Journal of Marine Research, 64, 881–914, 2006

Accounting for unresolved spatial variability in marine ecosystems using time lags

by Philip J. Wallhead[1,2], Adrian P. Martin[1], Meric A. Srokosz[1], and Mike J. R. Fasham[2]

ABSTRACT

The formulation and calibration of models is a vital method for probing and predicting the behavior of marine ecosystems. The ability to do this may suffer, however, if the calibrating data set is subject to significant spatial variability between samples that is not resolved in the model. We propose that some of this variability might be accounted for by variable time lags between sampled water masses which are otherwise assumed to follow a common pattern of ecosystem variability (dynamical trajectory). Using twin tests of fitting models to simulated data sets, we show that realistic levels of meso/sub-mesoscale variability in time lags may have significant distortion effects on the parameter fits from standard methods which do not account for it. The distortion is such as to 'smooth out' or underestimate the magnitude of temporal variability within sampled water masses, causing loss of accuracy and robustness of biological parameter estimates and functions thereof (e.g. gross primary production). A new method of model fitting is shown to avoid these effects, allowing improved estimates over a broad range of spatial time lag variability and measurement noise levels, assuming accurate estimation of the time lag variance, for which we also suggest a method.

1. Introduction

Marine ecosystems exhibit variability on a broad range of temporal and spatial scales. Temporal variability tends to be dominated by seasonal cycles, but may also be significant on timescales of days (e.g. biological interactions) and years (e.g. El Niño effects). Spatial variability is strongest over the largest scales of oceanic gyres ($O(10^3)$ km), but also pervades the meso/sub-mesoscale (1–100 km) and to a lesser degree yet smaller scales, giving rise to the 'patchy' distributions of plankton long-observed in the ocean (Bainbridge, 1957). This variability may have significant impact on estimates of quantities of practical interest such as primary productivity (Martin and Richards, 2002) which in turn affects higher trophic levels (e.g. fish recruitment), and nutrient uptake/export rates important in global nutrient cycles (Prunet and Minster, 1996).

[1] National Oceanography Centre, University of Southampton, Southampton, SO14 3ZH, United Kingdom.

[2] Corresponding author. ***email: pjw5@noc.soton.ac.uk***

The ability to predict and to test hypotheses concerning these important quantities requires the construction of ecosystem models. These models must be fitted to oceanographic data, of limited quantity and quality, with often a significant number of poorly known biological model parameters which must therefore be estimated from the data. Accurate hind-/forecasts and hypothesis tests rely on good model formulations with accurate estimates of the free parameters. Good formulations should contain a degree of complexity/resolution that is supported by the data: too little, and important processes may be misrepresented or not captured in the model, resulting in highly biased parameter estimates (underfit); too much, and surplus freedom in the model may be fitted to noise in the data set, resulting in highly variable, poorly constrained estimates (overfit—see e.g. Burnham and Anderson (1998) for further discussion).

Accounting for all the important influences of spatio-temporal variability on marine ecosystems is particularly challenging, because the advective/diffusive movements of the fluid, and limitations on our ability to track them, result in a mixing or confusion of spatial and temporal variability in the sampled data set. This applies to all marine data sets, whether acquired from cruise ships, fixed mooring stations, satellites (which only measure near-surface concentrations), or 'Lagrangian' sampling surveys (which cannot track the mixing and dispersion of water masses over sustained periods—even the weekly scale of a short cruise). Neither can we rely on physical circulation models to separate the two sources of variability, since these too are subject to limitations of resolution and data constraint (uncertainty in forcings and initial and boundary conditions, and model error due to subgrid-scale parameterizations). Yet, we would like to somehow filter out or allow for the effects of unresolved spatial variability on the data set, so that the fitted biological model may accurately represent the temporal variability within water masses (the 'Lagrangian trajectory'), rather than the combination of spatial and temporal variability represented in practical (non-Lagrangian) data sets. A biological model so-fitted may provide more accurate tests of biological hypotheses, and better hind-/forecasts of mean ecosystem variables when combined with a model of the spatial variability, perhaps involving a physical model of the mesoscale circulation or even a General Circulation Model.

A first approach to modeling marine ecosystems is to neglect spatial variability within a region of interest, and fit a 'single box model' to the regional data (i.e. a set of ODEs, one for each field of interest). This itself is a challenging problem: first, because the dearth of reliable *a priori* information about the biological model results in a significant number of poorly known parameters (typically O(10)), which require a high-dimensional search for optimal values; second, because the nonlinearity of the biological system can result in multiple 'local minima' within the range of *a priori* uncertainty in the parameter values when searching for optimal parameter sets. Nevertheless, progress has been made using nonlinear regression techniques, first applied in this context by Matear (1995) and Fasham and Evans (1995). Matear (1995) demonstrated the powerful Simulated Annealing method to estimate optimal parameters for ecosystem models with O(10) free parameters. The variance-covariance matrices of the optimal parameters were also estimated, revealing groups of covarying estimates and hence limitations of the observational data to constrain more than 10 free parameters independently in any of the 3, 4 and 7 compartment models used to fit the data. Such underconstraint was also observed in the single box model studies of Fennel *et al.* (2001) and Spitz *et al.* (1998). In Fasham and Evans (1995), by contrast, a single box model employing the 7-component formulation of Fasham *et al.* (1990) provided poor fit to data despite having 28 free parameters, although it was not optimized using such powerful techniques as Simulated Annealing. Thus no firm conclusions have been reached on how to optimally formulate and fit a model to available oceanographic data whilst not explicitly accounting for spatial variability.

The single box approach has been extended to account for large-scale variability between spatially separate regions of the ocean, using models consisting of multiple, non-interacting boxes with different ecosystem parameter sets. This was effectively the approach of Hemmings *et al.* (2004). In such model fits, at least some of the ecosystem parameters (coefficients, forcing or initial conditions) must be allowed to differ between regions (boxes) to allow for their different ecologies, whilst parameters common to several regions may be better constrained than for single data set fits. In general, such global models require higher complexity (measured by number of parameters) in order to maintain quality of fit to data in contrasting regions, whilst this level of complexity may not be optimal for fitting the data from one particular region (on a smaller spatial scale). For instance, the success of Hurtt and Armstrong (1999) in simultaneously fitting the biological model in Hurtt and Armstrong (1996) to data from the Bermuda Atlantic Time-series Station (BATS – 31.67N 64.17W, North Atlantic) and Ocean Weather Station India (OWSI – 59N 20W, North Atlantic) was partly attributed to the inclusion of iron limitation to accommodate conditions at OWSI. Hemmings *et al.* (2004) also achieved significant improvement in fit to validation data by allowing box model parameters to take different values in different spatial 'domains' within the total sampled area. However, whilst such multiple-box fits, employing variable ecologies, may address spatial variability on the broad scale (over O(1000) km or more), it would not seem reasonable or practical to use the same approach to account for meso/sub-mesoscale variability in ecologically homogeneous regions.

Finer scale spatial variability may be accounted for by subdividing the boxes into grid cells, and coupling the biology to physical circulation models (since advective and diffusive transports have significant mean effects on the meso/sub-mesoscale). For example, Schartau and Oschlies (2003) fitted a vertically-resolved 1D ecosystem model to data sets from 3 locations (BATS, OWSI and NABE – North Atlantic Bloom Experiment, 47N 20W, North Atlantic) using fixed physical forcing extracted from an independently fitted 3D circulation model. The authors inferred general deficiencies in their biological model—mainly, the use of fixed parameter values common to all three sites, and the formulation of light limitation on photosynthesis. However, it was deemed necessary to take monthly averages of data samples (sacrificing temporal resolution) in order to reduce the sensitivity of the model fit to 'small phase errors' in the biological dynamics, arising from errors in the physical circulation model such as misplaced eddies. This exemplifies the general problem that driving spatial variability with a high resolution physical model introduces a very large number of (possibly time-varying) parameters known with finite error, which, strictly speaking, should therefore be fitted to the total ecosystem data together with the biological parameters. Such a model fit would be not only computationally impractical, but the number of free parameters may be so high that overfit becomes inevitable, resulting in poorly constrained parameter estimates with high mean-square error. On the other hand, failing to allow any freedom in these parameters may result in optimal biological parameter estimates which are biased and spuriously precise.

Therefore, the above methods of fitting marine ecosystem models to available data struggle to robustly account for potential meso/sub-mesoscale variability, without overfitting the free parameters or sacrificing temporal resolution. We demonstrate that if spatial variability in the phase or 'dynamical time' of the ecosystem is significant, then the optimal biological trajectories and parameter sets obtained using existing techniques may be consistently biased as estimates of their 'true', small-scale, Lagrangian counterparts. We also demonstrate a strategy to account for this kind of unresolved (or poorly resolved) spatial variability by assuming random between-sample time lags imposed on a common biological dynamic (equation coefficients, and hence dynamical trajectory). The method is equivalent to allowing all the initial conditions of the biological variables in each sampled water mass to vary in the model fit, but only along a common dynamical trajectory, thus limiting the free parameter : sample size ratio to the number of state variables as the number of samples becomes large.

This study was largely inspired by a cruise data set from the eastern North Atlantic collected after a period of unsettled weather. It was suggested that a significant proportion of the scatter seen in the data might be accounted for by time lags, since the time series data appeared less scattered when plotted in 'phase space' (the space of state variables plotted against each other) in which scatter due to time lags is suppressed (see Fig. 1 and Srokosz *et al.* (2003) for more details). In fact, the latter study showed that time lags may even assist the trajectory delineation in phase space, compensating for any intermittency in the sampling times. Naïvely, this may suggest that models would be better fitted to the phase space trajectory whenever time lags are present. Doing so, however, effectively discards all temporal information, and consequently any timescale parameter estimates (e.g. the duration of a phytoplankton bloom) would have infinite variance. Our method is a more general 'middle way' between time series and phase space fitting, relaxing constraints on sampled time lags to a controlled extent in order to allow for this kind of variability between sampled water masses.

In this study we do not address the issue of to what extent spatial variability in real data can be attributed to time lags. Before answering that question it is necessary to develop the techniques needed to diagnose the phenomenon if it is there, and assess the potential dangers of failing to correct for it. This is done using twin tests to investigate the potential impact of between-sample time lags on standard model fits which do not account for them, then developing a fitting technique which allows for the presence of finite, random between-sample time lags. This new 'variable lag fit' is shown to be capable, given certain assumptions, of significantly outperforming the conventional 'zero lag fit' for a broad range of time lag and measurement noise levels, where performance is measured by the ability to recover, or hindcast, the true (Lagrangian) biological dynamics from practical (non-Lagrangian) data. The zero lag fit is shown to be biased by the confusion of temporal and spatial variability in the data set in such a way that the inferred temporal variation of biological variables is 'smoothed out'. The variable lag fit incurs significantly less bias in estimating the Lagrangian biological dynamics, assuming the true level of time lag variability is accurately estimated prior to fitting. We also discuss a possible method of estimating this level of variability from real oceanographic data sets, in order to realize the potential benefits of the variable lag fit demonstrated in this paper.

The paper is structured as follows. Section 2 details our assumptions (including the model equations), methods of generating artificial data and fitting the model to it, and how we evaluate 'fit performance'. Section 3 discusses results from the twin tests and their robustness to changes in 'truth', 'model' and sampling conditions, and potential extensions of the method for practical applications. Conclusions are drawn in Section 4.

2. Methods

a. *Lagrangian biological model*

We wish to make conservative assumptions about the quantity and quality of available data in practical applications. In these circumstances, a complex model would be in danger of overfitting the data set (fitting parameters controlling mean trends to noise variability in the data set). For this reason, and for convenience of analysis and numerical implementation, we choose to perform our twin tests on a simple (though potentially applicable) marine ecosystem model. A further requirement is that the modeled mean trajectory may yield at least some significant information about the model parameters. This rules out trajectories for which the temporal variability is not significant relative to measurement noise over the sampling period. Our chosen system executes free, periodic, stable nonlinear oscillations with a frequency ω, which is some unknown function of the internal model parameters. This behavior is observed in some more complex marine ecosystem models fitted to real data sets (Ryabchenko *et al.*, 1997). However it is not necessary to assume that such limit cycles occur in reality, nor are they a prerequisite for applying our technique; they merely provide a convenient regime for demonstration, and serve as a simple proxy for seasonal variability in the more complex and generic case of external seasonal forcing (to be explored in later work). The technique is essentially a 'type II regression' and as such may be applied to any system of ODEs with an independent variable (in our case time) which is assumed to be subject to finite noise variations. Whether in fact the between-sample variability in the dynamics may be robustly described by time lags concerns the particular system of interest (and is unlikely to be true, for instance, if the system dynamics are chaotic, or if the system behavior varies significantly over the sampled region).

Table 1. Model structural parameters, true values and relative sensitivities.

Parameter	Symbol	'True' value θ_i^t	Sensitivity
maximum P growth	u_0	1.0 g C m^{-3} day^{-1}	high
nutrient half-saturation constant	k_n	0.03 g C m^{-3}	medium
P self-shading coefficient	c	2.0 g C m^{-3}	medium
Z excretion fraction	γ	0.33	medium
maximum Z grazing rate	R	0.6 day^{-1}	high
Z search efficiency	Λ_z	20 g^{-1} m^3	high
P recycled losses	m_p	0.15 day^{-1}	medium
Z remineralization fraction	r_z	0.5	low
Z mortality rate	m_z	1.5 g^{-1} m^3 day^{-1}	high
cross-thermocline exchange rate	k	0.05 day^{-1}	medium
N concentration below mixed layer	N_0	0.6 g C m^{-3}	medium
P sinking loss rate	s_p	0.04 day^{-1}	low
Z assimilation efficiency	e_z	0.25	low
Nutrient initial condition	$N(0)$	0.131 g C m^{-3}	low
Phytoplankton initial condition	$P(0)$	0.0398 g C m^{-3}	medium
Zooplankton initial condition	$Z(0)$	0.0750 g C m^{-3}	medium

Hence we choose the following NPZ model, adapted from Edwards and Brindley (1996), adapted in turn from the model of Steele and Henderson (1981):

$$\frac{dN}{dt} = -u_0 \frac{N}{k_n + N} \frac{1}{1 + cP} P + \gamma R(1 - e^{-\Lambda_z P})Z + m_p P + r_z m_z Z^2 + k(N_0 - N) \quad (1)$$

$$\frac{dP}{dt} = u_0 \frac{N}{k_n + N} \frac{1}{1 + cP} P - R(1 - e^{-\Lambda_z P})Z - m_p P - (k + s_p)P \quad (2)$$

$$\frac{dZ}{dt} = e_z R(1 - e^{-\Lambda_z P})Z - m_z Z^2 \quad (3)$$

where N, P, Z are nutrient (N), phytoplankton (P) and zooplankton (Z) concentrations measured in g C m^{-3} (carbon currency). The model parameters and their 'true' values are listed in Table 1, again, adapted from Edwards and Brindley (1996), where the original set was chosen from realistic ranges based on a review of literature sources. The functional forms parameterize various biological processes and shall not be discussed at length here (see Edwards and Brindley (1996) for detailed argument). Briefly, the trophic transfers are driven by: Michaelis Menten-parameterized uptake of nutrient by phytoplankton limited by self-shading, Ivlev-parameterized grazing of phytoplankton by zooplankton (replacing the Hollings III function used in Edwards and Brindley (1996), which may produce excitable behavior and hence undesirable trajectory sensitivity, with the grazing form used by Franks *et al.* (1986)), and predation of zooplankton by higher trophic levels. Recycling terms account for remineralized products of phytoplankton mortality/respiration, and zooplankton excretion and mortality fractions. Non-recycled loss terms represent phytoplankton sinking and vertical mixing of nutrient and phytoplankton, the latter also allowing nutrient influx from below the seasonal pycnocline, with the model only explicitly representing shallower waters. The robustness of our results to changes in the choice of model formulation is discussed in Section 3a*v*.

The relative sensitivity of the true solution to changes in each of the parameters may be assessed by computing the maximum increase in model-data discrepancy for a standard time series fit to data generated by the true parameter set, incurred by varying each parameter in turn from 90 to 110% of their 'true' values. Table 1 identifies the model parameters, their true values and respective sensitivities. 'High' sensitivity parameters incurred more than 10^3 units of 'cost', 'medium' sensitivity parameters incurred 10^2–10^3 units, and 'low' sensitivity parameters less than 10^2 units.

Note that (1–3) describe the 'average' dynamics over a constant mixed layer depth and some horizontal scale, which should be reflected in the optimal parameter values. For zero lag fits, the minimum horizontal scale represented by the optimal parameter set will be given by the total area of fluid spanned by the samples. By allowing variable time lags, this lower limit may be reduced to a smaller scale on which the sampled dynamics may be described a common trajectory with variable time lags. This is a much weaker restriction on the model, allowing the data to express a finer scale of spatial resolution in the optimal parameter set. The model is therefore 'Lagrangian' on a scale determined by the data set, subject to this restriction.

b. Between-sample variability in time lag

We assume that spatial variability imposes time lags on the non-Lagrangian data set. These time lags are generated by a 'true' time lag model of independent random variables drawn from a stationary normal distribution with mean zero and variance σ_τ^2. Though motivated by simplicity, this model may serve as a good first approximation in arguably realistic circumstances. The assumption of normality is justified by a Central Limit Theorem argument, considering the net effect of the many independent time lagging mechanisms on the biological dynamics of Lagrangian water masses, such as fluxes of material due to meteorological disturbance, mixed layer depth variation, and between-sample disturbance (weather) events. The assumption of a stationary distribution requires that the sampling interval be small relative to possible changes in time lag variance, which seems reasonable for data sets spanning O(month), but may be unrealistic over a seasonal cycle. The assumption of sampled time lag independence is a reasonable approximation if the interval between successive samples is longer than the time for the largest coherent fluid structure (eddy) to pass through the sampler. For spatial surveys sampling from a cruise vessel (such as produced the data in Fig. 1, see Srokosz *et al.* (2003)), the interval of 1 day used in this study should be adequate at mid-latitudes (where eddies are less than ~100 km in size), although lower-latitude surveys may require a lower sampling frequency for this model to be realistic.

To choose a realistic range of time lag variability to use in our twin tests, we gain a very rough estimate from Figure 1 by examining the vertical scatter about the mean value over

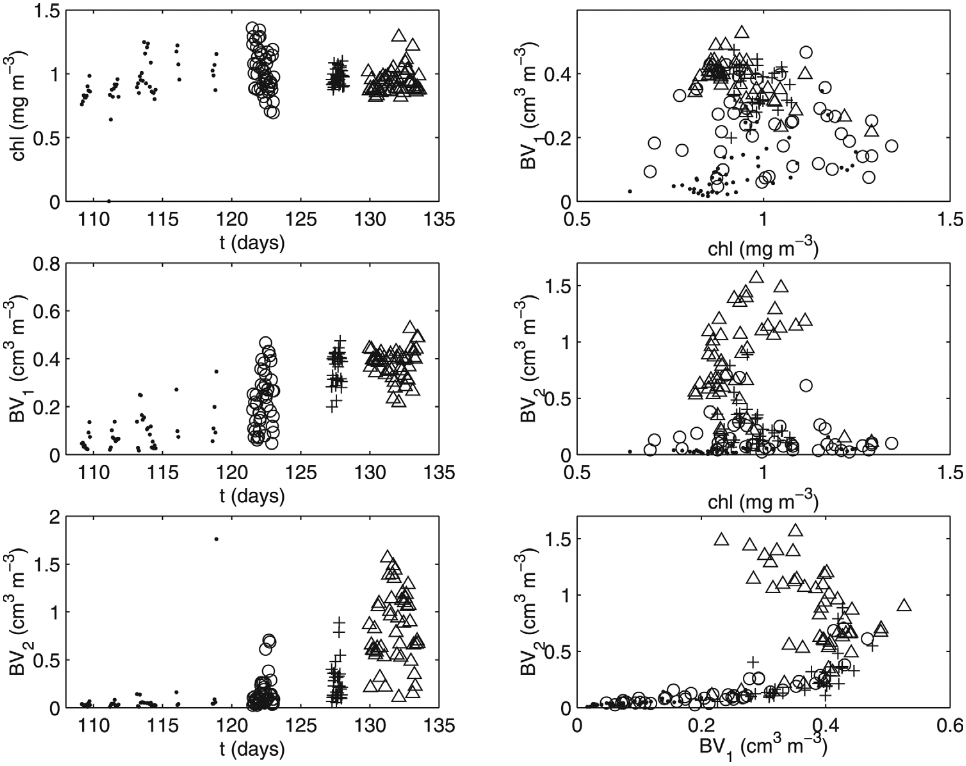

Figure 1. Data from RRS Discovery cruise D227 (with permission, from Srokosz *et al.* (2003)) displayed as time series (left), and in state variable phase space (right). Mixed layer concentrations of chlorophyll (*chl*) and two size classes of zooplankton (BV_1 = 250–500 μm and BV_2 = 500–1000 μm biovolume) are shown, with different symbols used for different sampling subintervals.

a short interval of one of the sampled variables (e.g. the circles), and estimating the time interval over which this mean value changes by an equal amount (cf. the triangles). We thereby estimate 10 days as an approximate upper limit to the time lag standard deviation relevant to the ocean. A more precise estimate would seem to require fitting an ecosystem model without allowing distortion due to time lags, which begs the question of this study. A possible method to allow estimation of this parameter, which is treated as fixed in this study, is discussed in Section 3b*iii*.

Note that even if these assumptions are not strictly met in reality, they may yet prove adequate as assumptions in estimating the Lagrangian trajectory and biological parameters from the data (see Section 2f). We return to this question in Section 3b*vi*.

c. *Measurement and model errors*

To generate the data, we adopt the 'true' error model that all measurement errors are independent random variables sampling from a normal distribution with zero mean and standard deviation proportional to the modeled mean value (after time lagging). Negative sampled values are disallowed, effectively truncating the true error distribution, although this was a rare occurrence for our chosen trajectories and noise levels. No covariance is assumed between measurement errors in different state variables in the same sample, between the same state variables in different samples, or between measurement errors and time lags. This is an extension of a standard model for measurement error often assumed in fitting (cf. Hurtt and Armstrong, 1996). The assumption of normality is justified again by the Central Limit Theorem; the stationarity, independence and zero mean assumptions imply that the measuring instruments perform consistently and without significant systematic error over the course of the sampling survey, which is a plausible (if slightly optimistic) assumption. The lack of covariance between measurement errors in state variables in the same sample effectively assumes independent measurements. Model error is assumed to be indistinguishable from measurement error viz. no covariance or non-Gaussianity contribution is included in the total error. This latter assumption might be challenged, since stochastic model errors may induce correlation between total errors and between these and the time lags. The error model assumed in fitting (Section 2f) is identical to the true error model barring the (small) truncation effect. Though this model may be questionable as a realistic true error model, yet again we argue that its use in fitting may not significantly impair the estimation method when the assumptions fall short of truth (see Section 3b*vi*).

d. Generating a noisy, non-Lagrangian data set

The ‘true’ biological dynamics used to generate the data are specified by (1–3) and the ‘true’ parameter set $\underline{\theta}^t$ shown in Table 1. Note that the superscript t will be used to denote the ‘true’ value throughout. The equations are integrated using fourth order Runge-Kutta with a step size of 1/128 days. The initial conditions $[N(0), P(0), Z(0)]$ result from running the model for 710 days from an arbitrary but fixed set of starting values to eliminate any transient, so that the model is executing repeated limit cycles of period roughly 40 days. The model is run for a further integration time of $\mathrm{T}_I = 80$ days and the output is recorded at intervals of 1/8 days, producing a stable trajectory over the sampling interval $\mathrm{T}_s = 40$ days, with 20 days of integration either side to allow for time lagging. A data set of observations for each field is then generated by sampling the model output at a sequence of $\mathrm{N}_s = 40$ true ‘dynamical times’ given by:

$$\underline{t}'^{t} = \underline{t} - \underline{\tau}^{t} \tag{4}$$

where $\underline{t}$ is a 40-dimensional vector of sampling times as measured by the sampler, starting at 20 days of integration time and advancing uniformly at intervals of 1 day, $\underline{\tau}^t$ is the vector of true time lags drawn from a Gaussian probability distribution with mean zero and standard deviation σ_τ at a resolution of 1/8 day, and $\underline{t}'^{t}$ is the resultant vector of true sampled dynamical times. Time lags of magnitude greater than 20 days are disallowed, effectively truncating the distribution to a range of roughly one period. Note that we assume simultaneous measurements of all modeled variables at each sampling time, which will allow time lagging to be most easily distinguished from other sources of noise, albeit this may be difficult to obtain from real data sets. To simulate measurement error, we add independent Gaussian noise to each variable datum, with constant proportional variance $(s_j y^t_j(t'^t_i))^2$ where $y^t_j(t'^t_i)$ is the true mean value of the j^{th} variable at true sampled dynamical time t'^t_i and s_j is the true fractional error in the j^{th} variable (following the measurement error model of Hurtt and Armstrong (1996)). There is no correlation between state variable measurement errors or between measurement errors and the time lag random variables.

e. Trajectory and parameter estimation

We start with a noisy, non-Lagrangian artificial data set where the spatial variability in biological dynamics is generated by time lags. Our problem is to use it to obtain optimal estimates of the true Lagrangian trajectory y^t and associated parameters $\underline{\theta}^t$ starting with plausible initial guesses $\underline{\theta}^i$, assuming that we have the correct biological model formulation given by (1–3), the correct level of *a priori* uncertainty in the dynamical time of samples represented by the modeled variance σ_τ^2 and the correct *a priori* fractional uncertainty s_j. We assume no prior information about y^t and $\underline{\theta}^t$ except that all trajectory values and parameters are non-negative.

First, we attempt to fit the data using a standard time series fitting technique, which we call the Zero Lag Fit (hereafter ZLF). This implicitly assumes the null hypothesis that there is no significant between-sample variability in time lag, hence the modeled time lag vector $\underline{\tau}^m$ is set to zero, and the fit is independent of the *a priori* uncertainty σ_τ. Then we repeat the procedure using the new technique, which we call the Variable Lag Fit (hereafter VLF). This assumes that $\underline{\tau}^m = \underline{\tau}$, where $\underline{\tau}$ is a (possibly non-zero) time lag vector constrained by the data and the *a priori* uncertainty σ_τ. The ZLF must interpret the scatter in the data set generated by time lags as measurement errors in the state variables, which are, by hypothesis, uncorrelated between state variables. The VLF can potentially distinguish time lags from measurement error, using the fact that time lags produce correlated changes in the state variables (along the phase space trajectory), and thereby achieve a better estimate of the Lagrangian dynamics. However, optimizing the VLF does involve added complications beyond those of optimizing the ZLF, as will be seen below.

We employ Bayesian Estimation (BE) in this study as a general method to infer model parameters from the data. We maximize the 'posterior probability' of the model parameter set by maximizing a product of the Likelihood, the probability of the data given the parameter set (hypothesis), and the 'prior probability' of the parameter set. The latter prior probability density is assumed to be uniform in all parameters except the time lags (see below). Thus we obtain Bayesian estimates $\hat{\underline{\theta}}$ of the 'true' parameter values $\underline{\theta}^t$. An alternative interpretation of our method is as Maximum Likelihood Estimation where the Likelihood function includes errors in the regressor variable (time)—hence it is a 'nonlinear model II regression' type with 'controlled' regressor errors in the terminology of Laws (1997), or a 'functional relationship without replication' in the formal statistical terminology (Seber and Wild, 2003). For the ZLF, our estimates will be classical Maximum Likelihood (ML) estimates. For the VLF, our Bayesian estimates may sacrifice some of the desirable properties of ML estimates as we now discuss.

Assuming that the model parameter set for the ZLF is minimal (containing no two parameters with identical effects), then the ML estimates should be unique, or 'identifiable' (Stuart *et al.*, 1999). As we are using the correct model formulation, these ML estimates will also be 'consistent', meaning that as the number of observations becomes large, we can expect them to converge on the true values. However, ML estimates are not in general 'unbiased' (bias being defined as the expected difference between the ML estimates and the true values given a finite data set). We would also like our estimates to be 'efficient', meaning that they have minimal variance over different data sets. For the ZLF, fitted to data without time lags, the ML estimates will be asymptotically (as sample size $n \rightarrow \infty$) efficient and normally distributed (with variance decreasing as n^{-1}) (Stuart *et al.*, 1999). In fact, the parameter variance-covariance matrix is given asymptotically by the inverse of the 'Fisher information matrix', which may be estimated in practice by the Hessian matrix for parameter perturbations about the ML solution for a single data set (Matear, 1995).

For the VLF, fitted to data with time lags, many of the above desirable properties are not guaranteed. The difficulties stem from the fact that as the number of observations increases, so does the number of 'incidental' free parameters (dimension of $\underline{\tau}^m$), thus theoretical results associated with large sample size may not be applicable. The incidental parameter estimates comprising $\hat{\underline{\tau}}$ may not be consistent, since each of the components only occurs in a finite number of observable variables (Neyman and Scott, 1948). The structural parameter estimates comprising $\underline{\theta}$ remain consistent, but only if the ratio of error variance coefficients s/σ_τ is assumed known (Seber and Wild, 2003). Even if the structural estimates retain consistency, their asymptotic variance-covariance matrix is not necessarily given by the inverse information matrix (Seber and Wild, 2003), and even if it is, the ML estimates may yet not be asymptotically efficient (Neyman and Scott, 1948). Therefore, the VLF parameter estimates may incur larger biases, and lose efficiency and accuracy of confidence interval estimates using the Hessian matrix method (although the latter technique is problematic in practice anyway due to ill-conditioning of the Hessian matrix for underdetermined parameter sets (Fennel *et al.*, 2001; Schartau and Oschlies, 2003), and breakdown of the linearity assumption due to inadequate data). Problems with the Bayesian estimates initially obtained by the VLF eventually prompted the use of a modified maximization function, as detailed below.

f. Maximization functions

The posterior probability quantifies the probability of a parameter set given the model formulation and the data set. It is a measure of the success of the model in fitting the data set for 'reasonable' parameter values, and is maximized by tuning the model parameters (adjusting the 'hypothesis'). This process is called 'calibration'. The two 'submodels' in this study—the ZLF and VLF—are calibrated using ‘calibration cost’ functions which assume that the model explains the dynamics of the noisy, non-Lagrangian data set described in Section 2a, and which are specified formally in this section. The calibrated models are then evaluated in terms of their ability to reproduce the noise-free, Lagrangian data set. This is quantified by a different ‘validation cost’ function described in Section 2h.

In order to construct the calibration cost functions, we specify a general hypothesis defining the VLF, of which the ZLF submodel is a restriction. Our composite hypothesis $\mathcal{H}$ asserts the following:

$\mathcal{H}^{(1)}$: The observational vector of (fixed) sampling times is given by $\underline{t} = \underline{t}'^{m} + \underline{\tau}^{m}$ where $\underline{t}'^{m}$ is a vector of modeled dynamical times (dimension $N_s = 40$) and $\underline{\tau}^{m}$ is the vector of modeled time lags, assumed to be independent Gaussian random variables with mean zero and variance σ_{τ}^{2}.

AND

$\mathcal{H}^{(2)}$: The j^{th} observed variable at sampling time t_i is given by $x_{ij} = y_j^m(t_i'^m|\underline{\theta}^m) + \epsilon_{ij}^m$, where $y_j^m(t_i'^m|\underline{\theta}^m)$ is the output of model (1–3) at time $t_i'^m = t_i - \tau_i^m$ given the set of parameters and initial conditions $\underline{\theta}^m$, and ϵ_{ij}^m is an independent Gaussian random variable with mean zero and variance $(s_j y_j^m(t_i'^m|\underline{\theta}^m))^2$, representing measurement noise and model error.

Note that the time lags are ‘controlled’ in this study because the sampling times comprising $\underline{t}$ are fixed, advancing uniformly at intervals of 1 day with negligible error, as one might expect in a plausible sampling strategy.

For the VLF, since the time lags are by-assumption random variables, their values inferred from one data set cannot be used to predict another. Rather, we should aim to fit the model such that it is a ‘best bet’ for any future set of time lags, viz. we should maximize the posterior probability density integrated (or ‘marginalized’) over all values of $\underline{\tau}^m$:

$$P(\underline{\theta}^m|D) = \int p^{(1)}(\underline{\tau})p^{(2)}(D|\underline{\tau}, \underline{\theta}^m)\mathrm{d}\underline{\tau} \tag{5}$$

where D denotes the observational data, and the assumed independence of $\underline{\tau}$ and the ϵ_{ij}^m allows the product decomposition of the integrand. However, integrating over the space of allowed time lags is computationally impractical even at coarse resolution for more than a few lags. Therefore, we make the standard step of approximating the integral by its maximal value (over $\underline{\tau}^m$) multiplied by some constant:

$$P(\underline{\theta}^m|D) \propto p^{(1)}(\hat{\underline{\tau}})p^{(2)}(D|\hat{\underline{\tau}}, \underline{\theta}^m) \tag{6}$$

Note, the ZLF restricts the optimization to ($\underline{\tau}^m = \underline{0}$), whilst the VLF allows ($\underline{\tau}^m = \hat{\underline{\tau}}$) for $\hat{\underline{\tau}}$ within the allowed range of ±20 days for each component (see Section 2d).

Therefore, our calibration cost function for the standard ZLF is given by Eq. (6), fixing $\hat{\underline{\tau}} = \underline{0}$:

$$M_{\mathrm{ZLF}}^{(C)} = p_{\mathrm{ZLF}}^{(1)} \cdot p_{\mathrm{ZLF}}^{(2)} \tag{7}$$

where

$$p_{\mathrm{ZLF}}^{(1)} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\tau^2}} \tag{8}$$

is the (restricted) prior probability density function of the time lags (a constant), and

$$p_{\mathrm{ZLF}}^{(2)} = \prod_{i=1}^{N} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(s_j y_j^m(t_i^m))^2}} e^{-\frac{(x_{ij}-y_j^m(t_i^m))^2}{2(s_j y_j^m(t_i^m))^2}} \tag{9}$$

is the Likelihood function, combining Likelihoods of the data over the *m* (independent) variables and over the *N* (independent) simultaneous samples of these variables. For the new VLF, the calibration cost function, derived from (6), is first formulated as:

$$M_{\mathrm{VLF}}^{(C)} = p_{\mathrm{VLF}}^{(1)} \cdot p_{\mathrm{VLF}}^{(2)} \tag{10}$$

where

$$p_{\mathrm{VLF}}^{(1)} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\tau^2}} e^{-\frac{(\tau_i^m)^2}{2\sigma_\tau^2}} \tag{11}$$

is the (variable) prior probability density function of the time lags, and

$$p_{\mathrm{VLF}}^{(2)} = \prod_{i=1}^{N} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(s_j y_j^m(t_i'^m))^2}} e^{-\frac{(x_{ij}-y_j^m(t_i'^m))^2}{2(s_j y_j^m(t_i'^m))^2}} \tag{12}$$

is the Likelihood function, now evaluating the model at the modeled dynamical times:

$$t_i'^m = t_i - \tau_i^m \tag{13}$$

In this study, we make no attempt to optimize the statistical structural parameters s and σ_τ. In all optimizations, we use the same value of s during fitting as that used to generate the data. For fitting to data 'generated on the null' ($\underline{\tau}^t = \underline{0}$) we set σ_τ equal to some fixed (possibly wrong) value for each optimization. For data generated 'on the alternative' ($\underline{\tau}^t \neq \underline{0}$) we use the same value of σ_τ as that used to generate the data. Whilst it may be unrealistic to set these *a priori* estimates equal to the true values given the difficulties of estimating these quantities in practice (especially σ_τ), it serves as a starting point which ensures consistent *a posteriori* estimates of $\underline{\theta}^t$ for the VLF (Seber and Wild, 2003), and which may later be generalized to include consistent *a posteriori* estimates of s and σ_τ, perhaps using replicated data or an iterative fitting technique (see Section 3b*iii*).

Now, maximizing $M_{\mathrm{VLF}}^{(C)}$ in (10–13) yields consistent structural parameter estimates given accurate *a priori* estimates of the error parameters (s, σ_τ), but these are not necessarily the 'best' as regards bias and efficiency. The problem is that by relaxing constraints on the sampled time lags we increase variance in the estimated frequency of oscillation, especially given the finite number of iterations in practical numerical optimization routines. To suppress this variance, we multiply the VLF maximization function by a 'lag drift penalty' term S which measures the probability, assuming independent Gaussian time lags with variance σ_τ^S, of obtaining a persistent 'drift' in the time lags:

$$S(\underline{\tau}|\sigma_\tau^S) = \frac{\max(p'_{(0)}(\underline{\tau}|\sigma_\tau^S))}{\max(p'_{(A)}(\underline{\tau}|\sigma_\tau^S))} \tag{14}$$

where $p'^{(A)}$, $p'^{(0)}$ are the maximum Likelihoods for fitting $\tau(t) = at + b$ and $\tau = c$ respectively to the series of time lags plotted against sampling time. With this penalty term our calibration cost function for the VLF becomes:

$$M'^{(C)}_{\mathrm{VLF}} = p^{(1)}_{\mathrm{VLF}} \cdot p^{(2)}_{\mathrm{VLF}} \cdot S(\underline{\tau}|\sigma_\tau^S) \tag{15}$$

$\hat{S} \rightarrow 1$ as number of sampling times $N_s \rightarrow \infty$, so the consistency of the structural parameter estimates is not spoilt by the maximization of $M'^{(C)}_{\mathrm{VLF}}$. Note that $M'^{(C)}_{\mathrm{VLF}}$ is not strictly (proportional to) a probability, since we have used each datum twice in $p^{(1)}_{\mathrm{VLF}}$ and S; nevertheless, maximizing $M'^{(C)}_{\mathrm{VLF}}$ using the entire data set in both $p^{(1)}$ and S was found to give lower validation cost (see Section 2h) and better frequency constraint than partitioning the data set between $p^{(1)}$ and S. Similarly a first choice of σ_τ^S might be σ_τ; yet we obtained better performance using a 'weighted' S by reducing the value of σ_τ^S to $0.1.\sigma_\tau$. We will see that the use of S with the aforementioned weighting improves validation fit and reduces the bias and variance of the structural parameter estimates. Note that such 'Likelihood modifiers' are nothing new in random regressor problems (e.g. Neyman and Scott, 1948). A possible alternative would be to fit a model with the timescales effectively fixed by non-dimensionalization, then estimate the timescales separately (Froda and Colativa, 2005). Our method, however, aims to constrain all structural parameters in a single estimation procedure.

In summary, our maximization functions are $M^{(C)}_{\mathrm{ZLF}}$ defined in (7–9) for the ZLF and $M'^{(C)}_{\mathrm{VLF}}$ defined in (10–15) for the VLF. Unless otherwise stated, we used $\sigma_\tau^S = 0.1.\sigma_\tau$ as a weighting for S in the VLF.

g. *Optimization*

Our task is to maximize (7) with respect to the $\underline{\theta}^m$ free parameter vector for the ZLF (13 rate constants + 3 initial conditions), and (15) with respect to the $(\underline{\theta}^m, \underline{\tau}^m)$ free parameter vectors for the VLF (13 rate constants + 3 initial conditions + 40 time lags). The data set consists of 40 simultaneous measurements of each of 3 variables (hence sample size $n = 120$). For the purposes of optimization, it is more convenient to minimize the 'cost functions': $cost_{\mathrm{ZLF}} = -\log M^{(C)}_{\mathrm{ZLF}}$ and $cost_{\mathrm{VLF}} = -\log M'^{(C)}_{\mathrm{VLF}}$. In the ZLF, the 16 free parameters are all varied by the search algorithm. However, for the 56 free parameters of the VLF we obtain a massive computational saving by performing a nested optimization (similar to 'concentrating the Likelihood') over the 40 incidental time lag parameters at each iteration. This is facilitated by the fact that the $cost_{\mathrm{VLF}}$ decomposes (except for the small contribution of $-\log S$) into a sum of contributions from each sample time, allowing each τ_i to be optimized by choosing the value from the allowed range for which the cost increment is minimal. The contribution of $-\log S$ is then calculated and added to give the (partially) optimized cost for the trial value of $\underline{\theta}^m$.

The search algorithm varies the 16 structural parameters comprising $\underline{\theta}$ in order to (fully) minimize the cost. We use the 'Downhill Simplex with Simulated Annealing' algorithm of Press *et al.* (1999). Simulated Annealing (SA) algorithms attempt to avoid trapping in local minima by adding random cost fluctuations chosen from a Gibbs distribution with a certain 'temperature' (rms fluctuation). The temperature is then cooled as the optimization proceeds on the expectation that the global minimum is being approached and less fluctuation is required. Given any Markovian transition matrix for exploring the parameter space, if the SA rule for accepting 'uphill' moves with a probability $\propto e^{-\Delta(cost)/T}$ is applied, the equilibrium probability state vector will be a Gibbs distribution over $cost$. It follows that for a slow 'cooling schedule' that varies as $1/\log k$ for k iterations, the search remains ergodic even as $T \to 0$, implying that the global minimum will be approached with certainty as $k \to \infty$. Unfortunately, such cooling schedules are found to be too slow for practical applications (Matear (1995)—although there may be ways to enable a speed-up without sacrificing ergodicity—see Ingber (1993)) and in any case would not necessarily work for our simplex algorithm since it is not a Markovian search (next position dependent at most on last position). After trial-and-error testing, the schedule we found most practically effective was to set $T(k) = \max\,(cost, T_0 e^{-k/\gamma})$, so that the temperature adapts to any improvements in $cost$ beyond a baseline exponential cooling rate of $1/\gamma$.

The optimizations were all limited by a maximum number of iterations. This number was set by the criterion that the maximal 'optimization error' due to finite iterations be much less than the typical between-data set variation for our chosen test statistic $\Lambda^{(V)}$ (see Section 2h for definition). The maximal optimization error was estimated by comparing with a single long optimization of 5×10^5 iterations. It was found that, for model fits to data generated by the alternative hypothesis ($\underline{\tau}^t \neq \underline{0}$), 10^5 iterations safely satisfied this criterion for initial guesses with 10% error in every parameter and σ_τ of up to 8 days. For model fits on the null ($\tau^t = 0$), we found that 10^4 iterations were sufficient as long as the initial guess was close to the optimum, i.e. at $\underline{\theta}^i = \underline{\theta}^t$. Note that because we have finite data with finite noise, the exact optimal (ML) solution $\underline{\theta}^0$ is in general not coincident with the 'true' generating solution $\underline{\theta}^t$, since inevitably some of the noise is fitted by the modeled mean variability in the exact optimal solution. For cases where we were not trying to demonstrate efficiency of the search algorithm for realistically 'wrong' initial guesses, but rather trying to make our numerically approximated optimal solution $\hat{\underline{\theta}}$ as close as possible to the exact optimal solution $\underline{\theta}^0$, we used $\underline{\theta}^i = \underline{\theta}^t$ for speed of convergence. The practical convergence time was finite for the ZLF and VLF in all cases. On the alternative ($\underline{\tau}^t \neq \underline{0}$), the ZLF was slower (necessitating 10^5 iterations) as a result of a larger discrepancy between $\underline{\theta}^t$ and $\underline{\theta}^0$, leading, as will be seen, to larger biases on the parameter estimates.

h. Evaluation of model fit

Our principal measure of fit performance is the expected accuracy in recovering (hindcasting) the true (Lagrangian) biological trajectory. Thus we estimate the relative fit performance using the test statistic:

$$\Lambda^{(V)} = \max\,(\log p^{(V)}_{\mathrm{VLF}}) - \max\,(\log p^{(V)}_{\mathrm{ZLF}}) \tag{16}$$

where $-\log p^{(V)}_{\mathrm{ZLF/VLF}}$ is the 'validation cost' of the ZLF/VLF. The function $p^{(V)}_{\mathrm{ZLF/VLF}}$ is identical in form to $p^{(2)}_{\mathrm{ZLF}}$ (equation (9)), but here the data are generated by $\underline{\theta}^t$ with zero measurement noise and no time lags (the true (Lagrangian) trajectory sampled at times $\underline{t}$), and the fitted model trajectory is generated by the estimates $\hat{\underline{\theta}}_{\mathrm{ZLF/VLF}}$ obtained by the ZLF/VLF in calibration without the inclusion of time lags. The validation cost is therefore a measure of a total squared predictive error, weighted to give greater importance to predictive error where the true values (and hence measurement errors) are low. Note that since (16) is evaluated over an independent (noise-free) validation data set, it does not require correction for the different noise-fitting capacities of the ZLF and VLF (the latter having N_s more free parameters), therefore we use $\Lambda^{(V)} = 0$ as our threshold for selecting the submodel with the highest fit performance.

It is also of interest to compare the accuracy and precision, as determined by the bias and variance, of the fitted parameter vector $\hat{\underline{\theta}}$ for each of the two fits. This defines a set of performance measures for hypothesis testing purposes. We generally avoid combining the biases and variances into single performance measures, since according equal importance to each parameter is an essentially arbitrary, subjective choice (whereas $\Lambda^{(V)}$ weights them somewhat more objectively by their sensitivities in determining the true trajectory).

3. Results and discussion

a. Visual model fit assessment

i. Dependence on time lag variance and measurement noise. For a visual assessment of fit performance, we ran zero and variable lag fits to several data sets generated with time lag standard deviations of $\sigma_\tau = 0, 4$ and 8 days and measurement error coefficients of $s = 0.05$ and 0.15, using $\underline{\theta}^i = 0.9$ and $1.1 \times \underline{\theta}^t$ as 'plausibly wrong' initial guesses. Two examples are shown, with $s = 0.05$ in Figure 2a and $s = 0.15$ in Figure 2b. In both cases, the generating trajectory is well recovered when $\sigma_\tau = 0$ days, but the VLF clearly does a better job when $\sigma_\tau = 4$ or 8 days. The effect of the time lag variance is in general to cause underestimates of the (N, P, Z) temporal variability when fitted using the standard ZLF technique, whilst the VLF seems to accurately recover the extent of variability even at

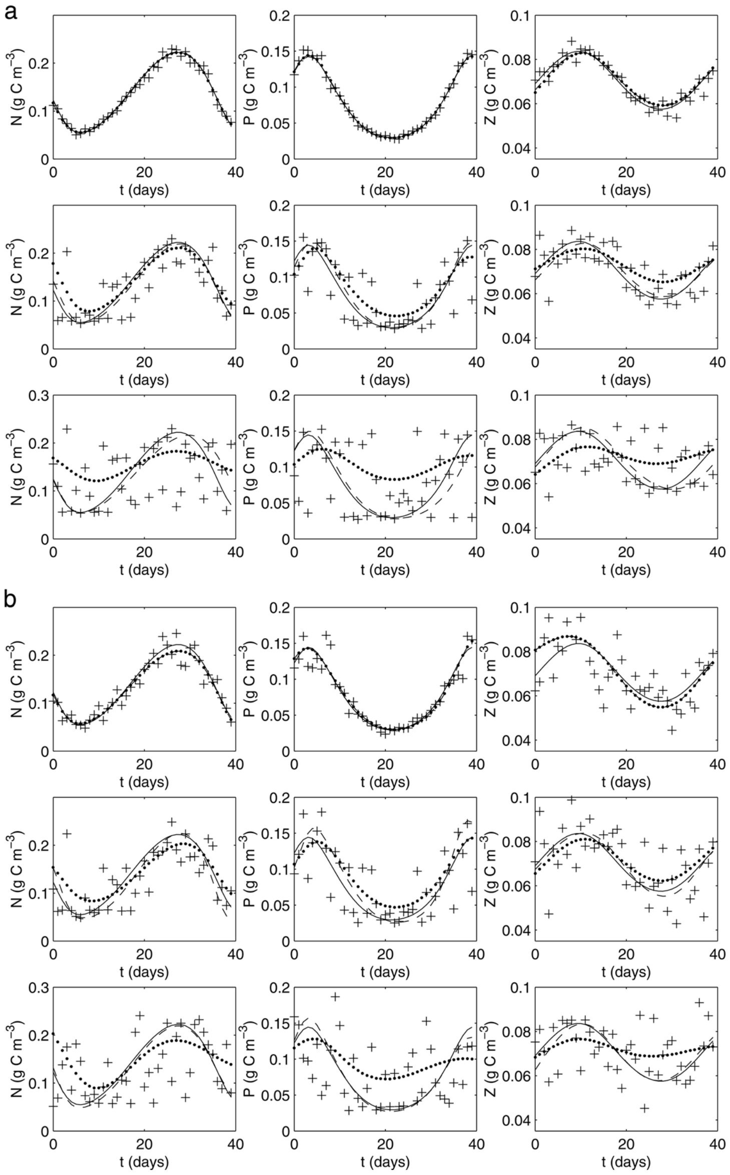

Figure 2. (a) Best fit trajectories to data (+) generated by 'true' (Lagrangian) trajectory (solid line) using standard zero lag method (ZLF, dotted) and using the new variable lag method (VLF, dashed), generating data with time lag standard deviations $\sigma_\tau = 0$ days (upper), 4 days (middle), and 8 days (lower) and 5% measurement noise imposed on each variable ($s = 0.05$). N = Nutrient (left), P = Phytoplankton (middle), and Z = Zooplankton (right). Initial guess parameter error was -10% in each model parameter (see Eqs. (1–3) and Table 1 for specification of the true model). (b) As in Figure 2a but with 15% measurement noise ($s = 0.15$) imposed on each state variable, and initial guess parameter error of $+10\%$ in every parameter.

$\sigma_\tau = 8$ days and s $= 0.15$. The distortion of the ZLF first becomes noticeable over intervals when the (N, P, Z) trajectory curvature $\left(\frac{d^2N}{dt^2}, \frac{d^2P}{dt^2}, \frac{d^2Z}{dt^2} \text{ respectively}\right)$ is high i.e. at the peaks and troughs (see the fits for $\sigma_\tau = 4$ days in Figs. 2a, b). This suggests that intervals in the seasonal cycle such as spring/autumn blooms and subsequent troughs due to flourishing grazers may be 'smoothed out' or underestimated in magnitude by standard model fits.

Note that at $\sigma_\tau = 8$ days the smoothing effect appears to result in a decaying oscillation in the ZLF rather than the true stable limit cycle, and indeed this was confirmed by running the ZLF solution over a longer integration time. It is clear that such mistaken decays to equilibrium would seriously impair the forecast accuracy for the true dynamics after the sampling interval—and may perhaps help to explain the dearth of free oscillations found in models fitted to real data sets in the literature. These errors in the topology/stability of the dynamics are shown more clearly in Figure 3, where the results for $\sigma_\tau = 0, 4$ and 8 days

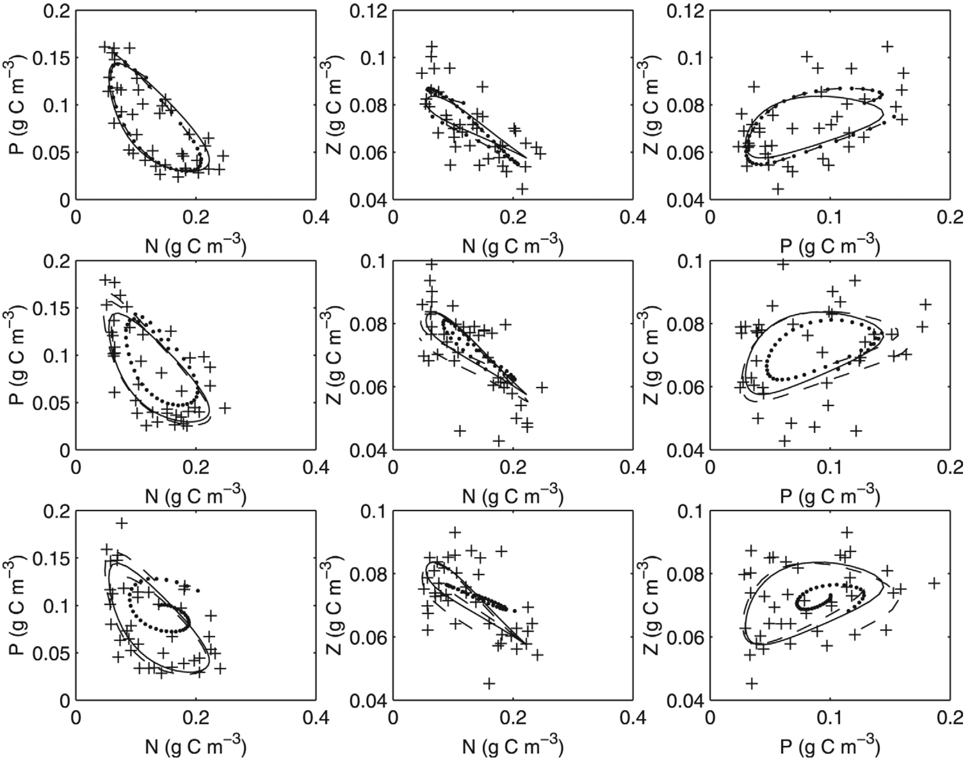

Figure 3. Phase space cross-sections showing best fit trajectories to data (+) generated by 'true' Lagrangian trajectory (solid line) using standard zero lag method (ZLF, dotted) and using the new variable lag method (VLF, dashed), generating data with time lag standard deviations $\sigma_\tau = 0$ days (upper), 4 days (middle), and 8 days (lower) and 15% measurement noise imposed on each variable ($s = 0.15$).

and $s = 0.15$ shown in Figure 2b are replotted in phase space cross-sections. Viewed in this way, dynamical transience is much more apparent: the ZLF seems to produce smaller, yet stable limit cycles at $\sigma_\tau = 4$ days, and spirals towards equilibrium at $\sigma_\tau = 8$ days, whilst the VLF maintains stable limit cycles with a small amount of transience at both $\sigma_\tau = 4$ and $\sigma_\tau = 8$ days.

Comparing Figures 2a and 2b, the increase in measurement noise increases the deviation of the inferred trajectory from the true trajectory of both fits at $\sigma_\tau = 0$ days, whilst it does not appear to significantly alter the trajectories at $\sigma_\tau = 4$ and 8 days. Thus the relative importance of errors due to spatial time lag variability must be a function of the measurement noise level as well as the amount of time lag variation. For high enough measurement noise level, the benefit of using the VLF will be insignificant for realistic levels of time lag variability (although from Figs. 2a, b this does not seem to be the case at $s \leq 0.15$ for $4 \leq \sigma_\tau \leq 8$ days). For any level of time lag variability, the improvement yielded by the VLF will be insignificant when measurement errors become comparable with the extent of mean temporal variability in the data set—in which case any kind of time-dependent model fit is a questionable exercise.

ii. Optimization vs. fit-by-eye. Looking at the data for $\sigma_\tau = 8$ days in Figures 2a, b, one might wonder how the optimizer is able to extract any information at all from such a scattered set. Note, however, that the optimizer considers the model-data discrepancy in all three state variables simultaneously, whilst the eye tends not to do this when examining a set of time series (although this is partly achievable by displaying 2D phase space cross-sections as in Fig. 3, perhaps using symbols or colors to retain some temporal information as in Fig. 1). Thus, in any of the optimizations, at least the correct patterns of variability and phase relationships between N, P and Z are robustly recovered (these being absent in the 'initial guess' trajectory), whilst the extent of variability is underestimated by the ZLF, which tries to follow a smoothed variation over sampling time (within the constraints imposed by the model formulation (1–3)). The VLF has the additional freedom to effectively shift data points forwards and backwards in time (left and right in Figs. 2a, b) to an extent roughly proportional to σ_τ in order to achieve a better fit. Crucially, however, the temporal shift applied to each sample must be the same for all three state variables (which, again, is difficult to perceive by eye in time series data sets).

b. Quantitative model fit evaluation

i. Recovery of the Lagrangian biological trajectory. First, we assess the potential benefits of the new technique when finite time lags are present in the data set. This is done by computing $\Lambda^{(V)}_{(alt)}$ over 5 different data sets, each requiring 10^5 iterations, for $\sigma_\tau = 0, 0.5, 1, 2, 4$ and 8 days and $s = 0.05$ and 0.15 (see Figs. 4a, b). Here the subscript 'alt' denotes use of the alternative hypothesis to generate the data ($\underline{\tau}^t \neq \underline{0}$), and recall that we use the correct value of σ_τ as an *a priori* estimate to explore the maximum potential benefit of the VLF. Also, we set $\underline{\theta}^i = \underline{\theta}^t$ for speed of convergence (note: convergence time is still non-zero for the VLF, since the 'true' solution is generally not equivalent to the optimal solution for finite data fitting). As σ_τ increases, the mean $\Lambda^{(V)}_{(alt)}$ increases as worsening performance of the ZLF, due to negative bias on the estimated extent of variability (the smoothing effect), outweighs the deterioration in the VLF due to timescale variance. This bias arises as the ZLF tries to minimize vertical scatter and hence fit to the smoothed variability over sampling time, which is a convolution of the true trajectory with the true time lag distribution.

From Figures 4a, b, we estimate that the maximum mean saving in validation cost achieved by the VLF (at $\sigma_\tau = 8$ days) is about 450 units for $s = 0.05$ and about 350 units

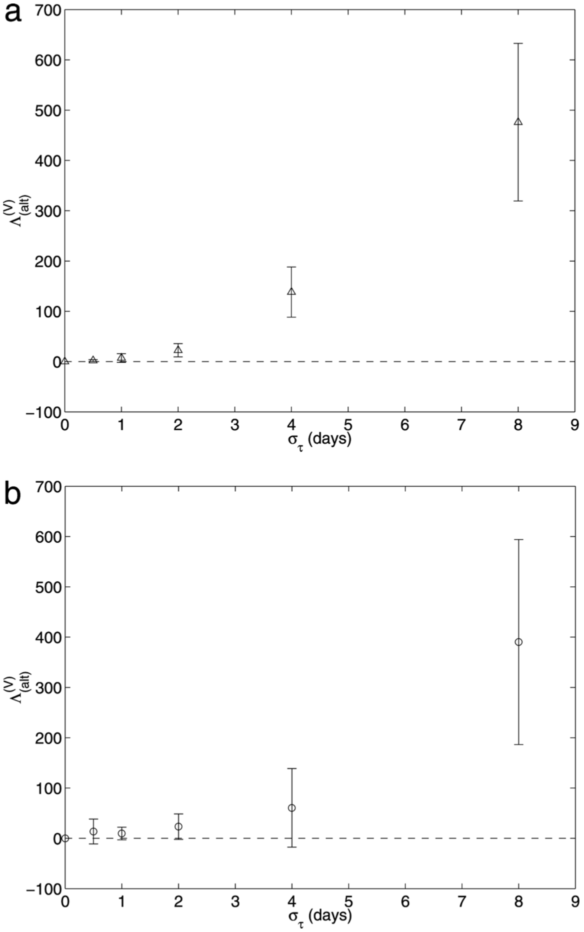

Figure 4. (a) Validation cost savings yielded by new technique $\Lambda^{(V)}_{(alt)}$ (triangles) vs. time lag standard deviation σ_τ when time lags in calibrating data set are non-zero (the 'alternative' hypothesis $\underline{\tau}^t \neq \underline{0}$) and 5% measurement noise is imposed on all three state variables ($s = 0.05$). (b) As in Figure 4a but with 15% measurement noise imposed ($s = 0.15$).

for $s = 0.15$—or 750% and 580% respectively of the expected true validation cost due to measurement noise ($n/2 = 60$ units). The mean $\Lambda^{(V)}_{(alt)}$ strays beyond one standard deviation of zero (VLF performs better in more than 70% of cases) at $\sigma^c_\tau \approx 1$ days for $s = 0.05$ and $\sigma^c_\tau \approx 4$ days for $s = 0.15$. It seems, therefore, that the minimum lag variability required for significant improvement in fit performance with the VLF scales roughly in proportion to the measurement noise level (at low s) and that these conditions are not unrealistic for marine ecosystem sampling.

Second, we assess the potential dangers of using the new technique by comparing with the ZLF when time lags are in fact absent in the data set ($\underline{\tau}^t = \underline{0}$). This was done by computing $\Lambda^{(V)}_{(null)}$ over 100 different data sets (each requiring 10^4 iterations) for $\sigma_\tau = 0, 0.5, 1, 2, 4$ and 8 days and $s = 0.05$ and 0.15 (see Figs. 5a, b). Here the subscript 'null' denotes data generated on the null hypothesis ($\underline{\tau}^t = \underline{0}$). We see from the means and standard deviations that the VLF performs persistently worse and is less robust than the ZLF when time lags are not present. This is to be expected, since in this case the ZLF suffers no smoothing effects whilst the VLF produces increasingly variable timescale estimates as σ_τ is increasingly overestimated.

From Figures 5a and 5b, we estimate that the maximum mean increase in validation cost incurred by using the VLF (at $\sigma_\tau = 8$ days) is about 1.2 units for $s = 0.05$ and about 35 units for $s = 0.15$—or 2% and 60% respectively of the true cost due to measurement noise. Thus the expected error of the VLF in recovering the true Lagrangian trajectory is a strong function of measurement noise s, which increases the tendency of the VLF to give distorted timescale estimates.

Note also, that since the variance in $\Lambda^{(V)}_{(null)}$ increases roughly in proportion to the decrease in the mean, the mean stays within one standard deviation of zero (over our realistic range of σ_τ), which implies that the ZLF never performs better in more than roughly 70% of data sets, even when time lags are truly absent.

Now we can summarize the potential risks vs. benefits of using a VLF in the more realistic circumstances where σ_τ is only known to be some value less than ~8 days. At 5% measurement noise ($s = 0.05$), we stand to incur an increase of roughly 2% but save potentially 750% on average of the true validation cost due to measurement error alone ($= n/2$), assuming ($\underline{\tau}^t = \underline{0}$) is a 'worst case scenario' for the VLF. At 15% measurement noise ($s = 0.15$), we stand to incur an increase of roughly 60% but save potentially 580% on average. In summary, the risks appear to be outweighed by the potential benefits of using the VLF, although to fully realize this potential, one must accurately estimate σ_τ, which we will discuss in Section 3b*iii*.

We found that qualitatively similar plots to Figures 4a, b were obtained using the logarithm of the ratio of the optimal Likelihoods $p^{(2)}_{\mathrm{ZLF/VLF}}$ (the Likelihood Ratio Test LRT) as a test statistic (see Figs. 6, 7):

$$\Lambda^C = \max\left(\log p^{(2)}_{\mathrm{VLF}}\right) - \max\left(\log p^{(2)}_{\mathrm{ZLF}}\right) \tag{17}$$

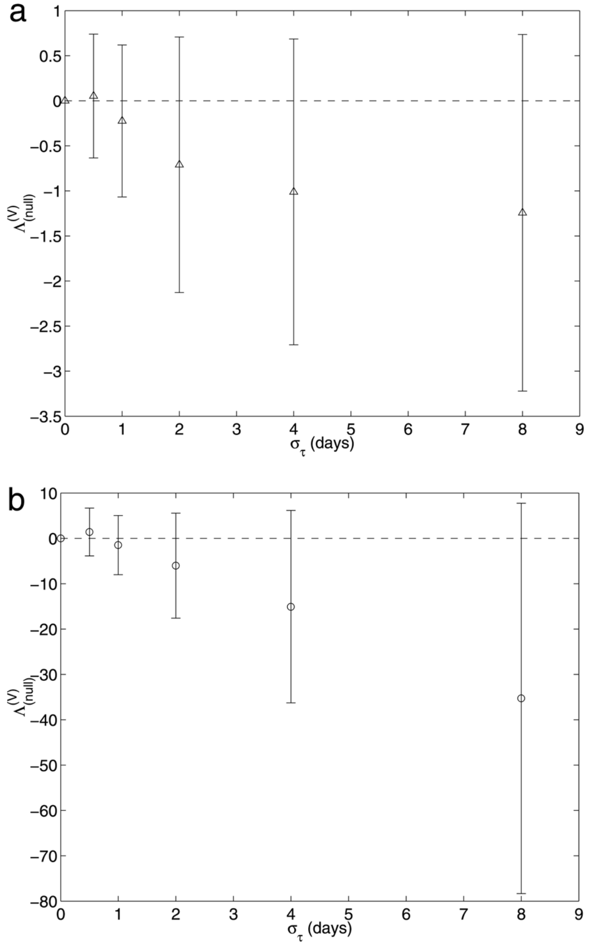

Figure 5. (a) Validation cost savings yielded by new technique $\Lambda^{(V)}_{(null)}$ vs. modeled time lag standard deviation σ_τ when time lags in calibrating data set are zero (the null hypothesis $\underline{\tau}^t = \underline{0}$) and 5% measurement noise is imposed on all three state variables ($s = 0.05$). Means (triangles) and standard deviations (error bars) are shown from 100 optimizations over different calibrating data sets for each value of σ_τ. (b) As in Figure 5a but with 15% measurement noise imposed ($s = 0.15$).

Comparing Figures 6 and 7 allows rejection of the null hypothesis ($\underline{\tau} = \underline{0}$) at similar threshold values of the time lag standard deviation σ^c_τ as those quoted above (and roughly agrees, for high σ_τ, with the classical result that the mean of the distribution of the LRT statistic on the null (Fig. 7) should lie near $K/2$ where K is the number of extra parameters in the ‘alternative’ model—in our case the 40 time lags (Stuart *et al.*, 1999)). Note also that calculating Λ^C does not require prior knowledge of the true Lagrangian trajectory, thus

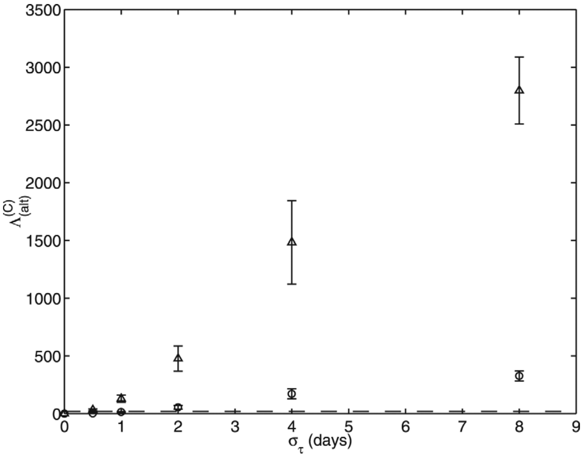

Figure 6. Calibration cost savings yielded by new technique $\Lambda^{(C)}_{(alt)}$ vs. time lag standard deviation σ_τ when time lags in calibrating data set are non-zero (the 'alternative' hypothesis $\underline{\tau}^t \neq \underline{0}$). Means (triangles/circles) and standard deviations (error bars) are shown for $s = 0.05/0.15$.

it is applicable to testing for the presence of time lags in real data sets. For our purposes, however, it is not as appropriate an evaluation statistic as $\Lambda^{(V)}$ (equation (16)), because it does not specifically test for accurate recovery of the noise-free Lagrangian trajectory (and in particular, the correct oscillation frequency), which is the goal of our twin tests. We shall however find further use for Λ^C when estimating σ_τ (see Section 3b*iii*).

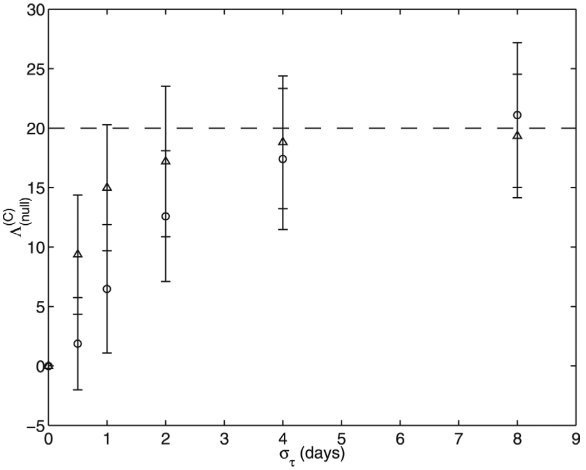

Figure 7. Calibration cost savings yielded by new technique $\Lambda^{(C)}_{(null)}$ vs. time lag standard deviation σ_τ when time lags in calibrating data set are zero (the null hypothesis $\underline{\tau}^t = \underline{0}$). Means (triangles/circles) and standard deviations (error bars) are shown for $s = 0.05/0.15$.

ii. Bias and variance in structural parameter estimates. Here we seek to compare VLF vs. ZLF structural parameter estimates in terms of their bias, defined as the expected error in a parameter estimate over all data sets, and their variance, defined as the mean square deviation of the estimate from its expected value. These quantities are estimated by performing 100 optimizations using $\sigma_\tau = 8$ days and $s = 0.15$ to generate and fit the data (see Table 2), initializing the optimization with $\underline{\theta}^i = \underline{\theta}^t$. This latter choice may lead to underestimation of optimal parameter variances, if the cooling is insufficiently gradual to prevent search localization near $\underline{\theta}^i$. Note, however, our main aim is not to assess the absolute performance of the model fits—rather the relative performance of the VLF vs. the ZLF, and we expect any search localization in $\underline{\theta}$-space to affect the VLF and ZLF equally.

First, note from Table 2 that the VLF has yielded smaller bias than the ZLF in almost all parameter estimates (except r_z, s_p and N_0, which are all low sensitivity parameters in determining the model trajectory), and smaller estimator variance in all cases. Therefore the VLF estimates of all the biological parameters to which the Lagrangian trajectory is sensitive (‘high’ and ‘medium’ sensitivity parameters) are both more accurate and robust than those of the ZLF, as one might have expected from Figure 4b. The improvement in parameter constraint is, we expect, a result of the fact that as sample size grows large, the VLF may fit to the true trajectory, which lies within the $(\underline{\tau}, \underline{\theta})$ search space, whilst the ZLF fits asymptotically to the convolution of the true trajectory with the time lag distribution,

Table 2. Bias and variance in structural parameter estimates using the (standard) zero lag fit (ZLF), variable lag fit (VLF), and variable lag fit with no lag drift penalty S (VLF(no S)), expressed as percentages of the true values. The data were generated with time lag standard deviation and fractional measurement noise levels of $\sigma_\tau = 8$ days and $s = 0.15$ respectively. Estimates highlighted in bold have biases more than three standard errors above or below zero.

		Estimator Bias (%)			Estimator Stdev (%)		
Parameter	Sensitivity	ZLF	VLF	VLF(no S)	ZLF	VLF	VLF(no S)
u_0	high	−0.8	0.0	**0.6**	4.8	1.4	0.6
R	high	−1.0	0.2	−0.1	6.4	1.1	0.4
Λ_z	high	**−18.8**	0.1	**−0.5**	12.5	1.5	0.6
m_z	high	**7.8**	−0.3	0.0	16.0	2.3	1.4
k_n	medium	**67.1**	0.2	**42.8**	62.0	16.7	29.8
c	medium	17.6	**−12.3**	**−34.4**	65.3	24.1	22.3
γ	medium	−6.5	1.3	**−16.4**	48.0	19.1	25.8
m_p	medium	11.1	0.4	**−24.9**	40.8	8.1	21.4
k	medium	**10.4**	−0.9	0.1	32.4	6.1	10.2
N_0	medium	**9.1**	**2.8**	**−1.9**	21.9	4.5	4.2
$P(0)$	medium	**54.7**	8.3	**56.5**	52.6	30.0	45.0
$Z(0)$	medium	−4.8	0.8	**8.5**	21.0	8.6	8.9
r_z	low	−4.0	14.8	**49.3**	48.2	42.7	46.3
s_p	low	−2.0	9.9	**25.9**	64.4	41.8	49.3
e_z	low	**6.2**	0.5	**1.9**	15.1	2.6	2.2
$N(0)$	low	4.2	−8.1	**−25.2**	46.7	36.5	35.9

Table 3. 'Derived parameters' (functions of the parameters in Table 1) and true values, where 'mean' denotes average over the sampling interval (roughly one cycle).

Derived Parameter	Symbol	'True' value
Oscillation frequency	ω	0.175 rad s^{-1}
Mean gross primary production	$\overline{GPP}$	0.0507 g C m^{-3} day^{-1}
Mean net nutrient export rate	$\overline{NNE}$	−0.0161 g C m^{-3} day^{-1}

which is a smaller signal relative to the measurement noise, and may not lie exactly in the $\underline{\theta}$ search space.

We can use Table 2 to infer the parameter biases that are likely most responsible for persistent distortion effects. Highlighted in bold are the significantly biased parameter estimates, defined by those with a bias more than three standard errors from zero. The ZLF produces several more significantly biased estimates than the VLF, including two among the high sensitivity parameters. This suggests that the persistent smoothing effect is achieved by a general slow-down in ecosystem conversion rates—decreasing grazer search efficiency Λ_z and increasing uptake saturation constant k_n—whilst mean concentrations over the sampling interval are roughly maintained by increasing diapycnal influx of nutrient (increasing k and N_0) and adjusting the initial conditions ($N(0)$, $P(0)$, $Z(0)$).

Tables 3 and 4 show true values and biases/variances for a few interesting 'derived parameters' i.e. functions of the structural parameters in Table 1, again for $\sigma_\tau = 8$ days, $s = 0.15$. First, we estimate the bias and variance in oscillation frequency ω by calculating the peak-to-trough separations in the true and optimal zooplankton trajectories (zooplankton showing the most symmetrical oscillation—see Figs. 2a, b) over the 100 optimizations. The bias in the VLF estimated frequency is very low (less than 1%) the standard deviation is significant ($\approx$15 %), the latter being responsible for significant increase in validation cost as the estimated trajectory drifts out of phase with the true Lagrangian trajectory (see e.g. Fig. 2a, lower). The estimated bias and variance in ZLF oscillation frequency is very poor, as we would expect since many of the ZLF trajectories are decaying to equilibrium (see Figs. 2a, b and Fig. 3).

Table 4. Bias and variance in 'derived parameter' estimates using the (standard) zero lag fit (ZLF), variable lag fit (VLF), and variable lag fit with no lag drift penalty S (VLF(no S)), expressed as percentages of the true values. The data were generated with time lag standard deviation and fractional measurement noise levels of $\sigma_\tau = 8$ days and $s = 0.15$ respectively. Estimates highlighted in bold have biases more than three standard errors above or below zero.

	Estimator Bias (%)			Estimator Stdev (%)		
Parameter	ZLF	VPF	VPF(no S)	ZLF	VPF	VPF(no S)
ω	24.2	−0.6	−6.7	166	14.7	32.8
$\overline{GPP}$	**8.5**	−1.3	−4.6	11.4	4.0	5.6
$\overline{NNE}$	**−3910**	**−4460**	**−5860**	−2100	−1740	−1860

Second, we consider estimates of mean (over sampling interval) gross primary production ($\overline{GPP}$), defined by the uptake term in equation (2). Here, the ZLF performs much worse (+9% bias, 11% standard deviation) than the VLF (bias -1%, standard deviation 4%). Note that underestimation of the extent of variability in biological variables (N, P, Z etc.) does not necessarily imply underestimation of fluxes between them, since errors in multiple fluxes may compensate each other. The net nutrient export rate $\overline{NNE}$, defined by the sum of sinking and mixing terms in equations (1) and (2), was poorly estimated by both methods—we suspect, because of its small absolute true value of -0.0161 g C m^{-3} day^{-1} (indicating a delicate balance), and because the parameters which determine it (k, N_0 and s) are not highly sensitive for the chosen true trajectory (see Table 2).

Finally, we briefly illustrate the importance of using the time lag drift penalty S (see Section 2f) by comparing the VLF parameter estimates with those obtained by maximizing $M^{(C)}_{\mathrm{VLF}}$ in (10) which has no lag drift penalty S ('VLF(no S)') for $\sigma_\tau = 8$ days, $s = 0.15$ (see Tables 2 and 4). For almost all parameters, the bias and variance is significantly increased by not using S, the main effect of which is to allow more variance in estimated oscillation frequency. In fact the variance almost entirely offsets any improvement in fit performance relative to the ZLF due to lack of smoothing and we get similar validation cost using the ZLF and the VLF(no S). The VLF(no S) also incurs about a factor of 10 larger bias, and roughly twice the standard deviation, in $\hat{\omega}$ relative to the VLF. The mean gross primary production estimate $\overline{GPP}$ is also impaired relative to estimates obtained using S.

iii. Methods of estimating time lag variance. Though we have not tested any methods for estimating time lag variance σ^2_τ in this study, here we suggest a method for doing so in practical applications. The idea is to exploit asymmetry in regard to under/over-estimation of σ_τ in calibration statistics. An example is the LRT statistic (equation 17). The results for $\Lambda^{(C)}_{(alt)}$ and $\Lambda^{(C)}_{(null)}$ are plotted in Figure 6 and Figure 7 respectively, showing sets for $s = 0.05$ and $s = 0.15$ together on both plots. The key observation is that increasingly overestimating σ_τ on the null, as in Figure 7, results in a small increase in $\Lambda^{(C)}$ (O(10) units) relative to the values of $\Lambda^{(C)}_{(alt)}$ on the alternative (O(10^3) units—see Fig. 6), for most realistic values of σ_τ. We expect this result to carry over to different models, sampling conditions etc. as long as there is enough temporal variability in the Lagrangian dynamics that the decrease in $p^{(2)}_{\mathrm{ZLF}}$ due to the smoothing effect of the ZLF, when σ_τ is underestimated, is large compared to the increase in $p^{(2)}_{\mathrm{VLF}}$ due to fitting measurement noise in the VLF when σ_τ is overestimated (roughly $N_s/2$ units—cf. Fig. 7). As such the LRT has a high probability of rightly rejecting the null and is therefore a 'powerful' statistical test in this context (Stuart *et al.*, 1999). When the σ^m_τ used in fitting is decreased below the true value σ^t_τ, $\Lambda^{(C)}$ should vary smoothly between a large value when $\sigma^m_\tau \approx \sigma^t_\tau$ (see Fig. 6) and zero when $\sigma^m_\tau = 0$. Therefore there should be a marked change in the gradient of Λ^C with respect to σ^m_τ at close to the true value ($\sigma^m_\tau \approx \sigma^t_\tau$), coinciding roughly with the onset of significant smoothing effect and the 'release' of the fit improvement potential of the time lags. Thus σ_τ might be roughly estimated from successive trial fits, allowing some practical realization of the VLF benefits discussed above.

Alternatively, if sufficient replicated data are available (multiple, independent samples at the same sampling time), as might be accumulated from multiple years of data (carefully chosen to minimize interannual variability), the nonlinear optimization methods detailed here may be extended to consistently estimate both the measurement error s and time lag variability σ_τ simultaneously (Seber and Wild, 2003). We can say nothing, however, about the relative biases and variances that might be incurred by this method, which could be a topic for future investigation.

iv. Robustness of results to different data sets/initial parameter guesses. In Figure 8 we illustrate the robustness of the results in Figure 2a (σ_τ = 8 days, s = 0.05) to different realizations of the data set and to different initial guesses $\underline{\theta}^i$ = 0.9 or 1.1 × $\underline{\theta}^t$. We show examples obtained from data sets/initial guesses for which errors in the optimal frequency

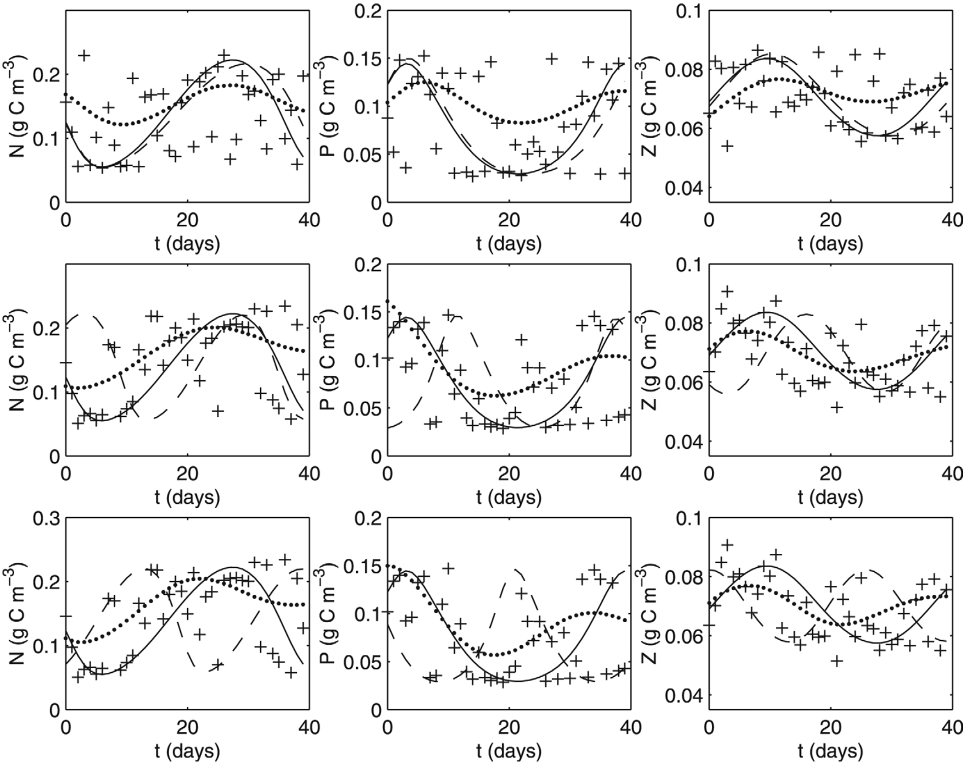

Figure 8. Variance associated with poor convergence in the variable lag fit for σ_τ = 8 days, s = 0.05. Upper panel reproduces Figure 2a lower. Middle and lower panels show examples of poor convergence in estimated frequency and phase of the oscillation respectively.

and phase of oscillation resulted in the worst fit performance by the VLF, using 10^5 iterations of the optimizer for convergence of the calibration cost. We see evidence that for high levels of assumed time lag variability ($\sigma_\tau = 8$ days) the VLF may incur high variance in both the estimated oscillation frequency $\hat{\omega}$ and overall oscillation phase despite use of the lag drift penalty S. This was mitigated to some extent by increasing the number of iterations, or using $\underline{\theta}^i = \underline{\theta}^t$, as with the results in Tables 2 and 3.

Diagnosing error due to poor convergence of frequency and phase estimates may be problematic in practical applications, but it is important as regards ensuring accurate hind-/forecasts and robust hypothesis tests. With the fits illustrated in Figure 8, for example, the final calibration cost $-\log M'^{(C)}_{\mathrm{VLF}}$ was not especially high, nor was there significant time lag drift measured by the S term. It may be possible to infer a frequency/phase deviation by comparison with the ZLF or a VLF with tighter timescale constraints (lower σ_τ), or perhaps subdividing the sampling interval to detect significant trends in the fitted time lags. Alternatively, convergence (and other) issues might be addressed by averaging estimated parameters over an ensemble of optimizations, using different initial guesses, or different data sets generated by replicated surveys, subsampling, or Bootstrap methods (e.g. see Schartau and Oschlies, 2003).

Given accurate estimation of ω, we found that $\underline{\tau}$ was well-estimated on the whole, except for some occasional 'slip' in the estimated time lags away from their true values (see Fig. 9 upper). We expect this result to depend on accurate specification of the measurement noise : time lag variability ratio (s/σ_τ) prior to fitting, in order to maintain consistency of $\hat{\underline{\tau}}$ as discussed in Section 2e. In Figure 9 we also see the effect on $\hat{\underline{\tau}}$ of VLF overestimation of ω. Here, the optimal time lags are accurate only over the second half of the sampling interval, when the fitted oscillation is roughly in phase with the true cycle (Fig. 8 middle). As we look further back in time the estimated and true time lags become increasingly divergent, resulting in overall loss of fit performance as discussed above.

v. Robustness of results to different Lagrangian 'truth'/models. The potential benefits of the VLF may be extended to a broad range of Lagrangian models by changing the assumptions of Section 2a (namely, the N, P, Z dynamics specified by (1–3)), whilst maintaining the assumption that the model formulation is a good approximation to the 'truth'. Consider varying the number m of simultaneously sampled state variables (ODEs) in the model. The VLF is more likely to be expedient for higher values ($m = 3$ gave substantial benefits in our example) because each time lag incurs errors in m state variables, and hence, if neglected, makes a larger contribution to the total cost for higher m. For the same reason, we would also expect the time lags to be better constrained for higher m.

Regarding the complexity of the model formulation, the VLF may only be expedient if the model produces nonlinear temporal variability. For linear models, because the errors in the random regressors are controlled ($\underline{t}$ is the same for all data sets) the ZLF will yield consistent estimates of all structural parameters (Laws, 1997), as long as the ratios of measurement error : time lag variability are assumed known (Seber and Wild, 2003). For higher complexity

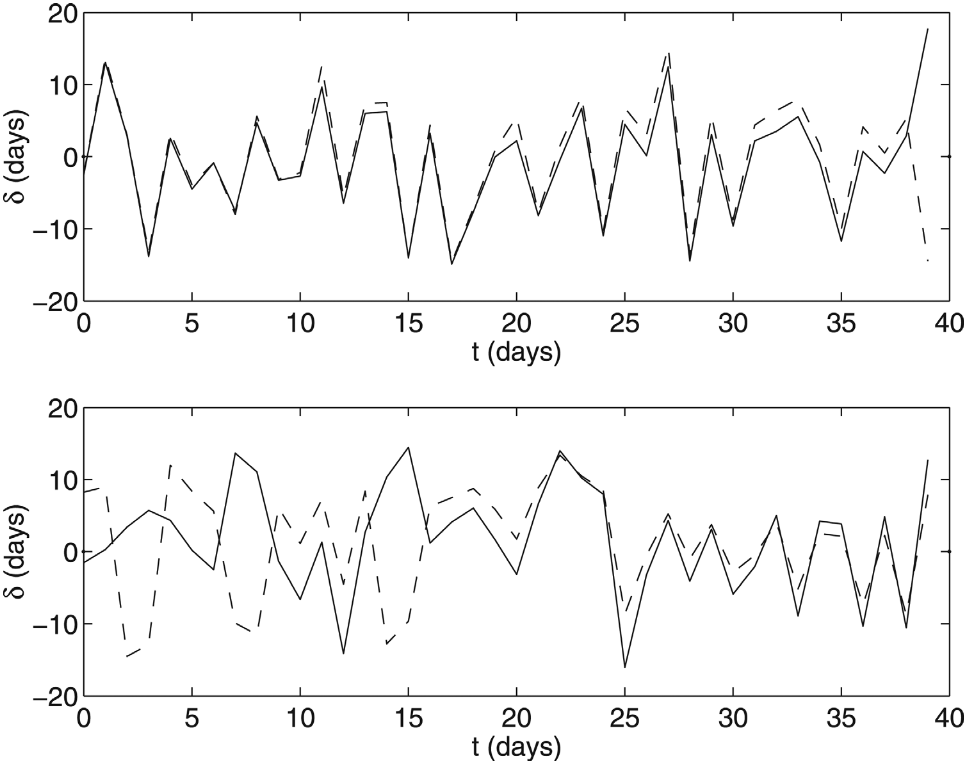

Figure 9. Best fit time lags (dashed) compared to true values (solid) for ($\sigma_\tau = 8$ days, $s = 0.05$), for optimizations with good convergence (upper) and poor convergence (lower) in estimated frequency and phase of oscillation (corresponding to fits shown in Figure 8 upper and middle panels).

models, with more free parameters, we expect VLF expediency to increase. This is because the higher complexity model trajectory may be distorted even further from the Lagrangian trajectory by fitting structural parameters to the measurement error inherited from neglected time lag variability. Consequently, we expect there will also be more total parameter bias and variance due to the fitting of structural parameters to random variables which change between data sets.

If we vary $\underline{\theta}^t$, maintaining significant temporal variability relative to measurement noise (as required to fit any dynamical model), we argue that the VLF will remain expedient as long as the errors due to neglected time lag variability in the true dynamics are statistically significant relative to measurement noise. It follows that values of $\underline{\theta}^t$ which generate rapid Lagrangian changes in biological variables increase the potential benefit of using a VLF.

Finally, consider increasing the discrepancy between the true dynamics ('truth') and the approximating model formulation used to fit the data, which was zero in our simulation study (a somewhat unrealistic scenario). This introduces the risk of using modeled time lags to compensate for deficiencies in the Lagrangian model. The VLF may substitute time lags for true rate variations in order to fit the data (the $p_{\text{VLF}}^{(2)}$ part), as long as no persistent rate change is imposed over the entire sampling interval (which would result in high lag drift penalty S). Model formulations which are insufficiently flexible in their rates of temporal variation may thereby escape penalty under the VLF, although one might hope to diagnose this by varying the sampling interval (subsampling the data set), so that persistent compensatory lag drifts may be detected by marked changes in S.

vi. Robustness of results to different spatial variability and sampling conditions. Consider relaxing the assumptions made in Sections 2a, b, c—namely: that the model of independent Gaussian time lags and measurement errors is accurate, and the sampling strategy of 40 samples at one-day intervals. In general, an estimation method fails not if its underlying assumptions are inaccurate, but if their inaccuracy spoils convergence on the true parameter values—in our case, if it fails to give the true model the minimum calibration cost (an important distinction). Suppose, for example, that there were significant correlation between successive time lags (e.g. due to mixing). Then our method would tend to over-penalize low-frequency and under-penalize high-frequency lag variations from the mean, because correlations would favor the occurrence of the former over the latter, whilst the independence assumption penalizes variations of all frequencies equally. Since only low frequency false time lag variations are likely to result in significant trajectory distortion, the assumption of zero correlation seems to be the safest in lieu of reliable *a priori* information. Similarly, one might argue that although correlated model error, non-Gaussian instrumental or time lag noise may in practice compromise our ability to estimate the true Likelihood of the data, they are yet unlikely to seriously impair the estimation method.

In realistic cruise sampling surveys, the interval between samples may be as short as a few minutes (e.g. see Fig. 1). Clearly using all these samples would violate the time lag independence assumption made in Section 2c. In case this does increase estimate biases, one might sub-sample the data using an interval of 1 day (or enough time for the largest coherent fluid structure to pass through the sampler), to generate a sub- data set for the VLF. Thus an ensemble of sub- data sets with different $\underline{t}$ vectors and corresponding model fits may be built up. This may also allow iterative estimation of σ_τ, in a similar fashion to 'replicated' data. On the other hand, if the between-sample interval is too long to track the temporal variability in the model (the ecosystem is undersampled, or the model is ill-suited) then all parameters will tend to be underconstrained, and probably more would be gained from a better sampling strategy (or new model, motivated by different questions) than from a new estimation method as described here.

In practical applications, one may of course expect a certain amount of model error (in addition to measurement error), due to unresolved spatial variability in the extent/pattern of variability (phase space trajectory, and hence $\underline{\theta}^t$); in such cases the optimal VLF trajectory should be interpreted as a 'mean Lagrangian trajectory'. Nevertheless, the total model error will be smaller for the VLF than for the ZLF if between-sample time lags do describe a significant portion of the unresolved spatial variability.

Another consideration is the effect of varying the number of sampling times N_s for fixed sampling interval T_s. As N_s is increased we expect the benefit yielded by the VLF to increase as the structural parameter estimates become better constrained, in spite of the increased risk of violating time lag independence (see above). Also, increasing the total survey interval T_s for constant sampling frequency would likely improve timescale constraints imposed by the data and the lag drift penalty S.

vii. Alternative methods. We do not claim that our method is the only sensible method by which one may attempt to allow for time lagging in a marine ecosystem data set. We do claim, however, that there is no obviously superior alternative at present. One might envisage a method based on decomposing the time series into (e.g. sinusoidal) component signals, filtering the unwanted components arising from phase modulation (time lagging), and then fitting an ecosystem model to the resulting 'de-modulated' time series. However, we expect that such methods will face difficulties owing to the fact that significant time lags may occur as frequently as between consecutive samples (as in this study), and the fact that the time series may not resolve many periods of the underlying signal (if indeed it is periodic, which our method does not require). Both of these factors, we suspect, may make it difficult to distinguish wanted and unwanted components. In any case, fitting to a filtered trajectory would sacrifice the ability to assess the importance or extent of time lagging in a real data set from a Bayesian or hypothesis-testing viewpoint.

4. Conclusions

We performed a simulation study to compare the effects of time lags in simultaneously sampled ecosystem variables, due to unresolved spatial variability, on two methods of fitting a simple marine ecosystem model to data. The first (standard) method did not account for time lags; the second was a new method which allows for time lags from an assumed statistical distribution. Our findings are:

1) Fitting a model using the standard time series approach leads to a 'smoothing out' or underestimation of ecosystem temporal variability when spatial variability in the form of random time lags is significant. This is because the practical, non-Lagrangian data set effectively samples from the true Lagrangian trajectory smoothed out by the distribution of random time lags. The smoothing first becomes apparent (as the level of time lag variability is increased) around periods of high temporal curvature, hence peaks and troughs are underestimated in magnitude. For large enough time lags, the estimated variability may be damped out entirely, such that periodic oscillations in the true dynamics are misfitted as exponential decays.

Though our twin tests used a free predator-prey oscillating system, we expect that similar smoothing out and associated parameter biases in standard model fits may occur in the more generic case of fitting models to seasonal data, where phytoplankton blooms and subsequent troughs due to grazing may be underestimated as a result. In general the most serious underestimation is liable to occur in the high frequency components of the temporal variability, therefore the standard fit biases on the seasonal trajectory, characterized by sudden blooms, may be larger than estimated for the (roughly sinusoidal) monthly oscillation used in this study. Consequently, gross and net primary production rates may also be significantly underestimated during blooms. Although we cannot predict the impact on estimates of annual averages of these rates, it is clear that the effects of neglected time lags on standard model fits may be serious, and warrant further investigation.

2) A new 'variable lag' model fitting technique performs significantly better in recovering the 'Lagrangian' biological dynamics, when spatial time lag variability is large enough relative to measurement noise. The new technique was estimated to perform better in more than 70% of cases (data sets) when the standard deviation in the time lag distribution was 1 day or more with 5% measurement noise imposed on all three state variables, and when the lag standard deviation was 4 days or more with 15% measurement noise, provided that the level of time lag variability was accurately estimated prior to fitting. Given only a realistic range of possible time lag variances, the potential losses/gains in recovery performance from using the new method were estimated as 2%/750% with 5% measurement noise and 60%/580% with 15% measurement noise.

Correspondingly, the biases and variances in most parameter estimates were significantly reduced using the new technique. The bias and variance of several 'derived estimates' (functions of the model parameters) such as time-averaged gross primary production were also reduced, although a certain amount of variance in oscillation frequency could not be avoided. We proposed a method of estimating the time lag standard deviation, assumed known in this study, by exploiting asymmetric effects of under/over-estimating the time lag variability on the relative model-data misfit in calibration. Given many years-worth of data with negligible interannual variability (a 'replicated' data set), we expect that the new method might be extended to optimize both time lag and measurement error variances, which otherwise may need to be fixed in optimization to help constrain the parameter estimates.

The expediency of the new method shown by our results should be robust to variations in the assumed biological model, as long as this remains a good approximation of the true Lagrangian dynamics. Accounting for time lags should be important in the general circumstances that the Lagrangian dynamics produce rapid, nonlinear changes in state variables, resulting in large changes relative to the measurement noise over the duration of typical time lags.

In summary, we argue that the optimal 'Lagrangian' parameter sets obtained by our new method may reflect a robust mean dynamics on a smaller scale of spatial averaging than standard model fits, without explicitly resolving those scales. This may help to improve estimation of the biological components (and hence the performance) of models with a wide range of resolutions, including the more complex, spatially-resolved marine ecosystem models.

Acknowledgments. The authors wish to thank Robin Hankin at NOC for useful advice on technical aspects of this work, and gratefully acknowledge sponsorship from NERC in the form of a PhD Studentship for PJW.

REFERENCES

Bainbridge, R. 1957. The size, shape and density of marine phytoplankton concentrations. Biol. Rev., *32*, 91–115.

Burnham, K. P. and D. R. Anderson. 1998. Model Selection and Inference: A Practical Information-Theoretic Approach, Springer, 320 pp.

Edwards, A. M. and J. Brindley. 1996. Oscillatory behavior in a three-component plankton population model. Dyn. Stability Systems, *11*, 347–370.

Fasham, M. J. R., H. W. Ducklow and S. M. McKelvie. 1990. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. J. Mar. Res., *48*, 591–639.

Fasham, M. J. R. and G. T. Evans. 1995. The use of optimization techniques to model marine ecosystem dynamics at the JGOFS station at 47N 20W. Phil. Trans. R. Soc. Lond. B, *348*, 203–209.

Fennel, K., M. Losch, J. Schroter and M. Wenzel. 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. J. Mar. Sys., *28*, 45–63.

Froda, S. and G. Colativa. 2005. Estimating predator-prey systems via ordinary differential equations with closed orbits. Aust. N.Z. J. Stat., *47*(2), 235–254.

Franks, P. J. S., J. S. Wroblewski and G. R. Flierl. 1986. Behavior of a simple plankton model with food-level acclimation by herbivores. Mar. Biol., *91*, 121–129.

Hemmings, J. C. P., M. A. Srokosz, P. Challenor and M. J. R. Fasham. 2004. Split-domain calibration of an ecosystem model using satellite ocean colour data. J. Mar. Sys., *50*(3–4), 141–179.

Hurtt, G. C. and R. Armstrong. 1996. A pelagic ecosystem model calibrated with BATS data. Deep Sea Res. II., *43*, No. 2–3., 653–683.

Hurtt, G. C. and R. Armstrong. 1999. A pelagic ecosystem model calibrated with BATS and OWSI data. Deep Sea Res. I., *46*, 27–61.

Ingber, L. 1993. Simulated Annealing: Practice versus theory. Mathematical and Computer Modeling, *18*, 11, 29–57.

Laws, E. 1997. Mathematical Methods for Oceanographers, John Wiley & Sons, Inc, 343 pp.

Martin, A. P. and K. J. Richards. 2002. Patchy productivity in the open ocean. Global Biogeochem. Cycles, *16*, 1025, 10.1029/2001GB001449.

Martin, A. P. 2003. Phytoplankton patchiness: the role of lateral stirring and mixing. Prog. Oceanogr., *57*, 125–174.

Matear, R. J. 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: A case study at Station P. J. Mar. Res. *53*, 571–607.

Neyman, J. and E. L. Scott. 1948. Consistent estimates based on partially consistent observations. Econometrica, *16*, 1–32.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. 1999. Numerical Recipes in Fortran 77, Second Edition, The Art of Scientific Computing, Cambridge University Press, 963 pp.

Prunet, P. and J. Minster. 1996. Assimilation of surface data in a one-dimensional physical-biogeochemical model of the surface ocean 1. Method and preliminary results. Global Biogeochem. Cycles, *10*, 111–138.

Ryabchenko, V. A., M. J. R. Fasham, B. A. Kagan and E. E. Popova. 1997. What causes short-term oscillations in ecosystem models of the ocean mixed layer? J. Mar. Sys., *13*, 33–50.

Schartau, M. and A. Oschlies. 2003. Simultaneous data-based optimization of a 1D-ecosystem model at three locations in the North Atlantic: Part I—Method and parameter estimates. J. Mar. Res., *61*, 761–793.

Seber, G. A. F. and C. J. Wild. 2003. Nonlinear Regression, John Wiley & Sons Inc., 792 pp.

Spitz, Y. H., J. R. Moisan, M. R. Abbott, J. G. Richman. 1998. Data assimilation and a pelagic ecosystem model: parameterization using time series observations. J. Mar. Sys., *16*, 51–68.

Srokosz, M. A., A. P. Martin and M. J. R. Fasham. 2003. On the role of biological dynamics in plankton patchiness at the mesoscale: An example from the eastern North Atlantic Ocean. J. Mar. Res., *61*, 517–537.

Steele, J. H. and E. W. Henderson. 1981. A simple plankton model. American Naturalist, *117*, 676–691.

Stuart, A., K. Ord and S. Arnold. 1999. Kendall's Advanced Theory of Statistics, Vol. 2A: Classical Inference and the Linear Model, Oxford University Press, NY.

Received: *1 November, 2005*; revised: *28 August, 2006.*