Statistics and Its Interface Volume 16 (2023) 419–431

Two-stage multivariate dynamic linear models to extract environmental and climate signals in coastal ecosystem data

Jacob Strock*, Gavino Puggioni, and Susanne Menden-Deuer

In environmental time series the presence of missing data, desire for multiple modeling structures, non-simultaneous data streams and computationally costly inference in highly parameterized model structures bring major challenges. In this work, we describe how multistage dynamic linear model (DLM) structures can be used to concomitantly describe long-term patterns, infer missing data, test predictive relationships, and altogether facilitate model development where multiple objectives and data streams may exist. We demonstrate the utility of this modeling approach with long-term data from Narragansett Bay (NB), Rhode Island, USA which has undergone major ecological changes including reductions in anthropogenic nutrient pollution. In a first stage, DLMs were used both to interpolate missing data and describe changes in both seasonality and long-term trend for nitrogenous nutrients and size structure of phytoplankton communities. These models revealed a long-term decline in large phytoplankton, and intensifying seasonal blooms for smaller phytoplankton. In a second modeling stage, parameters with associated uncertainty from stage 1 were used as covariates to test how features of the nitrogen series impacted phytoplankton. Conditional on the posterior inference of predictors modeled in stage 1, the dynamic regression revealed a newly discovered seasonal dependence of large phytoplankton on nitrogen sources.

Keywords and phrases: Time Series, Dynamic Linear Model, Pollution, Oceanography.

1. INTRODUCTION

In long-term environmental monitoring, scientists face a plethora of challenges which make statistical inference difficult and the use of a flexible yet strategic framework valuable. First, environmental monitoring data are inherently autocorrelated. While non-time series frameworks are common in environmental sciences, models which do not consider the autocorrelated structure of time series may have independence assumptions which are violated. Without addressing temporal traits, even important diagnostic tools like cross-correlation functions may have issues like being prone to type 1 errors [1]. Such spurious correlation between series is the result of inflated cross-correlation due to remaining autocorrelation in the series and may result in misinterpretation [2]. Second, as with most environmental series, missing data are common and may occur over prolonged periods. This amplifies the need for accurate interpolation, forecasting, and quantification of uncertainty [3, 4, 5, 6, 7]. Frequently, the autocorrelation and missing data are treated independently, for example, fitting static linear trends, omitting data, or aggregating so as to circumvent direct handling of missingness or autocorrelation [4, 5, 6, 7]. When time series methods are employed, concerns for missingness still remain in the context of model structures which will impact the information loss during forecasting and interpolation [8, 9]. Next, it is often desirable to isolate specific components such as seasonality, long-term trend, and the effect of predictors. This is crucial considering that these features are not only apparent but also interpretable. For example, often intra-annual patterns drive the ecological dynamics of the marine ecosystems [10, 13, 11, 12]. Last, in time series, temporally constant parameterizations are often assumed, but it is generally considered that ecological systems with feed-back mechanisms or tipping points are likely to see major changes in temporal patterns at non-linear rates [14]. We expect that data dependencies may evolve with changes in ecosystem state, requiring time-varying parameterizations.

State-space models offer the capacity to address these key features in long-term environmental monitoring data. In particular, the dynamic linear model (DLM) is a specific case of state-space models, where Gaussian distributed errors are assumed and there is a long-standing body of work for inferential algorithms and common structure types (e.g., [15, 9]). The general structure of DLMs allows an easy decomposition into their additive components such as long-term trend, seasonality, and regressive components. The dependence structure of latent states in a DLM is Markovian, e.g., the state variable at a generic time t depends only on the state at time $t-1$. The sum of a linear transformation of the state variables and a random error term specifies the model for the observed values. This Markovian structure is a key trait that allows time-specific parameter estimates, that is, parameters in each of these components can vary with

*Corresponding author.

time [9]. This allows scientists to understand time-variation in ecosystems. Because of these utilities, these models in general form have been applied to a broad range in topics including and not limited to marketing [16], finance [17], and transportation [18]. Despite long-standing use, there is opportunity to study how specific structures may serve for long-term environmental monitoring such as in marine phytoplankton ecology. There is limited but growing literature on the utility of DLMs in marine phytoplankton ecology demonstrating that these models may be critical in elucidating the driving relationships in ecology which are unlikely to be static in time [19, 20].

While DLM structures can accommodate multivariate series with diverse component specifications, we suggest that such single joint specifications may be unideal in many applied contexts, such as in the case of long-term environmental monitoring. To begin, in such cases there are often multiple modeling goals. For example, two common objectives, particularly for environmental time series, is the need to accurately describe long-term changes in conjunction with the need to explore and test predictors of series of interest [3, 4, 5, 6, 7]. Inference could be performed jointly in a multivariate structure, but this would be at great parametric cost particularly in the covariance structure. Too many parameters in a DLM model can result in identifiability issues, high variance in the estimates, and the computational concern of highly correlated MCMC chain sequences [21, 9]. This could require prohibitively long MCMC simulations to produce a sufficient effective sample size. Furthermore, it may become more practical to perform a first inferential stage as an exploratory measure, or to impute data before other analyses including exploratory measures as is common in analysis with missing data. In model development, where many variations may need to be run, it is also substantially more computationally efficient to sample from the posterior inference of other independent model components rather than to repeatedly make inferences on a highly parameterized, large joint model. Last, multistage structures may be practically advantageous if multiple data sources are not available simultaneously. This may be highly relevant where data sources have different processing times. Rather than waiting for all data to perform a full joint inference, it is possible to make inferences on data sources as they become available. These are practical issues in model development which can be accommodated in a multistage modeling framework. Different stages may hold different purposes, allow different rates of data availability, ameliorate computational costs during development, and yet share information as opposed to implementing multiple independent models.

In this paper, we apply such multistage DLMs to study photosynthetic microscopic organisms called phytoplankton in a coastal study site. We implement this application with the intention to highlight a practical approach to environmental time series and demonstrate new ecological insights through the multistage DLM modeling approach. Multidecadal marine monitoring of Narragansett Bay (NB) RI, USA both represents a wealth of multivariate monitoring data, and represents many of the aforementioned challenges in environmental data series. The data is autocorrelated, has extensive missingness, and dynamic structure that is expected to require time-varying parameterization in modeling. Scientists are driven both to understand long-term changes and seasonal patterns as well as a need study the dependence between series in exploratory analysis and in targeted regression structures. Further, although the proceeding analysis centers around a historical segment of published data, to scientists, data sources are often processed at different rates (e.g., batches of chemical analysis vs direct measurement), providing advantage to analyses that can be performed in stages as data become available.

The data series used in this analysis also has substantial applied practical significance. While phytoplankton are small, globally these organisms are highly abundant in the oceans and therefore responsible for primary production on the order of 36.5-48.5 Gt C yr^{-1} [23, 24]. This biological production is critical for everything from global biogeochemical cycling (e.g., carbon) [25], to the productivity of marine food webs [26]. And despite the magnitude of this biological production, it is not impervious to human impact with some direct forms of disturbance including nutrient pollution (especially nitrogen) which alters marine ecosystem function in ways such as excessive production (eutrophication) to changing the biological community structure [27]. While some regions are facing increasing eutrophication [28], in others, government policy has been enacted or proposed to better regulate nutrient pollution [29]. Most recently in NB, between 2005 and 2012, the nutrient loading by wastewater facilities were reduced by 50% through wastewater treatment [29]. Cross-sectional studies have shown this has measurably reduced the standing stock of dissolved inorganic nitrogen (DIN) and dissolved inorganic phosphorous (DIP) in the bay by $50-60\%$ [35]. There is concern for how this could affect phytoplankton communities, including in terms of cell sizes. These phytoplankton size features are known to impose a physical constraint on the potential rate of nutrient supply [30], but ecologically, cell size is also inversely related to maximum abundance [31], strength of the microbial recycling, and food chain length [32], and positively related to features including sedimentation rate, and export efficiency of material to the deep ocean [34].

In this study, we outline how multistage Bayesian DLMs can be used to meet the above statistical goals of ecological time series through the example of the Narragansett Bay Time Series. In section 2 we cover the basic DLM, introduce the general multistage structure, useful component specifications, and inference. In section 3 we apply this model structure to investigate long-term ecologic change in the size structure of phytoplankton and environmental predictors in Narragansett Bay from 2003-2019 through a weekly record of environmental features. In section 4 we describe inference

on long-term patterns as well as the relationship of the phytoplankton community as a function of these environmental variables. We conclude (5) with scientific interpretation and discussion of how biological function of marine communities may change over long spans of time, necessitating flexible time-series models and benefiting from multistage structures.

2. METHODOLOGY

2.1 Dynamic linear model structure

The standard, multivariate DLM can be represented as a system of two linear equations: the observation equation, and the state equation. The state equation is also called the evolution equation borrowing from terminology of dynamic equation literature and does not reference biologic definitions. The observation equation represents the $(r \times 1)$ observation vector (Y_t), $t = 1, \ldots, T$, from an $(n \times 1)$ unobserved state (θ_t), transformed by the $(r \times n)$ observation matrix (F_t) with $(r \times 1)$ error (v_t). This normally distributed observational error is considered as coming from a zero mean normal with $(r \times r)$ covariance $\boldsymbol{V}_t$. The second equation models the evolution of the latent state θ_t through time, according to the $(n \times n)$ evolution matrix $(\boldsymbol{G}_t)$ with $(n \times 1)$ evolution error w_t. This normally distributed evolution error is considered as coming from a zero mean normal with with $(n \times n)$ evolution covariance $(\boldsymbol{W}_t)$. The DLM is defined at time t by the set $\{F_t, \boldsymbol{G}_t, \theta_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$. In the DLM, it is assumed that both evolution and observational errors are normally distributed, and priors for unknown parameters —commonly $P(\theta_0), P(V_0), P(W_0)$ —are designed to suit these assumption.

$$
(1) \qquad \begin{cases} Y_t = F_t\theta_t + v_t, & v_t \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{V}_t) \\ \theta_t = \boldsymbol{G}_t\theta_{t-1} + w_t, & w_t \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{W}_t) \\ \theta_0 \sim P(\theta_0) & \\ \boldsymbol{V}_0 \sim P(\boldsymbol{V_0}) & \\ \boldsymbol{W}_0 \sim P(\boldsymbol{W_0}) & \end{cases}
$$

Through this general framework a wide choice of model structures are possible, including popular model structures such as ARIMA models [9, 33] to static regression in the case when $\boldsymbol{W}$ is set to 0. While potential component specification are expansive, some common components to θ include parameters for ARIMA class structures, seasonality, trend, and regression [9, 33]. In applied time series analysis, several of these model components and even multiple, independent model specifications may be important. For example, it may be valuable to use models with seasonality and trend to understand long-term changes in these features. In addition, regression models may also be desired to understand the relationship between the series. Ultimately, use of two or more models may be critical among these applications. In part, this is because while missingness is easily accommodated in response variables in standard updating algorithms [15], it does pose an issue in regression models as complete predictors are required. Therefore, preliminary imputation and analysis models may be necessary.

We define the general multistage model for stage $s = 1, \ldots, \mathcal{S}$. In the multistage model, the posterior parameters at stage $s = u$, are conditional on data and posterior parameters of all stages $s < u$. That is, specification of any given stage can be described as in equations 1, conditional on inference from any previous stage. While posteriors parameters of s can be marginalized over the posteriors of stages $s < u$, in the multistage they are required for inference. Otherwise, posterior inference could be carried out independently in separate models. Practical examples–as will be demonstrated–will be regression models where parameters of one model serve as predictors in the next stage.

In this analysis we exemplify a two-stage extension on the basic DLM structure that provides a framework for frequent tasks in applied time series analysis. Stage 1 serves to characterize long-term trend and changing seasonality as is important—for example —in climate and phenological studies [3, 4, 5, 6, 7]. Stage 1 also provides imputation for predictive series with missingness, and inference on features of interest. Here, for the length r set $Y^Q = \{Y_1^Q, \ldots, Y_r^Q\}$ of all variables of interest, a multivariate structure is used to model $Y_t^Q, t = 1, \ldots, T$, and provide long-term description and inference. Given these stage 1 results, in stage 2, we may perform a regression between any z length subset Y^Z, $Y^Z \subseteq Y^Q$, of response variables, and posterior parameters(Ψ^X_{stage1}) of x length subset Y^X, $Y^X \subseteq Y^Q$, of predictors such that $Y^X \neq Y^Z$. As is common for regression models, if the $z \times w$ length state vector θ_t^Z contains a regression coefficient β_X, then the predictor derived from Ψ^X_{stage1} is contained in the corresponding element $F^Z_{X,t}$ of F_t^Z (see 2.2.2). Given the preceding stage 1 inference, this regressor could either be raw data with imputation (Y_t^X), posterior quantity of interest (Ψ^X_{stage1}) generally, or any function thereof $(g(\Psi^X_{stage1}))$. In this way, stage 2 allows hypotheses tests of associations, which are also common in applied time series analysis. More formally the structure can be presented as below:

Stage 1

$$
(2) \qquad \begin{cases} \boldsymbol{Y}_t^Q = F_t^Q\boldsymbol{\theta}_t^Q + v_t^Q, & v_t^Q \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{V}_t^Q) \\ \boldsymbol{\theta}_t^Q = \boldsymbol{G}_t^Q\boldsymbol{\theta}_{t-1}^Q + w_t^Q, & w_t^Q \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{W}_t^Q) \end{cases}
$$

Stage 2

$$
(3) \qquad \begin{cases} \boldsymbol{Y}_t^Z = F_t^Z\boldsymbol{\theta}_t^Z + v_t^Z, & v_t^Z \sim \mathcal{N}_z(\mathbf{0}, \boldsymbol{V}_t^Z) \\ \boldsymbol{\theta}_t^Z = \boldsymbol{G}_t^Z\boldsymbol{\theta}_{t-1}^Z + w_t^Z, & w_t^Z \sim \mathcal{N}_{rw}(\mathbf{0}, \boldsymbol{W}_t^Z \otimes \boldsymbol{V}_t^Z) \\ F^Z_{X,t} \sim P(g(\Psi^X_{Stage1})) & \end{cases}
$$

$$
\Psi_0 \sim \Pi_{k=1}^K P(\Psi_{0,k})
$$

$$
\Psi_0 = \{\theta_0^Z, \theta_0^Q, V_0^Z, V_0^Q, W_0^Z, W_0^Q\}
$$

In this framework, stage 1 and stage 2 can be completed sequentially. Inference of the stage 1 parameters, Ψ^Q, is independent of stage 2 parameters Ψ^Z. This is because stage 1 parameters (Ψ^Q) are independent of stage 2 parameters Ψ^Q so that $P(\Psi^Q|\Psi^Z) = P(\Psi^Q)$. Developing this dependence structure carries the assumed causal relationship represented in the regression structure. This feature has practical value in that it allows posterior inference from the stage 1 descriptive models to better inform stage 2 model specification as is commonly done, for example in exploring lags of interest between series [36]. However, during stage 2 inference via MCMC algorithms, it is simple to sample the stage 1 model for parameters of interest or missing data, so that the posterior distribution of stage 2 parameters $p(\Psi^Z|\cdot)$ are integrated over uncertainty in the predictor $p(\Psi^X)$ as can be seen below:

$$\int P(\Psi^Z|\Psi^X,\cdot)P(\Psi^X)d\Psi^X = P(\Psi^Z|\cdot),$$
$$\Psi_Z = \{\theta^Z_{1:T}, \boldsymbol{V}^Z, \boldsymbol{W}^Z_t\}$$

2.2 Model structures

2.2.1 Stage 1 model structures

Stage 1 models (2) provide the opportunity to characterize the long-term patterns of the environmental series jointly. Through the multivariate structure and inference, the correlation between series can be leveraged in the inferential algorithm, so that for r series of $Y^Q_t = (Y^Q_{1,t}, \ldots, Y^Q_{r,t})$, if one series $Y^Q_{i,t}$, $i \subset 1, \ldots, r$, is missing data where another is not missing, the covariance between the series provides additional information. To best describe long-term change and seasonality among all r series, for each series i, $\theta^Q_{i,t}$ has two major components to describe long-term pattern and seasonal trend. Assuming this specification with n components is sufficient for all series, i, the complete latent state $(\theta^Q_t = (\theta^Q_{1,t}, \ldots, \theta^Q_{r,t}))'$ of the full model is $rn \times 1$, the observation matrix $(F^Q_t = (F^Q_{1,t}, \ldots, F^Q_{r,t}))$ is $(r \times rn)$ with $(r \times 1)$ error vector (v^Q_t) with $(r \times r)$ covariance $\boldsymbol{V}^Q_t$, $(rn \times rn)$ evolution matrix $(\boldsymbol{G}^Q_t = diag(\boldsymbol{G}^Q_{1,t}, \ldots, \boldsymbol{G}^Q_{r,t}))$ with $(rn \times 1)$ evolution error w^Q_t with $(rn \times rn)$ evolution covariance $(\boldsymbol{W}^Q_t)$

In this form, the first component for each series i is a dynamic intercept $\theta^Q_{i,\mu,t} = \mu^Q_{i,t}$ with corresponding components in $F^Q_{i,\mu,t}$ and $\boldsymbol{G}^Q_{i,\mu,t}$ of 1. Thereby, marginally, $f(\mu^Q_{i,t} \mid \mu^Q_{i,t-1}, \boldsymbol{W}^Q_{i,\mu}) \sim \mathcal{N}_{rn}(\mu^Q_{i,t-1}, \boldsymbol{W}^Q_{i,\mu})$, and so $(\mu^Q_{i,t})$ has the flexible structure of a random walk process. The second component is dynamic seasonality. Seasonal components for each i, $(\theta^Q_{i,S,t} = S^Q_{i,t})$ may be included by harmonics in Fourier form, for a parsimonious representation of annual cyclicity [9]. For example, with weekly resolution data and annual patterns, cyclicity could be modeled with a period 52.17 function. While any function of period s can be modeled by $s/2$ harmonics, in general a smaller number J is both more interpretable and aids in reducing over-parameterization. Within θ^Q_t, for each frequency, both the harmonic $S_{j,t}$ and its conjugate $S_{j,t}*$ are included, and evolve according to the subcomponent H_j, for Fourier frequencies $j = 1, \ldots, J$ of the evolution matrix $\boldsymbol{G}^Q_i = diag(1, \boldsymbol{H}_1, \ldots, \boldsymbol{H}_J)$.

$$\begin{cases} F^Q_i = \begin{bmatrix} 1 & (1 & 0) & (1 & 0) & \cdots & (1 & 0)_J \end{bmatrix} \\ \boldsymbol{G}^Q_i = \begin{bmatrix} 1 & \\ & \boldsymbol{G}^Q_s \end{bmatrix} \\ \boldsymbol{G}^Q_s = \begin{bmatrix} \boldsymbol{H}_1 & & \\ & \ddots & \\ & & \boldsymbol{H}_J \end{bmatrix} \\ \boldsymbol{H}_j = \begin{bmatrix} \cos(\omega_j) & \sin(\omega_j) \\ -\sin(\omega_j) & \cos(\omega_j) \end{bmatrix} \\ \omega_j = 2\frac{\pi_j}{s}, j = 1, \ldots, J \\ \theta^Q_{i,t} = \begin{bmatrix} \mu_i \\ S_{i,1,t} \\ S*_{i,1,t} \\ \vdots S_{i,J,t} \\ S*_{i,J,t} \end{bmatrix} \theta^Q_t = \begin{bmatrix} \theta^Q_{1,t} \\ \vdots \\ \theta^Q_{r,t} \end{bmatrix} \end{cases} \tag{4}$$

2.2.2 Stage 2 modeling structures

In addition to description and data exploration, the stage 1 models also provide opportunity for various imputation and parameter estimates for use in subsequent modeling steps. One option for the stage 2 models (3), rather than to use the observed cases and imputation of missing cases of predictors such as from the posterior predictive, the predictors may be a function of latent states such as $(F^X_t \boldsymbol{\theta}^X_t \mid Y^Q_t)$ which is designed to represent the latent state of the predictor without observational uncertainty. For multivariate or even matrix-variate response series, it may also be advantageous for the regression to take the form of a matrix-variate regression model, that is, $\boldsymbol{Y}^Z_t = (Y^Z_{1,t}, \ldots, Y^Z_{z,t})$ becomes dimension $(z \times p)$, $\boldsymbol{\theta}^Z_t$ as $(z \times m)$, and the evolution covariance as the Kronecker product of the $(m \times m)$ $\boldsymbol{W}^Z_t$ and $(z \times z)$ $\boldsymbol{V}^Z_t$. Specifying a matrix-variate state allows the ability to model correlated evolution among the series as well as the state variables [37].

For each series $l, l \subset 1, \ldots, z$, which take the same component form, regression models might be specified with three major components to the latent state, $\theta^Z_t = (\theta^Z_{1,t}, \ldots, \theta^Z_{z,t})$. The first component $\theta^Z_{l,\mu,t}$ is a dynamic intercept $(\mu^Z_{l,t})$ with corresponding components in $F^Z_{\mu,t}$ and $G^Z_{\mu,t}$ of 1. Thereby, marginally, $f(\mu^Z_{l,t} \mid \mu^Z_{l,t-1}, W^Z_{l,\mu}) \sim \mathcal{N}(\mu^Z_{l,t-1}, W^Z_{l,\mu})$, and so

($\mu^Z_{l,t}$) has the flexible structure of a random walk process. The second component is a regression coefficient ($\beta^Z_{l,t}$) on the appropriate lag k of the predictor X. For example, this lag may be chosen as the lag with highest cross-correlation between the predictor series and response. The corresponding multiplier in $F^Z_{\beta,t}$ is the predictor X_k, and the corresponding component in $G^Z_{\beta,t}$ is 1. Thereby marginally, $\beta^Z_{l,t}$ can also evolve with the flexibility of a random walk process. The final component of the model is a Fourier form seasonal component, with $j = 1$, that is with period s. To have a flexible yet parsimonious specification, we choose to have some of the components as constant such as the seasonality, especially if the predictor is already expected to describe some of the seasonal signal. To accomplish a static component, the corresponding elements of $\boldsymbol{W}$ are set to 0. By this structure, the interpretation of the regressive component, is not in describing the total variability attributable to the predictor, but rather, the anomaly from the long-term trend and regular seasonality.

$$
\begin{cases}
F^Z = \begin{bmatrix} 1 & g(X) & 1 & 0 \end{bmatrix} \\
\boldsymbol{G}^Z = \begin{bmatrix} 1 & & \\ & 1 & \\ & & \boldsymbol{G}^Z_s \end{bmatrix} \\
\boldsymbol{G}^Z_s = \boldsymbol{H} \\
\boldsymbol{H} = \begin{bmatrix} \cos(\omega_j) & \sin(\omega_j) \\ -\sin(\omega_j) & \cos(\omega_j) \end{bmatrix} \\
\omega_j = 2\frac{\pi}{s} \\
\theta^Z_t = \begin{bmatrix} \mu_{1,t} & \cdots & \mu_{z,t} \\ \beta_{1,t} & \cdots & \beta_{z,t} \\ S_{1,t} & \cdots & S_{z,t} \\ S*_{1,t} & \cdots & S*_{z,t} \end{bmatrix}
\end{cases} \tag{5}
$$

2.3 Posterior inference

In the two-stage DLM, for both stage 1 and stage 2 conditional on the posteriors of stage 1, inference can be carried out via Markov chain Monte Carlo (MCMC), and under specific semi-conjugate prior specifications described below, this can occur simply by sampling the conditionally conjugate posterior distributions for each unknown parameter. The following apply for DLMs in general as well as each stage of the multi-stage model.

2.3.1 Latent state variables, $\theta_{1:T}$

Because of the Markovian structure, solving for the posterior distribution of the latent states, $f(\theta_t|D_{1:T}, \cdot)$, 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 [15]. Multivariate and matrix-variate extensions have been developed and detailed by Wang and West [37]. 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 [15]. 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.

2.3.2 Covariance terms V & W

In all models, we assume static observational covariance, consistent with the static Bayesian inference for the Gaussian distribution. The inverse-Wishart ($\mathcal{IW}$) is semi-conjugate in the multivariate-Gaussian case [38]. The $\mathcal{IW}$ is also used to specify a static evolution covariance in the stage 1 models which describe the long-term trends and seasonal structure of the nitrogen and chl. a series. Alternate to the static covariance, to accommodate time-varying stochasticity in the regression model, $\boldsymbol{W}_t$ may be specified as time-varying through the use a discount factor. Discount factors 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 [9]. While an explicit state-space model could be specified for the covariance matrix, this can be disadvantageous in terms of both complexity, and non-conjugacy in inference. The discount factor applies to the one step ahead state covariance matrix $\boldsymbol{P}_t$, itself a function of the filtered state covariance at time $t-1$, $\boldsymbol{C}_{t-1}$.

$$
\begin{aligned}
\boldsymbol{W}_t &= (1-\delta)/\delta \boldsymbol{P}_t \\
\boldsymbol{P}_t &= \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t^{'}
\end{aligned}
$$

When missingness exceeded length 1, that is forecast k steps ahead needed to be greater than 1, the method of 'practical discounting' was used [9] which limits the loss of information to linear rather than exponential increase during missing data periods.

$$
R_t = \boldsymbol{G}^{k-1} \boldsymbol{C}_{t+1} \boldsymbol{G}^{'k-1}, \; k > 1,
$$

To avoid mixing issues that come with highly parameterized models [21], instead of specifying a prior and sampling discount factors (e.g., [39]), models with fixed discount factors were run in parallel and compared as is relatively common practice [40].

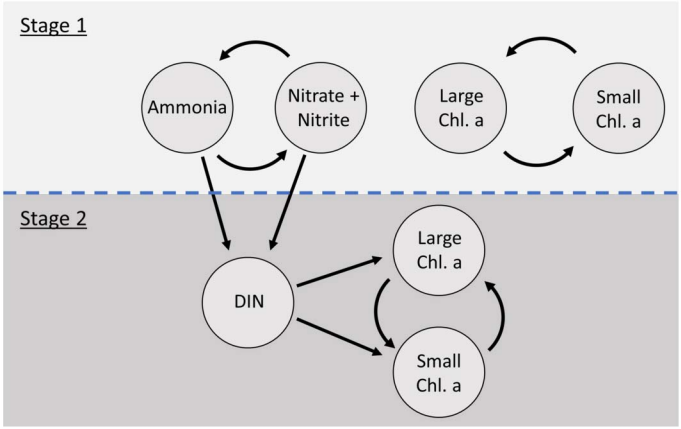

Figure 1. Dependence structure between components of the stage 1 and stage 2 model for the Narragansett Bay ecological model. A bivariate model was run for small and large chl. a as well as nitrate + nitrite to describe long-term patterns among all series. Examination of prewhitened cross-correlations between the imputed series after stage 1 led to the use of DIN as a predictor in stage 2 to explore the influence of nitrogen on size structure of phytoplankton. Stage 2 used latent levels ($F_{1:T}\theta_{NH_4,1:T}$, $F_{1:T}\theta_{NO_{3,2},1:T}$) of ammonia and nitrate + nitrite.

3. APPLICATION TO NARRAGANSETT BAY DATA

We applied this two-stage DLM with data from Narragansett Bay ecosystem in Rhode Island, USA. Chlorophyll <20 μm (small chl. a), chlorophyll >20 μm (large chl. a) and associated measurements of nitrate and nitrite ($NO_2^- + NO_3^-$), and ammonia (NH_4^+) were all obtained from the University of Rhode Island Long-Term Plankton Time Series of Narragansett Bay from January 2003 to March 2019 at weekly resolution from the publicly available dataset (https://web.uri.edu/gso/research/plankton/data/). An additional six months of data (September 2019 to March 2020) were used to evaluate out-of-sample accuracy. Dissolved inorganic nitrogen (DIN), a frequently limiting nutrient for growth in marine environments is represented here as the sum of nitrate + nitrite and ammonia. Among all series, it is expected that after natural log-transformation the assumption of Gaussian distributed errors, and linear associations are reasonably appropriate, thus conforming to the distributional and structural assumptions of the DLM case of state space models.

In stage 1, following the structure outlined above (eqns. 2, 4), bivariate models were run for log small chl. a, log large chl. a, and another for log ammonia and log nitrate + nitrite (table 1). To accommodate the complexity of the seasonal signal, 6 different seasonal harmonics are used beginning with that of the longest possible period, s=52.14 weeks. For the DIN and chl. a series, $j = (1, \ldots, 5)$ to accommodate a more complex seasonal cycle. With sufficiently long data series, weakly informative priors could be specified in all cases. For the covariance matrices $\boldsymbol{V}^Q$ and $\boldsymbol{W}^Q$ were modeled as time invariant and given priors with weakly informative, respectively $\mathcal{IW}(a_v^Q = 2, b_v^Q = 0.1 * \boldsymbol{I}_2)$, $\mathcal{IW}(a_W^Q = 2, b_W^Q = 0.1 * \boldsymbol{I}_2)$, where I_2 is the identify matrix with rank 2. θ_0^Q was given a $\mathcal{N}(m_0^Q, \boldsymbol{C}_0^Q)$ prior, where $m_{0,\mu}^Q = \bar{y}$, $m_{0,-\mu}^Q = 0$, and $\boldsymbol{C}_0^Q = \boldsymbol{I}_2$. Because θ_t^Q,$\boldsymbol{V}^Q$,and $\boldsymbol{W}^Q$ are conditionally conjugate, Gibbs sampling from the FFBS, $\mathcal{IW}$ distribution of $\boldsymbol{V}^Q$, and $\mathcal{IW}$ distribution of $\boldsymbol{W}^Q$ leads to a sample of the full joint posterior.

Stage 1

$$\boldsymbol{Y}_t^Q = F^Q \boldsymbol{\theta}_t^Q + v_t^Q, \qquad v_t^Q \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{V}^Q)$$
$$\theta_t^Q = \boldsymbol{G}^Q \boldsymbol{\theta}_{t-1}^Q + w_t^Q, \qquad w_t^Q \sim \mathcal{N}_{22}(\boldsymbol{0}, \boldsymbol{W}^Q)$$
$$\boldsymbol{\theta}_0^Q \sim \mathcal{N}_{22}(m_0^Q, \boldsymbol{C}_0^Q)$$
$$\boldsymbol{V}_0^Q \sim \mathcal{IW}(a_v^Q, \boldsymbol{b_v^Q})$$
$$\boldsymbol{W}_0^Q \sim \mathcal{IW}(a_w^Q, \boldsymbol{b_w^Q})$$

Stage 2

$$\boldsymbol{Y}_t^Z = F_t^Z \boldsymbol{\theta}_t^Z + v_t^Z, \qquad v_t^Z \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{V}^Z)$$
$$\boldsymbol{\theta}_t^Z = \boldsymbol{G}^Z \boldsymbol{\theta}_{t-1}^Z + w_t^Z, \qquad w_t^Z \sim \mathcal{N}_8(\boldsymbol{0}, \boldsymbol{W}_t^Z \otimes \boldsymbol{V}^Z)$$
$$F_{X,t}^Z \sim P(g(F_t^X \theta_t^X))$$
$$\boldsymbol{\theta}_0^Z \sim \mathcal{N}_8(m_0^Z, \boldsymbol{C}_0^Z)$$
$$\boldsymbol{V}_0^Z \sim \mathcal{IW}(a_v^Z, \boldsymbol{b_v^Z})$$
$$\boldsymbol{W}_t^Z = \frac{1-\delta}{\delta} \boldsymbol{P}_t$$
$$\boldsymbol{P}_t^Z = \boldsymbol{G}^Z \boldsymbol{C}_{t-1}^Z \boldsymbol{G}'^Z$$
$$R_t = \boldsymbol{G}^{Z,k-1} \boldsymbol{C}_{t+1}^Z \boldsymbol{G}'^{Z,k-1}, \; k > 1$$

After stage 1, the relationship between each chl. a fraction and DIN was explored through their cross-correlations to inform the structure of stage 2 (eqns. 3, 5). That is, the features of X and function $g(X)$ were identified to specify the regression component of the observation matrix ($F_{X,t}^Z$). Weakly informative priors were specified in all cases. $\boldsymbol{V}^Z$ was modeled as time invariant and given a prior with weakly informative, $\mathcal{IW}(2, 0.1 * \boldsymbol{I}_2)$, where $\boldsymbol{I}_2$ is the identify matrix with rank 2. $\boldsymbol{\theta}_0^Z$ was given a $\mathcal{N}(m_0^Z, \boldsymbol{C}_0^Z)$ prior, where where $m_{0,\mu}^Z = \bar{y}$, $m_{0,-\mu}^Z = 0$, and $\boldsymbol{C}_0^Z = I_p$. For the case of discount factors, for $\boldsymbol{W}^Z$, discrete discount factors (0.8, 0.85, 0.9, 0.95, 0.99, 0.999) were tested in parallel model runs. Because $\boldsymbol{\theta}_t^Z$ and V^Z are conditionally conjugate and $\boldsymbol{W}_t^Z$ specified by δ, updating W_t^Z, and Gibbs sampling from the FFBS and $\mathcal{IW}$ distribution of $\boldsymbol{V}^Z$. Given stage 2 inference is conditional on stage 1 posterior quantities $\theta_{1:T}^X$, these were sampled from the posterior MCMC samples of the stage 1 model so as to integrate over uncertainty. The complete model followed the structures outlined in methodology.

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

Stage	Model	Response Variables	Components	V Specification	W Specification
1	1	ammonia, nitrate + nitrite	Dynamic, Intercept μ Dynamic, Season (S_i), i=1,...,5	Static, $\mathcal{IW}$ prior	Static $\mathcal{IW}$ prior
	2	Small Chl. a, Large Chl. a	Dynamic, Intercept μ Dynamic, Season (S_i), i=1,...,5	Static, $\mathcal{IW}$ prior	Static $\mathcal{IW}$ prior
2	3	Small Chl. a, Large Chl. a	Dynamic, Intercept μ Dynamic, Regression on DIN Static, Season (S_i), i=1,...,5	Static $\mathcal{IW}$ prior	Dynamic, Fixed Discount Factor

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, simulation length was 10,000 iterations with a burn-in period of 2,000 to reach a minimum of 1,000 effective samples according to Gelman et al. [21], where W is the within sequence variance, and B is the between sequence variance for n samples and m MCMC chains of the parameter ϕ.

$$\hat{n}_{eff} = \frac{mn}{1 \pm 2\Sigma_{t=1}^{T}\hat{\rho}_t}$$

$$\hat{\rho}_t = 1 - \frac{V_t}{2\hat{var}}$$

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

$$\hat{var}(\phi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

The fixed discount factors are commonly chosen with mean one-step-ahead forecast error [40, 41]. RMSFE was calculated for each MCMC iteration from the one step ahead mean forecast. As this was accomplished for each MCMC iteration of each model, Monte Carlo sampling of the RMSFE distribution of each model was used to calculate, for each pair (i, j) of models $P(RMSFE_i < RMSFE_j)$. RMSFE can be defined as described below, allowing comparison of the distribution of RMSFE among models.

$$RMSFE = \sqrt{\frac{\Sigma_{t=1}^{n}(y_t - F_{t-1}\theta_{t-1,i})^2}{n}}$$

One-step-ahead RMSFE was also analyzed out of sample for the final six months (Sept. 2019-March 2020), and compared between the stage 1 DLM approach taken here and a standard ARIMA model fit using AIC for model selection.

To test some of the long-term changes in mean levels, we use Monte-Carlo method whereby the dynamic intercept pre-remediation (before 2005) and post-remediation (after 2012) are sampled from the posterior distribution and compared. In this way, the Bayesian framework provides a simple method of hypothesis testing any model parameter.

$$P(A > B) = \frac{\Sigma_r \delta_{a>b}}{r}$$

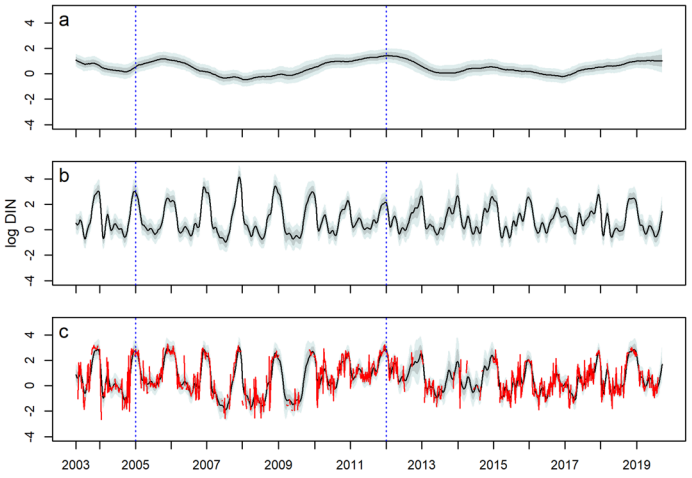

Figure 2. Decomposition of the DIN series DLM (2003–2019), fit with the stage 1 model structure. **a.** *the dynamic intercept,* **b.** *the seasonal trend,* **c.** *the posterior predicted mean with the true data (red). The median (black), 80% (dark grey shading), and 95% (light grey shading) pointwise credible intervals are shown. Blue dotted lines denote the beginning and end years of policy mandated nutrient remediation.*

In addition to these quantities, hereafter in results, we report mean and the 95% CI of the posterior distributions of interest.

4. RESULTS

4.1 Stage 1 inference

For ammonia and nitrate + nitrite although long-term monotonic changes were hypothesized for the dynamic intercept, the actual patterns from 2003–2019 were more complex, with multiyear patterns not following clear monotonic changes expected with the policy mandates for reduction in wastewater levels by 50% between 2005–2012 (fig. 2). 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. 2). Both series of N species show a high correlation of (0.50±0.074), with overall higher variability in the nitrate + nitrite (0.94±0.14), as compared to ammonia (0.74±0.10).

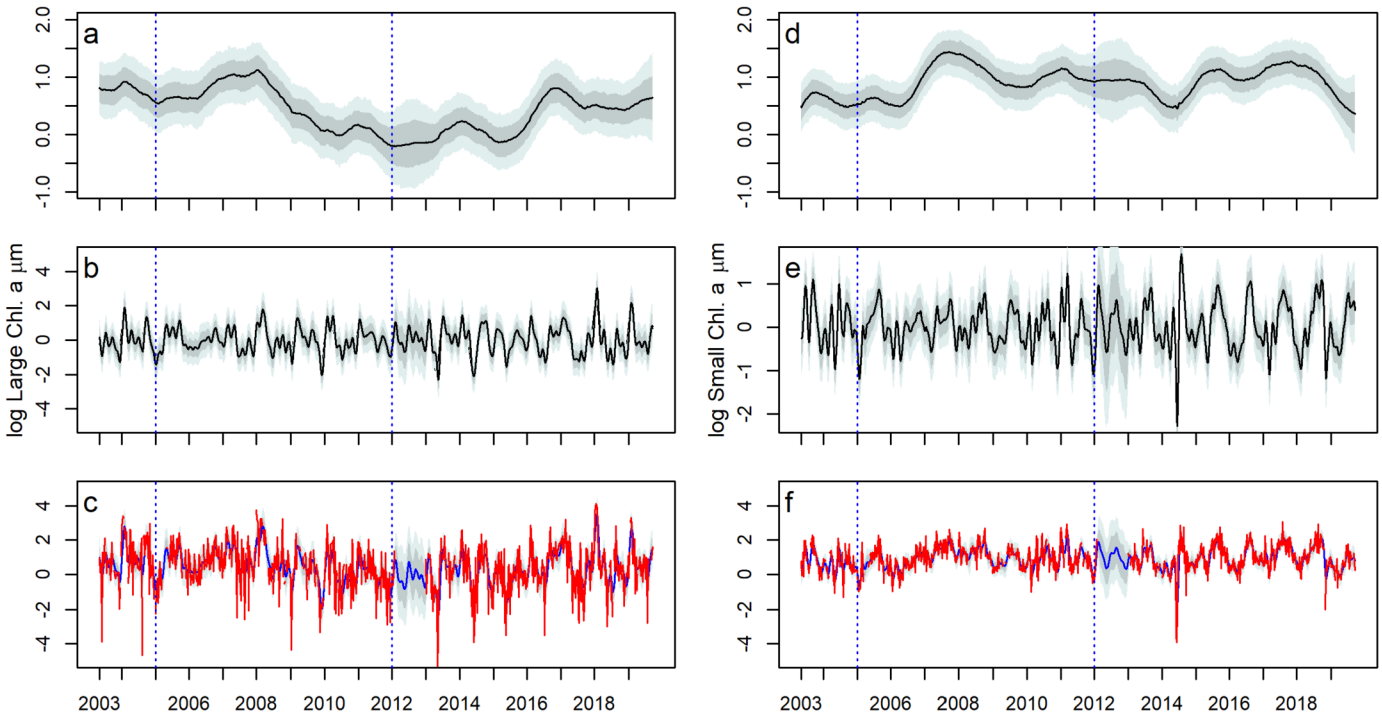

Figure 3. Decomposition of the small and large chl. a DLM (2003–2019), fit with the stage 1 model structure. **a,d.** *the dynamic intercept,* **b,e.** *the seasonal trend,* **c,f.** *the posterior predicted mean (blue) with the true data (red). For the dynamic intercept and the season, the median (black), 80% (dark grey shading), and 95% (light grey shading) pointwise credible intervals are shown. Blue dotted lines denote the beginning and end years of policy mandated nutrient remediation.*

Decomposition of the small chl. a series show variable levels in the dynamic intercept over the observation period, with an increase from pre-remediation (0.59±0.44) to post-remediation (0.91±0.66), comparing the posterior distributions from these two periods through Monte Carlo simulation, however, this is not a significant increase with posterior probability of post-remediation higher than pre-remediation equal to 0.15 (fig. 3). 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.

Decomposition of the large chl. a series show variable levels in the dynamic intercept over the observation period, with a marked decline beginning in 2008 (fig. 3). 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 pre-remediation (0.78±0.52) to after 2012 (0.31±76), the posterior probability that the later period is lower is equal to 0.989. 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. Beyond mean levels, the observational matrix of the bivariate DLM for chl. a shows several features. The large chl. a series shows inherently higher variability (1.1±0.04) than the small chl. a series (0.3±0.14).

Table 2. Mean one-step-ahead RMSFE calculated for ARIMA and first stage DLM models in 6-months of data September 2019-March 2020.

Data Series	ARIMA Fit			ARIMA	Stage 1 DLM
	p	d	q	RMSFE	RMSFE
Ammonia	5	0	0	1.4	0.57
Nitrate + Nitrite	3	0	0	1.2	0.57
Large Chl. a	4	1	0	1.3	0.82
Small Chl. a	2	0	0	1.1	0.76

The cross-correlations between the imputed DIN series and each chl. a series were examined. Imputation of the DIN series used the posterior mean when data were missing. This was done to characterize the relationship between the series across time lags. These cross-correlations helped inform the significantly associated lags, and the strength of association between the series.

4.2 Stage 2 inference

After examining the prewhitened cross-correlation between the series to reduce spurious cross–correlations [22] and testing preliminary model structures, it was decided to

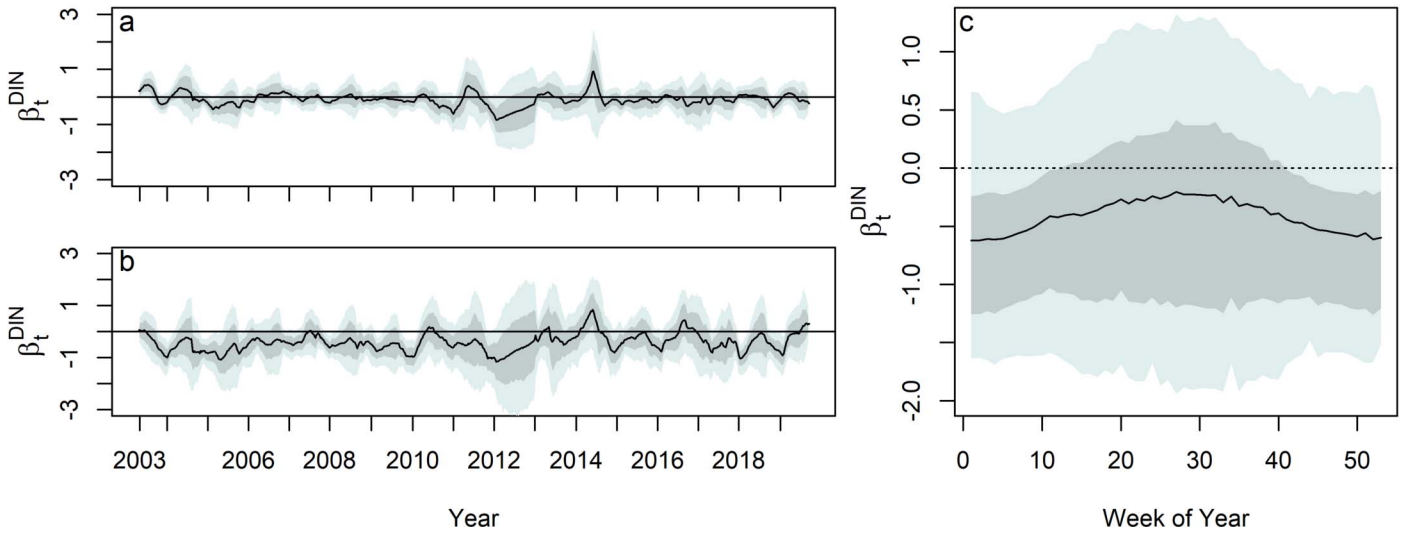

Figure 4. The dynamic regression coefficient, β_t^{DIN}, for both the **a.** *Small chl. a series* **b.** *Large chl. a series.* **c.** *Posterior distribution of the dynamic regression coefficient, β_t^{DIN}, on DIN for the large chl. a, plotted by week on the x–axis, and by year as denoted by color shading. The median (black), 80% (dark grey shading), and 95% (light grey shading) are shown.*

use dissolved inorganic nitrogen (DIN, ammonia + nitrate + nitrite as a predictor at lag 0, which showed the strongest cross-correlation to both chl. a series. Using DIN instead of the ammonia and nitrate + nitrite series as separate predictors statistically avoided multicollinearity issues considering the two series are correlated. 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 [21, 9], which hypothetically could require prohibitively long MCMC simulations to produce a sufficient effective sample size. Combined with scientific studies in Narragansett Bay that have shown no preference to N species by size fraction [42] it was decided DIN was an appropriate predictor for both chl. a series. With this structure, the F_{t,X_t} of the stage 2 model is specified in this model as follows:

$$F_{X,t}^{Z}\ P(\log(DIN_t) \mid Y_{1:T}^{X}) = \log(exp(F_{1:T}^{X}\theta_{NH_4,1:T}^{X}) + exp(F_{1:T}^{X}\theta_{NO_{3,2},1:T}^{X}))$$

$$\theta_{NH_4,1:T}^{X} \mid Y_{1:T}^{X} \ \sim \int P(\theta_{NH_4,1:T}^{X} \mid Y_{1:T}^{X}, \cdot)P(\cdot|Y_{1:T}^{X})d\cdot$$

$$\theta_{NO_{3,2},1:T}^{X} \mid Y_{1:T}^{X} \ \sim \int P(\theta_{NO_{3,2},1:T}^{X} \mid Y_{1:T,N}^{X}, \cdot)P(\cdot \mid Y_{1:T,N}^{X})d\cdot$$

For the bivariate regression model, the use of a single discount factor for μ and β was considered as was a separate discount factor for μ and β. With two discount factors, RMSFE suggested more flexible models, particularly for δ_μ, with δ_μ tending toward the lowest levels provided. In this case, δ_μ= 0.80, δ_β=0.90 was suggested. 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. 4, table 3). 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.

Table 3. Type 1 RMSFE calculated for each model with fixed, equal discount factors for μ and β. Each cell is the probability that the RMSFE of the row index exceeds that of the column index.

	0.8	0.85	0.9	0.95	0.99	0.999
0.8	<0.001	0.664	0.512	0.148	0.016	0.008
0.85	0.336	<0.001	0.32	0.04	<0.001	<0.001
0.9	0.488	0.68	<0.001	0.064	<0.001	<0.001
0.95	0.852	0.96	0.936	<0.001	0.004	<0.001
0.99	0.984	>0.999	>0.999	0.996	<0.001	0.008
0.999	0.992	>0.999	>0.999	>0.999	0.992	<0.001

For the small chl. a series, for the pre-remediation to the post-remediation period, there was no major change in the regression coefficient (fig. 4 a.). Considering the pointwise credible interval for the small chl. a series, the regression coefficient was also not significantly different from 0. For the large chl. a series, for the pre-remediation period to the

post-remediation period, the posterior probability the regression coefficient had increased was equal to 0.484. (fig. 4 b.). Considering the pointwise credible interval for the large chl. a series, the regression coefficient was periodically significantly different from 0. Graphical representation on the annual scale suggested this coefficient takes an annual cycle (fig. 4 c.). Aggregating the mean coefficient across all years and comparing across the annual cycle, the strongest associations occur in the winter, suggesting the large phytoplankton are seasonally tied to the ambient nutrient signal (fig. 4 c.). evolution covariance showed that there is a seasonal pattern to the evolution rate in the state components. The observational error, which also serves to represent cross-series-covariance shows similar results as 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. The mean of residuals was not significantly different from 0 and the ACF of the residuals showed there was no significant temporal structure left in the data. According to Ljung box in ACF, residuals are practically indistinguishable from white noise.

5. DISCUSSION

While Rhode Island state law mandated the reduction of 50% from wastewater effluent into the bay during the period of 2005–2012 [43, 44], the stage 1 inference of long-term trend in multiple nitrogen species combined show that at this observation station, net changes in the ambient DIN signal and its constituents 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 [35], and others have suggested that other sources like marine sediments may also be a major N source, which would potentially stabilize ambient levels [45]. Nevertheless, instead of a clear monotonic drawdown in nutrients, applying a flexible stage 1 structure to describe changing trend and seasonality showed there is multiyear variability apparent both in seasonal cycle and mean levels.

The DLM of the chl. a series shows a net change in the size structure of phytoplankton in Narragansett Bay, with phytoplankton >20 μm 0.47 $\mu g L^{-1}$ lower and phytoplankton <20 μm 0.32 $\mu g L^{-1}$ higher across the nutrient remediation period (2005–2012). Comparison of the dynamic intercepts shows net ecosystem shift toward smaller organisms which could have potentially rippling effects through food webs [46]. Again, graphical representation of the posterior of μ emphasizes that the dynamic intercept is valuable in capturing non-linear trends with multiyear variability. Further, in addition to changes in mean level represented by the dynamic intercept, the dynamic seasonal components show an increase in the intensity of the seasonal cycle of phytoplankton >20 μm following the end of nutrient reduction. This equates to higher summer maxima and lower winter minima for the small phytoplankton fraction.

Beyond state features, covariance of the DLM is interpretable and useful for understanding the variability in the environmental series. The DLM showed that the chl. a >20 μm are inherently more variable, with an observational variance of 1.05±0.06 as compared to small chl. a (0.31±0.02). This suggests that larger plankton in the bay are potentially much patchier, and inherently more stochastic in their population dynamics. This result aligns with hypotheses about bloom dynamics and predator control on growth. For example, smaller phytoplankton are more tightly coupled to predator control, and this results in more stable populations as compared to larger phytoplankton [47]. 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.

Altogether, stage 1 in the model provided several useful utilities. It provided a description of long-term and seasonal trend that was easily extractable from the stage 1 model structure. It provided an imputed data set for exploratory data analysis, that helped develop the second stage regression structure including through the calculation of cross-correlation and checks on linearity assumptions. This ultimately helped us select DIN as a regressor as opposed to the individual nitrogen species. It also aided computationally with the development of the regression model. It was not necessary to simultaneously model the regressors. Rather, the posterior quantities of the stage 1 could be sampled directly.

Ultimately, in the second stage of the model, 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 small phytoplankton 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.

Stage 2 inference showed that, in contrast to the phytoplankton <20 μm, phytoplankton >20 μm are often significantly tied to the DIN signal. The negative coefficient has an important scientific interpretation: N is typically the limiting nutrient for phytoplankton growth, including in Narragansett Bay [54]. The negative relationship suggests that the ambient nitrogen levels decline as the nutrient feeds large phytoplankton growth. Nevertheless, as expected, 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 $>20\ \mu m$ series. This suggests that for the phytoplankton $>20\ \mu m$, 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.

In this analysis we address several of the challenges in long-term environmental monitoring but also the importance and potential of state-space models for such series. In application on data from long-term series in Narragansett Bay, USA, variability from weeks to years in this marine ecosystem made the flexible structure of the DLM necessary for inference. Further, our two-stage modeling strategy fit several needs and showed several practically beneficial properties. First, stage 1 allowed characterization of seasonal and long-term change in the data, which poses a major challenge to environmental scientists. With inference into periods of missing data, it allowed further exploratory analysis into optimal model structures for stage 2, and provided completed data series. In the stage 2 dynamic regression models, it also allowed working with the latent levels of the predictors, which do not carry observational uncertainty. Last, in computation, the MCMC inference could be carried out independently for stage 1 and 2, meaning practically, the posteriors from stage 1 could be sampled in stage 2. This means the inference on missing data does not need to be repeated in the regression analysis. Beyond the importance of two stages, we found that to accommodate the flexibility necessary to fit the data and prevent over-parameterization, practical discount methods were critical for the evolution covariance matrix. Using a single discount factor for all dynamic components at fixed levels circumvented over-parameterization issues and mixing issues in the MCMC algorithm respectively, while allowing a time-varying covariance structure, necessary in these highly stochastic environmental series.

While the multistage DLM offers several practical advantages, there are some situations where its adoption would not be ideal. First, when the goal is to empirically investigate the correlation structure of a multivariate time series, by splitting the inference into different stages, information may be lost as compared to a joint multivariate approach. For joint inference on multiple series, in a standard multivariate DLM, the covariance between each series and between every state variable of every series would be explicitly modeled [15, 9, 37]. This means that any individual series, regardless of the specific components in its specification would gain from the information carried by all other series at each time step. This differs from our use of the multistage DLM where specific posterior parameters are passed to subsequent modeling stages. The second major consideration for multistage DLMs is the causal relationship between parameters in each stage of the study system. In the multistage approach, inference of stage 1 parameters is independent of data and parameters from stage 2. For instance, in the Narragansett Bay application discussed here, it was known scientifically that nutrients impact the phytoplankton abundance and therefore, the stage 2 regression might logically depend on stage 1 posterior parameters, but not the other way around. Together, these two points should be considered when deciding between a joint multivariate approach as opposed to a multistage structure.

Nevertheless, the two-stage implementation of DLMs used in this paper provided an important framework to make inferences from noisy, incomplete time-series, characterized by non-monotonic changes in both biology and chemistry, and changing temporal characteristics and dependencies. While this application of DLMs has important implications for the Narragansett Bay system, it may also serve as evidence for the value of DLMs and more general state-space models in long-term monitoring and environmental fields, where similar data structures and features might be expected in other data sets. Altogether, the multistage DLM structure provided a framework where modeling had several goals, substantial missing data, and disparate data streams. We hope this paper provides a multistep strategy and framework that might aid model development with analysis of other data series where one or more of these conditions may be met.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation (EPSCoR Cooperative Agreement OIA-1655221, Biological Oceanography OCE-1736635). Jacob P. Strock was supported by the National Space Grant College and Fellowship Program, Space Grant Opportunities (NASA STEM -NNX15AI06H).

Received 1 March 2021

REFERENCES

[1] Dean, R.T. Dunsmuir, W.T.M. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models *Behavior research methods* **48** (2) 783–802.

[2] Yule, G.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.

[3] Karentz, D. and Smayda, T.J. (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** 277–293.

[4] Li, Y. and Smayda T.J. (1998). Temporal variability of chlorophyll in Narragansett Bay, 1973–1990. *ICES Journal of Marine Science* **55** (4) 661–667.

[5] d'Alcalà, M. R., Conversano, F., Corato, F., Licandro, P., Mangoni, O., Marino, D., ... and Zingone, A. (2004). Seasonal patterns in plankton communities in a pluriannual time series at a coastal Mediterranean site (Gulf of Naples): an attempt to discern recurrences and trends. *Scientia Marina* **68** 65–83.

[6] Kane, J. (2009). A comparison of two zooplankton time series data collected in the Gulf of Maine *Journal of Plankton Research* **31** (3) 249–259.
[7] Kane, J. (2012). Temporal changes in plankton of the North Sea: community shifts and environmental drivers *Marine Ecology Progress Series* **462** 21–38.
[8] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control *Holden-Day* San Francisco MR0436499
[9] West, M. and Harrison, J. (1997). Bayesian forecasting and dynamic models Springer MR1482232
[10] Pratt, D.M. (1965). The winter-spring diatom flowering in Narragansett Bay *Limnology and Oceanography* **10** (2) 173–184.
[11] Karentz, D. and Smayda, T.J. (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** 277–293.
[12] Lawrence, C. and Menden-Deuer, S. (2012). Drivers of protistan grazing pressure: seasonal signals of plankton community composition and environmental conditions *Marine Ecology Progress Series* **459** 39–52.
[13] Durbin, E.G. Krawiec, R.W. and Smayda, T.J. (1975). Seasonal studies on the relative importance of different size fractions of phytoplankton in Narragansett Bay (USA) *Marine Biology* **32** (3) 271–287.
[14] Steffen, W. Richardson, K. Rockström, J. Cornell, S.E. Fetzer, I. Bennett, E.M. Biggs, R. and Carpenter, S.R. DeVries, W. and DeWit, C.A. et al. (2015). Planetary boundaries: Guiding human development on a changing planet *Science* **347** (6223).
[15] Kalman, R.E. (1960). A new approach to linear filtering and prediction problems *Journal of Basic Engineering* **83** (1) 95–108. MR3931993
[16] Ataman, M.B. Van Heerde, H.J. and Mela, C.F. (2010). The long-term effect of marketing strategy on brand sales *Journal of marketing research* **47** (5) 866–882.
[17] Karya, D.F Katias, P. Herlambang, T., and Rahmalia, D. (2019). Development of Uncented Kalman Filter Algorithm for stock price estimation *Journal of Physics:Conference Series* **1211** (1).
[18] Almanna, M. Elhenawy, M. and Rkha, H. (2018). Predicting bike availability in bikesharing systems using dynamic linear models *Transfportation Research Board 97th Annual Meeting, Washington DC, USA*.
[19] Arhonditsis, G.B. Paerl, H.W. Valdes-Weaver, L.M. Stow, C.A. Steinberg, L.J. and Reckhow, K.H. (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** 63–80.
[20] Jones, E. Parslow, J. and Murray, L. (2010). A Bayesian approach to state and parameter estimation in a Phytoplankton-Zooplankton model *Australian Meteorological and Oceanographic Journal* **59** 7–16.
[21] Gelman, A. Carlin, J.B. Stern, H.S. Dunson, D.B. Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis, 3rd Ed. CRC Press* MR3235677
[22] Katz, R.W. (1988). Use of cross-correlations in the search for teleconnections *Journal of Climatology* **8** (3) 241–253.
[23] Antoine, D. André, J.M. and Morel, A. (1996). Oceanic primary production: 2. Estimation at global scale from satellite (coastal zone color scanner) chlorophyll. *Global biogeochemical cycles* **10** 57—69.
[24] Field, C.B. Behrenfeld, M.J. Randerson, J.T. and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components *Science* **281** (5374) 237–240.
[25] Falkowski, P.G. (1994). The role of phytoplankton photosynthesis in global biogeochemical cycles *Photosynthesis research* **39** (3) 235–258.
[26] Steinberg, D.K., and Landry, M.J. (2017). Zooplankton and the ocean carbon cycle *Annual Review of Marine Science* **9** 413–444.
[27] Ryther, J.H. and Dunstan, W.M. (1971). Nitrogen, phosphorus, and eutrophication in the coastal marine environment *Science* **171** (3975) 1008–1013.
[28] Cloern, J.E. Foster, S.Q. and Kleckner, A.E. (2014). Phytoplankton primary production in the world's estuarine-coastal ecosystems *Biogeosciences* **11** (9) 2477–2501.
[29] Rhode Island Department of Environmental Management (RIDEM) (2005). Plan for Managing Nutrient Loadings to Rhode Island Waters
[30] Mei, Z.P. Finkel, Z.V. and Irwin, A.J. (2009). Light and nutrient availability affect the size-scaling of growth in phytoplankton *Journal of theoretical biology* **259** (3) 582–588.
[31] Irwin, A.J. Finkel, Z.V. Schofield, O.M.E and Falkowski, P.G. (2006). Scaling-up from nutrient physiology to the size-structure of phytoplankton communities *Journal of Plankton Research* **28** (5) 459–471.
[32] Sprules, W.G. and Munawar, M. (1986). Plankton size spectra in relation to ecosystem productivity, size, and perturbation*Canadian Journal of Fisheries and Aquatic Sciences* **43** (9) 1789–1794.
[33] West, M. and Prado R. (2010). Time Series: Modeling, Computation, and Inference *CRC* MR2655202
[34] Miklasz, K.A. and Denny, M.W. (2010). Diatom sinkings speeds: Improved predictions and insight from a modified Stokes' law *Limnology and Oceanography* **55** (6) 2513–2525.
[35] Oviatt, C. Smith, L. Krumholz, J. Coupland, C. Stoffel, H. Keller, A. McManus, C.M. and Reed, L. (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** 25–34.
[36] Cryer, J.D. and Chan, K.S. (2008). Time Series Analysis Springer
[37] Wang, H. and West, M. (2009). Bayesian analysis of matrix normal graphical models *Biometrika* **96** (4) 821–834. MR2564493
[38] Hoff, P.D. (2009). A first course in Bayesian statistical methods **Springer Texts in Statistics** MR2648134
[39] Rodriguez, A. and Puggioni, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction *International Journal of Forecasting* **26** (2) 293–311.
[40] Ameen, J.R.M. and Harrison, P.J. (1984). Discount weighted estimation. *Journal of forecasting* **3** (3) 285–296.
[41] Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation *Journal of business & economic statistics* **18** (3) 338–357.
[42] Furnas, M.J. (1983). Nitrogen dynamics in lower Narragansett Bay, Rhode Island. I. Uptake by size-fractionated phytoplankton populations *Journal of plankton research* **5** (5) 657–676.
[43] Monroy, E. Twichell, J. Kuhn, A. Charpentier,M. Cressman,J. Jordan,P. August,P. Swigor,J. and Schmidt,C. (2017). 2017 State of Narragansett Bay and its watershed-mapping drivers of change and variation.
[44] Oczkowski, A. Schmidt,C. Santos,E. Miller,K. Hanson,A. Cobb,D. Krumholz, J. et al (2018). How the distribution of anthropogenic nitrogen has changed in Narragansett Bay (RI, USA) following major reductions in nutrient loads *Estuaries and Coasts* **41** (8) 2260–2276.
[45] Nixon, S.W. Fulweiler, R.W. Buckley, B.A. Granger, S.L. Nowicki, B.L. and Henry, K.M. (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.
[46] Finkel, Z.V (2007). Does phytoplankton cell size matter? The evolution of modern marine food webs *Evolution of primary producers in the sea* 333–350.
[47] Irigoien, X. Flynn, K.J. and Harris, R.P. (2005). Phytoplankton blooms: a loophole in microzooplankton grazing impact? *Journal of Plankton Research* **27** (4) 313–321.

[48] Collos, Y. (1986). Time-lag algal growth dynamics: biological constraints on primary production in aquatic environments *Marine Ecological Progress Series* **33** 193–206.
[49] Dugdale, R.C. Wilkerson, F.P. Hogue, V.E. and Marchi, A. (2007). The role of ammonium and nitrate in spring bloom development in San Francisco Bay *Estuarine, Coastal and Shelf Science* **73** (1) 17–29.
[50] Probyn, T.A. (1987). Ammonium regeneration by microplankton in an upwelling environment *Mar. Ecol. Prog. Ser* **37** (5) 64.
[51] Goeyens, L. DeVries, R.T.P. Bakker, J.F. and Helder, W. (1987). An experiment on the relative importance of denitrification, nitrate reduction and ammonification in coastal marine sediment *Netherlands journal of sea research* **21** (3) 171–175.
[52] Leadbeater, B.S.C. and Green, J.C. (2000). Flagellates: Unity, Diversity and Evolution *CRC Press, New York*
[53] Barton, A.D. Taboada, F.G. Atkinson, A. Widdicombe, C.E. and Stock, C.A. (2020). Integration of temporal environmental variation by the marine plankton community. *Marine Ecology Progress Series* **647** 1–16.
[54] Sakshaug, E. (1977). Limiting nutrients and maximum growth rates for diatoms in Narragansett Bay *Journal of Experimental Marine Biology and Ecology* **28** (2) 109–123.

Jacob Strock
University of Rhode Island
215 South Ferry Road, Narragansett, RI
United States
E-mail address: jstrock@uri.edu

Gavino Puggioni
University of Rhode Island
45 Upper College Road, Kingston, RI
United States
E-mail address: gpuggioni@uri.edu

Susanne Menden-Deuer
University of Rhode Island
215 South Ferry Road, Narragansett, RI
United States
E-mail address: smenden@uri.edu