2021

INVESTIGATING LONG-TERM PHYSICAL, CHEMICAL AND BIOLOGICAL CHANGES IN NARRAGANSETT BAY USING BAYESIAN MULTIVARIATE DYNAMIC LINEAR MODELS

Jacob P. Strock
University of Rhode Island, jstrock@uri.edu

INVESTIGATING LONG-TERM PHYSICAL, CHEMICAL AND BIOLOGICAL CHANGES IN NARRAGANSETT BAY USING BAYESIAN MULTIVARIATE DYNAMIC LINEAR MODELS

BY

JACOB P. STROCK

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 IN STATISTICS THESIS

OF

JACOB P. STROCK

APPROVED:

Major Professor Gavino Puggioni

Thesis Committee: Susanne Menden-Deuer

Jing Wu

Brenton DeBoef
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2020

ABSTRACT

Within the past 50 years, Narragansett Bay has undergone major physical and chemical changes including climate-induced warming and policy-driven reductions in anthropogenic nutrient pollution. These long-term changes have the capacity to transform the ecological function of Narragansett Bay, but may also represent a case study for global oceanic changes. Despite its importance, the long-term response of ecosystems to climatological change remains uncertain, as well as the consistency of biological interactions with the environment over time. I explore these uncertainties here using Bayesian dynamic linear models (DLMs) to investigate the Narragansett Bay Long-Term Plankton Time Series. In a first stage, DLMs were used both to interpolate missing data and describe changes in seasonality and long-term trend for nitrogenous nutrients, water temperature, and size structure of phytoplankton communities. Among complex physical and chemical changes observed, these models revealed a long-term decline in large phytoplankton and intensifying seasonal blooms for smaller phytoplankton. These changes in size structure of biological communities were expanded through analysis of cross correlations and a second modeling stage where the imputed nitrogen series was used as a predictor of phytoplankton levels in a multivariate dynamic linear regression model (DLR). The DLR revealed a newly discovered seasonal dependence of large phytoplankton on nitrogen sources. Results suggested highly dynamic states and the need for discount specification of covariance matrices. This motivated more general analysis of model selection in time series with high stochasticity and long intervals of missing data. Through simulated data and metrics of model fit including information criteria and forecasting errors, I explored model selection as well as standard and practical discounting methods in series with long intervals of missingness. These analyses highlight one-step-ahead root mean square forecast error as a relatively consistent selection tool, but also evidence the uncertainty in accurate recovery of discount factors in general, and potential impacts on model inference.

ACKNOWLEDGMENTS

I sincerely thank my advisor Dr. Gavino Puggioni and committee members Drs. Susanne Menden-Deuer and Jing Wu for their insights, patience, and thoughtful guidance over the course of this project. This project would also not be possible without the diligence of those who have contributed to the Narragansett Bay Plankton Time Series, including current supervisor, Dr. Tatiana Rynearson, the late Dr. Theodore Smayda who began the Narragansett Bay time-series, and many others who have contributed to the collection, organization and sharing of the data.

PREFACE

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1 Scientific Background

Globally, phytoplankton are responsible for primary production on the order of 36.5-48.5 Gt C yr^{-1} (Antoine et al. 1996; Field et al. 1998). This biological production is critical for everything from biogeochemical cycling (e.g. carbon; Falkowski 1994), to the productivity of marine food webs (Steinberg and Landry 2017). However, with climate change, the ecological functions of plankton communities are at risk.

Cell size represents one such compositional feature, that both varies broadly, and is strongly associated with ecosystem traits. The size of phytoplankton from the unicellular to colony scale covers over 9 orders of magnitude, with a range from <1 µm to several centimeters (Beardall et al. 2009). These size differences manifest into predictive allometric relationships. In general, at the cellular level, size is negatively related to traits including metabolic rate (López-Urrutia et al. 2011), nutrient diffusion and uptake rates (Mei et al. 2009), and light absorption efficiency (Marra et al. 2007), but also positively related to features including biomass-based growth rate under nutrient replete conditions, and tolerance to light variability (Mei et al. 2009; Irwin et al. 2006; Key et al. 2010) . At the ecosystem level, these traits manifest so that cell size is also inversely related to maximum abundance (Irwin et al. 2006), strength of the microbial loop, and food chain length (Sprules and Munawar 1986), but also positively related to features including sedimentation rate, and export efficiency

(Miklasz and Denny 2010). These allometric relationships thereby make cell size a highly informative feature for inferences from cellular to ecosystem function.

While plankton size structure has the capacity to affect marine ecosystem function; biological, chemical, and physical ecosystem traits also feedback to affect the size structure. In the context of climate change, there is thus the potential that the community size structure of phytoplankton may be affected. The Intergovernmental Panel on Climate Change predicts that average global temperature is likely to increase to 1.5 °C above pre-industrial levels between 2030 and 2052, and that the current average rate of increase is 0.2 °C per decade with higher rates of warming towards polar regions (IPCC 2018). At the intraspecific level, cell size is known to scale inversely with temperature (Atkinson et al. 2003). Community level studies have also found community size distribution scaled inversely with temperature (Morán et al. 2010; Hilligsøe et al. 2011). These findings suggest that the global increases in temperature may also result in phytoplankton communities with size distributions tending toward smaller organisms.

Beyond temperature, the size distribution of phytoplankton communities may also be affected by limiting nutrients. At a global scale, stratification is expected to increase and thereby limit the nutrient supply to phytoplankton (Sarmiento et al. 2004; Behrenfeld et al. 2006). Cell size directly imposes a physical constraint on the potential rate of nutrient supply (Mei, Finkel, and Irwin 2009), and thereby community size distribution has been positively related to nutrient availability (Chisholm 1992; Irwin et al. 2006). In future climate scenarios, temperature and nutrient influences may act in concert to skew community size distributions to smaller cells.

These predictions give the potential that the biological function of the ocean may be drastically altered in future climate scenarios by shifting the size structure of phytoplanktonic communities. Notably, most climate predictions and observations of size structure have focused on open ocean environments. However, the impacts of climate change will not be limited to these regions. The coastal environment shows evidence of some of the highest rates of change. Moreover, coastal nutrient levels can be strongly influenced by anthropogenic inputs. While some regions are facing increasing eutrophication (Cloern et al. 2014), in others, government policy has been enacted or proposed to better regulate nutrient pollution (Saarman 2007).

In this study Iuse Narragansett Bay (NB) RI, USA as a study system. NB is a temperate, coastal estuary, where several decades long monitoring efforts have covered long-term anthropogenic influences from climate to nutrient pollution (Fulweiler et al. 2015). Linear regression analyses of long-term temperature record from 1960 to 2012 showed mean annual water temperature has increased 1.4-1.6°C, with warming as high as 2.2°C in the winter (Nixon et al. 2009; Fulweiler et al. 2015). Superimposed on these long-term climate trends are climate oscillations, such as the NAO, which in the positive phase can increase water temperatures by as much as 3°C from the average Oviatt 2004). Beyond temperature rise, Narragansett Bay has undergone major decline in nutrient loadings. Most recently, between 2005 and 2012, the nutrient loading by wastewater facilities were reduced by 50% through tertiary wastewater treatment (DEM 2005). These changes have measurably reduced the standing stock of DIN and DIP in the bay by 50-60% (Oviatt et al. 2017). While nutrient changes in the Bay have been confirmed by separate, discrete field

observations, they have not been considered in a time-series context, and their potential effects on the structure of the phytoplankton community has not been examined.

Previous Statistical Work

To date the published literature has generally focused on descriptive statistics at the annual scale and deterministic models such as simple linear regression. Such methods have worked to describe static, long-term patterns with large, monotonic, and approximately linear effects. However, such summary statistics can be biased or information lost because the full resolution of the data or its inherent features were not considered. Such features include autocorrelation, missingness, seasonality, as well as changing rates and associations amongst variables over time. To my knowledge, there is no published work on the Narragansett Bay Phytoplankton Time-Series using time series methods which may directly address these features. I emphasize that statistically, this may have left a wealth of information hidden. In example of limitations of previous work, while linear regression of temperature series may find a significant fit to long-term trends over decades, it imposes a strict, static relationship which might otherwise be expected to change over the course of the many decades being analyzed. On the multidecadal scale, this means that relevant expected multiyear cycles are not represented (Borkman and Smayda 2009), nor are the intra-annual seasonal cycles relevant on the scale of the biological components which drive the ecosystem function of the bay (Pratt 1965; Durbin, Krawiec, and Smayda 1975; Karentz and Smayda 1984; Lawerence and Menden-Deuer 2012). Of key concern is that constant rates of change are assumed, but it is generally considered that systems

with feed-back mechanisms (positive or negative) or tipping points, are likely to see non-linear rates of change (Steffen et al. 2015).

Models which do not consider the autocorrelated structure of time series data are prone to type 1 errors when the cross-correlation between series is considered. Such spurious correlation between series is the result of inflated cross-correlation due to remaining autocorrelation in the series (Yule 1926). Independent series may have inflated covariance when at least one of the series is not first filtered to white-noise (pre-whitening). This is an important process because it can otherwise lead to the interpretation of spurious correlation between series as meaningful associations (Katz 1988).

Another important feature of these time series data which has not been addressed are the missing data. Missingness is critical to address because it may represent both a loss of information and bias in results. To exemplify the potential for bias, consider the most common analysis method, where annual means are used. Because of the strong seasonal signature, if data are missing in a given season, then the annual metric may be biased. The loss of information also represents a major concern because it is the intra-annual patterns which drive the ecological dynamics of the bay (Pratt 1965; Durbin, Krawiec, and Smayda 1975; Karentz and Smayda 1984; Lawerence and Menden-Deuer 2012). It is also known specifically that size structure of phytoplankton have a distinct seasonal pattern (Durbin, Krawiec, and Smayda 1975). Further, the phytoplankton community is one with high turnover, and so their ecology is not represented by integrated statistics over long periods in time. In consultation with managers of the data series, it is believed that missingness is in

generally not conditional on the missing value (i.e. missing not at random, MNAR). Instead, the data missingness is more likely to be missing completely at random (MCAR, not conditional on any observed or unobserved value), or missing at random (MAR, conditional on some observed value such as missingness at a previous time-point).

State space models are one class of flexible time series model with the capacity to address key data features including and not limited to changing associations over time, missingness, seasonal structure, and autocorrelation. There is limited but growing literature on the utility of Bayesian state space models in plankton ecology, demonstrating that these models may be critical in elucidating the driving relationships in ecology which are unlikely to be static with time (Arhonditsis et al. 2007; Jones et al. 2010). The general structure of these models allows easy decomposition into their additive components. State space models, in general, model latent states through Markovian dependence structure (depending only on the previous state), assuming measured data are observations of these states with some error (Fig. 1). This Markovian structure is a key trait that allows time-specific parameter estimates. The proceeding methodology section outlines a specific case of state-space models, the dynamic linear model (DLM), where Gaussian distributed errors are assumed. The DLM alone is flexible in structure, and the cases represented here will be a few specific examples relevant to this application (e.g. West and Harrison 1997, Prado and West 2010).

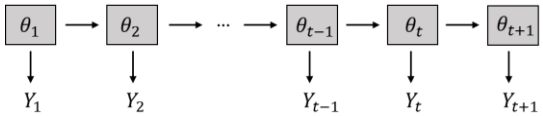

Figure 1. the Markovian dependence structure of state-space models, with each latent state at time i (θ_i) and corresponding observations of each state (Y_i).

In this study, I investigate change in the size structure of phytoplankton in Narragansett Bay from 2003-2019 through a weekly record of size fractioned chlorophyll (chl. a) and associated measurements of surface water temperature, nitrate and nitrite (NO_2^-+ NO_3^-), and ammonia (NH_4^+) concentration. A bivariate Bayesian dynamic linear regression is used to model changes in the chlorophyll concentration in two size groups (<20 µm), and (>20 µm) as a function of these environmental variables. In this thesis, I use the decompositions of these models to demonstrate the changes in size structure in Narragansett Bay, and test dependence on environmental traits.

1.2 Data Description

Surface water temperature, NO_2^-+ NO_3^- concentration, NH_4^+ concentration, chlorophyll a concentration less than 20 µm, and chlorophyll a concentration greater than 20 µm were all obtained from the University of Rhode Island Long-Term Plankton Time Series of Narragansett Bay from 2003-2019 at weekly resolution from the publicly available dataset (https://web.uri.edu/gso/research/plankton/data/, fig. 2, 3). When available, missing data from the temperature dataset were filled with surface temperature measurements from the University of Rhode Island Fish Trawl

(https://web.uri.edu/fishtrawl/data/), which are taken at the same location less than 1 hour later. Dissolved inorganic nitrogen (DIN), a frequently limiting nutrient for growth in marine environments is represented here as the sum of NO_2^-+ NO_3^- and NH_4^+.

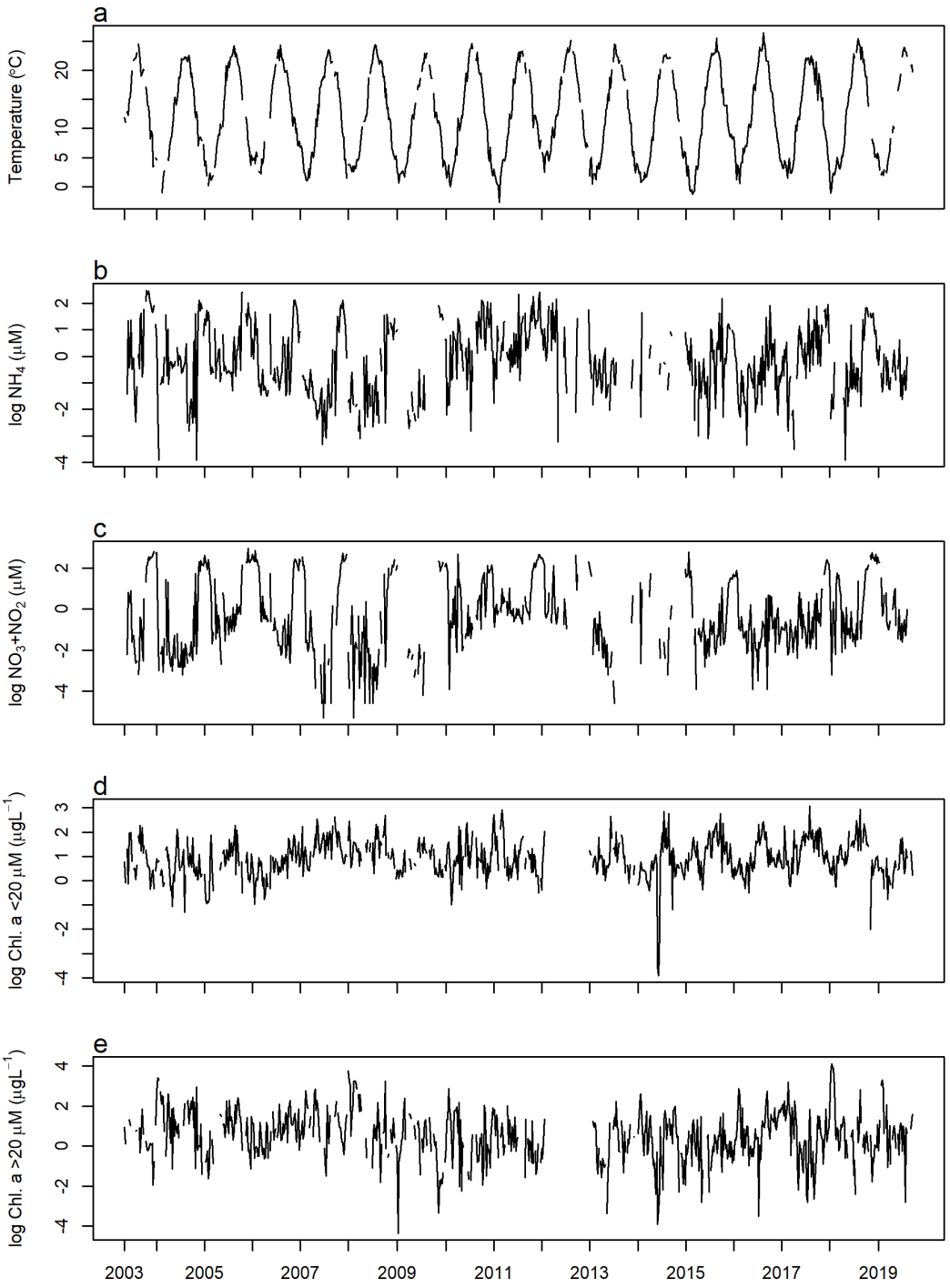

Figure 2. Temperature and nutrient records from the URI Phytoplankton Time-Series. From top to bottom: **a.** surface water temperature (°C), **b.** log NH_4^+(µg L^{-1}), **c.** log NO_3^-+NO_2^- (µg L^{-1}), **d.** log Chlorophyll a > 20 µm (µg L^{-1}), **e.** log chlorophyll a < 20 µm (µg L^{-1}).

It is expected that the innovations of the chlorophyll and nitrogen series are approximately normal after log transformation, and that the associations, between potential predictors and responses are reasonably log-linear for temperature-chlorophyll (Eppley 1972) and reasonably linear for the log chlorophyll, log nutrient associations. Therefore, the chlorophyll and nitrogen series were first log-transformed.

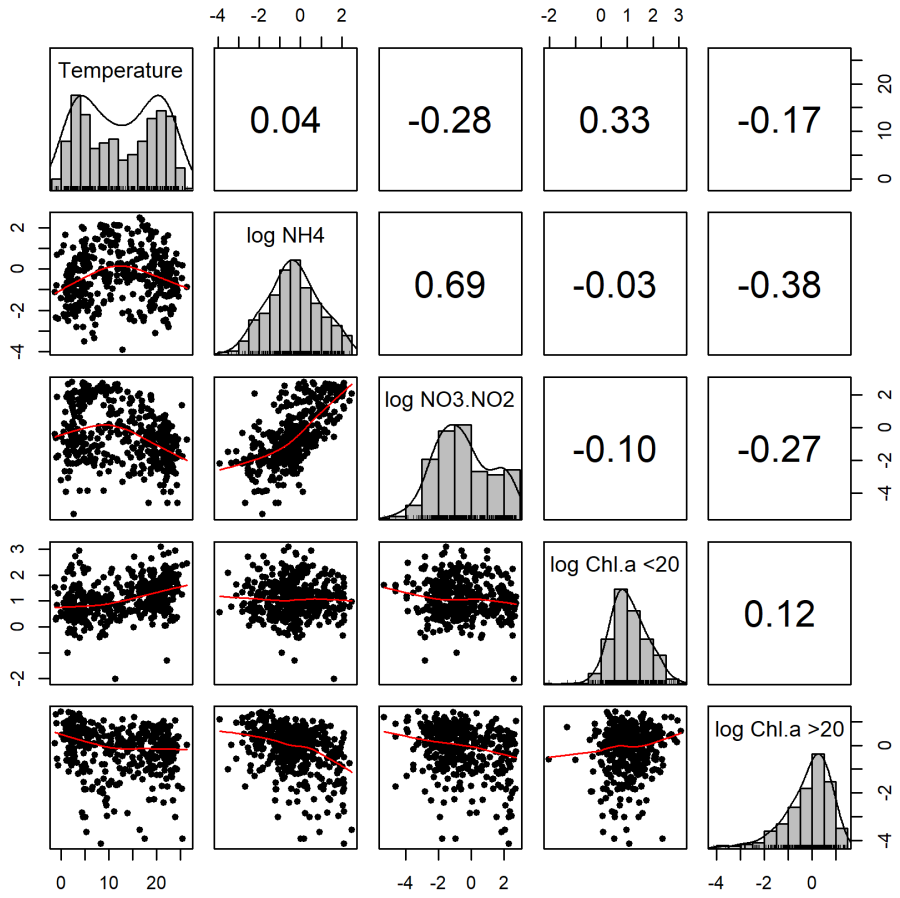

Figure 3. Histograms, Pearson Correlation, and pairwise scatterplots among the primary variables of interest in the model.

Table 1. Missingness Lengths in the Chl. a series

	Missingness Length	Frequency
Chl. a <20 µm	1	61
	2	6
	4	1
	48	1
Chl. a >20 µm	1	61
	2	5
	4	1
	48	1

While there are 573 (67%) complete cases of all data series, the remainder are of various missing data patterns with 38 (4%) time points without observation of any of the data series. Among the chlorophyll series, the duration of missing periods ranges from 1 to 48 observations. Notably, the missing data at 48 consecutive observations represents nearly one full year of missing data (table 1, fig. 4). It is expected that mostly, the missing data mechanisms is missing completely at random (MCAR) meaning missingness is not dependent on the missing value. Therefore, missingness in ignorable, meaning it does not bias inference.

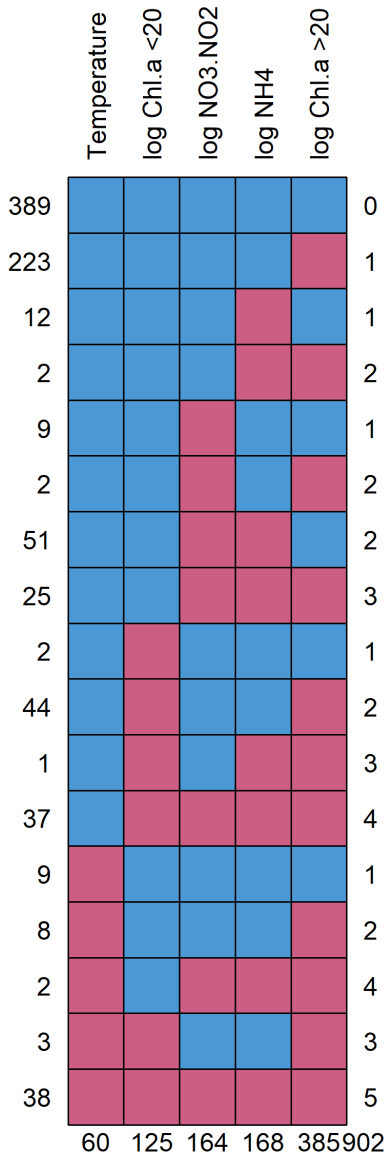

Figure 4. Missingness patterns in the data. Blue squares represent present in the temperature, Chl. a < 20µm, Chl. a >20µm, NO_3^- +NO_2^-, and NH_4^+ series. Maroon squares represent missing data. Counts on the bottom axis are the number of total missing values for the variable in that column specified at the top. Counts on the left side are the number of times that missing data pattern occurs in the series. Counts on the right side are the number of missing variables in that missingness pattern.

CHAPTER 2

METHODS

2.1 Dynamic Linear Models

The standard dynamic linear model (DLM) can be represented as two linear equations, the first of which (eqn. 1) represents observations (y_t) from a true unobserved state ($\boldsymbol{\theta}_t$), transformed by the observation matrix ($\boldsymbol{F}_t$) with error (v_t). The second equation (eqn. 2) models the evolution of the latent state $\boldsymbol{\theta}_t$ through time, according to the evolutional matrix ($\boldsymbol{G}_t$) with evolutional covariance ($\boldsymbol{W}_t$). The DLM is defined at time *t* by the set $\{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{\theta}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$ In the DLM, it is assumed that both evolutional and observational errors are normally distributed. This general form will carry through all the models shown here, with modifications to these parameters, dimensions, and the algorithms for their estimation.

$$y_t = \boldsymbol{F}_t\boldsymbol{\theta}_t + v_t, \qquad v_t \sim N(0, \boldsymbol{V}) \qquad (1)$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + w_t, \quad w_t \sim N(0, \boldsymbol{W}) \qquad (2)$$

$$\boldsymbol{\theta}_0 | D_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0)$$

$$\boldsymbol{V}_0 \sim IG(a_v, b_v)$$

$$\boldsymbol{W}_0 \sim IW(a_w, b_w)$$

For specifications of static $\boldsymbol{V}$, and $\boldsymbol{W}$, consistent with the static Bayesian inference for the Gaussian distribution, the inverse-gamma (*IG*) can serve as a semi-conjugate prior in the univariate case for variance, and the inverse-Wishart (*IW*) is semi-conjugate in the multivariate-Gaussian case. Alternately, in cases where model evolution or observational error are time varying and dynamic $\boldsymbol{V}_t$, and $\boldsymbol{W}_t$ must be

specified, discount factors can be used, which model the loss of information between time steps, whereby low discount factor levels correspond to more information lost per step ahead, and higher discount factors represent greater predictability between time-steps (eqn. 3). While an explicit state-space model could be specified, this can be disadvantageous in terms of both complexities, and due to non-conjugacy in inference.

$$\boldsymbol{W}_t = \frac{1-\delta_i}{\delta_i}\boldsymbol{P}_{i,t}, \qquad (3)$$

$$\boldsymbol{P}_{i,t} = \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t'$$

And when missingness exceeded length 1, that is forecast needed to be greater than 1, the method of 'practical discounting' was used as proposed by Harrison and West 1997 (eqn. 4).

$$if\ k > 1, \qquad \boldsymbol{R}_t = \boldsymbol{G}^{k-1}\boldsymbol{C}_{t+1}\boldsymbol{G}'^{\,k-1} \quad (4)$$

Because of the Markovian structure, solving for the posterior distribution of the latent states $f(\boldsymbol{\theta}_t|\boldsymbol{D}_T \cdot)$, where T=1,2,…,n, is a three step process conditional on the unknown variance and covariance parameters. In the case of the DLM, in the Bayesian framework this an iterative algorithm of forecasting, filtering, and smoothing, which has been derived in the Kalman Filter and Kalman Smoother (Kalman 1960).

One-step -ahead predictive distribution of the latent state, $f(\boldsymbol{\theta}_t|y_{1:t-1}) = N(\boldsymbol{a}_t, \boldsymbol{R}_t)$, where:

$$\boldsymbol{a}_t = E(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{G}_t\boldsymbol{m}_{t-1},$$
$$\boldsymbol{R}_t = Var((\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{W}_t$$

One-step -ahead predictive distribution of the observation, $f(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}) = N(\boldsymbol{f}_t, \boldsymbol{Q}_t)$, where:

$$\boldsymbol{f_t} = E(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{F}_t\boldsymbol{a}_t$$
$$\boldsymbol{Q_t} = Var(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}) = \boldsymbol{F}_t\boldsymbol{R}_t\boldsymbol{F}_t' + \boldsymbol{V}_t$$

The filtered distribution of the latent state, $f(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) = N(\boldsymbol{m}_t, \boldsymbol{C}_t)$, where:

$$\boldsymbol{m}_t = E(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) = \boldsymbol{a}_t + \boldsymbol{R}_t\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1}e_t,$$
$$\boldsymbol{C}_t = Var(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}) = \boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{F}_t'\boldsymbol{Q}_t'\boldsymbol{F}_t\boldsymbol{R}_t,$$
$$e_t = y_t - f_t$$

The smoothed distribution of the latent state, $f(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:T}) = N(s_t, \boldsymbol{S}_t)$, where:

$$s_t = E(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:T}) = m_t + \boldsymbol{C}_t\boldsymbol{G}_{t+1}'\boldsymbol{R}_{t+1}'(s_{t+1} - a_{t+1}),$$
$$\boldsymbol{S}_t = \boldsymbol{C}_t - \boldsymbol{C}_t\boldsymbol{G}_{t+1}'\boldsymbol{R}_{t+1}^{-1}(\boldsymbol{R}_{t+1})\boldsymbol{R}_{t+1}^{-1}\boldsymbol{G}_{t+1}\boldsymbol{C}_t$$

Multivariate and matrix-variate extensions have been developed and detailed by Wang and West (2009). The Kalman Filter and Smoother are used to sample the latent state conditional on the observed data and other model parameters in the forward-filter backward-sampling (FFBS) algorithm. The Kalman Filter can be derived both by Bayes theorem, and standard normal theory. From the Bayesian perspective, the derivation comes about from the one step ahead forecast which serves as a prior for filtering the next time interval.

Amongst all series, it is expected that after transformation the assumption of Gaussian distributed errors, and linear associations are generally appropriate, thus conforming to the distributional and structural assumptions of the DLM case of state space models. Therefore, it is anticipated that the FFBS will be conserved across the model variations included in this proposal. Notably, both the filter and smoother are conditionally conjugate on other parameters such as $\boldsymbol{V}$, and $\boldsymbol{W}$ or within $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ parameters.

The discount factors themselves can be sampled from a discrete distribution conditional on $\boldsymbol{\theta}_t, \boldsymbol{F}_t, \boldsymbol{G}_t$, with probabilities proportional to the log likelihood (e.g. Rodriguez and Puggioni 2010). This is equivalent to Bayesian model averaging where each sampled model has a fixed discount factor. In the DLM literature, this may also be referred to as a type 1 multi-process model, where the model structure is sampled

for all timepoints. However, it is also common practice to fix discount factors and compare models under different sets of fixed factors (Ameen and Harrison 1984). The fixed factors are commonly chosen with mean one-step-ahead forecast error (Ameen and Harrison 1984, Augilar West 2000). Depending the case, either the log likelihoods for discrete factors in the sampling model or full models themselves may be parallelized, respectively, to expedite computation.

While the usage of discrete sampling for the discount factor means only one model needs to be run to select discount factors, the calculation of the log likelihood for the discrete sampling distribution requires that filtering be calculated at all time-points for each discrete discount factor value. As mentioned, this can be done in parallel, but so too can several independent models with fixed discount factors. Sampling discount factors also adds the additional complexity for additional issues in MCMC sampling, and potentially contribute to highly autocorrelated samples and thus low effective sample sizes from the posterior distributions that can come with highly parameterized models (Gelman et al. 2014). Both discount factors and covariance matrices may be specified uniquely for independent blocks or components of the model (West and Harrison 1997).

2.2 Model Structures

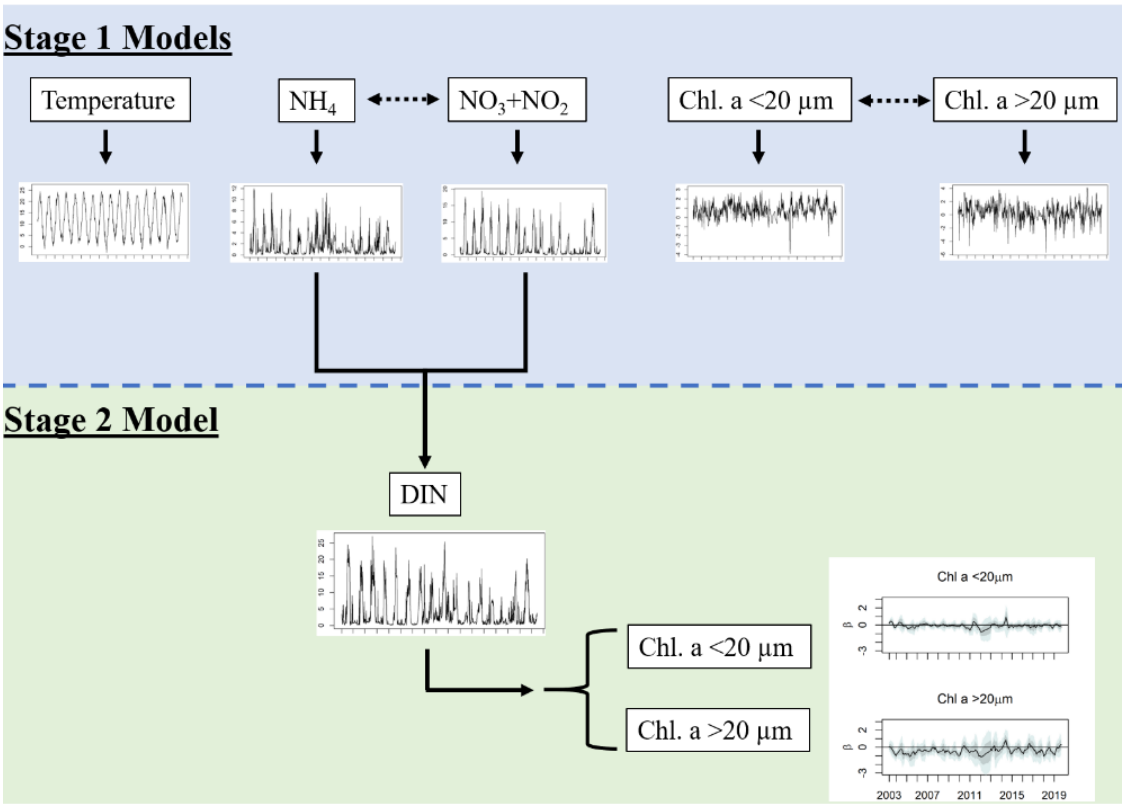

Figure 5. Analysis workflow for the Narragansett Bay Time Series analysis. First stage models are intended for imputation of missing data, exploration, and to characterize both seasonal change and long-term trends in the series. The second stage model is specifically designed to test how DIN, which was legally regulated for wastewater treatment facilities, may be tied to the phytoplankton size structure of Narragansett Bay.

In this project, the analysis of the NB data is structured in stages to investigate the potential associations and changing associations between chlorophyll size fractions and DIN and temperature (fig. 5). With this objective, a two-stage model is proposed. The first stage is a bivariate DLM for NH_4^+ and NO_3^-+NO_2^-, and independent

univariate DLMs for temperature, chl. a < 20 µm, and chl. a >20 µm. This stage has a scientific purpose to provide a descriptive measure of long-term and potentially seasonal change in each of these environmental features and explore associations of complete data series. As described above, this alone is novel for the NB time series. Statistically this first model stage also serves for data imputation, which is necessary for the second regression stage of the model which requires a complete series of the regressors, which have missingness (fig. 2, 4). The second component of the total model, is a dynamic linear regression of chlorophyll on DIN (NH_4 + (NO_3+NO_2)). This formulation is designed to test the hypothesized associations between the series and describe data patterns. The second component must sample from the posterior distributions of the smoothed latent states (i.e. $f(\boldsymbol{F}_t\boldsymbol{\theta}_t|\boldsymbol{y}_{1:T})$) of the first stage models. Because the posterior latent state is designed to represent the true level of the regressors without observational error, it was decided that this would represent a more accurate predictor for the chl. a series. The specification of these two models is as described below.

Stage 1 models

In stage 1, four total models were run, univariate models for temperature, log chl. a <20µm, log chl. a >20µm, and another bivariate for log DIN. Both these models have two major components to the latent state θ_t. The first is a dynamic intercept (μ_t) with corresponding components in $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ of 1 (eqns. 5, 6). Thereby, $f(\mu_t|\mu_{t-1}, \boldsymbol{W}_\mu) \sim N(\mu_{t-1}, \boldsymbol{W}_\mu)$, and so ($\mu_t$) has the flexible structure of a random walk process.

In all cases where seasonal components are included, they are represented by harmonics in Fourier form, for a parsimonious representation of annual cyclicity (eqns. 6,7). At weekly resolution, the data has long periodicity (52.17, the number of weeks in a year). While any function of period s can be modeled by s/2 harmonics, in general a smaller number is both more practical and effective. Depending on the complexity of the seasonal signal, j different harmonics are used beginning with that of the longest possible period, *s*. For the temperature series, *j*=1, for the DIN and chl. a series, *j*=(1,…,5) to accommodate a more complex seasonal cycle. Within $\boldsymbol{\theta}_t$, for each frequency, both the harmonic $S_{j,t}$ and its conjugate $S_{j,t}$* are included (eqn. 8), and evolve according to the subcomponent H of the evolutional matrix G (eqn. 6).

$$F = [11010 \ldots 01] \tag{5}$$

$$G = \begin{bmatrix} 1 & \\ & G_S \end{bmatrix}, \qquad G_S = \begin{bmatrix} H_j & & \\ & \ddots & \\ & & H_{\frac{S}{2}} \end{bmatrix} \tag{6}$$

$$H_j = \begin{bmatrix} cos(\omega_j) & sin(\omega_j) \\ -sin(\omega_j) & cos(\omega_j) \end{bmatrix}, \qquad \omega_j = \frac{2\pi j}{s} \tag{7}$$

$$\theta_{t,X} = \begin{bmatrix} \mu \\ S_{1,t} \\ S^*_{1,t} \\ \vdots \\ S_{\frac{S}{2},t} \\ S^*_{\frac{S}{2},t} \end{bmatrix} \tag{8}$$

For temperature, $\boldsymbol{\theta}_{t,x}$ was composed of μ_t and one dynamic season component with period s to describe the seasonal cyclicity and its interannual changes. Considering consistent observational uncertainty, error $\boldsymbol{V}$ was modeled as time

invariant and given a conjugate prior with low informativity, $IG(0.1, 0.1)$. Under the expectation that temperature patterns should be slow to evolve in time and that long-term and seasonal patterns should not obviously be related, a time invariant evolutional covariance matrix $\boldsymbol{W}$ was specified as diagonal, with each diagonal variance element given a conjugate inverse-gamma prior with low informativity $IG(0.1, 0.1)$. Because $\boldsymbol{\theta}_{t,x}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ are conditionally conjugate, Gibbs sampling from the FFBS, IG distribution of $\mathbf{V}$, and IG distributions of the diagonal elements in $\boldsymbol{W}$ lead to a sample of the full joint posterior.

For DIN, which is equal to the sum of the series of NH_4^+ and NO_3^-+NO_2^-, a bivariate model was specified so that each of its components is modeled with correlations permitted between series components. This bivariate structure allows the correlation between series to be leveraged in the FFBS, so that if one series is missing data where the other is not, the correlation between the series provides additional information. Both the data series are first log-transformed. $\boldsymbol{\theta}_{t,x}$ is composed of a μ_t component for each series and multiple dynamic season components for each series with periods $s, \frac{s-1}{2}$, ... to describe the seasonal cyclicity and its interannual changes. Considering consistent observational uncertainty, error $\boldsymbol{V}$ was modeled as time invariant and given a prior with low informativity, $IW(0.1, 0.1*\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the identify matrix with rank p. Because $\boldsymbol{\theta}_t$,$\boldsymbol{V}$, and components are conditionally conjugate, Gibbs sampling from the FFBS, IG distribution of $\boldsymbol{V}$, and IW distribution of $\boldsymbol{W}$ leads to a sample of the full joint posterior.

Table 2. Models run in stage 1 and stage 2 of this thesis, with general specifications.

Stage	Model	Response Variable(s)	Components	*V* Specification	*W* Specification
1	1	Temperature	Dynamic, Intercept (μ) Dynamic, Season (S_i), i=1	Static, *IG* prior	Static, *IW* prior
	2	NH_4 , NO_3 + NO_2	Dynamic, Intercept (μ) Dynamic, Season (S_i), i=1,...,6	Static, *IG* prior	Static, *IW* prior
	3	Chl. <20, Chl. >20	Dynamic, Intercept (μ) Dynamic, Season (S_i), i=1,...,6	Static, *IG* prior	Static, *IW* prior
2	4	Chl. <20, Chl. >20	Dynamic, Intercept (μ) Dynamic, Regression on DIN Static, Season (S_i), i=1	Static, *IW* prior	Dynamic, Fixed Discount Factor

Considering this thesis has specific hypotheses about temperature and nitrogen changes, the cross-correlation of the pre-whitened series was calculated to evaluate the magnitude, significance, and potentially meaningful cross-correlations, after accounting for the temporal structure of the data. In pre-whitening, an ARIMA (autoregressive integrated moving average) class is fit to one series so that only unstructured residuals are left. The same model is used the filter the second series. This is designed to reduce spurious correlations between the series in the CCF. Also, as an exploratory measure, before modeling the dependence structure in the second stage of the model, a wavelet analysis was performed on the univariate series to investigate patterns of cyclicity. Further, cohesion calculated between the series to visualize potential changes in seasonal structure, lagged relationships across timescales (phase differences through phase angle) and coherence between the series, coherence being analogous to cross correlation as a function of frequency. Wavelet analysis was performed using the *waveletcomp* package in R, which uses the Mortlet wavelet transform of the time series (Roesch and Schmidbauer 2018, eqn. 9).

$$\psi(t) = \pi^{-\frac{1}{4}} e^{6it} e^{-\frac{t^2}{2}} \quad (9)$$

$$Wave(\tau, s) = \Sigma_t x_t \frac{1}{\sqrt{s}} \psi^* \frac{t-\tau}{s}$$

$$Wave_{x,y} = \frac{1}{s} Wave_x(\tau, s) * Wave_y^*(\tau, s)$$

$$Power(t, s) = \frac{1}{s} |Wave(\tau, s)|^2$$

$$Coherence = \frac{\left|Wave_{x,y}\right|^2}{Power_x * Power_y}$$

Stage 2 Models

After fitting potential predictive series, exploring their structure and evaluating cross-correlative features between series, I examined the association between DIN and the chlorophyll series in a multivariate observational model (eqn. 10) with a matrix-variate state (eqn. 11). The advantage of allowing a matrix-variate state is in the ability to model correlated evolution among the series as well as the state variables.

$$\boldsymbol{Y}_t = \boldsymbol{F}_t \boldsymbol{\theta}_t + v_t, \qquad v_t \sim N(0, \boldsymbol{V}) \quad (10)$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + w_t, \quad w_t \sim N(0, \boldsymbol{W}_t \otimes \boldsymbol{V}) \quad (11)$$

$$\boldsymbol{\theta}_0 | D_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0)$$

$$\boldsymbol{V}_0 \sim IW(a_v, b_v)$$

$$\boldsymbol{W}_t = \frac{1-\delta_i}{\delta_i} \boldsymbol{P}_{i,t},$$

$$\boldsymbol{P}_{i,t} = \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t'$$

Both series have three major components to the latent state θ_t. The first is a dynamic intercept (μ_t) with corresponding components in $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ of 1 (eqns. 12, 13). Thereby, $f(\mu_t|\mu_{t-1}, W_\mu) \sim N(\mu_{t-1}, W_\mu)$, and so ($\mu_t$) has the flexible structure of a random walk process. The second component is a regression coefficient (β_t) on the appropriate lag k of the predictor, in this case, lag 0 of DIN. The corresponding multiplier in $\boldsymbol{F}_t$ is the predictor X_{t-k}, and the corresponding component in $\boldsymbol{G}_t$ is 1. Thereby, β_t can also evolve with the flexibility of a random walk process. The final component of the model is a Fourier form seasonal component, with *j*=1 (eqns. 13, 14). While the predictor is seasonal, and the dynamic intercept has the capacity to adapt to seasonal variability, including a static season component helps ensure that the dynamic intercept adapts to what is truly the long-term pattern and not necessarily residual seasonal patterns. By this structure, the interpretation of the regressive component, is not in describing the total variability attributable to the regressor, but rather, the anomaly from the long-term trend and regular seasonality.

$$\boldsymbol{F}_t = [1, X_{t-k}, 1, 0] \tag{12}$$

$$\boldsymbol{G} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & \boldsymbol{G}_s \end{bmatrix}, \qquad \boldsymbol{G}_S = \boldsymbol{H}_j \tag{13}$$

$$\boldsymbol{H_j} = \begin{bmatrix} cos(\omega_j) & sin(\omega_j) \\ -sin(\omega_j) & cos(\omega_j) \end{bmatrix}, \qquad \omega_j = \frac{2\pi j}{s} \tag{14}$$

$$\boldsymbol{\theta}_{t,X} = \begin{bmatrix} \mu_{t,1} & \mu_{t,2} \\ \beta_{t,2} & \beta_{t,2} \\ S_{1,t,1} & S_{1,t,2} \\ S^*_{1,\,t,1} & S^*_{1,\,t,2} \end{bmatrix} \tag{15}$$

Notably, there are other specifications of possible to study the influence of the regressor besides on the overall mean levels. For example, apart from the mean levels of the series, the seasonal impacts could be investigated directly by multiplying the regressor by a seasonal component. The impact of the regressor could also be studied on the covariance matrices, by specifying another level of the model (in itself a state-space model) for the covariance. While there would be no conjugate or semi-conjugate model for the posteriors of a model for covariance, alternative inferential algorithms like particle filtering make this possible.

The observational covariance $\boldsymbol{V}$ is conditionally conjugate and static, with an $IW(10, 10 * \sigma_{Y_1,Y_2})$ prior.

For $\boldsymbol{W}$, both static and dynamic specifications were tested. For the static case, a conditionally conjugate $IW(p, \boldsymbol{I}_p)$ prior was utilized. For the case of discount factors, for $\boldsymbol{W}$, discrete discount factors (0.8, 0.85, 0.9, 0.95, 0.99, 0.999) were tested in parallel model runs. Because $\boldsymbol{\theta}_{t,x}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ are conditionally conjugate, Gibbs sampling from the BS, posterior conditional IW distribution of $\boldsymbol{V}$, and posterior conditional IW distributions of $\boldsymbol{W}$ led to a sample of the full joint posterior.

2.3 Posterior Computational Methods

Posterior samples of unknown variance, covariance, and latent states were all iteratively sampled via Markov Chain Monte Carlo simulations. As many models were being run, the default simulation length was 5,000 iterations with a burn-in period of 2,000. As necessary, for models with high autocorrelation in the MCMC chain, the effective number of simulation draws was calculated as in Gelman et al. (2014), where $W = within\ sequence\ \sigma^2$, and $B = between\ sequence\ \sigma^2$ (eqn. 15).

$$\hat{n}_{eff} = \frac{mn}{1 + 2\Sigma_{t=1}^{\mathrm{T}}\ \hat{\rho}_t}, \qquad (15)$$

$$\hat{\rho}_t = 1 - \frac{V_t}{2\ \widehat{var}^{+}},$$

$$V_t = \frac{1}{m(n-t)}\Sigma_{j=1}^{m}\Sigma_{i=t+1}^{n}\left(\psi_{i,j}, -\psi_{i-t,j}\right)^2,$$

$$\widehat{var}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B,$$

The minimum effective number of samples was 1000. At a maximum this required 10,000 MCMC iterations, which was within capacity of computer memory so thinning was not required. After examining trace, density, and acf plots of posterior samples, as necessary, simulations were split into 5 chains of equal length to calculate effective sample size.

Considering temperature and nitrogen are expected to be inversely and positively correlated with cell size respectively, I analyzed the cross-correlation of the pre-whitened series to evaluate the magnitude, significance, and potentially meaningful cross-correlations, after accounting for the temporal structure of the data. Also, as an exploratory measure, before modeling the dependence structure in the second stage of the model, a wavelet cohesion analysis was performed in a pairwise manner between the series to visualize potential changes in seasonal structure and coherence between the series.

For the optimal models at each stage, analysis code are publicly available on Github (https://github.com/JacobPStrock/NBPTS_DynamicLinearModels)

CHAPTER 3

DATA SIMULATION AND MODEL SELECTION FOR δ

3.1 Model Selection with Fixed Discount Factors

In the environmental series included in this study, there are several features which make inference challenging. These features touch remaining uncertainties in model specification and selection for DLMs, despite the abundant usage and long-standing prevalence of these models (Kalman 1960; West and Harrison 1997). One such feature in the case of the series of this study is missingness periods of extended length ($\leq$48 consecutive observations). It is in these cases with extensive missingness, where it stands to question what is the appropriate discounting method, and which criteria accurately select the discount factor in these cases.

Usage of discount factors has primarily focused on one-step ahead forecasting $f(\theta_t|\boldsymbol{y}_{1:t-1})$, which is—as aforementioned—critical for calculating the posterior distribution of the state at a given time-point, imputation for missing time-points, and forecasting to future time-points (Harrison 1967). Practically, there is no standard for choice of discount factors for fixed discount models, which can be necessary due to mixing rates in adaption-rejection sampling algorithms for discrete distributions of discount factors. Typically, a small number of discount factors (>5) is supplied due to computational costs, and one-step ahead forecasts may be evaluated (Ameen and Harrison 1984; Aguilar and West 2000). To my knowledge, most of the foundational work in discount factors has been exemplified with complete or mostly complete series, and in cases where high resolution, or predictability of the data itself meant that

some high (>0.95) discount factor could be applied with constrained effects on the final model fit. Despite the relatively popular usage of discount factor models, to my knowledge, selection of discount factors in less predictable series (corresponding to low discount factors), and those with potentially long durations of missingness has not been evaluated. The uncertainty of model selection in these series may be a particularly important issue because uncertainty in the selection of potentially lower discount factors (<0.9) could have major effects on model fit and interpretation. This issue is compounded in the missing data periods, where, by strict definition of the discounting methods, the uncertainty in the model increases exponentially, until the next observation is made (eqn. 16). Especially if a low discount factor is needed to model the data, this results in unrealistically large credible intervals during missing data periods. Consider the explicit, k-steps ahead updating of $\boldsymbol{W}_t$ below (eqn. 16).

$$\boldsymbol{R}_t(k) = \frac{\boldsymbol{G}^k \boldsymbol{C}_t \boldsymbol{G}'^k}{\delta^k} \quad (16)$$

It has been suggested that in missing data periods greater than length 1, discounting should be halted so that in forecasting through longer missing data periods, the uncertainty should only grow linearly as in equation 17 below (West and Harrison 1998).

$$if\ k > 1, \qquad \boldsymbol{R}_t = \boldsymbol{G}^{k-1} \boldsymbol{C}_{t+1} \boldsymbol{G}'^{k-1} \qquad (17)$$

This method termed ‘practical discounting’ indeed seems more realistic than standard discounting for periods of greater missingness, but to my knowledge, model selection, and accuracy of discount factor selection have not been formally evaluated for this case. In this chapter, I examine standard and practical discounting in instances

in simulated series with large, artificial data gaps and consider multiple metrics for accurate recovery of discount factors.

3.2 Data Generation and Comparisons

To evaluate the accuracy of discount factor selection, particularly in noisy data with prolonged data gaps presented here, data was simulated from the posterior of model fits to real data with fixed discount factors. Specifically, the bivariate regression model of the small and large size fraction of chlorophyll series, with a single fixed seasonal component, static $\boldsymbol{V}$, and component discount specified $\boldsymbol{W}_t$ was used. The posterior mean of the seasonal component, $\boldsymbol{V}$, and $\boldsymbol{W}_t$ were used to generate new bivariate data series from models with a lower (0.95, 0.99), and higher (0.999, 0.99) discount factor set.

Data were generated by initializing the state at 0 and sequentially sampling the state equation (eqn. 18), so that in this simulation case, the true state was observed. Because these simulated data were based on model fits of the real data, the length of the simulated series was equal to the length of the original series. Missingness was also generated randomly in the new series, following the frequency of missing data patterns in the original data (table 1). That is the location of missingness did not mirror the original data, but the new data contained the same number of missing observations and with the same frequency of data gaps.

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, \boldsymbol{V} \otimes \boldsymbol{W}_t) \tag{18}$$

In order to test the recovery of the discount factors, the simulated data were fit with three variants of the data generation model. The first case was with standard

discounting formulation and no missing data. The second case was with standard discounting formulation, including in missing data gaps. The third case was with 'practical discounting' formulation, whereby with data gaps greater than length 1, the discount was not repeatedly replied at each timestep as in (eqn. 17). Rather, $\boldsymbol{R}_t$ was updated as in (eqn. 18, Harrison and West 1998), so that uncertainty increased linearly in the data gaps.

3.3 Model Selection and Criteria

To inform model selection, several existing indices were first considered: deviance information criterion (DIC, eqn. 19), Watanabe-Aikake Information Criterion (WAIC, eqn. 20), and root mean square forecast error (RMSFE). DIC is importantly based on the deviance, or the log predictive density of the data given the mean point estimate of the model (Spiegelhalter et al. 2002). Because of the missing data in these data a variant of DIC, DIC_4 (Celeux et al. 2006) is used which takes expectations over the missing data, in this case, our regressor X_t. $Y_{missing}$ does not impact likelihood calculations because $Y_{observed}$ is conditionally independent of $Y_{missing}$ given θ_t. As compared to DIC, WAIC uses pointwise estimates of log predictive density, expectation is found, and integrated over the posterior predictive density (Watanabe 2010). For WAIC as well, we take expectations over the missing data X_t.

$$DIC = -E_X[\log(p(y|E_\theta[\theta|y,X],X)p(X))] + 2p_{DIC} \quad (19)$$

$$p_{DIC} = -2E_{X,\theta}[\log(p(y|\theta,X)p(X))] + E_X[\log(p(y|E_\theta[\theta|y,X],X)p(X))]$$

$$WAIC = -2(llpd - p_{WAIC}) \quad (20)$$

$$llpd = 2\Sigma_{i=1}^{n} E_{X,\theta}[\log(p(y_i|\boldsymbol{\theta},X)p(X))]$$

$$p_{WAIC} = 2\Sigma_{i=1}^{n} \log\big(E_{X,\theta}[p(y_i|\boldsymbol{\theta},X)p(X)]\big) - E_{X,\theta}[\log(p(y_i|\boldsymbol{\theta},X)p(X))]$$

RMSFE was calculated for each MCMC iteration from the one step ahead mean forecast (eqn. 21).

$$RMSFE = \sqrt[2]{\frac{\Sigma_{t=1}^{n}\left(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i}\right)^2}{n}} \quad (21)$$

An alternative RMSFE was calculated for each timepoint, using all MCMC samples (eqn. 22), where r is the number of MCMC samples.

$$RMSFE_{t,2} = \sqrt[2]{\frac{\Sigma_{i=1}^{r}\left(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i}\right)^2}{r}} \quad (22)$$

In the artificially induced missing data periods, the root mean square error (RMSE, eqn. 23) was calculated, integrating over the full posterior distribution, as compared to the RMFSE, which in addition to using the one step ahead forecast, used the posterior mean of this forecast.

$$RMSE = \sqrt[2]{\frac{\Sigma_{t=1}^{n}\left(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i}\right)^2}{n}} \quad (23)$$

An alternative RMSE was calculated for each timepoint, using all MCMC samples (eqn. 24).

$$RMSE_{t,2} = \sqrt[2]{\frac{\Sigma_{i=1}^{\mathrm{r}}\left(y_t - \boldsymbol{F}_{t-1}\boldsymbol{\theta}_{t-1,i}\right)^2}{r}} \quad (24)$$

3.3 Simulation Comparisons

$\boldsymbol{\delta_{\mu}}$ = 0.999, $\boldsymbol{\delta_{\beta}}$ =0.99 data generation model

For the data series, where the data generation model used discount factors $\delta_{\mu} = 0.999$, $\delta_{\beta} = 0.99$ and standard discounting, the exact discount factor set was not

recovered in any case of although the estimates were close. DIC and WAIC both suggested a model of $\delta_\mu = 0.99$, $\delta_\beta = 0.999$; RMSFE suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.99$; $RMSFE_2$ suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.999$; RMSE suggested $\delta_\mu = 0.999$, $\delta_\beta = 0.999$; $RMSE_2$ suggested $\delta_\mu = 0.999$, $\delta_\beta = 0.999$ (appx. tables 1-5). Notably, RMSFE and RMSE had higher power than $RMSFE_2$ and $RMSE_2$ respectively (appx. tables 1-5, appx. fig. 1), still many models did not produce significantly different RMSFE or RMSE. Further, the RMSE measures in this case with missingness and without the use of practical discounting tended to select more static models as opposed to true model and that selected by the DIC, WAIC, and RMSFE measures.

For the data series, where the data generation model used discount factors $\delta_\mu = 0.999$, $\delta_\beta = 0.99$ and practical discounting, the exact discount factor set was not recovered in any case of although the estimates were close. DIC and WAIC both suggested a model of $\delta_\mu = 0.99$, $\delta_\beta = 0.999$; RMSFE suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.99$; $RMSFE_2$ suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.99$; RMSE suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.999$; $RMSE_2$ suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.999$ (appx. tables 6-10). Notably, RMSFE and RMSE had higher power than $RMSFE_2$ and $RMSE_2$ respectively (appx. tables 6-10), still many models did not produce significantly different RMSFE or RMSE. Further, the RMSE measures in this case with missingness and with the use of practical discounting tended to select more static models as opposed to true model and that selected by the DIC, WAIC, and RMSFE measures.

$\delta_\mu = 0.95$, $\delta_\beta = 0.99$ data generation model

For the data series, where the data generation model used discount factors $\delta_\mu = 0.95$, $\delta_\beta = 0.99$, and standard discounting during missingness, the exact discount factor set was optimal for some indices. DIC and WAIC both suggested a model of $\delta_\mu = 0.99$, $\delta_\beta = 0.999$; RMSFE suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; $RMSFE_2$ suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; RMSE suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; $RMSE_2$ suggested $\delta_\mu = 0.99$, $\delta_\beta = 0.999$ (appx. tables 11-15). Notably, consistent with the other simulated series, RMSFE and RMSE had higher power than $RMSFE_2$ and $RMSE_2$ respectively (appx. tables 12-15), still many models did not produce significantly different RMSFE or RMSE. Further, the $RMSE_2$, DIC, and WAIC measures in this case with missingness and without the use of practical discounting tended to select more static models as opposed to true model and that selected by the RMSFE measures.

In the same simulated series with practical discounting during missingness, the exact discount factor set was optimal for some indices in the single simulation, and other indices such as DIC and RMSE more strongly favored models closer to the data generation model as compared to the method without practical discounting. DIC and WAIC both suggested a model of $\delta_\mu = 0.95$, $\delta_\beta = 0.999$; RMSFE suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; $RMSFE_2$ suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; RMSE suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$; $RMSE_2$ suggested $\delta_\mu = 0.95$, $\delta_\beta = 0.99$ (appx. tables 16-20, appx. fig. 6). RMSFE and RMSE had higher power than $RMSFE_2$ and $RMSE_2$ respectively (appx. tables 16-20, appx. fig. 6), still many models did not produce significantly different RMSFE or RMSE. The RMSE suggest that practical discounting will

optimize performance in long-periods of missingness. DIC and WAIC also supports that practical discounting improves the model fit within sample.

Last considering that DIC and WAIC did not exactly pick the correct model with missingness, in either case of practical discounting or non-practical discounting, to test the possible usage of DIC and WAIC in complete or nearly complete series, the model was fit with the complete data series (table 21). In this case, the DIC and WAIC did not pick the exact data generation model, however, similar to the practical discounting model, the selection was for a model near the data generation model.

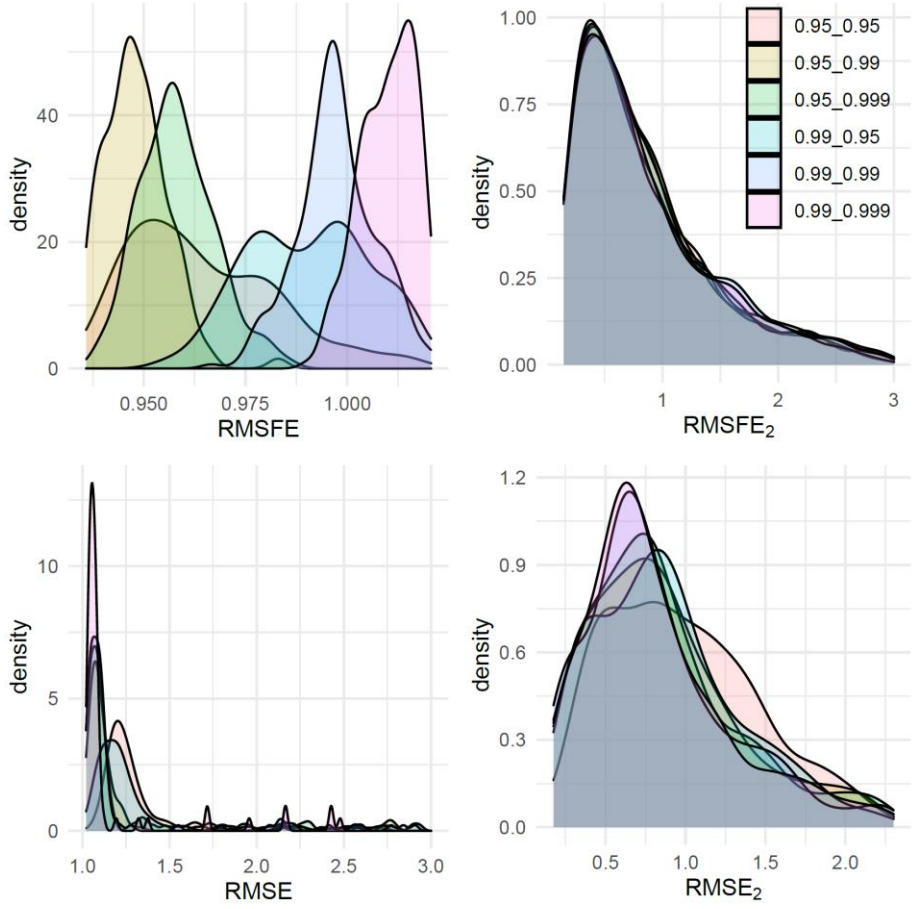

Figure 6. Posterior distributions for **a.** RMSFE **b.** $RMSFE_2$ **c.** RMSE **d.** $RMSE_2$ for each model with fixed discount factors for μ and β (fill color), fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$.

In terms of model selection RMSFE is the most accurate and highest power metric to use for model comparison (fig. 6). DIC and WAIC did not match the data generation model in any circumstance. $RMSFE_2$ and $RMSE_2$ both had lower power than RMSFE and RMSE respectively. This suggests that variability in MCMC iterations were more clearly pronounced than variability at given timepoints. RMSE tended to penalize the high uncertainty produced during discounting with large periods of missingness. This was apparent in all simulations. The RMSFE was optimal for the data generation model in the low discount factor data generation case, and selected slightly lower discount factors (0.99, 0.99) as compared to the data generation model in the case of the high discount factor generated data.

Concerning practical discounting and model selection, practical discounting did not impact the model selection according to RMSFE. However, RMSE, which is calculated during periods of long-missingness, identified the data generation model when practical discounting was used, but not when standard discounting methods were used. Together, this suggests that RMSFE in itself may be an optimal index for model selection. Further, this suggests that practical discounting will produce the optimal inferences when missingness is greater in length than 1 missing observation.

3.3 Discussion

Considering WAIC, DIC, and two variants each of RMSE and RMSFE, RSMFE was the most consistent estimator in recovering the discount factors of simulated data. While RMSE during imputed missing data periods might be a good gauge for forecast or imputation accuracy, particularly with standard discounting

methods, it tended toward more static models than the actual data. WAIC and DIC also tended toward more static models than the original simulated data also particularly during standard discounting. RMSFE tended to be closest to the data generation model in cases without missing data as well as those with missing data and practical or standard discounting. Considering the relative consistency of RMSFE, it was ultimately the chosen criterion for model selection in our multivariate dynamic regression. Notably, longer series of simulated data may have aided distinguishing the exact discount factor set of the data generation model. Further, simulating more series could allow calculation of how often RMSFE and other criterion select the correct model, however, both circumstances add significant computation time and cost, and the single simulation provides evidence in our case, that RMSFE will choose close to the 'true' values. It is also important to note, in part that discount factors are a synthetic latent variable, and potentially different levels along with different observational covariance matrices could produce similar performing models.

Although RMSE was a biased metric for model selection, particularly during prolonged periods of missingness, it still had utility in evaluating the performance of practical discounting in data with long period missingness. This is important because although the method of practical discounting has been proposed (West and Harrison 1997), its performance during missing data imputation has not been evaluated in published literature. While RMSFE may be the optimal method for discount factor selection, it does not account for performance during long-period missingness as our metric of RMSE does. Therefore, results of RMSE in comparable models with practical and standard discounting provide an evaluation for this imputation method.

This is important, not only to understand the results of this study and to interpret forecasts during missing periods but also because for time series data with missingness, DLMs are a popular method for imputation and within DLMs, discount factors are a popular method for specifying evolutional covariance. Therefore, it is important to understand for series with long periods of missingness if practical discounting will provide optimal imputation representative of the data generation model. RMSE in cases of practical discounting showed that the set of discount factors coincident with the data generation model was optimal. This contrasts with the standard discounting method, for which the RMSE suggested a higher discount model. This is because in data with missingness length greater than one, under standard discounting uncertainty increases exponentially, a trait which to the reason of West and Harrison is unreasonable. Instead, the linear increase in uncertainty propagated with practical discounting results in a more accurate imputation. Notably, while we used these two discounting functions, theoretically, other discounting functions are possible, with the potential to impact both model selection and missing data imputation.

CHAPTER 4

NARRAGANSETT BAY SERIES RESULTS

4.1 Stage 1 models

Temperature

While marine temperature rise has been recorded and published about Narragansett Bay, at a rate of 0.3°C decade^{-1} (Fulweiler et al. 2015),the DLM fit of the temperature data 2003-2019 shows non-monotonic changes at multi-year scale (fig. 7). The 95% CI for the intercept is 10.70 to 14.02 °C, which has a range 3.32°C, exceeding the total documented change in Narragansett Bay. Therefore, while there was a prior expectation that the 17 years used in this study might demonstrate a slight mean temperature increase of 0.5°C coincident with the linear patterns which have been described, this was not found. Instead, it was revealed that multiyear patterns could drive changes in temperature dependent biogeochemistry.

In addition to long-term but non-monotonic changes in the trend, the seasonal signal shows interannual variation, with minimum winter temperatures of the posterior mean predicted (0.68°C$\pm$1.6), more variable than summer maxima (24.15$\pm$°C1.12, fig. 7). The ratio of the variances (2.03) is F-distributed with a p-value of 0.077, suggesting significance at the α=0.10 level. Though this is a highly predictable series in relation to other environmental data such as the other data series included in this study, there is an observational variance of (0.69$\pm$0.07), which has potential contributions from weather, tidal signal, and riverine input (appx. fig. 2).

Overall, the wavelet analysis supports the interpretation of dominant seasonal cyclicity, with some multiyear variability and finer scale variation (appx. fig. 3).

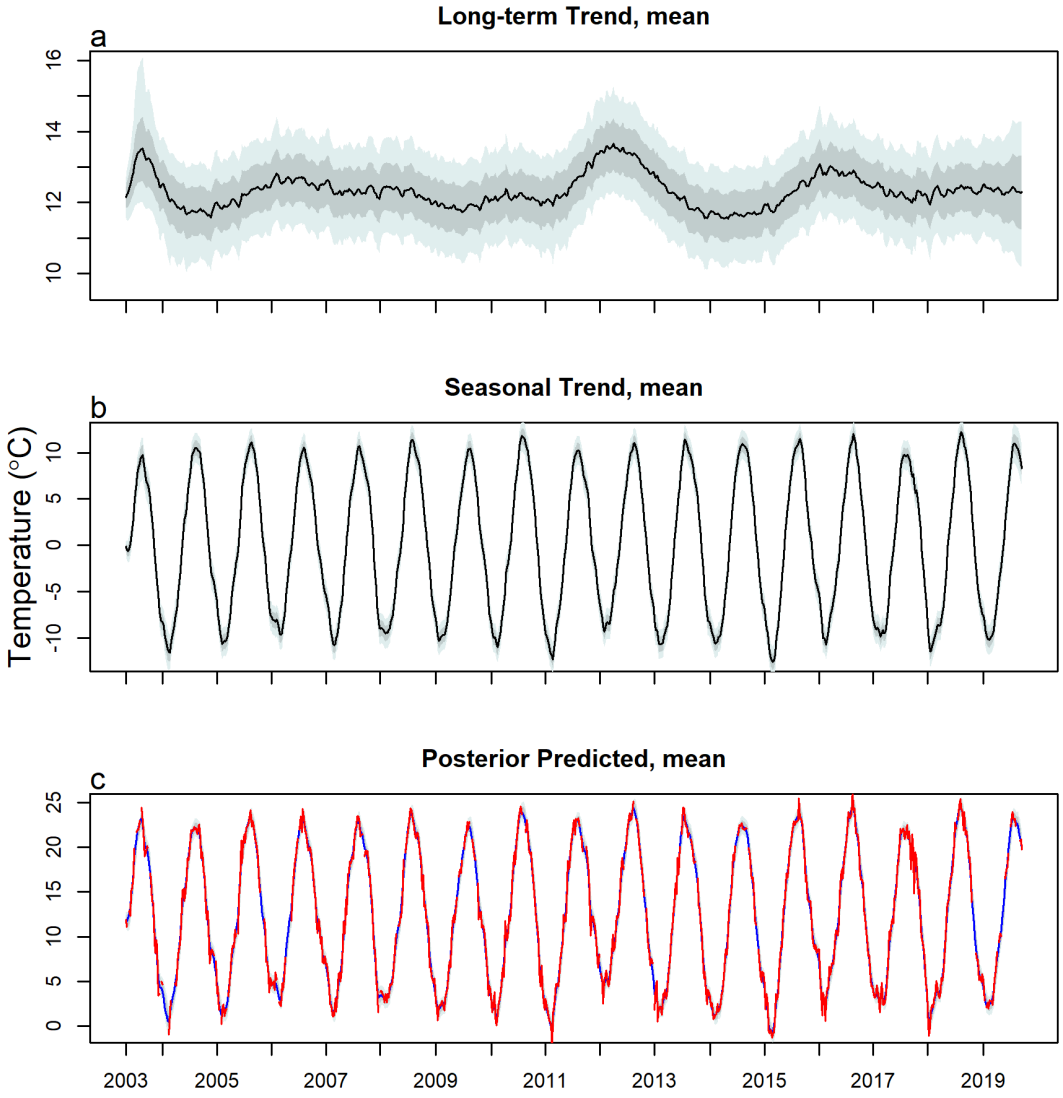

Figure 7. Decomposition of the temperature series DLM (2003-2019), fit with a dynamic intercept and seasonal component with a yearly (period = 52.17 week) seasonal frequency. **a.** the dynamic intercept, **b.** the seasonal trend, **c.** the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

NH_4 & NO_3+NO_2

Much like the temperature series, for NH_4 and NO_3+NO_2 although long term monotonic changes were expected in the dynamic intercept, the actual patterns from 2003-2019 were more complex, with multiyear patterns at odds with the policy mandates for reduction in wastewater levels by 50% between 2005-2012 (fig. 8,9). Further, while the seasonal cycle is quite variable year to year, there are now clear trends in the features of this annual cycle such as levels at the annual maxima (fig. 8, 9). Both series of N species show a high correlation of (0.50 ± 0.037), with overall higher variability in the NO_3+ NO_2 (0.94 ± 0.07), as compared to NH_4 (0.74 ± 0.05, appx. fig. 4).

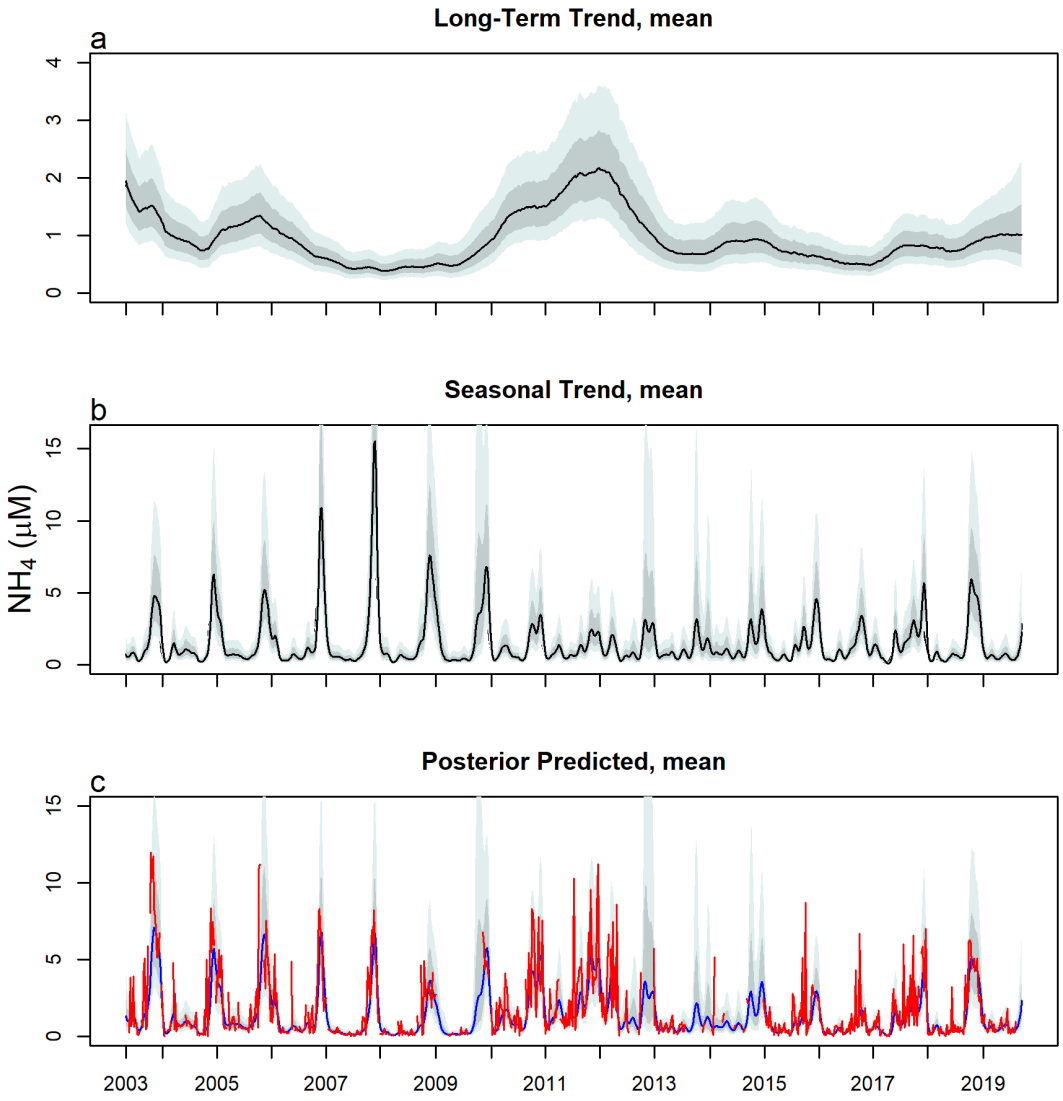

Figure 8. Decomposition of the NH_4 series DLM (2003-2019), fit with a dynamic intercept and 5 seasonal components with periods of (52.17,…,48.17 weeks). Original data were fit on the log scale. Posterior states were back-transformed by exponentiation. **a.** the dynamic intercept, **b.** the seasonal trend, **c.** the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

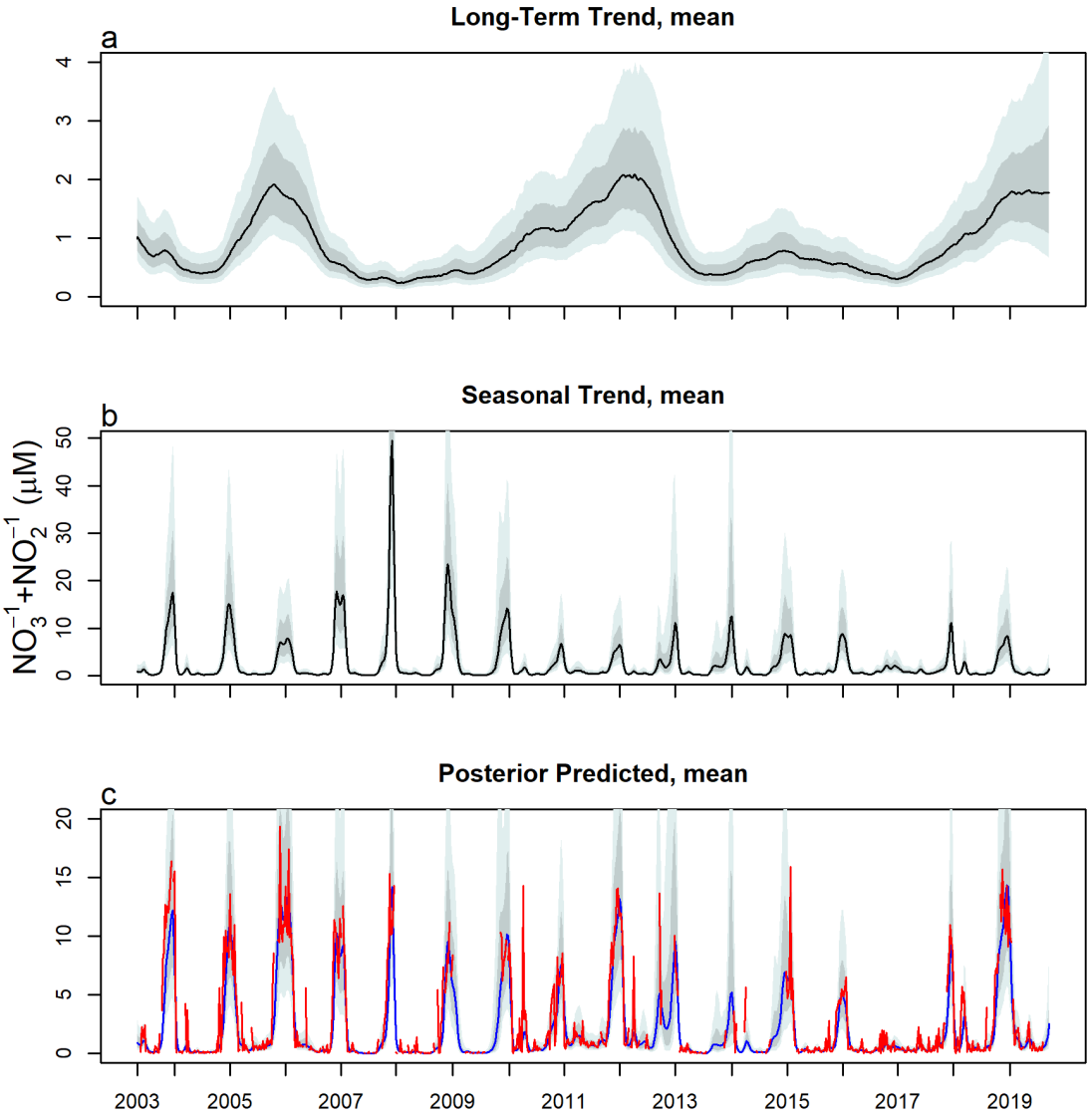

Figure 9. Decomposition of the NO_3+NO_2 series DLM (2003-2019), fit with a dynamic intercept and 5 seasonal components with periods of (52.17,…,48.17 weeks). Original data were fit on the log scale. Posterior states were back-transformed by exponentiation. **a.** the dynamic intercept, **b.** the seasonal trend, **c.** the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

Because of the high correlation of the series and similar patterns, series were combined into a single DIN series. Still, the temporal relationship between the two N series was explored with the CCF and wavelet coherence. In calculating the CCF, 'pre-whitening' was used, that is fitting a model to one variable to reduce it to white noise and using the same model to filter the second series. After pre-whitening, the CCF was calculated, showing NO_3+NO_2 is significantly correlated as per the $CCF_{critical}$ ($1.96/\sqrt{n}$) with NH_4 at lags 1:6, with highest correlation (0.42) at lag 0 (fig. 10). While the wavelet coherence requires smoothing, potentially masking dynamics on finer scales, significant coherence is shown between the series on the scale from several weeks to several years. The angle of coherence for spectral frequencies up to and around 1 year also show that NO_3+NO_2 dynamics lead NH_4 (appx. fig. 5). The wavelet analysis of individual series shows relatively consistently high power around the annual cycle at period 52 weeks, and in a biannual cycle around period 25 weeks for both series (appx. fig. 6, 7).

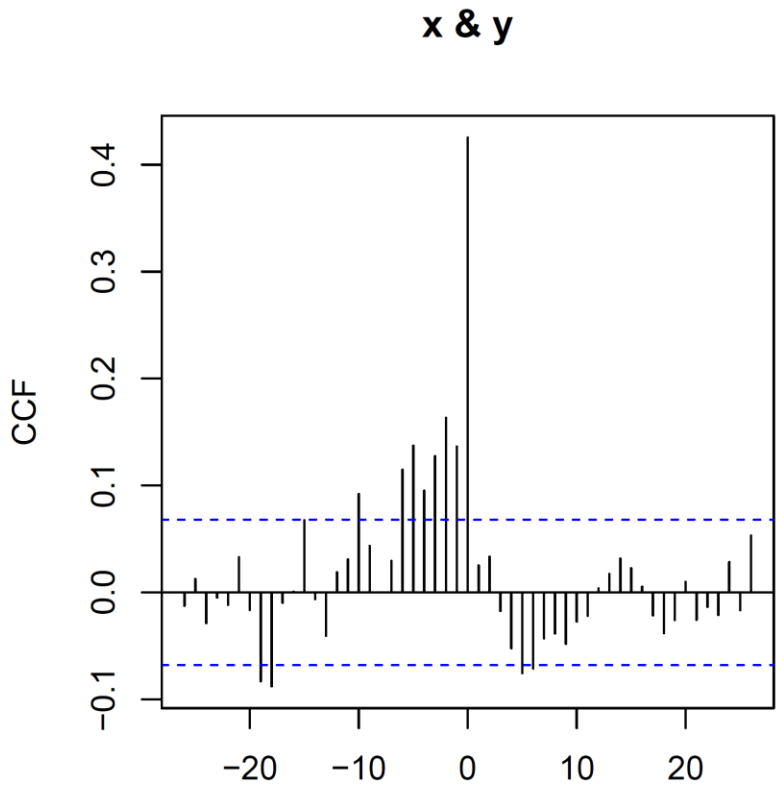

Figure 10. Cross correlation between NH_4 and NO_3+NO_2 after prewhittening. At the 95% confidence level, according to $CCF_{critical}$ ($\frac{1.96}{\sqrt{n}}$, blue dotted line), NO_3+NO_2 is significantly correlated with NH_4 at lags 1:6, with highest correlation at lag 0.

Last, although identifying predictors of DIN dynamics was not the focus of this thesis, the unexpected dynamics of the signal prompted some comparison to a monthly NAO index. NAO is not a single factor of environmental change but instead represents a complex of ecosystem traits, with positive years tending to have stronger southern wind, cloudier, and milder winters (Hurrell 1995). Negative NAO years tend to have colder, sunnier, and windier winters (Hurrell 1995). Investigation of plankton dynamics have—in some cases—been associated with NAO indices, including in Narragansett Bay (D. G. Borkman and Smayda 2009). Thereby, there was some

justification to explore how the NAO index may be related to DIN in an exploratory sense. Both show multiyear patterns in terms of their variability as well. At the annual and multiyear scale, significant coherence is seen between the NAO signal and DIN, suggesting an interesting subject for future investigation would be to investigate why DIN signal shows multiyear variability and not long-term patterns parallel to anthropogenic forcing (appx. fig. 8, 9). Because the NAO is a complex of environmental changes, potentially causal mechanisms are not obvious, but the coherence between the series suggests there are at least associations between this and the nutrient signal that are worth future investigation.

Chlorophyll a <20 & >20µm

Decomposition of the <20 chlorophyll series show variable levels in the dynamic intercept over the observation period, with an increase from before 2005 (0.59 ± 0.22) to after 2012 (0.91 ± 0.33), comparing the posterior distributions from these two periods, however, this is not a significant increase ($p=0.15$, fig. 11, 12). The seasonal signal shows clear annual periodicity, with some complex features likely due to bloom events. Overall, the amplitude of the seasonal signal increases with time, particularly in the years after 2012. This feature of increasing seasonal amplitude is also evidenced in the wavelet analysis performed on the imputed data series (fig. 13), where the annual cycle strengthens following the 2012 period.

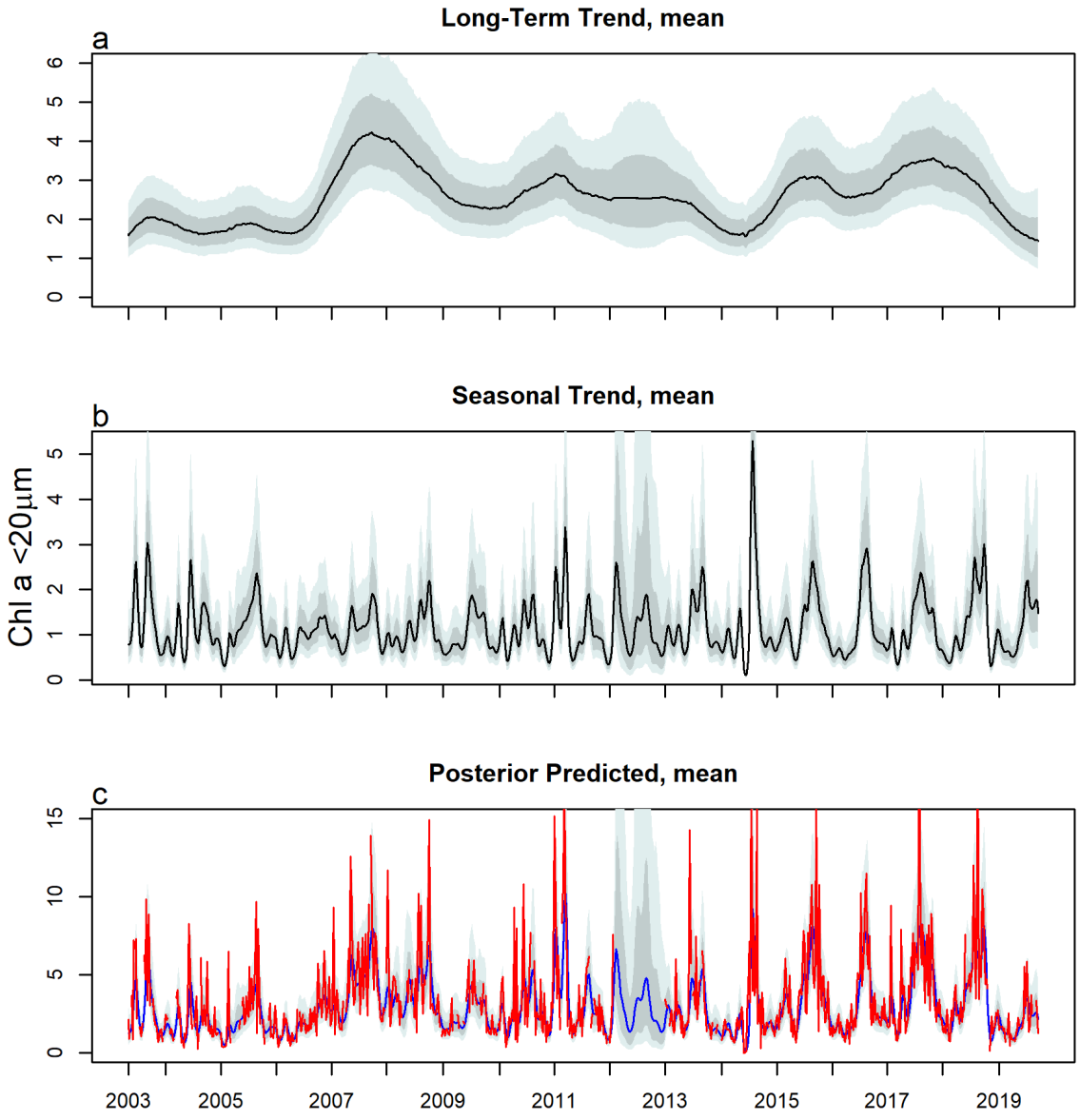

Figure 11. Decomposition of <20 µm Chl. a DLM (2003-2019), fit with a dynamic intercept and 5 seasonal components with periods of (52.17,…,48.17 weeks). Original data were fit on the log scale. Posterior states were back-transformed by exponentiation. **a.** the dynamic intercept, **b.** the seasonal trend, **c.** the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

Decomposition of the >20 chlorophyll series show variable levels in the dynamic intercept over the observation period, with a marked decline beginning in 2008 (fig. 14). While there is a slight uptick 2017-2019, the intercept of this period is still below that of the 2003-2008 period before declining levels. Comparing the posterior distribution of the intercepts from before 2005 (0.78±0.26) to after 2012 (0.31±38), the later period is significantly lower (p=0.011, fig. 12, 14). The seasonal signal shows no clear pattern, with some complex features likely due to bloom events. Overall, the amplitude of the seasonal signal is variable in time, with lowest seasonal signals represented during the period when the intercept was also at its lowest. This feature of low seasonal amplitude during overall lower levels is also evidenced in the wavelet analysis performed on the imputed data series (fig. 15), where the annual cycle is weakest 2010-2016. Further, beyond the predictable annual cycle, lower period signals intermittently show strong power in this series, suggesting the importance of finer scale events such as blooms.

Beyond mean levels, the observational matrix of the bivariate DLM for chlorophyll shows sever features. The >20 chlorophyll series shows inherently higher variability (1.1±0.02) than the <20 chlorophyll series (0.3±0.07, appx. fig. 10).

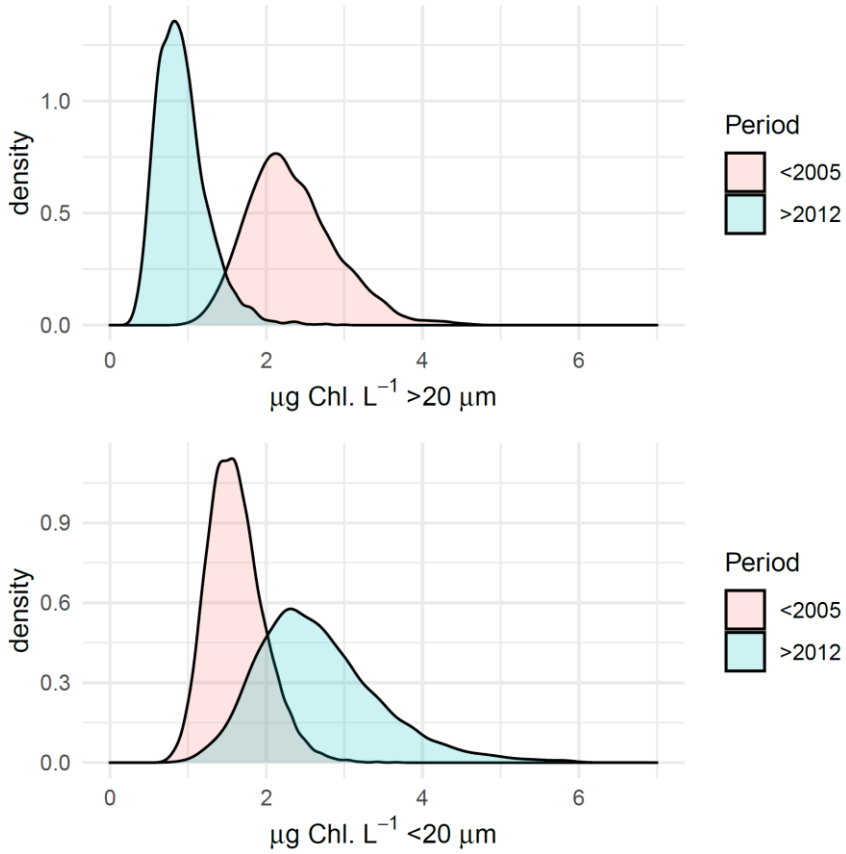

Figure 12. Density of the aggregated posterior dynamic intercept levels for the periods <2005, and greater than 2012 for the >20 and <20 µm Chl. a series.

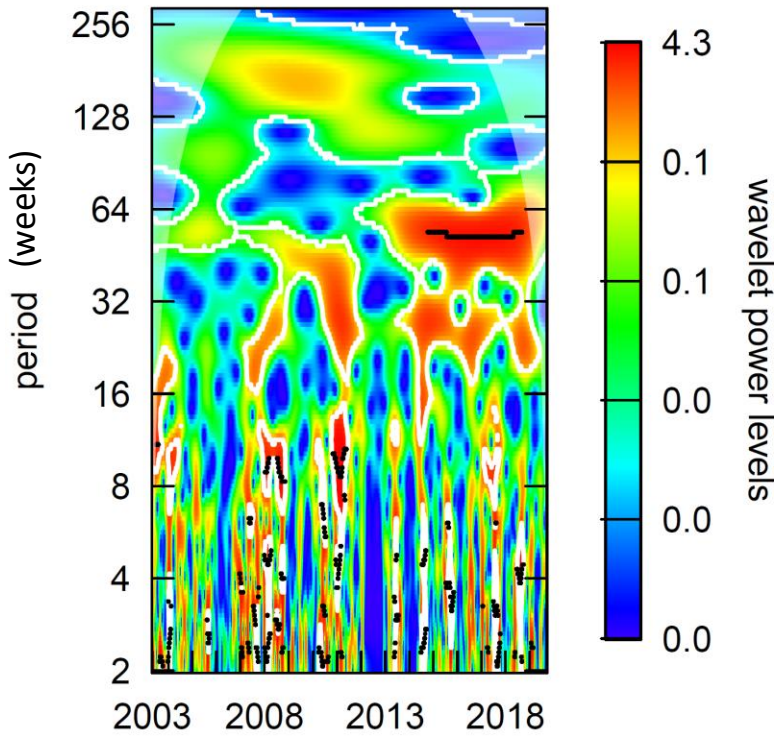

Figure 13. Wavelet analysis of the <20 µm Chl. a with missing data periods imputed with the mean latent state $\sum_{i=1}^{r} \frac{F_{i,T}{}'\theta_{i,T}}{r}, r = MCMC\ iterations$. Dominant variability occurs at the annual frequency though at lower and higher periodicity (to multiyear), variability is observed.

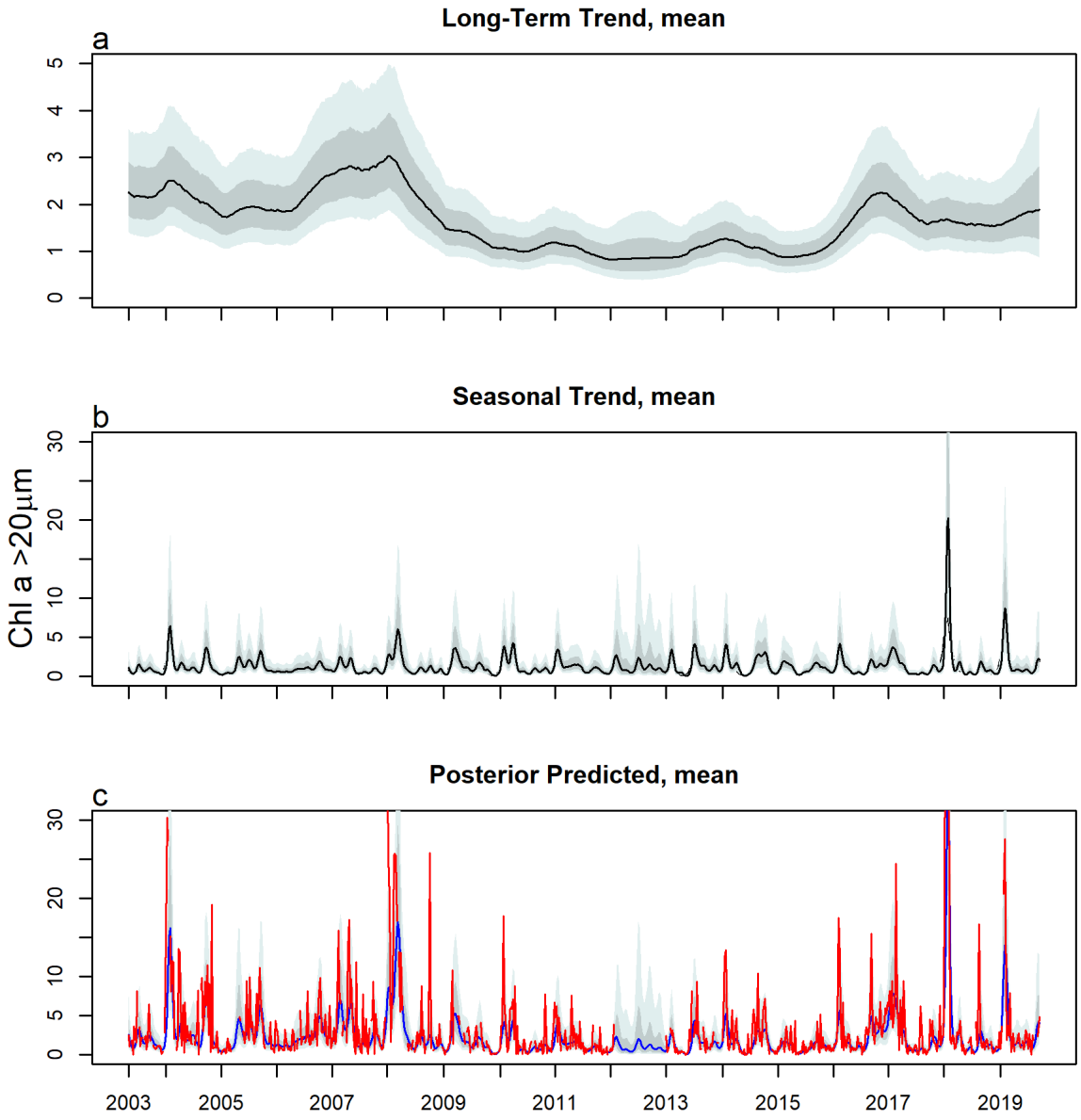

Figure 14. Decomposition of >20 µm Chl. a DLM (2003-2019), fit with a dynamic intercept and 5 seasonal components with periods of (52.17,…,48.17 weeks). Original data were fit on the log scale. Posterior states were back-transformed by exponentiation. **a.** the dynamic intercept, **b.** the seasonal trend, **c.** the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

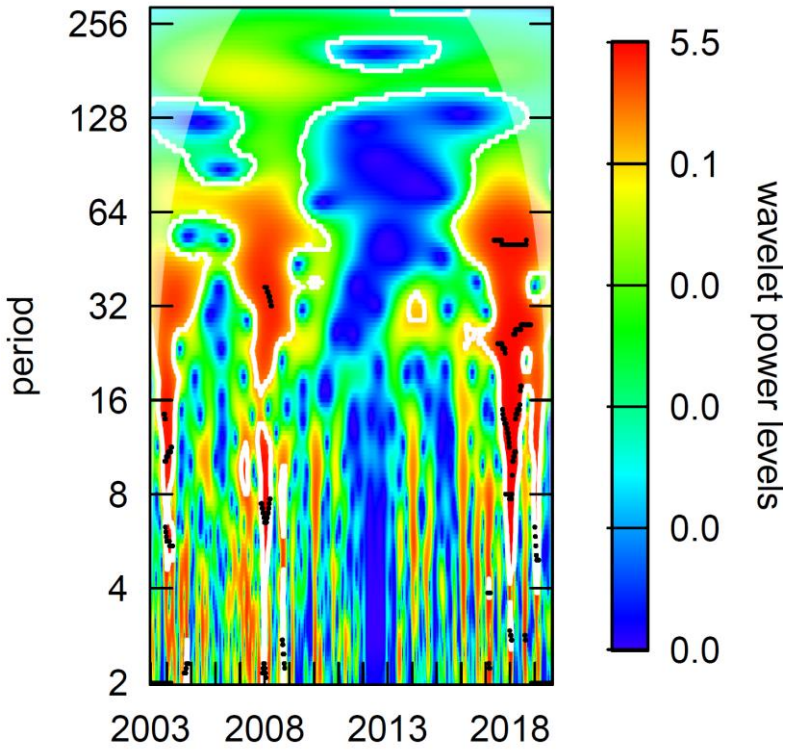

Figure 15. Wavelet analysis of the >20 µm Chl. a with missing data periods imputed with the mean latent state $\sum_{i=1}^{r} \frac{F_{i,T}{}'\theta_{i,T}}{r}, r = MCMC\ iterations$. Dominant variability occurs at the annual frequency though at lower and higher periodicity (to multiyear), variability is observed.

Using the above completed series of DIN, temperature, <20 µm Chl. a, and >20 µm Chl. a, the series were then compared with cross correlations, to characterize the relationship between the series across time lags (fig. 16). These cross-

correlations help inform the significantly associated lags, and the strength of association between the series.

After examining the cross correlation between the series and testing preliminary model structures, it was decided to use DIN as a regressor at lag 0, which shows the strongest cross-correlation to the chl. series (fig. 16). Using DIN instead of the NH_4 and NO_3+NO_2 series as separate regressors statistically avoided multicollinearity issues considering the two series are correlated. Scientifically, they also both represent sources of N to phytoplankton, and studies in Narragansett Bay have shown no preference to N species by size fraction (Furnas 1983). Ultimately, this allowed testing of how the nitrogen signal is associated to different size fractions of bulk chlorophyll. Notably, the cross-correlations between chlorophyll and DIN are negative, which has an important scientific interpretation: considering the expectation is that high ambient nitrogen can produce high biomass as indicated by chlorophyll. This is motivated by studies which show N is typically the limiting nutrient for phytoplankton, including in NB (Sakshaug 1977). However, the negative relationship suggests that the ambient nitrogen levels may in fact be partly determined by the phytoplankton, and thereby changes in the DIN signal are describing growth in the phytoplankton biomass. The lag of this feature is also important. The highest cross-correlation is at lag 0, suggests that the phytoplankton and nutrients are tightly coupled in time. For this reason, the lags 0 relations are explored to see how the relationship between phytoplankton and their nutrients could be changing with time. If the association between chl. and DIN increases, it would suggest that nutrients and biomass are becoming more tightly coupled. If the opposite occurs, it may mean that

there is an ecosystem shift to where other processes are dominating nutrient cycling and phytoplankton dynamics. In the second stage of the model, I fit a bivariate DLM with a static seasonal component, dynamic intercept, and dynamic regression with the DIN series. In this way, the regression with DIN captures anomalies in the chl. a series.

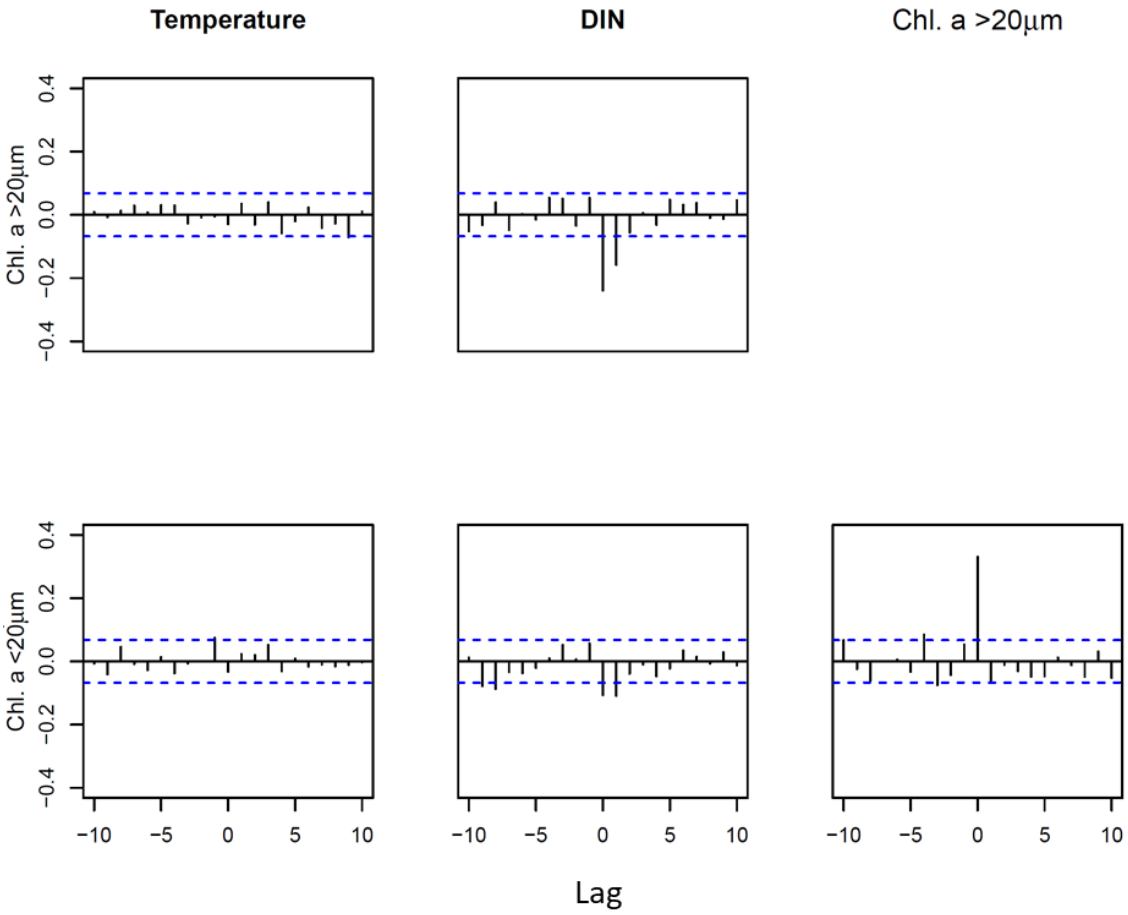

Figure 16. Cross correlations (y-axis) between the row and column variables in respective order, at different time lags (x-axis). The blue dotted line is the critical value for the 95% confidence level ($1.96/\sqrt{n}$), which represents the threshold for significant cross-correlations.

4.2 Stage 2 Model

For the bivariate regression model, with the true data fit with practical discounting and induced artificial missingness, while DIC, WAIC, and RMSE suggest the highest discount level model ($\delta_\mu = 0.99$, $\delta_\beta = 0.999$, appx. tables 22-23), RMSFE suggested a more flexible model ($\delta_\mu = 0.99$, $\delta_\beta = 0.99$, appx. table 24). Considering the simulation study, RMSFE is evidenced as a reliable metric with both practical and standard discounting in data series with long period missingness. As was seen in simulation, DIC, WAIC, and RMSE tended toward more static models, particularly with long-period missingness.

For final model interpretation, no artificial missingness was introduced. Again, DIC and WAIC suggested $\delta_\mu = 0.9$, $\delta_\beta = 0.999$ and $\delta_\mu = 0.99$, $\delta_\beta = 0.999$ respectively as the optimal model (appx. table 25). RMSFE suggested $\delta_\mu = 0.80$, $\delta_\beta = 0.90$ as more flexible model (appx. table 26). In total in terms of the discount factors selected, this model selection suggests there is a long-term trend in the chlorophyll series, varying slowly over the years as evidenced by the dynamic intercept. Further the selection of the discount factor for the regression coefficient suggests that the association between chlorophyll and the environment is time-varying.

Considering the low discount factor set selected in this case, with potentially too much adaptability from the selection of separate discount factors, the subset of models with identical discount factors was considered. In this case, the optimal model according to RMSFE was with $\delta_{\mu,\beta} = 0.85$ (fig. 17, appx. table 27). While this still suggests a highly flexible model, and relatively low signal in the data, it does still make apparent some of the meaningful dynamics in relation to each chl. a series and

the DIN series. For the chl. a <20 μm series, for the period prior to 2005 to that after 2012, there was no significant difference in the regression coefficient (fig. 18, p= 0.31). Considering the confidence intervals for the chl. a <20 μm series, the regression coefficient was also not significantly different form 0. For the chl. a >20 μm series, for the period prior to 2005 to that after 2012, there was no significant difference in the regression coefficient (fig. 19, p= 0.485). Considering the confidence intervals for the chl. a >20 μm series, the regression coefficient was periodically significantly different from 0. Wavelet analysis suggested this coefficient takes an annual cycle (fig. 20). Aggregating the mean coefficient across all years and comparing across the annual cycle, the strongest association tends to occur in the winter, suggesting the large phytoplankton are seasonally tied to the ambient nutrient signal (fig. 21). Evolutional covariance showed that there is a seasonal pattern to the evolutional rate in the state components (appx. fig. 11). The observational error, which also serves to represent cross series covariance shows similar results as the non-regression models in stage 1, where the larger size fraction of chl. a is more variable than the smaller size fraction, with a positive covariance between the series (appx. fig. 12). ACF of the residuals showed there was no significant temporal structure left in the data (appx. fig. 13). With this model, residuals were reduced to white noise.

Discount factor sampling was implemented with the full data series, however, there was no mixing in the sampling for the discount factor.

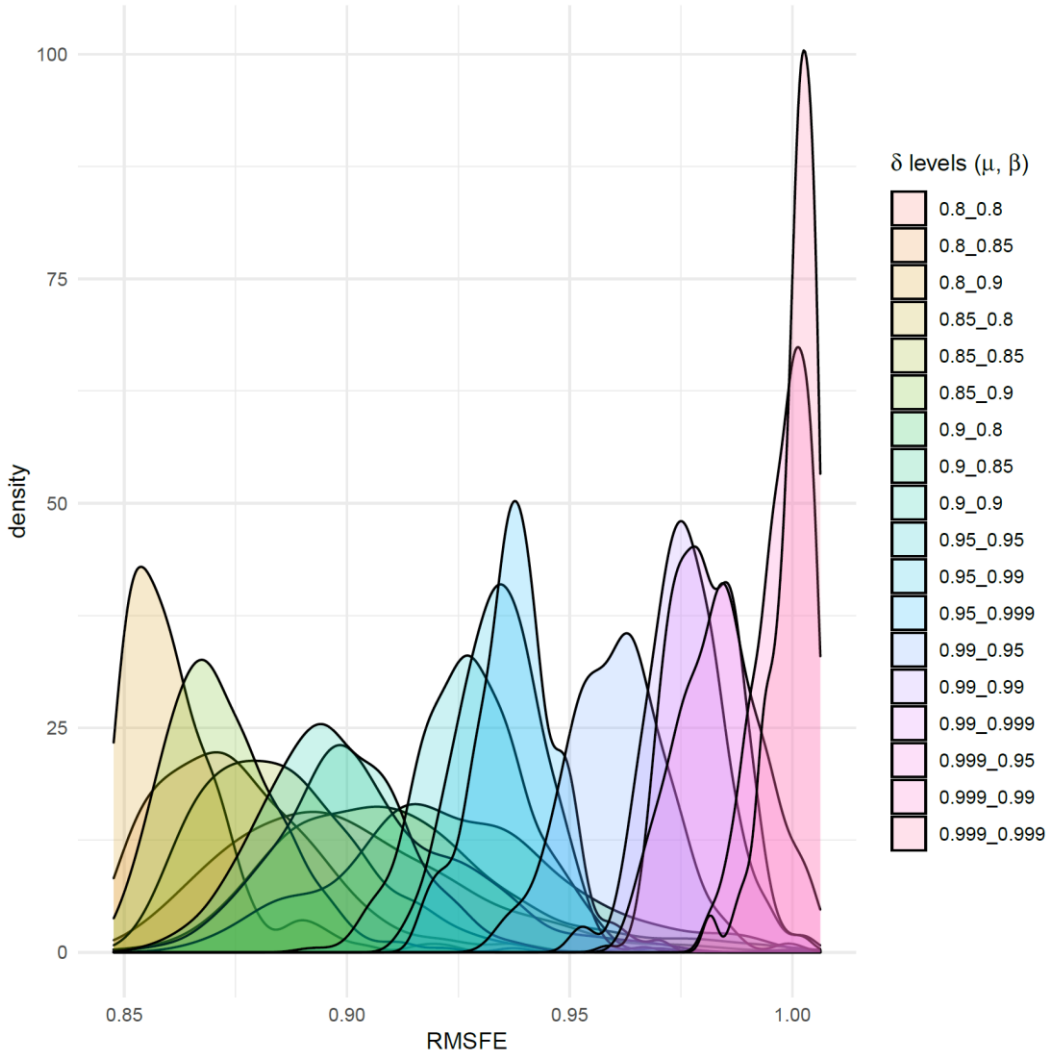

Figure 17. Posterior distributions for RMSFE for each model with fixed discount factors for μ and β (fill color), fit to true data in the case with no artificial missingness.

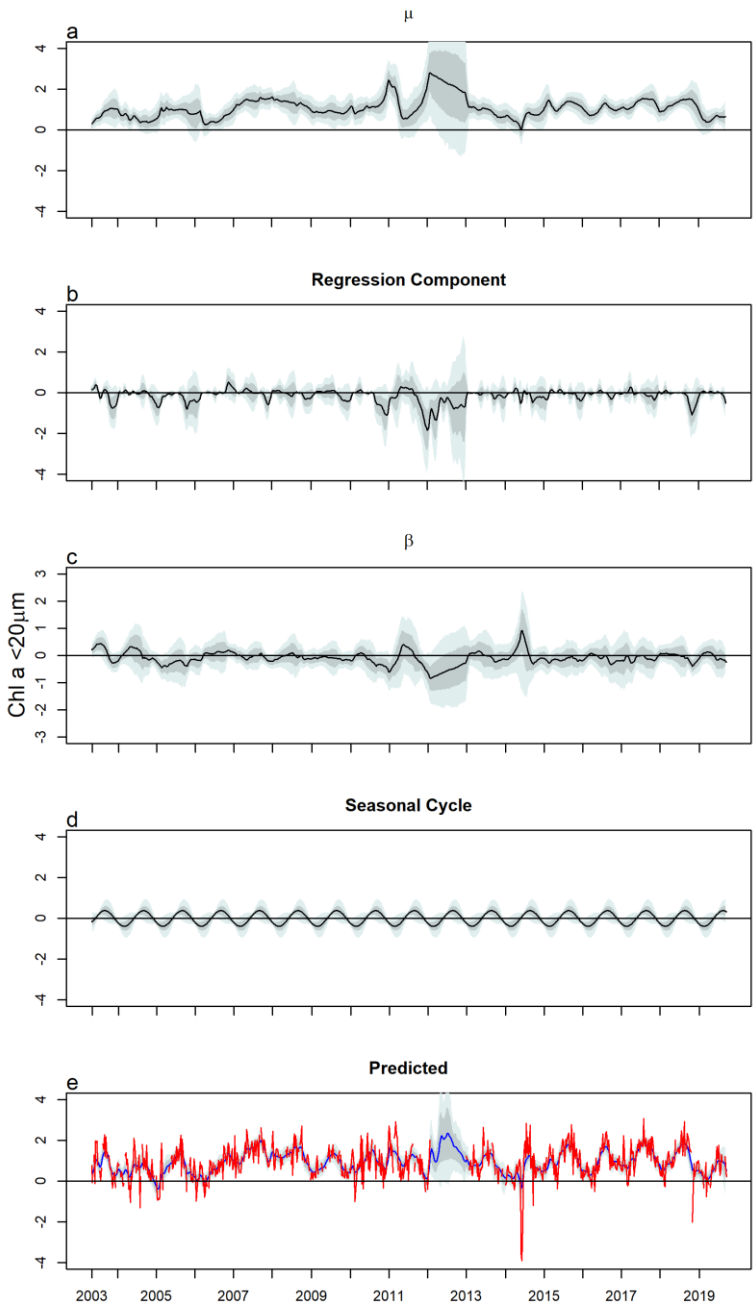

Figure 18. Decomposition of <20 µm Chl. a DLRM fit with a dynamic intercept, static seasonal component with a period of 52.17 weeks and regression component. **a.** the dynamic intercept, **b.** the regression component ($\beta_t X_t$), **c.** the regression coefficient (β), **d.** the seasonal cycle **e.** the posterior predicted mean (blue) with the true data (red). The median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

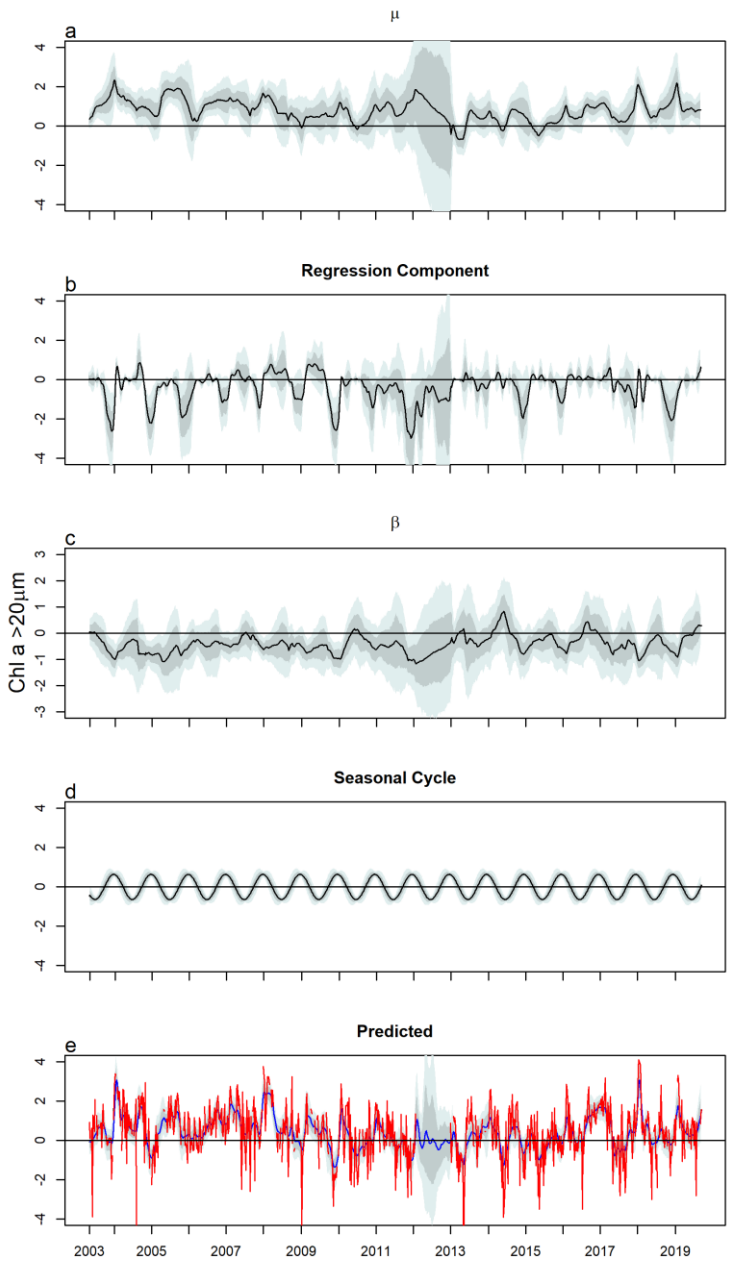

Figure 19. Decomposition of >20 µm Chl. a DLRM fit with a dynamic intercept, static seasonal component with a period of 52.17 weeks, and regression component. **a.** the dynamic intercept, **b.** the regression component ($\beta_t X_t$), **c.** the regression coefficient (β), **d.** the seasonal cycle **e.** the posterior predicted mean (blue) with the true data (red). The median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

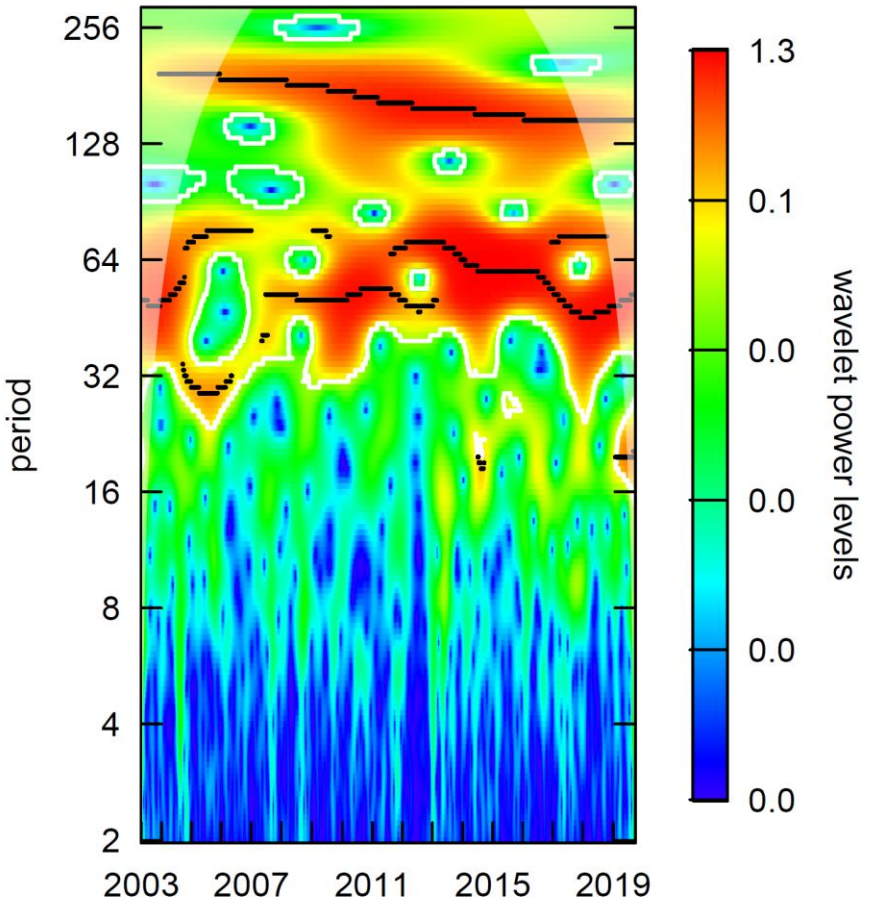

Figure 20. Wavelet decomposition of the regression coefficient on the latent state of DIN for the chl. a >20 µm series.

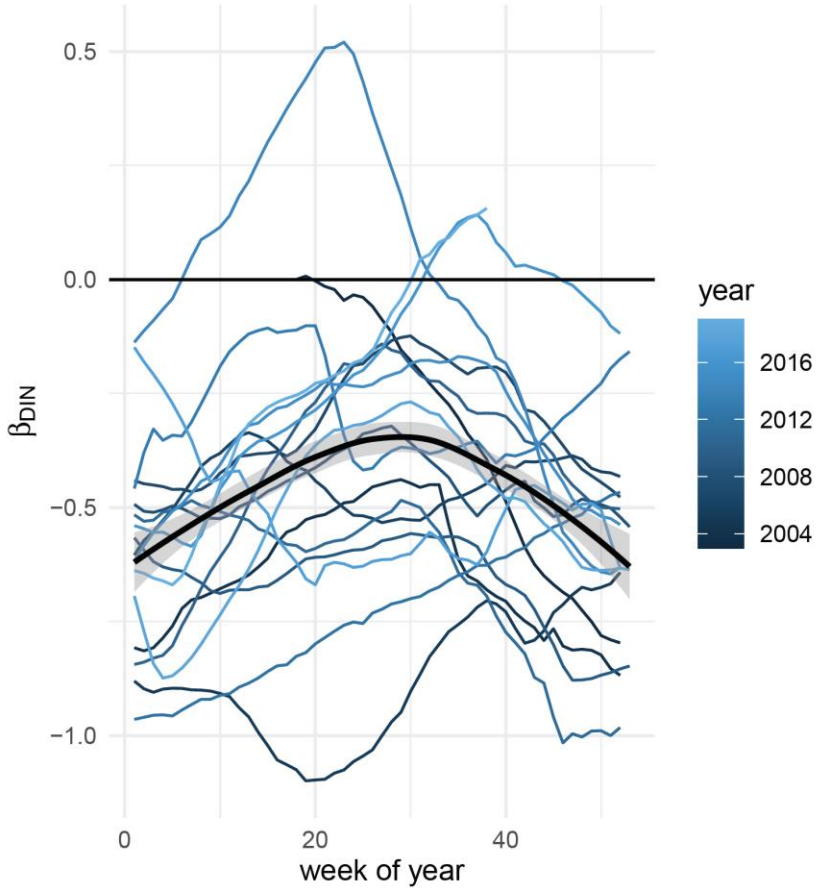

Figure 21. Dynamic regression coefficient on DIN for the chl. a > 20 µm, plotted by week on the x-axis, and by year as denoted by color shading. The seasonal pattern in the regression coefficient is exemplified by the loess smooth fit and 95% confidence interval.

CHAPTER 5

DISCUSSION

5.1 Physical and Chemical Change in Narragansett Bay

First, analysis of the DIN and temperature series with DLMs showed that the hypothesized patterns of change did not hold. In terms of temperature, although a linear trend has been repeatedly used to describe long-term change in the bay, which may generally hold on longer time scales, the dynamic intercept had a 95% CI range of 3.32°C as compared to the mean 1.5°C warming documented since 1959 (Fulweiler et al. 2015). This suggests that multiyear cycles have 66% greater magnitude than the past 60 years of climate warming. While this is not to suggest long-term warming is not evident, multiyear variations may dominate the temperature signal when trying to map environmental change to biological responses. Multiyear patterns could drive changes in temperature dependent biogeochemistry.

Beyond long-term mean levels in the temperature series, the seasonal signal shows anisotropic variability. That is, winter and summer are not equally variable. Winter temperatures of the posterior mean predicted (0.68 ± 1.6), more variable than summer maxima (24.15 ± 1.12). This is consistent with the previous finding that winter temperatures may be more susceptible to warming as compared to summer temperatures (Fulweiler et al. 2015). Biologically this represents a concern for the cold-water species of Narragansett Bay which may not be able to tolerate intermittently warm winters(Prog et al. 1984; Borkman and Smayda 2009). In terms of the size structure of the bay, altogether, this result in combination with the finding of

multiyear variability suggest that the expected pattern of net warming inducing a shift toward smaller organisms is unfounded. While multiyear variations in temperature could potentially influence the community structure, net switching toward smaller organisms does not match the temperature signal, which shows no net increase over this period of study from 2003 to 2019.

While Rhode Island state law mandated the reduction of 50% from wastewater effluent into the bay during the period of 2005-2012 (NBEP 2017, Oczkowski et al. 2018), net changes in the ambient DIN signal or its constituents at NBSII were not apparent. Notably, this does not mean that reduction in nutrient pollution did not take place and was not effective. Cross-sectional studies with spatial resolution showed net decline in nutrient levels at more northern locations closer to point sources of nutrient influx (Oviatt et al. 2017, fig. 22). Further, it has also been suggested that the benthos of the bay may be a major N source, which would potentially stabilize ambient levels (Nixon et al. 2009). Instead of a clear monotonic drawdown in nutrients, there is multiyear variability apparent both in seasonal cycle and mean log levels. Cross-correlation of DIN, showed significant correlation to both series, suggesting an interesting linkage to the biological feature of chlorophyll of all sizes. The high correlation between DIN and each chlorophyll signal at lag 0 suggested that these series are tightly linked in time, and that both the biology and ambient nutrients respond quickly to each other. This is consistent with laboratory work in phytoplankton which has showed that phytoplankton growth is impacted by nutrient variation within 24 hours, much finer than the weekly resolution of the data presented here (Collos 1986).

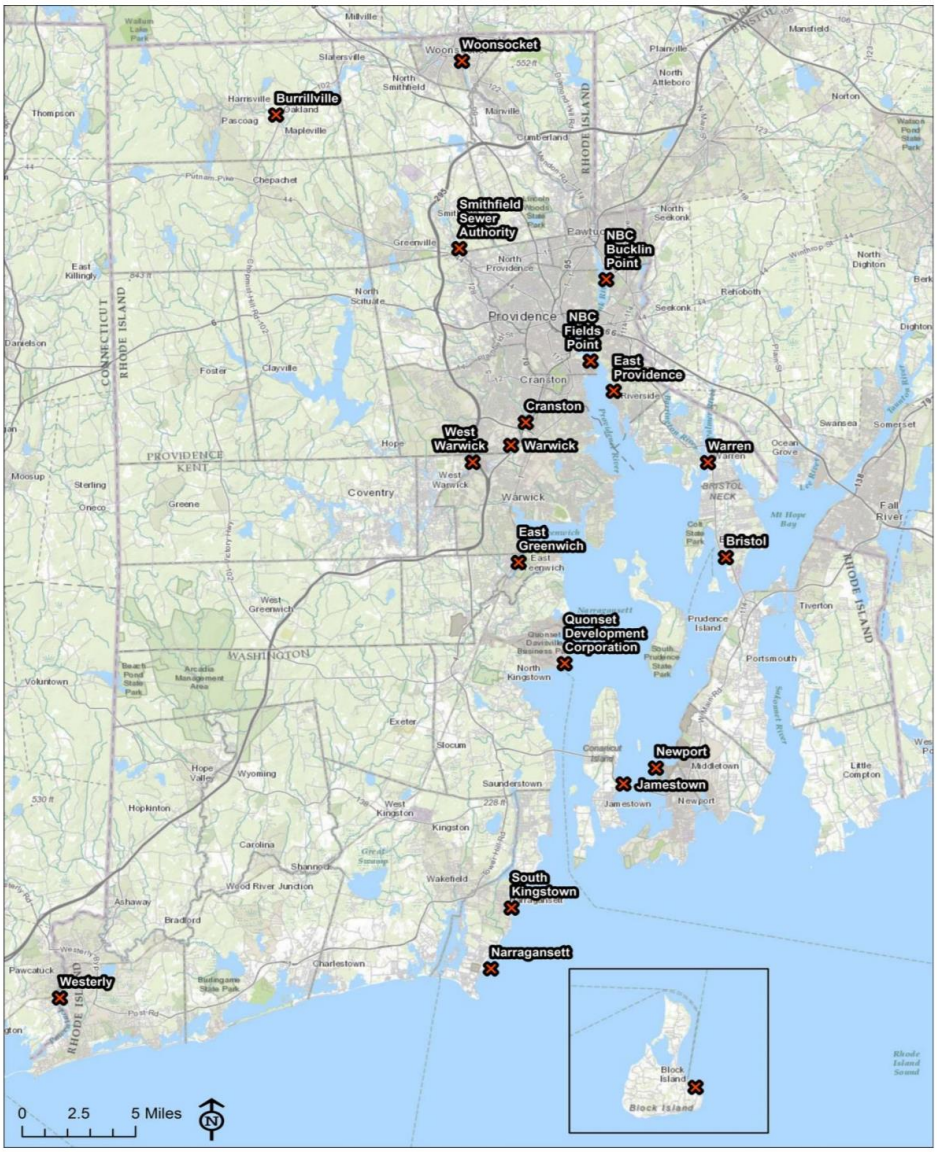

Figure 22. Map of all 19 discharge locations for wastewater in RI, reproduced from (RI DEM, 2016). Point sources are concentrated in northern reaches in the bay which also experience less flushing from the ocean.

Because the constituents of DIN (NH_4 and NO_3+NO_2) are highly correlated (0.50±0.037) and phytoplankton are generally capable of using both forms of N (Glibert and Garside 1992), it was decided to relate DIN to the chlorophyll series. The

high correlation between the covariates was of particular concern in part because of identifiability in DLMs and bias variance tradeoff. Too many parameters in the DLM model, where each variable is also 'flexible' in time can result in identifiability issues, high variance in the estimates, and a more minor but computation concern of highly correlated MCMC chain sequences (Gelman et al. 2014; West and Harrison 1997), which hypothetically could require prohibitively long MCMC simulations to produce a sufficient effective sample size. Considering the DIN constituents, both series of N species show a high correlation of (0.50+0.037), with overall higher variability in the NO_3+ NO_2 (0.94+0.07), as compared to NH_4 (0.74+0.05). The CCF between NO_3+NO_2 and NH_4 showed the strongest correlation at lag 0 (0.42), but also some significant positive cross-correlations with NH_4 at lags 1:6, suggesting NH_4 can trail behind NO_3+NO_2. Biologically, this could be representative of processes like preferential uptake in bloom dynamics (Dugdale et al. 2007) remineralization of NH_4 which occurs from both benthic and pelagic organisms (Goeyens et al. 1987; Probyn 1987). This cross-correlation by itself is interesting as it suggests variation in the NO_3+NO_2 may influence other nutrient species (i.e. NH_4) as much as 6 weeks later. Again, the significant correlation at lag 0 and 1 for DIN with chlorophyl suggests phytoplankton are consistently responding rapidly to nutrient changes in the bay whereas other components of the microbial loop, which are thought to fuel nutrient regeneration (Laybourn-Parry and Parry 2000), may take weeks to respond. These rates of response between different members of marine microbial ecosystems are an active area of investigation with some conflicting results between studies depending the stressor and dataset. For example, spectral decomposition of time-series of

phytoplankton and zooplanktonic predators in the Western English Channel, have shown evidence that phytoplankton have whiter power spectra, suggesting dominant variability on short timescales as compared to zooplankton (Barton et al. 2020). Nevertheless, the choice to use DIN captures both of these series, as different phytoplankton are capable of utilizing both these nutrients and the chlorophyll series captures numerous species with different nutrient preferences (Dortch 1990).

While the practical applications of this work center around the size structural data and their relationship to the DIN signal, the exploratory comparison with NAO yielded interesting correlations between the series as represented by the wavelet coherency. While NAO is itself an index for a broad range in environmental features, each capable of affecting phytoplankton and their physical environment (Hurrell 1995), it still provides an interesting subject for what could be influencing the multiyear patterns in all of the N series. The wavelet coherency shows that there is significant correlation between the DIN levels and the NAO index at both the annual and multi-year scales, suggesting that some of the long-term dynamics in nutrients may be related to this and possibly other multiyear climate patterns, and not just anthropogenically driven change, even as dramatic as 50% reduction in the largest pollutant source (NBEP 2017, Oczkowski et al. 2018).

5.2 Size Structural Changes of Phytoplankton in Narragansett Bay

Analysis of the size structural data with DLM's show that there is a marked change in the size structure of phytoplankton in Narragansett Bay, with phytoplankton >20 μm 0.47 $\mu g\ L^{-1}$ lower ($p= 0.011$), and phytoplankton <20 μm 0.32

μg L^{-1} higher, though, not significantly so (p=0.15). This suggests that across the period of nutrient reduction in the wastewater effluent of the bay, the phytoplankton of the Bay has undergone a major shift toward smaller organisms, mostly as a consequence of declining stocks of larger phytoplankton. Ecologically this is important because cell size of the phytoplankton community can have rippling effects through food webs (Finkel 2007). This size structural change aligns with hypotheses of nutrients in the bay, though those reductions were not detectible in the NBSII series. While phytoplankton >20 μm were more abundant than phytoplankton <20 μm prior to 2005, this reversed after 2012, with phytoplankton <20 μm more abundant than those > 20 μm. This suggests that, although there are still seasonal differences, Narragansett Bay has become dominated by smaller phytoplankton over a geochemically short span of time (7 years).

Further, in terms of abundance patterns, while the <20 μm fraction of phytoplankton was not significantly increased in their mean, they do show an increase in their seasonal cycle following the end of nutrient reduction. This equates to higher summer maxima and lower winter minima. Though, after back transformation through exponentiation, these effects are most pronounced in the summer.

In addition to long-term patterns of change, the univariate DLMs show that the phytoplankton >20 μm are inherently more variable, with an observational variance of 1.05 $\pm$0.06. The phytoplankton <20 μm show much lower variability with a variance of 0.31$\pm$0.02. This suggests that larger plankton in the bay are potentially much patchier, and inherently more stochastic in their population dynamics. In contrast, stocks of smaller phytoplankton are much less stochastic. The difference in variance

between the two series has direct biological interpretation in line with hypotheses about bloom dynamics and grazing control such as the loophole hypothesis (Irigoien, Flynn, and Harris 2005). It has been suggested that in general smaller phytoplankton are more tightly coupled to predator control, and this results in more stable populations as compared to larger phytoplankton (Irigoien, Flynn, and Harris 2005). The higher variance of the phytoplankton >20 µm, suggests this may be true locally in some of the fine scale dynamics in the standing stock. Overall, considering these differences in stochasticity, seasonal cycle, and long-term trend, it stands to question which phytoplankton species, might be driving the chlorophyll dynamics, and what might their specific ecology inform about these long-term changes.

5.3 Selected Model and Dynamic Regression Signal with DIN

Ultimately, model selection tended toward low discount factors for μ and β, suggesting not only dynamic levels of each series over time, but also that the association with DIN is variable. As indicated graphically by the 95% CI, for most of the series the phytoplankton <20 µm are not significantly associated with DIN signal. This suggests, both that phytoplankton <20 µm are relatively invariant to ambient DIN signals and that DIN levels are not shaped by the phytoplankton <20 µm community in NB. Considering that after the nutrient reduction period, phytoplankton <20 µm are on average dominant in the phytoplankton community, it is thus surprising that they are still non-significantly related to the DIN signal.

In contrast to the phytoplankton <20 µm, phytoplankton >20 µm are often significantly tied to the DIN signal. However, the relationship between phytoplankton >20 µm and DIN is non-static and exhibits evidence of annual

cyclicity. In general, the regression coefficient is largest in magnitude and has the largest effect in winter periods. The lowest effect is in the summer. This suggests that the larger phytoplankton and DIN levels are more closely tied in the colder month periods when blooms are known to occur. There is no clear long-term shift in the regression coefficient for the chl. a > 20 µm series. This suggests that for the phytoplankton > 20 µm, dependence on DIN has not shifted after nutrient reductions, and further that the potential role of larger phytoplankton as biogeochemical engineers has not been impacted despite the apparent declines in the representation of this size class.

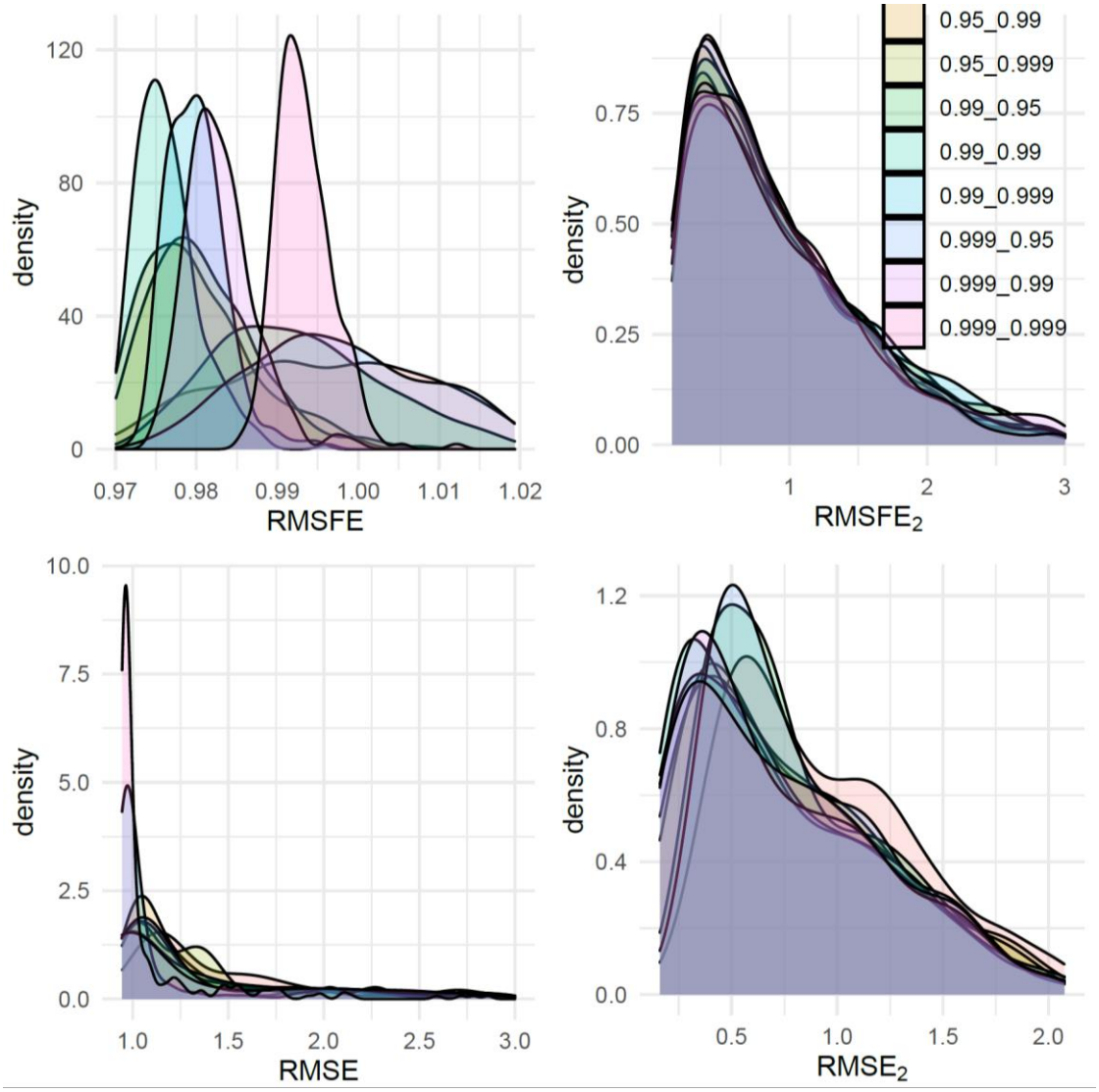

Figure 1. Posterior distributions for **a.** RMSFE **b.** RMSFE$_2$ **c.** RMSE **d.** RMSE$_2$ for each model with fixed discount factors for μ and β (fill color), fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$.

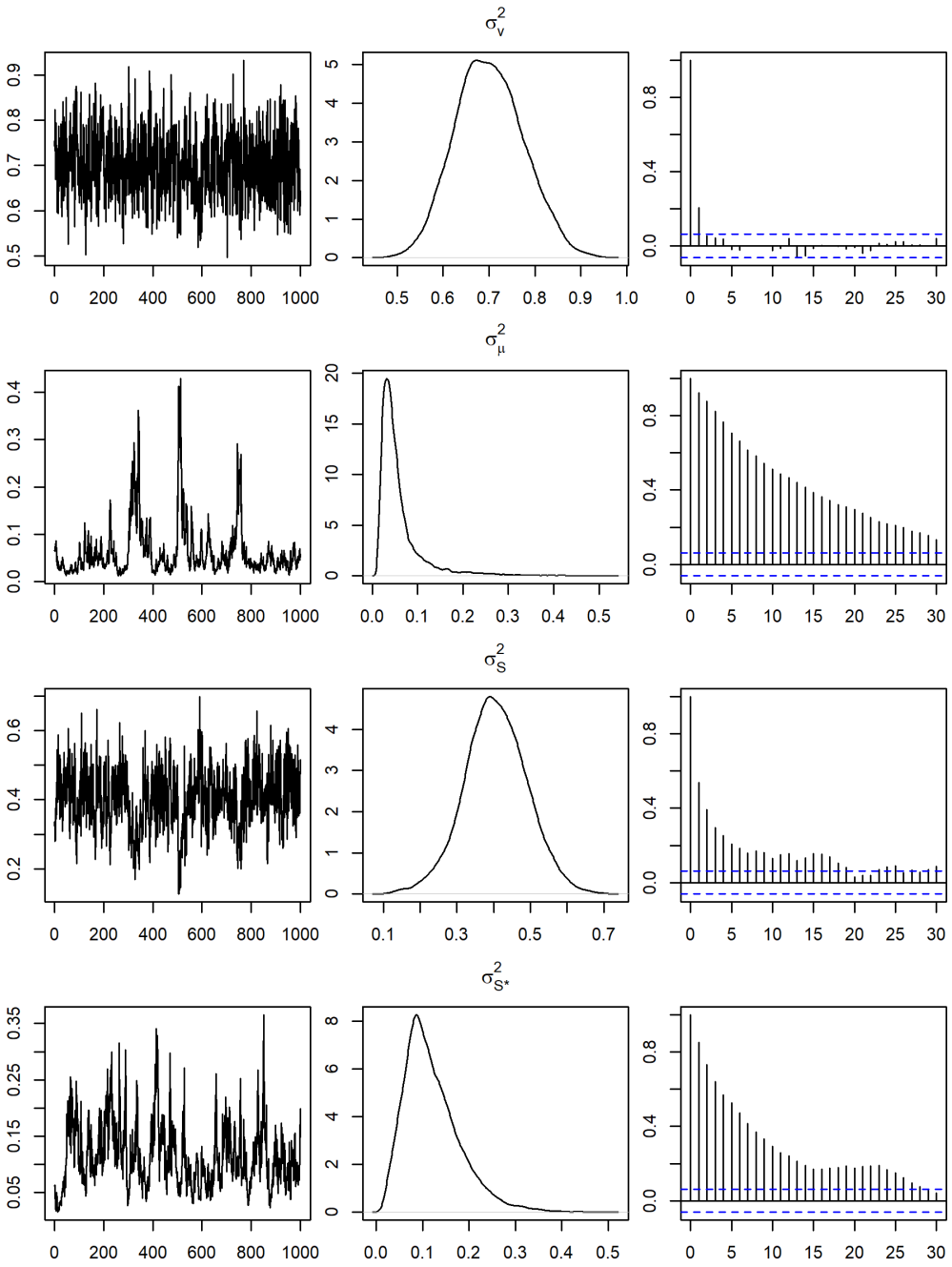

Figure 2. Trace and density plots of the observational and evolutional variance components in the state and observation equation fits. **a.** Evolutional variance of the dynamic intercept. **b.** Evolutional variance of the seasonal frequency **c.** Evolutional variance of the conjugate of the seasonal frequency. **d.** Observational variance of the series.

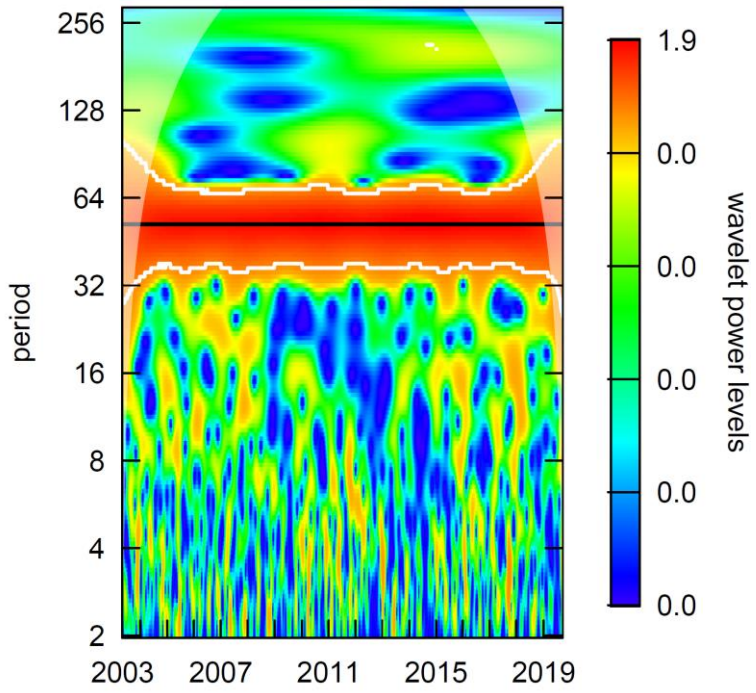

Figure 3. Wavelet analysis of the temperature series with missing data periods imputed with the mean latent state $\sum_{i=1}^{r} \frac{F_{i,T}{}'\theta_{i,T}}{r}, r = MCMC\ iterations$. Dominant variability occurs at the annual frequency, though at lower and higher periodicity (to multiyear), variability is observed.

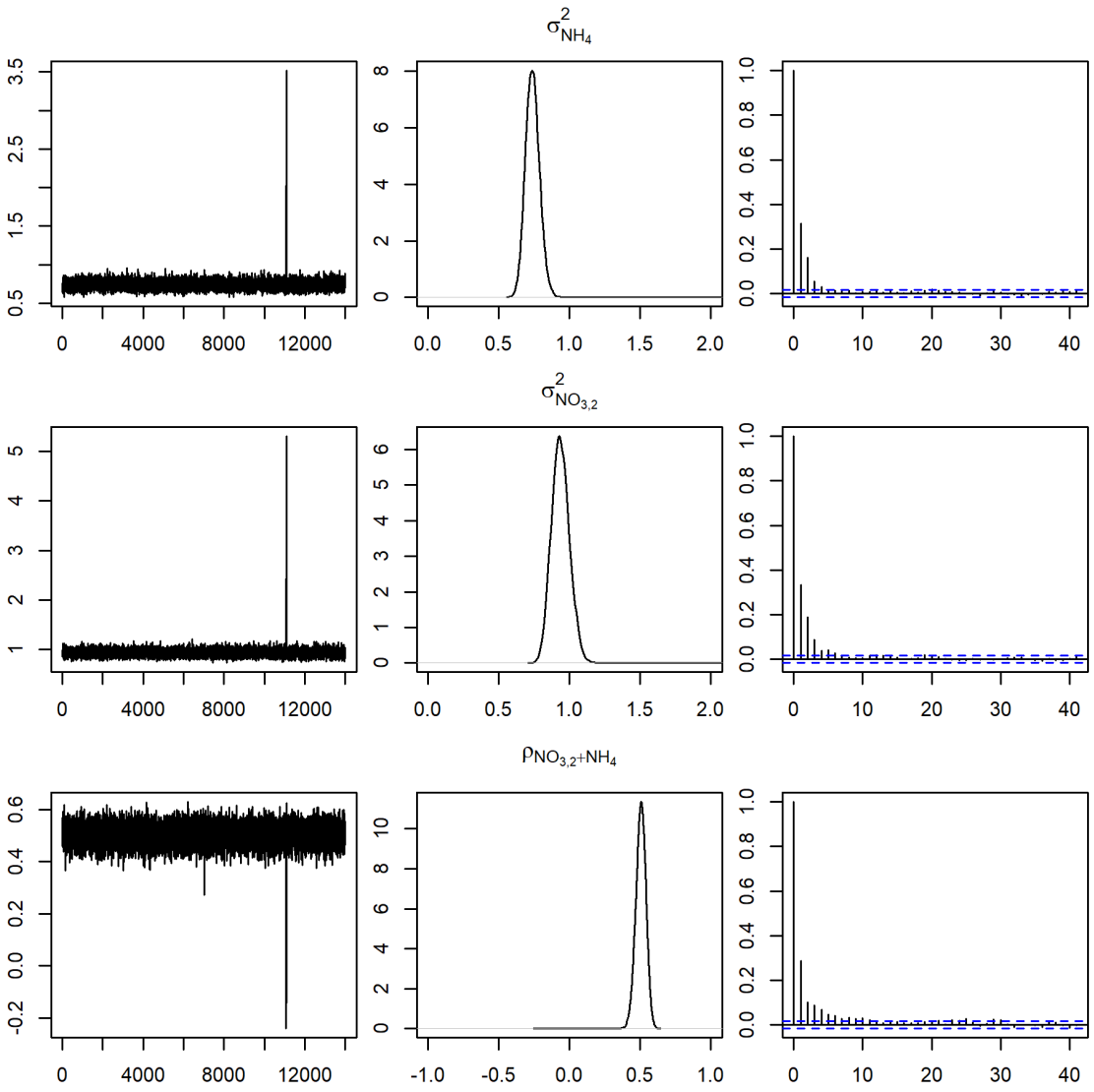

Figure 4. Trace and density plots of the observational variance and correlation of the NH_4 and NO_3+NO_2 series **a.** Observational variance of the NH_4 dynamic intercept. **b.** Observational variance of the NO_3+NO_2 seasonal frequency **c.** Observational covariance of the NH_4 and NO_3+NO_2 series.

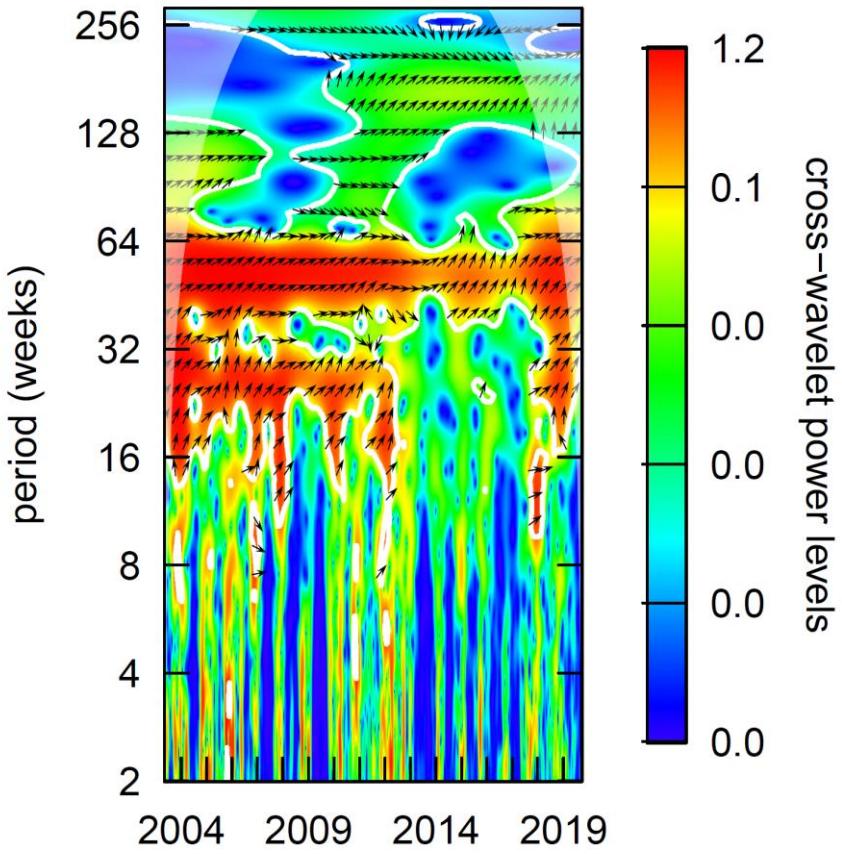

Figure 5. Coherence between NH_4 and NO_3+NO_2. Arrows indicate the angle of cohesion, with those pointing right and up show that NO_3+NO_2 is leading in the dynamics at that scale. This suggest the annual cycle of NH_4 is lagged behind NO_3+NO_2 dynamics. While the magnitude of coherence decreases with decreasing period, notably, the method uses smoothing which may obfuscate finer scale dynamics.

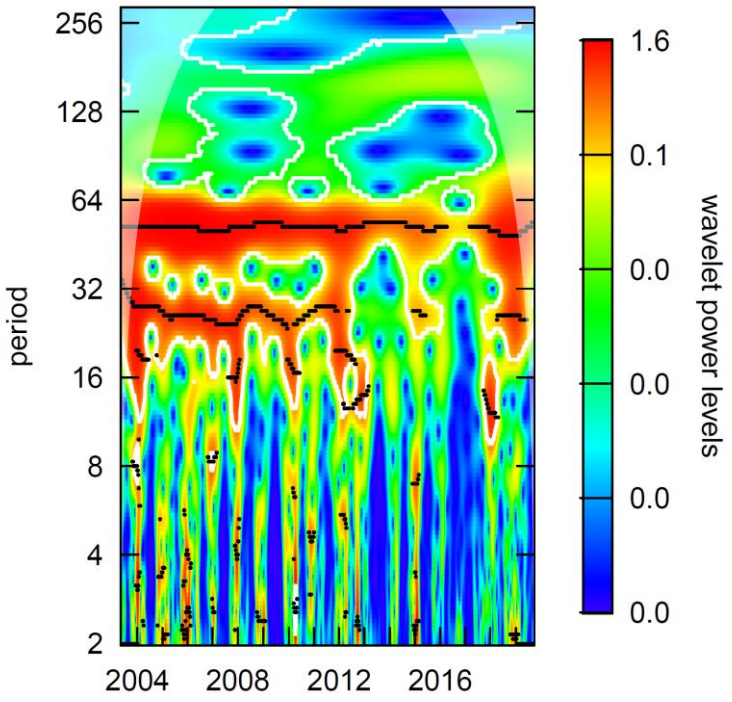

Figure 6. Wavelet analysis of the NH_4 series with missing data periods imputed with the mean latent state $\sum_{i=1}^{r} \frac{F_{i,T}{}'\theta_{i,T}}{r}, r = MCMC\ iterations$. Dominant variability occurs at the annual frequency though at lower and higher periodicity (to multiyear), variability is observed.

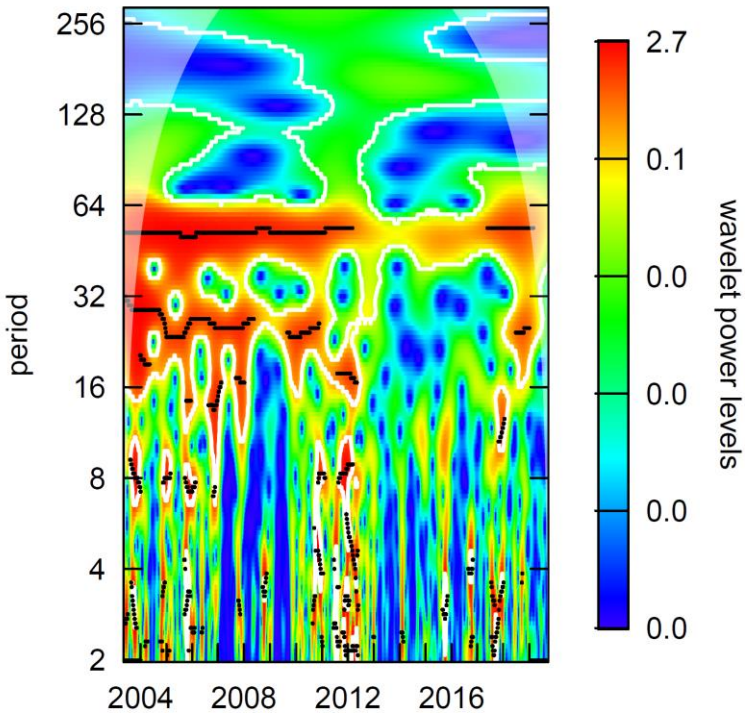

Figure 7. Wavelet analysis of the NO_3+NO_2 with missing data periods imputed with the mean latent state $\sum_{i=1}^{r} \frac{F_{i,T}{}'\theta_{i,T}}{r}, r = MCMC\ iterations$. Dominant variability occurs at the annual frequency though at lower and higher periodicity (to multiyear), variability is observed.

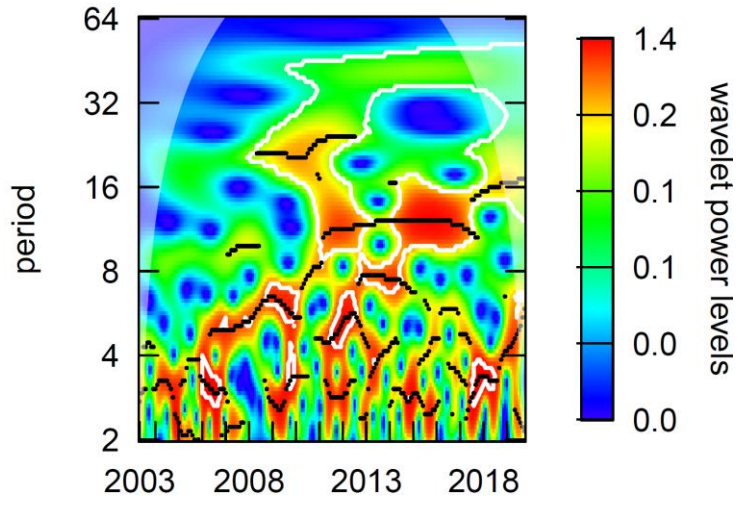

Figure 8. Wavelet analysis of the NAO index series.

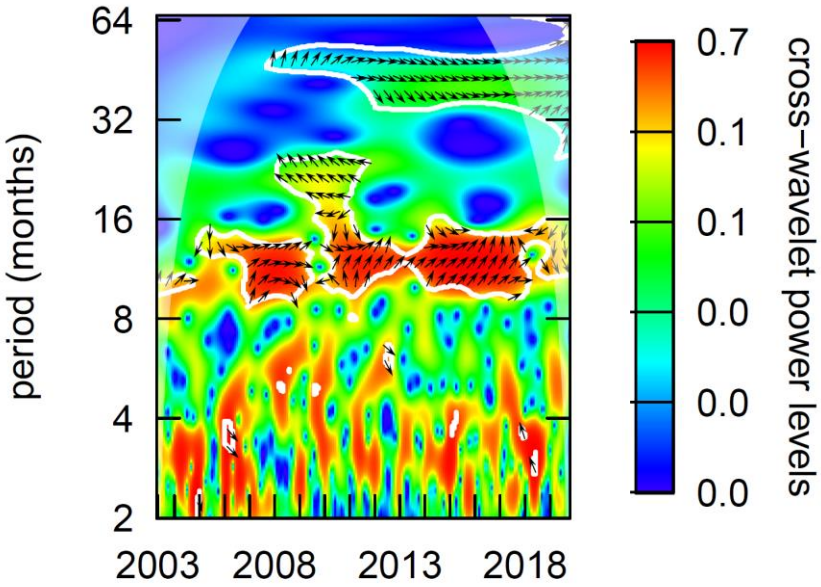

Figure 9. Coherence between DIN and the NAO index. Arrows pointing right and up indicate the angle of cohesion and that the NAO index is leading in the dynamics at that scale. This suggest the annual cycle of DIN is lagged behind NAO index at the annual scale and potentially synchronous at the multiyear scale. While the magnitude of coherency decreases with decreasing period, notably, the method uses smoothing which may obfuscate finer scale dynamics.

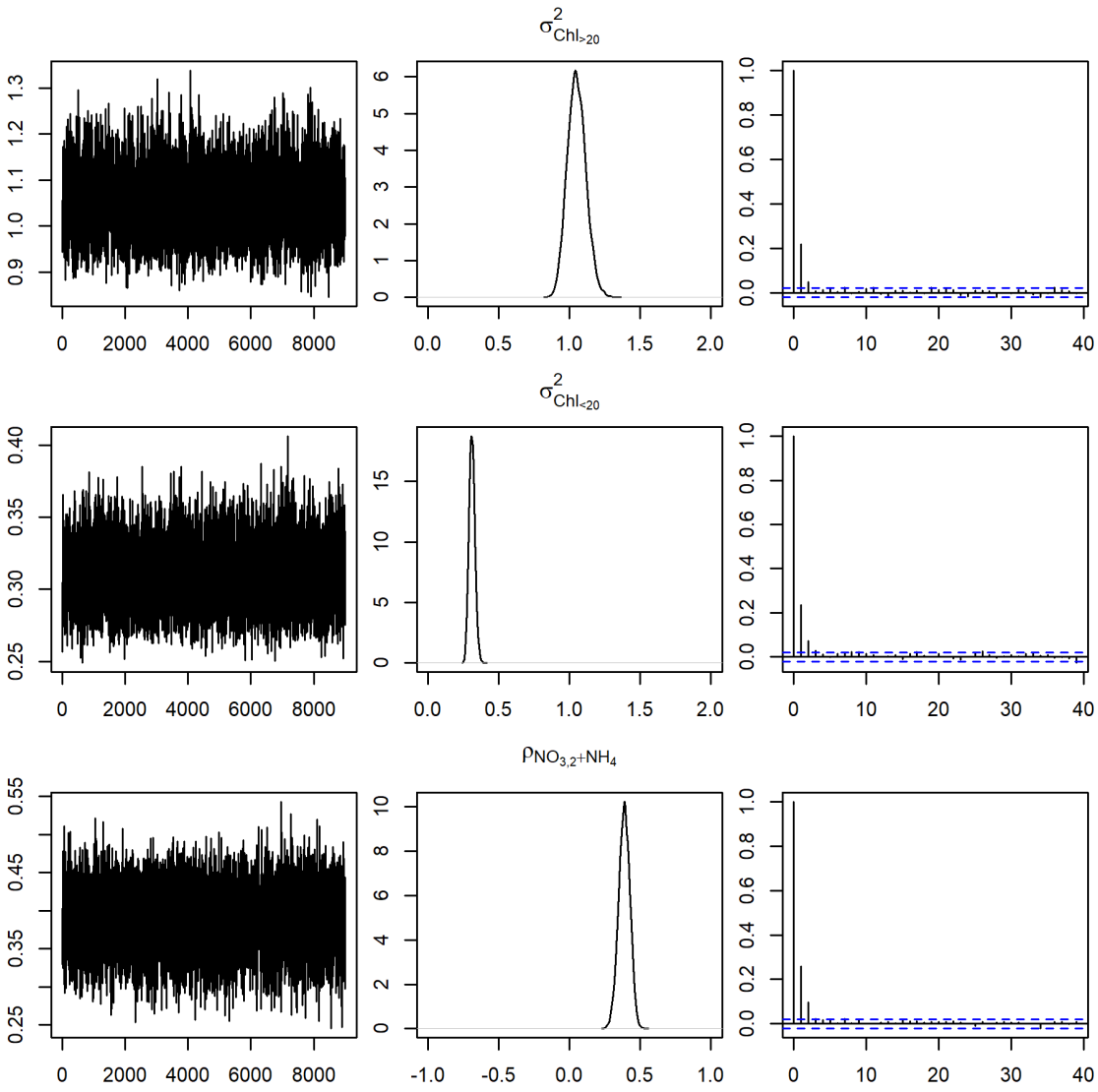

Figure 10. Trace and density plots of the observational variance and correlation of the >20 µm and <20 µm Chl. a series **a.** Observational variance of the >20 µm series. **b.** Observational variance of the <20 µm series **c.** Observational correlation of the >20 µm and <20 µm series.

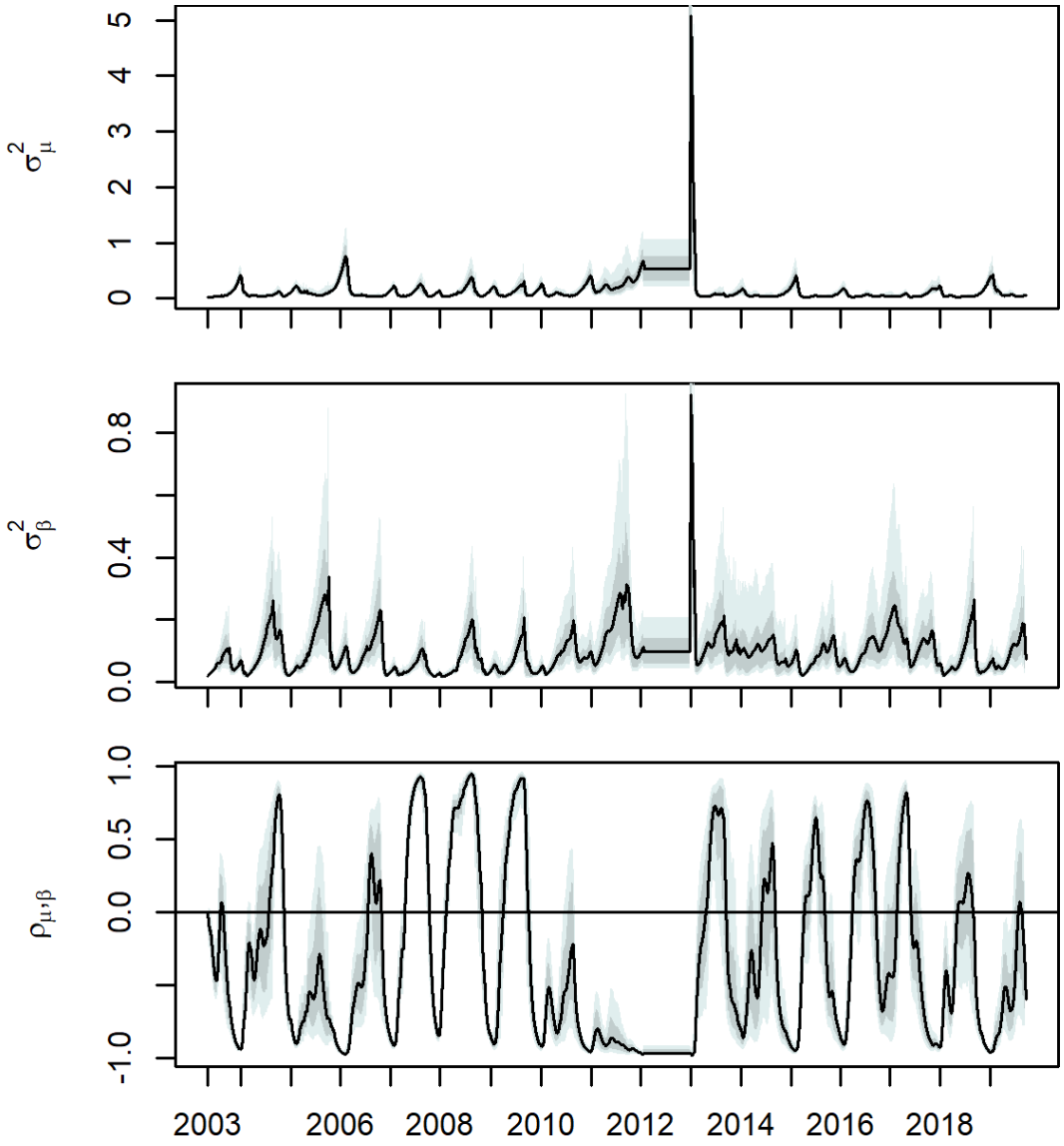

Figure 11. Trace and density plots of the evolutional variance and correlation of the dynamic intercept and regression coefficients. **a.** Evolutional variance of the dynamic intercept **b.** Evolutional variance of the regression coefficient. The median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.

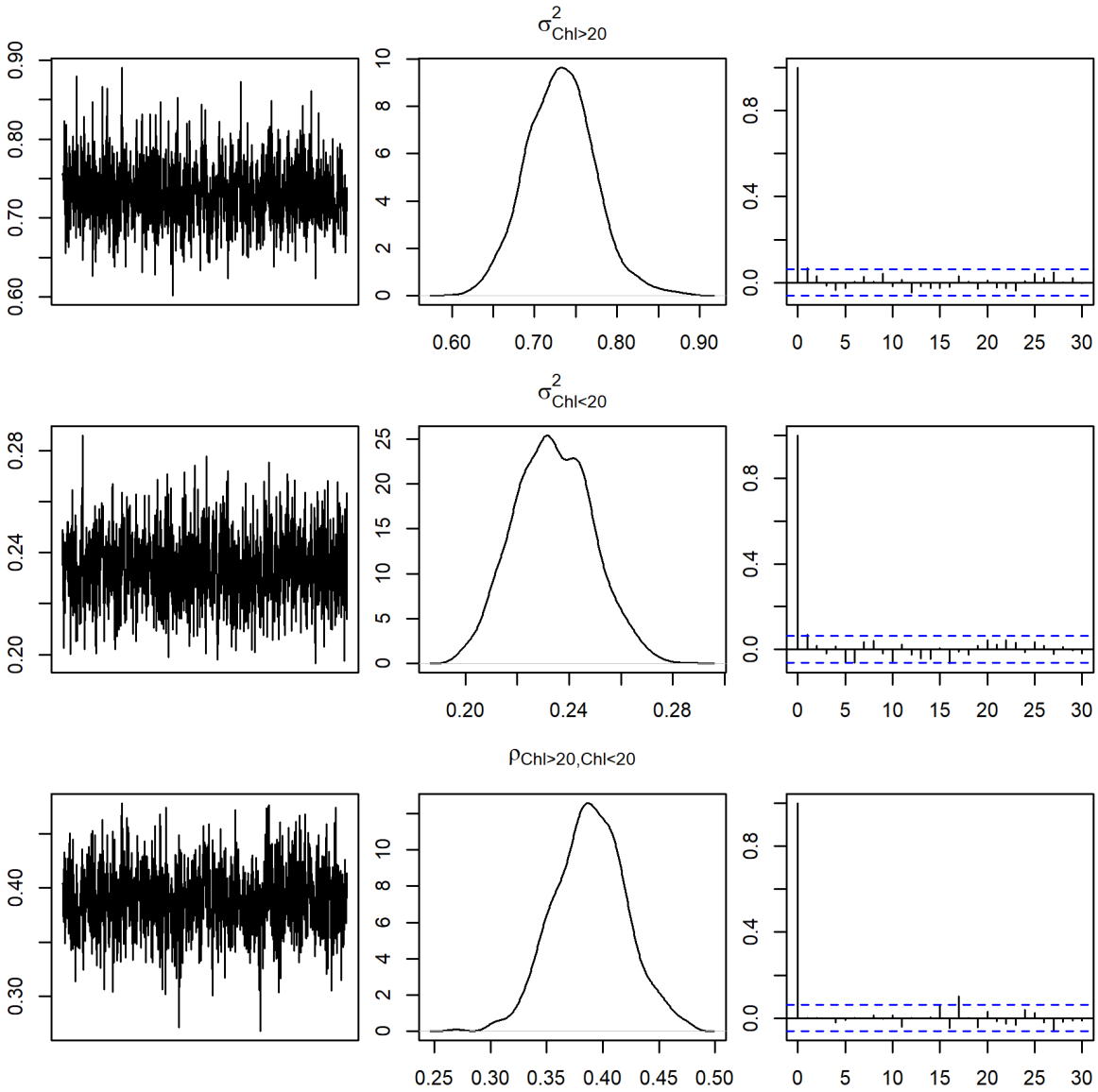

Figure 12. Trace and density plots of the observational variance and correlation of the >20 µm and <20 µm Chl. a series **a.** Observational variance of the >20 µm series. **b.** Observational variance of the <20 µm series **c.** Observational correlation of the >20 µm and <20 µm series.

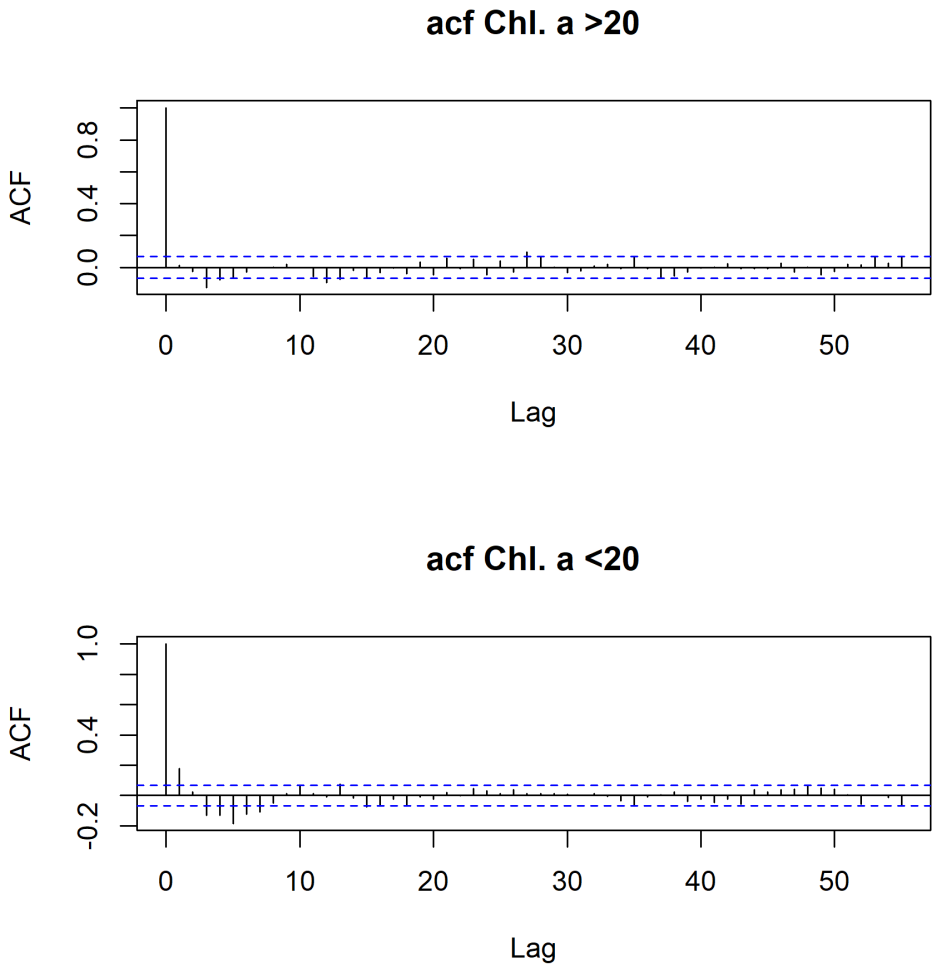

Figure 13. Autocorrelation in the residuals of the dynamic regression model of the >20 µm and <20 µm Chl. a series.

Table 1. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$.

δ_μ	δ_β	DIC	WAIC
0.95	0.95	4112.363	4590.61
0.95	0.99	3477.201	3621.86
0.95	0.999	3328.274	3413.224
0.99	0.95	3860.554	4165.332
0.99	0.99	3417.666	3509.669
0.99	0.999	3314.376	3348.018
0.999	0.95	3998.256	4340.894
0.999	0.99	3509.048	3610.245
0.999	0.999	3370.816	3392.343

Table 2. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0.000	0.891	0.851	0.617	0.919	0.871	0.472	0.827	0.645
0.95_0.99	0.109	0.000	0.460	0.125	0.589	0.403	0.065	0.290	0.060
0.95_0.999	0.149	0.540	0.000	0.153	0.726	0.500	0.073	0.367	0.065
0.99_0.95	0.383	0.875	0.847	0.000	0.964	0.891	0.355	0.810	0.492
0.99_0.99	0.081	0.411	0.274	0.036	0.000	0.218	0.008	0.085	0.004
0.99_0.999	0.129	0.597	0.500	0.109	0.782	0.000	0.065	0.298	0.008
0.999_0.95	0.528	0.935	0.927	0.645	0.992	0.935	0.000	0.923	0.685
0.999_0.99	0.173	0.710	0.633	0.190	0.915	0.702	0.077	0.000	0.048
0.999_0.999	0.355	0.940	0.935	0.508	0.996	0.992	0.315	0.952	0.000

Table 3. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0.000	0.552	0.581	0.544	0.548	0.593	0.520	0.589	0.605
0.95_0.99	0.448	0.000	0.560	0.540	0.540	0.573	0.480	0.577	0.569
0.95_0.999	0.419	0.440	0.000	0.492	0.528	0.573	0.452	0.569	0.560
0.99_0.95	0.456	0.460	0.508	0.000	0.492	0.548	0.480	0.556	0.548
0.99_0.99	0.452	0.460	0.472	0.508	0.000	0.552	0.492	0.504	0.565
0.99_0.999	0.407	0.427	0.427	0.452	0.448	0.000	0.407	0.472	0.464
0.999_0.95	0.480	0.520	0.548	0.520	0.508	0.593	0.000	0.581	0.597
0.999_0.99	0.411	0.423	0.431	0.444	0.496	0.528	0.419	0.000	0.544
0.999_0.999	0.395	0.431	0.440	0.452	0.435	0.536	0.403	0.456	0.000

Table 4. Type 1 RMSE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0.000	0.613	0.742	0.484	0.754	0.847	0.565	0.714	0.911
0.95_0.99	0.387	0.000	0.593	0.399	0.605	0.798	0.444	0.601	0.819
0.95_0.999	0.258	0.407	0.000	0.319	0.569	0.722	0.339	0.532	0.819
0.99_0.95	0.516	0.601	0.681	0.000	0.718	0.827	0.548	0.766	0.867
0.99_0.99	0.246	0.395	0.431	0.282	0.000	0.621	0.319	0.464	0.637
0.99_0.999	0.153	0.202	0.278	0.173	0.379	0.000	0.210	0.319	0.504
0.999_0.95	0.435	0.556	0.661	0.452	0.681	0.790	0.000	0.702	0.863
0.999_0.99	0.286	0.399	0.468	0.234	0.536	0.681	0.298	0.000	0.702
0.999_0.999	0.089	0.181	0.181	0.133	0.363	0.496	0.137	0.298	0.000

Table 5. Type 2 RMSE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0.000	0.944	0.931	0.746	0.895	0.911	0.649	0.847	0.859
0.95_0.99	0.056	0.000	0.706	0.395	0.782	0.831	0.363	0.641	0.730
0.95_0.999	0.069	0.294	0.000	0.375	0.669	0.750	0.319	0.581	0.685
0.99_0.95	0.254	0.605	0.625	0.000	0.903	0.891	0.431	0.738	0.786
0.99_0.99	0.105	0.218	0.331	0.097	0.000	0.694	0.198	0.460	0.565
0.99_0.999	0.089	0.169	0.250	0.109	0.306	0.000	0.177	0.423	0.540
0.999_0.95	0.351	0.637	0.681	0.569	0.802	0.823	0.000	0.940	0.944
0.999_0.99	0.153	0.359	0.419	0.262	0.540	0.577	0.060	0.000	0.778
0.999_0.999	0.141	0.270	0.315	0.214	0.435	0.460	0.056	0.222	0.000

Table 6. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$, and practical discounting was used.

δ_μ	δ_β	DIC	WAIC
0.95	0.95	3922	4279
0.95	0.99	3467	3602
0.95	0.999	3337	3414
0.99	0.95	3755	4020
0.99	0.99	3420	3493
0.99	0.999	3317	3352
0.999	0.95	3871	4151
0.999	0.99	3508	3605
0.999	0.999	3378	3402

Table 7. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$ and practical discounting was used. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0	0.808	0.788	0.536	0.888	0.776	0.36	0.7	0.344
0.95_0.99	0.192	0	0.428	0.192	0.52	0.328	0.064	0.168	0.036
0.95_0.999	0.212	0.572	0	0.2	0.568	0.3	0.068	0.192	0.04
0.99_0.95	0.464	0.808	0.8	0	0.88	0.768	0.304	0.644	0.308
0.99_0.99	0.112	0.48	0.432	0.12	0	0.22	0.02	0.1	0
0.99_0.999	0.224	0.672	0.7	0.232	0.78	0	0.056	0.28	0.008
0.999_0.95	0.64	0.936	0.932	0.696	0.98	0.944	0	0.86	0.516
0.999_0.99	0.3	0.832	0.808	0.356	0.9	0.72	0.14	0	0.044
0.999_0.999	0.656	0.964	0.96	0.692	1	0.992	0.484	0.956	0

Table 8. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$ and practical discounting was used. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0	0.903226	0.854839	0.725806	0.875	0.883065	0.58871	0.758065	0.814516
0.95_0.99	0.096774	0	0.693548	0.366935	0.745968	0.810484	0.375	0.552419	0.657258
0.95_0.999	0.145161	0.306452	0	0.310484	0.608871	0.709677	0.350806	0.5	0.600806
0.99_0.95	0.274194	0.633065	0.689516	0	0.895161	0.907258	0.447581	0.697581	0.778226
0.99_0.99	0.125	0.254032	0.391129	0.104839	0	0.681452	0.209677	0.467742	0.560484
0.99_0.999	0.116935	0.189516	0.290323	0.092742	0.318548	0	0.201613	0.407258	0.479839
0.999_0.95	0.41129	0.625	0.649194	0.552419	0.790323	0.798387	0	0.923387	0.935484
0.999_0.99	0.241935	0.447581	0.5	0.302419	0.532258	0.592742	0.076613	0	0.75
0.999_0.999	0.185484	0.342742	0.399194	0.221774	0.439516	0.520161	0.064516	0.25	0

Table 9. Type 1 RMSE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$ and practical discounting was used. Each cell is the probability that the RMSE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0	0.68	0.876	0.512	0.684	0.908	0.532	0.708	0.928
0.95_0.99	0.32	0	0.716	0.364	0.536	0.868	0.392	0.536	0.836
0.95_0.999	0.124	0.284	0	0.176	0.356	0.772	0.204	0.356	0.764
0.99_0.95	0.488	0.636	0.824	0	0.66	0.892	0.524	0.724	0.888
0.99_0.99	0.316	0.464	0.644	0.34	0	0.784	0.38	0.5	0.772
0.99_0.999	0.092	0.132	0.228	0.108	0.216	0	0.124	0.172	0.328
0.999_0.95	0.468	0.608	0.796	0.476	0.62	0.876	0	0.644	0.92
0.999_0.99	0.292	0.464	0.644	0.276	0.5	0.828	0.356	0	0.78
0.999_0.999	0.072	0.164	0.236	0.112	0.228	0.672	0.08	0.22	0

Table 10. Type 2 RMSE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.999$, $\delta_\beta = 0.99$ and practical discounting was used. Each cell is the probability that the RMSE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.95_0.95	0	0.903226	0.854839	0.725806	0.875	0.883065	0.58871	0.758065	0.814516
0.95_0.99	0.096774	0	0.693548	0.366935	0.745968	0.810484	0.375	0.552419	0.657258
0.95_0.999	0.145161	0.306452	0	0.310484	0.608871	0.709677	0.350806	0.5	0.600806
0.99_0.95	0.274194	0.633065	0.689516	0	0.895161	0.907258	0.447581	0.697581	0.778226
0.99_0.99	0.125	0.254032	0.391129	0.104839	0	0.681452	0.209677	0.467742	0.560484
0.99_0.999	0.116935	0.189516	0.290323	0.092742	0.318548	0	0.201613	0.407258	0.479839
0.999_0.95	0.41129	0.625	0.649194	0.552419	0.790323	0.798387	0	0.923387	0.935484
0.999_0.99	0.241935	0.447581	0.5	0.302419	0.532258	0.592742	0.076613	0	0.75
0.999_0.999	0.185484	0.342742	0.399194	0.221774	0.439516	0.520161	0.064516	0.25	0

Table 11. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$ and standard discounting was used.

δ_μ	δ_β	DIC	WAIC
0.95	0.95	4878.65	5805.761
0.95	0.99	3884.765	4256.285
0.95	0.999	3760.02	4054.729
0.99	0.95	4457.628	5061.727
0.99	0.99	3759.542	3966.759
0.99	0.999	3659.232	3816.543

Table 12. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.752	0.524	0.1	0.084	0.02
0.95_0.99	0.248	0	0.212	0.012	0.004	0.004
0.95_0.999	0.476	0.788	0	0.028	0.004	0.004
0.99_0.95	0.9	0.988	0.972	0	0.424	0.212
0.99_0.99	0.916	0.996	0.996	0.576	0	0.124
0.99_0.999	0.98	0.996	0.996	0.788	0.876	0

Table 13. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0.000	0.614	0.568	0.490	0.552	0.553
0.95_0.99	0.386	0.000	0.488	0.422	0.479	0.500
0.95_0.999	0.432	0.512	0.000	0.451	0.497	0.506
0.99_0.95	0.510	0.578	0.549	0.000	0.602	0.591
0.99_0.99	0.448	0.521	0.503	0.398	0.000	0.512
0.99_0.999	0.447	0.500	0.494	0.409	0.488	0.000

Table 14. Type 1 RMSE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.652	0.58	0.524	0.612	0.524
0.95_0.99	0.348	0	0.428	0.38	0.492	0.42
0.95_0.999	0.42	0.572	0	0.404	0.568	0.456
0.99_0.95	0.476	0.62	0.596	0	0.632	0.54
0.99_0.99	0.388	0.508	0.432	0.368	0	0.428
0.99_0.999	0.476	0.58	0.544	0.46	0.572	0

Table 15. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0.000	0.903	0.879	0.645	0.806	0.823
0.95_0.99	0.097	0.000	0.581	0.355	0.621	0.641
0.95_0.999	0.121	0.419	0.000	0.347	0.597	0.637
0.99_0.95	0.355	0.645	0.653	0.000	0.895	0.895
0.99_0.99	0.194	0.379	0.403	0.105	0.000	0.617
0.99_0.999	0.177	0.359	0.363	0.105	0.383	0.000

Table 16. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit with practical discounting to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$.

δ_μ	δ_β	DIC	WAIC
0.95	0.95	90139.9	150694.7
0.95	0.99	72077.36	122939.5
0.95	0.999	34421.1	51847.42
0.99	0.95	58866.3	90461.28
0.99	0.99	36624.13	59817.27
0.99	0.999	50947.16	96132.38

Table 17. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit with practical discounting to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.616	0.324	0.064	0.008	0
0.95_0.99	0.384	0	0.14	0.004	0.004	0.004
0.95_0.999	0.676	0.86	0	0.028	0.004	0.004
0.99_0.95	0.936	0.996	0.972	0	0.284	0.136
0.99_0.99	0.992	0.996	0.996	0.716	0	0.108
0.99_0.999	1	0.996	0.996	0.864	0.892	0

Table 18. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit with practical discounting to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.568213	0.563438	0.466576	0.508186	0.523192
0.95_0.99	0.431787	0	0.52251	0.444065	0.497271	0.510232
0.95_0.999	0.436562	0.47749	0	0.444748	0.45839	0.489086
0.99_0.95	0.533424	0.555935	0.555252	0	0.593452	0.579809
0.99_0.99	0.491814	0.502729	0.54161	0.406548	0	0.527967
0.99_0.999	0.476808	0.489768	0.510914	0.420191	0.472033	0

Table 19. Type 1 RMSE calculated for each model with fixed discount factors for μ and β, fit with practical discounting to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.472	0.524	0.496	0.564	0.512
0.95_0.99	0.528	0	0.532	0.504	0.564	0.548
0.95_0.999	0.476	0.468	0	0.452	0.52	0.512
0.99_0.95	0.504	0.496	0.548	0	0.568	0.564
0.99_0.99	0.436	0.436	0.48	0.432	0	0.488
0.99_0.999	0.488	0.452	0.488	0.436	0.512	0

Table 20. Type 2 RMSFE calculated for each model with fixed discount factors for μ and β, fit with practical discounting to simulated data with missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.95_0.95	0	0.858871	0.842742	0.58871	0.745968	0.794355
0.95_0.99	0.141129	0	0.653226	0.306452	0.600806	0.633065
0.95_0.999	0.157258	0.346774	0	0.245968	0.548387	0.572581
0.99_0.95	0.41129	0.693548	0.754032	0	0.907258	0.887097
0.99_0.99	0.254032	0.399194	0.451613	0.092742	0	0.620968
0.99_0.999	0.205645	0.366935	0.427419	0.112903	0.379032	0

Table 21. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to simulated data with no missingness where the data generation model was $\delta_\mu = 0.95$, $\delta_\beta = 0.99$.

δ_μ	δ_β	DIC	WAIC
0.95	0.95	90139.9	150694.7
0.95	0.99	72077.36	122939.5
0.95	0.999	34421.1	51847.42
0.99	0.95	58866.3	90461.28
0.99	0.99	36624.13	59817.27
0.99	0.999	50947.16	96132.38

Table 22. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to the true data with artificial missingness.

δ_μ	δ_β	DIC	WAIC
0.9	0.95	4005.839	4920.432
0.9	0.99	3226.983	3700
0.9	0.999	3049.024	3407.732
0.95	0.95	3856.14	4628.743
0.95	0.99	3223.991	3595.045
0.95	0.999	3064.359	3370.367
0.99	0.95	3507.952	3997.436
0.99	0.99	3163.408	3434.308
0.99	0.999	3066.24	3286.251

Table 23. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit to the true data with artificial missingness. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.9_0.95	0.9_0.99	0.9_0.999	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.9_0.95	0	0.7	0.732	0.656	0.812	0.86	0.816	0.932	0.96
0.9_0.99	0.3	0	0.512	0.504	0.668	0.712	0.668	0.848	0.888
0.9_0.999	0.268	0.488	0	0.508	0.66	0.704	0.688	0.884	0.888
0.95_0.95	0.344	0.496	0.492	0	0.608	0.668	0.68	0.848	0.84
0.95_0.99	0.188	0.332	0.34	0.392	0	0.608	0.544	0.748	0.784
0.95_0.999	0.14	0.288	0.296	0.332	0.392	0	0.468	0.716	0.708
0.99_0.95	0.184	0.332	0.312	0.32	0.456	0.532	0	0.7	0.688
0.99_0.99	0.068	0.152	0.116	0.152	0.252	0.284	0.3	0	0.488
0.99_0.999	0.04	0.112	0.112	0.16	0.216	0.292	0.312	0.512	0

Table 24. Type 1 RMSE calculated for each model with fixed discount factors for μ and β, fit to the true data with artificial missingness. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.9_0.95	0.9_0.99	0.9_0.999	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999
0.9_0.95	0	0.7	0.696	0.588	0.776	0.776	0.724	0.88	0.876
0.9_0.99	0.3	0	0.564	0.4	0.612	0.664	0.524	0.7	0.796
0.9_0.999	0.304	0.436	0	0.352	0.544	0.584	0.452	0.648	0.7
0.95_0.95	0.412	0.6	0.648	0	0.728	0.764	0.644	0.804	0.872
0.95_0.99	0.224	0.388	0.456	0.272	0	0.56	0.372	0.592	0.68
0.95_0.999	0.224	0.336	0.416	0.236	0.44	0	0.332	0.512	0.64
0.99_0.95	0.276	0.476	0.548	0.356	0.628	0.668	0	0.712	0.772
0.99_0.99	0.12	0.3	0.352	0.196	0.408	0.488	0.288	0	0.612
0.99_0.999	0.124	0.204	0.3	0.128	0.32	0.36	0.228	0.388	0

Table 25. DIC and WAIC calculated for each model with fixed discount factors for μ and β, fit to the true data with no artificial missingness.

δ_μ	δ_β	DIC	WAIC
0.8	0.8	1344299	2249522
0.8	0.85	589180.7	1153076
0.8	0.9	649766.1	845180.3
0.85	0.8	885615.1	1246872
0.85	0.85	794300.9	1276092
0.85	0.9	368430.2	491328.9
0.9	0.8	633079.4	1105831
0.9	0.85	398712.5	621549.6
0.9	0.9	282162	427507
0.95	0.95	51921.3	65436.39
0.95	0.99	42697.75	76718.33
0.95	0.999	34009.99	63533.18
0.99	0.95	77564.23	82744.16
0.99	0.99	37995.08	54691.75
0.99	0.999	37224.55	54232.08
0.999	0.95	67989.94	160291.4
0.999	0.99	37867.53	85436.49
0.999	0.999	28503.83	58615.79

Table 26. Type 1 RMSFE calculated for each model with fixed discount factors for μ and β, fit to the true data with no artificial missingness. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.8_0.8	0.8_0.85	0.8_0.9	0.85_0.8	0.85_0.85	0.85_0.9	0.9_0.8	0.9_0.85	0.9_0.9	0.95_0.95	0.95_0.99	0.95_0.999	0.99_0.95	0.99_0.99	0.99_0.999	0.999_0.95	0.999_0.99	0.999_0.999
0.8_0.8	0	0.788	0.952	0.372	0.664	0.828	0.232	0.376	0.512	0.148	0.112	0.112	0.044	0.016	0.02	0.02	0.012	0.008
0.8_0.85	0.212	0	0.776	0.108	0.328	0.468	0.06	0.1	0.148	0.008	0.012	0.008	0	0	0	0	0	0
0.8_0.9	0.048	0.224	0	0.032	0.092	0.192	0.012	0.016	0.04	0.008	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
0.85_0.8	0.628	0.892	0.968	0	0.756	0.92	0.344	0.556	0.652	0.204	0.152	0.156	0.068	0.036	0.04	0.032	0.012	0.012
0.85_0.85	0.336	0.672	0.908	0.244	0	0.764	0.112	0.212	0.32	0.04	0.024	0.02	0.004	0	0	0	0	0
0.85_0.9	0.172	0.532	0.808	0.08	0.236	0	0.032	0.068	0.116	0.004	0.004	0.004	0	0	0	0	0	0
0.9_0.8	0.768	0.94	0.988	0.656	0.888	0.968	0	0.756	0.824	0.44	0.34	0.304	0.148	0.08	0.06	0.044	0.028	0.024
0.9_0.85	0.624	0.9	0.984	0.444	0.788	0.932	0.244	0	0.62	0.184	0.112	0.076	0.02	0.008	0.004	0.004	0	0.004
0.9_0.9	0.488	0.852	0.96	0.348	0.68	0.884	0.176	0.38	0	0.064	0.024	0.012	0.004	0	0	0	0	0
0.95_0.95	0.852	0.992	0.992	0.796	0.96	0.996	0.56	0.816	0.936	0	0.352	0.248	0.02	0.004	0	0	0	0
0.95_0.99	0.888	0.988	0.996	0.848	0.976	0.996	0.66	0.888	0.976	0.648	0	0.372	0.028	0	0	0	0	0
0.95_0.999	0.888	0.992	0.996	0.844	0.98	0.996	0.696	0.924	0.988	0.752	0.628	0	0.064	0.004	0	0	0	0
0.99_0.95	0.956	1	0.996	0.932	0.996	1	0.852	0.98	0.996	0.98	0.972	0.936	0	0.172	0.064	0.036	0.004	0
0.99_0.99	0.984	1	0.996	0.964	1	1	0.92	0.992	1	0.996	1	0.996	0.828	0	0.36	0.26	0.024	0.008
0.99_0.999	0.98	1	0.996	0.96	1	1	0.94	0.996	1	1	1	1	0.936	0.64	0	0.348	0.044	0.008
0.999_0.95	0.98	1	0.996	0.968	1	1	0.956	0.996	1	1	1	1	0.964	0.74	0.652	0	0.144	0.08
0.999_0.99	0.988	1	0.996	0.988	1	1	0.972	1	1	1	1	1	0.996	0.976	0.956	0.856	0	0.3
0.999_0.999	0.992	1	0.996	0.988	1	1	0.976	0.996	1	1	1	1	1	0.992	0.992	0.92	0.7	0

Table 27. Type 1 RMSFE calculated for each model with fixed, equal discount factors for μ and β, fit to the true data with no artificial missingness. Each cell is the probability that the RMSFE of the row index exceeds that of the column index. The optimal model is highlighted in light grey.

	0.8_0.8	0.85_0.85	0.9_0.9	0.95_0.95	0.99_0.99	0.999_0.999
0.8_0.8	0	0.664	0.512	0.148	0.016	0.008
0.85_0.85	0.336	0	0.32	0.04	0	0
0.9_0.9	0.488	0.68	0	0.064	0	0
0.95_0.95	0.852	0.96	0.936	0	0.004	0
0.99_0.99	0.984	1	1	0.996	0	0.008
0.999_0.999	0.992	1	1	1	0.992	0

a.) *Kalman Filter & Smoothing:*

- One-step -ahead predictive distribution of the latent state, $f(\theta_t|y_{1:t-1}) = N(a_t, R_t)$, where:

$$a_t = E(\theta_t|y_{1:t-1}) = G_t m_{t-1},$$
$$R_t = Var((\theta_t|y_{1:t-1}) = G_t C_{t-1} G_t' + W_t$$

- One-step -ahead predictive distribution of the observation, $f(Y_t|y_{1:t-1}) = N(f_t, Q_t)$, where:

$$ft = E(Y_t|y_{1:t-1}) = F_t a_t$$
$$Qt = Var(Y_t|y_{1:t-1}) = F_t R_t F_t' + V_t$$

- The filtered distribution of the latent state, $f(\theta_t|y_{1:t}) = N(m_t, C_t)$, where:

$$m_t = E(\theta_t|y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t,$$
$$C_t = Var(\theta_t|y_{1:t}) = R_t - R_t F_t' Q_t' F_t R_t,$$
$$e_t = Y_t - ft$$

- The smoothed distribution of the latent state, $f(\theta_t|y_{1:T}) = N(s_t, S_t)$, where:

$$s_t = E(\theta_t|y_{1:T}) = m_t + C_t G_{t+1}' R_{t+1}' (s_{t+1} - a_{t+1}),$$
$$S_t = C_t - C_t G_{t+1}' R_{t+1}^{-1} (R_{t+1}) R_{t+1}^{-1} G_{t+1} C_t$$

b.) *Semi-conjugacy of inverse-gamma+ inverse-Wishart*

$$p(x|\mu, \sigma^2) \propto (\sigma^2)^{\frac{1}{2}} e^{-\frac{\left(\sum_{i=1}^n x_i - \mu\right)^2}{2\sigma^2}}$$
$$p(\sigma^2|\alpha, \beta) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{\frac{-\beta}{\sigma^2}}$$
$$p(\sigma^2|x, \mu, \alpha, \beta) \propto p(x|\mu, \sigma^2) p(\sigma^2|\alpha, \beta)$$
$$\propto (\sigma^2)^{\left(-\frac{n}{2}\right)} e^{-\frac{\left(\sum_{i=1}^n x_i - \mu\right)^2}{2\sigma^2}} (\sigma^2)^{-\alpha-1} e^{\frac{-\beta}{\sigma^2}}$$
$$= (\sigma^2)^{-\left(\alpha + \frac{n}{2}\right)} e^{-\left(\beta + \frac{\left(\sum_{i=1}^n x_i - \mu\right)^2}{2\sigma^2}\right)}$$

which is the form of an inverse-gamma with parameters $\left(\alpha + \frac{n}{2}\right)$ and $\left(\beta + \frac{1}{2}(\sum_{i=1}^n x_i - \mu)^2\right)$.

For the inverse-Wishart:

$$p(X|\mu, \sigma) \propto |\Sigma|^{-\frac{n}{2}} e^{-\frac{\sum (X-\mu)' \Sigma^{-1} (X-\mu)}{2}}$$
$$\propto |\Sigma|^{-\frac{n}{2}} e^{-tr\left(\frac{S_\mu \Sigma^{-1}}{2}\right)}$$
$$p(\Sigma|\nu_0, S_0^{-1}) \propto |\Sigma|^{\frac{(\nu_0 + p + 1)}{2}} e^{-\frac{tr\left(S_0 \Sigma^{-1}\right)}{2}}$$
$$p(\Sigma|X, \mu) \propto p(\Sigma|\nu_0, S_0^{-1}) p(X|\mu, \sigma)$$
$$\propto |\Sigma|^{-\frac{n}{2}} e^{-tr\left(\frac{S_\mu \Sigma^{-1}}{2}\right)} |\Sigma|^{\frac{(\nu_0 + p + 1)}{2}} e^{-\frac{tr\left(S_0 \Sigma^{-1}\right)}{2}}$$

$$\propto |\Sigma|^{\frac{-\nu_0+n+p+1}{2}} e^{-\frac{tr(S_0+S_\mu)\Sigma^{-1}}{2}}$$

Which is the form of an inverse-Wishart with parameters $(\nu_0 + n)$ and $\left(S_0 + S_\mu\right)^{-1}$

c.) *Discount factors*

Representation as loss of information between time steps:

$$R_t = P_t + W_t,$$

$$W_t = \frac{1-\delta}{\delta} P_t$$

$$P_t = G_t C_{t-1} G_t' = V(G_t \theta_{t-1} | D_{t-1})$$

Sampling discount factors from a discrete probability distribution with probabilities equal to:

$$p\left(\delta_{\mu,i}, \delta_{\beta,i} \middle| \theta_{1:T}, C_{1:T}, G_{1:T}\right) = \frac{\prod_{t=1}^{T} N\left(\theta_t \middle| G_t, W_t\left(\delta_{\mu,i}, \delta_{\beta,i}\right)\right)}{\sum_{i=1}^{k} \prod_{t=1}^{T} N\left(\theta_t \middle| G_t, W_t\left(\delta_{\mu,i}, \delta_{\beta,i}\right)\right)}$$

BIBLIOGRAPHY

Aguilar, Omar, and Mike West. 2000. “Bayesian Dynamic Factor Models and Portfolio Allocation.” *Journal of Business and Economic Statistics* 18 (3): 338–57. https://doi.org/10.1080/07350015.2000.10524875.

Ameen, J. R. M., and P. J. Harrison. 1984. “Discount Weighted Estimation.” *Journal of Forecasting* 3 (3): 285–96. https://doi.org/10.1002/for.3980030306.

Antoine, David, Jean-Michel André, and André Morel. 1996. “Oceanic Primary Production: 2. Estimation at Global Scale from Satellite (Coastal Zone Color Scanner) Chlorophyll.” *Global Biogeochemical Cycles* 10 (1): 57–69. https://doi.org/10.1029/95GB02832.

Arhonditsis, G. B., H. W. Paerl, L. M. Valdes-Weaver, C. A. Stow, L. J. Steinberg, and K. H. Reckhow. 2007. “Application of Bayesian Structural Equation Modeling for Examining Phytoplankton Dynamics in the Neuse River Estuary (North Carolina, USA).” *Estuarine, Coastal and Shelf Science* 72 (1–2): 63–80. https://doi.org/10.1016/j.ecss.2006.09.022.

Atkinson, David, Benjamin J Ciotti, and David J S Montagnes. 2003. “Protists Decrease in Size Linearly with Temperature: Ca. 2.5% °C.” https://doi.org/10.1098/rspb.2003.2538.

Barton, AD, F González Taboada, A Atkinson, CE Widdicombe, and CA Stock. 2020. “Integration of Temporal Environmental Variation by the Marine Plankton Community.” *Marine Ecology Progress Series* 647 (August): 1–16. https://doi.org/10.3354/meps13432.

Beardall, John, Drew Allen, Jason Bragg, Zoe V. Finkel, Kevin J. Flynn, Antonietta

Quigg, T. Alwyn V. Rees, Anthony Richardson, and John A. Raven. 2009. "Allometry and Stoichiometry of Unicellular, Colonial and Multicellular Phytoplankton." *New Phytologist* 181 (2): 295–309. https://doi.org/10.1111/j.1469-8137.2008.02660.x.

Behrenfeld, Michael J, Robert T O'malley, David A Siegel, Charles R Mcclain, Jorge L Sarmiento, Gene C Feldman, Allen J Milligan, Paul G Falkowski, Ricardo M Letelier, and Emmanuel S Boss. 2006. "Climate-Driven Trends in Contemporary Ocean Productivity." *Marine Sciences* 444: 4469–5741. https://doi.org/10.1038/nature05317.

Borkman, D. G., and T. J. Smayda. 2009. "Gulf Stream Position and Winter NAO as Drivers of Long-Term Variations in the Bloom Phenology of the Diatom Skeletonema Costatum "Species-Complex" in Narragansett Bay, RI, USA." *Journal of Plankton Research* 31 (11): 1407–25. https://doi.org/10.1093/plankt/fbp072.

Borkman, David G., and Theodore J. Smayda. 2009. "Gulf Stream Position and Winter NAO as Drivers of Long-Term Variations in the Bloom Phenology of the Diatom Skeletonema Costatum 'Species- Complex' in Narragansett Bay, RI, USA." *Journal of Plankton Research* 31 (11): 1407–25. https://doi.org/10.1093/plankt/fbp072.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. 2006. "Deviance information criteria for missing data models". Bayesian analysis, 1(4), 651-673.

Chisholm, Sallie W. 1992. "Phytoplankton Size." In *Primary Productivity and Biogeochemical Cycles in the Sea*, 213–37. Boston, MA: Springer US.

https://doi.org/10.1007/978-1-4899-0762-2_12.

Cloern, J. E., S. Q. Foster, and A. E. Kleckner. 2014. “Phytoplankton Primary Production in the World’s Estuarine-Coastal Ecosystems.” *Biogeosciences* 11 (9): 2477–2501. https://doi.org/10.5194/bg-11-2477-2014.

Collos, Yves. 1986. “Time-Lag Algal Growth Dynamics: Biological Constraints on Primary Production in Aquatic Environments.” *Marine Ecology Progress Series* 33: 193–206.

DEM, RI. 2005. *Plan for Managing Nutrient Loadings to Rhode Island Waters*.

Dortch, Quay. 1990. “The Interaction between Ammonium and Nitrate Uptake in Phytoplankton.” *Marine Ecology Progress Series* 61: 183–201.

Dugdale, Richard C., Frances P. Wilkerson, Victoria E. Hogue, and Albert Marchi. 2007. “The Role of Ammonium and Nitrate in Spring Bloom Development in San Francisco Bay.” *Estuarine, Coastal and Shelf Science* 73 (1–2): 17–29. https://doi.org/10.1016/j.ecss.2006.12.008.

Durbin, E. G., R. W. Krawiec, and T. J. Smayda. 1975. “Seasonal Studies on the Relative Importance of Different Size Fractions of Phytoplankton in Narragansett Bay (USA).” *Marine Biology* 32 (3): 271–87. https://doi.org/10.1007/BF00399206.

Falkowski, Paul G. 1994. “The Role of Phytoplankton Photosynthesis in Global Biogeochemical Cycles.” *Photosynthesis Research* 39 (3): 235–58. https://doi.org/10.1007/BF00014586.

Field, Christopher B., Michael J. Behrenfeld, James T. Randerson, and Paul Falkowski. 1998. “Primary Production of the Biosphere: Integrating Terrestrial

and Oceanic Components." *Science (New York, N.Y.)* 281 (5374): 237–40. https://doi.org/10.1126/SCIENCE.281.5374.237.

Finkel, Zoe V. 2007. "Does Phytoplankton Cell Size Matter? The Evolution of Modern Marine Food Webs." In *Evolution of Primary Producers in the Sea*, 333–50. Academic Press.

Fulweiler, R. W., A. J. Oczkowski, K. M. Miller, C. A. Oviatt, and M. E.Q. Pilson. 2015. "Whole Truths vs. Half Truths - And a Search for Clarity in Long-Term Water Temperature Records." *Estuarine, Coastal and Shelf Science* 157 (May): A1–6. https://doi.org/10.1016/j.ecss.2015.01.021.

Furnas, Miles J. 1983. "Nitrogen Dynamics in Lower Narragansett Bay, Rhode Island. I. Uptake by Size-Fractionated Phytoplankton Populations." *Journal of Plankton Research* 5 (5): 657–76. https://doi.org/10.1093/plankt/5.5.657.

Gelman, Andrew, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton: Taylor & Francis.

Glibert, Patricia M., and Chris Garside. 1992. "Diel Variability in Nitrogenous Nutrient Uptake by Phytoplankton in the Chesapeake Bay Plume." *Journal of Plankton Research* 14 (2): 271–88. https://doi.org/10.1093/plankt/14.2.271.

Goeyens, L., R. T.P. De Vries, J. F. Bakker, and W. Helder. 1987. "An Experiment on the Relative Importance of Denitrification, Nitrate Reduction and Ammonification in Coastal Marine Sediment." *Netherlands Journal of Sea Research* 21 (3): 171–75. https://doi.org/10.1016/0077-7579(87)90010-X.

Hilligsøe, Karen Marie, Katherine Richardson, Jørgen Bendtsen, Lise-Lotte Sørensen, Torkel Gissel Nielsen, and Maren Moltke Lyngsgaard. 2011. "Linking

Phytoplankton Community Size Composition with Temperature, Plankton Food Web Structure and Sea–Air CO2 Flux." *Deep Sea Research Part I: Oceanographic Research Papers* 58 (8): 826–38. https://doi.org/10.1016/J.DSR.2011.06.004.

Hurrell, James W. 1995. "Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation." *Science* 269 (5224): 676–79. https://doi.org/10.1126/science.269.5224.676.

Irigoien, X., K.J. Flynn, and R.P. Harris. 2005. "Phytoplankton Blooms: A 'Loophole' in Microzooplankton Grazing Impact? | Journal of Plankton Research | Oxford Academic." *Journal of Plankton Research* 27 (4): 313–21. https://academic.oup.com/plankt/article/27/4/313/1508420.

Irwin, Andrew J., Zoe V. Finkel, Oscar M. E. Schofield, and Paul G. Falkowski. 2006. "Scaling-up from Nutrient Physiology to the Size-Structure of Phytoplankton Communities." *Journal of Plankton Research* 28 (5): 459–71. https://doi.org/10.1093/plankt/fbi148.

Jones, Emlyn, John Parslow, and Lawrence Murray. 2010. "A Bayesian Approach to State and Parameter Estimation in a Phytoplankton-Zooplankton Model." *Article in Australian Meteorological and Oceanographic Journal* 59: 7–16. https://doi.org/10.22499/2.5901.003.

Kalman, R. E. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Fluids Engineering, Transactions of the ASME* 82 (1): 35–45. https://doi.org/10.1115/1.3662552.

Karentz, Deneb, and Theodore J Smayda. 1984. "Temperature and Seasonal

Occurrence Patterns of 30 Dominant Phytoplankton Species in Narragansett Bay over a 22-Year Period (1959-1980).” *Marine Ecology Progress Series* 18: 277–93. https://www.int-res.com/articles/meps/18/m018p277.pdf.

Katz, Richard W. 1988. “Use of Cross Correlations in the Search for Teleconnections.” *Journal of Climatology* 8 (3): 241–53. https://doi.org/10.1002/joc.3370080303.

Key, Tim, Avery McCarthy, Douglas A. Campbell, Christophe Six, Suzanne Roy, and Zoe V. Finkel. 2010. “Cell Size Trade-Offs Govern Light Exploitation Strategies in Marine Phytoplankton.” *Environmental Microbiology* 12 (1): 95–104. https://doi.org/10.1111/j.1462-2920.2009.02046.x.

Lawerence, C, and S Menden-Deuer. 2012. “Drivers of Protistan Grazing Pressure: Seasonal Signals of Plankton Community Composition and Environmental Conditions.” *Marine Ecology Progress Series* 459 (July): 39–52. https://doi.org/10.3354/meps09771.

Laybourn-Parry, Johanna, and Jacqueline Parry. 2000. *Flagellates: Unity, Diversity and Evolution* . Edited by Barry Leadbeater and John Green. 1st ed. London: Taylor and Francis. https://books.google.com/books?hl=en&lr=&id=GURZDwAAQBAJ&oi=fnd&pg=PA216&dq=microbial+loop+nutrient+regeneration+zooplankton&ots=V1FliINah6&sig=H9w4UKf2JAVpRWHRyCAZRHq6J5w#v=onepage&q=microbial loop nutrient regeneration zooplankton&f=false.

López-Urrutia, Angel, Elena San Martin, Roger P Harris, and Xabier Irigoien. 2011. “Scaling the Metabolic Balance of the Oceans´Angel.” Vol. 103.

www.pnas.orgcgidoi10.1073pnas.0601137103.

Marra, John, Charles C. Trees, and John E. O'Reilly. 2007. "Phytoplankton Pigment Absorption: A Strong Predictor of Primary Productivity in the Surface Ocean." *Deep-Sea Research Part I: Oceanographic Research Papers* 54 (2): 155–63. https://doi.org/10.1016/j.dsr.2006.12.001.

Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.). 2018. "Summary for Policymakers — Global Warming of 1.5 ºAn IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate PovertyC." https://www.ipcc.ch/sr15/chapter/spm/.

Mei, Zhi-Ping, Zoe V Finkel, and Andrew J Irwin. 2009. "Light and Nutrient Availability Affect the Size-Scaling of Growth in Phytoplankton." *Journal of Theoretical Biology* 259: 582–88. https://doi.org/10.1016/j.jtbi.2009.04.018.

Miklasz, Kevin A., and Mark W. Denny. 2010. "Diatom Sinkings Speeds: Improved Predictions and Insight from a Modified Stokes' Law." *Limnology and Oceanography* 55 (6): 2513–25. https://doi.org/10.4319/lo.2010.55.6.2513.

Morán, Xosé Anxelu G., Ángel López-Urrutia, Alejandra Calvo-Díaz, And William K. W. Li. 2010. "Increasing Importance of Small Phytoplankton in a Warmer Ocean." *Global Change Biology* 16 (3): 1137–44. https://doi.org/10.1111/j.1365-

2486.2009.01960.x.

Nixon, Scott W., Robinson W. Fulweiler, Betty A. Buckley, Stephen L. Granger, Barbara L. Nowicki, and Kelly M. Henry. 2009. “The Impact of Changing Climate on Phenology, Productivity, and Benthic-Pelagic Coupling in Narragansett Bay.” *Estuarine, Coastal and Shelf Science* 82 (1): 1–18. https://doi.org/10.1016/j.ecss.2008.12.016.

Oviatt, Candace A. 2004. “The Changing Ecology of Temperate Coastal Waters During a Warming Trend 1.” *Estuarine Research Federation Estuaries*. Vol. 27. www.noaa.gov.

Oviatt, Candace, Leslie Smith, Jason Krumholz, Catherine Coupland, Heather Stoffel, Aimee Keller, M. Conor McManus, and Laura Reed. 2017. “Managed Nutrient Reduction Impacts on Nutrient Concentrations, Water Clarity, Primary Production, and Hypoxia in a North Temperate Estuary.” *Estuarine, Coastal and Shelf Science* 199 (December): 25–34. https://doi.org/10.1016/j.ecss.2017.09.026.

Pratt, David M. 1965. “The Winter-Spring Diatom Flowering in Narragansett Bay.” *Limnology and Oceanography*, 173–84. https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1965.10.2.0173.

Probyn, Trevor. 1987. “Ammonium Regeneration by Microplankton in an Upwelling Environment.” *Marine Ecology Progress Series* 6: 53–64.

Prog, Ser, Deneb Karentz, and Theodore J Smayda. 1984. “Temperature and Seasonal Occurrence Patterns of 30 Dominant Phytoplankton Species in Narragansett Bay over a 22-Year Period (1959-1980).” *Marine Ecology Progress Series* 18: 277–93.

Pryor Emily Saarman, Donald. 2007. “Nitrogen Loading From Wastewater Treatment Plants To Upper Narragansett Bay Recommended Citation.” http://digitalcommons.uri.edu/nbcollection/2.

Rodriguez, Abel, and Gavino Puggioni. 2010. “Mixed Frequency Models: Bayesian Approaches to Estimation and Prediction.” *International Journal of Forecasting* 26 (2): 293–311. https://doi.org/10.1016/j.ijforecast.2010.01.009.

Roesch, Angi, and Harald Schmidbauer. 2018. “WaveletComp: Computational Wavelet Analysis.” CRAN. https://cran.r-project.org/package=WaveletComp.

Sakshaug, Egil. 1977. “Limiting Nutrients and Maximum Growth Rates for Diatoms in Narragansett Bay.” *Journal of Experimental Marine Biology and Ecology* 28 (2): 109–23. https://doi.org/10.1016/0022-0981(77)90110-1.

Sarmiento, J. L., R. Slater, R. Barber, L. Bopp, S. C. Doney, A. C. Hirst, J. Kleypas, et al. 2004. “Response of Ocean Ecosystems to Climate Warming.” *Global Biogeochemical Cycles* 18 (3): n/a-n/a. https://doi.org/10.1029/2003GB002134.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. 2002. “Bayesian Measures of Model Complexity and Fit.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64 (4): 583–616. https://doi.org/10.1111/1467-9868.00353.

Sprules, W. G., and M. Munawar. 1986. “Plankton Size Spectra in Relation to Ecosystem Productivity, Size, and Perturbation.” *Canadian Journal of Fisheries and Aquatic Sciences* 43 (9): 1789–94. https://doi.org/10.1139/f86-222.

Steffen, W., K. Richardson, J. Rockstrom, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, et al. 2015. “Planetary Boundaries: Guiding Human Development on a

Changing Planet." *Science* 347 (6223): 1259855–1259855. https://doi.org/10.1126/science.1259855.

Steinberg, Deborah K., and Michael R. Landry. 2017. "Zooplankton and the Ocean Carbon Cycle." *Annual Review of Marine Science* 9 (1): 413–44. https://doi.org/10.1146/annurev-marine-010814-015924.

Watanabe, Sumio. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research*. Vol. 11.

West, Mike, and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd ed. Verlag New York: Springer. https://doi.org/10.1007/b98971.

Yule, George U. 1926. "Why Do We Sometimes Get Nonsense Correlations between Time Series? A Study in Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society* 89: 1–64.