ANALYSIS OF MISSING DATA IN MARINE DISSOLVED OXYGEN TIME SERIES USING DYNAMIC LINEAR MODELS

Shajratul Alam
University of Rhode Island, shajratu_alam@uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/theses

Recommended Citation
https://digitalcommons.uri.edu/theses/1903

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.
ANALYSIS OF MISSING DATA IN MARINE DISSOLVED OXYGEN TIME SERIES USING DYNAMIC LINEAR MODELS

BY

SHAJRATUL ALAM

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

UNIVERSITY OF RHODE ISLAND

2020
MASTER OF SCIENCE THESIS

OF

SHAJRATUL ALAM

APPROVED:

Thesis Committee:

Major Professor     Jing Wu
Gavino Puggioni
Yichi Zhang
Candace Oviatt
Brenton DeBeof
DEAN OF THE GRADUATE SCHOOL
The marine ecosystems cannot survive without dissolved oxygen (DO). Low oxygen events (hypoxia) in the ocean cause stress on the benthic community and, hampers their growth rate initiating mortality. To monitor the concentration of oxygen, different water quality monitoring sites have been established across the globe. The Narragansett Bay fixed-site water quality monitoring network (NBFWQMN) is a facility that regularly measures oxygen level as well as other important water parameters (temperature, salinity, pH level, and Chlorophyll) at different locations of Narragansett Bay (NB). Missing observation is a common phenomenon for this times-series dataset and, can occur for various reasons. In this study, we analyzed time-series data of dissolved oxygen (DO) after taking into account the missing data. Variability of DO across any water-body depends on diffusion from the atmosphere, respiration of organic matter in the water column and in the sediment and advection of saltwater. The oxygen concentration in water also depends on instantaneous temperature, salinity, and freshwater inputs from nearby rivers. In this study, we used time-series data of temperature, salinity, and river discharge as covariates for DO time-series. In addition to the response variable, some of the covariates also have missing data. In this thesis, we applied dynamic linear model to handle the time-series data with ignorable missing response and covariates.
ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my major advisor Dr. Jing Wu and co-major advisor Dr. Gavino Puggioni for their relentless guidance and encouragement throughout the project. They have also been very supportive of my future career path. I cannot thank them enough for their words of encouragement and continuous patience. I am grateful to my outside committee member, Dr. Candace Oviatt, for her advise and suggestion. I also want to extend my gratitude to Heather Stoffel for tending to my curiosity and help me understand the dataset. I would also like to thank Dr. Yichi Zhang for agreeing to be my internal committee member and sharing his thought. I deeply appreciate Dr. Tatiana Rynearson for chairing my defense and providing me with insightful comments. My heartfelt thanks to Dr. Liliana Gonzalez for showing great confidence in me.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>List of Reference</td>
<td>5</td>
</tr>
<tr>
<td>2 Methodology</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Data Source and Study Area</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Dynamic Linear Models</td>
<td>12</td>
</tr>
<tr>
<td>2.2.1 Basic Theory for The Inference of DLM</td>
<td>14</td>
</tr>
<tr>
<td>2.2.2 Bayesian Computation for DLM</td>
<td>16</td>
</tr>
<tr>
<td>2.2.3 Data Analysis Plan</td>
<td>18</td>
</tr>
<tr>
<td>List of Reference</td>
<td>19</td>
</tr>
<tr>
<td>3 Results</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Preliminary Data Analysis</td>
<td>22</td>
</tr>
<tr>
<td>3.1.1 Exploratory Data Analysis of Temperature</td>
<td>22</td>
</tr>
<tr>
<td>3.1.2 Exploratory Data Analysis of Salinity</td>
<td>24</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Narragansett Bay fixed site water quality monitoring network locations (image source: <a href="http://www.dem.ri.gov/programs/emergencyresponse/bart/stations.php#map">http://www.dem.ri.gov/programs/emergencyresponse/bart/stations.php#map</a>)</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Distribution of missing values in daily measurement of temperature, salinity and DO% of GB (left panels) and TW (right panels).</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Conceptual framework of data analysis plan</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>ACF and PACF plot of $\Delta_{365} Temperature_t$ with blue lines representing 95% confidence intervals</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>Time series plot of river discharge data with salinity for both sites</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>Cross-correlation analysis between river discharge data and Salinity</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>ACF and PACF plot of Salinity</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>Cross-correlation analysis between temperature and DO%</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>Cross-correlation analysis between salinity and DO%</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>Decomposition of the temperature DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI)</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>Posterior distribution of AR(1) coefficient, $\phi$, variance, $v$ and discount factor $\delta$ for temperature DLM</td>
<td>31</td>
</tr>
<tr>
<td>12</td>
<td>Decomposition of the salinity DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI)</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>Dynamic regression coefficients for salinity DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI)</td>
<td>33</td>
</tr>
<tr>
<td>14</td>
<td>Posterior distribution of AR(1) coefficient, $\phi$, variance, $v$ and discount factor $\delta$ for salinity DLM</td>
<td>34</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>15</td>
<td>Decomposition of the DO% DLMs of the two sites (red line represents arithmetic</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>mean of the samples and blue lines represent 95% CI)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Dynamic regression coefficients for DO% DLMs of the two sites (red line</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>represents arithmetic mean of the samples and blue lines represent 95% CI)</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Posterior distribution of AR(1) coefficient,$(\phi)$, variance $(\nu)$ and</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>discount factor $\delta$ for DO% DLM</td>
<td></td>
</tr>
<tr>
<td>B.1</td>
<td>Time series plot of precipitation data with salinity for both sites</td>
<td>42</td>
</tr>
<tr>
<td>B.2</td>
<td>Cross-correlation analysis between river discharge data and Salinity</td>
<td>42</td>
</tr>
<tr>
<td>B.3</td>
<td>Cross-correlation analysis of river discharge data and DO% with blue lines</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>representing 95% confidence intervals</td>
<td></td>
</tr>
<tr>
<td>B.4</td>
<td>ACF and PACF plot of $\Delta_{365} DO_t$ with blue lines representing 95%</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>confidence intervals</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

Non-compound dissolved oxygen (DO) in water is essential for the existence of the benthic community. Warm water fish, for example, require DO concentrations of at least 5 mg/L for optimum health condition [1]. Hypoxia or low level of oxygen in water occurs in the event of oxygen concentration dropping to 2-3 mg/l [2]. Although some aquatic species can survive a brief period of hypoxia, persistent duration of this condition significantly decreases their growth rate [3, 4, 5]. Mortality is observed in many sensitive species within 4-7 hours if oxygen level remains below 2 mg/l[6]. From 1984 to 2002, 64.4% of the fish in South Florida died resulting from DO concentration dropping below 3 mg/l [7]. In the Chesapeake Bay, many blue crabs died due to low oxygen level [8]. Fish kills from hypoxia have occurred periodically in Narragansett Bay (NB) due to hypoxic events [9].

Sources of oxygen in water include photosynthesis of aquatic plants, atmosphere and mixing with more-oxygenated waters through circulation. Other than these inputs, there are multiple factors that are responsible for spatial and temporal variability of oxygen level across the water body [10]. As temperature increases, water can hold less oxygen, that is oxygen level in water and temperature are inversely related. The solubility of oxygen also decreases exponentially with increased level of salinity. Further, stratification of water column causes poor mixing of oxygen level at different depths of water. Stratification occurs when less dense freshwater from rivers surges into more concentrated ocean water - this creates a difference in density of water column that prevents vertical mixing of surface oxygen to the lower level. Generally, stratification remains during spring and summer weather but is disrupted by storms during fall and winter storms [11]. These factors alone, however, cannot trigger the
depletion of oxygen to hypoxic level. Occasional availability of overabundance nutrients from river discharge plays an important role in this whole process. With ample nutrients to feed upon, algae bloom occurs and death follows when nutrients are no longer available. This mass of dead algae decompose at the bottom of water using up oxygen and replenishment is very slow due to the stratification of water columns. Thus follow the onset of hypoxia at bottom level of the water body [12, 13].

Hypoxia is a hazardous phenomenon for any aquatic ecosystems. Aiming to study and monitor occurrences of low oxygen level, water quality monitoring systems have been established across different coastal systems around the world [14]. By recording continuous measurement of DO at regular intervals and analysing them, researchers want to understand the dynamics of oxygen level across the water body over time and predict possible onset of future hypoxic events. Existence of missing values in this kind of dataset is more or less inevitable. Sometimes instruments fail due to trapped sludge/weed or extreme weather conditions. This can halt continuous measurement over days or weeks. Sometimes calibration of the tools are not done properly - resulting in corrupted data that are removed. Analysis of such incomplete data, that have natural temporal ordering (i.e. time series data), is not that straightforward. The unique characteristics of time series data, that is observations close together in time will be correlated, makes the handling of missing data quite challenging. We cannot simply discard the missing part of the series during analysis, because we can infer the distribution of these missing values using the structure of the time series data and its variability. Ignoring missing information may lead to loss of information, biased estimates, and, eventually, poor prediction.

Types of missing data can be categorized into three according to their generation process [15]. Missing completely at random (MCAR) refers to the scenario that observations being missing independent of any other values in the dataset. An example
for MAR would be loss of data due to machine malfunction if it was hit by a boat. If occurrences of missing data depend on other observed values in the dataset, then it is classified as missing at random (MAR). The tools that measure water quality start to have a steady decline in their performance after sometime and occurrence of missing data is most probable during that time. This is an example of MAR. The final category is missing not at random (MNAR), in which case, the probability of the missing values depends on the missing values itself. For example, if measurements could not be collected because they are beyond the instrumental detection limit, then we can encounter MNAR. Another important property concept for missing data analysis is *ignorability*. It refers to the fact that distribution of the missing data mechanism is independent of the unobserved part of the data. When missing data mechanism is *ignorable*, we do not require specific assumption while modeling. To handle MNAR and some instances of MAR, we need to account for how the observation went missing to calibrate the model accordingly and, hence, is "non-ignorable"[16].

The methods for handling missing data can be grouped into four general categories - (1) complete case analysis, (2) weighting of the observed value, (3) imputation-based methods, and (3) likelihood-based method. However, this grouping is not mutually exclusive. The methods belonging to the first two groups are not applicable for the analysis of missing data in time-series. Among the imputation methods, mean/median/mode imputation, linear interpolation, arithmetic smoothing, last observation carried forward and next observation carried backward are the simplest ones and included in popular statistical software packages [17, 18, 19]. However, the resulting estimates are often biased [20, 21] and cannot estimate the uncertainty due to missing values. Forecasting methods, that incorporates time series components such as auto-regressive, moving average process and seasonality, are often used to provide reliable estimates of sequential missing values of a short interval or random instances
of missing values over the entire series - subject to correct specification of data generation process [22, 23, 19]. When time series data on multiple variables are available, the inter-dependencies between these values can be utilized to for the imputation of missing values via multiple linear regression [21]. This method assumes that the values at one time point are independent of the values at previous time points, which is not true for time series data. It fails completely when all variables are missing for a specific time interval. Likelihood based iterative method, expectation-maximization (EM) algorithm, can address this problem and provide better estimates of missing values for multivariate time series [24]. However, this algorithm is not suitable for variables with non-stationarity of variance.

A class of models that can incorporate versatile structure of time series process, also allowing for the inherent treatment of missing observations is dynamic models [25, 26, 27, 28]. Under this approach, each observations are treated as signals with additive noise emitting from unobserved and varying states. Through Kalman filter [29], we try to estimate these states by updating the current knowledge with each new observation. Simultaneously, we assign a value of uncertainty for our updated information. In case of missing value, we keep our previous estimation of the state but increase our level of uncertainty according to the variability the state values. Thus, even if we are missing all values for a portion of time, we can use our current knowledge to fill in the gap.

Use of dynamic models for handling missing data in time series is not new [26, 30, 31]. This methodology can incorporate complex model structures despite being conceptually simple. However, it is computationally intensive and often requires finding inverses of large matrices for each data point. This hindered it’s popularity in the 90’s [32]. With advancement in computers, state-space modeling is gaining more acceptance across different field of research. Successful use of this methodology in
recent years include, but are not limited to, handling loss of data in sensor network
[33, 34], analysis of incomplete panel data (observations collected from multiple lo-
cations over time) [35] and modeling dynamics of population in presence of missing
observation [36].

In this study, we focused on the use of a particular class of dynamic models,
that is dynamic linear models (DLM), for handling missing information in water
quality time series data. Chapter 2 includes detailed description of data source and
methodology of the model. Chapter 3 enlists results and important figures. In the
final chapter, we present our concluding remarks with possible directions for futher
research.

List of References


[2] USEPA, “Ambient aquatic life water quality criteria for dissolved oxygen (salt-
water): Cape cod to cape hatteras,” Environmental Protection Agency, Office of

environmental oxygen tension on the growth of the deposit-feeding polychaete,
capitella species 1,” Limnology and Oceanography, vol. 35, no. 7, pp. 1535–1544,
1990.


(oncorhynchus kisutch) and sockeye (o. nerka) salmon at 15 c.” Canadian Journal
of Fisheries and Aquatic Sciences, vol. 38, no. 4, pp. 399–404, 1981.

L. Quéméner, “Effects of repeated hypoxic shocks on growth and metabolism of

[7] M. V. Hoyer, D. Watson, D. Willis, and D. Canfield, “Fish kills in florida’s canals,
creeks/rivers, and ponds/lakes,” Journal of Aquatic Plant Management, vol. 47,


[22] B. Abraham, “Missing observations in time series,” *Communications in Statistics-Theo-

[23] Y. Walter, J. Kihoro, K. Athiany, and H. Kibunja, “Imputation of incomplete non-


[26] R. H, “Maximum likelihood fitting of arma models to time series with missing ob-

[27] M. West and J. Harrison, *Bayesian forecasting and dynamic models*. Springer Science & Busi-
ness Media, 2006.


[31] R. H. Shumway and D. S. Stoffer, “An approach to time series smoothing and fore-


CHAPTER 2

Methodology

2.1 Data Source and Study Area

Dataset for this study has been obtained from Narragansett Bay fixed-site water quality monitoring network (NBFWQMN). This network of fixed-site monitoring stations (Figure 1) across the bay area has been put together by Rhode Island Department of Environmental Management (RIDEM), Narragansett Bay Commission (NBC), University of Rhode Island Graduate School of Oceanography (URI GSO), Narragansett Bay National Estuarine Research Reserve (NBNERR) and Massachusetts Department of Environmental Protection (MADEP) - with a sole purpose of assessing the water quality of Narragansett Bay.

This temperate medium-sized (370 km$^2$) estuary is home to a versatile benthic community and a low level of DO poses a major threat to its ecosystem. Periodic occurrences of hypoxic events during summer were observed in the mid-1990s through the analysis of continuous DO measurements obtained from 5 separate monitoring stations (3 from URI GSO & 2 from NBNERR) [1]. Though the purpose of this earliest setup was to monitor bay-wide water quality as a supplement for analyzing - how physical and/or chemical properties of water would affect different benthic species, soon researchers realized the importance of such surveillance and added more sampling sites across the bay. Fishkill event of Greenwich Bay, a shallow embayment located on the western side of Narragansett Bay (F5 station in Figure 1), in August 2003 [2] aided toward this extension. Currently, 15 active monitoring stations are continuously collecting water quality data for improved monitoring of hypoxic events. Some of these stations collect data all through the year-round and some are deployed seasonally (mid-May to October).

Aggregation of data from multiple stations was initiated in 2004 and made pub-
Figure 1: Narragansett Bay fixed site water quality monitoring network locations (image source: http://www.dem.ri.gov/programs/emergencyresponse/bart/stations.php#map)

Each site records measurements within an equal interval of 15 minutes at two depths: near-surface level and near-bottom level (0.5 or 1 m above the seafloor). Along with DO, data are collected on other water quality parameters, which are: temperature, salinity, pH, chlorophyll level, and turbidity. All these parameters are needed to be monitored because changes in their values can greatly affect the overall structure of the marine ecosystem [3].

Like other tool-based measurement systems, this dataset consists of missing information occurring due to fouling growth of machine, incorrect calibration, or extreme weather conditions. Several researchers used this dataset to analyze episodic occur-
Figure 2: Distribution of missing values in daily measurement of temperature, salinity and DO% of GB (left panels) and TW (right panels).

References of hypoxia and its spatio-temporal variability across Narragansett Bay [4, 5]. The effect of nutrient loadings coming from major freshwater inputs of this estuary has also been analyzed [6] using this dataset. For all these studies, however, linear interpolation has been the tool for handling missing information. Our study will be a stepping stone for statistical analysis of missing values for this time-series data.

DO is measured by recording percentage saturation of oxygen in water compared to air, that is, \((\text{millivolt signal proportional to the concentration of oxygen in water/millivolt signal proportional to the concentration of oxygen in air}) \times 100\). This value is multiplied with a correction factor from DO manual charts corresponding to specific temperature and salinity calculated at the time of measurement. These charts
enlist the values of DO concentration at 100% air saturation for different values of temperature and salinity. By multiplying these charted values to DO%, we finally obtain the measurement of DO concentration (mg/l) which is a direct indicator for hypoxic events. This implies that the temperature and salinity values are required to calculate DO mg/l from DO%. Therefore, the imputation of missing data is a must for these two time-series as well. In addition, by incorporating temperature and salinity data into our model we can obtain better estimates of the missing DO measurements, as the variability of oxygen level depends on these two physical properties of water - among other factors. For imputation, it makes more sense to work with DO% time series because this measurement is collected directly from water while DO mg/l is calculated based on the measurement of temperature and salinity.

We focused our analysis on handling missing data of saturated DO measurements for bottom level sonde in Greenwich Bay (GB) and T-Wharf (TW) sites (F5 and F3 in Figure 1 respectively). The dataset consists of daily averaged values of water quality parameters for these sites from 04/14/2004 to 12/31/2014. The distributions of missing values (represented by shaded background) for temperature, salinity, and DO% are shown in Figure 2. Percentages of missing data in these time series are 12.8, 23.5 and 24.6 respectively for the GB site. These numbers are higher compared to those in the TW site. For this site, 5.2%, 5.3% and 7.5% values are missing for temperature, salinity, and DO% measurements respectively. As seen in Figure 2, sometimes the time intervals of missingness overlap for all three series, and sometimes they don’t.

2.2 Dynamic Linear Models

Dynamic linear model (DLM) is a special class of dynamic stochastic models, where latent state variables change over time. We can describe DLM as linear models with time-varying (dynamic) coefficients. These coefficients are the so-called latent
states that are estimated using observed data in a recursive algorithm. The following
description of the methodology for DLM has been adopted from Prado & West (2010) [7]. The simplest form of linear regression is the intercept only model. From DLM perspective, we can assume that the dynamic intercepts, say $\theta_t$, are the underlying stochastic process varying over time while we observe $y_t = \theta_t + \nu_t$, where $\nu_t$ is a random error term with mean zero. By adding independent variables in the model similar to multiple linear regression, we consider dynamic slope parameters for each of them. These dynamic coefficients represent changes in the linear dependency between the dependent variables and the corresponding independent variables over time. The model can be formally defined using the following equations:

**Observation equation:**

$$y_t = F_t'\theta_t + \nu_t$$  \hspace{1cm} (1)

**Evolution equation:**

$$\theta_t = G_t\theta_{t-1} + w_t$$  \hspace{1cm} (2)

Given below are the different components and assumptions of DLM:

- $\theta_t = (\theta_{t,1}, ..., \theta_{t,p})'$ is the $p \times 1$ state vector at time $t$.
- $F_t$ is a $p$-dimensional vector of known regressors at time $t$.
- $\nu_t$ is the observation noise following Gaussian distribution with mean 0 and variance $\nu_t$.
- $G_t$ is a known $p \times p$ matrix, usually referred to as the state evolution matrix at time $t$.
- $w_t$ is the state evolution noise at time $t$, distributed as zero mean multivariate Gaussian distribution with covariance matrix $W_t$.
- The noise sequences $\nu_s$ and $w_t$ are independent of each other.

A DLM is completely defined by the four components, $\{F_t, G_t, \nu_t, W_t\}$. 
Auto-regressive (AR) process of the series, if any, can be modeled by considering $\nu_t$ as a zero mean AR($q$), that is AR process of order $q$. Such process is defined as,

$$\nu_t = \phi_1 \nu_{t-1} + \phi_2 \nu_{t-2} + \ldots + \phi_q \nu_{t-q} + \epsilon_t$$

where $\epsilon_t$ has a Gaussian distribution with variance $\nu_t$. Our goal is to infer about the unknown state $\theta_t$ including the AR components $\Phi = \{\phi_1, \phi_2, \ldots, \phi_q\}$ and, variance components $\nu_t$ and $W_t$. In this study, we are assuming that the variance of the observational error is constant across time, i.e. $\nu_t \equiv v$.

### 2.2.1 Basic Theory for The Inference of DLMs

First step for inference in DLM is finding parameters of the joint distribution of all the state vectors, $\theta_t$, given data. This is done in two steps: sequential updating and Retrospective updating. These steps are fundamental for both classical or Bayesian inference. The details about the steps are described below.

**Sequentially Updating: Filtering**

At initial time point, $t = 0$, we assume Gaussian distribution as prior distribution of state vector $\theta_0$ with mean vector $m_0$ and covariance matrix $C_0$. For $t > 0$, prior distributions of state vectors is updated sequentially by calculating mean vector, $a_t$ and covariance matrix, $R_t$, using the following equations:

$$a_t = G_t m_{t-1}$$

$$R_t = G_t C_{t-1} G_t' + W_t$$

Following the property of Gaussian distribution, one-step-ahead predictive distributin at time $t - 1$ given prior observations, $D_{t-1} = \{y_{t-1}, D_{t-2}\}$ is $N(y_t | f_t, q_t)$ with

$$f_t = F'_t a_t \text{ and } q_t = F'_t R_t F_t + \nu_t$$

The posterior distribution of state vector $\theta_t$ is $N(\theta_t | m_t, C_t)$, obtained using
Bayes’ theorem, with parameters calculated from the following equations:

\[ m_t = a_t + A_t e_t \quad \text{and} \quad C_t = R_t - A_t A'_t q_t \]  

(5)

where \( A_t = R_t F_t / q_t \) and \( e_t = y_t - f_t \).

This recursive algorithm allows us to estimate \( \theta_t \) even if \( y_t \) is missing by simply assuming \( D_t = D_{t-1} \). Hence, equation (5) reduces to the following: \( m_t = a_t \) and \( C_t = R_t \). That means, we are updating missing information using the observed value at hand and assumed structure of the series determined by \( F_t \) and \( G_t \) only. This is, otherwise, accepting that the reason for a missing value do not provide information relevant to the inference on the data (ignorable missing data mechanism) given observed \( y_t \)'s, \( F_t \) and \( G_t \). Therefore, the new information about missing observations are extracted from the data itself. Using the equations (3)-(5) for each time point, we obtain the distribution of \( p(\theta_t | D_t) \).

**Retrospectively Updating: Smoothing**

Forward filtering provides information on \( \theta_t \) by summarising current information through \( D_t \). Sequential updating of equation in this way also provides information about the past. For example, the observed value of \( y_{t+1} \) not only contain information about \( \theta_t \) but also about \( \theta_t \). We can use this information to do backward sampling of current state \( \theta_t \) given future state \( \theta_{t+1} \) from \( p(\theta_t | \theta_{t+1}, D_T) \). This leads to an algorithm that provide us with the joint distribution of \( p(\theta_0, \theta_1, \theta_2, ..., \theta_T | y_1, y_2, ..., y_T) \). We start off with sampling state vector \( \theta_T \) of end time point \( T \) from \( \mathcal{N}(m_T, C_T) \). For \( t = T - 1, T - 2, ..., 0 \) we sample \( \theta_t \) from \( \mathcal{N}(a_{t}(t - T), R_{T}(t - T)) \) recursively where

\[ a_T(t - T) = m_t - B_t [a_{t+1} - a_T(t - T + 1)] \]  

(6)

\[ B_t = C_t G'_{t+1} R_t^{-1} \]  

(7)

\[ R_T(t - T) = C_t - B_t [R_{t+1} - R_T(t - T + 1)] B'_t \]  

(8)
2.2.2 Bayesian Computation for DLM

Statistical inference can be done with frequentist approach or Bayesian approach. In the first approach, we consider that the model parameters as constant terms to be estimated from data. In the second approach, we treat the model parameters as random and try to find their distribution given data. By using forward filtering and backward sampling in DLM, we either obtain parameters of the assumed distribution of the latent states (frequentist approach) or obtain samples of latent states from the assumed distribution (Bayesian approach) defined by the components \{F_t, G_t, \nu_t, W_t\}. From the perspective of handling missing data, the frequentist approach is only concerned with making inference about model parameters after integrating out the missing information, which is not trivial when model is complicated, while missing observations are generated in the Bayesian approach. Thus the latter approach provides us with multiple 'imputed' samples of the unobserved part of the series along with a band for uncertainty. For this reason, Bayesian approach is more practical for our time series of interest as we want to be able to monitor oxygen level regularly even if instrument failure occurs.

Inference of Unknown State Variables

Equations (3)-(8) already provide us functions to carry out a Markov Chain Monte Carlo (MCMC) sampling procedure called Gibbs sampling to obtain required number of samples of \(\theta_{0:T}\). With known \(\nu_{1:T}\) and \(W_{1:T}\), we use the following steps to obtain our required samples of state vectors:

1. We start from equations (3) - (5) to obtain \(\mu_{1:T}\) and \(C_{1:T}\).

2. These vectors and matrices are then used in equations (6) - (8) for the calculation of \(a_T(t - T)\), \(R_T(t - T)\) recursively.

3. We sample \(\theta_t\) from \(N(a_T(t - T), R_T(t - T))\) for each data point \(t =\)
4. We repeat steps 1 to 3 to obtain our required number of samples of $\theta_{0:T}$.

Inference of AR components and corresponding variance, $v$

Let us consider that $\nu_t$ has an AR(1) process. Then, under model structure, $\nu_t = y_t - F_t' \theta_t$ and $\epsilon_t = \nu_t - \phi \nu_{t-1}$. Given $\theta_{0:T}$, the joint distribution of $\phi, \nu_{0:T}, \epsilon_{1:T}$ and $v$ is proportional to

$$p(\phi)p(v)p(\nu_0|\phi, v) \prod_{t=1}^T p(\nu_t|y_t)p(\epsilon_t|\epsilon_{t-1}, \phi, v)$$

By considering $\mathcal{N}(c, C)$ and $\mathcal{G}(a, b)$ as priors for $\phi$ and $v^{-1}$, we get the following full conditional distributions:

$$\phi|\nu, \nu_{1:T} \sim \mathcal{N}\left(\frac{c/C + \sum_{t=1}^T \nu_t \nu_{t-1}}{1/C + \sum_{t=1}^T \nu_t^2/v}, \frac{1}{1/C + \sum_{t=1}^T \nu_t^2/v}\right)$$  \hspace{1cm} (9)

$$v|\phi, \nu_{0:T} \sim \mathcal{IG}\left(a + T/2, b + \sum_{t=1}^T \nu_t - \phi \nu_{t-1}\right)$$  \hspace{1cm} (10)

These functions can be used to obtain samples of $\phi$ and $v$ after sampling for $\theta_t$ at each step of MCMC algorithm described above.

Inference of Evolution Variances, $W_t$

When data is missing at time point, $t$, we have $C_t = R_t$. For next data point

$$R_{t+1} = G_t R_t G_t' + W = G_{t-1}[G_{t-1} R_{t-1} G_{t-1}' + W_{t-1}] G_t' + W_t$$

. With increasing number of sequentially missing observations, $R_t$ becomes increasingly large. Larger value correspond to less precise estimate for the state variables. To counter it, we can use method of discounting $W_t$. Let $P_t = G_t C_t G_t'$. Then $P_t$ can be regarded as the prior variance in a DLM having constant state vector, i.e. $W_t = 0$. This assumption can be made flexible if we consider $R_t = P_t/\delta$, where $\delta \in (0, 1]$. That is by choosing a value of $\delta$ further from 1, we can control the variability of state variables. Under this structure, we have
\[ W_t = \frac{1 - \delta}{\delta} P_t \]

Hence, we just need to sample \( \delta \) which can be computationally consuming. In practical situation, we select a range of possible values for \( \delta \) where \( \delta > 0.9 \). After assigning a discrete uniform prior for the chosen values, we obtain a posterior distribution proportional to

\[
p(\delta)p(y_{1:T}|D, \delta) = p(\delta) \sum_{t=1}^{T} p(y_t|D_{t-1}, \delta)
\]

(11)

This distribution is used to sample \( \delta \) at each step of Gibbs sampler.

### 2.2.3 Data analysis plan

We want to analyze the time-series data of DO\% \( (y_t) \) using temperature \( (z_{1,t}) \), salinity \( (z_{2,t}) \) and river discharge data \( (z_{3,t}) \). However, temperature and salinity needs to be modeled as well because these time-series contain missing data. The variability of salinity in water depends on fresh water inputs either from rainfall \( (x_{1,t}) \) or river discharge \( (z_{3,t}) \). Hence, these dataset can be used to model salinity data and obtain estimates for the missing parts. The joint distribution of the data is,

\[
P(y, z_1, z_2 | z_3, x_1) =
\]

\[
P(y|z_1, z_2, z_3)P(z_1)P(z_2|z_3, x_1)
\]

First we define the structures of the three times series DLMs by their respective model components \( \{F_t, G_t, \nu_t, W_t\} \). This is done by analysing characteristics of each series and their interdenpendencies. The variability of DO\% depends on both temperature and salinity. Therefore, we sample temperature and salinity first, and then include these series as independent variables for DO\% DLM. A conceptual framework of the entire procedure is provided in Figure 3. The Gibbs sampling procedure for obtaining samples of the model parameters, \( \{\theta_{0:T}, \phi, v, \delta\} \), is described below:
1. We start by setting up initial values for $\phi^{(0)}, v^{(0)},$ and $\delta^{(0)}$. We use these to obtain $\nu_{1:T}^{(0)}$ and $W_{1:T}^{(0)}$.

2. Using forward filtering and backward sampling algorithm we obtain $\theta_{0:T}$ given $\nu_{1:T}$ and $W_{1:T}$.

3. Given $\theta_{0:T}$, we use the full conditional distributions in (9), (10) and (11) to obtain samples of $\phi, v$ and $\delta$ respectively.

4. We repeat step 2 and 3 until we have our required number samples.

For model inference, we can use arithmetic means of these samples as corresponding parameters estimates. For state vector, $\theta_t$, this is simply the expectation of the posterior distribution, that is $E(\theta_t|y_1, y_2, y_3, ..., y_T)$. We can use 2.5% & 97.5% quantiles of these samples to calculate 95% credible intervals (CI) for $\theta_{0:T}$. At each step of the sampling process we can use equation (1) and (2) to obtain samples of $\tilde{y}_t$ representing posterior predictive distribution of the series given our observation.

List of References


Figure 3: Conceptual framework of data analysis plan
CHAPTER 3

Results

3.1 Preliminary Data Analysis

In this section, we explore the characteristics of each time series. Our goal is to determine the structure of the model components, \( \{F_t, G_t, \nu_t, W_t\} \), after looking into the autocorrelation (ACF), partial autocorrelation functions (PACF) of the series, and cross-correlation function (CCF) between the series. Details about the methodology of ACF, PACF, and CCF are given in Appendix A.

3.1.1 Exploratory Data Analysis of Temperature

If we look at the time series plots of the temperature in Figure 2, we notice a periodicity of one year. By taking the difference of the series by 365 days, \( \Delta_{365} \text{Temperature}_t = \text{Temperature}_t - \text{Temperature}_{t-365} \), we remove the seasonality so that we can determine the order of auto-regressive or moving average process of \( \nu_t \). Figure 4 shows the ACF and PACF plots of \( \Delta_{365} \text{Temperature}_t \) for both sites. The ACF values are decreasing exponentially while PACF at lag 1 has the highest significant value. Therefore, it is plausible that \( \nu_t \) might be well described by an AR(1) process. Considering these results, we assume the following structure of the temperature DLM components \( \{F_t, G_t, \nu_t, W_t\} \) for both sites:

\[
G_t = G = \begin{pmatrix}
1 & 0 & 0 \\
0 & \cos\left(\frac{2\pi}{365}\right) & \sin\left(\frac{2\pi}{365}\right) \\
0 & -\sin\left(\frac{2\pi}{365}\right) & \cos\left(\frac{2\pi}{365}\right)
\end{pmatrix},
\]

\[
F'_t = F = (1, 1, 0),
\]

\[
W_t = \begin{pmatrix}
w_t & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}.
\]
\[ \nu_t = \phi \nu_{t-1} + \epsilon_t \]

The Fourier components in \( G_t \) captures the seasonality of the series. By defining \( W_t[2, 2] = W_t[3, 3] = 0 \), we assume that the coefficients corresponding to the Fourier components are non-stochastic and dynamic intercept, \( \theta_t = \{\beta_{0,t}\} \) is the only unknown state variable. We control the variability of the state variable by sampling for \( \delta = \{0.99, 0.992, 0.995\} \) weighted by the posterior predictive distribution given by (11).

Figure 4: ACF and PACF plot of \( \Delta_{365} Temperature_t \) with blue lines representing 95% confidence intervals
3.1.2 Exploratory Data Analysis of Salinity

The salinity of the water is highly influenced by freshwater input either from the nearby rivers or from precipitation. One of the major rivers opening to Narragansett Bay is Moshassuck river. A side by side comparison of daily river discharge data, shown in Figure 5, reveals that salinity decreases around the same time river discharge increases. This is further confirmed by cross-correlation analysis (details of this method in given in Appendix A) of the series in Figure 6. We see that there is a negative correlation between river discharge and salinity at lag 0 for both sites. Similar analysis (figures B.1 & B.2) reveals that significant negative correlation exist between precipitation series and salinity at lag 0. By including these series in the model, we can obtain more information about the missing part of the series. For this study, we are only considering precipitation and river discharge data measured on the same day (lag 0) as salinity to be included as independent variables into our model. Hence we have the following components of the Salinity DLMs:

\[ F_t' = [1, X_{1t}, X_{2t}] \]

\[ G_t = I_3 \]

\[ W_t = \begin{pmatrix} w_{1t} & 0 & 0 \\ 0 & w_{2t} & 0 \\ 0 & 0 & w_{3t} \end{pmatrix} \]

Here \( X_{1t} \) and \( X_{2t} \) represent water discharge and precipitation at time \( t \) respectively. The variability of each states, \( \theta_t = \{\beta_{0t}, \beta_{1t}, \beta_{2t}\}' \), are controlled by sampling discount factor from \( \{0.975, 0.98, 0.99\} \) using the posterior predictive distribution of \( \delta \) in (11).

\(^1\)Data source: USGS National Water Information System (https://waterdata.usgs.gov/nwis/rt)

Figure 5: Time series plot of river discharge data with salinity for both sites

Figure 6: Cross-correlation analysis between river discharge data and Salinity

The salinity time series don’t have a periodic structure as the temperature time series. This is also apparent in the ACF and PACF plots shown in Figure 7. The ACF values of both series are decreasing exponentially indicating mean stationarity. The PACF values of both series are highest at lag 1. Although values at lag 2,3,4,5 are significant, the magnitude is much smaller than the value at lag 1. For our study, we model $\nu_t$ as an AR(1) process.

3.1.3 Exploratory Data Analysis of DO%

As mentioned in chapter 2, we need measurements of temperature and salinity to calculate DO concentration from saturated DO measured by the tool. Further, both of these variables directly affect the solubility of oxygen level in the water. Cross-correlation analysis between temperature and DO% in Figure 8 reveals that daily
temperature measured in the same day as DO%, (lag 0) has a positive correlation while temperature recorded in the previous day (lag -1) has a negative correlation. Similarly, between salinity and DO%, we find that cross-correlation function (CCF) is significant at lag 0 for both sites 9. At first, we included temperature series at lag 0 and at lag -1 in the model. However, by doing this we were introducing strong multicollinearity in the model. To avoid this situation, we took the average of temperature at lag 0 and lag -1, and included the resulting series as independent variable into the DLMs. In addition to this, we also want to account for the changes in DO% due to freshwater inputs from the Moshassuck river by including this series into our model. After analysing the cross-correlation functions between the two series (see Figure B.3), we included river discharge data at lag 0 as independent variable in
Finally, we have our model components as:

\[
F_t' = [1, Z_{1t}, Z_{2t}, Z_{3t}], \quad G_t = I_4 \quad \text{and} \quad W_t = \begin{pmatrix}
    w_{1t} & 0 & 0 & 0 \\
    0 & w_{2t} & 0 & 0 \\
    0 & 0 & w_{3t} & 0 \\
    0 & 0 & 0 & w_{4t}
\end{pmatrix}
\]

Here \( Z_{1t}, Z_{2t} \) and \( Z_{3t} \) represent average temperature at time \( t \) & \( t - 1 \), salinity and river discharge data at time \( t \) respectively. Hence, our state variables consist of four coefficients, \( \theta_t = \{\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}\} \). The discount factor is sampled from \{0.988, 0.989, 0.99, 0.995\} using the posterior predictive distribution of \( \delta \) in equation (11). The observational error is modeled as an AR(1) process (see Figure B.4).

---

**3.2 Model Results**

For our study we generated 15000 samples of each parameters and took every 15\(^{th}\) one to allow for proper representation of posterior density. In this following section, we show model outputs of temperature, salinity and DO% time series. We draw posterior inference for each model parameters from the MCMC samples.

**3.2.1 Temperature Model Results**

The mean functions of both temperature time-series, \( (F_t \theta_t) \), are modeled with a seasonal component representing annual fluctuation and dynamic intercept repre-
senting trend. The observational variance is assumed to have AR(1) distribution. Decomposition of the model by different components are shown in Figure 10. In the top row we have predicted series by the model along with the 95% CI for both sites. We see that the predicted series closely resembles the actual time series showed in 2. It also produces estimates for missing part of the series with larger 95% CI. An upward trend is noted in both of the series from 2011 to 2012. This trend reflects the event that winter 2011-2012 was the second warmest for Rhode Island [1]. While Fourier component captures the annual fluctuation (fourth row in Figure 10), the AR(1) noise process captures the dependency of values between the consecutive time points (third row in Figure 10). Figure 11 shows the estimated mean (red vertical line) and 95% credible intervals (shaded area under the curve) obtained from the posterior density of AR(1) coefficient, $\phi$, and the variance, $v$. The posterior means of $\phi$ for GB site and TW site are very close, 0.92 and 0.94 respectively, with no inclusion of null value in the corresponding credible intervals. The posterior mean of the variance for GB site is 0.020 while the variance of actual series is 67.71. This means that the model is able to capture most of the variability of the data. For TW site, estimated variance is 0.016 compared to overall variance of 42.70. In the same figure, we also
have the posterior distribution of the discount factors, $\delta$, corresponding to the dynamic intercepts. Of the three values being sampled, 0.99 has the highest posterior probability and 0.995 has the lowest for both series.

### 3.2.2 Salinity Model Results

The decomposition of the DLM for Salinity by dynamic intercept and other regression components are shown in Figure 12. Here, the dynamic intercept provides us with an estimation of the overall trend of the series after we have removed the daily variability (using AR(1) component), and the variability occurred with river flow. The estimated value of dynamic coefficients of the river discharge data, $\beta_{1t}$ and precipitation, $\beta_{2t}$ is shown in 13. If we compare the trend of $\beta_{1t}$ for both sites, we see some similarity between them. For example, both plots show convexity around 2004 and 2009. The 95% credible intervals for $\beta_{1t}$ rarely includes the null value 0. This confirms the fact that salinity of the water for both sites are largely influenced by the fresh water inputs from the Moshassuck river. Also, the value of this coefficient is negative indicating fresh water input decreases salinity. The contribution of precipitation for this series is negligible. Figure 14 shows the posterior inference of AR(1) coefficients and variances. We obtain similar estimation of the AR(1) coefficients for both sites (0.83). The unexplained variability of the GB site data is estimated to be 0.316 while the actual variance is 2.27. Similarly, for TW site, we have 0.70 as the series variance compared to 0.10 as estimated error variance. The posterior probability of the discount factor is also shown in Figure ?? We see that, for each sample only 0.975 is chosen. Sampling of discount factor is not so straightforward. We may need values more close to each other to trigger some variability in the posterior sample.
Figure 10: Decomposition of the temperature DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI).
Figure 11: Posterior distribution of AR(1) coefficient ($\phi$), variance ($v$) and discount factor $\delta$ for temperature DLM. For $\phi$ and $v$ we have posterior density plot where the vertical line represents estimated mean and the shaded region represents 95% CI. For $\delta$ we have the posterior probability.

3.2.3 DO% Model Results

Figure 15 displays the estimated dynamic intercept and other regression components for DO% DLM with 95% CI. The first row shows posterior estimation of the
Figure 12: Decomposition of the salinity DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI).
Figure 13: Dynamic regression coefficients for salinity DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI).

overall series including the missing observations. The model estimates the series with small CI for the observed part and with large CI for missing part. In the first plot, we notice that the model is able to estimate the periodicity of the series quite well for the year 2005 and 2006. These estimates are clearly better than doing linear interpolation between the two observed values. The dynamic intercepts of both sites don’t show major fluctuation between the years 2004 and 2012. Beyond that point, we see an upward trend in GB and a downward trend in TW. The AR(1) observational error removes the daily variability from the series completely. Regression component corresponding to temperature capture the seasonal variability of the series quite well. If we see the posterior estimation of the coefficient corresponding to this variable in Figure 16, we notice that the 95% CIs don’t contain 0 and all the values are negative. This makes sense because temperature has a negative effect on oxygen level. The salinity time-series fail to explain any the variability of the DO% time-series for both sites. This is also confirmed by the fact that 95% CI corresponding to this coefficient contains the null value (Figure 16). Contribution of river discharge data is significant.
Figure 14: Posterior distribution of AR(1) coefficient, $\phi$, variance, $v$ and discount factor, $\delta$ for salinity DLM. For $\phi$ and $v$ we have posterior density plot where the vertical line represents estimated mean and the shaded region represents 95% CI. For $\delta$ we have the posterior probability.

for these series, although not so salient compared to the contribution of temperature time series. Estimated value of AR(1) coefficient is larger for TW compared to GB.
(see Figure 17). This can be seen in the ACF and PACF plots of the differenced series in Figure B.4. Overall variance of the DO% in GB is much larger than that of TW. Among the four values of discount factor, 0.988 has the highest posterior probability.

List of References

Figure 15: Decomposition of the DO% DLMs of the two sites (red line represents arithmatic mean of the samples and blue lines represent 95% CI).
Figure 16: Dynamic regression coefficients for DO% DLMs of the two sites (red line represents arithmetic mean of the samples and blue lines represent 95% CI).
Figure 17: Posterior distribution of AR(1) coefficient (φ), variance (v) and discount factor δ for DO% DLM. For φ and v we have posterior density plot where the vertical line represents estimated mean and the shaded region represents 95% CI. For δ we have the posterior probability.
CHAPTER 4

Discussion

The focus of this study is to analyze the time-series data with missing observation for Narragansett Bay water quality measurements. Instead of using the conventional methods for handling missing data for this time-series, we used DLM. Under this methodology, we were able to model distinct characteristics of each time-series as well as the interdependencies between them.

The time-series of temperature displayed the most predictable pattern. We used Fourier component to model seasonal variability and AR(1) observational error to model the auto-regressive process of the series. The dynamic intercept for this model was able to show the overall trend of the temperature for both sites. To model the salinity time series, we incorporated the time series data of freshwater inputs from the Moshassuck river and precipitation. We found that salinity in water for both of the sites was controlled mostly by river discharge data. Finally, to model DO% time-series data, we used the sampled values from the temperature DLM and the salinity DLM. We also included river discharge time-series data in the model because freshwater has high concentration of oxygen. We found that the variability of oxygen level largely depends on the water temperature, and then on river discharge. Influences of salinity were found to be insignificant. After excluding the variability of oxygen due to temperature and river discharge we were left with an overall trend of the series. While modeling these time-series, we also obtained estimations for the missing part of the time series along with a measure of uncertainty.

A limitation of this model in applying to large scale time-series data, such as our data, is that the computation time is very long. To draw single set of samples for all the parameters of the three DLMs, it took around 55 seconds. It took around
10 days to generate 15000 samples of parameters using single core. Sampling of the
discount factor also adds to its complexity. If we choose more values to sample from,
for $\delta$, the longer it takes to complete one iteration. For fast computation, we can
adopt Intergrated Nested Laplace Approximation (INLA) [1]. Under this framework,
recursive updating of state variables is not required and we can directly approximate
the posterior distibution. In this study, we selected a few values of discount factor
and sampled them using the posterior predictive distribution after repeatedly running
the algorithm for different combinations of values. The set of discount factor chosen
is not common for all three DLMs because variances are different for each three time
series.

Nutrient loadings from rivers play a major role in the variability of oxygen level
in water [2]. A potential study could be the inclusion of nutrient loadings time-series
data as an indepedent variable in the model for DO%. Moreover, we only analysed
data from two sites and modeled them separately. It would be interesting to do a
multivariate DLM by including more sites and model them simultaneously. If we
look at the variability of DO% for GB and TW site, we see that sometimes they show
similar structure although the former has a large variability than the latter. Modeling
them simultaneously would allows us to explore the interdepeencies between the sites.
By doing this we can also model spatio-temporal dynamics of dissolved oxygen across
NB.

List of References


McManus, and L. Reed, “Managed nutrient reduction impacts on nutrient con-
centrations, water clarity, primary production, and hypoxia in a north temperate
APPENDIX A

A.1 ACF

The auto-correlation function (ACF) is a measurement of correlation between observations of a time series, $Y_t$ and $Y_s$. This is defined by,

$$
\rho_{t,s} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}}
$$

where $Cov(Y_t, Y_s) = E[(Y_t - E(Y_t))(Y_s - E(Y_s))]$ and $Var(Y_t) = E[(Y_t - E(Y_t))^2]$.

ACF at lag $q$ is simply the correlation between $Y_t$ and $Y_{t-q}$.

A.2 PACF

The partial-correlation function (PACF) is a measurement of correlation between observation of a time series after removing the correlation between other shorter lags. For example, PACF at lag $k$ is defined as,

$$
\psi_{kk} = Corr(Y_t, Y_{t-k}|Y_{t-1}, Y_{t-1}, ..., Y_{t-k+1})
$$

To achieve this, we first regress $Y_t$ and $Y_{t-k}$ over $Y_{t-1}, Y_{t-1}, ..., Y_{t-k+1}$ and, then calculate the correlation from the resulting residuals.

A.3 CCF

The cross-correlation function (CCF) is a measurement of correlation between two time series observation, say $Y_t$ and $X_t$ where $t = 1, 2, ..., T$. The CCF at lag $k$ is defined as,

$$
\rho_k(X,Y) = Corr(X_t, Y_{t-k}) = Corr(X_{t+k}, Y_t)
$$
Figure B.1: Time series plot of precipitation data with salinity for both sites

Figure B.2: Cross-correlation analysis between river discharge data and Salinity
Figure B.3: Cross-correlation analysis of river discharge data and DO% with blue lines representing 95% confidence intervals.
Figure B.4: ACF and PACF plot of $\Delta_{365} DO_t$ with blue lines representing 95% confidence intervals.
Bibliography


