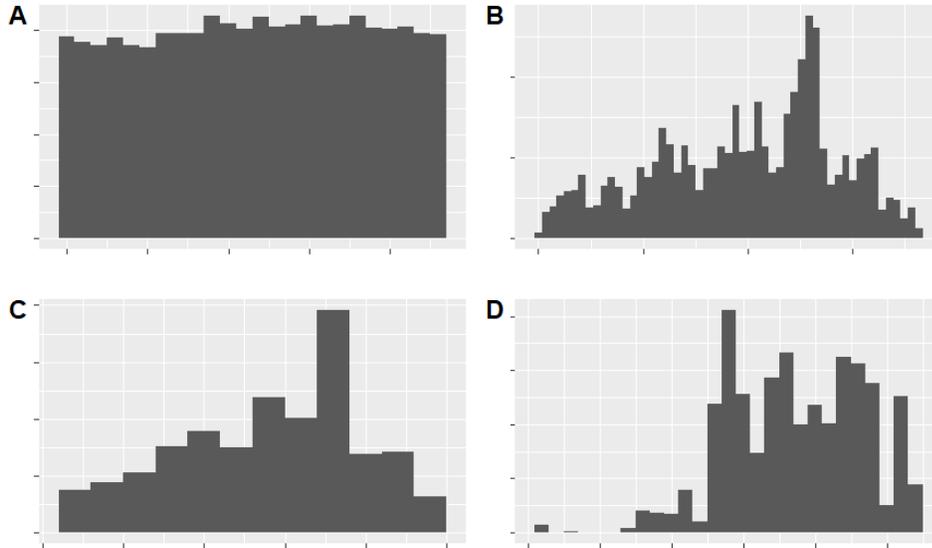


Figure 2.3
Histograms of grouped SOCAT data. A: time of day, B: day of year, C: month, D: year.

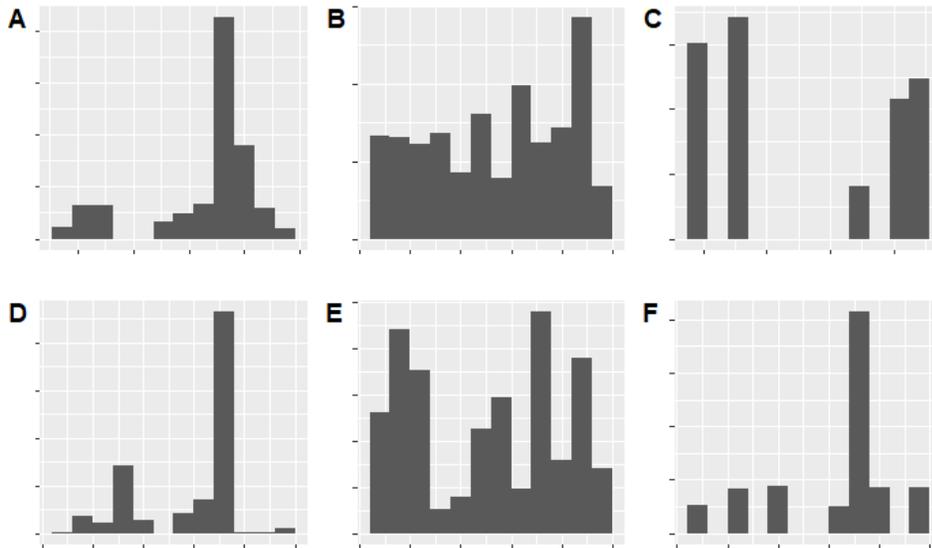


Figure 2.4
Histograms of daily-aggregated data by month. A: 2003, B: 2005, C: 2008, D: 2013, E: 2018, F: 2022.

3. METHODS

A plot of observed monthly SST at the EFGB from 1989-2022 is depicted in Figure 3.1. As seen in the plot, SST has one minimum and one maximum value each year. Over the observed time domain, yearly maxima occur on average in August and yearly minima occur in February. Also of note is the visible warming trend throughout the years, most notably shown in the winter months. These two features of the data are the modeling methods' focus - an overall linear trend and a seasonal yearly pattern, including an error term for modeling randomness not captured in these elements.

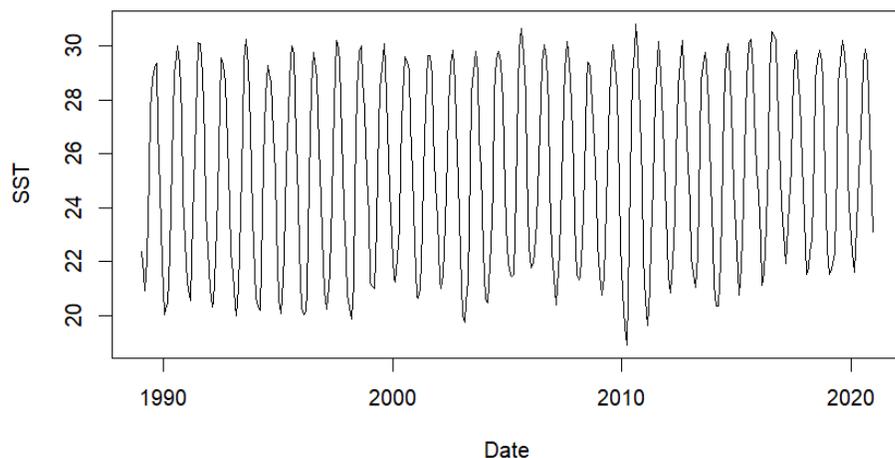


Figure 3.1
Monthly SST from January 16, 1989, to December 16, 2020, at the East Flower Garden Bank.

SST temporal variation is generally composed of seasonal and spatial effects, a yearly trend, and randomness. Hence, the initial model for SST is as follows:

$$Y_i = g(t_i) + \beta t_i + e_i, \quad i = 1, \dots, n. \quad (3.1)$$

In this model, Y_i refers to the SST of the i -th observation in temporal order, t_i refers to the number of years elapsed since t_1 , β refers to the linear yearly trend of SST, $g(\cdot)$ refers to the nonlinear seasonal effects within each year, and e_i refers to the dependent errors. This model has the form of a partially linear model (PLM). For more information about PLMs, refer to Hardle et al. (2000).

An advantage of using a PLM for modeling SST over time is its ability to isolate an overall linear warming trend while also including nonlinear aspects of the data. Relationships between variables can be complex and nonlinear, especially with environmental data. The use of PLMs allows for the inclusion of both linear and nonlinear relationships in the same model. Across the observed locations, SST behaves in a similar yet nonlinear fashion month-to-month, as depicted in Figure 3.2. This suggests a strong seasonal trend within each year. This trend will be captured in the nonlinear term $g(\cdot)$.

Typically, the error terms of a PLM comprise of independent and identically distributed (i.i.d.) random variables with mean zero and finite variances $\sigma_i^2 = E e_i^2$. Due to the autocorrelation of SST data seen in Figure 4.2, the errors in the proposed model will be modeled by an ARMA time series.

The overall slope, β , in a model, can be used to quantify the overall warming trend of SST each year. Different nonlinear modeling approaches such as cyclic splines, monthly means, and sinusoidal regression can be used to estimate the nonlinear seasonal components $g(\cdot)$.

3.1 Statistical models

3.1.1 Dynamic seasonal mean model

First, the dynamic seasonal mean model (DSMM) will be considered for SST. Dynamic regression models are regression models in which the error terms are assumed to be dependent rather than i.i.d. white noise terms (R. Hyndman & Athanasopoulos, 2021). This approach is particularly useful in time series where it is known that previous values have an impact on further values, which is the case with SST. In other words, previous SST values are known to affect future SSTs in some way due to the thermal retention of seawater. For this analysis, the terms are modeled using an auto-regressive integrated moving average (ARIMA) process.

The dynamic seasonal mean model for the monthly SST data is

$$Y_i = \mu_{i^*} + \beta t_i + e_i, \tag{3.2}$$

where Y_i is the observed SST at i -th month, i is the number of months after the beginning of the time series, $\mu_i = \mu_{i^*}$, in which $i^* = (i - 1) \bmod (12)$ represents the overall mean temperature at the i^* -th month of the year, β is the yearly linear slope of SST, t is the number of years elapsed since t_1 ,

and e_i represents error terms modeled by an ARIMA process. In this model, $t_i = i$. Therefore, the parameters $\mu_0, \mu_1, \dots, \mu_{11}$ are the unknown true average temperature across time for each of the 12 months.

For computational implementation, a model in matrix form equivalent to Equation (3.2) is

$$Y = X\beta + e, \quad (3.3)$$

where $Y = (y_1, y_2, \dots, y_i)^T$, an $n \times 13$ matrix

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 12 \\ 1 & 0 & 0 & \cdots & 0 & 13 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & n \end{pmatrix},$$

in which $\beta = (\mu_1, \mu_2, \dots, \mu_{12}, \beta)^T$, and $e = (e_1, e_2, \dots, e_i)^T$, where N is the total number of months. If the starting month is not January, then the first few rows will be removed accordingly.

ARIMA models are adopted to model the errors e_i in Equation (3.2). ARIMA models combine auto-regressive (AR) and moving-average models (MA) to estimate time series data, and differencing (I) can be included to center the data (Box et al., 2015). The ARMA model is the basic model to analyze classical time series, popularized by Whittle (1953). Both AR and MA models use previous terms of a time series to predict future observations. In an AR model, a combination of previous terms' values estimates the following value. Conversely, in MA models, those previous terms' errors are used to estimate future terms.

In general, an ARMA model of errors e_i is notated by

$$e_i = \sum_{k=1}^p \phi_k e_{i-k} + \sum_{k=1}^q \theta_k + \epsilon_{i-k}, \quad (3.4)$$

where $\epsilon_{t-k}, \epsilon_{t-k+1}, \dots, \epsilon_i$ are i.i.d. white noise terms, ϕ are the parameters of the AR components, θ are the parameters of the MA components, and p and q refer to the orders of the model. p

indicates the AR order, and q indicates the MA order. For example, if the model for the error term is ARMA(1,1), the term e_i has the form of

$$e_i = \phi_1 e_{i-1} + \theta_1 + \epsilon_{i-1}.$$

3.1.2 Cyclic spline model

Second, the cyclic spline model will be considered. A spline is a component of a generalized additive model where an interpolated piecewise polynomial function is used to produce a smooth model (Hastie, 2017). The major advantage of the use of splines in model creation is their ability to provide highly accurate interpolations of many functions (Ahlberg et al., 1967). Additionally, splines produce continuous functions that can model discrete data, such as the monthly FGB data.

The most common form of spline models when modeling the seasonal component of SST in literature is B-spline. The defining characteristic of this spline is the use of "knots" for the function being approximated. These are based off of the Bézier curve (Racine, n.d.) and use basis functions.

The formula for a B-spline model of a variable t with τ knots is as follows:

$$g(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t - \tau_1)^3 + \beta_5 (t - \tau_2)^3 + \dots \quad (3.5)$$

The sequence of knots for K splines as defined in Perperoglou et al. (2019) is as follows:

$$\xi_1 \leq \dots \leq \xi_d \leq \xi_{d+1} < \xi_{d+2} < \dots < \xi_{d+K+1} < \xi_{d+K+2} < \xi_{d+K+3} < \dots < \xi_{2d+K+2}. \quad (3.6)$$

At these knots, the difference between the spline function and the observed function is 0. This allows for the approximation to be accurate and anchored to the observed data at multiple points within the function's domain, and the spline interpolates between them.

A special case of the spline is the cyclic spline. A cyclic spline model of a variable t_i , $i = 1, 2, \dots$ with a defined cycle of T present throughout t is defined as a spline in which

$$g(t_1) = g(t_{nT}), g(t_2) = g(t_{nT+1}), \dots, \quad n \in \mathbb{N}.$$

An advantage of cyclic B-splines is that they are able to be smoothly expanded across a domain, given that a subset of it repeats itself in a regular pattern throughout the domain. Because this

pattern is assumed as rigid and static, using a cyclic spline allows for the model's time domain to be extended freely using the subset as a model for the remainder of the data. In the context of time series data, the use of cyclic splines is an efficient way to create a model with fixed seasonal trends that span the entire time frame of observations.

A plot of monthly SST observations averaged throughout the years is pictured in Figure 3.2. The behavior of this pattern is smooth, and the average SST in January is very similar to the average SST in December. These two aspects of the data provide evidence that the seasonal effects of SST are justifications for the data being accurately modeled through a cyclic spline.

Monthly SSTs cycle regularly each year throughout time at a set pattern, so utilizing a cyclic spline produces an approximation of each month's expected effect on the yearly average SST independent of yearly warming trends. For example, the monthly effect of December 1996 is equal to that of December 2022.

For SST, a cyclic spline mean model is as follows:

$$Y_i = s(m_i) + \beta t_i + e_i, \tag{3.7}$$

where Y_i and t_i are equatable terms to Equation 3.2, s is a cyclic spline to model the seasonal effects of SST within each year, m_i represents the number of months elapsed of the i th observation within its respective year, and e_i are i.i.d. errors. For an observation that was taken on the first day of March, $m_i = 2$.

3.1.2.1 Dynamic sinusoidal regression model

Next, the dynamic sinusoidal regression model will be considered. This model is a simple approach that produces a smooth estimate for the seasonal component using sinusoidal functions. This can also be interpreted as a dynamic harmonic regression model with an order of 1. Because sinusoidal functions are periodic with a singular local minimum and maximum each period, this process reflects the yearly pattern of SST.

A general form of a sinusoidal mean model as described in R. Hyndman & Athanasopoulos

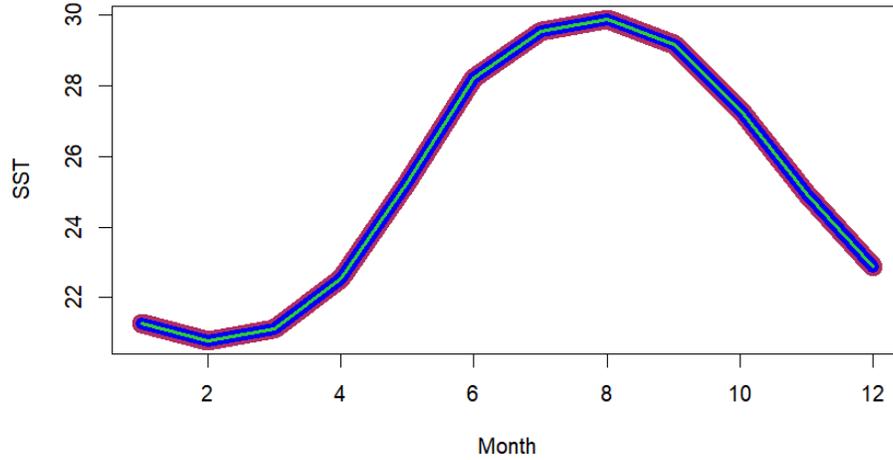


Figure 3.2
Monthly-aggregated SST observations in the three FGB locations. Maroon: EFGB; Blue: WFGB; Green: STET.

(2021) is

$$Y_i = A + B \sin\left(\frac{2\pi t_i}{m}\right) + C \cos\left(\frac{2\pi t_i}{m}\right) + e_i,$$

where A is an intercept, B and C are coefficients, m is the seasonal period, and e_i is a term assumed to represent i.i.d. error terms.

Because the errors of SST modeling are established not to be i.i.d., a dynamic model will instead be created. The seasonal period of monthly SST observations is 12 as there are 12 months in one year, therefore $m = 12$. A dynamic sinusoidal regression model of monthly SST data t is notated as follows:

$$Y_i = A + B \sin\left(\frac{2\pi t_i}{12}\right) + C \cos\left(\frac{2\pi t_i}{12}\right) + \beta t_i + e_i, \quad (3.8)$$

where A is the mean observed SST value, B and C are coefficients, and e_i are ARIMA-modeled errors.

For computational implementation, a model in matrix form for the dynamic sinusoidal regression model is

$$Y = X\beta + e, \quad (3.9)$$

where $\mathbf{Y} = (y_1, y_2, \dots, y_i)^T$, \mathbf{X} is a $n \times 4$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & \sin(2\pi \cdot \frac{1}{12}) & \cos(2\pi \cdot \frac{1}{12}) & 1 \\ 1 & \sin(2\pi \cdot \frac{2}{12}) & \cos(2\pi \cdot \frac{2}{12}) & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \sin(2\pi \cdot \frac{i}{12}) & \cos(2\pi \cdot \frac{i}{12}) & i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \sin(2\pi \cdot \frac{n}{12}) & \cos(2\pi \cdot \frac{n}{12}) & n \end{pmatrix},$$

$\boldsymbol{\beta} = (A, B, C, \beta)^T$, $\mathbf{e} = (e_1, e_2, \dots, e_i)^T$, and n is the total number of months. As with the dynamic seasonal mean model, if the starting month is not January, then the first few rows will be removed. The `auto.arima` function in the `forecast` R package (R. J. Hyndman & Khandakar, 2008) uses this matrix form.

3.1.2.2 Regression model with sandwich method

The final SST time series model considered is the regression model with sandwich method. The dynamic seasonal mean model in Equation 3.2 can be simplified by assuming that the errors are i.i.d. The calculations would be performed via ordinary-least-squares (OLS) methods, but there would be model assumption failures in the correlation structure of the error terms. The sandwich estimator is typically used to protect against these incorrect specifications of the correlation structure. An OLS linear regressive model's assumed errors are independent and homoskedastic, with stable variance throughout the domain. Because the residuals of a linear regression model of SST are known to be correlated, using the sandwich method is useful as this assumption is not necessary to produce a valid estimator.

The sandwich estimator, also known as the robust covariance matrix estimator, of a regression model is as follows:

$$\boldsymbol{\Psi} = \left(\frac{1}{n} \mathbf{X}^T \mathbf{X}\right)^{-1} \frac{1}{n} \boldsymbol{\Phi} \left(\frac{1}{n} \mathbf{X}^T \mathbf{X}\right)^{-1}, \quad (3.10)$$

where $\boldsymbol{\Phi}$ is the covariance matrix of score functions $\boldsymbol{\beta}$ and n is the sample size (White, 1980).

Due to the robust nature of the estimate, the resulting confidence intervals are usually wider than typical OLS estimators. In this analysis, HC4 estimates are used. These are White's

heteroskedasticity-consistent estimators with the formula of

$$HC4 = \text{diag}\left(\frac{\hat{u}_1^{\delta_1}}{1-h_1}, \dots, \frac{\hat{u}_n^{\delta_n}}{1-h_n}\right), \quad (3.11)$$

where $\delta_i = \min(4, \frac{nh_i}{h})$, $i = 1, \dots, n$, u represents a matrix of errors, and h represents the diagonal of the hat matrix $H = X(X^T X)^{-1} X^T$ (Cribari-Neto, 2004). This process is applied to each model using the sandwich R package Zeileis et al. (2020).

This estimate is less sensitive to high leverage points (observations with extreme x-values) than other estimates(Cribari-Neto & da Silva, 2011).

3.1.2.3 Thin spline spatial model

To include the spatial variation of SST seen in the SOCAT observations to the PLM, a proposed model is as follows:

$$Y_i = s(u_i, v_i) + g(t_i) + \beta t_i + e_i, \quad (3.12)$$

where (u_i, v_i) refer to the latitude and longitude rather than the month of the i -th observation, s is a spatial function that estimates the effect of SST for each location (u, v) , and the remaining terms correspond to those in Equation 3.2.

Thin plate splines are used to analyze spatial differences across a domain. As a generalized additive model, the smoothed parameters are spatial - in the case of SST, latitude and longitude. For additional information regarding methodology, refer to Wood (2003). The thin plate splines in this work are not temporal.

Producing splines in the same location in different years can illustrate the spatial variation of the data over time. This was performed in the western GoM due to frequent sampling as previously stated. The observations used for the splines are shown in Figure 3.3, and the resulting contour plots of those splines are shown in Figure 4.3.

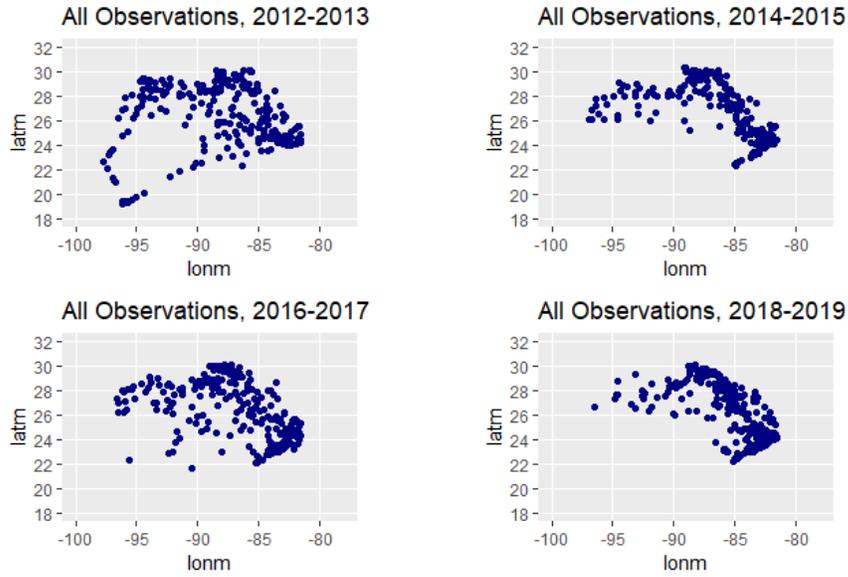


Figure 3.3
Maps of observational data used to create the thin plate spline model. Clockwise from top left: 2012-2013; 2014-2015; 2018-2019; 2016-2017.

4. RESULTS

4.1 Simulation

Statistical tests and simulations will be performed to determine the models' performance. Statistical simulation is a practical approach to modeling otherwise difficult-to-handle data (Boulesteix et al., 2020). Additionally, statistical simulation is used for comparative studies, as more variation present in noisy observed data can be held constant. This allows for more of a focus on model performance.

For data with unknown processes, using simulation data can help control for certain signals that are of interest for modeling. Other factors, such as data that do not adhere to certain modeling assumptions, can lead to statistical simulation being beneficial. Using simulation to test model accuracy is advantageous due to the ability to control aspects of the data that are not possible when using collected data.

Because the true behavior of SST is unknown and likely fluctuates over time, using simulated data allows for certain aspects of the observed data to be controlled. In the case of this project, the true observed β in the EFGB data set cannot be determined, so it is important to understand how the different models perform with different β values. As seen later, this approach is useful due to different model types performing better than others depending on the value of β .

The simulation process for this project entailed creating many simulated time series in which its key features such as β were controlled, and then different models were produced numerous times to try to model the simulated data. A simulated time series was created for each iteration, and the models in each simulation were fitted to that singular respective time series before re-simulating the time series in the next iteration.

The basis of the simulated time series is the monthly EFGB data. Before the simulation can begin, certain features of the basis data need to be selected as these alter the properties of the simulated time series. These included β , or the overall linear trend, as well as monthly averages (a part of g). An error term similar to the original data's natural variance was then added in order to preserve some randomness for the models to try to capture.

The specified β values for each simulation range from 0.00 (no linear overall trend) to 0.2°C each year. According to Wang et al. (2023), the Gulf of Mexico has been warming at a yearly rate of approximately 0.019°C between 1970 and 2020, with an error of 0.005°C.

Each month’s average SSTs were calculated through a linear model, passed in R by

```
lm(sst_{sim} ~ fmonth + t).
```

The resulting coefficients of monthly averages listed in Table 4.1. In the table, it is important to note that the intercept in the model is 1, or January. The other monthly values are relative to January’s estimated temperature. The resulting estimated yearly warming trend is 0.028°C.

1	2	3	4	5	6
20.86	-0.52	-0.19	1.24	3.94	6.92
7	8	9	10	11	12
8.24	8.60	7.86	6.01	3.62	1.56

Table 4.1

Estimates of monthly coefficients from the linear model of EFGB observed SST data. The estimated slope is 0.028, and the intercept of the model refers to January.

To produce the error model, an ARIMA model was created for the EFGB data with monthly averages and linear trends removed to stabilize the time series. The simulated series resulted from the `forecast::arima.sim` R function in which the orders of the simulated ARIMA time series as well as its standard deviation were equal to the EFGB ARIMA model’s orders. The length of the simulated time series can vary, but $n = 500$ simulated monthly observations were selected.

The simulation was iterated 1,000 times for each specified β value and seed. For each of the 1,000 iterations of the simulation process, a simulated time series based on the EFGB monthly data’s ARIMA and error behaviors was created and a specified β slope term was added. From this, a model from each type was produced (dynamic seasonal mean, cyclic spline, dynamic sinusoidal regression, and regression with sandwich method) for the simulated data. The fits for the models were then extracted, including $\hat{\beta}$ and seasonal estimates. Deviation from the known “true” values of the simulated time series was then calculated and stored in a vector for each simulation iteration. Finally, a 95% confidence interval (CI) for each simulated model’s $\hat{\beta}$ was calculated - if the CI captured the known β value, a count would be added to be totaled after the simulation process.

When the simulation process is completed, performance metrics are calculated for the models. These give insight into how well each model can estimate β . The performance metrics used for each model include mean absolute error (MAE) and variance of $\hat{\beta}$, as well as the proportion of simulations in which the specified β value was captured by the models' resulting 95% CIs.

Simulation results are given in Table 4.2. Both spline and seasonal methods performed similarly in estimating the specified β when assessed by error and variance. However, spline models more consistently captured the specified β value throughout varying specified β s.

The ARIMA error models fit to the simulated data most commonly resulted in an order of (1,0,0), or an AR(1) model. Although each automatically-fitted model had the same order, any order ARIMA model is possible. Orders were selected by the least conditional sum of squares and maximum likelihood (R. J. Hyndman & Khandakar, 2008). A regression model with an AR(1) error possesses the formula

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \tag{4.13}$$

where β_0 represents the intercept, β_1 represents the linear trend over time, and e_i is an AR(1) model behaving as follows:

$$e_i = \phi_1 e_{i-1} + \epsilon_i. \tag{4.14}$$

ϵ represents a series of residuals not accounted for in the regression or AR(1) model of errors.

For the mean square error (MSE) of seasonal components, some lower β values resulted in NA values for the ARIMA models. Because ARIMA models are more complex than the other modeling techniques used, it is possible that this is the reason that mean components could not be accurately calculated.

Overall, dynamic modeling resulted in the greatest CI capture proportion in each specified β value. Both the dynamic seasonal mean model and sinusoidal models performed the most consistently, being the only simulated models in which over 90% of the simulated CIs contained the specified β value. Other models captured β less than 80% of the time. The worst-performing

model was the linear method with sandwich estimator - this is a less complex method, so variance not captured in the model likely resulted in a lower accuracy despite a low variance of $\hat{\beta}$ values.

The warming trend value in Wang et al. (2023), 0.019, was simulated, and the performance of the models is listed in Table 4.2. If taken as the true β value for observed SST in the GoM, spline and seasonal methods would produce the least errors and variance, but ARIMA and sinusoidal models most consistently capture β in a 95% CI.

Due to the strong performance of the sinusoidal models, it can be concluded that seasonal trends are constant throughout time. This is a differing conclusion from what is being observed in the Bohai Sea as reported by Yuan et al. (2023).

4.1.1 Data analysis

The models used in the simulation analysis were performed on observed data as well. Both the EFGB data (the basis for the simulated time series), as well as SOCAT data, give insight into how the models perform with simulated compared to observed data.

Because the SOCAT data are not regularly spaced, time series modeling including the ARIMA modeling process were not able to be applied to the data set. Similarly, the FGB data are not applicable to a thin plate spline model as the data are not spatial. The models' estimates and 95% CIs for β of the EFGB data set are listed together in Table 4.3 for convenience.

First, the monthly means of the EFGB data were produced by creating a linear model of SST predicted by the month of the year and the numeric date. This date variable was produced by combining an integer of the number of years since 1989 with a decimal corresponding to the proportion of the year that has passed as of the date and time that the observation was recorded. The process was completed by using the `lubridate` R package (Grolemund & Wickham, 2011).

The resulting monthly means are listed in Table 4.1 and were used for time series simulation. This series of means is relative to the mean value of January for a given year. The model estimates the average difference of monthly means in a given year, excluding yearly mean changes due to β . The average $\hat{\beta}$ from each model is 0.0283, and the 95% CI is given in Table 4.3.

Second, the dynamic seasonal mean model was fitted to the EFGB data. Because the SOCAT data are not uniformly spaced, the ARIMA modeling process does not apply as it relies on lags. These lags are defined by previous observations - since the time difference between observations is not constant, this type of model would not give useful insight with respect to modeling a time series.

First, dummy variables for each observation's month as explained in Equation 3.3 were produced. Then, using the `forecast::auto.arima` R function, an ARIMA model was fit to the data frame, and a matrix of the monthly dummy variables were passed as external regressors. The function iteratively fits differently-ordered ARIMA models to the data, selecting the model with an optimized AIC value.

The model coefficients are listed in Table 4.4, and the estimated $\hat{\beta}$ and the resulting CI are listed in Table 4.3. The intercept of this model is the monthly mean of December, or `fmonth_12`. Note that this is different from the previous model. The other `fmonth` values' estimates are relative to the December estimated SST. For example, the estimated SST of `fmonth_2`, or February, is 2.1069°C less than December's estimate: this would result in an SST of 20.3597°C. The coefficient of the `ar1` term is the coefficient of the autoregressive component of the ARIMA time series. Finally, the $\hat{\beta}$ value is stored in the `t` variable, and the 95% CI of β is listed in Table 4.3. This model's CIs are the least conservative, with the most narrow range of all dynamic model types. This could be what resulted in a lower CI capture rate than the other methods.

Next, the cyclic spline model was used. This model was applied to both the EFGB and SOCAT data in slightly different ways. Because the EFGB data are collected monthly, producing a model based on months results in the data being discrete. However, SOCAT data are collected irregularly and on a finer scale than monthly, so these observations are widely dispersed within the domain in a continuous manner. This is interesting because the SOCAT fitted model's marginals can be calculated throughout each month, rather than only in the middle of the month. In the EFGB data, there are exactly 12 observations each year. In the SOCAT, there are considerably more.

The model was fitted to the EFGB data by creating a spline based on the month of the year and the numeric date using a similar process to the dynamic seasonal mean model. The model was fitted to the SOCAT data by creating a new column, the numeric month, akin to the numeric date. This was created by combining an integer of a month with a decimal representing the day of observation out of the total number of days in the given month. The fitted models are displayed in Figure 4.1.

An autocorrelation function (ACF) plot of the residuals residuals of the cyclic spline performed on the EFGB data is shown in Figure 4.2. The significantly large, moderately strong autocorrelation seen at lag 1 indicates that these residuals are not independent and a given month's SST is correlated with the month prior. This is justification for producing a dynamic model, as an ARIMA model for errors would be able to capture this aspect of the data.

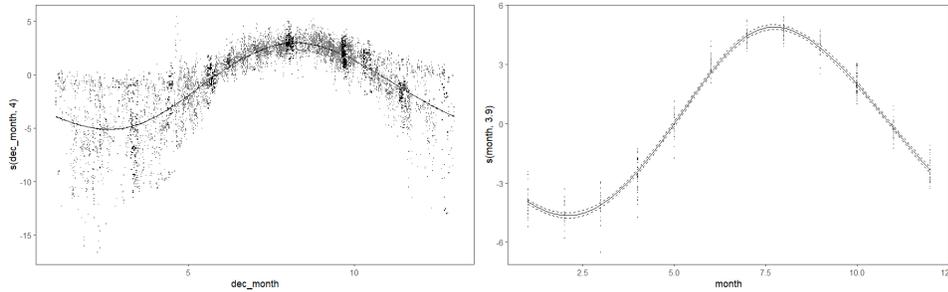


Figure 4.1
Fitted cyclic spline models to observed data. Left: SOCAT; right: EFGB.

The dynamic sinusoidal regression model was then applied to the EFGB data. Two dummy variables were created as external regressors, equal to $\sin(\text{num.date})$ and $\cos(\text{num.date})$, where num.date corresponds to the numeric date as described in the dynamic seasonal mean model. Similarly, `forecast::auto.arima` is used with the external regressors to select the best-performing model.

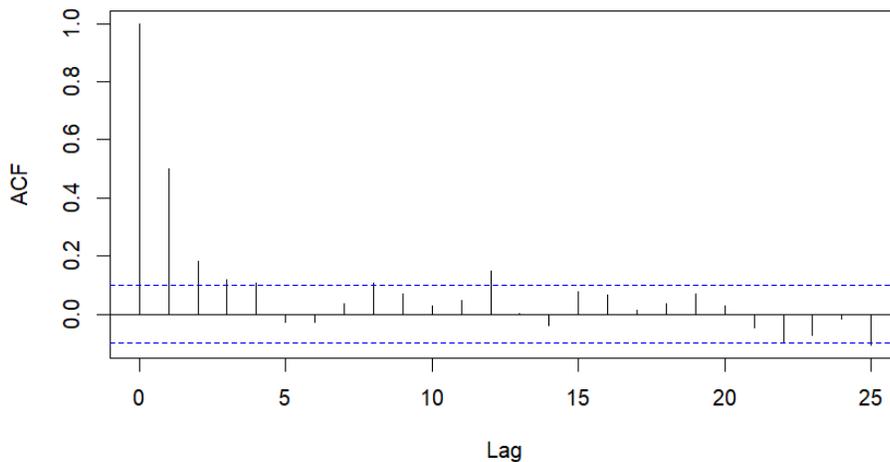


Figure 4.2
ACF plot of residuals from cyclic spline model of centered observed data.

The model's estimates are shown in Table 4.5, and the resulting 95% CI for β is included in Table 4.3. This model's estimate and CI are comparable to that of the monthly means model and its sandwich estimator. The external regressors do not include easily interpretable insights when compared to the monthly external regressors, but these results provide a simpler model to produce as seen in Equation 3.8.

The final temporal model produced was the regression model with sandwich estimator. This model was applied to the EFGB data, as it relates to a previously-generated ARIMA model. The sandwich estimator for $\hat{\beta}$ was created from the monthly means estimate. The HC4 estimator was calculated to provide an updated variance of the $\hat{\beta}$ value estimated in the monthly means model, and a new 95% CI was produced from the sandwich variance.

The new CI range from the sandwich process is 0.0133, which is less than the monthly means CI range of 0.0240. Using the sandwich method allows, in this case, for a narrower interval that is still as confident as the previous CI, as seen in Table 4.3.

For additional analysis, thin plate spline models of the SOCAT data were created. From Figure 4.3, it can be concluded that spatial effects vary throughout time. This is a phenomenon that has been explored, with others attributing this to the prevalence of eddies and currents (Foster et al., 2014), (Wang et al., 2023).

$\beta = 0.00$, seed = 12345	MAE	Variance	CI Capture Proportion
Spline	9.451e-05	4.029e-06	0.791
DSMM	0.003	0.005	0.944
Sinusoidal	0.068	2.477	0.924
Seasonal	5.021e-05	4.029e-06	0.734
Seasonal t Sandwich			0.734
$\beta = 0.01$, seed = 1234	MAE	Variance	CI Capture Proportion
Spline	1.9988e-05	4.357e-06	0.774
DSMM	6.9557e-05	4.344e-06	0.944
Sinusoidal	0.0002	4.351e-06	0.923
Seasonal	6.4758e-05	4.358e-06	0.703
Seasonal t Sandwich			0.706
$\beta = 0.019$, seed = 2345	MAE	Variance	CI Capture Proportion
Spline	5.190e-05	4.725e-06	0.744
DSMM	0.002	0.003	0.927
Sinusoidal	0.030	0.883	0.908
Seasonal	9.643e-05	4.725e-06	0.673
Seasonal t Sandwich			0.678
$\beta = 0.05$, seed = 23456	MAE	Variance	CI Capture Proportion
Spline	8.591e-05	4.201e-06	0.785
DSMM	0.002	0.003	0.944
Sinusoidal	5.5669e-05	4.207e-06	0.915
Seasonal	4.140e-05	4.201e-06	0.721
Seasonal t Sandwich			0.734
$\beta = 0.15$, seed = 45678	MAE	Variance	CI Capture Proportion
Spline	6.848e-05	4.468e-06	0.766
DSMM	1.992e-05	4.433e-06	0.933
Sinusoidal	8.085e-05	4.457e-06	0.919
Seasonal	2.388e-05	4.471e-06	0.699
Seasonal t Sandwich			0.697
$\beta = 0.20$, seed = 456789	MAE	Variance	CI Capture Proportion
Spline	0.0001	4.547e-06	0.762
DSMM	9.911e-05	4.556e-06	0.925
Sinusoidal	4.529e-06	4.564e-06	0.907
Seasonal	0.0001	4.548e-06	0.685
Seasonal t Sandwich			0.682

Table 4.2

Simulation results: MAE, variance, and the CI capture proportion of various models. Model names: Spline (cyclic spline), DSMM (dynamic seasonal mean), Sinusoidal (dynamic sinusoidal regression), Seasonal (seasonal mean regression), Seasonal t Sandwich (seasonal mean regression with sandwich estimator).

Method	$\hat{\beta}$	95% CI
Monthly means	0.0283	(0.0151, 0.0391)
Sandwich	0.0283	(0.0216, 0.0349)
DSMM	0.0271	(0.0114, 0.0429)
Sinusoidal	0.0278	(0.0175, 0.0381)
Spline	0.0282	(0.0194, 0.0370)

Table 4.3

Estimated linear yearly trend and associated 95% CIs of EFGB observed data.

ar1	12	t	1	2	3	4
0.5633	22.4666	0.0271	-1.5910	-2.1069	-1.7739	-0.3473
5	6	7	8	9	10	11
2.3525	5.3364	6.6499	7.0176	6.2771	4.4279	2.0490

Table 4.4

Estimates of coefficients from the dynamic monthly means model of EFGB observed data. The errors are modeled by a ARIMA(1,0,0) model. Please note the intercept of the model refers to December.

ar1	ar2	intercept	t	sin.x	cos.x
0.6120	-0.1718	24.8475	0.0278	-3.1143	-3.5257

Table 4.5

[Estimates of coefficients from the dynamic sinusoidal regression model of EFGB observed data. The errors are modeled by an ARIMA(2,0,0) model.

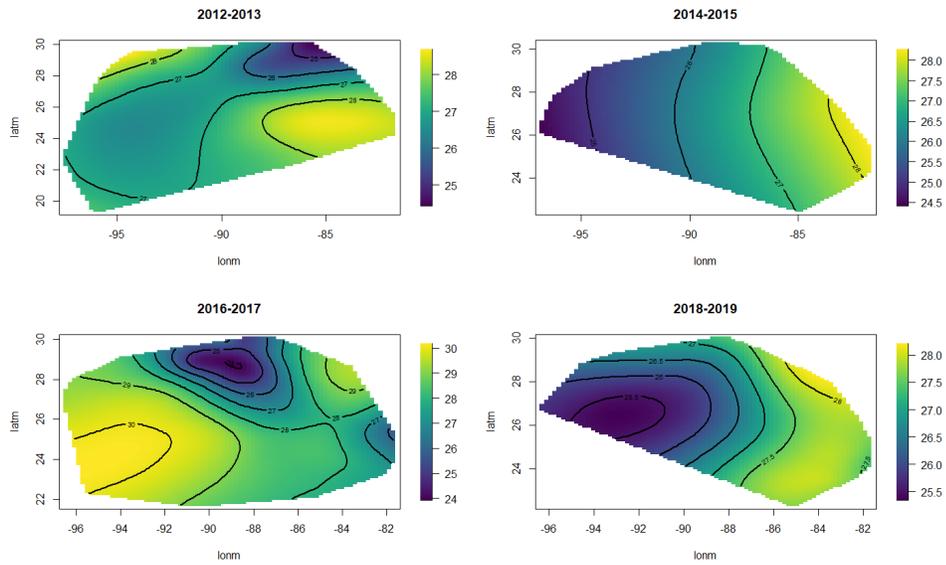


Figure 4.3
Contour plots of thin plate spline fit for SST in the Western GoM. Clockwise from top-left: 2012-2013; 2014-2015; 2018-2019; 2016-2017.

5. CONCLUSION

5.1 Summary

This work sought to provide an analysis of observed sea surface temperature (SST) data in the Gulf of Mexico (GoM). Different model types were statistically simulated based on time series data in the Flower Garden Banks National Marine Sanctuary (FGB). From these simulations and supporting models of observed data, a recommendation of modeling strategies for estimating SST in the region could be made. Understanding monthly seasonal effects and providing the slope for an overall linear warming trend, β , was the goal of the modeling process.

Regarding estimating observed data, a study conducted by Wang et al. (2023) concluded that the GoM is warming by about 0.019°C each year on average. At the East Flower Garden Bank (EFGB), a site known for its temperature-sensitive coral population, it was estimated that this trend is comparable. Model outputs estimated the EFGB β to be between 0.0114 and 0.0429°C at 95% confidence, with an average $\hat{\beta}$ of 0.02785°C . Additionally, thin plate splines were explored to verify that SST variation is also spatial in nature, with SST distributions altering over time spatially.

For simulated data, dynamic models performed much better than non-dynamic models. In particular, the dynamic seasonal mean model outlined in Equation 3.2 outperformed across all specified β values for simulated data sets in terms of consistently capturing the correct value inside a 95% confidence interval (CI). However, the resulting CIs were noticeably wider than other models, such as the slightly under-performing dynamic sinusoidal regression model. Due to the continuous and smooth nature of the spline model, using dynamic sinusoidal regression for modeling SST over time is recommended.

5.2 Discussion

The recommendation of the dynamic sinusoidal regression model stems from the balance of accuracy and specificity. Because the dynamic seasonal mean models treat monthly SSTs as static, this method would lose accuracy as non-discrete data are fit to the model. For this reason, the use of the continuous sinusoidal model would be advantageous in many instances, specifically for data that are collected more frequently than one time each month.

Modeling a complex environmental process such as SST can have many approaches, and it is important to note that covariates such as wind, CO₂, or other variables were not utilized in the modeling process for this work. Other aspects of the data such as spatial variability were also not included in the central FGB data and should be explored in future research. To extend this work, the simulation of the proposed spatial model in Equation 3.12 using SOCAT observations can be performed to estimate the spatio-temporal behavior of SST in the GoM.

REFERENCES

- Ahlberg, J., Nilson, E., & Walsh, J. (1967). The theory of splines and their applications. In *Mathematics in science and engineering* (Vol. 38, pp. vv–viii, 1–284). Elsevier. doi: doi:10.1016/S0076-5392(08)61984-0
- Bakker, D. C. E., Alin, S. R., Bates, N., Becker, M., Feely, R. A., Gkritzalis, T., . . . Wimart-Rousseau, C. (2023). *Surface ocean co2 atlas database version 2023 (socatv2023) (ncei accession 0278913)* (Tech. Rep.). NOAA National Centers for Environmental Information. doi: doi:10.25921/r7xa-bt92
- Bashevkin, S. M., Mahardja, B., & Brown, L. R. (2022). Warming in the upper san francisco estuary: Patterns of water temperature change from five decades of data. *Limnology and Oceanography*, 67(5), 1065–1080. doi: doi:10.1002/lno.12057
- Boulesteix, A.-L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., . . . Sauerbrei, W. (2020). Introduction to statistical simulations in health research. *BMJ open*, 10(12), e039921. doi: doi:10.1136/bmjopen-2020-039921
- Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics Data Analysis*, 45(2), 215-233. doi: doi:10.1016/S0167-9473(02)00366-3
- Cribari-Neto, F., & da Silva, W. B. (2011). A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *AStA Advances in Statistical Analysis*. doi: doi:10.1007/s10182-010-0141-2
- Dunstan, P. K., Foster, S. D., King, E., Risbey, J., O’Kane, T. J., Monselesan, D., . . . Thompson, P. A. (2018). Global patterns of change and variation in sea surface temperature and chlorophyll a. *Scientific reports*, 8(1), 14624. doi: doi:10.1038/s41598-018-33057-y
- Foster, S. D., Griffin, D. A., & Dunstan, P. K. (2014). Twenty years of high-resolution sea surface temperature imagery around australia: Inter-annual and annual variability. *PLoS one*, 9(7), e100762. doi: doi:10.1371/journal.pone.0100762

- Gil-Alana, L. A., Monge, M., & Rojo, M. F. R. (2019). Sea surface temperatures: Seasonal persistence and trends. *Journal of Atmospheric and oceanic technology*, 36(12), 2257–2266. doi: doi:10.1175/JTECH-D-19-0090.1
- Glenn, E., Comarazamy, D., González, J. E., & Smith, T. (2015). Detection of recent regional sea surface temperature warming in the caribbean and surrounding region. *Geophysical Research Letters*, 42(16), 6785–6792. doi: doi:10.1002/2015GL065002
- Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3), 1–25. doi: doi:10.18637/jss.v040.i03
- Hardle, W., Liang, H., & Gao, J. (2000). *Partially linear models*. Physica. doi: doi:10.1007/978-3-642-57700-0
- Hastie, T. J. (2017). *Generalized additive models*. Routledge. doi: doi:10.1201/9780203753781
- Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd ed.). OTexts. Retrieved from [OTexts.com/fpp3](https://otexts.com/fpp3)
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3), 1–22. doi: doi:10.18637/jss.v027.i03
- Johnston, M. A., O’Connell, K., Blakeway, R. D., Hannum, R., Nuttall, M. F., Hickerson, E. L., & Schmahl, G. P. (2022). *Long-term monitoring at east and west flower garden banks: 2020 and 2021 annual report* (Tech. Rep.). U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Flower Garden Banks National Marine Sanctuary. Retrieved from <https://sanctuaries.noaa.gov/science/conservation/2020-2021-long-term-monitoring-east-and-west-flower-garden-banks-annual-report.html>
- Kealoha, A. K., Shamberger, K. E., DiMarco, S. F., Thyng, K. M., Hetland, R. D., Manzello, D. P., ... Enochs, I. C. (2020). Surface water co2 variability in the gulf of mexico (1996–2017). *Scientific Reports*, 10(1), 1–13. doi: doi:10.1038/s41598-020-68924-0
- Liu, J., Jin, B., Yang, J., & Xu, L. (2021). Sea surface temperature prediction using a cubic b-spline interpolation and spatiotemporal attention mechanism. *Remote Sensing Letters*, 12(5), 478–487. doi: doi:10.1080/2150704X.2021.1897182

- O'carroll, A. G., Armstrong, E. M., Beggs, H. M., Bouali, M., Casey, K. S., Corlett, G. K., . . . others (2019). Observational needs of sea surface temperature. *Frontiers in Marine Science*, 6, 420. doi: doi:10.3389/fmars.2019.00420
- Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in r. *BMC medical research methodology*, 19(1), 1–16. doi: doi:10.1186/s12874-019-0666-3
- Racine, J. S. (n.d.). A primer on regression splines [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf
- Wang, Z., Boyer, T., Reagan, J., & Hogan, P. (2023). Upper-oceanic warming in the gulf of mexico between 1950 and 2020. *Journal of Climate*, 36(8), 2721 - 2734. doi: doi:10.1175/JCLI-D-22-0409.1
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. doi: doi:10.2307/1912934
- Whittle, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(1), 125–139. doi: doi:10.1111/j.2517-6161.1953.tb00131.x
- Wood, S. N. (2003, 01). Thin Plate Regression Splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1), 95-114. doi: doi:10.1111/1467-9868.00374
- Yuan, C., Kuang, X., Xu, J., Li, R., & Wang, C. (2023). Phase variations of the summer and winter seasons in the bohai sea during the last four decades. *Frontiers in Marine Science*, 10, 427. doi: doi:10.3389/fmars.2023.1095792
- Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95(1), 1–36. doi: doi:10.18637/jss.v095.i01