UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Ocean and Earth Sciences

Accounting for Unpredictable Spatial Variability in Plankton Ecosystem Models

by

Philip John Wallhead

Thesis for the degree of Doctor of Philosophy

March 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS

SCHOOL OF OCEAN & EARTH SCIENCES

Doctor of Philosophy

ACCOUNTING FOR UNPREDICTABLE SPATIAL VARIABILITY IN PLANKTON ECOSYSTEM MODELS

by Philip John Wallhead

Limitations on our ability to predict fine-scale spatial variability in plankton ecosystems can seriously compromise our ability to predict coarse-scale behaviour. Spatial variability which is deterministically unpredictable may distort parameter estimates when the ecosystem model is fitted to (or assimilates) ocean data, may compromise model validation, and may produce mean-field ecosystem behaviour discrepant with that predicted by the model. New statistical methods are investigated to mitigate these effects and thus improve understanding and prediction of coarse-scale behaviour e.g. in response to climate change. First, the standard model fitting technique is generalised to allow model-data 'phase errors' in the form of time lags, as has been observed to approximate mesoscale plankton variability in the open ocean. The resulting 'variable lag fit' is shown to enable 'Lagrangian' parameter recovery with artificial ecosystem data. A second approach employs spatiotemporal averaging, fitting a 'weak prior' box model to suitably-averaged data from Georges Bank (as an example), allowing liberal biological parameter adjustments to account for mean effects of unresolved variability. A novel skill assessment technique is used to show that the extrapolative skill of the box model fails to improve on a strictly empirical model. Third, plankton models where horizontal variability is resolved 'implicitly' are investigated as an alternative to coarse or higher explicit resolution. A simple simulation study suggests that the mean effects of fine-scale variability on coarse-scale plankton dynamics can be serious, and that 'spatial moment closure' and similar statistical modelling techniques may be profitably applied to account for them.

Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, Philip John Wallhead, declare that the thesis entitled Accounting for Unpredictable Spatial Variability in Plankton Ecosystem Models and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at the University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

  Wallhead, P.J., Martin, A.P., Srokosz M.A., 2006. Accounting for unresolved spatial variability in marine ecosystems using time lags. J. Mar. Res. 64, 881–914.

Signed:...

Date:...

Acknowledgements

I am grateful to the Natural Environment Research Council (NERC) for funding my PhD studentship at the National Oceanography Centre, Southampton (NOCS), and to the World University Network (WUN) for sponsoring my six-month fellowship at Scripps Institution of Oceanography (SIO). I also thank Woods Hole Oceanographic Institution (WHOI) for providing funding for me to attend the two stimulating Skill Assessment workshops at the University of North Carolina (UNC).

I am also grateful to Meric Srokosz for data from the cruise RRS Discovery D227 (Chapter 2), and to the following people for the Georges Bank data used in Chapter 3: D. Townsend for the 'Phytoplankton + Nutrient studies' database and photosynthesis-irradiance data, D. Mountain for the biovolume data, and R. Payne for the 'MET 60 minute averaged data...' files of surface irradiance measurements — all made available on the Global Ocean Ecosystem Dynamics (GLOBEC) website.

Otherwise, I am indebted to the following people:

- My main supervisor Adrian Martin, thank you for providing continual intellectual and personal support, for always making time to listen and debate my ideas and concerns, for always taking an interest, for helping me to manage this project, and for being a constant source of inspiration as well as diligent and exacting criticism. To my second supervisor, Meric Srokosz, thank you for inspiring my PhD studentship, for all your time, patience, encouragement, scientific and personal advice, and continual good humour. To my third supervisor, Mike Fasham, thank you for also inspiring my PhD, and indeed for practically inventing a field of science which has so thoroughly engaged my interest for these years gone by and I expect many more to come.

- My WUN supervisor Peter Franks at SIO, for guidance, inspiration and encouragement, for pushing me to formulate new scientific questions, and for giving me the opportunity to experience SIO for 6 months and attend the Skill Assessment workshops at UNC.

- Robin Hankin, for convincing me that statistics is hard and worth thinking about, and for providing valuable technical advice.

- Dan Lynch, Dennis McGillicuddy and Francisco Werner for inviting me to attend the Skill Assessment workshops at UNC, which have proved greatly influential on this work.

- Thomas (Zack) Powell at Berkeley, for introducing me to the subject of plankton modelling, and for a good few memorable quotes!

- All my friends at NOCS. In particular, thanks to my climbing partners Doug McNeall, Jo Hopkins and Chris Nicholle for getting me out of Southampton on weekends to places with fresh air, rocky outcrops, and (usually) sunny weather. Thanks to Kirsteen MacKenzie for being a fun and marvellous housemate, to Rebecca Bell for always brightening the office with her style and humour, and to all my other officemates Tim Adey, Lisa Weber and Nadie Suarez-Bosché for putting up with the intermittent fist-slamming and gentle pleas of persuasion to my computer. Thanks to Kirsty Edgar for keeping lunch (and dinner) socially alive, and to Chris Jeffreys and Mark Vardy for help with computer glitches. Thanks also to Emily Dolan and Sarah Bennett for helping me to settle in Southampton when I was a stranger to this town.

- Tracie Gooch for running her fantastic Latin dance club in Southampton.

- John Allen, for giving me the chance to experience ‘real oceanography’ on a cruise of the North Atlantic.

- Family members for their unflagging support and encouragement over the years: my mother Margaret Wallhead, my father Melvyn Wallhead and my brothers Nick and Chris Wallhead.

- Emma Guirey. Thank you for all your help. Thank you for being you.

Chapter 1

Introduction

1.1 Problem addressed in this thesis

This thesis addresses the following general problem in plankton ecosystem modelling:

- Fine-scale spatial variability in plankton ecosystems which we may not want to or be able to predict and/or reproduce can seriously compromise our ability to understand and predict coarse-scale ecosystem behaviour.

The overall objective of this work is to devise and test new statistical methods for dealing with this problem. In this Introduction, it is first explained what plankton ecosystems are and why they are important, what plankton ecosystem models are and why they are necessary for understanding and prediction, as well as the importance of model validation or 'skill assessment'. Second, current thinking is summarised regarding the nature and origins of spatial variability in plankton ecosystems over the full range of scales at which it is observed. Third, the predictive capacity of models is discussed, including limitations on how much spatial variability they can and should be designed to predict in a deterministic sense. Fourth, the effects of spatial variability which is 'unpredictable' in a deterministic sense are explained and existing approaches to mitigate them are reviewed. Fifth, and finally, the statistical methods investigated in this thesis are outlined, explaining why they are worth investigating as well as how they differ from and build upon previous methods.

1.2 Plankton ecosystems and models

In most general terms, 'plankton' are aquatic organisms with very limited ability to swim against currents (from the Greek for 'drifter'). The vast majority of plankton, by numbers and biomass, consists of microscopic protists, bacteria, animals and viruses. Plankton are often categorised as 'phytoplankton', which perform photosynthesis and make their own food ('autotrophy'), 'zooplankton', protists and animals which feed on other organisms ('heterotrophy'), or 'bacterioplankton' to describe aquatic bacteria. Referring to plankton in combination with their fluid and chemical environment, we speak of 'plankton ecosystems'. In seas and oceans, plankton ecosystems form part of larger marine ecosystems, including swimming organisms such as fish (the 'nekton') and organisms fixed to or residing in the sea/ocean floor (the 'benthos'). Marine plankton ecosystems play a key role in regulating the global climate by sequestering carbon from the atmosphere in the deep ocean (the biological pump), and by releasing aerosols which promote cloud-formation (4^{th} IPCC report, Chapter 7: Denman et al., 2007). They also support marine species of great economic and intrinsic value. The ability to accurately *predict* their behaviour is therefore essential for understanding, forecasting, and mitigating the effects of climate change and for sustainable management of natural resources and biodiversity.

In order to make such predictions, it is necessary to synthesise knowledge and assumptions about how plankton populations respond to changes in their physical environment, nutrient supply and predation pressure, all of which may be influenced by human activities such as waste disposal, fishing and carbon emissions. This synthesis is encapsulated in plankton ecosystem 'models' — systems of mathematical relationships designed to make accurate quantitative predictions. It is also important to estimate the likely magnitude of errors in the model predictions due to erroneous information used by the model (model parameter values, forcings, and initial/boundary conditions) and errors in model formulation, in lieu of any firm underlying theory (analogous to the Navier-Stokes equations in fluid dynamics) from which to formulate plankton ecosystem models.

Another goal is to *understand* plankton ecosystems: to be able to explain their behaviour in terms of true causal relationships. A model which encapsulates a good understanding (high mechanistic veracity) need not necessarily have good predictive ability, in the sense of providing accurate inter-/extrapolation from data (predictive skill), because the model may be highly sensitive to errors in required input information provided by the data (via model fitting or 'calibration'). For example, a model of a well-understood system such as the double pendulum may yet make poor predictions due to extreme sensitivity to initial conditions or 'chaos'. Similarly, predictive skill need not necessarily imply high mechanistic veracity, since a model may mimic the behaviour of a real system without accurately describing the underlying causal relationships. In general, however, mechanistic veracity is evidenced by predictive skill. Such evidence is stronger for large, nonlinear extrapolations from fitted data (beyond the decorrelation scale of the true signal), since in such cases predictive accuracy is much less likely to arise 'simply as an artefact of data fitting'. Predictive skill may be measured, given sufficient data, by careful model testing ('skill assessment' or 'validation').

Nevertheless, useful predictions in marine ecology and climate science are sometimes made with 'strictly empirical' models which do not aim for any mechanistic veracity, and describe statistical correlations rather than causes (e.g. Sarmiento et al., 2004). However, such models are usually a stop-gap for accurate mechanistic models: their predictive accuracy is likely to be limited to a narrow range of conditions similar to those in which calibration data were collected, and they do not aid the development of understanding. Mechanistic models have the potential to extrapolate well beyond the conditions of calibration, and promote understanding by allowing hypotheses to be tested about how ecosystems behave and why they do so. Therefore, the pursuit of predictive tools in marine ecology is assisted by the simultaneous pursuit of understanding via mechanistic modelling. Similarly, mechanistic understanding is evidenced to the extent that predictive skill is demonstrated. In lieu of any such validation, modelling studies exploring how different ecosystem processes might interact lose both credibility and relevance.

Plankton ecosystem models are thus essential tools for marine ecology. The two modelling goals of prediction and understanding are complementary and require good techniques for model calibration and validation (skill assessment). In this study, methods are investigated by which plankton models might be better formulated, calibrated and validated to account for a fundamental characteristic of plankton ecosystems: spatial variability which is deterministically unpredictable over the timescale of interest. The focus is on understanding and predicting ecosystem behaviour on seasonal to decadal timescales, and on spatial scales of perhaps 100km or more, although some of the findings will also pertain to the development of 'marine ecosystem weather forecasting' on smaller spatio-temporal scales. The methods are demonstrated mainly using simple plankton models 'mechanistic' at the level of 'NPZ' (Nutrient-Phytoplankton-Zooplankton), but more empirical models are also used where necessary.

1.3 Spatial Variability in Plankton Ecosystems

1.3.1 Large-scale spatial variability ($>$ 500km in the horizontal)

Perhaps the most obvious source of large-scale plankton variability is their biological response to astronomical forcing. The diel (day-night) cycle in solar irradiance drives a diel cycle in phytoplankton growth (primary production) and is a control on zooplankton 'diel vertical migration' — a behaviour which allows zooplankton to graze fertile surface waters at night and retreat to darker depths by day, thereby avoiding predation. The seasonal cycle in solar irradiance induces variability between north and south hemispheres, and latitudinal time lags in ecosystem responses such as the phytoplankton 'spring bloom'.

However, the form of the seasonal plankton cycle also varies on the large scale. Biogeographical patterns can be identified which evolve on relatively slow interannual and interdecadal timescales (associated with e.g. El Niño Southern Oscillations and climate change). Much of this pattern may be attributed to large-scale gradients in physical conditions, such as latitudinal or vertical variations in solar radiation. However, some important physical factors, such as the depth of mixing in the fertile upper ocean, may change over shorter scales across 'fronts' separating water masses with distinct properties (e.g. the ∼200km-wide fronts comprising the subtropical convergence zones, or the ∼10km-wide fronts between stratified deep ocean water and a well-mixed shallow sea). Consequently, the concepts of distinct oceanic 'biomes', and within these 'biogeochemical provinces' (Longhurst, 1998), are widely considered to be useful for describing large-scale variability.

The minimal set of factors by which to define ocean biomes is disputable (cf. Longhurst, 1998, Sarmiento et al., 2004, and Lévy, 2005), but current emphasis is on physical rather than biological factors, partly because the sparseness of current biological data sets and the difficulty of categorising plankton prohibit a useful small set of biological factors (Longhurst, 1998). Also, the large-scale variations in the seasonal plankton cycle are thought to be largely determined by variations in the seasonal physics controlling light and nutrient supply to phytoplankton, which over time have structured variations in plankton community composition via competition and natural selection. Light supply is mainly controlled by changes in surface irradiance, affected by latitude and ice cover, and the depth of upper ocean mixing, affected by surface heat transfer and wind stress. Nutrient supply is mainly controlled in the open ocean by wind-driven upwelling flows and turbulent mixing of surface and nutrient-rich deep water, and nearer land by lateral flows from riverine sources, tidally-induced mixing, and coastal upwelling due to wind and bottom topography.

Biomes may be subdivided into 'provinces' according to the ocean/hemisphere in question, smaller-scale upwelling and downwelling patterns, as well as other nutrient supply factors such as airborne dust deposition. Note that these biomes do not strictly coincide with the boundaries of large-scale gyre circulations. Horizontal fluid transport or 'advection', on the large scale, is generally too slow to stir the large-scale plankton patterns e.g. between the north and south sides of the northern subtropical gyres, which are generally considered to lie in different biomes.

Given these large-scale horizontal variations in light and nutrient availability, and possibly turbulence intensities, it is widely held that some aspects of the plankton community in the surface ocean can be predicted using the principles of competition and natural selection. By this paradigm, turbulent and nutrient-rich 'eutrophic' waters tend to favour larger phytoplankton species (diatoms) which are less efficient at absorbing nutrients but are more resilient and respire less relative to their body size when mixed to depth. These produce more intense blooms, shorter food chains, more 'mesozooplankton' (relatively big herbivores of size 0.2–20mm) and more export of biogenic material to the ocean depths. Calmer and nutrient-poor 'oligotrophic' waters tend to favour smaller phytoplankton which are less resilient to turbulent and dark water but more efficient at absorbing nutrients. These produce lesser blooms, less mesozooplankton and longer food chains, with more recycling and less export (Azam et al., 1983). Consequently, the definition of biomes is roughly supported by large-scale data on mesozooplankton abundance (Haury et al., 1978) as well as satellite chlorophyll data (Longhurst, 1998; Sarmiento et al., 2004). As with terrestrial communities, there is also a general latitudinal trend in plankton species diversity moving from the poles to the tropics, as a result of increased light energy and greater competition for nutrients nearer the equator.

Vertical large-scale plankton variability, on scales 100m–1000m, consists of a general decline in total biomass moving from the well-lit 'euphotic' surface ocean into the dark intermediate ocean, essentially due to the lack of light for photosynthetic primary production. The boundary between the two is an O(100m)-thick vertical front known as the permanent thermocline. Plankton do survive below this boundary in the mesopelagic or 'twilight' zone, where food webs of bacteria and zooplankton are largely fuelled by organic matter sinking from the surface ocean. Chemotrophic primary production at hydrothermal vents supports plankton communities in yet deeper waters on the ocean floor.

1.3.2 Meso-/submesoscale spatial variability (1–500km in the horizontal)

Over smaller scales, the magnitude of plankton variability is generally smaller (the spectral slope of plankton spatial variability is negative over all wavenumbers), as is the timescale over which structure evolves. At the oceanic mesoscale (10–500km in the horizontal, weeks to months in time), plankton spatial variability includes intricate evolving patterns of 'whorls' and 'tendrils' (see Fig. 1.1). These complex patterns are thought to be largely shaped by highly energetic mesoscale flows. Such flows provide the 'weather of the ocean', in the form of 'turbulent' eddy and front structures roughly on the scale of the first internal Rossby radius, which decreases with increasing latitude and decreasing stratification. Like their large-scale counterparts, mesoscale flows are mostly two-dimensional, but do induce small but ecologically important vertical velocities which upwell and downwell nutrients into and out of the euphotic zone, thereby affecting mesoscale plankton variability. In fact, the combination of 2D, time-dependent advection with 100km-scale variability in nutrient input can produce plankton spectral slopes consistent with observations over length scales 10–100km even in highly simplified plankton ecosystem models (Abraham, 1998).

Mesoscale eddies can induce upwelling due to upward deformation of the seasonal thermocline during the formation/intensification phase (Flierl and McGillicuddy, 2002), eddy-wind interactions (McGillicuddy et al., 2007) and perturbations to the circular flow (Martin and Richards, 2001). Nutrients upwelled by these flows may stimulate increased large-scale primary production, especially in the subtropical oligotrophic biome (Falkowski et al., 1991; McGillicuddy and Robinson, 1997; Oschlies and Garçon, 1998; McGillicuddy et al., 1998), provided there is a mechanism to restore the deep nutrient concentrations (Oschlies, 2002). Similarly, the horizontal motion of eddies across biome boundaries may induce large-scale fluxes of nutrients (Oschlies and Garçon, 1998; Garçon et al., 2002; Oschlies, 2002). As well as net abundance, plankton species composition may be altered inside eddies by the nutrient-rich local environment (Lochte and Pfannkuche, 1987; McGillicuddy et al., 2007). Eddy distribution and intensity is strongly inhomogeneous on the large-scale (Woods, 1988; Garçon et al., 2002), which may also affect large-scale variability.

Mesoscale fronts are also strongly associated with plankton patchiness. The largest of these flow structures occur in deep ocean waters where major ocean water masses converge/diverge and at the boundaries of large-scale jets such as the Gulf Stream, and may produce plankton patches of scale 100km in the horizontal and 100m in the vertical (e.g. Fernandez and Pingree, 1996). Such patches may result from accumulation due to plankton swimming or floating/sinking vertically against the slow cross-frontal flows in order to maintain a preferred depth for nutrient/light availability (Franks, 1992a,b; Fernandez and Pingree, 1996). Despite their name, most plankton can swim or move by altering their buoyancy at speeds of order 0.1mm s^{-1} (for phytoplankton, perhaps 1–10mm s^{-1} for zooplankton), which is often comparable with cross-frontal vertical velocity components (Franks, 1992a). Smaller patches are associated with coastal fronts. In these cases, it is commonly found that a plankton patch of horizontal scale 5–50km resides subsurface and sloping down along the front, but does not reflect the scale of the front itself (Franks, 1992b). This implies a decoupling of the biology from the physical front formation processes, perhaps suggesting that 'diffusion' (fine-scale mixing) is important (Franks, 1992b, and references therein; Franks and Chen, 1996), or perhaps wind forcing (Franks and Walstad, 1997).

Fronts may also become unstable, developing 'meanders' and large cross-frontal vertical velocity components of order up to 10mm s^{-1} (100m d^{-1}) which in turn upwell/downwell nutrients and biota (e.g. Spall and Richards, 2000, and references therein), with potentially serious consequences for net primary production (Mahadevan and Archer, 2000; Lévy et al., 2001). The resulting plankton variability ranges from O(100km) coherent eddies to thin submesoscale 'filaments' of a few kilometres in width. Plankton community structure is also thought to be strongly affected as plankton are displaced outside their ecological niches with respect to nutrient and light availability (Martin et al., 2001). Finally, mesoscale Rossby waves may also be important sources of mesoscale plankton variability (Uz et al., 2001;

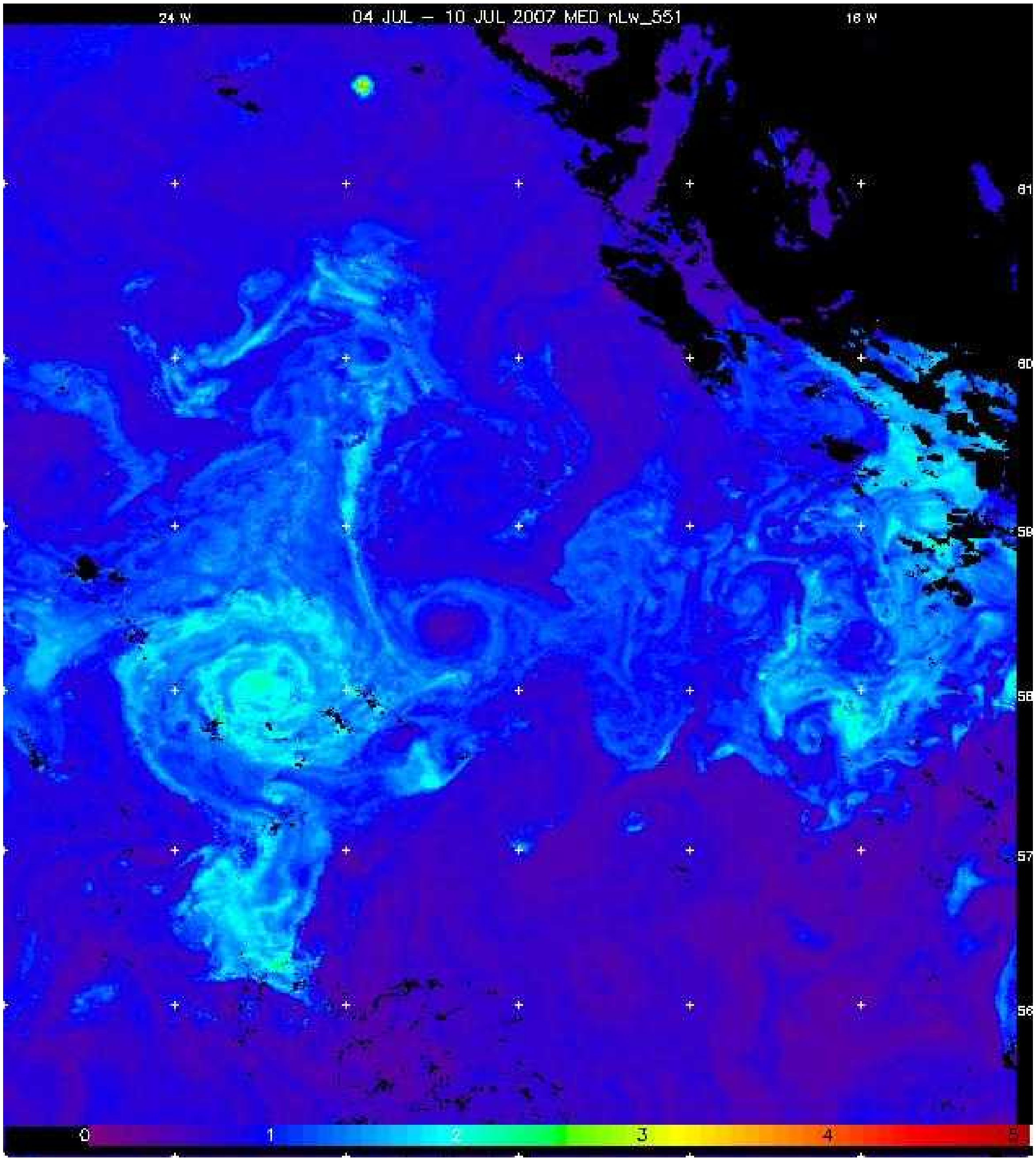

Figure 1.1: Normalised water leaving radiance at 551nm, a measure of coccolithophore (a type of phytoplankton) concentration, from a 7-day composite of MODIS (Moderate Resolution Imaging Spectroradiometer) satellite data for the North Atlantic in summer. Units are mW cm^{-2} um^{-1} sr^{-1}. The image was produced by NERC Earth Observation Data Acquisition and Analysis Service (NEODAAS) at Plymouth Marine Laboratory.

Cipollini et al., 2001; Dandonneau et al., 2003; Killworth et al., 2004, Charria et al., 2006), although the importance and dominant mechanism of their interactions with plankton are still very much in dispute.

1.3.3 Small-/microscale spatial variability ($<$ 1km in the horizontal)

Early theoretical investigations argued that horizontal plankton variability due to growth acting on initial heterogeneities of scale $<$ 1km should be smoothed out by the 'effective diffusion' of small-/microscale turbulent advection (Kierstead and Slobodkin, 1953; Okubo, 1978). Observations of small-/microscale plankton variability are relatively scarce, since satellites cannot resolve these scales, but significant small and microscale ($<$ 1m) variance has been observed by other methods (Platt 1972; Derenbach et al., 1979; Mitchell and Fuhrman, 1989; Davis et al., 1992; Franks and Jaffe, 2001; Doubell et al., 2006). This may be due to a host of factors neglected in early studies, such as nonlinear biological interactions (Petrovskii 1999a,b), predator swarming or mating aggregation behaviour (Folt and Burns, 1999), and physical effects of realistic small-/microscale ocean turbulence not captured by the effective diffusivity parameterisation used by models in these studies. These latter may include: buoyant interactions with vertical eddy structures such as Langmuir circulations (Bees, 1998), the intermittent stirring effects of convective overturns and shear instabilities (Cowles et al., 1998), or the action of inertial waves, shown to be capable of creating strong vertical gradients or 'thin layers' (Franks, 1995). Generally, these physical flow structures vary on timescales of a day or less. A more complete discussion of the mechanisms which may generate microscale variability is given in Yamazaki et al. (2002).

Early observations (Platt, 1972) also suggested that the spatial power spectrum of small-scale plankton variability follows a $k^{-5/3}$ shape, similar to that predicted for a passive tracer in the 'inertial subrange' of 3D, high Reynolds number turbulence (Kolmogorov, 1941). Numerous theoretical attempts have been made to explain such spectral slope observations at small and larger scales, necessarily neglecting some or all of the factors mentioned above (see Martin, 2003, and references therein). Early studies assumed linear biology and 'similarity' relations premissed on an inertial subrange separating coarse-scale forcing from fine-scale dissipation (Denman et al., 1977; Powell and Okubo, 1994). However, the neglect of intermediate-scale forcing may be invalid for the ocean (Martin, 2003), and in many cases the observed spectra may not actually resolve the inertial subrange even if it does exist (usually over 1mm–1m — Franks, 2005). Therefore, given the limited data and understanding of small-/microscale plankton behaviour and ocean turbulence, as well as the inherent ambiguities of spectral analysis (Armi and Flament, 1985; Martin, 2003; Franks, 2005), it is likely premature to interpret observations such as Platt's as evidence for a 'physics dominated' regime for plankton variability at small scales, as is often argued. Whatever the cause, the observed small-/microscale patchiness has serious implications for local plankton growth and grazing rates (Davis et al., 1992) and poses severe challenges for observing and modelling plankton ecosystems.

1.3.4 Summary remarks

- Plankton spatial variability occurs on a wide range of spatial and temporal scales. Partly, this is driven by spatial variability in physical variables, including ocean flows, which usually vary on an approximate timescale of 1 day per 1km of spatial structure (tidal flows being the notable exception). The other driver is the biologically encoded behaviour of plankton populations. On timescales shorter than those of climatic changes, population behaviour generally 'reacts' to physical variables rather than the reverse (a possibly important exception being biotically induced warming — Oschlies, 2004).

- The drivers of spatiotemporal variability in plankton ecosystems are often phase-related in the sense that similar changes occur in different places at different times, separated by phase or time lags (e.g. the seasonal irradiance/mixed layer depth cycle, propagating patterns in upwelling due to wind stress, or propagating mesoscale flow structures such as fronts and eddies).

- Different spatial scales of variability in plankton ecosystems interact. For example, mesoscale variability in vertical fluid transports is thought to upwell nutrients to the euphotic zone where they are then utilised by plankton (thus 'rectifying' the nutrient flux), resulting in large-scale nutrient fluxes and changes in plankton abundance. Such mesoscale variability in nutrients/plankton also drives finer-scale variability under the action of fluid stirring.

1.4 Predicting Spatial Variability in Plankton Ecosystems

1.4.1 To what extent *can* spatial variability be predicted?

The word 'prediction' can imply different levels of stringency in plankton ecology. At the 'softer' end, models can be constructed to 'predict' spatial features in a semi-quantitative manner. For example, we might ask: is a deep chlorophyll maximum predicted by the model? Does the model predict significantly enhanced concentrations within the eddy? Are different biomes, if not their spatial extent, predicted by the model? Or somewhat harder: does the model predict spectral slopes consistent with observations? Naturally, as a relatively new science, plankton modelling has first concerned itself with this level of prediction (Arhonditsis, 2004), and with some success. On the large scale, General Circulation Models coupled to plankton ecosystem models can reproduce many of the general patterns in satellite-based estimates of seasonal and annual-mean chlorophyll and primary production (e.g. Fasham et al., 1993; Sarmiento et al., 1993; Six and Maier-Reimer, 1996; Oschlies et al., 2000; Palmer and Totterdell, 2001; Rothstein et al., 2006; Popova et al., 2006). On the meso-/submesoscale, most of the qualitative features seen in data have been reproduced in high-resolution regional models such as in Spall and Richards (2000) or Franks and Walstad (1997), and even basin-ranging models such as that of Oschlies (2002) or McGillicuddy et al. (2003), simulating most of the Atlantic ocean at a resolution of roughly 0.1^o. Less effort has been deployed to reproduce small-/microscale features, partly because fewer data are available at this scale to allow model-data comparisons (current satellite resolution being insufficient).

In recent years, the predictive ambitions of plankton modellers have increased to build on earlier qualitative successes and to meet growing demands for quantitatively accurate predictions of 'harmful algal blooms' and ecosystem responses to climate change, pollution and aquaculture (Rothstein et al., 2006). For example, we might ask: does the model correctly *forecast* the time, location and severity of algal blooms? Unfortunately, the development and application of quantitative skill assessment or validation methods for plankton models does not appear to have kept pace (Arhonditsis and Brett, 2004). Where quantitative validation is attempted at all, modelling studies often merely quantify model-data discrepancy in somewhat arbitrary ways based largely on convention. A much preferable approach would be to first provide an explicit definition of 'skill' and then proceeding to estimate it. Definitions will inevitably vary, since different models may be designed to fulfil different predictive purposes (a point we return to in the next section). Of particular relevance to this thesis, the predictive aims of models may vary with respect to the spatial scales of variability which they are designed to predict.

At the large scale, almost all of the global ocean modelling studies mentioned above make visual comparisons between spatial model output and satellite observations from the Coastal Zone Colour Scanner (CZCS) and/or the Sea Wide Field-of-view Sensor (SeaWiFS). None, however, provides a quantitative metric to estimate how accurate the predicted large-scale variability is on average, or which which would allow us to say which model is most skillful at predicting present regional patterns as a whole.

At the coastal meso-/submesoscale, Di Lorenzo (2007) investigated the forecast skill, defined by a predictive average squared error, for physical variables of the Regional Ocean Modeling System (ROMS) model fitted to artificial data using the '4DVAR' assimilation method. Given a realistic set of data, as might be obtained from an oceanographic cruise taking CTD casts, they found that forecast skill was better than climatology for roughly 20 days beyond the end of the cruise for alongshore velocity, and for roughly 15 days for temperature. Ecosystem variables were not investigated, but given the more noisy sampling and larger model errors expected with plankton dynamics, due to stronger nonlinearity and less accurate model formulations, a two week timescale of predictability is likely an upper bound.

Similar mesoscale twin experiments have been performed on open-ocean physical models employing a widely-used simplification of the 'optimal interpolation' method (Dombrowsky and De Mey, 1992). Their conclusion was that 'model predictability limits the reliability of the forecast beyond 20 days' — although their experiments seem to suggest a forecast decorrelation time of 50–100 days beyond the last assimilation. The same method, used in the Harvard Ocean Prediction System (HOPS), has also been tested for assimilating real data into an open-ocean model (Robinson, 1996), with the conclusion that 'predictability of the region is about 30 days'. Again, since these are for predictions of physical variabilities with better-known dynamics, we should take 30 days as an upper bound for mesoscale plankton predictability in the open ocean. Results from using HOPS to predict mesoscale plankton variability in the open ocean (Popova et al., 2002a,b) do not seem to support a predictability timescale as long as this, although forecast accuracy is not quantitatively estimated or compared with climatology in these studies.

There do not appear to be any predictability studies for small-/microscale variability in physical or biological variables. However, the meso-/submesoscale studies suggest that predictability is limited roughly by the timescale of variability in the ocean flows at that scale. Hence a timescale of a few days or less is to be expected for small-/microscale predictability.

1.4.2 To what extent *should* spatial variability be predicted?

Fine-scale spatial variability in plankton ecosystems can have significant coarse-scale effects (e.g. Mahadevan and Archer, 2000; Lévy et al., 2001; McGillicuddy et al., 2003). Therefore, one might argue that, for making predictions, as much spatial resolution as is computationally feasible should be included in plankton models. Although we may not be able to exactly predict/forecast the fast, fine-scale behaviour over a useful timeframe (as discussed above), the *statistical* effect of fine-scale variability on the slow, coarse-scale behaviour may yet be accurately represented in a model.

This predictive strategy raises two issues. First, the computational feasibility of a particular spatial resolution depends on how the model is used. Model-fitting or 'data assimilation' has been shown to generally improve the accuracy of plankton model predictions (Friedrichs, 2001; Friedrichs et al., 2006, 2007) (Appendices 1A and 1B provide technical summaries of model fitting and testing methods in the context of plankton modelling). If the model is used to estimate poorly known biological parameters by fitting to data, limited computational resources might be better committed to performing more runs of a faster, lower-resolution or lower-dimensional model with different parameter values, allowing a more rigorous parameter fit. The fitted parameters may subsequently be employed in a single run of a higher-resolution or higher-dimensional model to make predictions. For example, Schartau and Oschlies (2003) used a 1D ecosystem model to estimate biological parameters which were later used in the 3D basin-scale model of Oschlies and Schartau (2005). Necessarily, horizontal fluxes could not be accounted for in the 1D model, but it was fast enough to be run $O(10^4)$ times, therefore allowing parameter optimisation via a powerful 'micro-genetic' algorithm. A faster run time also allows more rigorous skill assessment by exploring the robustness of parameter estimates/predictions to errors in model 'inputs' (forcings and initial/boundary conditions) and fitted data, perhaps using bootstrap methods (see Schartau and Oschlies, 2003 and Appendix 1B). A danger of this method, however, is that biological parameters may be distorted to account for the lower spatial resolution/dimension of the fitted model, and consequently yield poor predictions in the higher spatial resolution/dimension predictive model (c.f. Oschlies and Schartau, 2005).

Second, one might ask: how accurately will the statistical effects of finer scales be captured by the higher-resolution model? Suppose we have a coarse and a fine resolution model for making coarse-scale predictions, and suppose computational resources are unlimited. The same biological model is used in both models. As argued above, the exact pattern of fine-scale plankton variability cannot be predicted over long timescales (in the same way atmospheric weather forecasts are limited to short period — generally a few days — beyond the last data fit or 'analysis'). Moreover, forecasting this fine-scale variability is not, by assumption, within the remit of either model. Therefore, an optimal strategy to fit biological parameters might be to fit both models to data averaged over the coarse scale (see section 1.5.4). Which model will give better coarse-scale predictions? If the high-resolution model formulation and inputs are good enough, the extra resolution will accurately capture the statistical effect of fine-scale variability on the large-scale dynamics, and predictions will be less biased (here, 'bias' is a mean discrepancy with truth at given times/locations if the whole experiment were repeated many times over an 'ensemble'). If, however, the model produces inaccurate statistical behaviour at the fine scale, because the physical model misrepresents the fine scale, or the semi-empirical parameterisations of biological processes are less accurate at the fine scale, the prediction biases could in fact be worse than those of the coarse-resolution model. Alternatively, the additional information used by the high-resolution model, in the form of high-resolution forcings and initial/boundary conditions, may be sufficiently noisy to increase bias or, which is perhaps more likely, to increase the 'variance' of predictive errors (variation at a given time/location over an ensemble of repeat experiments). Typically, we are interested in predictions which are 'accurate' in the sense of low 'mean-square error', a sum of (bias)^2 and variance. Hence there can be a 'bias-variance trade-off' between the more-complex, higher resolution model (lower bias, higher variance) and the simpler low-resolution model (higher bias, lower variance). A simple illustration of the bias-variance trade-off is given in Figure 1.2.

The over-arching factor in deciding model spatial resolution will be the particular scientific purpose or remit that the model is designed to fulfil. Figure 1.2 demonstrates how the best choice of model resolution for a specific predictive purpose defining 'skill' can be an intermediate-complexity model, when the same model is used for parameter fitting and prediction. Of course, if the purpose of the model is to test hypotheses regarding the statistical effects of fine-scale variability, this variability must be explicitly resolved in the model. Similarly, in some applications, the exact spatiotemporal evolution of fine-scale variability might be crucial. A high-resolution coastal ocean model of a harmful algal bloom which predicts algal growth just 1km further up the coast could (in theory) influence a management decision to close a beach to public bathing. Alternatively, a model which is envisaged to guide fishing vessels to plankton-rich and therefore nekton-rich frontal areas needs to be deterministically precise about when and where fine-scale plankton patches will occur. However, in other applications e.g. to understanding and predicting the larger-scale effects of El Niño Southern Oscillations or climate change on plankton ecosystems, we do not particularly care about the fact that our models cannot predict the exact configuration of eddies and other mesoscale features on annual or even seasonal timescales. We *do* care, however, about the effects this deterministically-unpredictable spatial variability may have on biological parameter estimates obtained from fitting plankton models, and about its statistical effects on larger-scale variability. These problems, and current solutions, are further discussed in the next section.

1.4.3 Summary remarks

- The ability of current plankton models to predict spatial variability on different scales is understudied. However, it seems that the timescale of deterministic predictability is roughly bounded by the timescale of change in ocean flows at that scale (about 1 day per 1km of scale). This correlation may be weakened by variation with scale of the degree of system sensitivity to forcings and initial/boundary conditions (including chaoticity), or system dimensionality (number of non-negligible influences on the dynamics).

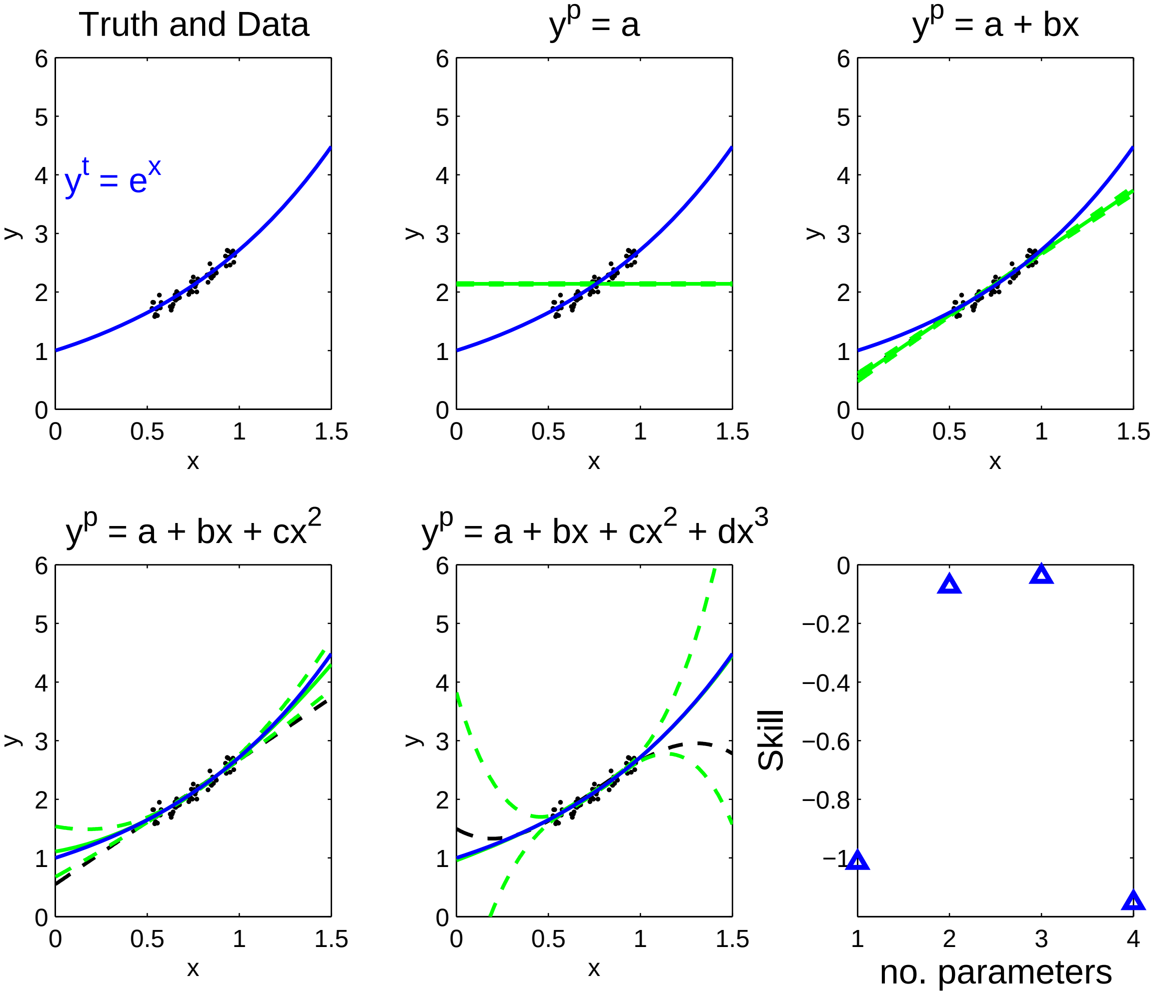

Figure 1.2: Bias-variance trade-off in model complexity. Data sets are created by sampling a 'true' variation $y^t(x) = e^x$ non-uniformly in the range $0.5 < x < 1.0$ with 10% Gaussian error superposed. A series of increasingly complex polynomial models $y^p(x)$ (containing more and more terms in $y^t(x)$) are fitted by Ordinary Least Squares linear regression, and then used to predict values in the range $0 < x < 1.5$. Predictive 'skill' is defined by average predictive mean-square error: $Skill = -\sum_i E[(y^t(x_i) - y^p(x_i))^2]$ where the expectation E denotes an average over an ensemble of 5000 'repeat experiments' with different data sets but the same truth. The quadratic model is an optimal compromise between high prediction bias (difference between solid green ensemble-mean prediction and solid blue truth) for low model complexity and high prediction variance (dashed green shows ±1 ensemble standard deviation in the prediction) for high model complexity

- The amount of spatial variability that should be resolved in a plankton model depends strongly on the purpose that the model is designed to fulfil. Models designed to predict variability at a specified scale of averaging may need to explicitly resolve finer scales in order to reduce bias in predictions resulting from the statistical effects of fine-scale variability on coarse-scale averages (unless this can be achieved by sub-grid scale parameterisation). However, errors in fitted data and model inputs (forcings and initial/boundary conditions) at the fine scale of averaging may increase the variance of predictions, which in turn may mitigate improvements in predictive skill gained from bias reduction (bias-variance trade-off). A trade-off in computational feasibility should also be considered: faster, lower resolution models may be run more times (larger ensemble), allowing more rigorous calibration (which may improve predictive skill) and skill assessment/validation (including better uncertainty estimates).

1.5 Effects of unpredictable spatial variability and existing methods to mitigate them

1.5.1 The need for model calibration/validation with ocean data

Plankton models generally contain a large number of poorly-known biological parameters. This is partly simply because they represent processes which are poorly understood and poorly measured in general. Also, the parameters are associated with ‘average’ rates of growth and grazing in a patchy and time-varying ocean environment and community of species which is almost impossible to reproduce in a laboratory experiment. Furthermore, they are applied to model grid cells which are usually several orders of magnitude larger than laboratory containers and 10m-scale mesocosm enclosures (e.g. Vallino, 2000). Hence laboratory parameter estimates may require adjustment to describe the concentration dynamics of a much larger scale of averaging (although there may be preferable solutions, as we will see). Consequently, the standard approach is to try to estimate these parameters, or some subset of them, by fitting ecosystem models to ocean data (assuming a sufficiently fast model is available for this purpose). This process of ‘calibration’ will therefore be assumed to be an important if not vital initial stage of a plankton modelling study.

However, it is worth noting that alternatives have been and are being sought, in view of the problems of inadequate model constraint by ocean data and the semi-empirical (and therefore less extrapolatable) nature of the standard approach. For example, some investigators favour the development of more mechanistic plankton models, including more complexity/detail but basing it on more accurate plankton cell physics/chemistry/biology which is better constrained by physical laws and laboratory experiments (e.g. Baird and Emsley, 1999; Flynn, 2001, 2005). By this approach, if successful, little or no ‘tuning’ to ocean data would be required. It is questionable, however, that there will ever be a sufficient number of realistic laboratory experiments to constrain parameters for all plankton species that significantly affect marine ecosystems. Furthermore, even if such an approach could deal with the ‘aggregation’ or ‘integration over species’ problem (a view challenged in Anderson, 2005, 2006), the ‘integration over space’ problem, due to all the fluid flow structures and patchiness contained within each model grid cell, remains. Consequently, the benefits of highly mechanistic multi-species models might only be realised with very high model spatial resolutions (e.g. Baird and Emsley, 1999). This would compound the problems of computational speed and robustness to errors in model dynamics/inputs associated with higher model complexity, even if it is not fitted to ocean data. Nevertheless, it is a promising line of enquiry which, perhaps in moderation, might allow better use to be made of the information gathered by laboratory plankton experimentalists in more mechanistic plankton models.

Another alternative worthy of note is the idea of using principles of competition and natural selection to constrain the relative abundances of species at different locations in the ocean (Follows et al., 2007). By this approach, ecological 'niches' and their associated abundances are naturally selected from an initial state uniform in space and random in biological parameter space, and the process is repeated over an ensemble to identify robust patterns. This is an innovative new approach, but it is not clear yet whether it can be made computationally practical for general usage, or how close the similarity really is between such simulations and the real, dynamic, ocean biogeography, which represents a single evolution obtaining variety from genetic variation among offspring rather than a random initial state. Initial comparisons with data are nevertheless quite promising (Follows et al., 2007).

Finally, note that even if these alternative approaches become preferable, we will still need to validate model predictions with ocean data, in order to show that they are successful and to measure predictive skill. Hence the effects of deterministically-unpredictable spatial variability on model-data discrepancies will still need to be considered in model testing or 'skill assessment' (see Appendix 1B).

1.5.2 Effects of unpredictable model inputs and undersampling

Consider the fine-scale concentration of a single plankton ecosystem variable $y(x, l, t)$ at a point x, averaged over a fine sampling scale l at a time t (the discussion easily generalises to multiple variables). Typically, l might represent the 1–10 centimetres scale of a fluid mass collected in a Niskin bottle or sampled by a continuous plankton recorder in a single measurement. Larger-scale averages are available from satellite measurements (with pixel sizes of order 1km), but it is not generally thought that this surface chlorophyll data is on its own sufficient to constrain a plankton ecosystem model. Consequently, the smallest scale l associated with plankton data is generally the order of centimetres. Observations, $o(x, l, t)$, are made of the true sample-scale concentration $y^t(x, l, t)$ with *observational error* due to instrumental noise (superscript 't' denotes 'truth'). The observational error in the i^{th} fine-scale observation may be defined as:

$$\epsilon_i^o \equiv o(x_i, l, t_i) - y^t(x_i, l, t_i) \tag{1.1}$$

Typically, these observations need to be somehow used to maximise the 'skill' of model predictions for certain ecosystem variables over certain spatio-temporal ranges over certain scales of averaging. Alternatively, only estimates of unknown model parameters may be required, in order to test some ecological hypothesis, or to provide inputs for future predictive models. Since a model is only optimal for a specified purpose (see Figure 1.2), it is important and helpful to specify the objectives of the modelling study as precisely as possible at the outset (perhaps using a 'skill function' — see Chapter 3). A general way to achieve these goals is model fitting or 'data assimilation'.

In practice, the fitted model is unlikely to resolve $y^t(x,l,t)$: the computational demands of using model grid cells of scale l to cover a region traced by fluid masses over the timescales of plankton growth is almost certainly prohibitive. Rather, the fitted model attempts to reproduce the true coarse-scale concentration $Y^t(X,L,t) \equiv \overline{y^t(x,l,t)}$ where the bar denotes spatial averaging over a larger model grid cell of scale L centred at X at time t. We may then define the true *unresolved spatial deviation* δy^t as:

$$\delta y^t(x,l,X,L,t) \equiv y^t(x,l,t) - Y^t(X,L,t) \tag{1.2}$$

Hence $\overline{\delta y^t(x,l,X,L,t)} \equiv 0$. In order to fit and test the model, we might compare model output to observational estimates of the true grid-scale concentrations $Y^t(X,L,t)$. The coarse-scale observation O for model grid cell X_i at time t_j may be written generally as a function of the full set of fine-scale observations: $O(X_i,L,t_j) \equiv h_{ij}(o)$. In the simplest approach, the 'function' h could be a simple synoptic (same time) average over observations lying within the grid cell X_i at time t_j (or $O(X_i,L,t_j) = o(x_i,l,t_j)$ if there is only one such observation). In practice, better estimates may be obtained by averaging over asynoptic samples, or over 'objective analysis' fields generated by inter-/extrapolating the sparse biological data over space/time, using statistical ('kriging') and deterministic (trend surfaces, 'feature models') empirical models (Thiébaux and Pedder, 1987; Anderson et al., 2000). The coarse-scale observational error ϵ^O_{ij} may then be defined as:

$$\epsilon^O_{ij} \equiv O(X_i,L,t_j) - Y^t(X_i,L,t_j) \tag{1.3}$$

Now, the coarse-scale plankton model output Y^m can be written generally as a function $Y^m(a,\hat{\kappa},X,t)$ of some 'adjustable' or 'control' parameter set a, a set of fixed model inputs $\hat{\kappa}$ (which may include include e.g. fixed biological parameters and/or fixed forcings such as velocity fields from a physical model), spatial coordinates of the grid cell centre X, and time t. The hat notation is used in this thesis to denote estimates subject to significant error (such as $\hat{\kappa}$). There are many alternative ways to fit a, but they can usually be categorised, in Bayesian language, as either 'posterior mode' estimation (or 'variational assimilation'), or 'posterior mean' estimation, often employed in 'sequential assimilation' methods (see Appendix 1A). Perhaps the simplest estimation method, of the posterior mode type, is Ordinary Least Squares. Here, we adjust a such that it minimises *cost* $\equiv \sum_{ij}(\epsilon^r_{ij})^2$, where the misfit or 'residual' ϵ^r_{ij} at model coordinates (X_i,t_j) is given by:

$$\begin{aligned}
\epsilon^r_{ij} &\equiv Y^m(a,\hat{\kappa},X_i,t_j) - O(X_i,L,t_j) \\
&= Y^m(a,\hat{\kappa},X_i,t_j) - Y^t(X_i,L,t_j) - \epsilon^O_{ij} \\
&= \epsilon^m_{ij} - \epsilon^O_{ij} \qquad (1.4)
\end{aligned}$$

where, in the last line, we have substituted the 'fitted model prediction error':

$$\epsilon^m_{ij} \equiv Y^m(a,\hat{\kappa},X_i,t_j) - Y^t(X_i,L,t_j) \tag{1.5}$$

Note, this is not to be confused with the 'ultimate predictive error', which may involve a different model to the fitted model, or 'model error' which is conventionally used to described stochastic corrections to model dynamics.

From (1.4) and (1.5), it is clear that unpredictable model inputs (κ) will make spurious contributions to the ϵ_{ij}^{m} which will compromise the fit of a. From (1.3) and (1.4), it is also clear that undersampling will contribute to the ϵ_{ij}^{O} and further compromise the fit. There is no easy solution to these problems. The simple estimation method *might* be improved by added complexity in the form of extra statistical parameters to account for correlations in the ϵ_{ij}^{O}, or extra adjustable parameters in a to account for errors in the model inputs $\hat{\kappa}$, or by integrating estimates over a modelled probability distribution conditional on the data (posterior mean approach). However, such attempts to improve 'realism' do not necessarily improve estimation, essentially because the extra parameters invariably need to be estimated directly or indirectly from data (cf. Fig. 1.2). Further complexity, whether in the dynamical model or the estimation method, should always be added with care.

One strategy to reduce the mean size and hence effects of the ϵ_{ij}^{O} is to increase the spatial resolution of the fitted ecosystem model (smaller L). As we decrease the grid cell size L, the total spatial variance of the unresolved variability decreases. Therefore, *assuming that the number of samples per grid cell does not change*, the ϵ_{ij}^{O} should get smaller. However, as discussed above, the predictability of grid cell averages also decreases with grid cell size, hence the predictive errors ϵ_{ij}^{m} in smaller grid cells are liable to increase more rapidly with time due to errors in the ecosystem model inputs $\hat{\kappa}$, or stochastic errors in the physical or biological model dynamics, which are likely to be larger at smaller averaging scales. Similarly, if we attempt to validate the high-resolution model with fine-scale data, we may get an excessively pessimistic skill assessment using misfits such as (1.4) because the model fails to reproduce exact fine-scale patterns without phase error (see Appendix 1B), even if it does a good job of reproducing them in a statistical sense (including coarse-scale patterns and spatial power spectra).

Alternatively, by sufficiently decreasing the resolution of the fitted model (larger L), we may improve the predictability of model inputs $\hat{\kappa}$, and perhaps also reduce the ϵ_{ij}^{O} if more 'raw data' are employed to constrain the coarse-scale observational estimates. However, $\hat{a}$ may then be distorted to account for the statistical effect of unresolved variability on the dynamics of the true coarse-scale averages $Y^{t}(X, L, t)$, as discussed in the next section. This may lead to inaccurate predictions when used in higher resolution models, and also in coarse resolution models if these are required to predict beyond the range of conditions in which they were calibrated (extrapolation failure due to excessive empiricism or 'data-fitting').

1.5.3 Effects of inadequate resolution

Subject to various assumptions, discussed in Chapter 4, plankton ecosystem dynamics may be written as a system of advection-reaction equations:

$$\frac{\partial y_i}{\partial t} = -\nabla.(v y_i) + F_i(y) \tag{1.6}$$

where y_i is the concentration of the i^{th} ecosystem component, averaged over a suitably fine scale, v is the fine-scale fluid velocity vector, and $F_i(y)$ describes the fine-scale biological dynamics (nutrient uptake, growth, grazing etc.). For simplicity, plankton motility is neglected in (1.6). Note that since v is the fine-scale velocity, and motility is neglected, a diffusion term is not included (Abraham, 1998). From (1.6), the dynamics of grid cell averages $Y(X, L, t) = \overline{y(x, l, t)}$ are given by the Reynolds equations (Reynolds, 1895; Lévy, 2006):

$$\frac{\partial Y_i}{\partial t} = -\nabla.(VY_i) + F_i(Y) - \overline{\nabla.(\delta v \delta y_i)} + \overline{\delta F} \tag{1.7}$$

where $V \equiv \overline{v}$, $\delta v \equiv v - V$, $\delta y \equiv y - Y$ and $\delta F(y) \equiv F(y) - F(Y)$. Hence the grid cell dynamics differ from the fine-scale dynamics by two terms: the third term on the RHS of (1.7) is the 'eddy-flux' or 'physical Reynolds term' due to nonlinear transport coupling and sub-grid scale correlations between v and F (e.g. due to mesoscale eddies upwelling nutrients); the fourth term is the 'biological Reynolds term' and arises due to nonlinear biological dynamics and sub-grid scale correlations between the components of F. The combination of these terms is the statistical effect of unpredictable spatial variability on coarse-scale dynamics. Clearly, if this effect is significant and inadequately resolved by the model, the parameter estimates $\hat{a}$ may be distorted to account for it, which again may lead to loss of predictive accuracy. Moreover, even if we do not fit the model to ocean data, and use accurate values of fine-scale biological parameters in F (obtained e.g. from systematic laboratory studies), a model which does not adequately resolve these Reynolds terms will produce inaccurate predictions.

1.5.4 Existing methods

A good strategy to minimise ecosystem model forcing error, and hence its distortion effect on $\hat{a}$, is to fit the physical model to physical data. In order avoid spurious physical model adjustments to account for biological data, and vice versa, and to maintain a computationally tractable model fit, it may be expedient to fit the physics first with fixed biology (e.g. Schartau and Oschlies, 2003). However, this entails the neglect of potentially important biological adjustments on the physics, such as alterations to photon fluxes (Oschlies, 2004; Miller et al., 2006). It also neglects potentially useful information in the biological data for constraining the physics, and may lead to inconsistencies when the biology is subsequently adjusted (Anderson et al., 2000).

Even if the physical model fit is successful in minimising ecosystem model forcing error, there remains the problem of unpredictable ecosystem initial/boundary conditions. One solution to the latter is to apply the same variational techniques used to fit initial/boundary conditions in physical models to the ecosystem model fit. However, the fit of the ecosystem model initial/boundary conditions is clearly dependent on the values of the biological parameters used, and the neglect of this dependence would seem much harder to justify than for the physical variables. Consequently, by this approach, the ecosystem initial/boundary conditions may need to be fitted simultaneously with the biological parameters, resulting in a much larger control parameter space dimension. This tends to prohibit a thorough parameter search using 'non-greedy' optimisation methods such as simulated annealing and genetic algorithms, which avoid 'trapping' in local minima of *cost* resulting from the nonlinearity of the regression (Press et al., 1999).

Such fits have in fact been achieved. For example, Li et al. (2006) simultaneously fitted spatially-variable 'forcings' in the form of zooplankton 'sources', spatially-variable parameters in the form of mortality rates, and spatially-variable initial conditions using the greedy-but-efficient variational adjoint method. It is not clear, however, that such large numbers of model input errors can be skillfully estimated in combination with poorly-known biological parameters, i.e. without overfitting the data and losing predictive accuracy (see Figure 1.2), or fitting the data for the wrong reasons and losing extrapolative skill. Similar concerns may be felt about the use of 'weak constraint' methods to fit ecosystem model forcings along with biological parameters (Losa et al., 2003, 2004). This approach aims to mitigate the combined effect of forcing errors and errors in the model dynamical formulation by fitting stochastic 'model errors'. Again, it is not established that biological parameter estimates can be properly constrained along with model errors (although the twin experiment in Losa et al., 2004, gives some evidence that they can). The alternative approach of fitting ecosystem model input errors *instead* of biological parameters (e.g. Nerger and Gregg, 2008; Gregg, 2008 and references therein) should help to constrain the model input errors, but seems to hold little hope of obtaining extrapolative skill in view of the large prior uncertainty in biological parameter values. In all of these approaches, errors in the assumed *a priori* statistics of ecosystem model input error may also result in loss of predictive skill (though use of error covariance estimates from a physical model-data assimilation should help to constrain the ecosystem forcing error statistics). Although certainly interesting lines of enquiry, they perhaps do not sufficiently acknowledge the differences between biological and physical model fits: biological model parameters are much more poorly known, model input error statistics are more uncertain, model dynamics are potentially more nonlinear, and the data sets are much poorer.

For these reasons, and computational practicality, errors in ecosystem model forcings/dynamics and initial/boundary conditions have generally not been fitted along with biological parameters, even for 0D and 1D ecosystem model data assimilations (see e.g. Evans (1999), Fasham and Evans (1995), Fennel et al. (2000), Hemmings et al. (2004), Hurtt and Armstrong (1996, 1998), Kuroda and Kishi (2004), Matear (1995), Schartau et al. (2001), Spitz et al. (1998, 2001) for single box studies; Dadou et al. (2004), Friedrichs et al. (2006, 2007), Prunet and Minster (1996), Prunet et al. (1996), Schartau and Oschlies (2003) for 1D studies). In such studies, biological initial conditions are usually assumed to be related to the forcing via a steady state assumption imposed by a period of model 'spin-up'. The physical forcing is usually taken as given from climatological estimates or the output of physical models, perhaps fitted separately to physical data. Even so, robust fits with reasonably constrained biological parameter estimates are often unobtainable for fitting more than a small number of biological parameters (Matear 1995; Schartau et al., 2001; Garcia-Gorriz et al., 2003; Hemmings et al., 2004; Friedrichs et al., 2006, 2007). And of course, if the model grid is made coarse to improve the accuracy of these model inputs, we return to the problem of large-scale dynamics being perturbed by the action of sub grid-scale variability.

Another strategy to mitigate these problems is to use a high-resolution ecosystem model, and simply average both model output and data over larger spatiotemporal scales in the model-data comparison (calibration and validation). In theory this might take care of the unpredictable forcing and inadequate resolution effects, and also the undersampling effect by including more data in the 'observations' at the comparison scale $O(X_i, L_c, t_j)$, thereby reducing the observational errors ϵ_{ij}^O. For example, time-averaging of data and fitted model output was applied in Schartau and Oschlies (2003) to allow for the effect of 'misplaced eddies' in their 1D ecosystem model fit due to 'physical phase errors' in the forcing provided by a 3D physical model. Presently, there are relatively few examples in the literature of 2/3D plankton model fits to real data where biological parameters are objectively optimised (Garcia-Gorriz et al., 2003; Huret et al., 2007; Li et al., 2006; Tjiputra et al., 2007) but many more are sure to appear in coming years. Perhaps the first question of such studies should be: can the model reproduce the more-predictable, large-scale patterns in the data when only these are used to constrain it? Run-time constraints may limit the rigour of the high-resolution ecosystem model fit and skill assessment, although it seems likely that such constraints will be substantially relaxed by future improvements in processing power. A more serious problem with this approach is the implicit disposal of information, or degrees of freedom, in the data that could potentially help to constrain the model. In many areas of the ocean, fitting e.g. monthly or mixed layer averages may simply not provide enough signal to constrain biological parameters.

1.5.5 Summary remarks

- Spatial variability which is 'unpredictable' in a deterministic sense causes error in ecosystem model dynamics, forcings and initial/boundary conditions and observational estimates at the averaging scales at which model and data are compared. This results in distortion of adjustable parameters during calibration, potentially over-punitive skill assessment, and ultimately inaccurate predictions.

- Many methods for mitigating these effects have been explored, but no clearly effective or 'best' solution has been demonstrated. Degrees of freedom in the model or estimation method need to be added strategically and rigorously tested for improvements in predictive skill. Differences need to be acknowledged between plankton modelling and the relatively well-sampled, dynamically linear (Hsieh et al., 2005) and well-known physical modelling fields from which many of the new methods originate. Perhaps more urgently required than new methods are methods of objectively and independently comparing different methods/models by estimating their predictive skill in realistic scenarios.

1.6 Methods investigated in this thesis

Here the statistical methods investigated in this thesis are outlined and motivated, and how they differ from and build upon previous methods is explained. Chapter 2 starts with the idea that the combined effect of errors in fine-scale plankton model forcings and initial/boundary conditions might be jointly accounted for (modelled) as effective time lags. The idea came from a mesoscale survey of the North Atlantic, for which the data in 'phase space' (the space of different plankton variables plotted against each other) showed a much clearer trajectory than the individual time series (Srokosz et al., 2003). This suggested that some of the mesoscale plankton variability could be described by displacements along a common dynamical trajectory, or equivalently, by 'phase errors' or 'time lags'. This 'phase relation'[1]

[1] Phase relation could be defined as a type of low-dimensional behaviour whereby at different locations, similar changes occur with different timing. A complementary low-dimensional phenomenon is 'synchronisation', whereby, at different locations, possibly-different changes occur with similar timing. Synchronisation in

phenomenon might be anticipated from the drivers of mesoscale plankton variability discussed above. Hence, if we can infer these time lags from the data, we might fit all the data from a large area using a 0D ecosystem model representing the fine-scale 'Lagrangian' dynamics following a single 'typical' fluid mass. The resulting parameter estimates would then be suitable for use in future high-resolution plankton models. In theory, this would mitigate the effects mentioned above by allowing time lags to 'mop-up' the net model-data discrepancy due to model input and dynamical error, hence allowing the model to effectively fit parameters to the sample scale ($L = l$) without explicitly resolving this variability. Similar methods of allowing for 'phase error' could be used to make 'suitably forgiving' calibrations and skill assessments of plankton models at various levels of spatial resolution.

A simple method is presented to estimate time lags within a Bayesian framework, whereby the time lags are extra adjustable parameters in the ecosystem model fit constrained by prior assumptions about their statistical distribution. This 'variable lag fit' is tested using simple experiments generating artificial data with a 0D NPZ model and fitting the data with the same model.

However, to assess the extrapolative skill of a method for fitting plankton models to real data, twin tests are seldom adequate because, in reality, the plankton model will almost certainly not fit all the data. Chapter 3 explores a method to estimate extrapolative skill by fitting a strictly-empirical 'data simulation' model to all the data, then using it to generate ensembles of data sets for fitting and testing different 0D plankton models, allowing estimation of their predictive skill as well as bias and variance as functions of time. The method is a kind of replicated 'sibling' experiment (Evans, 1999) or 'Observational System Simulation Experiment' (McGillicuddy et al., 2001) which also employs the technique of dividing the data into calibration ('active') and validation ('passive') subsets (cf. Friedrichs et al., 2007; Smith et al., submitted). The method is used to estimate the skill of a 'weak prior bulk plankton model' — the standard approach of aggregating space and allowing extensive retuning of biological parameters (i.e. using a broad or 'weakly informative' Bayesian prior pdf or '*a priori* bias function') to account for the statistical effect of unpredictable spatial variability on coarse-scale dynamics. The question is: Can this approach yield better skill than a strictly empirical or 'inductive' model, when asked to extrapolate from real Georges Bank data?

Although the coarse-resolution modelling approach of Chapter 3 may, in some cases, 'shield' the biological parameter fit from the effects of unpredictable model inputs and fine-scale observational error, the extrapolative skill will likely always be limited by inability to account for the Reynolds terms in (1.7)) without distorting parameter estimates. In recent years, investigators have begun to seek ways of parameterising the 'physical Reynolds terms' or eddy-fluxes in (1.7) to improve the predictive accuracy of coarse-resolution plankton models, mostly by modification of the standard 'eddy-diffusivity' approach to account for the finite 'reaction' timescale of plankton (Pasquero, 2005; Lévy, 2006). None, it seems, have attempted to parameterise the contributions due to nonlinear biological dynamics, even though studies have shown that these contributions can be very significant (Brentnall et al., 2003).

To address this problem, Chapter 4 investigates methods of accounting for the 'biological Reynolds term' (see section 1.5.3) by modelling the evolution of the phase space distribution of

plankton ecosystem models has been studied in Hillary and Bees (2004) and Guirey et al. (2007).

plankton concentrations. The main method investigated is to dynamically model the spatial moments (covariances) of plankton variability — a technique which has been successfully used for many years in 'turbulence closure' models of coarse-scale velocity dynamics. To this end, a simulation study is performed whereby coarse-scale output from a simple, spatially-resolved, reaction-diffusion spatial plankton model is compared to the predictions of the standard 'mean field approximation', which neglects biological Reynolds terms, and various 'spatially-implicit' plankton models. These latter parameterise the Reynolds terms by making suitable assumptions about the evolution of the distribution of plankton concentrations in phase space (e.g. the neglect of third and higher central moments).

Finally, in the Conclusions Chapter 5 a brief summary is given of the most important findings. Limitations of the methods investigated and their potential importance in a broader context is discussed, including the most promising lines of future enquiry to build on this work.

Appendix 1A: Model Fitting

Model fitting (or 'data assimilation') offers a more rigorous and objective way to adjust model parameters than 'fit-by-eye' approaches. However, there are limits to this objectivity, particularly when a fitting method is claimed to be 'optimal'. In this Appendix, the main types of estimation methods are summarised and their 'optimality' or otherwise is discussed.

Irrespective of the statistical methodology or paradigm adopted by the modeller, the Bayesian framework provides a convenient means of categorising different approaches to model fitting. Bayes' rule defines the 'posterior probability' $p_{12}(a|D,\mathcal{M})$ of adjustable parameter values a given the data D and model formulation $\mathcal{M}$ in terms of the 'Likelihood' or probability $p_{21}(D|a,\mathcal{M})$ of the data D given a and $\mathcal{M}$, the 'prior probability' $p_1(a|\mathcal{M})$ of the adjustable parameters a given $\mathcal{M}$, and the total probability $p_2(D|\mathcal{M})$ of the data D given $\mathcal{M}$ (see e.g. O'Hagan and Forster, 2004):

$$p_{12}(a|D,\mathcal{M}) = \frac{p_{21}(D|a,\mathcal{M})p_1(a|\mathcal{M})}{p_2(D|\mathcal{M})} \tag{1A-1}$$

In order to compare different model formulations, the Bayesian framework also allows the assignment of prior probabilities $p(\mathcal{M})$ to the model formulations themselves. However, these are irrelevant to the present discussion of how to fit any single model (and we are not considering multi-model predictions), hence the argument $\mathcal{M}$ will be dropped. Referring to (1A-1), the various approaches used by plankton modellers to fit a set of adjustable parameters a may be broadly categorised, in Bayesian language, as 'posterior mode' or 'posterior mean' types of estimation[2].

1) 'Posterior mode' type estimation In this category of approach, we choose $\hat{a}$ such that it maximises or minimises an objective function $f(a,D)$ of the adjustable model parameters a and the data D. We may choose to interpret this as maximising a posterior probability $p_{12}(a|D,\mathcal{M})$, or more usually, minimising $cost = -\log p_{12}(a|D,\mathcal{M})$. However, the probability

[2] Posterior *median* estimation appear not to have been tried in plankton modelling. Perhaps this could be a more robust estimate than the mode and mean, which latter may be distorted by inaccurate modelling of the distribution tails?

interpretation is not prerequisite for a valid estimation method. For example, more accurate predictions may in some cases result from using a 'sum of squares' (or 'least squares') objective function even if a normal Likelihood function (which is consistent with a 'sum of squares' *cost* function) is not regarded as an accurate description of model/data errors (see section 3.2.6 and the lognormal example below).

Setting $p_1 = const$, the Bayesian method reduces to Maximum Likelihood Estimation (MLE) (Edwards, 1972). In the frequentist paradigm of 'objective' probability, 'true' probability distributions of events, estimates and predictions are defined via an infinite 'ensemble' of idealised 'trials' or 'repeat experiments' (see section 3.2.1 for discussion of this in the context of plankton modelling). Within this paradigm, if MLE is applied with 'perfect' formulations of the model and of the data error distribution, and with an asymptotically-large sample size (number of data), the MLE estimates $\hat{a}_{MLE}$ will be unbiased and have the minimum possible variance for *any* non-Bayesian estimate (Stuart et al., 1999). Hence predictions made using $\hat{a}_{MLE}$ will be 'optimal' in the sense of minimal mean-square error over the 'true' ensemble. Note, however, that such conditions are rather unlikely to pertain in plankton modelling. With finite sample size, MLEs are not, in general, unbiased (e.g. the MLE of the standard deviation of a normal distribution) or indeed minimal-variance (e.g. when the data are 'truly' lognormal, the simple sample mean obtained by Ordinary Least Squares estimation has a lower mean-square error than the MLE of the mean for small sample size, even if the Likelihood is correctly specified).

More accurate Bayesian estimates can always be made by incorporating accurate prior information (for example: the posterior mode given a prior centred on the true value with zero width!). In general, however, Bayesian posterior mode estimates obtain lower variance at the price of higher bias, due to the prior constraints. It is often argued that Bayesian estimates are 'best' because they make use of all available information, including subjective prior information (e.g. O'Hagan and Forster, 2004). This is not true, however, if the prior information is sufficiently biased (an alternative term for 'prior information' is *prejudice*). For example, an 'unfair' coin might have a relative frequency of landing 'heads' $\theta^t = 0.7$ in the long term (after $n \to \infty$ tosses). In this case, a rational Bayesian might impose a subjective prior centred on $\theta = 0.5$ (a fair coin). Given the same binomial Likelihood model, Bayesian estimates of θ^t based on n tosses may (depending on the form of the prior) be less accurate than the MLE for moderate n, because the posterior distribution of θ converges more slowly than the Likelihood function to a delta-function on the 'true' value θ^t as n increases. For small n, however, even biased prior information may be better than no prior information at all, viz. the higher variance in $\hat{\theta}_{MLE}$ may overcompensate for the higher bias in $\hat{\theta}_{MB}$.

Posterior mode estimation, or 'variational data assimilation', has been the more usual method of plankton model fitting to date (Gregg, 2008). In these studies, usually only biological model parameters and possibly initial/boundary conditions were fitted. These are 'strong constraint' data assimilations in the physical oceanography jargon, since model forcings were not adjusted. A 'weak constraint' plankton model fit, with forcings adjusted as well as initial/boundary conditions and biological parameters, was attempted using posterior mode estimation in Losa et al. (2004).

2) 'Posterior mean' type estimation Here, we choose $\hat{a} = \int a' f(D, a') da'$ for some integrating function f. In the Bayesian interpretation, where f is the posterior pdf p_{12}, this yields the

Bayesian posterior mean estimate $\hat{a}_{MEB}$. This estimate may be more robust (lower variance) than $\hat{a}_{MB}$, since information about the fit for a range of parameter values is used, although it may still be biased due to bias in the prior pdf. Suppose, however, that the parameter or variable a involved in the Bayesian prior can be regarded as a frequentist random variable with known pdf $p_1(a)$ (e.g. a might be an initial condition, or the outcome of a coin toss). If the conditional pdf $p_{21}(D|a)$ of data D given a is also known (e.g. D could be the result of a plankton measurement, or the output of a mechanical 'coin reader'), then the 'optimal' minimum mean-square error estimate $\hat{a}_o$, over the 'true' ensemble, of a given D, p_1 and p_{21} is the 'posterior mean' $\hat{a}_{MEB}$:

$$\begin{aligned} \hat{a}_o &= \frac{\int a p_{21}(D|a)p_1(a)\mathrm{d}a}{\int p_{21}(D|a)p_1(a)\mathrm{d}a} \\ &= \int a p_{12}(a|D)\mathrm{d}a \\ &= \hat{a}_{MEB} \end{aligned} \tag{1A-2}$$

which is easily derived by minimising the mean-square error:

$$MSE = \int (\hat{a} - a)^2 p_{21}(D|a)p_1(a)\mathrm{d}a \tag{1A-3}$$

with respect to $\hat{a}$ for each value of D. The posterior mean $\hat{a}_{MEB}$ is then unbiased, minimal variance and distributed according to p_{12} over the 'true' ensemble. Hence a Bayesian optimal estimate can also be optimal within the frequentist paradigm of 'true' probability. The difficulty for the frequentist paradigm comes when a is a strictly fictitious quantity, such as a structural parameter θ within a biological model. In this case, θ only strictly has a meaning within the model. The prior estimate $\hat{\theta}^{pr}$ might be thought of as a random variable — a function of prior model fits — varying over an objective ensemble of repetitions of all previous experiments with a 'prior pdf' $p_1(\hat{\theta}^{pr})$. However, this prior estimate is, in reality, independent of the data D. Hence the above optimality argument does not apply, if $a = \theta$, within the frequentist paradigm. Nevertheless, the posterior mean can still be justified as a robust estimate within either paradigm. Without prior constraints, the resulting Mean Likelihood Estimate (MELE) shares the asymptotic unbiased and efficiency (minimal variance) properties of the MLE, but is often more robust for finite samples (McLeod and Quenneville, 2001).

Posterior mean estimation has been commonly used to estimate biological parameters, but less for making predictions in studies where biological parameters are adjusted. This may be because plankton model posterior pdfs are 'nasty' functions of the biological parameters: the regression is highly nonlinear and local maxima in the posterior pdf are common. Consequently, predictions generated by the posterior mean parameters may be obviously poorer fits than posterior mode predictions, and in fact poorer estimates of the true fields. Also, integrating over the posterior pdf is nontrivial for the 5–10 biological parameters typically adjusted, and is likely to require Monte Carlo methods (Harmon and Challenor, 1997). When posterior mean estimation is applied to model forcings as well as initial/boundary conditions (weak constraint fit) the integral becomes of too high dimension even for Monte Carlo methods, and a simplification is necessary — typically, to only allow information to propagate forwards in time. This results in a 'filter' or 'sequential data assimilation' which is amenable to Monte Carlo methods (Losa et al., 2003). Sequential data assimilation has been applied quite commonly to fit plankton models (Gregg, 2008), but almost always without adjusting biological parameters — the exception here being Losa et al. (2003).

Appendix 1B: Model Testing

Model testing or validation is a crucial part of any modelling study which seeks to draw inferences from model output. In oceanography, an emerging community jargon for this is 'skill assessment'. Assessing how wrong a model's predictions are (validation or skill assessment) is a separate, additional estimation problem to the one of estimating true fields or making model predictions (calibration or model fitting — see Appendix 1A). In general, we want to know how wrong a model's predictions are (in some specified sense e.g. root-mean-square error) in a range of conditions or spatio-temporal coordinates which may extend beyond those of the fitted data ('intra-' and 'extra-sample' accuracy, in the statistical modelling jargon). Furthermore, it is useful to distinguish systematic error (bias) from random error (variance). As discussed in section 1.2, empirical evidence for the mechanistic veracity of a model is provided by demonstrated predictive accuracy at high levels of extrapolation from fitted data.

The specific predictive purpose of the model may be mathematically defined as maximising a 'skill function'. This skill function then informs the choice of fitting method (cost function for calibration) and suggests 'skill metrics' which might be applied to estimate the skill function, or relative skills of different models, with available data (see Chapter 3). With the toy problem presented in Figure 1.2, it is possible to test different types of skill metric and measure their 'performance' in terms of the true skill of the model chosen by the skill metric, on average over repeat experiments (see Table 1B.1). The take-home message seems to be that when extrapolative skill is demanded of a model, skill metrics which only account for noise-fitting in the calibration data, such as Likelihood Ratio Tests (Stuart et al., 1999) or Akaike's Information Criterion (Akaike, 1973; Burnham and Anderson, 1998), do not perform as well as 'cross-validation' or 'division' skill metrics which demand extrapolation by the model between disjoint calibration and validation data subsets (Stone, 1974). Though not tested here, similar shortfalls would be expected of Bayesian skill metrics based on calibration data, such as the Bayesian Information Criterion or the 'Evidence' $p_2(D|\mathcal{M})$ in (1A-1). Data simulation or 'parametric bootstrap' techniques (Young, 1994) may allow more robust 'simulation and division' skill metrics, particularly when the number of different disjoint data divisions is small (resulting in a 'noisy' cross-validation metric — see Hastie et al., 2001 and Table 1B.1). In such cases, simulation may also improve the robustness of predictive bias and variance estimates.

The distinction between calibration and skill assessment seems to have gone under-appreciated in plankton modelling. Usually, the calibration *cost* (typically, a spatiotemporally-averaged squared misfit) is used as a skill metric to compare different model formulations. This is, in general, a poor estimate of intra-sample accuracy, because it neglects observational error, favours more complex models which are more able to fit observational errors, and fails to test extrapolative accuracy (see Table 1B.1 and Chapter 3). Taylor diagrams (Taylor, 2001) are very widely used to display average squared misfit in geometric relation to model-

Table 1B.1: Skill and skill metric performance for the toy problem simulation study (Figure 1.2). 5000 simulations were performed to evaluate the frequency with which each model obtained the highest $Skill^{(1)} = -\sum_i (y^t(x_i) - y^p(x_i))^2$ and the average value of $Skill^{(1)}$ for each model over the ensemble: $Skill^{(2)} = E[Skill^{(1)}]$ (first three rows). Also, for each of the 5000 data sets, a different skill metric was used to select the model with the highest $Skill^{(2)}$. The relative frequency with which each model was selected by each metric is shown, along with the resulting average true value of $Skill^{(2)}(choice)$ over all 5000 model choices, this being a measure of the overall performance of the skill metric. The 'Naïve' metric was the calibration *cost* (row 6). Akaike's Information Criterion (AIC) was used to correct the calibration *cost* for noise fitting (row 7). The Likelihood Ratio Test (LRT, row 8) was applied to reject successively more complex models with a significance threshold $\alpha = 0.15$ (the more standard threshold of $\alpha = 0.05$ yielded even more parsimonious model selection). Simulation (row 9) involved fitting a statistical data simulation model to the given data set and using it to create a 'simulated ensemble' of 50 auxiliary data sets, each used to fit the models and estimate skill with the simulation model as a proxy for 'truth' (sometimes called the 'parametric bootstrap' method). Skill estimates were averaged over a new simulated ensemble for each of the 5000 true ensemble data sets. Division$^{(n)}$ (rows 10–12) involved using only n of the 5 data clumps to fit the models and estimating skill using the remaining $5-n$ clumps. Here, skill estimates were averaged over all possible calibration/validation combinations. Simulation&Division$^{(n)}$ (rows 13–15) involved applying both data simulation and division techniques simultaneously.

	Constant	Linear	Quadratic	Cubic	
Freq. Best $Skill^{(1)}$ (%)	0	16	80	4	
$Skill^{(2)}$	-1.015	-0.073	-0.040	-1.153	
	Freq. Selected (%)				
Skill Metric	Constant	Linear	Quadratic	Cubic	E[Skill$^{(2)}$(choice)]
Naïve	0	0	0	100	-1.12
AIC	0	46	40	14	-0.20
LRT (α=0.15)	8	92	0	0	-0.15
Simulation	0	24	45	31	-0.38
Division$^{(4)}$	0	55	36	9	-0.16
Division$^{(3)}$	0	85	15	0	-0.07
Division$^{(2)}$	0	100	0	0	-0.07
Simulation&Division$^{(4)}$	0	77	22	1	-0.08
Simulation&Division$^{(3)}$	0	99	1	0	-0.07
Simulation&Division$^{(2)}$	0	100	0	0	-0.07

data spatiotemporal correlation and model/data spatiotemporal ‘variance’ (after correcting for average misfit or spatiotemporal ‘bias’). Although certainly an effective visualisation tool, the utility of Taylor diagrams for skill assessment is perhaps overestimated, and they seem to have instilled a rather specialised use of the terms ‘bias’ and ‘variance’ in the plankton modelling community. Bias and variance in classical statistical modelling are defined by true fields and an ensemble, and may therefore be defined at single points in space/time. The average (over space/time) misfit is an unbiased estimate of average ensemble bias (for unbiased data), although this is not a very informative quantity (since biases at different places/times can compensate each other). However, the spatiotemporal ‘variance’ is not an unbiased estimate of average (over space/time) ensemble variance (we can make predictions with little spatiotemporal variability but high predictive error variance), nor is the average squared misfit an unbiased estimate of average ensemble mean-square error (or intra-sample accuracy). Furthermore, simple linear correlation between model and data does not account for time lags or phase error (explored in Chapter 2). For example, the Taylor diagram would attribute equal ‘skill’ to a model which predicts all features perfectly except for a single phase error (causing zero correlation) and a model which is entirely random in space and time but with the correct spatiotemporal variance.

Chapter 2

Time Lag Model Fitting

Published article: Wallhead, P.J., Martin, A.P., Srokosz M.A., 2006. Accounting for unresolved spatial variability in marine ecosystems using time lags. J. Mar. Res. 64, 881–914.

2.1 Introduction

Accounting for spatio-temporal variability in fitting marine ecosystem models is a challenging problem. Limitations on our ability to track turbulent fluid motions, by observations and/or physical circulation models, inevitably result in a mixing or confusion of spatial and temporal variability in 'Lagrangian' ocean data (data attributed to 'individual' water parcels). A solution might to be somehow filter out these effects, so that a biological model can be fitted to the temporal variability within a 'typical' moving water mass (the 'Lagrangian biological trajectory'[1]), rather than the combination of spatial and temporal variability represented in most practical (non-Lagrangian) data sets. A model thus-fitted could provide more accurate tests of biological hypotheses, and better predictive skill when coupled to a physical model of mesoscale circulation or even a General Circulation Model.

This chapter demonstrates that if unpredictable 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 biased as estimates of their 'true', small-scale, Lagrangian counterparts. A method is demonstrated to account for this spatial variability by assuming random between-sample time lags imposed on a common biological trajectory (temporal variation). The method is equivalent to allowing 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.

The method was largely inspired by a cruise data set from the eastern North Atlantic collected after a period of unsettled weather. It 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. 2.1 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 intermittency in the sampling times. Naïvely, this may suggest that models would be better fitted to the phase space trajectory

[1] A dynamical trajectory in phase space, not to be confused with a 'physical' or 'spatial' Lagrangian trajectory describing the motion of a water parcel in real space.

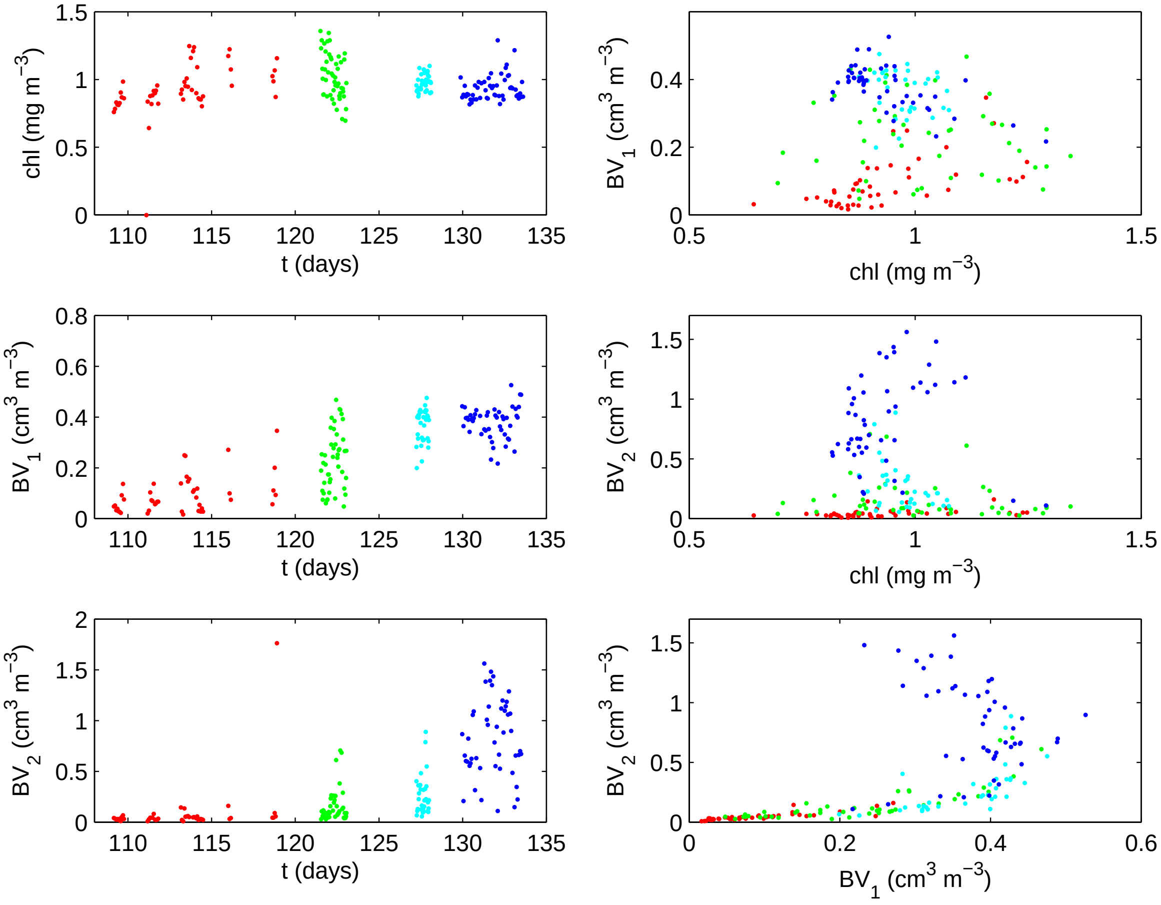

Figure 2.1: Data from RRS Discovery cruise D227 (with permission, from Srokosz et al. (2003)) displayed as time series (left), and in state variable phase space (right). Mixed layer concentrations of chlorophyll (chl) and two size classes of zooplankton ($\mathrm{BV}_1 = 250 - 500\mu$m and $\mathrm{BV}_2 = 500 - 1000\mu$m biovolume) are shown, with different symbols used for different sampling subintervals.

rather than the time series 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. The method described here 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 the issue of the extent to which spatial variability in real data can be attributed to time lags is not addressed. Before answering that question it is necessary to develop the techniques needed to diagnose the phenomenon if it is there, and assess the importance of correcting it. This is done here using twin tests. The potential impact of between-sample time lags on standard model fits which do not account for them is investigated, then a fitting technique is developed 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' or 'time series' 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. The measurement error and time lag statistics are assumed to be known in this study, but methods of estimating these from real oceanographic data sets are also discussed.

The chapter is structured as follows. Section 2.2 details assumptions (including the model equations), methods of generating artificial data and fitting the model to it, and how 'fit performance' is evaluated (skill assessment). Section 2.3 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 2.4.

2.2 Methods

2.2.1 Lagrangian Biological Model

For convenience of analysis and numerical implementation, a simple (though potentially applicable) marine ecosystem model is used for the twin tests. A basic requirement is that the modelled mean trajectory may yield 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. The chosen system executes free, periodic, stable nonlinear oscillations with a frequency ω, which is some unknown function of the internal model parameters. This behaviour is observed in some more complex marine ecosystem models (e.g. Ryabchenko et al., 1997), but it is not a prerequisite for the technique, and serves rather as a simple proxy for the more generic case of external seasonal forcing. The variable lag fit is essentially a 'type II regression' and as such may be applied to any system of ODEs with an independent variable (in this 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 behaviour varies significantly between sampled water masses).

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

$$\begin{aligned}
\frac{dN}{dt} &= -u_0 \frac{N}{k_n+N}\frac{1}{1+cP}P \; + \gamma R(1-e^{-\Lambda_z P})Z + m_p P \\
&\quad + r_z m_z Z^2 + k(N_0 - N) \qquad (2.1) \\
\frac{dP}{dt} &= u_0 \frac{N}{k_n+N}\frac{1}{1+cP}P \; - R(1-e^{-\Lambda_z P})Z - m_p P \\
&\quad -(k+s_p)P \qquad (2.2) \\
\frac{dZ}{dt} &= e_z R(1-e^{-\Lambda_z P})Z - m_z Z^2 \qquad (2.3)
\end{aligned}$$

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 Tabel 2.1, 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

Table 2.1: Model structural parameters, true values and relative sensitivities.

Parameter	Symbol	'True' value θ_i^t	Sensitivity
maximum P growth	u_0	1.0 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})$^{-1}$	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 C m^{-3})$^{-1}$	high
P recycled losses	m_p	0.15 day^{-1}	medium
Z remineralisation fraction	r_z	0.5	low
Z mortality rate	m_z	1.5 (g C m^{-3})$^{-1}$ 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

forms parameterise 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-parameterised uptake of nutrient by phytoplankton, phytoplankton self-shading, Ivlev-parameterised grazing of phytoplankton by zooplankton (replacing the Hollings III function used in Edwards and Brindley (1996), which may produce excitable behaviour 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 remineralised 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. Vertical mixing also allows influx of nutrient (only) from below the seasonal pycnocline, where the nutrient concentration is N_0 (the model only explicitly represents shallower waters). The robustness of our results to changes in the choice of model formulation is discussed in section 2.3.4.

In order to assess the relative sensitivity of the chosen true solution to changes in each of the parameters, artificial Lagrangian data are generated with 5% measurement error (see section 2.2.3). Then each parameter is varied in turn from 90 to 110% of their 'true' values, and the maximum increase in model-data discrepancy (*cost*) is computed. Tabel 2.1 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 (2.1–2.3) 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 horizontal scale represented by the optimal parameter set will be the total area of fluid spanned by the samples. By allowing variable time lags, this scale may be reduced to a scale on which the sampled dynamics may be described by a common trajectory with variable time lags. Imposing a weaker restriction on the model thus allows 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.

2.2.2 Between-Sample Variability in Time Lag

Unresolved spatial variability is assumed to impose time lags on the time series data. For the twin test, 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, including fluxes of nutrient/plankton and stratification changes due to horizontally propagating features such as eddies. 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 moving at ~400km/day (such as produced the data in Figure 2.1, 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 assumption to be realistic.

To choose a realistic range of time lag variability to use in our twin tests, a rough estimate is obtained from the time series in Figure 2.1 by examining the vertical scatter about the mean value of one of the sampled variables (e.g. BV_1) over a subinterval (e.g. the green dots), and estimating the time interval over which this mean value changes by an equal amount (cf. the blue dots). An approximate upper limit of 10 days is thereby estimated for a 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. Estimation of this parameter, which is assumed to be known in the twin tests, is discussed in section 2.3.3.

Note that even if these assumptions are not strictly met in reality, they may yet prove adequate as assumptions in estimating the Lagrangian biological trajectory and biological parameters from the data (see section 2.2.6). Dealing with the effects of violation of these assumptions is discussed in section 2.3.4.

2.2.3 Measurement and Model Errors

To generate the data, a 'true' error model is adopted in which all measurement errors are independent random variables sampling from a normal distribution with zero mean and standard deviation proportional to the modelled mean value (after time lagging). Negative sampled values are disallowed, effectively truncating the true error distribution, although this was a rare occurrence for the 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 unaccounted for by time lags is assumed to be indistinguishable from measurement error viz. no covariance or non-Gaussianity contribution is included in the total error. The error model assumed in fitting (section 2.2.6) is identical to the true error model barring the (small) truncation effect. Again, section 2.3.4 discusses ways to deal with violation of these assumptions.

2.2.4 Generating a Noisy, Non-Lagrangian Data Set

The 'true' biological dynamics used to generate the data are specified by (2.1–2.3) and the 'true' biological parameter set θ^t shown in Tabel 2.1. 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 $[N(0), P(0), Z(0)]$ are chosen from a point on the repeating cycle, of period roughly 40 days, that results from running the model for 710 days from an arbitrary but fixed set of starting values to eliminate any transient. The model is run for a further integration time of 80 days and the output is recorded at intervals of 1/8 days, producing a non-transient trajectory over the sampling interval of 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 $N_s = 40$ true 'dynamical times' given by:

$$t'^t = t^s - \tau^t \tag{2.4}$$

where t^s is a 40-component vector of sampling times as measured by the sampler, starting at 20 days of integration time and advancing uniformly at intervals of 1 day, τ^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 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 simultaneous measurements are assumed of all modelled variables at each sampling time. This allows time lagging to be most easily distinguished from other sources of noise, although it may be difficult to obtain for real data sets. To simulate measurement error, independent Gaussian noise is added to each variable datum, with constant proportional variance $(s_j y_j^t(t_i'^t))^2$ where $y_j^t(t_i'^t)$ is the true mean value of the j^{th} variable at the i^{th} true sampled dynamical time $t_i'^t$ 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.

2.2.5 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. The problem is to use it to obtain optimal estimates of the true Lagrangian biological trajectory y^t and associated true biological parameters θ^t starting with plausible initial guesses θ^i, given the correct biological model formulation given by (2.1–2.3), the correct level of *a priori* uncertainty in the dynamical time of samples represented by the modelled variance σ_τ^2 and the correct *a priori* fractional uncertainty s_j. No prior information about y^t and θ^t is assumed except that all trajectory values and parameters are non-negative.

First, the data are fitted using a standard time series fitting technique: 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 modelled time lag vector τ is set to zero, and the fit is independent of the *a priori* uncertainty σ_τ. The data are then fitted using the new technique: the 'Variable Lag Fit' (hereafter VLF). This assumes that τ 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 a product of the true trajectory and measurement errors. 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, optimising the VLF does involve added complications beyond those of optimising the ZLF, as will be seen below.

To infer model parameters from the data, the 'posterior probability' of the model parameter set is maximised (posterior mode estimation). This is 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 estimates $\hat{a} = \{\hat{\theta},\hat{\tau}\}$ of the full set of 'true' adjustable parameter values $a^t = \{\theta^t,\tau^t\}$. The method can be interpreted as Bayesian estimation or Maximum Likelihood Estimation where the Likelihood function includes errors in the regressor variable (time) — hence a 'nonlinear model II regression' type with 'controlled' regressor errors in the terminology of Laws (1997), or a 'functional relationship without replication' in statistical jargon (Seber and Wild, 2003). For the ZLF, estimates will be classical Maximum Likelihood (ML) estimates without regressor errors.

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). Given the correct model formulation and hypothesis, these ML estimates will also be 'consistent', meaning that as the number of observations becomes large, they can be expected 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 over an ensemble of repeat experiments). It is also desirable for estimates to be 'efficient', meaning that they have minimal variance over the ensemble. For the ZLF, fitted to data without time lags, and given the correct model formulation, the ML estimates will be asymptotically (as sample size $n \to \infty$) 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 inverse of the Hessian matrix for parameter perturbations about the ML solution for a single data set (Thacker, 1989; 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' parameters (dimension of t'), thus theoretical results associated with large sample size may not apply. The incidental parameter estimates $\hat{t'} = t^s - \hat{\tau}$ 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 $\hat{\theta}$ may remain consistent, but perhaps only if the ratio of the coefficients s/σ_τ is known, and even then, the MLEs $\hat{\theta}$ may not be asymptotically efficient (Neyman and Scott, 1948; Seber and Wild, 2003). This forewarns us that the VLF parameter estimates might not be robust. As it turns out, a modified maximisation function is required to address this problem (see below).

2.2.6 Objective Functions

The objective or cost functions for the VLF and ZLF are specified by a general hypothesis defining the VLF, of which the ZLF submodel is a restriction. The composite hypothesis $\mathcal{H}$ asserts the following:

$\mathcal{H}^{(1)}$: The observational vector of (fixed) sampling times is given by $t^s = t' + \tau$ where t' is a vector of modelled dynamical times (dimension $N_s = 40$) and τ is the vector of modelled 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^s is given by $o_{ij} = y_j(t'_i|\theta) + \epsilon_{ij}$, where $y_j(t'_i|\theta)$ is the output of model (2.1–2.3) at time $t'_i = t_i^s - \tau_i$ given the set of parameters and initial conditions θ, and ϵ_{ij} is an independent Gaussian random variable with mean zero and variance $(s_j y_j(t'_i|\theta))^2$, representing measurement noise and remaining model error.

The estimation method now consists of maximising the posterior pdf with respect to the adjustable parameters $a = \{\theta,\tau\}$, given the data D and under the hypothesis $\mathcal{H}$:

$$P(a|D) \propto p^{(1)}(\tau)p^{(2)}(D|\tau,\theta) \tag{2.5}$$

where $p^{(1)}$ is a 'prior' pdf of the time lags, and $p^{(2)}$ is the 'Likelihood' of the data given the model with parameter values $a = \{\theta,\tau\}$. Note that the assumed independence of τ and the ϵ_{ij} allows us to decompose the posterior pdf in this way.

Now, the ZLF restricts the optimisation to ($\tau = 0$), whilst the VLF allows non-zero τ within the allowed range of ±20 days for each component (see section 2.2.4). Therefore, the posterior pdf for the standard ZLF is given by:

$$M_{ZLF} = p^{(1)}_{ZLF} \cdot p^{(2)}_{ZLF} \tag{2.6}$$

where

$$p^{(1)}_{ZLF} = \prod_{i=1}^{N_s} \frac{1}{\sqrt{2\pi\sigma_\tau^2}} \tag{2.7}$$

is the (restricted) prior pdf of the time lags (a constant), and

$$p^{(2)}_{ZLF} = \prod_{i=1}^{N_s}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(s_j y_j(t_i^s))^2}} e^{-\frac{(o_{ij}-y_j(t_i^s))^2}{2(s_j y_j(t_i^s))^2}} \tag{2.8}$$

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 posterior pdf is given by:

$$M_{VLF} = p^{(1)}_{VLF} . p^{(2)}_{VLF} \tag{2.9}$$

where

$$p^{(1)}_{VLF} = \prod_{i=1}^{N_s} \frac{1}{\sqrt{2\pi\sigma_{\mathcal{T}}^2}} e^{-\frac{(\tau_i)^2}{2\sigma_{\mathcal{T}}^2}} \tag{2.10}$$

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

$$p^{(2)}_{VLF} = \prod_{i=1}^{N_s} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi(s_j y_j(t'_i))^2}} e^{-\frac{(o_{ij} - y_j(t'_i))^2}{2(s_j y_j(t'_i))^2}} \tag{2.11}$$

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

$$t'_i = t^s_i - \tau_i \tag{2.12}$$

In this study, no attempt is made to optimise the statistical parameters s and $\sigma_{\mathcal{T}}$. In all optimisations, the same value of s is used during fitting as that used to generate the data. For fitting to data 'generated on the null' ($\tau^t = 0$), $\sigma_{\mathcal{T}}$ is set equal to some fixed (possibly wrong) value for each optimisation. For fitting data generated 'on the alternative' ($\tau^t \neq 0$), the same value of $\sigma_{\mathcal{T}}$ is used 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 $\sigma_{\mathcal{T}}$), it serves as a starting point which ensures consistent *a posteriori* estimates of θ^t using the VLF (Seber and Wild, 2003), and which may later be generalised to include consistent *a posteriori* estimates of s and $\sigma_{\mathcal{T}}$ where 'replicated' data are available (see section 2.3.3).

Now, maximising M_{VLF} in (2.9–2.12) yields consistent structural parameter estimates given accurate *a priori* estimates of the error parameters (s, $\sigma_{\mathcal{T}}$), but these are not necessarily the 'best' as regards bias and variance. The problem is that by relaxing constraints on the sampled time lags, variance is increased in the estimated frequency of oscillation. To suppress this variance, the VLF maximisation function is multiplied by a 'lag drift penalty' term S which measures the probability, assuming independent Gaussian time lags with variance $\sigma^S_{\mathcal{T}}$, of obtaining a persistent 'drift' in the time lags:

$$S(\tau|\sigma^S_{\mathcal{T}}) = \frac{\max(p'_{(0)}(\tau|\sigma^S_{\mathcal{T}})}{\max(p'_{(A)}(\tau|\sigma^S_{\mathcal{T}})} \tag{2.13}$$

where $p'^{(A)}$, $p'^{(0)}$ are the maximum Likelihoods for fitting mean trends $\overline{\tau}(t) = At + B$ and $\overline{\tau} = C$ respectively to the series of time lags plotted against sampling time. S can be thought of as an additional prior on the time lags which strengthens the hypothesis of independence — a kind of automated 'residual analysis' on the time lags. With this penalty term the calibration cost function for the VLF becomes:

$$M'_{VLF} = p^{(1)}_{VLF} . p^{(2)}_{VLF} . S(\tau|\sigma^S_{\mathcal{T}}) \tag{2.14}$$

$\hat{S} \to 1$ as number of sampling times $N_s \to \infty$, so the consistency of the structural parameter estimates is not spoilt by the maximisation of M'_{VLF}. A first choice of σ^S_τ might be σ_τ; however, better performance was obtained using a 'weighted' S with a reduced variance: $\sigma^S_\tau = 0.1\sigma_\tau$. The use of S with this 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-dimensionalisation, then estimate the timescales separately (Froda and Colativa, 2005). The method described here aims to constrain all structural parameters in a single estimation procedure.

In summary, the maximisation functions are M_{ZLF} defined in (2.6–2.8) for the ZLF and M'_{VLF} defined in (2.10–2.14) for the VLF. Unless otherwise stated, $\sigma^S_\tau = 0.1.\sigma_\tau$ as a weighting for S in the VLF.

2.2.7 Optimisation

The task is to maximise (2.6) with respect to the θ free parameter vector for the ZLF (13 rate constants + 3 initial conditions), and (2.14) with respect to the $\{\theta,\tau\}$ 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 total sample size $n = 120$). For the purposes of optimisation, it is more convenient to minimise the 'cost functions': $cost_{ZLF} = -\log M_{ZLF}$ and $cost_{VLF} = -\log M'_{VLF}$. In the ZLF, the 16 free parameters are all varied by the search algorithm. However, for the 56 free parameters of the VLF, a massive computational saving is obtained by performing a nested optimisation (similar to 'concentrating the Likelihood') over the 40 time lag parameters at each iteration. This is facilitated by the fact that the $cost_{VLF}$ decomposes (except for the small contribution of $-\log S$) into a sum of contributions from each sample time, allowing each τ_i to be optimised by choosing the value from the allowed range for which the cost increment is minimised. The contribution of $-\log S$ is then calculated and added to give the (partially) optimised cost for the trial value of θ.

The search algorithm varies the 16 structural parameters comprising θ in order to (fully) minimise the cost, using 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 optimisation 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 found to be 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 optimisations were all limited by a maximum number of iterations. This number was set by the criterion that the maximal 'optimisation error' due to finite iterations be much less than the typical between-data set variation for our chosen test statistic $\Lambda^{(V)}$ (see section 2.2.8 for definition). The maximal optimisation error was estimated by comparing with a single long optimisation of 5×10^5 iterations. It was found that, for model fits to data generated by the alternative hypothesis ($\tau^t \neq 0$), 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 ($\tau^t = 0$), 10^4 iterations were sufficient as long as the initial guess was close to the optimum, i.e. at $\theta^i = \theta^t$. Note that with finite noisy data, the exact optimal (ML) solution θ^0 is in general not coincident with the 'true' generating solution θ^t, since inevitably some of the noise is fitted by the modelled mean variability in the exact optimal solution. Since the main purpose of this study is to test an estimation method (cost function) rather than an optimisation algorithm, $\theta^i = \theta^t$ was used in some cases for speed of convergence. For all fits on the alternative ($\tau^t \neq 0$), the ZLF was slower (necessitating 10^5 iterations) as a result of a larger discrepancy between θ^t and θ^0, leading, as will be seen, to larger biases on the parameter estimates.

2.2.8 Skill Metrics

The principal measure of 'fit performance' or 'skill' in this study is the expected accuracy of the model fit in recovering (hindcasting) the true (Lagrangian) biological trajectory. Thus we estimate the relative skill using the following skill metric:

$$\Lambda^{(V)} = \max(\log p^{(V)}_{VLF}) - \max(\log p^{(V)}_{ZLF}) \qquad (2.15)$$

where $-\log p^{(V)}_{ZLF/VLF}$ is the 'validation cost' of the ZLF/VLF. The function $p^{(V)}_{ZLF/VLF}$ is identical in form to $p^{(2)}_{ZLF}$ (equation (2.8)), but here the 'validation data' are generated by θ^t with zero measurement noise and no time lags (the true (Lagrangian) trajectory sampled at times t^s), and the model prediction is generated by the estimates $\hat{\theta}_{ZLF/VLF}$ obtained by the ZLF/VLF in calibration. Of course, no time lags are included in the model prediction since the Lagrangian biological trajectory is being predicted. $\Lambda^{(V)}$ is a measure of the difference in total squared predictive error, weighted to give greater importance to predictive error where the true values (and hence measurement errors) are low. Note that the superscript (V) denotes a 'validation' statistic only accessible in the context of a twin test where the true trajectory is known. Also note that since (2.15) 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 most skillful prediction method. In addition, we will also estimate the bias and variance of the fitted parameter vectors $\hat{\theta}_{ZLF/VLF}$ over an ensemble of data fits.

2.3 Results and Discussion

2.3.1 Visual Skill 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 $\sigma_\mathcal{T} = 0$, 4 and 8 days and measurement error coefficients of $s = 0.05$ and 0.15, using $\theta^i = 0.9\theta^t$ and $1.1\theta^t$ as initial guesses. Two examples are shown, with $s = 0.05$ in Figure 2.2 and $s = 0.15$ in Figure 2.3. In both cases, the generating trajectory is well recovered when $\sigma_\mathcal{T} = 0$ days, but the VLF clearly does a better job when $\sigma_\mathcal{T} = 4$ 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 $\sigma_\mathcal{T} = 8$ days and s = 0.15. The distortion of the ZLF first becomes noticeable over intervals when the (N,P,Z) trajectory curvature ($\frac{d^2N}{dt^2}$, $\frac{d^2P}{dt^2}$, $\frac{d^2Z}{dt^2}$ respectively) is high, i.e. at the peaks and troughs (see the fits for $\sigma_\mathcal{T} = 4$ days in Figs. 2.2,2.3). This suggests that intervals in the fine-scale 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 $\sigma_\mathcal{T} = 8$ 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 predator-prey 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 2.4, where the results for $\sigma_\mathcal{T} = 0$, 4 and 8 days and $s = 0.15$ shown in Figure 2.3 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 $\sigma_\mathcal{T} = 4$ days, and spirals towards equilibrium at $\sigma_\mathcal{T} = 8$ days, whilst the VLF maintains stable limit cycles with a small amount of transience at both $\sigma_\mathcal{T} = 4$ and $\sigma_\mathcal{T} = 8$ days.

Comparing Figures 2.2 and 2.3, the increase in measurement noise is seen to increase the deviation of the inferred trajectory from the true trajectory of both fits at $\sigma_\mathcal{T} = 0$ days, whilst it does not appear to significantly alter the trajectories at $\sigma_\mathcal{T} = 4$ 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. 2.2 and 2.3 this does not seem to be the case at $s \leq 0.15$ for $4 \leq \sigma_\mathcal{T} \leq 8$ 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. Optimisation vs. fit-by-eye. Looking at the data for $\sigma_\mathcal{T} = 8$ days in Figures 2.2 and 2.3, one might wonder how the optimiser is able to extract any information at all from such a scattered set. Note, however, that the optimiser considers the model-data discrepancy in all three state variables simultaneously, whilst the eye tends not to do this when examining

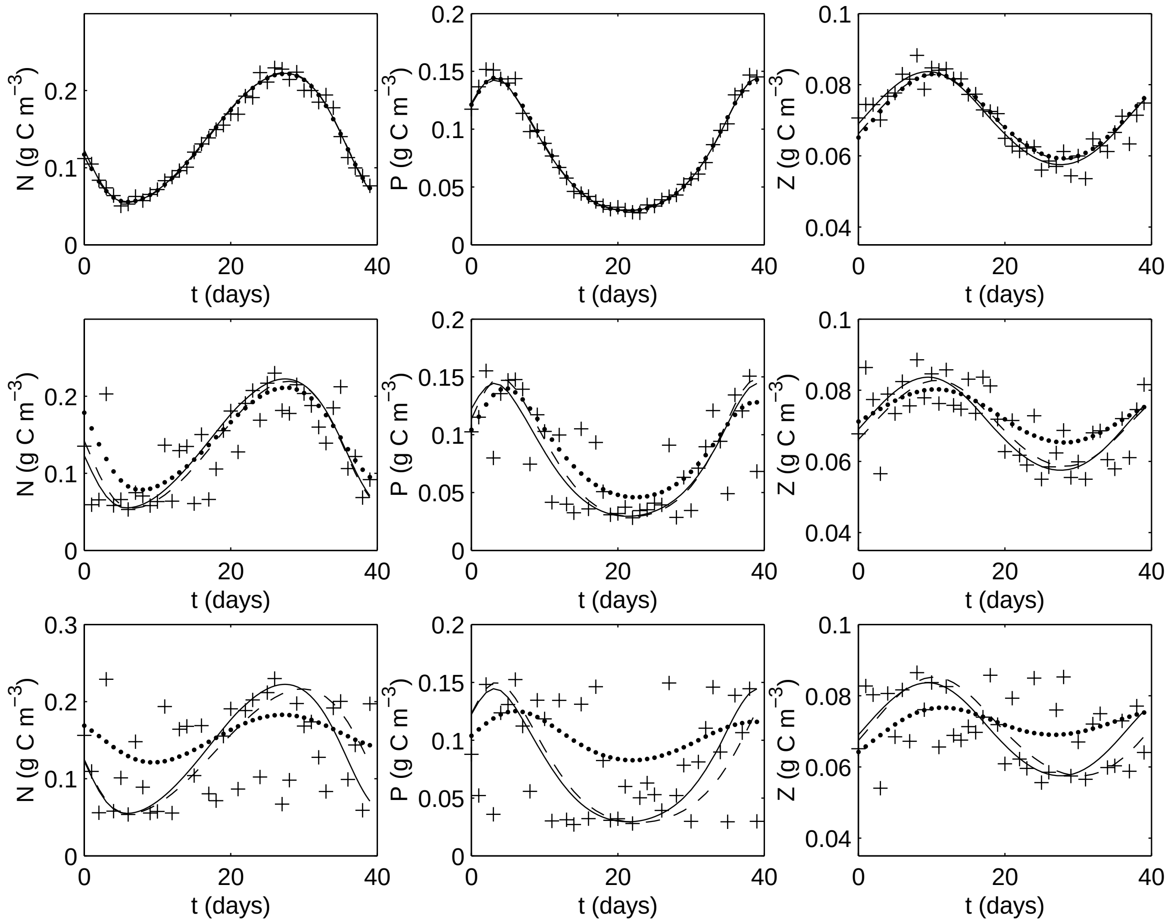

Figure 2.2: Best fit trajectories to data (+) generated by 'true' (Lagrangian) trajectory (solid line) using standard zero lag method (ZLF, dotted) and using the new variable lag method (VLF, dashed). Data were generated with time lag standard deviations $\sigma_\tau = 0$ days (upper), 4 days (middle), 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 equations (2.1–2.3) and Tabel 2.1 for specification of the true model).

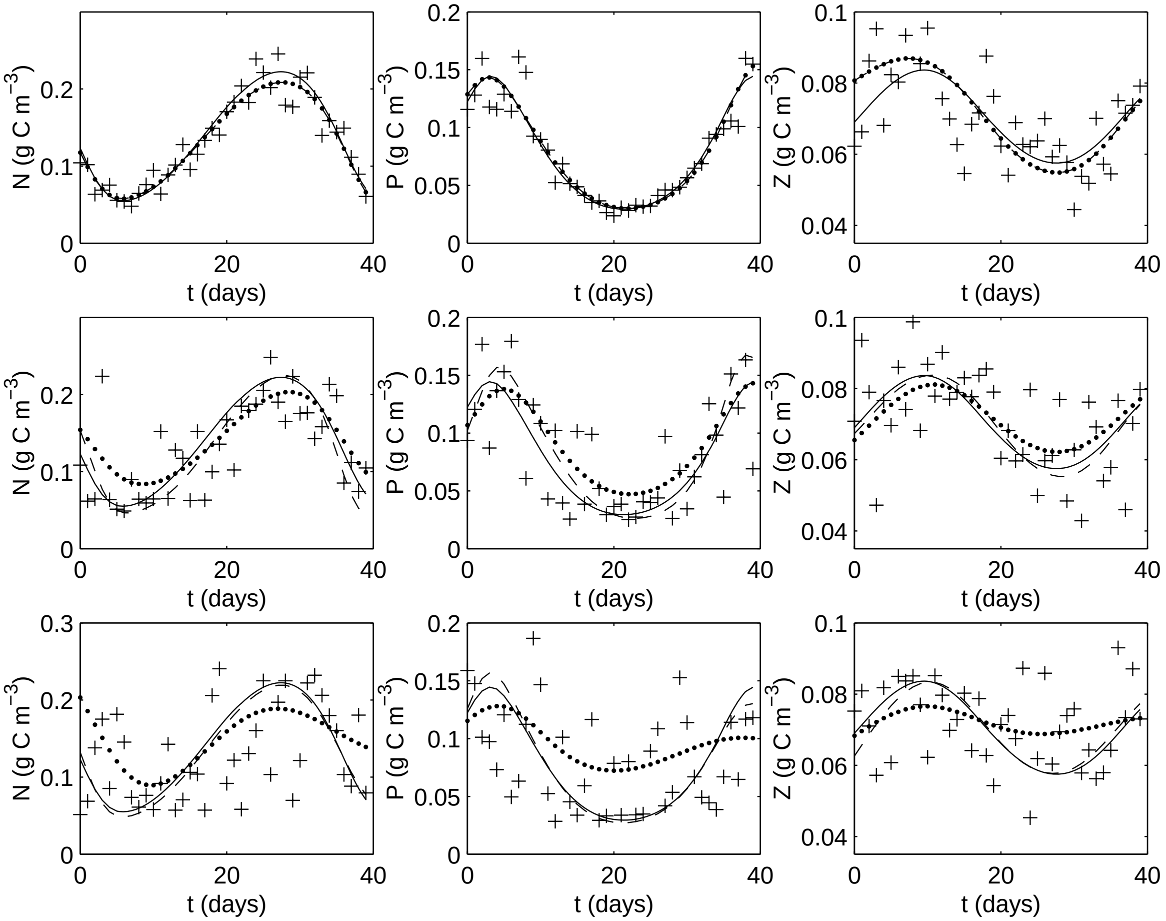

Figure 2.3: As in Figure 2.2 but with 15% measurement noise ($s = 0.15$) imposed on each state variable, and initial guess parameter error of +10% in every parameter.

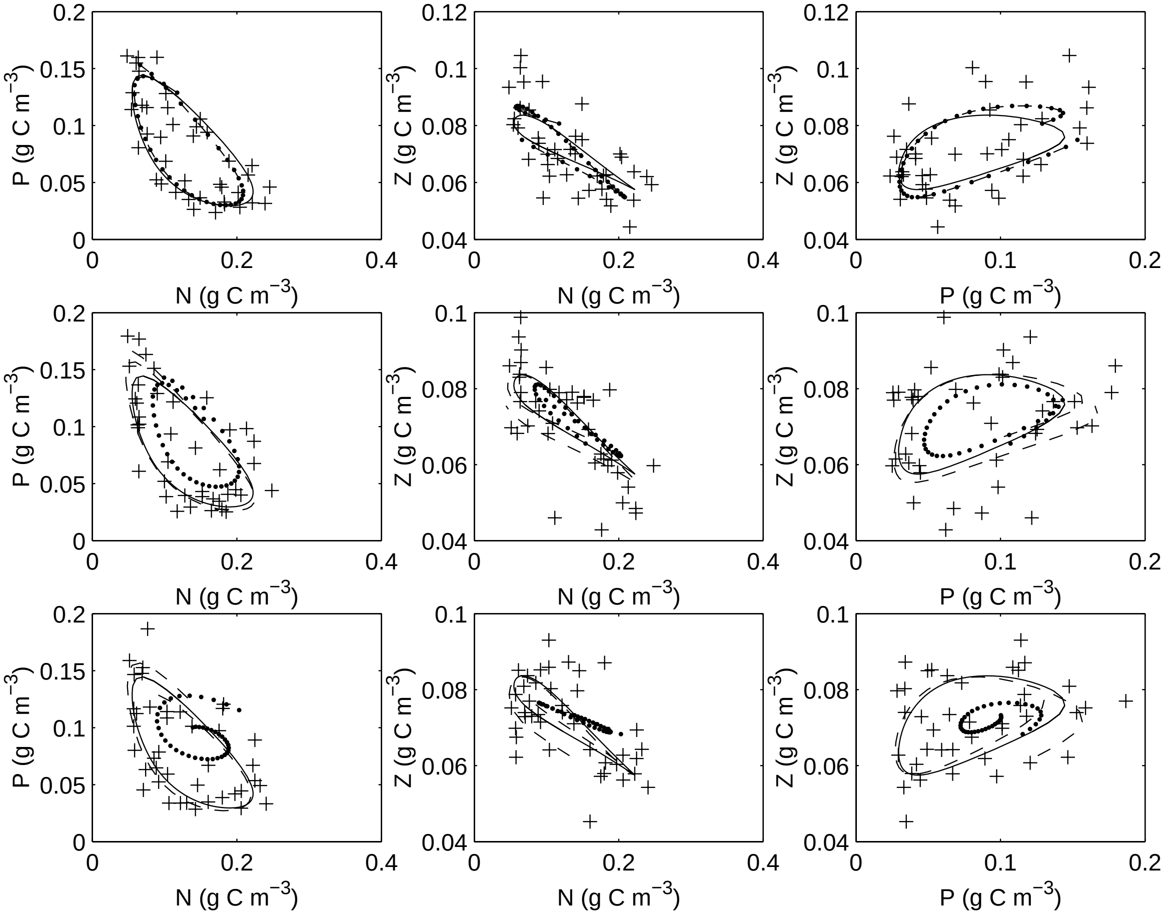

Figure 2.4: Phase space cross-sections showing best fit trajectories to data (+) generated by 'true' Lagrangian biological trajectory (solid line) using standard zero lag method (ZLF, dotted) and using the new variable lag method (VLF, dashed). Data were generated with time lag standard deviations $\sigma_\tau = 0$ days (upper), 4 days (middle), and 8 days (lower) and 15% measurement noise imposed on each variable ($s = 0.15$).

a set of time series (although this is partly achievable by displaying 2D phase space cross-sections as in Fig. 2.4, perhaps using colours to retain some temporal information as in Figure 2.1). Thus, in all of the optimisations, 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). 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 (2.1–2.3)). The VLF has the additional freedom to effectively shift data points forwards and backwards in time (left and right in Figs. 2.2 and 2.3) to an extent roughly proportional to $\sigma_\mathcal{T}$ 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).

2.3.2 Quantitative Skill Assessment

i. Recovery of the Lagrangian biological trajectory. First, the potential benefits of the new technique are assessed 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 $\sigma_\mathcal{T} = 0$, 0.5, 1, 2, 4 and 8 days and $s = 0.05$ and 0.15 (see Figs. 2.5 and 2.6). Here the subscript 'alt' denotes use of the alternative hypothesis to generate the data ($\tau^t \neq 0$), and recall that the correct value of $\sigma_\mathcal{T}$ is used as an *a priori* estimate to explore the maximum potential benefit of the VLF. Also, $\theta^i = \theta^t$ is used 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 fitting finite, noisy data). As $\sigma_\mathcal{T}$ 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 minimise 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 2.5 and 2.6, the maximum mean saving in validation cost achieved by the VLF (at $\sigma_\mathcal{T} = 8$ days) is about 450 units for $s = 0.05$ and about 350 units 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 $\sigma^c_\mathcal{T} \approx 1$ days for $s = 0.05$ and $\sigma^c_\mathcal{T} \approx 4$ 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, the potential dangers of using the new technique are assessed by comparing with the ZLF when time lags are in fact absent in the data set ($\tau^t = 0$). This was done by computing $\Lambda^{(V)}_{(null)}$ over 100 different data sets (each requiring 10^4 iterations) for $\sigma_\mathcal{T} = 0$, 0.5, 1, 2, 4 and 8 days and $s = 0.05$ and 0.15 (see Figs. 2.7,2.8). Here the subscript 'null' denotes data generated on the null hypothesis ($\tau^t = 0$). The means and standard deviations show 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 $\sigma_\mathcal{T}$ is increasingly overestimated.

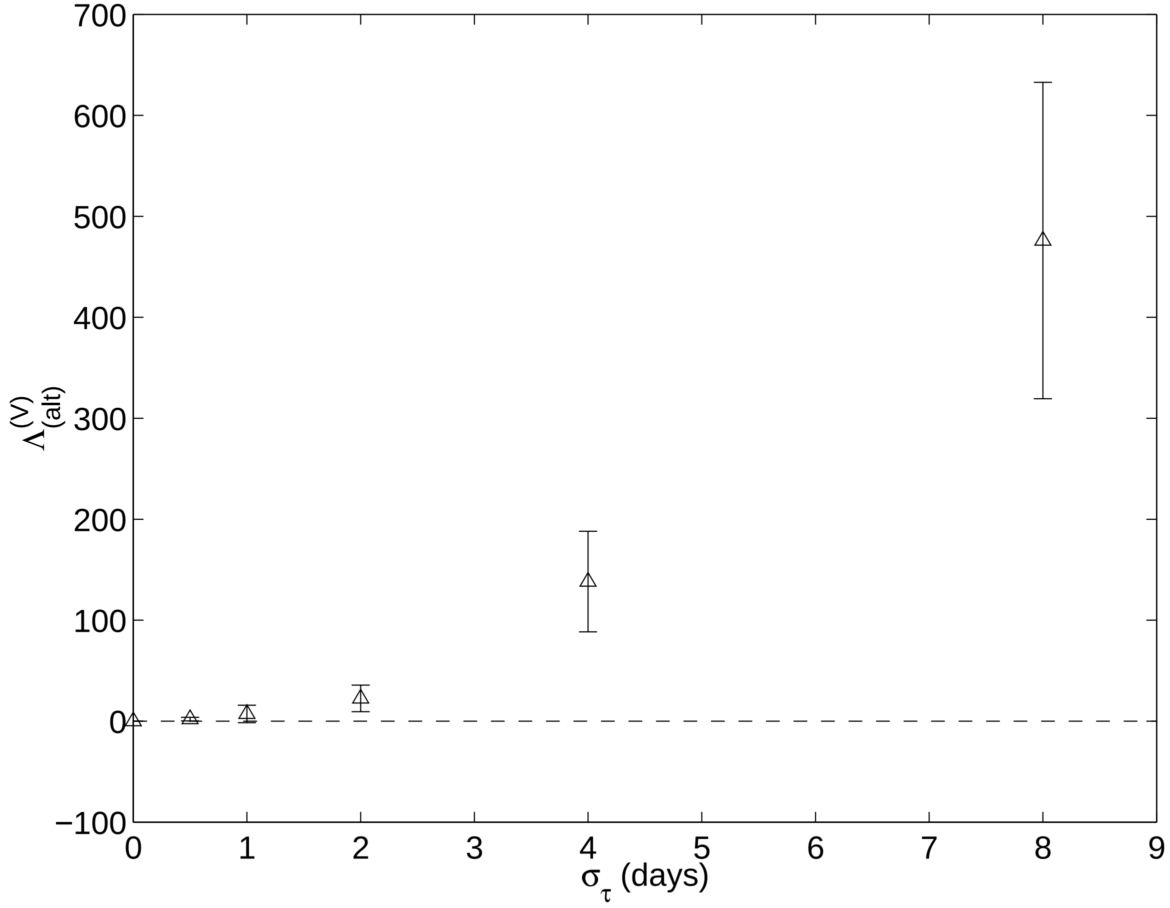

Figure 2.5: Validation cost savings yielded by new technique $\Lambda^{(V)}_{(alt)}$ (triangles) vs. time lag standard deviation σ_τ when time lags in calibrating data set are non-zero (the 'alternative' hypothesis $\tau^t \neq 0$) and 5% measurement noise is imposed on all three state variables ($s = 0.05$). Means (triangles) and standard deviations (error bars) are shown from 5 optimisations over different calibrating data sets for each value of σ_τ.

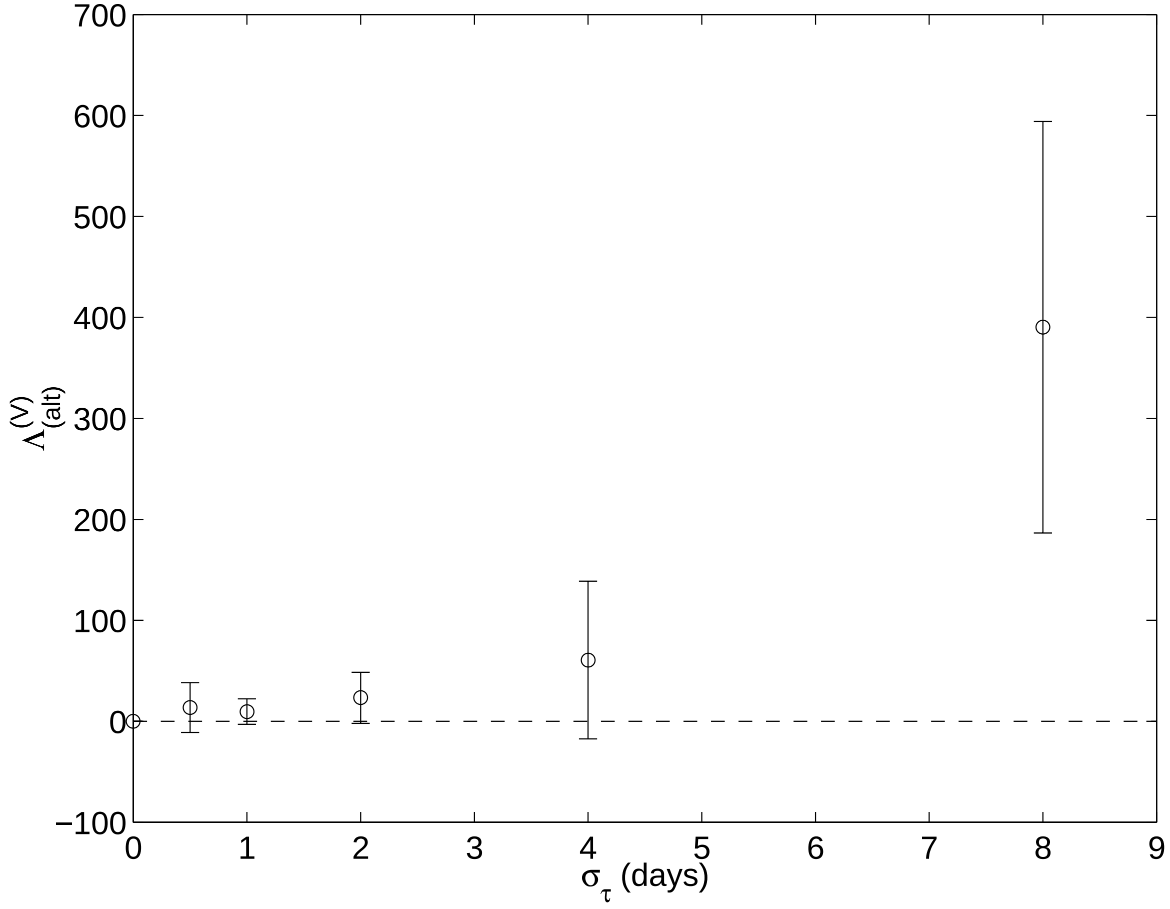

Figure 2.6: As in Figure 2.5 but with 15% measurement noise imposed ($s = 0.15$).

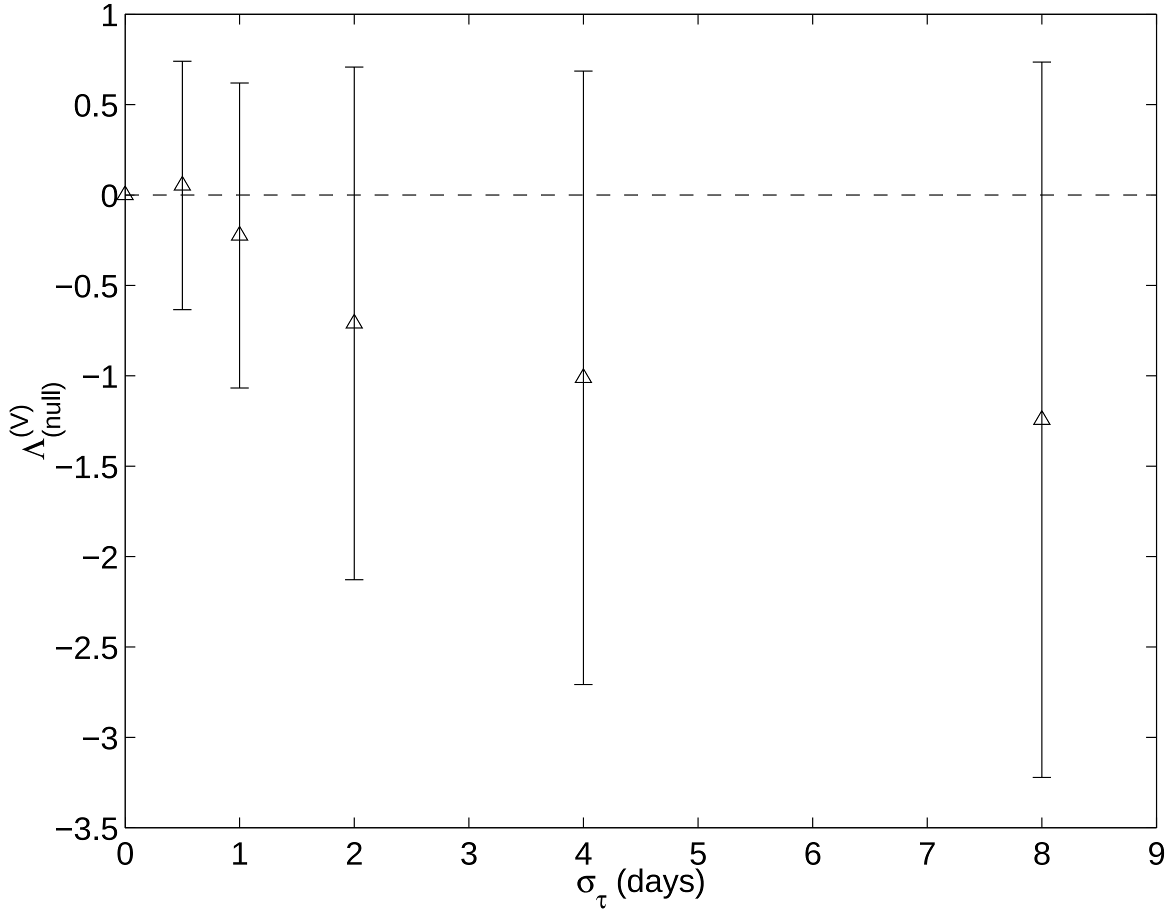

Figure 2.7: Validation cost savings yielded by new technique $\Lambda^{(V)}_{(null)}$ vs. modelled time lag standard deviation σ_τ when time lags in calibrating data set are zero (the null hypothesis $\tau^t = 0$) and 5% measurement noise is imposed on all three state variables ($s = 0.05$). Means (triangles) and standard deviations (error bars) are shown from 100 optimisations over different calibrating data sets for each value of σ_τ.

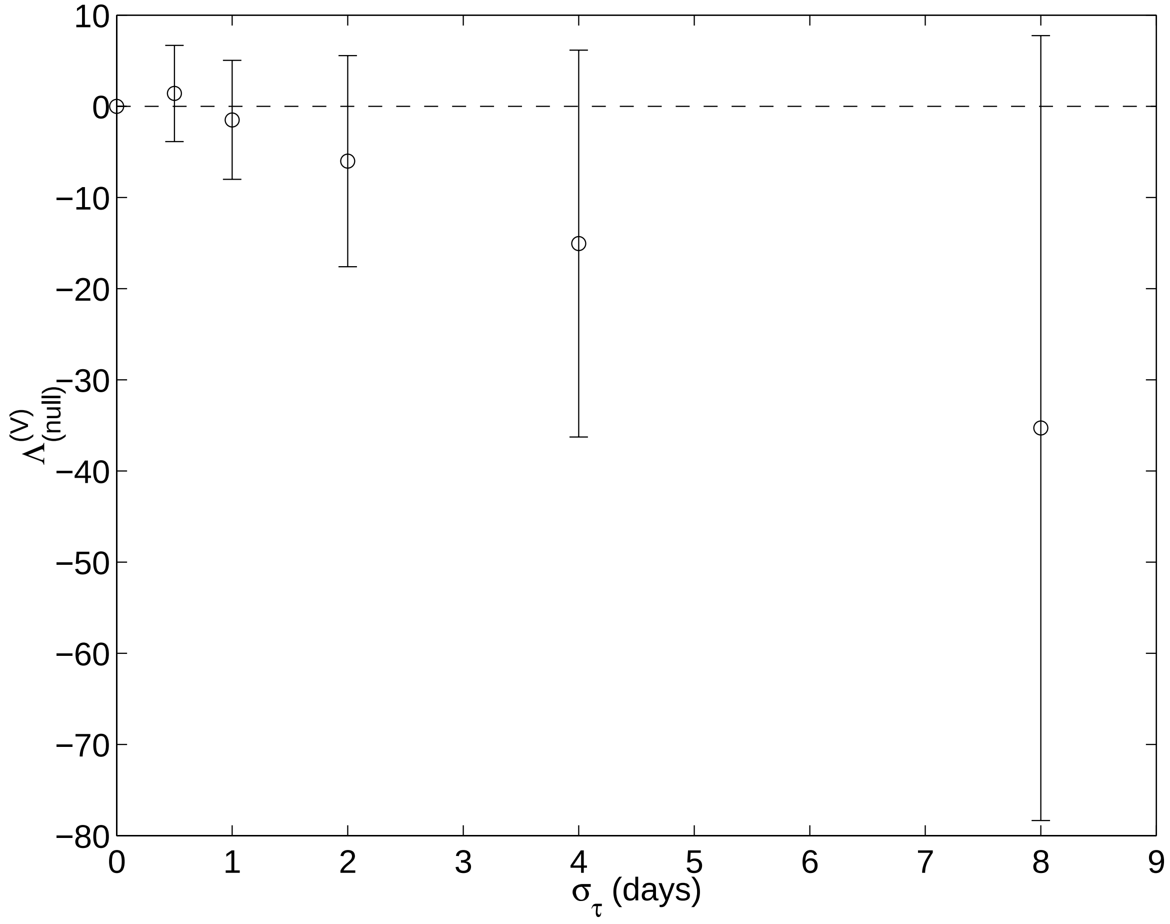

Figure 2.8: As in Figure 2.7 but with 15% measurement noise imposed ($s = 0.15$).

From Figures 2.7 and 2.8, 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 biological 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 ensemble 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.

The potential risks vs. benefits of using a VLF may now be summarised 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 ($\tau^t = 0$) 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 realise this potential, one must accurately estimate σ_τ, which is discussed in section 2.3.3.

Qualitatively similar plots to Figures 2.5 and 2.6 were obtained using the logarithm of the ratio of the optimal Likelihoods $p^{(2)}_{ZLF/VLF}$ (the Likelihood Ratio Test LRT) as a test statistic (see Figs. 2.9,2.10):

$$\Lambda^{(C)} = \max\left(\log p^{(2)}_{VLF}\right) - \max\left(\log p^{(2)}_{ZLF}\right) \tag{2.16}$$

where the superscript (C) denotes a 'calibration' statistic accessible when the true Lagrangian biological trajectory is unknown (as for fitting real data sets). Comparing Figures 2.9 and 2.10 allows rejection of the null hypothesis ($\tau^t = 0$) at similar threshold values of the time lag standard deviation σ^c_τ as those quoted above. Fig. 2.10 also confirms, for high σ_τ, the classical result that the asymptotic distribution of 2× the LRT statistic on the null, given correct model and data error formulations, should be chi-square with K degrees of freedom, where K is the number of extra parameters in the 'alternative' model (in our case the 40 time lags), hence $\Lambda^{(C)}$ should have a mean value of $K/2 = 20$ (Stuart et al., 1999).

ii. Bias and variance in structural parameter estimates. Bias and variance are estimated by performing 100 optimisations using $\sigma_\tau = 8$ days and $s = 0.15$ to generate and fit different data sets (see Table 2.2), initialising the optimisation with $\theta^i = \theta^t$ each time. This latter choice may lead to underestimation of optimal parameter variances, if the cooling is insufficiently gradual to prevent search localisation near θ^i. 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 localisation in θ-space to affect the VLF and ZLF equally.

First, note from Table 2.2 that the VLF yields 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 biological trajectory

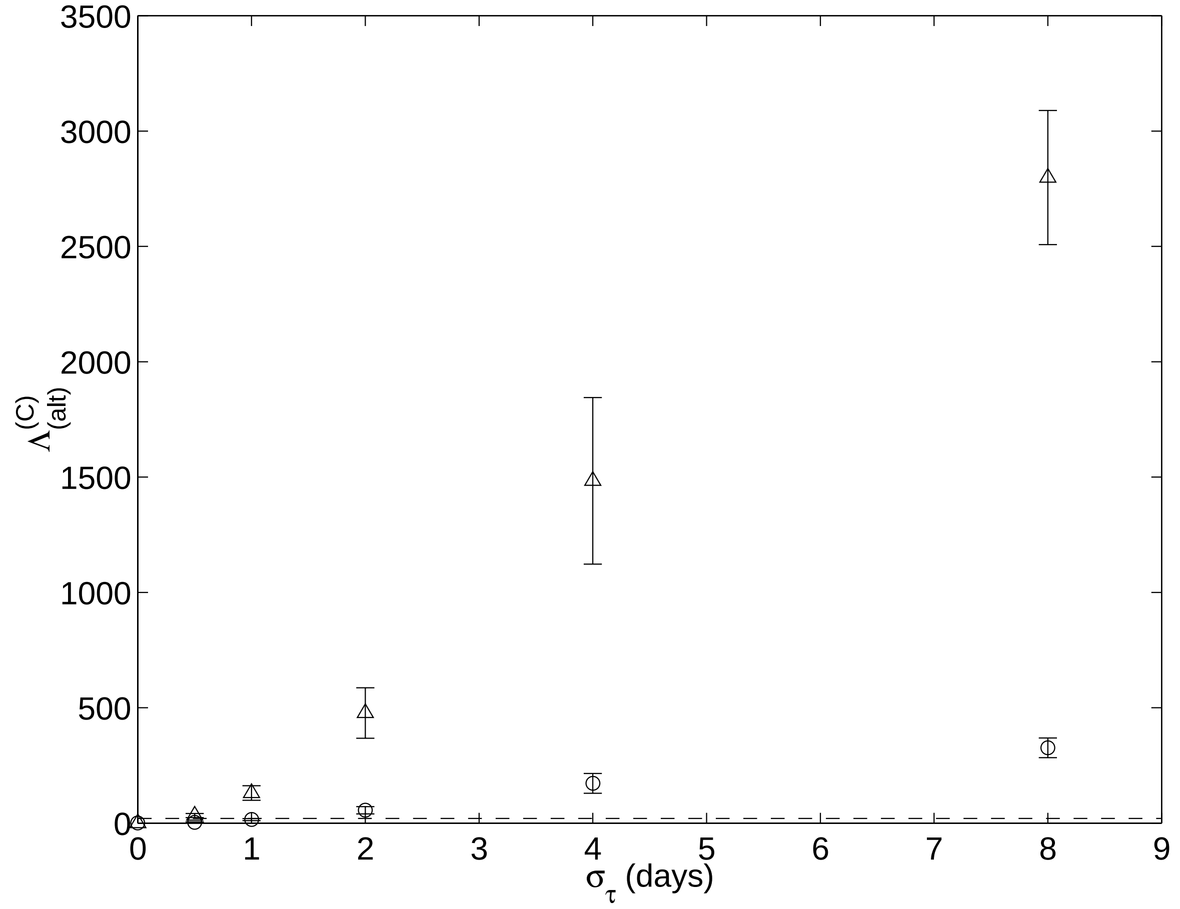

Figure 2.9: Calibration cost savings yielded by new technique $\Lambda^{(C)}_{(alt)}$ vs. modelled = true time lag standard deviation σ_τ when time lags in calibrating data set are non-zero (the 'alternative' hypothesis $\tau^t \neq 0$). Means (triangles/circles) and standard deviations (error bars) are shown for s = 0.05/0.15. The dashed line marks the theoretical expected value of $\Lambda^{(C)}_{(null)}$ for $\sigma_\tau \to \infty$ (see Figure 2.10)

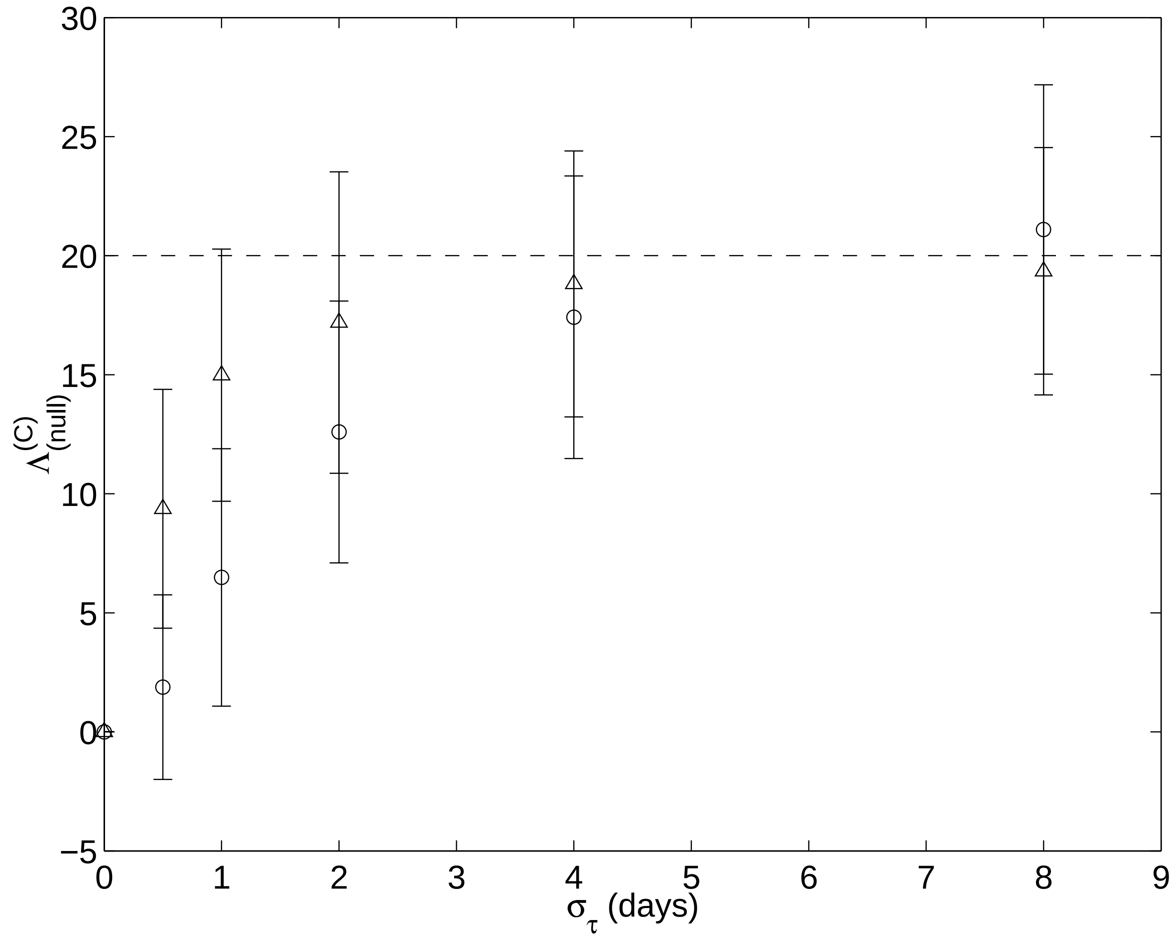

Figure 2.10: Calibration cost savings yielded by new technique $\Lambda^{(C)}_{(null)}$ vs. modelled time lag standard deviation σ_τ when time lags in calibrating data set are zero (the null hypothesis $\tau^t = 0$). Means (triangles/circles) and standard deviations (error bars) are shown for $s = 0.05/0.15$. The dashed line marks the theoretical expected value of $\Lambda^{(C)}_{(null)}$ for $\sigma_\tau \to \infty$

Table 2.2: 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.

Parameter	Sensitivity	Estimator Bias (%)			Estimator Stdev (%)		
		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

is sensitive (‘high’ and ‘medium’ sensitivity parameters) are more accurate than those of the ZLF, as one might have expected from Figure 2.6. The lower variance of the VLF estimates may be a result of the fact that the true trajectory lies within the (θ, τ) search space, whilst the ZLF effectively fits to the convolution of the true trajectory with the time lag distribution, which is a smaller signal relative to the measurement noise, and may not lie exactly in the θ search space.

Table 2.2 can be used to infer the parameter biases that are likely most responsible for persistent distortion effects. Highlighted in bold are the significantly biased parameter estimates, defined as 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 2.3 and 2.4 show true values and biases/variances for a few interesting ‘derived parameters’ i.e. functions of the structural parameters in Tabel 2.1, again for $\sigma_\tau = 8$ days, $s = 0.15$. First, the bias and variance in oscillation frequency ω are estimated by calculating the peak-to-trough separations in the true and optimal zooplankton trajectories (zooplankton showing the most symmetrical oscillation — see Figs. 2.2 and 2.3) over the 100 optimisations. Although the bias in the VLF estimated frequency is very low (less than 1%), the standard deviation is significant ($\approx$ 15 %), and causes significant increase in validation cost as the

Table 2.3: 'Derived parameters' (functions of the parameters in Table 2.1) and true values, where 'mean' denotes average over the sampling interval (roughly one cycle).

Derived Parameter	Symbol	'True' value
Oscillation frequency	ω	0.18 rad s^{-1}
Mean gross primary production	$\overline{GPP}$	0.05 g C m^{-3} day^{-1}
Mean net nutrient export rate	$\overline{NNE}$	-0.02 g C m^{-3} day^{-1}

Table 2.4: 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.

Parameter	Estimator Bias (%)			Estimator Stdev (%)		
	ZLF	VLF	VLF(no S)	ZLF	VLF	VLF(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

estimated trajectory drifts out of phase with the true Lagrangian biological trajectory (see e.g. Fig. 2.2, lower). The estimated bias and variance in ZLF oscillation frequency is very poor, as expected since many of the ZLF trajectories are decaying to equilibrium (see Figs. 2.2, 2.3 and 2.4).

Second, estimates of mean (over sampling interval) gross primary production ($\overline{GPP}$), defined by the uptake term in equation (2.2), are considered. 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 (2.1) and (2.2), was poorly estimated by both methods — probably because of its small absolute true value of -0.02 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 2.2).

Finally, the importance of using the time lag drift penalty S (see section 2.2.6) is illustrated by comparing the VLF parameter estimates with those obtained by maximising M_{VLF} in equation (2.9) which has no lag drift penalty S ('VLF(no S)') for $\sigma_\tau = 8$ days, $s = 0.15$ (see Tables 2.2 and 2.4). 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 similar validation cost is obtained 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.

2.3.3 Estimating time lag variance.

Although the true measurement error coefficient s^t and time lag variance σ_τ^t were assumed known in this twin test study, in practical applications this will not be the case. Knowledge of the measuring apparatus/method should give reasonable constraints on s^t, but σ_τ^t will be less well known *a priori*. It would probably be best to fix s *a priori* so that σ_τ can be adjusted to data with maximal constraint. If both s and σ_τ are adjusted, estimates $\hat{s}$ and $\hat{\sigma}_\tau$ are liable to covary and may even be jointly degenerate (individually unconstrainable) if the trajectory y is largely linear in time. With s fixed, there should be no problem fitting σ_τ as long as it is not allowed to be so large that timescale constraints are entirely relaxed, such that timescale distortion via $\hat{\theta}$ is not detected by the lag drift penalty S. Also, a large σ_τ may make the fit indifferent to time lags shifting model output to different sides of peaks/troughs in y. To some extent this is unavoidable near peaks/troughs of the unlagged trajectory, but the implied degeneracy in the fit of individual time lags may cause problems if it becomes excessive (particularly if successive lags covary — see section 2.3.4*iii*). Therefore, it is probably wise to make a 'conservative' choice of allowed range for σ_τ, or of parameters for a Bayesian prior on σ_τ. It may in fact be preferable to fit an empirical model which is by-design capable of fitting the Lagrangian data as a first step for deconvolving the data and/or estimating time lag variance, after which a mechanistic model may be fitted (see section 2.3.5).

Note also that Figures 2.9 and 2.10 show that $\Lambda^{(C)}$ is a *powerful* statistic, in the sense that it is very sensitive to σ_τ^t, hence it has a high probability of correctly rejecting the null when used as a hypothesis test statistic (Stuart et al., 1999). It might therefore be used as follows to estimate time lag variance without treating σ_τ as an adjustable parameter. First, the null hypothesis is as $\sigma_\tau^t = 0$ and ZLF/VLF model fits are performed using a suitably conservative modelled value of σ_τ, say $\sigma_\tau^{(1)} = 1$ day. Using $\hat{\theta}_{ZLF}$ from the real data fit to generate data sets, the null distribution of $\Lambda^{(C)}$ for $\sigma_\tau = \sigma_\tau^{(1)}$ may be computed and compared with $\Lambda^{(C)}$ from the real data fit. If the latter lies in the 5% upper tail of the null distribution, the null is rejected. Then a new null of, for example, $\sigma_\tau^t = 1$ day is proposed, and $\hat{\theta}_{VLF}$ from the real data fit is used to generate the new null distribution of $\Lambda^{(C)}$ for $\sigma_\tau = \sigma_\tau^{(2)} = 2$ days, and again this is compared with $\Lambda^{(C)}$ from a real data fit. The process is repeated until the null cannot be rejected, at which point the null value of σ_τ^t is taken as our estimate $\hat{\sigma}_\tau$, and the last value of $\hat{\theta}_{VLF}$ is our best estimate of the Lagrangian biological parameters. A conservative shortcut to computing the null distribution might be to assume the classical chi-squared distribution for $\sigma_\tau \to \infty$, this being the expected result when overestimated lag variance is only improving the fit to measurement errors.

Alternatively, if sufficient 'replicated' data are available i.e. multiple samples within the same fluid mass, the nonlinear optimisation methods detailed here might be extended to consistently fit both the measurement error s and time lag variability σ_τ simultaneously (Seber and Wild, 2003).

2.3.4 Robustness of the method

i. Different data sets/initial parameter guesses. In Figure 2.11 the robustness of the results in Figure 2.2 ($\sigma_\tau = 8$ days, $s = 0.05$) is illustrated with different realisations of the data set

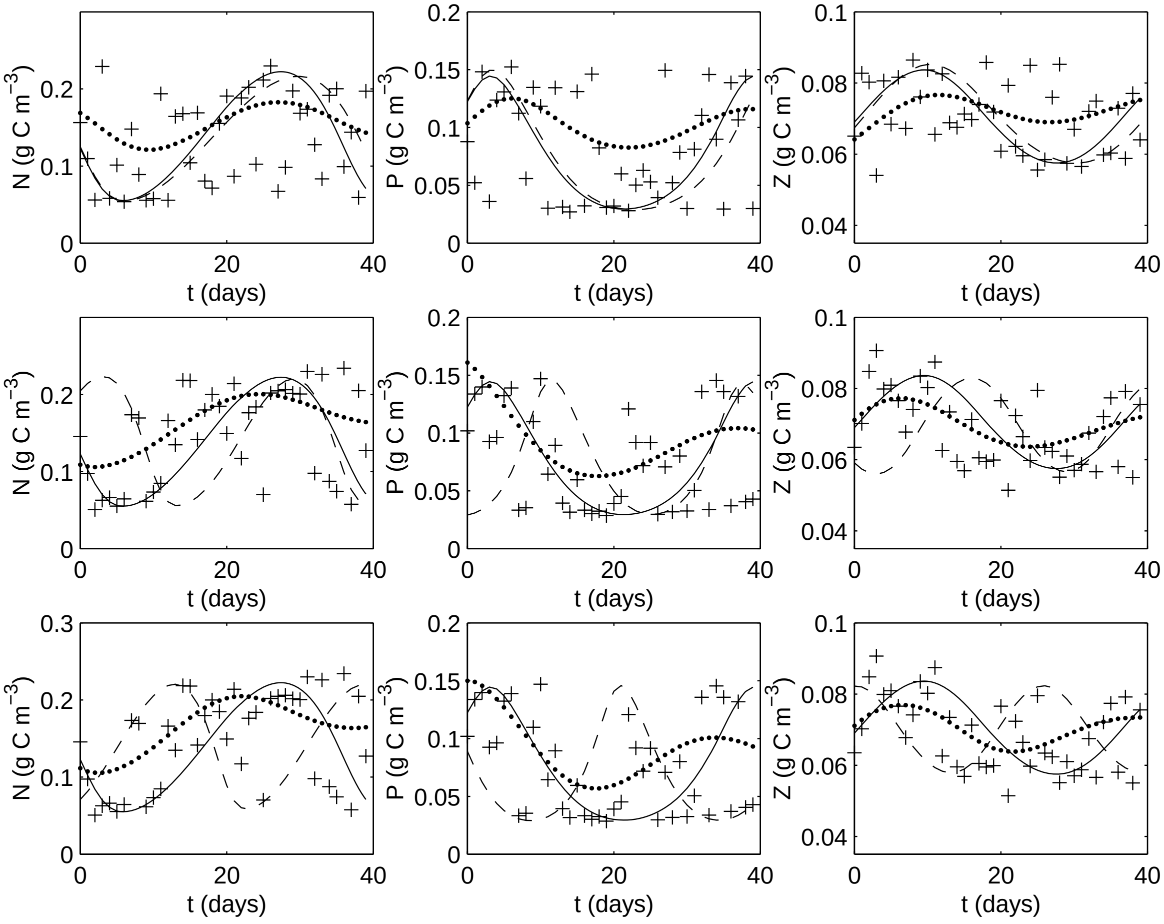

Figure 2.11: Variance associated with poor convergence in the variable lag fit for $\sigma_\tau = 8$ days, $s = 0.05$. Upper panel reproduces Figure 2.2 lower. Middle and lower panels show examples of poor convergence in estimated frequency and phase of the oscillation respectively.

and different initial guesses $\theta^i = 0.9$ or $1.1 \times \theta^t$. Examples are shown for which errors in the optimal frequency and phase of oscillation resulted in the worst fit performance by the VLF, using 10^5 iterations of the optimiser for convergence of the calibration cost. They show 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 variance is mainly responsible for the large error bars in Figures 2.5 and 2.6 for high σ_τ.

It may be possible to diagnose a frequency/phase deviation in a real data fit by comparing 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, robustness issues might be addressed by strengthening prior constraints on θ (e.g. from small-scale laboratory studies), or averaging estimated parameters over an ensemble of optimisations, using different data sets generated by subsampling replicated data or by bootstrap methods (e.g. see Schartau and Oschlies, 2003).

Given accurate estimation of ω^t, τ^t was well-estimated on the whole, except for some occasional 'slip' in the estimated time lags away from their true values (see Fig. 2.12 upper). This positive result may depend to some extent on the accurate specification of the measurement noise s and time lag variability σ_τ prior to fitting. In Figure 2.12 the effect on $\hat{\tau}$ of VLF overestimation of ω^t is evident. Here, the optimal time lags are accurate only over the

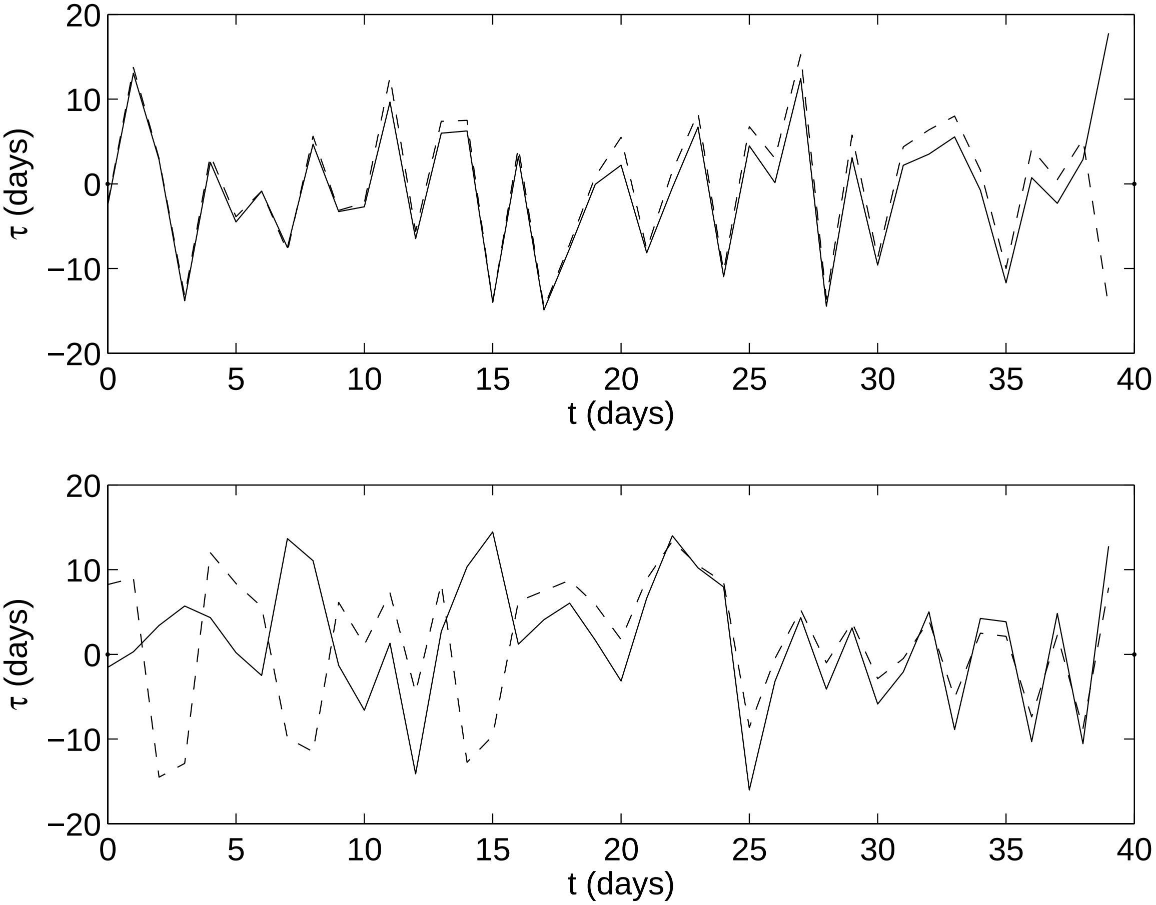

Figure 2.12: Best fit time lags (dashed) compared to true values (solid) for ($\sigma_\tau = 8$ days, $s = 0.05$), for optimisations with good convergence (upper) and poor convergence (lower) in estimated frequency and phase of oscillation (corresponding to fits shown in Figure 2.11 upper and middle panels).

second half of the sampling interval, when the fitted oscillation is roughly in phase with the true cycle (Fig. 2.11 middle). Looking further back in time the estimated and true time lags become increasingly divergent, reflecting the overestimation of ω^t.

ii. Different Lagrangian 'truth'/models. The results should be largely robust to different Lagrangian models provided: 1) The fitted model formulation is close to truth (it was perfect in this study), otherwise the fitted time lag parameters are liable to compensate for inadequacies in the model formulation; 2) the Lagrangian biological trajectory is strongly nonlinear in time, so that the effect of time lags on the time series is not equivalent to uncorrelated measurement noise (which would also help to constrain σ_τ if this were fitted — see section 2.3.3). In addition, models with more state variables all sharing the same time lags should allow better constraint on τ, so long as simultaneous samples in all state variables are provided. The fact that many state variables are in theory available to jointly constrain dynamical phase noise is a potential strength for biological modelling, with the strong proviso that these need to be somehow measured simultaneously.

iii. Different spatial variability and sampling conditions. Consider relaxing the assumption that true time lags between samples are uncorrelated, as may be necessary to apply the technique to real cruise survey data. As long as the decorrelation timescale is shorter than the total sampling interval, these correlations will not be penalised by the lag drift penalty S in (2.13). If the correlation is underestimated in the time lag prior (2.10), multiple correlated lags will be overpenalised in the model fit, and the biological parameters θ may be distorted to compensate. Similarly, overestimating the correlation would allow trends in the time lags to compensate for errors in the ecosystem dynamical tendency (wrong θ). The same considerations apply to possible correlations in the 'measurement error' representing remaining model error and instrumental noise. Therefore, correlations should either be estimated in the model fit (possible with 'replicated' data — see Seber and Wild, 2003) or else avoided by subsampling or averaging the data (which may also allow ensembles of data sets to be generated).

2.3.5 Alternative methods.

The VLF may be regarded as an attempt to do two things simultaneously: first to reverse the smearing or 'convolving' effects of phase noise — i.e. to 'deconvolve' the data set by a 'parametric' (model-fitting) or 'Bayesian' method; second to fit a dynamical model to the deconvolved data by nonlinear regression. It might be better to use a strictly empirical interpolation model (e.g. splines) for the first step, so that deconvolution and time lag variance estimation may be carried out 'unfettered' by mechanistic assumptions, then fit a mechanistic model to a set of deconvolved Lagrangian data, or fit the original data with the time lag variance estimate provided by the empirical model fit. An obvious alternative is to use nonparametric deconvolution methods to remove the suspected phase noise in Fourier space (e.g.: divide the frequency transform of the data by the frequency transform of the time lag distribution, and inverse transform the result to obtain a deconvolved 'Lagrangian' time series). Note, however, that parametric methods generally tend to work better with short time series (Laws, 1997), as are often obtained from cruise surveys. Also, the model-fitting approach may allow estimation of the time lag variance as a free parameter (see section 2.3.3).

2.4 Conclusions

A simulation study was performed 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. The experiments showed that:

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 non-Lagrangian data set effectively samples from a true Lagrangian biological trajectory smoothed out by (or convolved with) the distribution of random time lags. The smoothing is most apparent 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 Lagrangian dynamics are misfitted as exponential decays.

Although these twin tests used a free predator-prey oscillating system, similar smoothing out and associated parameter biases may occur in the more generic case of fitting to seasonal data, where Lagrangian phytoplankton blooms and subsequent troughs due to grazing may be underestimated as a result. Consequently, gross and net primary production rates during blooms may be significantly underestimated. Although the impact on estimates of annual averages of these rates cannot be predicted from this study, it is clear that the effects of unresolved time lags or 'phase noise' on standard model fits may be serious, and warrant further investigation.

2) A new 'variable lag' model fitting technique was shown to perform 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 time lag variance 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 (in comparison with the Lagrangian 'truth') 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 was also reduced, although a certain amount of variance in oscillation frequency could not be avoided. Although the time lag and measurement error statistics were assumed to be known in the twin tests, it is possible that these may be estimated in practical applications from knowledge of the measuring methods and by fitting time lag (co)variance parameters, possibly employing an empirical (deconvolving) model as a first step. Given sufficiently frequent sampling spanning nearby water masses (a 'replicated' data set), the variable lag fit might be extended to fit both time lag and measurement error variances.

In summary, if unresolved spatial variability in plankton ecosystems can be modelled as 'phase noise', the optimal 'Lagrangian' parameter sets obtained by a new 'variable lag fit' method may reveal a robust mean dynamics on a smaller scale of spatial averaging than standard model fits, without explicitly resolving those scales. The method may help to improve estimation of the biological parameters for high-resolution plankton models, and allow accounting for sub grid-scale variability in coarse-resolution plankton models by convolving a fitted Lagrangian biological trajectory with a distribution of time lags.

Chapter 3

Bulk Model Fitting and Testing

Submitted article: Wallhead, P.J., Martin, A.P., Srokosz M.A. Skill Assessment by Cross-Validation and Monte Carlo Simulation: An Application to Georges Bank Plankton Models. J. Mar. Sys.

3.1 Introduction

In this chapter, highly-aggregated 0D plankton models are fitted to real data, allowing liberal tuning of biological parameters to account for the effects of unresolved spatial variability and biological diversity/adaptation. Such 'weak prior' bulk plankton models were also fitted to real data in Evans (1999), Fasham and Evans (1995), Hurtt and Armstrong (1996, 1998), Kuroda and Kishi (2004), Schartau et al. (2001) and Spitz et al. (1998, 2001) (see Gregg, 2008, for a broader summary of plankton model fitting studies). In all these studies, model predictions were tested or 'skill-assessed' solely in terms of how well they fitted the calibration data. This is, ultimately, an inadequate criterion to test ecological hypotheses and guide plankton model development. The weak prior constraint on plankton model parameters, combined with the sparseness and noisiness of plankton data, means that improvements in data fit may be partly due to improved fit to data noise, rather than convergence on truth. Statistical hypothesis testing has been applied to determine if improvements in data fit are significant beyond expected improvements in noise-fitting due to greater model freedom (see e.g. Stock et al., 2005, and Chapter 2). However, even if data noise is not fitted, true trends in the data may yet be fitted for the wrong dynamical reasons, especially with a weakly constrained plankton model. Much stronger evidence of mechanistic veracity is the ability of a fitted model to predict true variability beyond the conditions in which it was fitted — in other words, to *extrapolate.*

This chapter investigates methods of testing extrapolative 'skill', suitably defined, which also account for noise fitting. New skill metrics are used to compare a weak prior bulk plankton model to a strictly empirical model, with the aim of testing the weak prior, coarse-resolution approach to plankton modelling in general. The skill assessment methodology should however be applicable to a broad range of plankton models, given sufficient computational power.

The data are a set of chlorophyll samples and associated 'forcing' data for 1997–1999 derived from the GLOBEC database. The empirical model is a simple temporal interpolation of the chlorophyll data pooled from all sampled years, termed 'inductive' because it assumes that the variability in future years will be the mean of observed variability during sampled years in the past. The inductive model serves as a baseline of model skill. The mechanistic 'process' model is a simple bulk plankton model forced by observed nutrient and mesozooplankton variability for the fitted and predicted years.

The term 'skill' generally refers to predictive accuracy. A precise definition is problem-dependent: it depends on the data available for model-fitting, the variables we want to predict and their relative importance, and the range of fixed 'model inputs' (information supplied to the model and not fitted to the data, such as forcing parameter values or spatio-temporal coordinates) we want to predict for. Also, the definition of 'accuracy' depends on the relative importance assigned to prediction 'bias' (expected deviation of the prediction from truth over repeat experiments) and prediction 'variance' (variance in the prediction over repeat experiments)[1]. Often, accuracy is conceived in terms of the 'mean-square error' between predictions and truth (this being a sum of (prediction bias) squared and prediction variance) but alternative skill definitions might focus on, for example, mean absolute error or mean absolute percentage error. Consequently, the choice of 'most skillful' model will depend critically on the details of the predictive task.

Therefore, skill assessment of marine ecosystem models should start by specifying the predictive objectives as precisely as possible. Ideally, these are defined mathematically in a 'skill function' which compares predictions to true values. The most skillful prediction maximises this 'skill function'. It will not be possible to directly evaluate this skill function, since it will involve the true values of the quantities we wish to predict, possibly under different conditions to those for which the model was calibrated. It is therefore necessary to estimate the skill — or more realistically, the relative skill — of different model predictions using appropriate 'skill metrics' which approximate the skill function and *can* be evaluated with available data.

Skill functions and metrics should not be confused with the 'cost functions' used in calibration to define model-data misfit, though there may be formal similarities between them, and the choice of skill function may inform the choices of skill metrics and calibration cost functions. Skill functions and metrics express objectives and estimates of how well the model meets them. Cost functions only detail the method of calibration, typically measuring how well the model fits the calibration data, which is not the final objective. The final objective is to predict true fields, and fitting the data is simply a means to this end.

Nevertheless, the first question asked by a modeller is usually: 'Can the model fit the calibration data?'. Failure to do so may indicate model *underfit* — bias in the model predictions due to inadequacy in the model formulation (assuming a sufficiently rigorous fitting method and sufficiently broad allowed parameter ranges). Underfit is an acute problem in marine ecosystem modelling, where no firm mechanistic theory (cf. the Navier-Stokes equations in fluid dynamics) offers guidance in deriving biological dynamics, and models persistently underfit the data despite apparently large numbers of adjustable parameters[2]. However: prediction bias does not only occur at the calibration data ('underfit' or 'intra-sample bias' in the statistical modelling jargon). Prediction bias also occurs at times and/or locations beyond

[1] Note that the classical statistical definitions of bias, variance and mean square error — with respect to truth and an infinite ensemble — are used (Laws, 1997; Stuart et al., 1999).

[2] This is a general interpretation on the part of the author. Too few studies explicitly acknowledge the full extent of underfit; many do not adequately expose model deficiencies by graphical model-data comparison.

the calibration data ('extra-sample bias'). Extra-sample bias may well make a larger contribution to the skill function, but it is not measured at all by the calibration misfit metrics often used in marine ecosystem modelling.

If and when the model does fit the data, the modeller might then ask: 'Is the model *overfitted*?'. When a model has too much freedom to fit to limited data, the model predictions may covary strongly with the noise in the data, resulting in a loss of predictive accuracy. If this covariance is excessive (a subjective judgement) then we may say that the model is 'overfitted'. A clear sign of overfit is if the variance in the misfit between model predictions and some of the data is less than the variance of the observational error. If the difference between the overfitted data and the corresponding fitted model output is used naïvely as a skill metric, the covariance will bias the skill estimates in favour of more complex models with more degrees of freedom. There are metrics based on calibration misfit which attempt to quantify and correct for this effect, such as the Akaike Information Criterion (Akaike, 1973). However, they generally presuppose that the model can fit all the data reasonably well (Burnham and Anderson, 1998), which is unlikely to be the case in marine ecosystem modelling. Moreover, the covariance between model and noise is likely to be strongest outside the calibration data set. Again, this 'extra-sample variance' cannot be accounted for by skill metrics based on calibration misfit, yet it is likely to be important when the model must significantly *generalise* (interpolate between or extrapolate beyond the data) for some of the model input values required by the skill function.

Two general methods are applied in this chapter in order to produce skill metrics which measure both intra- and extra-sample bias and variance. The first is 'simulation' of the data set, whereby surrogate repeat samples are produced artificially at each sampling time, using a statistical simulation model for the data. Simulated ensembles are then used to test different predictive models, in a similar manner to a replicated 'sibling experiment' (Evans, 1999) or an 'Observational System Simulation Experiment' (McGillicuddy et al., 2001). The data simulation model is chosen to accurately simulate repeat samplings of the original data, and is itself fitted to these data. The second method is 'division' of the full data set into calibration (active) and validation (passive) subsets of years and associated model inputs (cf. Hemmings et al., 2004; Friedrichs et al., 2007; Smith et al., submitted). These two methods allow the use of skill metrics based on replicated tests for different divisions of the sampled years. The metrics are used to gauge the relative ability of the models to make accurate, extrapolative predictions of the chlorophyll mean field in future years (cf. application of this method to the toy problem in Appendix 1B).

The chapter is structured as follows. Section 3.2 describes the skill function, skill metrics, data, model formulations and calibration methods. Results are presented in section 3.3 and a Discussion follows in section 3.4. Conclusions are drawn in section 3.5.

3.2 Methods

3.2.1 Skill Function

The objective of this study is to identify the bulk model formulation which best predicts the mean near-surface chlorophyll concentration on Georges Bank for January to June in future years (attention is restricted to the first half of the year in lieu of data for the second half of any year). The 'best' model minimises predictive mean-square error, integrated over the period $T_{int} = 180$ days, given the available data and model inputs for the sampled years 1997–1999, and model inputs for the unsampled 'target' years in the future. Equal importance is assigned to all A_T target years. Hence the following skill function is to be minimised:

$$Skill \equiv \frac{-1}{A_T T_{int}} \sum_{i=1}^{A_T} \int_0^{T_{int}} E[(Y^t_{T_i}(t) - Y^p_{T_i}(D_S, \hat{\kappa}_S, \hat{\lambda}_{T_i}, t))^2]\mathrm{d}t \tag{3.1}$$

where $Y^t_{T_i}(t)$ is the true spatial mean chlorophyll concentration for target year T_i at time t, and $Y^p_{T_i}(D_S, \hat{\kappa}_S, \hat{\lambda}_{T_i}, t)$ is the corresponding prediction, written here as a function of the assimilated data set D_S (constraining some finite adjustable parameter set a), the assimilation model inputs $\hat{\kappa}_S$, the (interannually-variable) target year model inputs $\hat{\lambda}_{T_i}$, and time t. Here, the hats recall the fact that the fixed model inputs $\hat{\kappa}_S$ and $\hat{\lambda}_{T_i}$ generally include uncertain estimates, which therefore vary over an 'ensemble' of 'repeat experiments'. The assimilation model inputs $\hat{\kappa}_S = \{V_a, \hat{\eta}, \hat{\lambda}_S, t_S\}$ include: prior distribution or allowed range parameters for the adjustable model parameters (V_a), interannually-constant model parameter estimates ($\hat{\eta}$), interannually-variable model inputs for the assimilation years ($\hat{\lambda}_S$), and highly certain temporal coordinates for the samples (t_S).

The expectation E is an average over a conceptual infinite ensemble of repeat experiments, representing an idealisation of how skill might be evaluated in absence of any practical constraints. A single 'repeat experiment' involves the same truth Y^t (hence 'resetting' the ocean) but different data and model input errors, hence different predictions Y^p. Obviously, we cannot practically perform such repeat experiments, but we can approximate them by resampling from assumed distributions of data and model input errors (generating a simulated ensemble). Arguably, what really matters is prediction error in a single experiment — the real one! One might therefore prefer a 'single-experiment' skill function similar to (3.1) but without the expectation operator. However, prediction errors in the single experiment depend on the particular values of data/model input errors which are by definition unpredictable. Consequently, the best estimate of 'ensemble skill' is in practice also the best estimate of single-experiment skill. For example, in the toy problem of Chapter 1 (Fig. 1.2), the overfitting cubic model did in fact yield the best single-experiment skill in 4% of repeat experiments, but no practical skill metric could diagnose these 'flukes' without knowledge of y^t (Table 1B.1).

Furthermore, estimates of ensemble statistics (such as bias and variance) may guide improvements in the model/prediction method. If the ensemble is defined such that all data and model input errors are randomised, then ensemble bias indicates where predictions are *systematically* in error, *suggesting* inadequacies in model formulation. Variance indicates where the model is *randomly* in error, suggesting inadequate accuracy of data and model inputs. It should be remembered, however, that variance at one time can produce bias at other times (via dynamical nonlinearity), and bias at one time can also affect the subsequent variance produced by forcing errors.

Generally, the prediction Y^p need not be an exact solution of the model Y^m: for example, it could be an average of Y^m over a Bayesian posterior pdf. In this study, the inductive model $Y^p = Y^m$ is formed by fitting a set of interpolation 'nodes' (see section 3.2.4), and no model inputs are required (except for a set of allowed range parameters V_a). The process model Y^m

is fitted by adjusting only the bulk biological parameters θ (including bulk initial conditions), hence $a = \theta$ (see section 3.2.5). Interannually-constant model inputs $\hat{\eta}$ account for the effect of changing surface irradiance and mixed layer depth. Interannually-variable model inputs $\hat{\lambda}$ account for nutrient and mesozooplankton forcings. The process model predictions Y^p are formed by re-running the fitted model once with the interannually-variable model inputs for the target year, hence: $Y^p = Y^m(\hat{\theta}(D_S, \hat{\kappa}_S), \hat{\lambda}_{T_i}, t)$.

Maximisation of *Skill* requires the model predictions to have minimal predictive mean-square error. Negative skill may result from failure of the model to accurately generalise (inter-/extrapolate) the chlorophyll mean field over the model input space (λ, t) due to inadequacies in the model formulation or prediction method. Alternatively, error may result if the estimated model inputs $(\hat{\kappa}_S, \hat{\lambda}_{T_i})$ are inaccurate or incomplete, i.e., missing information that is necessary to determine the true mean field. The model is required to 'interpolate' the chlorophyll field over a model input if the target value of that input lies on or between the maximum and minimum values in the assimilated set, and 'extrapolate' if it lies beyond — this being generally a more stringent test of the model.

In this application, mainly interpolation in seasonal time t (varying within years) is required with relatively little extrapolation (O(weeks) over the six-month interval T_{int}). However, both interpolation and significant extrapolation are required in the forcings λ, since these vary significantly over the sampled years and may vary beyond the range of the three sampled years in future years. Skill metrics must therefore be devised to test for this specific kind of predictive ability.

3.2.2 Skill Metrics

In order to estimate *Skill*, predictive mean-square error (MSE) must be estimated. A naïve estimate of MSE is provided by comparing data with the model output fitted to the same data:

$$\widehat{MSE}^n \equiv \frac{1}{N} \sum_{k=1}^{N} ((O_k - Y_k^p(D, \hat{\kappa}, t_k))^2 - (\hat{\sigma}_k^O)^2) \tag{3.2}$$

where N is the number of chlorophyll samples, O_k is the mean-field chlorophyll datum at time t_k, and $Y_k^p(D, \hat{\kappa}, t_k)$ is the model prediction at sampling time t_k, after calibration to the data set D with the model inputs $\hat{\kappa}$, and the superscript n denotes 'naïve'. The estimated observational error variances $(\hat{\sigma}^O)^2$ are independent of the tested models and are therefore irrelevant to estimating relative skill. To estimate absolute skill or MSE, their contribution must be subtracted from the squared misfit (see Appendix 3A). In general $(\hat{\sigma}^O)^2$ may be estimated from prior knowledge of the observational method, but here they are estimated by the data simulation model as discussed below.

The above estimate is naïve for two reasons. First, it is biased downward as an estimate of the intra-sample MSE (at the assimilation model inputs $\hat{\kappa}$), due to the covariance of the prediction Y^p with the test data O, since $O_k \in D$ (see Appendix 3A). This covariance depends on the model formulation, and in general increases with formulation complexity, therefore favouring more complex models even if they are not in fact more skillful. Note that if the $(\hat{\sigma}^O)^2$ are accurately estimated, this covariance may drive the naïve mean-square error estimate *negative*, especially if the model is overfitted (see section 3.1). Second, the naïve metric makes no attempt to estimate the extra-sample accuracy (accuracy at unsampled times/years) which is required by (3.1). What is needed, therefore, are alternative metrics that test against data which are *independent* of the calibration data, to remove spurious covariances, and which are *representative* of the generalising task defined by (3.1).

One method of achieving independent and representative tests is to divide the data set into disjoint calibration and validation data sets ('active' and 'passive' data as in Friedrichs et al., 2007; Smith et al., submitted). These calibration and validation data sets should be sufficiently separated that their errors are decorrelated and the task of extrapolating between them is repesentative of the task defined by (3.1). Many possible divisions may be applied in the search for independent and representative tests, but in this study only divisions with respect to the sampled years are considered. Hence, a subset S_i comprising $A_C < A_S$ years of data D_{S_i} is used to calibrate the model, and predictions are tested against data D_j from validation years j where $j \notin S_i$, so that the validation data D_j are independent of the predictions Y^p. Hence an extra-sample MSE estimate for year j after fitting to the i^{th} subset of A_C years is:

$$\widehat{MSE}_{ij}^{d(A_C)} \equiv \frac{1}{N_j} \sum_{k=1}^{N_j} ((O_{jk} - Y^p_{jk}(D_{S_i}, \hat{\kappa}_{S_i}, \hat{\lambda}_j, t_{jk}))^2 - (\hat{\sigma}^O_{jk})^2) \tag{3.3}$$

where N_j is the number of chlorophyll samples in year j and the superscript $d(A_c)$ denotes 'division' and 'fitting to A_C years'. The hope is that the task of generalising from $(\hat{\lambda}_{S_i}, t_{S_i})$ to $(\hat{\lambda}_j, t_j)$ is representative of the task defined by $Skill$. Such tests become more stringent as A_C is reduced. A more robust estimate of $Skill$ is formed by averaging over all possible A_C-year division experiments, fitting to $A_C < A_S$ years and testing predictions for each of the remaining $(A_S - A_C)$ years. There are $(A_S - A_C)B_C$ such tests, where B_C is the binomial coefficient $A_S!/((A_S - A_C))!A_C!)$. Hence:

$$\widehat{Skill}^{d(A_C)} \equiv \frac{-1}{(A_S - A_C)B_C} \sum_{i=1}^{B_C} \sum_{j \notin S_i} \widehat{MSE}_{ij}^{d(A_C)} \tag{3.4}$$

In general, this may yet be unsatisfactory for several reasons. First, the least stringent metric $\widehat{Skill}^{d(A_S-1)}$, based on fitting to $(A_S - 1)$ years, may still be too stringent to represent $Skill$. For example, suppose that the forcings in the future target years were known to be within the ranges of values explored in 1997–1999. In this case, the least stringent division metric, based on fitting two years and predicting the third, is liable to underestimate skill, because it will test for extrapolative abilities which are not actually required by the skill function. Or, suppose the objective were just to estimate mean fields during the sampled months in 1997–1999 — again, no division metric would approximate the skill function, because division skill tests cannot measure intra-sample accuracy (accuracy at the sampled data). Second, the number of calibration-validation tests $(A_S - A_C)B_C$ may not be large enough to obtain a robust skill metric or robust estimates of predictive bias and variance (Hastie et al., 2001; Appendix 1B). These two problems are especially serious if A_S is small.

These shortcomings are addressed by combining the division method with a data simulation technique sometimes referred to as 'Monte Carlo Simulation' (Press et al. 1999) or the 'parametric bootstrap' (Young, 1994). Here, a simulation model is fitted to the original full data set, and used to produce new data with realistic random errors. Thus, independent ensembles of simulated calibration and validation data sets $\{D^C\}$ and $\{D^V\}$ are produced for testing predictive models, with the fitted data simulation model Y^{sim} playing the role of the unknown truth Y^t. The model predictions Y^p are compared with the 'truth' Y^{sim}_j for the validation year j, after fitting to the calibration data subset $D^C_{S_i}$ and supplying independent validation forcing estimates $\hat{\lambda}^V_j$ derived from D^V. By using the same simulation model to create $\{D^C\}$ and $\{D^V\}$, it is implicitly assumed that the target year forcing estimates $\hat{\lambda}_T$ will have similar uncertainty to those used in calibration $\hat{\lambda}_S$. The resulting tests will not be biased by prediction covariance, even if the validation year is contained in the calibration subset. They may of course be biased by infidelities in the data simulation. Twin experiments suggested however that this information loss was not serious enough to prohibit accurate estimation of the tested model prediction statistics, assuming that the simulation model is reasonably accurate (see section 3.4.1).

By this approach, a matrix of A_C-year mean-square predictive error estimates ($MSE^{sd(A_C)}$) is formed from all possible combinations of A_C calibration years and single validation years, which may now include those in the calibration set:

$$\widehat{MSE}^{sd(A_C)}_{ij} \equiv \frac{1}{N_j} \sum_{k=1}^{N_j} \overline{(Y^{sim}_{jk} - Y^p_{jk}(D^C_{S_i}, \hat{\kappa}^C_{S_i}, \hat{\lambda}^V_j, t_{jk}))^2} \tag{3.5}$$

where the superscript sd denotes 'simulation and division' and the overline here denotes an average over an ensemble of simulations with different realisations of the noise in $D^C_{S_i}$, $\hat{\kappa}^C_{S_i}$ and $\hat{\lambda}^V_j$. By averaging over an ensemble of independent skill tests (50 in our case) a more robust estimate is achieved. Note that the data simulation model need only be accurate *at the data* — it does not require any ability to generalise. An estimate of $Skill$ is now obtained by averaging over all components of this matrix:

$$\widehat{Skill}^{sd(A_C)} \equiv \frac{-1}{A_S B_C} \sum_{i=1}^{B_C} \sum_{j=1}^{A_S} \widehat{MSE}^{sd(A_C)}_{ij} \tag{3.6}$$

These metrics are likely to become increasingly stringent as A_C is decreased from A_S down to the minimum calibration set size of $A_C = 1$. In theory, there is an (unknown) optimal value $A_C = A^*_C$ which provides the best 'gauge' of $Skill$, in the sense that the expected $Skill$ of the model selected by the skill metric is optimal. Since only $A_S = 3$ sampled years are available in this case, and the precise degree of extrapolation required is not specified in (3.1), all possible values of A_C will be considered as potentially optimal gauges of $Skill$.

Hence the primary skill metrics of this study are the matrices $\widehat{MSE}^{sd(1)}$, $\widehat{MSE}^{sd(2)}$, and $\widehat{MSE}^{sd(3)}$ and corresponding scalar metrics $\widehat{Skill}^{sd(1)}$, $\widehat{Skill}^{sd(2)}$ and $\widehat{Skill}^{sd(3)}$. For comparison, the naïve metric $\widehat{Skill}^{n(3)} = -\widehat{MSE}^{n(3)}$ and the two division-only skill metrics $\widehat{Skill}^{d(1)}$ and $\widehat{Skill}^{d(2)}$ are also evaluated, using the data simulation model to estimate $(\hat{\sigma}^O_{jk})^2$ (see Appendix 3B). For each of these metrics, the relative skill metric for a model formulation $\mathcal{M}$ is defined as:

$$\Delta\widehat{Skill}(\mathcal{M}) \equiv \widehat{Skill}(\mathcal{M}) - \widehat{Skill}(ind) \tag{3.7}$$

where $\widehat{Skill}(ind)$ is any skill metric obtained using the inductive model (see section 3.2.4). Hence positive $\Delta\widehat{Skill}(\mathcal{M})$ is good, negative $\Delta\widehat{Skill}(\mathcal{M})$ is bad, and the inductive model acts

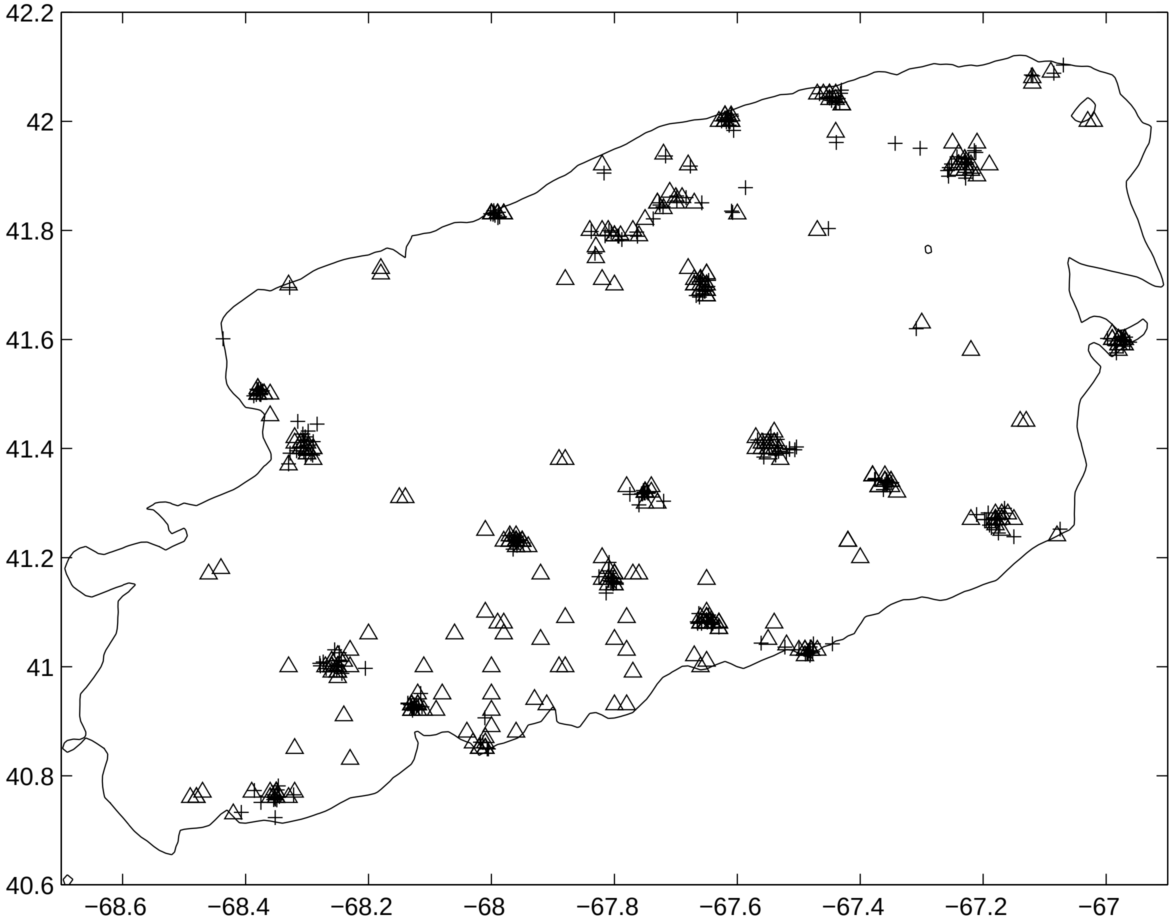

Figure 3.1: Raw sample locations on Georges Bank within the 60m isobath (longitude horizontal, latitude vertical). Triangles denote nutrient and phytoplankton samples, crosses denote mesozooplankton samples.

as a baseline ($\Delta\widehat{Skill}(ind) = 0$).

3.2.3 Data

The data were all taken from the GLOBEC website for Georges Bank[3] — see Acknowledgements for specific data providers. Attention is restricted to a volume of study comprising the near-surface mixed layer and constrained horizontally within the 60m isobath. This maximum water depth was chosen as the smallest value that allowed sample sizes large enough for robust mean field estimation, as a strategy to include a minimal volume of water which is not permanently well-mixed (Steele et al., 2007). Extracted chlorophyll-a and 'nutrient' (nitrate + nitrite + ammonium) concentrations within this region were sampled at near-surface (<5m) depths. Mesozooplankton abundance estimates were also obtained as 'displacement volume' samples (the total volume of matter in cm^3 per 10m^3 of filtered water) using a towed Bongo net with a mesh size of 333 μm. The locations of all samples within the study region are shown in Figure 3.1.

The raw data are clustered into roughly 10-day cruise intervals for each month from January up to and including June (with a few omissions). The data are treated as mean-field estimates with 'data errors' due to instrument noise and spatial variability (see section 1.5.2).

[3] http://globec.whoi.edu/jg/dir/globec/gb/

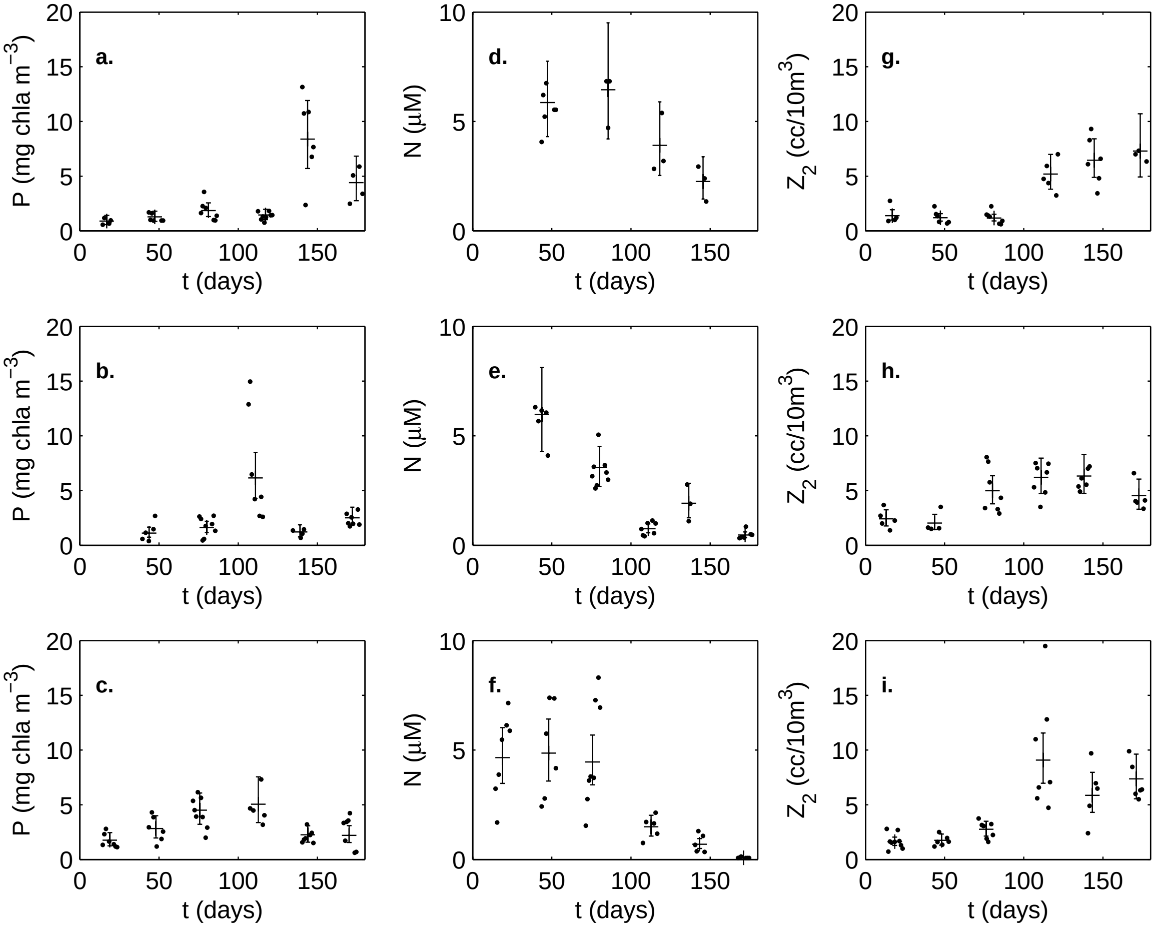

Figure 3.2: Georges Bank time series of diel-averaged data (dots) for phytoplankton (a,b,c), nutrient (d,e,f) and mesozooplankton data (g,h,i) during 1997 (a,d,g), 1998 (b,e,h) and 1999 (c,f,i). Error bars show simulation model means with 95% confidence intervals.

Estimating the along-track temporal autocorrelation function for each sampled variable reveals significant error correlations due to spatial structure between points sampled less than roughly 1 day apart. All samples are therefore diel-averaged to remove these correlations, which might otherwise lead us to underestimate the variance of our predictions. This averaging also removed any significant error correlations detectable between cruise samples of different variables sampled at the same locations and times, as well as any day-night signals (although none were apparent).

In the resulting diel-averaged data set (see Figure 3.2), the data errors are independent and generally much larger than any mean trend during each cruise. Linear regressions on the time series for each cruise revealed significant mean trends (at the 5% level) in at most 2 cruises out of 15-18 for each variable. This motivated use of a data simulation model consisting of independent samples from stationary lognormal distributions for each month and each variable. The data simulation model and its selection process are detailed in Appendix 3B. The data simulation model also helped to identify outlying data which were subsequently discarded (see Appendix 3C).

For the purpose of fitting and testing predictive models, the data were finally averaged again to minimise noise in the mean-field signal. In keeping with the assumption of negligible mean-field variation during each cruise in the simulation model, the running-average window was chosen to cover each roughly 10-day cruise period — in other words, the data were averaged over each cruise (month). Note that this two-step averaging is not equivalent to a simple average of the monthly data: multiple samples within the same day (and therefore correlated) have the same combined weight in the two-step average as a single sample from another day of the cruise sampled only once. The same cruise-averaging was applied to any 'new' diel-averaged data sets simulated by the data simulation model.

3.2.4 Inductive Model Formulation

The 'inductive'[4] model in this study is a six-point piecewise-linear interpolation of the chlorophyll-a data. It acts a baseline of predictive skill. The inductive model makes no use of forcing parameter information for the calibration years or target year viz. it is constant over forcing parameter space. It is formally specified by:

$$Y^p(ind)(t) = \begin{cases} a_1 + \frac{(t-p_1)}{(p_2-p_1)}(a_2 - a_1) & t < p_1 \\ a_i + \frac{(t-p_i)}{(p_{i+1}-p_i)}(a_{i+1} - a_i) & p_i \leq t < p_{i+1}, i = 1, ..., 5 \\ a_6 + \frac{(t-p_6)}{(p_6-p_5)}(a_6 - a_5) & t \geq p_6 \end{cases}$$

where $a_{1,...,6}$ are free parameters (the 'node amplitudes') and $p_{1,...,6}$ are node times fixed at the averages of the sampling times of the diel-averaged data for each month, so that the nodes are positioned approximately in the regions of maximum data constraint. Negative values of Y^p were set to zero. For computational purposes, the allowed range for each parameter in this model-fit was set as $0 < a_i < 20$ mg m^{-3} for $a_{1,...,6}$, but this did not restrict the fit.

3.2.5 Process Model Formulation

The process (dynamical) model used in this study is essentially a Lotka-Volterra predator-prey interaction between chlorophyll (P) and microzooplankton (Z_1) mean fields, driven by a dimensionless factor $\gamma(t)$ to account for seasonal changes in mean surface irradiance and mixed layer depth, and by the mean concentrations of limiting nutrients $N(t)$ and mesozooplankton $Z_2(t)$, which eat the microzooplankton:

$$\begin{aligned} \frac{dP}{dt} &= u_0\gamma(t)\frac{N(t)}{k_N + N(t)}\frac{P}{1+cP} - gPZ_1 \\ \frac{dZ_1}{dt} &= gPZ_1 - fZ_1Z_2(t) \end{aligned} \tag{3.8}$$

The equations (3.8) were integrated using the 4^{th}-order Runge-Kutta scheme over the 180-day January to June interval with a step of 1/4 day. The seven free parameters are: the maximal growth rate u_0, the nutrient half-saturation constant k_N, the self shading parameter c, the microzooplankton grazing rate g, the mesozooplankton carnivory rate f, the initial chlorophyll concentration $P(0)$ and the initial microzooplankton concentration $Z_1(0)$. These parameters are identified with their units and allowed ranges (defining the allowed volume V_θ in parameter space) in Table 3.1. Note that the grazing rate g and concentration $Z_1(t)$ in (3.8) are scaled by $1/e'$ and e' respectively, where e' accounts for microzooplankton growth

[4] From the Bertrand Russell's 'Inductivist Turkey' which successfully predicted its feeding time based on an average of past observations, right up until the morning of Christmas Eve...

Table 3.1: Process model parameters, symbols, units and allowed ranges

Parameter	Symbol	Units	Allowed range
maximum P growth rate	u_0	day^{-1}	0.001 – 1
nutrient half-saturation constant	k_N	mg chla m^{-3}	0.001 – 3
self-shading constant	c	(mg chla m^{-3})$^{-1}$	0.05 – 0.3
Z_1 grazing rate	g	(mg chla m^{-3})$^{-1}$ day^{-1}	0.001 – 0.5
Z_2 carnivory rate	f	(cc/(10m^3))$^{-1}$ day^{-1}	0.001 – 0.5
Initial P concentration	$P(0)$	mg chla m^{-3}	0.01 – 1.5
Initial Z_1 concentration	$Z_1(0)$	mg chla m^{-3}	0.01 – 5.00

efficiency and unit conversion factors (see below). The process model allowed parameter ranges are discussed in Appendix 3D.

The dimensionless factor $\gamma(t)$ in (3.8) is computed as follows. First, a relationship is assumed between cell photosynthesis (Ph) and irradiance (I) explored in Platt et al. (1990):

$$Ph(I) = A(1 - e^{I/I_s})e^{I/I_i} \tag{3.9}$$

where I_s and I_i denote saturation and inhibition intensities respectively. These are estimated using the photosynthesis-irradiance data sets collected by D. Townsend in Massachusetts Bay during February and August 1990 (Ji et al., 2006). Fitting (3.9) to each of these two data sets yields significantly lower $\hat{A}$, $\hat{I}_s$ and $\hat{I}_i$ in February than in August. Therefore, to obtain estimates ($\hat{I}_s$, $\hat{I}_i$) applicable to the period January to June, the two data sets are renormalised by their respective values of $\hat{A}$ and (3.9) is refitted to the resulting combined data set. An ensemble of such estimates is generated by bootstrap resampling[5] the combined data and refitting (3.9). From this ensemble, the uncertainty in the estimates is estimated, hence: $\hat{I}_s = (0.128 \pm 0.014) \times 10^{-3}$ Ein/m^2/s and $\hat{I}_i = (5.45 \pm 1.38) \times 10^{-3}$ Ein/m^2/s.

Second, the form of the diel variability in surface irradiance $I_0(x, y, t)$ is determined by fitting functions of the form $I_0(t') = I_0^{max} \sin^n(\pi t'/D(t))$ (for $0 < t' < D(t)$, zero otherwise) to R. Payne's 60-minute averaged data for Georges Bank in 1997–1999, using NOAA's 'low accuracy' solar position equations (http://www.srrb.noaa.gov/highlights/sunrise/solareqns.PDF) to calculate the daylength $D(t)$, sunrise ($t' = 0$) and sunset ($t' = D(t)$) times for each of the 180 days. The square exponent $n = 2$ was found to give the best fit. The resulting spatial estimates $\hat{I}_0^{max}(x, y, t)$ provided a time series which was highly variable due to changing sampling locations and the high-frequency spatio-temporal variability in cloud cover. By fitting a linear trend to this time series, a smooth increase in the areal-mean noon surface irradiance $I_0^m(t)$ is assumed (data from all three years were fitted together in lieu of any significant interannual variability). Again, bootstrap resampling is used to generate an ensemble of initial and final mean intensity estimates, hence: $\hat{I}_0^m(0) = (1.31 \pm 0.14) \times 10^{-3}$ Ein/m^2/s and $\hat{I}_0^m(180) = (4.00 \pm 0.19) \times 10^{-3}$ Ein/m^2/s.

Third, mixed layer depths are estimated by applying the Levitus criterion ($\Delta\sigma_\theta = 0.125$ kgm^{-3} with respect to the surface) to potential temperature profiles calculated from temper-

[5] Here, samples are chosen at random *with replacement* from the original data set, thereby generating a new 'bootstrap' data set (Young, 1994; Efron and Tibshirani, 1993). This simple nonparametric technique works well for linear model fits where the (Gauss-Markov) assumptions of standard linear regression may invalidate the standard approach to estimating parameter uncertainty. However, twin tests suggested that it was not as effective as a 'parametric bootstrap' method for simulating resamplings of the chlorophyll, nutrient and mesozooplankton data sets (see section 3.2.2)

ature and salinity data. The resulting time series for each year are very highly variable due to spatially-variable stratification and are relatively poorly-sampled in later months. Linear trends are therefore fitted to robustly estimate the areal-mean variations $M(t)$. Although a significantly steeper slope is observed in the 1999 fit, the uncertainty in $\hat{M}(t)$ defrays any improvement in predictive skill gained by using $\hat{M}(t)$ as an interannually-variable forcing. Therefore the data from all three years are combined to give an interannually-constant forcing $\hat{M}(t)$, again using bootstrap resampling to generate an ensemble of linear trend parameter estimates, yielding: $\hat{M}(0) = (55.6 \pm 1.6)$ m and $\hat{M}(180) = (41.5 \pm 2.2)$ m.

Bulk primary productivity averaged over horizontal space, the mixed layer $0 < z < \hat{M}(t)$ and the diel cycle $0 < t' < 1$, is now approximated by the first term in (3.8), with:

$$\gamma(t) = \frac{1}{\hat{M}(t)} \int_0^{D(t)} \int_0^{\hat{M}(t)} (1 - e^{-\hat{I}(z',t')/\hat{I}_s}) e^{-\hat{I}(z',t')/\hat{I}_i} \mathrm{d}z' \mathrm{d}t' \tag{3.10}$$

where $\hat{I}(z', t') = \hat{I}_0^m(t) \sin^2(\pi t'/D(t)) e^{-\hat{k}_e z'}$. As in other cases, the effects of horizontal variability on the areal average of the nonlinear integrand (the 'biological Reynolds fluxes' — see section 1.5.3 and Chapter 4) are assumed to be accounted for by retuning the control parameters (see Appendix 3D). Note that any diel signals in mixed layer depth $M(t')$, phytoplankton $P(t')$ or nutrient concentrations $N(t')$ are neglected (although none were apparent). Also, the effect of self shading is not accounted for by $\gamma(t)$, but rather by a separate factor $(1 + cP)^{-1}$ in (3.8). This is valid because $\gamma(k_e, t) \propto k_e^{-1}$ to a very good approximation. Hence accounting for self shading by replacing k_e with $(k_e + k_p P)$ only introduces the factor $(1 + cP)^{-1}$ with $c \equiv k_p/k_e$.

A first attempt to compute $\gamma(t)$ involved expanding the exponents as power series and integrating each term as in Platt et al. (1990). However, this was found to be numerically inaccurate for our largest estimates of the relative mean noon surface irradiance $\hat{I}_0^m(t)/\hat{I}_s > 40$ (requiring over 50 terms for convergence). Therefore simple mid-point rule integration over z' and t' was used with step sizes small enough for convergence ($\Delta z' = 0.2$m and $\Delta t' = 0.02$ days). This numerical expense is affordable because the integration does not need to be repeated when running or fitting the process model — rather, $\gamma(t)$ is approximated by a polynomial (linear was found to be adequate) and the parameters are exported to the process model fit. Note that the strong linearity of $\gamma(t)$ does not imply that saturation/inhibition effects may be neglected: $e^{-k_e M(t)} \ll 1$, implying $\gamma(k_e, t) \propto k_e^{-1}$, but $\hat{I}_0^m(t)/\hat{I}_s \sim \mathrm{O}(10)$, so many more terms in the exponential expansion beyond the first are required for convergence.

Integration of (3.10) requires estimation of k_e, the light attenuation due to water and non-phytoplankton suspended matter. An estimate $\hat{k}_e = 0.15\mathrm{m}^{-1}$ is used for consistency with Ji et al. (2007), with an allowed (uniform prior uncertainty) range $0.1 < k_e < 0.2\mathrm{m}^{-1}$. An ensemble of parameter estimates is generated by repeating the integration and linear fit with an ensemble of randomly-chosen $\hat{k}_e$ combined (independently) with the ensembles of parameters for $\hat{I}_0^m(t)$, $\hat{M}(t)$ and $\{\hat{I}_s, \hat{I}_i\}$. This finally yields: $\gamma(0) = 0.074 \pm 0.015$ and $\gamma(180) = 0.257 \pm 0.043$.

Nutrient (nitrate + nitrite + ammonium) limitation is represented by a simple Michaelis-Menten term, neglecting any limitation by silicate concentrations as these were not observed to vary much between years, surely not enough to explain the interannual variability in chlorophyll. Phytoplankton mortality due to causes other than grazing is neglected. The primary consumers are assumed to be microzooplankton (heterotrophic protozoans, which were not sampled in the GLOBEC program) which are then consumed by the mesozooplankton (mainly the copepod *Calanus finmarchicus*) (Ji et al., 2006). A constant bulk C (or N):Chl ratio is implicitly assumed for the primary producers, since the consumption by microzooplankton Z_1 depends on biomass rather than chlorophyll content. Herbivory by mesozooplankton is neglected. The grazing rates for microzooplankton on phytoplankton, and for mesozooplankton on microzooplankton, are both assumed to be non-saturating. Entrainment and losses from the mixed layer due to cross-pycnal mixing and sinking are neglected (Evans and Parslow, 1985). The initial concentrations $P(0)$ and $Z_1(0)$ are assumed to be invariant between years.

Readers uncomfortable with these assumptions should note: first, the primary concern of this study is with methods for rigorously comparing the skill of two models, not finding the best model to explain the observations; second, the aim is to model concentrations averaged over a coarse scale (O(100km)), whose behaviour may differ widely from laboratory-scale population dynamics (see Chapter 4); third, the robustness of the ecological conclusions is tested by fitting various extensions of (3.8) (see section 3.4.1).

Finally, one might wonder why the RHS of the Z_1 equation in (3.8) does not read: $egPZ_1 - fZ_1Z_2(t)$ with e the product of an ingestion efficiency α and a units conversion factor β (units(Z_1)/units(P)). This is because, in lieu of data for Z_1, fitting e would result in a degenerate or 'non-identifiable' triplet of parameters (g, e, $Z_1(0)$). The only data constraint on Z_1 is due to its effect on P via the grazing rate $G(t) = gZ_1(t) = g(Z_1(0) + \int_0^t egP(s)Z_1(s) - fZ_1(s)Z_2(s)\mathrm{d}s)$. This rate is invariant under the parameter rescalings: $g' = \zeta g$, $Z_1(0)' = Z_1(0)/\zeta$, $e' = e/\zeta$, which only result in a rescaling: $Z_1(t)' = Z_1(t)/\zeta$ and no effect on the data fit. Hence: using any arbitrary value of e to perform the model fits, exactly the same $P(t)$ can be recovered for a different value of e simply by applying the appropriate rescalings to g and Z_1. Therefore, for simplicity, the minimisation problem is solved using $e = 1$ with Z_1 measured in units of equivalent chlorophyll concentration.

3.2.6 Prediction method

The method adopted for fitting the inductive and process models, and for estimating forcing parameters, is Ordinary Least Squares. Hence the *cost* function is defined by:

$$cost = \sum_{j=1}^{A_C} \sum_{k=1}^{N_j} (O_{jk} - Y_{jk}^m(\theta, \hat{\kappa}_j, t_{jk}))^2 \tag{3.11}$$

where A_C is the number of calibration years ($\leq A_S$), O_{jk} is the k^{th} datum for year j and $Y_{jk}^m(\theta, \hat{\kappa}_j, t_{jk})$ is the corresponding fitted model value, written here as a function of the model parameters θ, the model inputs $\hat{\kappa}_j$, and the time t_{jk}. Note that variables are *not* log-transformed. Although the diel-averaged data are lognormal, the cruise-averaged data used for fitting are sample means of lognormal variables. For the cruise sample sizes (3–9) these sample means are intermediate between lognormal and normal variables. Moreover, even if they were lognormal, minimising the sum of squares in log space will fit the Y^m to the *median* of the population (true distribution) rather than the population mean, which is the true value (for unbiased data). In other words, the exponent of the sample mean in log space is a *biased* (under)estimate of the population mean, and needs to be corrected (unbiased) by a less-robust estimate of the population variance (Campbell, 1995). Also, *un*-weighted Least

Squares is used even though the cruise-averaged data uncertainty varied strongly month-to-month (see Fig. 3.2). Fitting the inductive model using Weighted Least Squares was found to yield biased and consequently less skillful estimates, as months with high error variances were 'smoothed over' in the piecewise-linear interpolation. The *cost* in (3.11) is minimised using a non-local optimisation algorithm discussed in Appendix 3E.

The process model forcings $N(t)$ and $Z_2(t)$ were determined prior to fitting (3.8) by fitting piecewise-linear models, as described in section 3.2.5, to nutrient and mesozooplankton biovolume data from each year individually. The resulting forcings comprised the interannually-variable model inputs: $\hat{\lambda} = \{\hat{N}(t), \hat{Z}_2(t)\}$. The interannually-constant model inputs $\hat{\eta}$ included the parameters defining $\gamma(t)$ in (3.8) (see section 3.2.5). Having fitted the process model to obtain $\hat{\theta}(D, \hat{\kappa})$, the model is re-run with the forcings $\hat{\lambda}_j$ for the prediction (test) year j to give the model predictions $Y^p = Y^m(\hat{\theta}(D, \hat{\kappa}), \hat{\lambda}_j, t)$.

3.3 Results

3.3.1 Inductive Model Predictions

Figure 3.3 shows a selection of inductive model skill tests. The least challenging are shown in Figs. 3.3a–c, where the ability of the model to predict the mean chlorophyll concentration in one year given data from the same year is tested. The fact that the mean of the predictions over the 50-member ensemble coincides with the true values (as defined by the data simulation model) shows that the predictions are unbiased. The width of the ensemble envelope shows the prediction standard deviation due to errors in the calibration data. This increases with the true mean field, due to the lognormality of the diel-averaged data, and as expected covers roughly half the widths of the cruise-average confidence intervals shown in Figs. 3.2a–c. The naïve mean-square error estimates (not shown) are all negative, showing that the model is overfitted (inevitable for these tests given the freedom of the inductive model). However, the simulation-division mean-square error estimates shown in Figs. 3.3a–c are small but positive, since they account for the covariance of predictions with data errors, and are therefore not biased by overfit.

Obviously, when the one-year inductive model fits shown in Figs. 3.3a–c are used to predict different years, the predictions suffer high extra-sample bias (outside the calibration data), reflecting the large interannual variability. The resulting mean-square error estimates for different validation years, represented by the 'off-diagonal' components of the $\widehat{MSE}^{sd(1)}$ matrix, are high (see Table 3.2), especially those where 1997 is used to predict 1998 or 1999 and vice versa. These latter errors tend to dominate the matrix-average, yielding $\widehat{Skill}^{sd(1)}(ind) = -6.3 \pm 0.2$ (suppressing units). The result obtained using the division-only skill metric $\widehat{Skill}^{d(1)}(ind) = -9.3 \pm 2.2$ is necessarily more pessimistic because same-year tests must be excluded, and less robust because it is formed by an average only over the 6 possible values of $\widehat{MSE}_{ij}^{d(1)}$ for $i \neq j$, using the original data and no ensemble.

The results for two-year and three-year tests of the inductive model skill are unsurprising. Figs. 3.3d–f show the two-year tests for which the validation year is not included in the calibration set (and to which the division-only metric is restricted). The predictions are obviously biased although there is a slight decrease in variance due to increased calibration sample size. The corresponding skill estimate is less pessimistic than the one-year estimate:

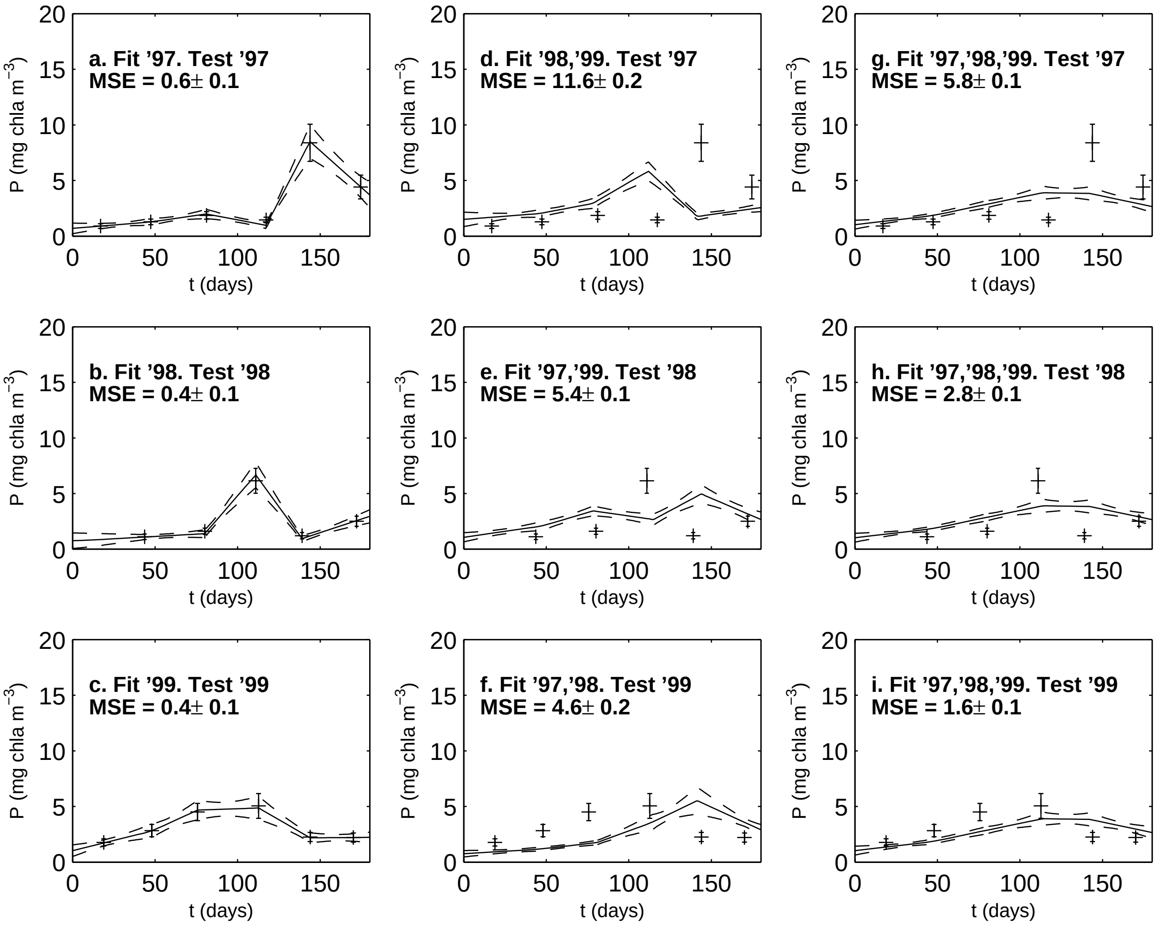

Figure 3.3: Inductive model forecast skill tests: fitting to one year and testing against the same year (a,b,c), fitting to two years and testing against the remaining year (d,e,f), and fitting to all three years (g,h,i). Predictions are for 1997 (a,d,g), 1998 (b,e,h) and 1999 (c,f,i). Solid and dashed lines show prediction ensemble means ± 1 standard deviation. Error bars show the true values ± 1 ensemble standard deviation of the cruise-averaged data. Numbers in bold show time-averaged mean-square error (MSE) estimates based on the ensembles of simulation-division skill tests.

Table 3.2: Mean-square error estimates ($\widehat{MSE}_{ij}^{sd}$) and skill metrics ($\widehat{Skill}^{sd}$ and $\Delta\widehat{Skill}^{sd}$) for the process model, with inductive model results shown in brackets for comparison. The validation index j runs over the three years: 1997 (j=1), 1998 (j=2) and 1999 (j=3), as does the calibration index i in $\widehat{MSE}_{ij}^{sd(1)}$. In $\widehat{MSE}_{ij}^{sd(2)}$, the calibration index i runs over the sets of years: (1997,1998) (i=1), (1997,1999) (i=2), (1998,1999) (i=3).

$\widehat{MSE}^{sd}$	$j=1$	$j=2$	$j=3$	$\widehat{Skill}^{sd}$	$\Delta\widehat{Skill}^{sd}$
$\widehat{MSE}_{1j}^{sd(1)}$	2.1 ± 0.2 (0.6 ± 0.1)	8.9 ± 0.9 (12.1 ± 0.4)	4.6 ± 0.5 (11.6 ± 0.5)		
$\widehat{MSE}_{2j}^{sd(1)}$	5.6 ± 0.4 (13.0 ± 0.3)	1.6 ± 0.4 (0.4 ± 0.1)	5.6 ± 0.4 (3.1 ± 0.1)	-5.2 ± 0.2 (-6.3 ± 0.2)	+1.1 ± 0.3 (0.0 ± 0.0)
$\widehat{MSE}_{3j}^{sd(1)}$	12.3 ± 0.3 (11.5 ± 0.3)	5.5 ± 0.3 (4.0 ± 0.2)	0.8 ± 0.2 (0.4 ± 0.1)		
$\widehat{MSE}_{1j}^{sd(2)}$	4.7 ± 0.5 (3.4 ± 0.2)	3.8 ± 0.6 (4.4 ± 0.2)	6.9 ± 0.8 (4.6 ± 0.2)		
$\widehat{MSE}_{2j}^{sd(2)}$	2.9 ± 0.5 (3.8 ± 0.1)	11.2 ± 0.9 (5.4 ± 0.1)	2.7 ± 0.4 (2.6 ± 0.1)	-5.0 ± 0.2 (-4.2 ± 0.1)	-0.8 ± 0.3 (0.0 ± 0.0)
$\widehat{MSE}_{3j}^{sd(2)}$	7.7 ± 0.4 (11.6 ± 0.2)	3.1 ± 0.5 (1.2 ± 0.1)	2.2 ± 0.3 (0.8 ± 0.1)		
$\widehat{MSE}_{j}^{sd(3)}$	4.6 ± 0.4 (5.8 ± 0.1)	5.5 ± 0.6 (2.8 ± 0.1)	5.3 ± 0.6 (1.6 ± 0.1)	-5.1 ± 0.4 (-3.4 ± 0.2)	-1.7 ± 0.5 (0.0 ± 0.0)

$\widehat{Skill}^{sd(2)}(ind) = -4.2 \pm 0.1$, reflecting the smaller extrapolative demands. The division-only metric gives a similar but less certain result: $\widehat{Skill}^{d(2)}(ind) = -7.3 \pm 2.6$. The predictions based on fitting to three years (Figs. 3.3g–i) are again biased but are now maximally robust, and the skill metric is again less pessimistic: $\widehat{Skill}^{sd(3)}(ind) = -3.4 \pm 0.2$. No division-only metric is available when fitting to all three years of data, since divisions of the data within each year are not considered. In this case, the naïve metric gives a similar result: $\widehat{Skill}^{n(3)}(ind) = -3.1 \pm 0.7$. This reflects the fact that when fitted to a large amount of data the inductive model yields robust predictions with little covariance between predictions and data errors.

3.3.2 Process Model Predictions

Figs. 3.4a–c show the process model predictions fitted to one year and tested against the same year. The prediction bias and variance are both clearly greater than for the inductive model, resulting in greater mean-square error. Again, the mean-square error increases markedly when we try to predict a different year (off-diagonal elements in Table 3.2), showing that the given $(N(t), Z_2(t))$ forcing information (Fig. 3.5) is inadequate to allow the process model to extrapolate accurately from a single year of data. The skill estimate derived from such tests ($\widehat{Skill}^{sd(1)}(pr) = -5.2 \pm 0.2$) is nevertheless slightly better than the inductive model skill ($\Delta\widehat{Skill}^{sd(1)}(pr) = +1.1 \pm 0.3$). A similar but less certain result is obtained by the division-only metric ($\widehat{Skill}^{d(1)}(pr) = -5.5 \pm 1.3$, $\Delta\widehat{Skill}^{d(1)}(pr) = +3.8 \pm 2.5$).

Using two years to calibrate the process model (as in Figs. 3.4d–f) results in a similar measure of process model skill ($\widehat{Skill}^{sd(2)}(pr) = -5.0 \pm 0.2$), this time significantly worse than the inductive model ($\Delta\widehat{Skill}^{sd(2)}(pr) = -0.8 \pm 0.3$). The division-only metric

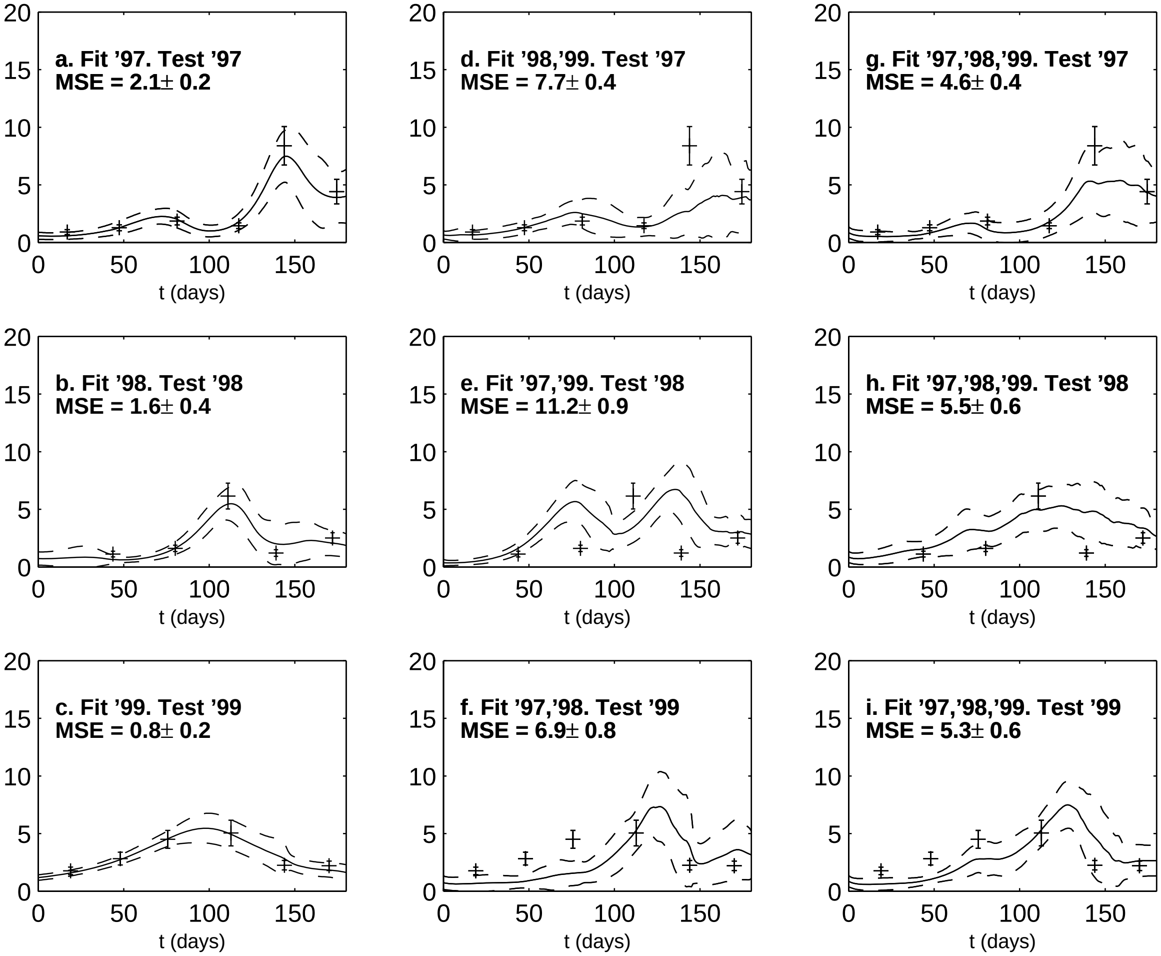

Figure 3.4: Process model forecast skill tests: fitting to one year and testing against the same year (a,b,c), fitting to two years and testing against the remaining year (d,e,f), and fitting to all three years (g,h,i). Predictions are for 1997 (a,d,g), 1998 (b,e,h) and 1999 (c,f,i). Solid and dashed lines show prediction ensemble means ± 1 standard deviation. Error bars show the true values ± 1 ensemble standard deviation of the cruise-averaged data. Numbers in bold show time-averaged mean-square error (MSE) estimates based on the ensembles of simulation-division skill tests.

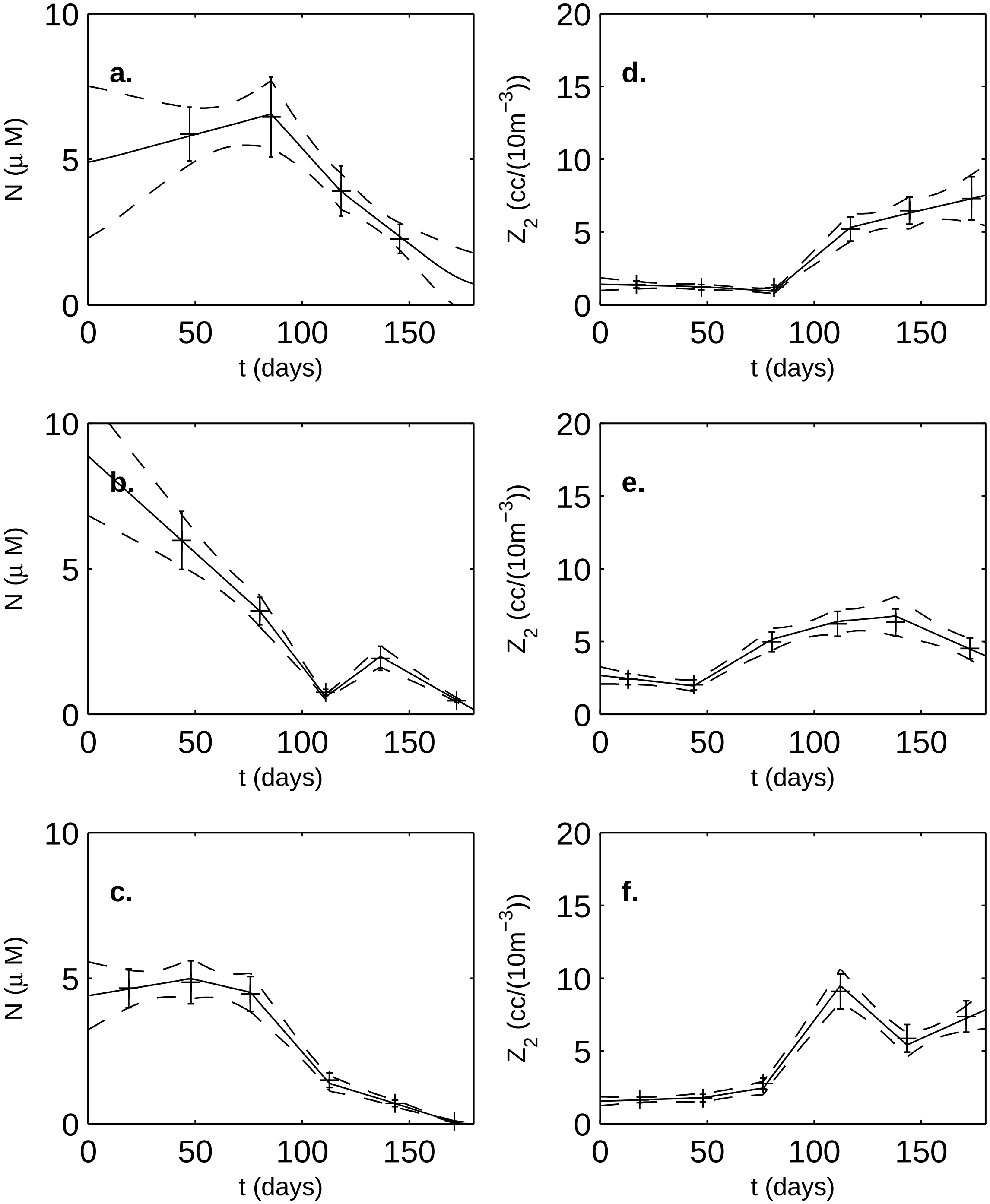

Figure 3.5: Process model forcing estimates for nutrient $N(t)$ (a,b,c) and mesozooplankton $Z_2(t)$ (d,e,f) mean fields during 1997 (a,d), 1998 (b,e) and 1999 (c,f). Solid and dashed lines show prediction ensemble means ± 1 standard deviation. Error bars show the true values ± 1 ensemble standard deviation of the cruise-averaged data.

gives an inconclusive result ($\widehat{Skill}^{d(2)}(pr) = -7.4 \pm 0.2$, $\Delta\widehat{Skill}^{d(2)}(pr) = -0.1 \pm 2.6$). The three-year tests (Figs. 3.4g–i) also favour the inductive model: $\widehat{Skill}^{sd(3)}(pr) = -5.1 \pm 0.4$, $\Delta\widehat{Skill}^{sd(3)}(pr) = -1.7 \pm 0.5$. By contrast the naïve metric gives $\widehat{Skill}^{n(3)}(pr) = -0.6 \pm 0.4$ and $\Delta\widehat{Skill}^{n(3)}(pr) = +2.5 \pm 0.8$, hence strongly favouring the process model.

3.4 Discussion

3.4.1 Limitations of the Process Model

In the skill tests, the inductive model performance was quite predictable, but the process model performance was rather disappointing. For example, the one-year tests shown in Figs. 3.4a–c might have resembled 'curvy' versions of Figs. 3.3a–c. Instead, the process model predictions are found to have slightly higher bias and higher variance, resulting in significantly higher mean-square error. The reason lies in the process model forcings, which as Figure 3.5 shows are estimated with significant error. If the ensemble of predictions is re-run using the *same* forcing parameter estimates as those used in calibration (see Fig. 3.6), the resulting 'predictions' are significantly more accurate (cf. Figs. 3.6a–c and Figs. 3.4a–c), but they are not relevant to the forecast-based skill function (3.1). Figure 3.6 shows the skill of *hindcasts* generated by $Y^p = Y^m(\hat{\theta}(D^C, \hat{\kappa}^C), \hat{\lambda}_j^C, t)$ (the corresponding naïve estimates of hindcast skill are all negative, showing that, as with the inductive model, the process model is overfitted to single years of data). To measure *forecast* skill, the validation year must be treated as a 'target' year in the future which happens to resemble one in the past, and for which the model must be given an independent set of forcing parameter estimates. Hence, the forecasts in Figure 3.4 were generated by: $Y^p = Y^m(\hat{\theta}(D^C, \hat{\kappa}^C), \hat{\lambda}_j^V, t)$. In forecasting, a greater penalty is incurred by inaccurate forcings, because errors occur independently in both assimilation and forecast intervals. Figs. 3.6a–c demonstrate that this is the main cause of the loss of skill in Figs. 3.4a–c.

Another likely source of variance in the process model predictions is the weak prior constraints imposed by the wide allowed parameter ranges in Table 3.1 (a measure to allow the model to fit the coarse-scale dynamics — see Appendix 3D). Table 3.3 shows that the estimated uncertainty in all parameter estimates was generally between 50 and 100%, even when all three years of chlorophyll data were used to constrain the model. Also, the ensemble parameter fits revealed a strong positive correlation (> 0.95) between the microzooplankton grazing rate $\hat{g}$ and the mesozooplankton feeding rate $\hat{f}$. Hence only six of the seven parameters were 'constrainable' by the (correlation < 0.8) criterion of Matear (1995). The main reason for this is surely the lack of microzooplankton data, which also results in very poorly constrained (at least 100% standard deviation) predictions of microzooplankton abundance (not shown).

Twin tests were also performed in which the 'true ensemble' variance of parameter estimates was calculated using the simulation model as a surrogate for truth. Then, an artificial 'original' data set was produced and used to fit a new data simulation model, which was subsequently used to produce 'simulated ensemble' estimates of 'true ensemble' parameter (co)variances as described in section 3.2.2. The simulated ensemble estimates were found to be much more accurate (20–30% error) than those given by the inverse Hessian method, which almost always underestimated the uncertainty. These tests also validated the simu-

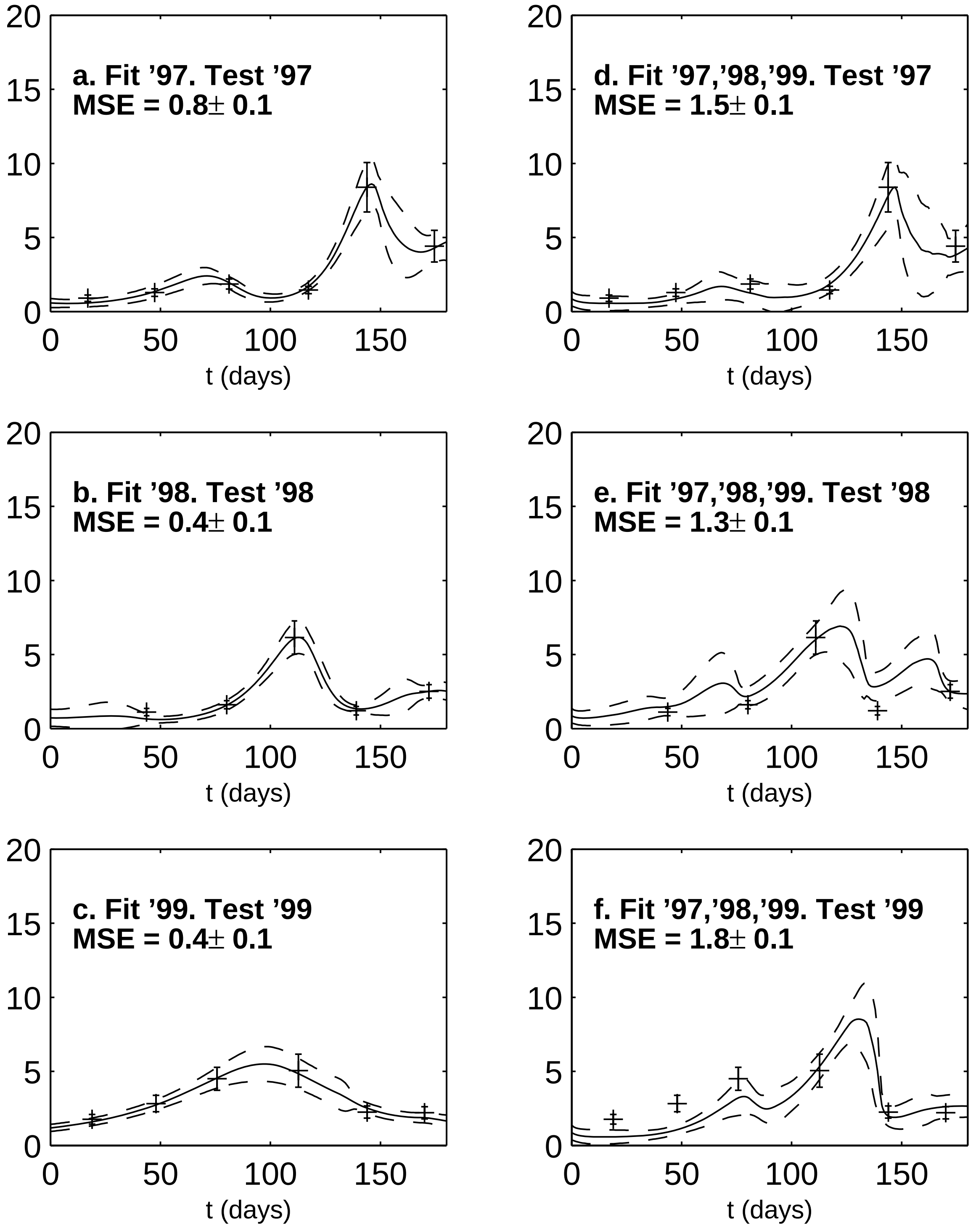

Figure 3.6: Process model *hindcast* skill tests: fitting to one year and testing against the same year (a,b,c), and fitting to all three years (d,e,f). Predictions are for 1997 (a,d), 1998 (b,e) and 1999 (c,f). Solid and dashed lines show prediction ensemble means ± 1 standard deviation. Error bars show the true values ± 1 ensemble standard deviation in the cruise-averaged data. Numbers in bold show time-averaged mean-square error (MSE) estimates based on the ensembles of simulation-division skill tests.

Table 3.3: Parameter estimates for the process model fitted to each year individually and to all three years simultaneously (ensemble mean ± one standard deviation)

S	$\hat{u}_0$	$\hat{k}_N$	$\hat{c}$	$\hat{g}$	$\hat{f}$	$\widehat{P(0)}$	$\widehat{Z_1(0)}$
1	0.8 ± 0.3	0.7 ± 1.2	0.09 ± 0.05	0.04 ± 0.04	0.03 ± 0.02	0.6 ± 0.3	2.0 ± 1.3
2	0.5 ± 0.2	0.3 ± 0.6	0.06 ± 0.03	0.05 ± 0.02	0.02 ± 0.01	0.7 ± 0.6	1.3 ± 1.4
3	0.3 ± 0.2	0.6 ± 1.1	0.13 ± 0.10	0.02 ± 0.01	0.01 ± 0.01	1.2 ± 0.2	0.4 ± 1.1
(1,2,3)	0.6 ± 0.4	1.2 ± 1.0	0.09 ± 0.07	0.21 ± 0.14	0.15 ± 0.11	0.9 ± 0.5	0.8 ± 0.9

lated ensemble (parametric bootstrap) estimates of bias, variance and trajectory-averaged mean-square error. Note that the simpler nonparametric method of bootstrap resampling (sampling with replacement from the original data set — see Efron and Tibshirani, 1993) used in Evans (1999) and herein for estimating uncertainty in $\gamma(t)$ (see section 3.2.5) was found to yield much poorer estimates of ensemble statistics for predictions and parameter estimates in this case.

The process model predictions also suffered from bias because of the model's inability to fully account for the interannual chlorophyll variability using the given interannual nutrient and mesozooplankton variability (underfit). This is best observed in the hindcast predictions (Fig. 3.6). Comparing the hindcasts for 1997 and 1998 where all three years are fitted (Figs. 3.6d,e), the model seems able to explain the earlier primary bloom in 1998 as a consequence of the earlier mesozooplankton increase (Figs. 3.5d,e) which suppresses the microzooplankton population (not shown), thereby releasing the phytoplankton. This hindcast skill is better if 1999 is excluded from the fit (Figs. 3.7a,b). Here, the sudden crash in phytoplankton between April and May 1998 is largely accounted for by a combination of microzooplankton grazing and nutrient limitation (the ensemble-mean nutrient saturation constant for these fits $\hat{k}_N = 0.74$ is comparable with the low nutrient concentrations in Fig. 3.5b).

However, the year 1999 is poorly accounted for by the model in hindcasting the three years (Fig. 3.6f). The model is unable to reproduce the early, gradual rise in the chlorophyll mean field during the first three months of 1999. In the single-year fit to 1999 (Fig. 3.6c), a good fit is achieved by assuming a lower initial microzooplankton population, slower grazing/feeding, and slower phytoplankton growth than in the previous two years, as shown by the fitted parameter statistics in Table 3.3. Looking at Table 3.3, the extra-sample bias of the process model might be anticipated from the significant differences in mean parameter estimates between different single-year fits (most obviously between 1997 and 1999). The extra-sample variance might also be anticipated from the large ensemble standard deviations in the parameter estimates, which are partly a result of the uncertainty in the forcings. When all three years are fitted, the mean estimates $\hat{u}_0$ can be understood as a compromise between the fast-growth conditions of 1997 and the slow-growth conditions of 1999. The model recovers as much as possible of the required interannual chlorophyll variability from the given interannual nutrient and mesozooplankton variability by increasing the half-saturation constant $\hat{k}_N$, grazing rate $\hat{g}$, and feeding rate $\hat{f}$ over any of the single-year results.

To investigate the source of underfit in 1999 when all three years are fitted, several extensions of the process model were fitted, relaxing some of the strong assumptions detailed in section 3.2.5. Since this is only a test for underfit, rather than an attempt to estimate model bias, variance or skill, an ensemble testbed is not necessary. A single fit to the full three

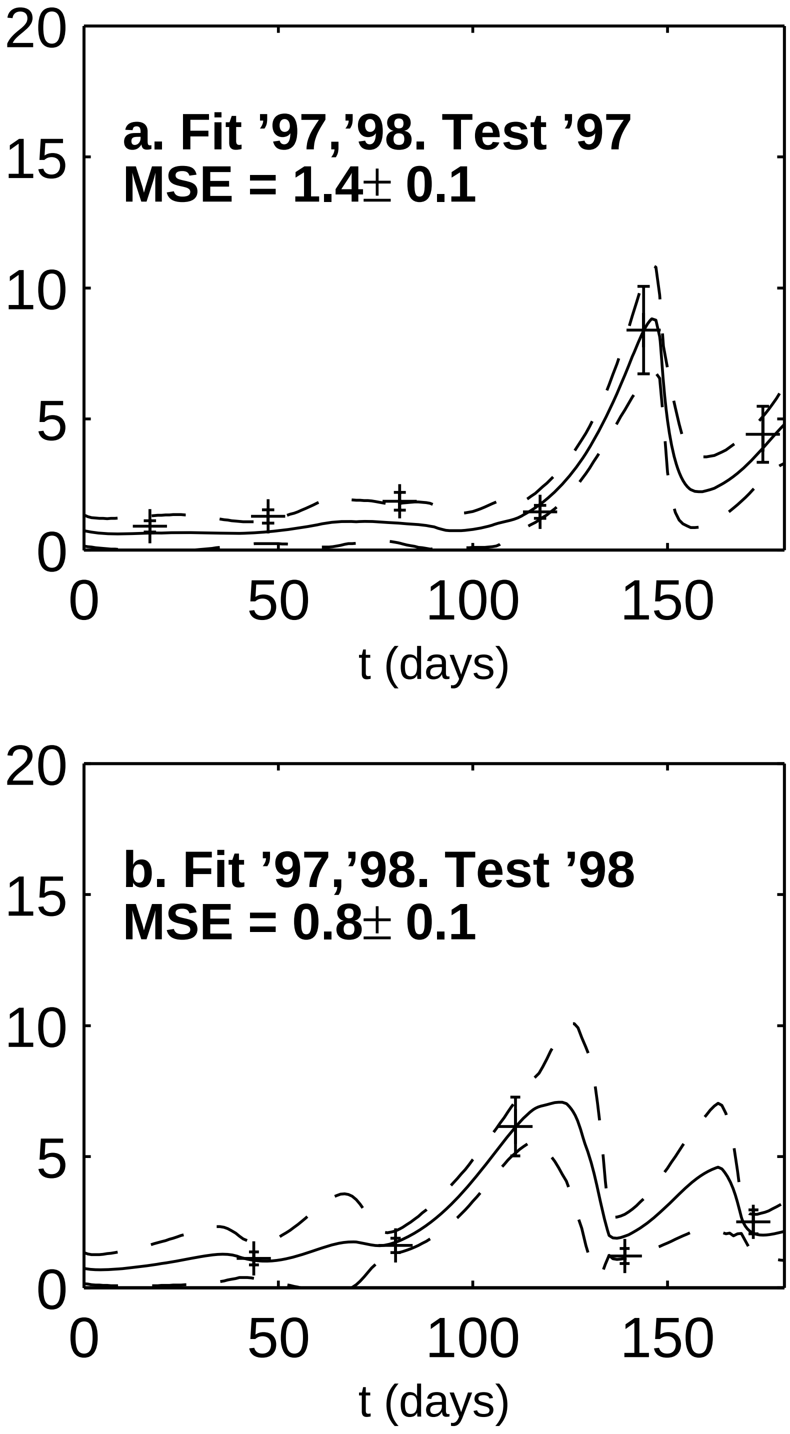

Figure 3.7: Process model *hindcast* skill tests, fitting to two years: 1997 and 1998. Predictions are for 1997 (a), and 1998 (b). Solid and dashed lines show prediction ensemble means ± 1 standard deviation. Error bars show the true values ± 1 ensemble standard deviation of the cruise-averaged data. Numbers in bold show time-averaged mean-square error (MSE) estimates based on the ensembles of simulation-division skill tests.

years of original (cruise-averaged) data does the job. Each model extension was tested in turn rather than in combination, as this latter would greatly increase the necessary optimisation time. The tested extensions include: interannually-variable mixed layer depth changes, interannually-variable initial conditions as model inputs, herbivory by mesozooplankton, a constant specific phytoplankton mortality rate, 2-parameter Ivlev grazing of phytotplankton by microzooplankton and mesozooplankton and Ivlev feeding on microzooplankton by mesozooplankton. Implicit resolution of two size classes of phytoplankton was attempted by assuming a small size fraction given by a saturating function of the total, consistent with the 'background/bloom hypothesis' (Denman 2003). Parameterisation of phytoplankton losses to the microbial loop was attempted using a function of the phytoplankton abundance scaled by a saturating function of available nutrient, as in Steele (1998).

None of these extensions relieved the serious underfit in the first three months of 1999, or indeed gave much improvement in overall data fit. Perhaps most surprising is the lack of support for mesozooplankton herbivory (the data fit consistently forces this grazing rate towards zero), since this is usually assumed to be the primary grazing loss for the large phytoplankton (diatoms) which dominate phytoplankton biomass on Georges Bank (Ji et al., 2006; Steele et al., 2007). Therefore the underfit is probably best explained by one or more of the following:

1) The zero-dimensional process model is incapable of capturing the mean-field behaviour on Georges Bank due to the effects of unresolved horizontal spatial variability, even allowing for substantial 'retuning' of model parameters.

2) The assumption of zero net boundary fluxes of chlorophyll on/off Georges Bank is invalid.

3) A combination of the individually-tested model extensions is required to improve data fit, such that any single extension on its own yields little improvement.

3.4.2 Comparative skill assessment and model selection

The skill tests revealed that the inductive model predictions tend to be quite biased, especially outside the calibration years (see Figs. 3.3d–f), due to the significant interannual variability in the mean chlorophyll field, and probably also between cruise periods, due to likely nonlinear variation between them. On the other hand, the predictions are at least robust and not very susceptible to overfit, due to the strong linearity assumption. As a result, estimates of the rms predictive error ($= (-Skill)^{1/2}$) of the inductive model were: 1.85 ± 0.05 mg chla m^{-3} fitting to all three years, 2.05 ± 0.03 mg chla m^{-3} fitting to two years, and 2.51 ± 0.03 mg chla m^{-3} fitting to individual years. Hence the uncertainty in the inductive model predictions for future years may be estimated as in the range 1.8–2.5 mg chla m^{-3}, on average over the period January through June.

The process model was able to make less biased three-year hindcasts than the inductive model, and generally made more skillful predictions when fitted to individual years, as long as 1999 is excluded (see Table 3.2). This suggests that (3.8) does at least partly capture the response of the chlorophyll mean field to changes in the nutrient and mesozooplankton mean fields. However, the process model forecasts suffer from a much greater variance compared to those of the inductive model, largely due to errors in model forcings which the predictive method allowed to perturb the biological model parameter estimates. This may be a rather general problem in marine ecosystem model-fitting (cf. Schartau and Oschlies 2003). In addition, the process model was clearly unable to fit (underfitted) the data during January–March 1999, indicating a failure of the model assumptions. The net effect was that the process model rms predictive error estimates were only better than those of the inductive model in one-year tests: 2.26 ± 0.09 mg chla m^{-3} fitting to all three years, 2.24 ± 0.05 mg chla m^{-3} fitting to two years, and 2.29 ± 0.05 mg chla m^{-3} fitting to individual years. Hence the average uncertainty in the process model predictions for future years may be estimated as in the range 2.2–2.3 mg chla m^{-3}.

In section 3.2.2, an optimal skill metric $\widehat{Skill}^{sd(A_C^*)}$ was defined as that which optimally gauges the *Skill* defined by (3.1) using tests within the full sampled set of model inputs. However, there is no way of determining the size of the optimal calibration subset A_C^*. This must rely on intuition, informed by any available information on the likely distribution of model inputs for the target years. For example, if the nutrient and mesozooplankton abundances in the future target years were expected to lie roughly within the ranges observed during 1997–1999, $A_C^* = A_S$ might be favoured as an optimal gauge, and hence the inductive model chosen. Alternatively, nutrient and/or mesozooplankton variability might have changed persistently over the ten years since the data were collected, e.g., due to changing fishing practices. In this case, substantial extrapolation is required of the model, so $A_C^* = 1$ might be favoured as a representative test, hence the process model is chosen.

In lieu of target year information, however, the inductive model might be chosen to predict the mean chlorophyll variability in future years, since the metrics based on simulation and division usually gauge it as having higher *Skill* than the process model in its current form. By contrast, a naïve skill assessment based on calibration *cost*, which ignores prediction variance due to data and forcing errors, would strongly favour the process model on the basis of the three-year fits (measuring an rms predictive error or uncertainty of 1.76 ± 0.20 mg chla m^{-3} and 0.79 ± 0.26 mg chla m^{-3} for the inductive and process models respectively). A division-only or 'cross-validation' metric similar to that used in Friedrichs et al. (2007) gave skill estimates consistent with our simulation-and-division metric, but which were not robust enough to distinguish between the two models.

Finally, note that even though the process model failed to consistently outperform the inductive model in terms of prediction, it revealed more about the mean-field chlorophyll dynamics on Georges Bank than the strictly-empirical inductive model, by allowing mechanistic hypotheses to be tested. Also, the successes and failures of the process model suggest future developments to improve process model predictions (e.g., more robust forcings, accurate net boundary flux inputs). By contrast, the inductive approach is something of a modelling cul-de-sac: it gives no understanding, and its predictions can only get more robust as more data becomes available, never addressing the bias due to interannual variability which it cannot account for.

3.4.3 Potential Improvements

In terms of data, the objective of finding representative skill tests was compromised by the small number of years ($A_S = 3$) for which nutrients, phytoplankton and mesozooplankton had all been sampled. For the years which were sampled, the data constraint would probably have been stronger if, practical considerations aside, samples had been taken more evenlyspaced in time. Sampling for the whole year rather than just the first half would obviously aid constraint, although whether the same number of cruises distributed over the whole year would improve matters is an open question. In terms of spatial distribution, the samples were very irregularly-spaced over the bank (Fig. 3.1), and a more uniform distribution would probably have been better for mean field estimation. Measurements of microzooplankton concentrations would greatly improve data constraint, especially of microzooplankton parameters. Remotely-sensed data might also be blended with the in situ data to improve estimates of surface chlorophyll concentrations.

A better mean field estimation method might be to fit a spatial trend surface to the cruise data and integrate over it (Thiébaux and Pedder, 1987), perhaps after correcting for asynopticity using an 'advective correction' as in McGillicuddy et al. (2001). Or, the simple two-step, unweighted cruise-averaging used in this study could be replaced with an optimal weighted average based on fitting spatio-temporal covariance parameters to the raw data. Alternatively, the raw data might be treated as estimates of the mean field with error correlations accounted for by extra statistical parameters within a Maximum Likelihood calibration method (e.g. Stock et al., 2005).

Variance in process model predictions may by reduced by narrowing the allowed ranges of adjusted parameters, but this may be at the expense of greater prediction bias, resulting in no improvement in prediction accuracy. One strategy to reduce variance without increasing bias is *a priori* fixing of 'insensitive' parameters which have little effect on the fit to calibration data when varied by small amounts. This does not guarantee, however, that their variation has little effect on extra-sample bias. Also, this 'sensitivity' would be better defined 'globally' over the allowed volume of parameter space, rather than locally along 1D 'transects' as is common practice (e.g. Garcia-Gorriz et al., 2003; Friedrichs et al., 2007). An alternative might be to fix parameters *a priori* as the values measured in laboratory experiments. Unfortunately, there is a general lack of relevant, multispecies laboratory 'microcosm' or field 'mesocosm' experiments to constrain the parameters of any species-aggregated model (indeed, this might be an argument in favour of using less aggregated models). A compromise might be to allow moderate retuning (e.g. ±30%) of parameter values based on semi-relevant laboratory studies and previous model fits.

Note that Bayesian methods of estimating *a posteriori* uncertainties may be computationally cheaper than the simulated ensemble (parametric bootstrap) approach, if an efficient way of integrating over the posterior distribution can be found (e.g. Harmon and Challenor, 1997). However, these should first be validated by a realistic twin test (see section 3.4.1).

As regards model formulation, the inductive model could be improved with a more sophisticated interpolation method (e.g., splines, Gaussian processes, etc.) but the basic predictive properties would probably not differ significantly from those measured herein. For the process model, it is possible that a reformulation of the biology, such as the explicit resolution of large and small phytoplankton types and use of silicate data, as in Ji et al. (2006), could address the serious underfit problem for 1999 and perhaps improve model skill (if model-noise covariance does not increase too much with the increased number of parameters). Prediction variance due to forcing uncertainty might be reduced by dynamically modelling other trophic levels (e.g. NPZD as in Ji et al., 2007) if the resulting extended model uses better-known forcing inputs. As noted above, bias in the process model predictions may be the result of neglected net boundary fluxes (see section 3.4.1). This could potentially be remedied by embedding the process model within a larger-scale model which provides accurate input boundary fluxes.

Of particular interest is the effect of increased spatial resolution. Several studies of the Georges Bank ecosystem have employed high-resolution dynamic circulation models (Ji et al., 2006, 2007) and model-derived climatological fluid transports (McGillicuddy et al., 1998, 2001; Li et al., 2006). It would be interesting to use the same Georges Bank data to fit e.g. an NPZD model coupled to one of these circulation models, and compare the skill of mean field predictions, accounting for all sources of error, with that of the equivalent bulk model. This study demonstrates a general methodology for making such a comparison.

3.5 Conclusions

There is a pressing need to objectively establish and inter-compare the extrapolative skill of mechanistic marine ecosystem models and prediction methods. When computational expense is not too restrictive, the following is a useful method:

1) Define a 'skill function' to specify the precise predictive objectives. Skill assessment is then focussed on estimating this skill function.

2) Find a statistical model for simulating repeat samplings of the data set, and fit it to the data. This 'data simulation model' need only simulate repeat samplings with reasonable accuracy and need not interpolate or extrapolate accurately beyond the data. It should not therefore be limited by ecological assumptions.

3) Use the fitted data simulation model to produce an 'ensemble testbed' of repeat data sets for the predictive models. For each new simulated data set, recalculate estimates of forcings, initial/boundary conditions and adjustable model parameters in keeping with prior uncertainty, and recompute predictions. The resulting distributions of predictions, and the truth estimated by the simulation model, allow estimation of the 'bias', 'variance' and 'mean square error' of model predictions in the classical statistical modelling sense. Restricting the predictive model calibration to subsets ('divisions') of the full data set allows skill tests which are representative of the inter-/extrapolative task defined by the skill function. Skill metrics defined over this ensemble testbed can then be used to estimate the skill function, and hence the relative skills of the different predictive models or prediction methods.

This method was applied to estimate the mean-square predictive error of two models for predicting the mean chlorophyll variation on Georges Bank in future years, based on data from the GLOBEC sampling program in the years 1997–1999. The first 'inductive' model was a simple piecewise-linear interpolation of chlorophyll data, the second a simple bulk process model assuming Lotka-Volterra grazing by an unsampled microzooplankton population and interannually-variable forcing by nutrient and mesozooplankton mean fields.

The ensemble skill tests showed that the process model could account for some of the interannual variability in mean chlorophyll concentrations. However, errors in forcing estimates caused high variance in model forecasts, due to distortion of the fitted biological parameters in calibration, and further model input error for the forecasted year. This variance was exacerbated by the weak prior constraints on biological model parameters, which may be necessary for fitting coarse-scale, mean-field plankton dynamics. Uncertainty on most parameters was 50–100% even when all three half-years of chlorophyll data were fitted. Two of the seven parameters could not be independently constrained, due to the lack of microzooplankton data. In addition, the process model was consistently unable to fit the data in the first three months of 1999, possibly due to failure of the zero net boundary flux assumption.

Consequently, the skill of the process model was estimated as generally worse than the inductive model, despite strong bias in the latter's predictions due to interannual variability (rms prediction error or uncertainty of 2.2–2.3 mg chla m^{-3} for the process model compared with 1.8–2.5 mg chla m^{-3} for the inductive model).

By contrast, a naïve skill assessment based on calibration cost for fitting all three years gave a qualitatively wrong result, strongly favouring the process model (which can fit the data better) mainly because of the neglect of prediction variance. The use of data division without simulation produced skill metrics consistent with the above, but the number of available divisions was insufficient to allow robust estimation of skill and of bias and variance as functions of time.

Appendix 3A: Bias in naïve skill metrics based on fitted data misfit

Consider estimating the mean-square error $MSE = \sum_i E[(Y_i^t - Y_i^p(D))^2]$ using the naïve estimate $\widehat{MSE}^n = \sum_i (O_i - Y_i^p(\hat{\theta}(D)))^2$, where the model used to produce Y^p is fitted to the data $D = \{O_i\}$ via the adjustable parameter set θ. Unbiased data are assumed, so that $O_i = Y_i^t + \epsilon_i$ where $E[\epsilon_i] = 0$. The bias in $\widehat{MSE}^n$ is then given by:

$$E[\widehat{MSE}^n] - MSE = \sum_i E[\epsilon_i^2] - 2\sum_i E[Y_i^p(\hat{\theta}(D))\epsilon_i] \tag{3A.1}$$

The first term on the RHS is just the sum of observational error variances, which can often be estimated (and subtracted from $\widehat{MSE}^n$) from knowledge of the measuring apparatus and data processing method, or empirically from the data (see Appendix 3B). The second term is a sum of covariances between the model prediction and the fitted data errors. This term is harder to estimate, but is generally positive and increases with model complexity (dimension of θ). If the size of this covariance is judged to be excessive, the model may be said to be 'overfitted'.

Appendix 3B: Data simulation model

The data simulation model represents our best efforts to simulate repeat samplings of the original data set. It need not be designed to provide estimates of mean fields beyond the sampling times and years specified by the original data, and therefore requires a different skill function and different skill metrics to those discussed in the text (see below).

The simulation model consists of statistical distributions, so to assess its skill the accuracy both of mean values and variance parameters must be considered. But how should the relative importance of means vs. variances be weighted in a skill function for the data simulation model? One answer is to consider the *Likelihood* (L), defined as the probability of the data given the fitted model with parameters θ^{sim} ($L(\hat{\theta}^{sim}|D) \equiv P(D|\hat{\theta}^{sim})$, with the conventional use of '|' to denote 'given'). However, the data used to evaluate this Likelihood should be independent of the calibration data, in order to negate the covariance effect discussed in Appendix 3A. Hence the following skill function should be maximised:

$$Skill^{sim} = E[\log L(\hat{\theta}^{sim}(D^C)|D^V)] \tag{3B.1}$$

where D^C and D^V are independent calibration and validation data sets respectively, and the expectation E is over the product of the pdfs of D^C and D^V (equivalently, the 'information loss' due to approximating the truth with the statistical model is to be minimised — Akaike, 1973). An estimate of (3B.1) might be obtained by dividing the data into independent calibration and validation sets, but in this case (as is fairly typical) there are insufficient data to make this a representative test (for some months there are only 3 nutrient samples from which to estimate a mean value). Fortunately, however, (3B.1) only requires intra-sample accuracy (accuracy 'at the data'), and the data simulation model is by-design capable of fitting all the data reasonably well. Therefore, analytical results can be used to correct for the noise-fitting bias incurred by using $L(\hat{\theta}^{sim}(D^C)|D^C)$ to estimate (3B.1). Specifically, the following skill metric is used:

$$\widehat{Skill}^{sim} = AIC_c = \log L(\hat{\theta}^{sim}(D^C)|D^C) - \frac{nK}{n-K-1} \tag{3B.2}$$

where n is the total sample size, K is the number of fitted model parameters, and AIC_c denotes the *Akaike Information Criterion* (Akaike, 1973) in its second-order form (Hurvich and Tsai, 1989; Burnham and Anderson, 1998).

The candidate data simulation models consist of stationary probability distributions for each variable and each month, with zero correlation between diel-averaged samples of the same variable on different days and between different variables on the same day. Looking at the data (Figure 3.2), it is clear that the variance increases with the mean value for each month for all variables. This suggests either lognormal distributions (cf. Campbell, 1995) or normal distributions with standard deviations proportional to the mean value (cf. Hurtt and Armstrong, 1996) and a single truncation point at zero. Given the small number of samples for each month (<10), it seems plausible to fit either separate variance parameters for each year ($\approx$ 30 samples per year per variable), or one variance parameter for all three years ($\approx$ 90 samples per variable). Hence, for each variable P, N, Z_2, there are four candidate models corresponding to different combinations of distributional form and number of variance parameters.

Before fitting the candidate models, outlier data common to outlier-screenings by all four candidate models are removed (see Appendix 3C). Having committed to the skill function (3B.1), it is logical to then use Maximum Likelihood Estimators (MLEs) to fit the data simulation model parameters. For the non-truncated lognormal distributions, the necessary mean and variance in log space parameters (μ_i, σ_i^2) for month i are estimated using the MLEs (m_i, s^2) where m_i is a sample mean in log space of the data in month i, and s^2 is the sample variance in log space calculated over the data in one or all three years, depending on the candidate model. For the truncated normal distributions with means μ_i and coefficients of variation α, the truncation at zero prohibits closed analytical expressions for the MLEs of (μ_i, α^2). Therefore, sample statistics (m_i, a^2) are used, where a^2 is the variance of the

Table 3B.1: Relative skill metric ΔAIC_c for each variable and each candidate data simulation model, assuming lognormal or normal (with truncation at zero and constant coefficient of variation) distributions with one variance parameter for all three years ($n_\sigma = 1$) or one variance parameter per year ($n_\sigma = 3$). p-values for a Monte-Carlo version of the chi-squared hypothesis test are shown in parentheses for each model.

	Lognormal		Normal	
Variable	$n_\sigma = 1$	$n_\sigma = 3$	$n_\sigma = 1$	$n_\sigma = 3$
P	0 (0.26)	-0.7 (0.08)	-0.1 (0.68)	-1.6 (0.13)
N	0 (0.28)	-0.2 (0.36)	-0.9 (0.01)	-3.3 (0.02)
Z	0 (0.29)	-2.0 (0.15)	-8.4 (0.15)	-11.8 (< 0.001)

data errors scaled by the appropriate mean estimates, averaging over one or three years as required. The distribution means and variances ($\mu_i, \sigma_i^2 = \mu_i^2\alpha^2$) are then written as functions ($f_i(\tilde{\mu}_i, \tilde{\sigma}_i^2, x_{min}, x_{max}), g(\tilde{\mu}_i, \tilde{\sigma}_i^2, x_{min}, x_{max})$) of the mode $\tilde{\mu}_i$ and non-truncated variance $\tilde{\sigma}_i^2$ parameters, for given truncation points $x_{min} = 0$ and $x_{max} = \infty$. These functions are inverted numerically by Gauss-Newton iteration to yield the parameters required to reproduce data sets ($\tilde{\mu}_i, \tilde{\sigma}_i^2$), using the sample estimates ($m_i, s^2 = m_i^2 a^2$) of ($\mu_i, \sigma_i^2 = \mu_i^2\alpha^2$).

Having fitted the models, the relative skill metric ΔAIC_c is calculated with respect to the maximum value across each of the four alternatives (see Table 3B.1). Note that any sub-optimality in the Likelihood for the truncated normal fits due to the lack of MLEs is neglected. The results suggest that the lognormal distributions with one variance parameter for all 3 years offer the best skill for all variables. Without the noise-fitting bias correction term in (3B.2), the lognormal distributions with 3 variance parameters would have been selected, resulting in higher $L(\hat{\theta}^{sim}(D^C)|D^C)$ but poorer skill as defined by (3B.1).

In order to check that the chosen model is a good fit, a Monte-Carlo generalisation of the classical chi-squared hypothesis test is performed as follows. First, a null distribution is calculated (assuming the data simulation model is true) of total squared deviation of the residual histogram (obtained from fitting to a single artificial data set) from the expected residual histogram (Pearson statistic). Then the Pearson statistic is calculated for the original data and the percentage of all values in the null distribution more extreme than the value given by the original data is found, yielding the p-values shown in Table 3B.1. The results indicate that the lognormal, single variance parameter models selected by the AIC_c metric are a reasonable fit to the diel-averaged data (Figure 3.2), since the relevant p-values are well above the significance threshold of 0.05. By contrast, the normal distribution with 3 variances, identified by AIC_c as the worst model for all variables, is rejected by the hypothesis test as having produced the original N and Z_2 data. This gives some confidence that the most skillful data simulation model as defined by (3B.1) has been selected.

As well as providing an ensemble testbed for the inductive and process models, the data simulation model allows estimation of confidence intervals for the monthly mean field estimates in the data simulation model, by computing the distribution and variance of these estimates over the simulated ensemble. Estimates for the error variances of the cruise-averaged data used to fit the predictive models (see section 3.2.2) are similarly obtained. These two error variances were broadly equal to 1 decimal place, reflecting the fact the Maximum Likelihood mean assuming lognormality and the sample mean are equally skillful estimators of the true mean fields for the observed means/variances and sample sizes of 3–9.

Appendix 3C: Outlier data

It is useful to distinguish two different types of outlier. The first is 'bad data' in the truest sense, as may occur due to e.g. apparatus malfunction or human recording error. One such datum was identified in the raw data set as the surface nitrate concentration of 0.6 μ M in June 1999, which was exactly a factor of 10 larger than the 9 similar measurements made in subsequent days, and a factor of 5 larger than the next largest measurement in preceding days. This datum was discarded at the outset after it was found to spoil the system of outlier detection described below, because it distorted the variance parameter estimates for the year(s) of data errors by such a large amount (see Appendix 3B). Similarly, all zero measurements of chlorophyll and mesozooplankton abundance were assumed to be the result of instrument malfunction or human error, and discarded them from the analysis. For a valid nutrient (=nitrate+nitrite+ammonium) measurement, both (nitrate+nitrite)>0 and ammonium$\geq$0 were required, since there were in fact a few negative readings for each! Zero measurements of ammonium were deemed usable because there were a large number of them and ammonium concentrations were generally very low over the sampled period.

The second type of outlier is not necessarily a 'wrong' measurement, but represents a rare extreme event (significant at the 5% level) in an ensemble of repeat samplings of the diel-averaged data simulated by the data simulation model. In general it is predictively expedient (reduces prediction variance) to discard such data ('All's fair in love and prediction'). However, having discarded such an outlier, the data simulation model must be refitted, threshold extreme data errors recalculated, and the data retested. The process is reiterated until no new outliers are found.

Applying this procedure to the diel-averaged data using the most skillful data simulation model (see Appendix 3B), one outlier datum is identified for each variable: a low P value of 0.19 mg chla m^{-3} during June 1999, a low N value of 1.07 μ M during April 1997, and a high Z value of 35.6 cc/10m^3 during June 1997. These data were therefore removed from the data set used for fitting the simulation model and testing the models discussed in the text. For testing the data simulation models (see Appendix 3B), only those data identified as outliers by all data simulation models were discarded — those being the above values for P and Z only.

Appendix 3D: Allowed Parameter Ranges for the Process Model

The allowed range of each parameter had to encompass the values adopted in previous studies using simple biological models for the Georges Bank ecosystem (Franks and Chen, 2001; Ji et al., 2006; Ji et al., 2007). However, the larger scale of spatial averaging in this study was accounted for by allowing substantial 'retuning' of parameters perhaps formerly derived from measurements of laboratory-scale averages (this is a 'weak prior mean-field model'). Usually, larger-scale growth/grazing/feeding rates are slower due to negative spatial covariance between consumer and resource arising from the action of consumption itself (see Chapter 4).

Therefore, the minimum allowed values of these rate parameters are reduced. For the same reason, the maximum allowed value of the nutrient half-saturation concentration is increased from $\sim 1\mu M$ in previous studies to $3\mu M$. The allowed range for the self shading parameter $c \equiv k_p/k_e$ is based on assuming $0.02 < k_p < 0.06$ (mg chla m^{-3})$^{-1}$m^{-1} and $0.1 < k_e < 0.2$ m^{-1} (consistent with Ji et al., 2007), which implies a 'safe' range of $0.05 < c < 0.3$ (mg chla m^{-3})$^{-1}$.

The allowed range of carnivory rates $0.001 < f < 0.500$ (cc/(10m^3))$^{-1}$ day^{-1} is consistent with Ji et al. (2006), using the ratio of January concentrations to convert the units ($Z_2(Jan) \approx$ 1.7 biovolume units in this study, whilst $Z_2(Jan) \approx 0.2\mu$M N in Ji et al. (2006)). The allowed range of initial chlorophyll concentration $0.01 < P(0) < 1.50$ mg chla m^{-3} is consistent with the mean field estimates for January 1st obtained by linearly extrapolating the trends over the first two sampled months (0.7 mg chla m^{-3} in 1997, 0.5 mg chla m^{-3} in 1998 and 1.1 mg chla m^{-3} in 1999).

The maximum allowed value of the scaled microzooplankton grazing rate $g = \alpha\beta g' =$ 0.5 (mg chla m^{-3})$^{-1}$day^{-1} assumes a maximum allowed unscaled grazing rate $g' = 0.5$ $(\mu MN)^{-1}$day^{-1} (consistent with Ji et al., 2006), a maximum plausible assimilation efficiency $\alpha = 0.8$ and a maximum plausible unit conversion factor $\beta = 1.2$ (mol N)/(g chla) (using a maximum Redfield molar N:C ratio of = 1/5.5 and a maximum mass ratio C:chla = 80 from Townsend and Thomas (2002)). Similarly, the scaled microzooplankton initial condition $Z_1(0) = Z_1'(0)/(\alpha\beta)$ was based on a plausible maximum value $Z_1'(0) < 0.50$ μM N (consistent with Ji et al., 2006) and plausible minimum values $\alpha = 0.2$ and $\beta = 0.4$ (mol N)/(g chla) (using a minimum Redfield molar N:C ratio of = 1/6.5 and a minimum mass ratio C:chla = 30 from Townsend and Thomas (2002)).

Appendix 3E: Optimisation algorithm

The *cost* in (3.11) is minimised by varying the adjustable parameters a (which may include initial conditions) using the Downhill Simplex with Simulated Annealing optimisation schedule of Press et al. (1999) with a fixed number of iterations and an exponential cooling schedule (a similar method to that of section 2.2.7). The parameter search was restricted to a 'reasonable' volume of parameter space by adding to the *cost* in (11) a penalty (or an '*a priori* bias' term) of 10^5 cost units for every trial parameter value outside its allowed range. Although this is a 'non-greedy' parameter search algorithm (i.e. one that takes occasional uphill steps to avoid trapping in local minima), single optimisations still tended to yield local minima solutions due to the finite cooling time, and therefore exhibited dependence on the initial parameter guess a_I. To reduce this dependence, optimisations each of length n_{it} iterations were repeated for n_I randomly chosen a_I for each data set and the solution with the lowest *cost* was selected. The a_I were chosen by *Latin Hypercube Sampling* from the allowed volume of parameter space V_a with uniform probability (see Press et al., 1999). This ensured that a reasonable coverage of V_a is achieved with a much smaller number of samples, by preventing samples having any of their parameter values in the same segment of the allowed range (where no. of segments = $5n_I$). When an initial guess resulted in a very high initial cost ($cost_I > 10^5$) the random selection was repeated.

Each optimisation required input of initial and final temperatures (T_I and T_F) and number of iterations n_{it} as cooling schedule parameters. Hence values of four 'optimisation parameters' including the number of initial guesses $\{T_I$, T_F, n_{it} and $n_I\}$ needed to be chosen to maximise fitting performance. First, the parameters n_{it} and n_I were varied for a fixed product $n_{it}n_I$, which determined the overall run-time, for a reasonable choice of T_I and T_F, and the mean costs achieved over an ensemble of 5 different simulated data sets were compared. It was found that the best choice for a given run-time involved a trade-off of cooling slowness and number of restarts. For the process model fit: $\{n_{it} = 1 \times 10^4$, $n_I = 50\}$ yielded best results; however for the 'less non-linear' regression problems of the inductive model fit, fewer repeats were necessary and $\{n_{it} = 2 \times 10^4$, $n_I = 3\}$ gave good fitting performance. Then $\{T_I$, $T_F\}$ were varied for fixed $\{n_{it}$, $n_I\}$. The process model fit generally required higher temperatures than the piece-wise linear regressions: $\{T_I = cost_I$, $T_F = 0.1\}$ gave good results for the process model fit, whilst $\{T_I = 1$, $T_F = 0.1\}$ was better for the piecewise-linear fits. The schedule was tested by repeating the optimisation over an ensemble of 50 *identical* data sets, varying only the set of initial guesses between each data set. The algorithm found the same optimal solution in all 50 cases for piecewise-linear and process model fits.

Chapter 4

Spatially-Implicit Modelling

Accepted article: Wallhead, P.J., Martin, A.P., Srokosz M.A. Spatially-Implicit Plankton Population Models: Transient Spatial Variability. J. Theor. Biol.

4.1 Introduction

Plankton population models have had a degree of success in explaining the dynamics of laboratory cultures (Fussman et al., 2000; McCauley et al., 1999). In such experiments, the plankton are kept *well mixed* by stirring and zooplankton swimming. This means that they move or are moved around the experimental volume swiftly and randomly enough that they experience the volumetric average concentrations of nutrients, prey etc. over their generation timescales (Dieckmann et al., 2000). Consequently, spatial variability within the laboratory experiment has by-design a negligible influence on the population dynamics.

At slightly larger scales (tens of metres or less) in the ocean surface mixed layer, turbulent currents might be expected to do a good job of creating well-mixed conditions. Tracer experiments (Okubo, 1971) suggest an order-of-magnitude turbulent 'eddy-mixing' timescale of between 0.01 and 1 day for mixed layer volumes of scale 10m. Plankton populations with typical generation timescales of 1 day are therefore likely to be well-mixed on these scales.

A crucially important feature of the oceans, however, is that they mix material slower over larger spatial scales (Okubo and Levin, 1980). In fact, Okubo (1976) proposes a mixing rate $\lambda_a \approx 120L^{-0.85}$ day^{-1}, plus or minus an order of magnitude, for an ocean scale of L m (obtained from the Okubo (1976) relation: $K(L) = 0.068L^{1.15}$ cm^{2}s^{-1} for L cm, by dividing by L^2 and converting the units). Consequently, at scales larger than $L = 10$ m it is not generally safe to assume well-mixed plankton populations. A plankton population that is not well-mixed may not obey the same dynamics as a well-mixed one, because of dynamical *nonlinearity*: laboratory population growth rates usually do not depend just on population size but also on (population size)^2 and perhaps higher powers. Dynamical nonlinearity acts on spatial variability to distort the coarse-scale (mean-field) population dynamics. This may lead to failure of the *mean field approximation* (MFA), whereby the population dynamics are extrapolated from the fine to the coarse scale. For plankton populations, this failure could be very serious, as demonstrated for a simplified plankton model by Brentnall et al. (2003) (hereafter B03). Failures of the MFA could be endemic in plankton modelling, where well-mixed laboratory population dynamics are often assumed to hold true over ocean model grid cells ranging in scale from 100 m in regional models to 100 km in global ocean models.

The most obvious, and currently prevalent, method of accounting for fine-scale variability in plankton ecosystems is to increase the explicit spatial resolution of the model (number of grid cells). However, this places heavier demands on data and model inputs (forcings, initial/boundary conditions) to accurately constrain the extra degrees of freedom when fitting the model and making predictions. Moreover, the added computational burden may impose practical restrictions e.g. on the spatio-temporal range of the model run, or the number of times it may be re-run with different parameter values, which in turn may limit our ability to make, and assess the robustness of, model predictions for different scenarios e.g. changes in climate.

An alternative, long-familiar to the turbulence modelling community, is to model higher moments of the unresolved variability as well as the mean fields (Reynolds, 1895). Hence, fine scales are resolved 'implicitly' in terms of their effect on larger scales. Over recent decades, a host of 'turbulence closure' models, themselves derived from models used in engineering, have been applied to modelling the ocean mixed layer and horizontal boundary layers (see Sander (1998) and Umlauf and Burchard (2005) for general reviews, or Carniel et al. (2007) for an oceanographic summary). These account for the effect of advective nonlinearity acting on unresolved variability in physical/biological fields to produce mean turbulent fluxes, also known as the 'transport Reynolds terms' (Lévy, 2006). Similar techniques have also been applied to coarse-resolution large-scale ocean models to parameterise the effects of unresolved mesoscale eddies (e.g. Gent and McWilliams, 1990; Gent et al., 1995). Modification of the standard eddy-diffusivity approach to parameterising turbulent fluxes to allow for the finite lifetime of reactive scalars such as plankton has been suggested by Pasquero (2005).

However, apparently no attempt has been made to implicitly resolve the statistical effect of *biological* nonlinearity on unresolved variability in plankton ecosystems, i.e. the 'biological Reynolds term' (Lévy, 2006). This is surprising, given that the biological dynamics usually involves nonlinearity that is of higher order than the quadratic nonlinearity of advective transport, and the spatial variability in plankton concentrations seen in data is usually very high (e.g. Martin et al., 2005), suggesting that the size of these neglected biological Reynolds terms may be very significant.

In this chapter, simple two-component, 2D 'reaction-diffusion' plankton models are used as high-resolution simulations (HRSs). The grid-cell scale populations react via a set of continuous-time, deterministic biological dynamics, identical in every grid cell, and interact with nearest-neighbour cells by standard linear mixing, representing eddy-diffusion. Spatial heterogeneity is only provided by variable initial conditions, and decays due to mixing and transience of the fine-scale biological dynamics. The accuracy of extrapolating the fine-scale biological dynamics to the domain-scale mean-field dynamics (the MFA) is investigated, as well as the possibility of improving on this approximation with implicit spatial resolution (ISR) models. These include 'spatial moment closure' (SMC) models, which assume central dominance of the phase space distribution, and a '2-spike approximation' (2SA), which assumes strong bimodality. The biological models and parameter sets are chosen to explore different strengths of biological nonlinearity, and mixing rates are varied to explore a range of scales relevant to practical ocean model grid cells.

Moment closure models have been used extensively in recent years in terrestrial ecology (e.g. Bolker and Pacala, 1997; Dieckmann et al., 2000; Keeling, 2000a,b; Keeling et al., 2002;

Law and Dieckmann, 2000a,b; Law et al., 2003; Lewis and Pacala, 2000; Murrell et al., 2004; Ovaskainen and Cornell, 2006; Pascual and Levin, 1999) and terrestrial epidemiology (e.g. Ferguson et al., 2001; Filipe and Gibson, 2001; Filipe and Maule, 2003). Readers familiar with this literature should note that the problem treated here is different on several accounts. First, as argued above, plankton populations may be considered well-mixed in the ocean at the laboratory scale ($<$ 10 m). Therefore, rather than starting with an individual-based model (IBM), we start with a population model for the laboratory scale, which then needs to be modified somehow to work on the (much larger) ocean grid cell scale.

Second, land plants are tethered to a solid substrate and act spatially only when they reproduce, dispersing seeds or pollen broadly via air or animal movement on a timescale much shorter than that of plant growth. By contrast, phytoplankton are fully committed to their liquid environment: they are shifted around continuously by ocean currents ('plankton' deriving from the Greek for 'drifter'), and reproduce locally.

Third, in contrast to terrestrial systems, marine primary producer/consumer populations are generally large ($>$ 100 individuals) at the scale of the laboratory, at least if the consumers are microzooplankton. This suggests that *deterministic* models may be applicable to phytoplankton/zooplankton population dynamics (Renshaw, 1991). If the zooplankton are interpreted as mesozooplankton, which are only moderately-numerous (O(10–100)) at the laboratory scale, it may be necessary to include demographic stochasticity at the level of fine-scale populations (Vainstein et al., 2007). The methods used in this chapter may be generalised to account for this (e.g. Rodriguez and Tuckwell, 1996).

Hence deterministic reaction-diffusion equations are used as 'simulation' models in this study. This is not to deny that stochastic IBMs with local (Martin, 2004) and non-local interactions (Birch and Young, 2006; Flierl and Grünbaum, 1999; Hernández-García and López, 2004) may be necessary for deriving population dynamics at higher trophic levels — perhaps mesozooplankton and higher. However, simulating realistic numbers of individual mesozooplankton in even a single ocean model grid cell may be computationally unfeasible. IBMs might be used to derive fine-scale, deterministic population behaviour using 'moment closure' as understood in terrestrial applications (see next paragraph). However, there remains the problem of extrapolating the fine-scale deterministic population dynamics to coarser scales for ocean models. An alternative approach might be to let simulated 'super-individuals' or 'agents' each represent large numbers of real individuals (Scheffer et al., 1995; Batchelder et al., 2002; Woods, 2005).

Consequently, the use of moment closure in this study has a more 'classical' flavour. In terrestrial ecology, it has been used to approximate the average behaviour of a stochastic IBM over a computationally-expensive 'ensemble' of runs. The result is typically a closed set of partial integro-differential equations for the first moments (as functions of time) and second moments (as functions of time and spatial lag), derived using 'closure assumptions' to approximate the effects of third-order moments on the dynamics of the second moments (Murrell et al., 2004). By contrast, in the problem treated here, little is gained by just transforming from one set of finely-resolved, partial differential equations (PDEs) in (time, real space) to another set of finely-resolved PDEs in (time, lag space), even if the higher-order moments can be successfully approximated. However, it is shown herein that the effect of non-zero lag second moments on the zero-lag second moments can often be neglected without much loss of accuracy. The result is an SMC model of complexity comparable to the MFA.

The chapter is ordered as follows. In section 4.2 (and the Appendices) two ISR methods are discussed in a general context: spatial moment closure (SMC) and the 'n-spike approximation' (nSA). SMC is applied to a general, two-component reaction-diffusion model, and numerical methods are detailed. In section 4.3 three examples are investigated with different strengths of biological nonlinearity, obtained by varying the functional forms and parameter values of the general model, comparing HRS model output with the MFA and ISR models. In the Discussion section 4.4 limitations and possible extensions of this work are discussed. Conclusions are drawn in section 4.5.

4.2 Methods

4.2.1 Spatial Moment Closure (SMC)

Consider a discrete 1D chain of deterministic, nonlinear plankton subsystems (well mixed sub-populations) interacting by Fickian mixing, representing eddy-diffusion. The method easily generalises to 2D and 3D grids and continuous space — see Appendices 4A and 4B. To include stochasticity at the subsystem level, see Rodriguez and Tuckwell (1996). Hence, the n^{th} subsystem is governed by a set of m ODEs:

$$\dot{y}_i^{(n)} = F_i(y^{(n)}) + \lambda_i(y_i^{(n-1)} + y_i^{(n+1)} - 2y_i^{(n)}) \tag{4.1}$$

for $i = 1,\ldots,m$ subsystem components with mixing rates λ_i, using the dot to denote d/dt. F describes the fine-scale biological reactions, perhaps derived from laboratory experiments. The spatial means are defined by $Y_i \equiv N^{-1}\sum_{n=1}^{N} y_i^{(n)}$ and the zero spatial lag central covariances by $C_{ij}(0,t) \equiv N^{-1}\sum_{n=1}^{N} \delta y_i^{(n)}\delta y_j^{(n)}$ where $\delta y_i^{(n)} \equiv y_i^{(n)} - Y_i$ and the summation is over all N subsystems. Taylor-expanding the biological interactions F_i about the spatial means, averaging over space, and dropping third and higher central moments, results in second-order moment closure approximation for the evolution of the spatial means in terms of the spatial means and covariances (see Appendix 4A):

$$\dot{Y}_i \approx F_i(Y) + \frac{1}{2}\frac{\partial^2 F_i(Y)}{\partial y_j \partial y_k} C_{jk}(0,t) \tag{4.2}$$

using the convention that identical indices are implicitly summed over.

A few comments are in order. First, note that the 'standard' assumption in plankton models is that the net effect of terms involving the second and higher central moments on the mean concentrations over the grid cell is negligible (the MFA). By contrast, (4.2) proposes that the net effect of terms involving all third and higher central moments is negligible. These are different *closure assumptions.* Importantly, note that neither the spatial deviations nor any of the moments themselves are assumed to be small relative to the mean (since the coefficients in the expansion may vary by orders of magnitude), nor is it assumed that the neglected terms are individually small relative to the mean field dynamics. Rather, their *net* effect — the sum of all neglected terms — is assumed to be negligible.

Second, note that the coupling between the spatial means and higher moments in (4.2) results from the nonlinearity in the biological interactions (nonzero second derivatives of F). This nonlinearity also couples the dynamics of the higher moments, but always such that the dynamics of the k^{th} moments depend on terms multiplying the k^{th} and higher-order moments, with coefficients dependent on the mean field (see Appendix 4A). Thus we generate an infinite *hierarchy* of moment dynamics which must be truncated by a suitable closure assumption.

Third, note that in deriving (4.2), net boundary fluxes or external mixing is neglected (see Appendix 4A). The linear internal mixing has no direct effect on the dynamics of the mean concentrations. Within the MFA, mixing has no effect at all on the mean concentrations. However, to higher order in moment closure, the nonlinearity allows the mixing to affect the mean field dynamics via the coupling to zero-lag higher moments, which are directly affected by mixing. In fact, the dynamics of the zero-lag covariances are given, neglecting the net effect of third and higher central moments, by:

$$\dot{C}_{ij}(0,t) \approx \frac{\partial F_i(Y)}{\partial y_k} C_{kj}(0,t) + \frac{\partial F_j(Y)}{\partial y_k} C_{ki}(0,t) - 2(\lambda_i + \lambda_j)(C_{ij}(0,t) - C_{ij}(1,t)) \qquad (4.3)$$

(see Appendix 4A). Note, however, that different closure assumptions for the zero-lag covariance dynamics can be made without necessarily being inconsistent with (4.2).

In (4.3) the zero-lag covariance dynamics depend on the covariance dynamics at one unit lag ($C_{ij}(1,t) \equiv N^{-1} \sum_{n=1}^{N} \delta y_i^{(n)} \delta y_j^{(n+1)}$), where the spatial unit is the separation between adjacent subsystems (L). In this study, these terms are neglected in order to obtain a closed system (4.2) and (4.3). This may seem a rather drastic assumption. However, due to the scale dependence of mixing rate in the ocean, a situation is possible where sub grid-scale populations are sufficiently mixed to obey (4.1), but grid-scale mixing (λ_i, λ_j) is slow enough that covariances between adjacent grid cells may be neglected in a first approximation to the local covariance dynamics.

The neglect of third and higher order moment contributions is consistent with the assumption of a normal distribution — one which is sufficiently compact that the fourth and higher even-order central moments make a negligible net contribution. However, for biological variables, the assumption of normal distributions in phase space may well be inaccurate. An alternative is to neglect the net effect of third and higher central moments *in log space*, consistent with the assumption of a compact lognormal distribution (as is often observed in chlorophyll data — see Campbell, 1995; Appendix 3B). This closure assumption is most-easily imposed by log-transforming the fine-scale dynamics (4.1) into logarithmic variables $\tilde{y}_i = \log y_i$, hence:

$$\dot{\tilde{y}}_i^{(n)} = \tilde{F}_i(\tilde{y}^{(n)}) + \lambda_i (e^{\tilde{y}_i^{(n-1)}} + e^{\tilde{y}_i^{(n+1)}} - 2e^{\tilde{y}_i^{(n)}}) e^{-\tilde{y}_i^{(n)}} \qquad (4.4)$$

Then the means $\tilde{Y}_i \equiv N^{-1} \sum_{n=1}^{N} \tilde{y}_i^{(n)}$ and covariances $\tilde{C}_{ij}(0,t) \equiv N^{-1} \sum_{n=1}^{N} \delta\tilde{y}_i^{(n)} \delta\tilde{y}_j^{(n)}$, $\tilde{C}_{ij}(1,t) \equiv N^{-1} \sum_{n=1}^{N} \delta\tilde{y}_i^{(n)} \delta\tilde{y}_j^{(n+1)}$ may be defined in log space, where $\delta\tilde{y}_i^{(n)} \equiv \tilde{y}_i^{(n)} - \tilde{Y}_i$. Now Taylor-expanding about $\tilde{Y}$ and averaging over space, dropping third and higher-order moments, yields:

$$\dot{\tilde{Y}}_i \approx \tilde{F}_i(\tilde{Y}) + \frac{1}{2} \frac{\partial^2 \tilde{F}_i(\tilde{Y})}{\partial \tilde{y}_j \partial \tilde{y}_k} \tilde{C}_{jk}(0,t) + 2\lambda_i (\tilde{C}_{ii}(0,t) - \tilde{C}_{ii}(1,t)) \qquad (4.5)$$

and

$$\dot{\tilde{C}}_{ij}(0,t) \approx \frac{\partial \tilde{F}_i(\tilde{Y})}{\partial \tilde{y}_k}\tilde{C}_{kj}(0,t) + \frac{\partial \tilde{F}_j(\tilde{Y})}{\partial \tilde{y}_k}\tilde{C}_{ki}(0,t) - 2(\lambda_i + \lambda_j)(\tilde{C}_{ij}(0,t) - \tilde{C}_{ij}(1,t)) \qquad (4.6)$$

Hence the Fickian mixing, being nonlinear in log space, acts on $\tilde{Y}$ by decreasing the variance, thereby forcing an increase in the mean in log space $\tilde{Y}$ via the mixing term in (4.5) (recall that for a lognormal distribution: $\log Y_i = \tilde{Y}_i + \tilde{C}_{ii}(0,t)/2$). Note that the approximations (4.5) and (4.6) neglect more terms than (4.2) and (4.3) (namely, third and higher-order lag covariances). Nevertheless, when lognormality does significantly skew the spatial statistics, (4.5) and (4.6) are likely to be more accurate approximations.

Finally, note that in general it is possible that some transformation of variables $\tilde{y}_i = s(y_i)$ could linearise the dynamics $\tilde{F}$. However, to do so, the function $s(y)$ must itself be nonlinear (as in the log transform) which implies that the mean is not preserved under the transformation ($E[s(X)] \neq s(E[X])$). Therefore, higher moments would still require tracking in order to be able to convert back to the mean fields in non-transformed space, under suitable distributional assumptions. For example, in order to convert between moments in log space and moments in normal space, lognormality is assumed and the following standard conversion formulae are applied:

$$\begin{aligned} \tilde{Y}_i &= 2\log Y_i - \frac{1}{2}\log{(C_{ii} + Y_i^2)} \\ \tilde{C}_{ij} &= \log{(1 + \frac{C_{ij}}{Y_i Y_j})} \end{aligned} \qquad (4.7)$$

and:

$$\begin{aligned} Y_i &= e^{\tilde{Y}_i + \tilde{C}_{ii}/2} \\ C_{ij} &= e^{\tilde{Y}_i + \tilde{Y}_j + (\tilde{C}_{ii} + \tilde{C}_{jj})/2}(e^{\tilde{C}_{ij}} - 1) \end{aligned} \qquad (4.8)$$

where, as in the rest of this paper, C_{ij} is shorthand for $C_{ij}(0,t)$.

4.2.2 n-Spike Approximation (nSA)

The nSA forms a natural complement to the SMC method. Whereas SMC assumes a centrally-dominated distribution in phase space, the nSA approximates the phase space distribution as a sum of n delta functions (a multi-modal distribution), and may therefore attribute dominance to extreme values. The delta functions or 'spikes' comprise constant fractions $\{q_1, ..., q_{n-1}, (1 - \sum_s^{n-1} q_s)\}$ of the total number N of subsystems in the HRS (obviously, to yield advantages over the HRS, we must have $n << N$). Again, lag-covariances are neglected. Under these assumptions, averaging over all subsystems in the k^{th} spike (a subset of the N subsystems in the HRS) yields m equations for each of the $i = 1,\ldots,m$ subsystem components:

$$\dot{y}_i^{(k)} = F_i(y^{(k)}) + 2^D \lambda_i (\sum_{s=1}^{n} q_s y_i^{(s)} - y_i^{(k)}) \qquad (4.9)$$

where $q_n = 1 - \sum_{s=1}^{n-1} q_s$. Be careful not to think of the spikes as spatial entities or 'patches'—rather they are subsets of the the N subsystems distributed randomly in space, and as such each member of a subset may mix with members of the same subset as indicated by (4.9). The spatial mean fields are given by $Y_i = \sum_{s=1}^{n} q_s y_i^{(s)}$ and local covariances by $C_{ij} = \sum_{s=1}^{n} q_s (y_i^{(s)} - Y_i)(y_j^{(s)} - Y_j)$. Note that the spikes mix faster for higher spatial dimensions with more nearest neighbours (factor 2^D). Also, note that the mixing is asymmetric between spikes: the higher the proportion of subsystems q in the spike, the less it differs from the spatial mean, hence the less net mixing flux it experiences.

4.2.3 SMC model for a general plankton system

This study considers fine-scale phytoplankton-zooplankton (pz) population dynamics of the general form:

$$\begin{aligned} \dot{p} &= f(p)p - g(p)z \\ \dot{z} &= rg(p)z - h(z) \end{aligned} \tag{4.10}$$

where $f(p)p = a(1-bp)p$ represents phytoplankton growth, assumed to be logistic; $g(p)z$ describes grazing by the zooplankton, where different forms of $g(p)$ are considered; r is a constant accounting for zooplankton feeding and assimilation efficiency; $h(z) = dz$ describes zooplankton mortality and other losses, assumed to be proportional to z. The subsystem (4.10) has a single 'coexistence' equilibrium at ($p^* = g^{-1}(d/r)$, $z^* = (r/d)f(p^*)p^*$) which exists in the positive quadrant provided p^* is real and positive and less than the carrying capacity ($g^{-1}(d/r) < b^{-1}$). This equilibrium is stable provided $K \equiv (f'(p^*)p^* + f(p^*) - g'(p^*)z^*) < 0$ (primes denoting derivatives), in which case it is a stable spiral (eigenvalues complex) if $K^2 < 4rg'(p^*)g(p^*)z^*$, otherwise it is a stable sink (eigenvalues real). The positivity requirements for g and g' rule out a saddle coexistence equilibrium. An important parameter in our study will be the *linear transience time* defined as $\tau_{lin} \equiv -2/K$, the constant $K/2$ being the oscillation decay rate in the case of a stable spiral and the average decay rate in the case of a stable sink equilibrium.

The MFA simply replaces (p, z) with their spatial means (P, Z). Applying the second moment closure approximation for the mean field dynamics (4.2) yields the SMC mean field dynamics:

$$\begin{aligned} \dot{P} &= fP - gZ + \frac{1}{2}((2f' - g''Z)C_{pp} - 2g'C_{pz}) \\ \dot{Z} &= rgZ - h + \frac{r}{2}(g''ZC_{pp} + 2g'C_{pz})) \end{aligned} \tag{4.11}$$

where f, g and h and their derivatives are evaluated at (P, Z) and C_{ij} is short for the covariance at zero lag ($C_{ij}(0,t)$). Applying the second moment closure approximation for the zero-lag covariance dynamics (4.3), and neglecting lag covariances, yields the SMC covariance dynamics:

$$\begin{aligned}
\dot{C}_{pp} &= 2(f + f'P - g'Z)C_{pp} - 2gC_{pz} - 2^{D+1}\lambda_p C_{pp} \\
\dot{C}_{pz} &= rg'ZC_{pp} + (f + f'P - g'Z + rg - h')C_{pz} - gC_{zz} - 2^D(\lambda_p + \lambda_z)C_{pz} \\
\dot{C}_{zz} &= 2rg'ZC_{pz} + 2(rg - h')C_{zz} - 2^{D+1}\lambda_z C_{zz}
\end{aligned} \tag{4.12}$$

where D is the number of spatial dimensions ($D = 2$ in our study). Thus (4.11) and (4.12) form a closed system. In log space, the biological model (4.10) is restated as:

$$\begin{aligned}
\dot{\tilde{p}} &= \tilde{f}(\tilde{p}) - \tilde{g}(\tilde{p})e^{\tilde{z}} \\
\dot{\tilde{z}} &= r\tilde{g}(\tilde{p})e^{\tilde{p}} - d
\end{aligned} \tag{4.13}$$

where $\tilde{p} = \log p$ and $\tilde{z} = \log z$. Note that the transformed function $\tilde{f}(\tilde{p}) = a(1-e^{\tilde{p}})$ is no longer linear. Applying the second moment closure approximation for the mean field dynamics (4.5) and zero-lag covariance dynamics (4.6) in log space, again neglecting lag covariances, yields the SMC mean field dynamics:

$$\begin{aligned}
\dot{\tilde{P}} &= \tilde{f} - \tilde{g}e^{\tilde{Z}} + \frac{1}{2}((\tilde{f}'' - \tilde{g}''e^{\tilde{Z}})\tilde{C}_{\tilde{p}\tilde{p}} - 2\tilde{g}'e^{\tilde{Z}}\tilde{C}_{\tilde{p}\tilde{z}} - \tilde{g}e^{\tilde{Z}}\tilde{C}_{\tilde{z}\tilde{z}}) + 2^D\lambda_p\tilde{C}_{\tilde{p}\tilde{p}} \\
\dot{\tilde{Z}} &= r\tilde{g}e^{\tilde{P}} - d + \frac{r}{2}(\tilde{g}'' + 2\tilde{g}' + \tilde{g})e^{\tilde{P}}\tilde{C}_{\tilde{p}\tilde{p}} + 2^D\lambda_z\tilde{C}_{\tilde{z}\tilde{z}}
\end{aligned} \tag{4.14}$$

and the SMC covariance dynamics:

$$\begin{aligned}
\dot{\tilde{C}}_{\tilde{p}\tilde{p}} &= 2(\tilde{f}' - \tilde{g}'e^{\tilde{Z}})\tilde{C}_{\tilde{p}\tilde{p}} - 2\tilde{g}e^{\tilde{Z}}\tilde{C}_{\tilde{p}\tilde{z}} - 2^{D+1}\lambda_p\tilde{C}_{\tilde{p}\tilde{p}} \\
\dot{\tilde{C}}_{\tilde{p}\tilde{z}} &= r(\tilde{g}' + \tilde{g})e^{\tilde{P}}\tilde{C}_{\tilde{p}\tilde{p}} + (\tilde{f}' - \tilde{g}'e^{\tilde{Z}})\tilde{C}_{\tilde{p}\tilde{z}} - \tilde{g}e^{\tilde{Z}}\tilde{C}_{\tilde{z}\tilde{z}} - 2^D(\lambda_p + \lambda_z)\tilde{C}_{\tilde{p}\tilde{z}} \\
\dot{\tilde{C}}_{\tilde{z}\tilde{z}} &= 2r(\tilde{g}' + \tilde{g})e^{\tilde{P}}\tilde{C}_{\tilde{p}\tilde{z}} - 2^{D+1}\lambda_z\tilde{C}_{\tilde{z}\tilde{z}}
\end{aligned} \tag{4.15}$$

4.2.4 Numerical methods

The high-resolution simulations require solution of the discrete reaction-diffusion equations. Note that since the subsystems are assumed to be of finite size in the problem, on a scale at which continuous-time population dynamics is applicable, diffusive mixing is implemented simply as stated in (4.1). Hence we have a coupled (for non-zero mixing rate) system of 512^2 ODEs in 2D. Since the model is defined in continuous time, the finite numerical time step is compensated by a fifth-order, adaptive step-size Runge-Kutta solver from Press et al. (1999). Error tolerance is set to 0.01% of the maximum absolute value of each variable over recording intervals of 0.1, 0.02 and 0.1 days for examples 1, 2 and 3 respectively. As an accuracy check, results were repeated at a higher error tolerance (0.1%), resulting in no perceptible difference. These integrations were coded in Fortran for speed purposes.

For the moment closure models, requiring integration of 5 ODEs, the 'ode45' solver in the Matlab mathematical software package (http://www.mathworks.com/) was used, but only after validating against output from the Fortran solver, since the non-negativity requirement for mean field and variance variables could not be strictly imposed at all intermediate 'trial' steps in the fixed Matlab routines. The two methods showed no significant disagreement.

4.3 Examples

4.3.1 Example 1a: Weak nonlinearity with no mixing ($\lambda = 0$)

Consider a small piece of ocean surface mixed layer where phytoplankton concentrations (p) obey self-limited, logistic growth but are too low to saturate a linear grazing response by zooplankton (z), and where these zooplankton are subject to a constant specific mortality rate due to senescence and higher predation. Nutrient limitation and layer-averaged light exposure are assumed to be constant. Omitting mixing interactions, the population dynamics in this volume might be represented by the Lotka-Volterra predator-prey equations:

$$\begin{aligned} \dot{p} &= ap(1-bp) - cpz \\ \dot{z} &= rcpz - dz \end{aligned} \tag{4.16}$$

The concentrations (p, z) are assumed to be measured in the same units (e.g. μM Nitrogen), hence the constant r accounts only for imperfect feeding and assimilation efficiency. Since this study is concerned with model behaviour rather than fitting real data, it makes sense to rescale the equations in order to allow us to explore model behaviour more efficiently. Applying the rescalings: $t = t'/a$, $p = p'/b$, $z = az'/c$ yields:

$$\begin{aligned} \dot{p} &= p(1-p) - pz \\ \dot{z} &= \alpha pz - \beta z \end{aligned} \tag{4.17}$$

where primes have been dropped, $\alpha = rc/(ab)$ and $\beta = d/a$. The system (4.17) has a single coexistence equilibrium at ($p^* = \beta/\alpha$, $z^* = (1 - \beta/\alpha)$), which is stable while it is in the positive quadrant ($\alpha > \beta$). As an example, plausible values ($\alpha = 0.2$, $\beta = 0.1$) are chosen — hence the mortality rate of z is a factor of 10 slower than the maximum growth rate of p, and, if $b \sim 1$ pz-unit^{-1}, the zooplankton growth is an order of magnitude slower than the phytoplankton growth. This results in a stable sink equilibrium (see section 4.2.3).

In this study, spatial heterogeneity is provided exclusively by the initial conditions: ($p(0)$, $z(0)$) are assumed to be independent (zero lag-covariance) lognormal variables with mean values ($P(0) = p^*$, $Z(0) = 0.78z^*$ — chosen for parity with B03) and standard deviations 100% of the mean values (hence $C_{pp}(0,0) = P(0)^2$, $C_{pz}(0,0) = 0$, and $C_{zz}(0,0) = Z(0)^2$). The initial coefficients of variation (standard deviations/mean values) are high (1, 1) but not implausible in view of the range of values found in plankton data from the Celtic Sea (Adrian Martin, pers. comm.). Subsequently, the HRS is run for 60 time units i.e. 60 × the growth timescale of phytoplankton, allowing convergence on the homogeneous equilibrium ($p(x,y) = p^*$, $z(x,y) = z^*$).

The MFA model for the domain is made simply by replacing (p, z) by (P, Z) in (4.17) and evolving the system of 2 ODEs from the HRS mean initial conditions ($P(0)$, $Z(0)$). The SMC approximation is given by the general model (4.14 and 4.3.1), substituting for the

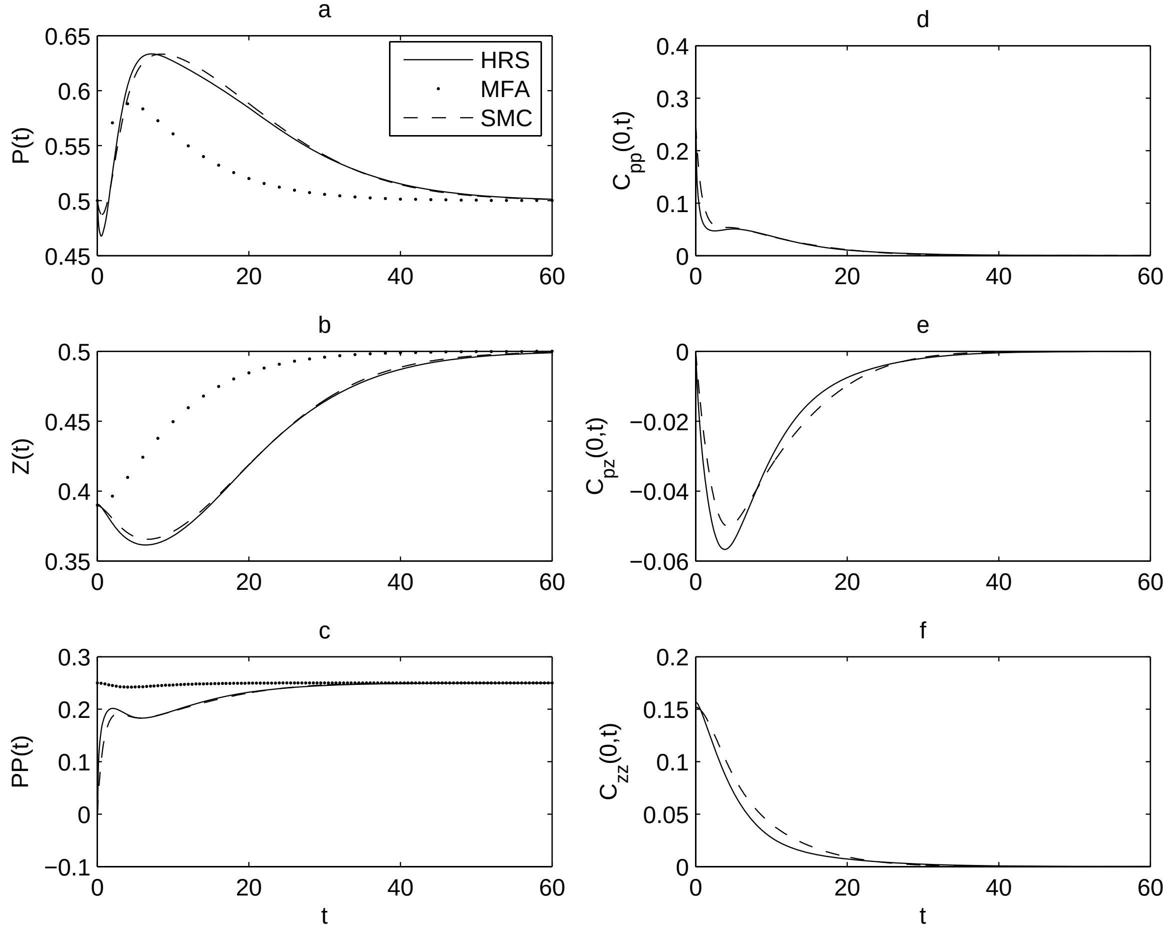

Figure 4.1: Time Series for Example 1a (weak nonlinearity, no mixing), showing High Resolution Simulation (HRS), Mean Field Approximation (MFA) and Spatial Moment Closure (SMC) output for phytoplankton and zooplankton spatial mean fields (a, b), mean primary productivity (c) and spatial covariances (d–f). Units are non-dimensional (see section 4.3.1).

functions ($\tilde{f} = (1 - e^{\tilde{P}})$, $\tilde{g} = 1$) and the constants ($r = \alpha$, $d = \beta$) and setting mixing rates (λ_p, λ_z) to zero. This model, which neglects third and higher central moments in log space, is a better choice than (4.11 and 4.12) given that the imposed initial variability is lognormal. The resulting system of 5 ODEs is evolved from ($\tilde{P}(0)$, $\tilde{Z}(0)$, $\tilde{C}_{\tilde{p}\tilde{p}}(0,0)$, $\tilde{C}_{\tilde{p}\tilde{z}}(0,0)$, $\tilde{C}_{\tilde{z}\tilde{z}}(0,0)$), using the formulae (4.7 and 4.8), which assume lognormality, to convert moments to/from log space.

The resulting behaviour of (P, Z, C_{pp}, C_{pz}, C_{zz}) is shown for HRS, MFA and SMC models in Figure 4.1. Note that the ensemble variability (between repeat runs due to different initial configurations) is negligible in the spatial moments, thanks to the large number of subsystems (512^2) in 2D (hence the simulation is 'ergodic' in the sense that spatial averages are equivalent to ensemble averages). Therefore single-run results rather than ensemble averages are shown. Also shown are the model predictions for bulk primary productivity (PP), defined as the spatial mean self-limited phytoplankton growth rate, as this is often a quantity of interest to plankton modellers. Note that: $PP = P(1 - P) - C_{pp}$ with no approximation — spatial variance enhances coarse-scale logistic self-limitation.

The spatial heterogeneity amplifies the coarse-scale phytoplankton 'bloom' and slightly delays the timing of the peak relative to the MFA (Fig. 4.1a). The HRS zooplankton mean field approaches equilibrium slower than in the MFA, and undergoes an initial decrease not predicted by the MFA (Fig. 4.1b). The bulk primary productivity is slightly suppressed relative to the MFA for about the first third of the transient interval, implying that in this case the deficit in primary production due to local variance (C_{pp}) outweighs the increase in P (Fig. 4.1c).

The time series for the local covariances also allow us, in this case, to anticipate the qualitative evolution of the pz-distribution in phase space (Figure 4.2). After 5 time units $C_{pp}(0,t)$ has reduced dramatically, $C_{pz}(0,t)$ has dropped to a negative trough, and $C_{zz}(0,t)$ has reduced slightly (Figs. 4.1d–f). Hence, starting from the initial 'blob' of lognormal variability (Fig. 4.2a), the distribution contracts into a 'sausage' shape sloping steeply towards lower z with increasing p at $t = 5$ (Fig. 4.2b). The differential contraction is largely a linear effect of stronger dynamical flow convergence in the direction of the more stable manifold (reverse-time trajectory of a subsystem perturbed from equilibrium along the eigenvector associated with the more negative eigenvalue) leaving the distribution elongate along the direction of the least stable manifold (associated with the less negative eigenvalue). The dynamical nonlinearity acts on the distribution width to perturb the mean fields with respect to the MFA. Subsequently, the covariances decay gradually (Figs. 4.1d–f) as the distribution becomes 'needle' shaped, slowly contracting in the long direction as the mean inches towards the equilibrium (Figs. 4.2b,c).

All features of the HRS mean field behaviour are accurately predicted by the SMC model (Figs. 4.1 and 4.2). In addition, the SMC does a good job of reproducing the transient variability in local covariances (see Figs. 4.1d–f, the MFA prediction for each being zero for all time). This latter is more surprising because the SMC model includes first-order corrections to the mean field dynamics (due to local covariances) but only retains the zeroth-order terms in the local covariance dynamics (again due to local covariances).

The success of the SMC model allows us to explain the results of the HRS model in terms of interactions between the mean fields and local covariances. This is done most simply by referring to (4.11) and (4.12) rather than (4.14) and — the latter is more accurate given lognormal variability but requires us to convert to/from log space. First, a negative pz-covariance develops due to grazing and relatively slow zooplankton growth (an imbalance between the C_{pp} and C_{zz} terms in $\dot{C}_{pz}$). This negative covariance reduces the domain-scale grazing rate because the zooplankton exhaust their local food supply (C_{pz} term in $\dot{P}$), thereby allowing an enhanced mean-field bloom due to local predator release (outweighing the reduction of PP due to C_{pp}). The predator-prey decoupling (negative C_{pz}) also delays the domain-scale increase in zooplankton and resulting depletion of the phytoplankton crop (C_{pz} term in $\dot{Z}$). Subsequently, all local covariances decay due to dynamical convergence of (p, z) phase space trajectories towards the stable equilibrium (C_{pp}, C_{pz} and C_{zz} terms in their respective time derivatives).

However, in this weakly nonlinear example the effect of nonlinearity is after-all rather small. To quantify it, Δ_P^{MFA} is defined as the root-mean-square discrepancy between the MFA and HRS model P fields, averaged over a nonlinear transient interval and normalised by p^*. The nonlinear transient interval for P is defined to last from $t = 0$ until the largest time at which $P^{HRS}(t)$ is more that 5% away from the equilibrium p^*. This yields an rms error of $\Delta_P^{MFA} = 11\%$, compared to $\Delta_P^{SMC} = 1\%$. Similarly, for zooplankton rms error we

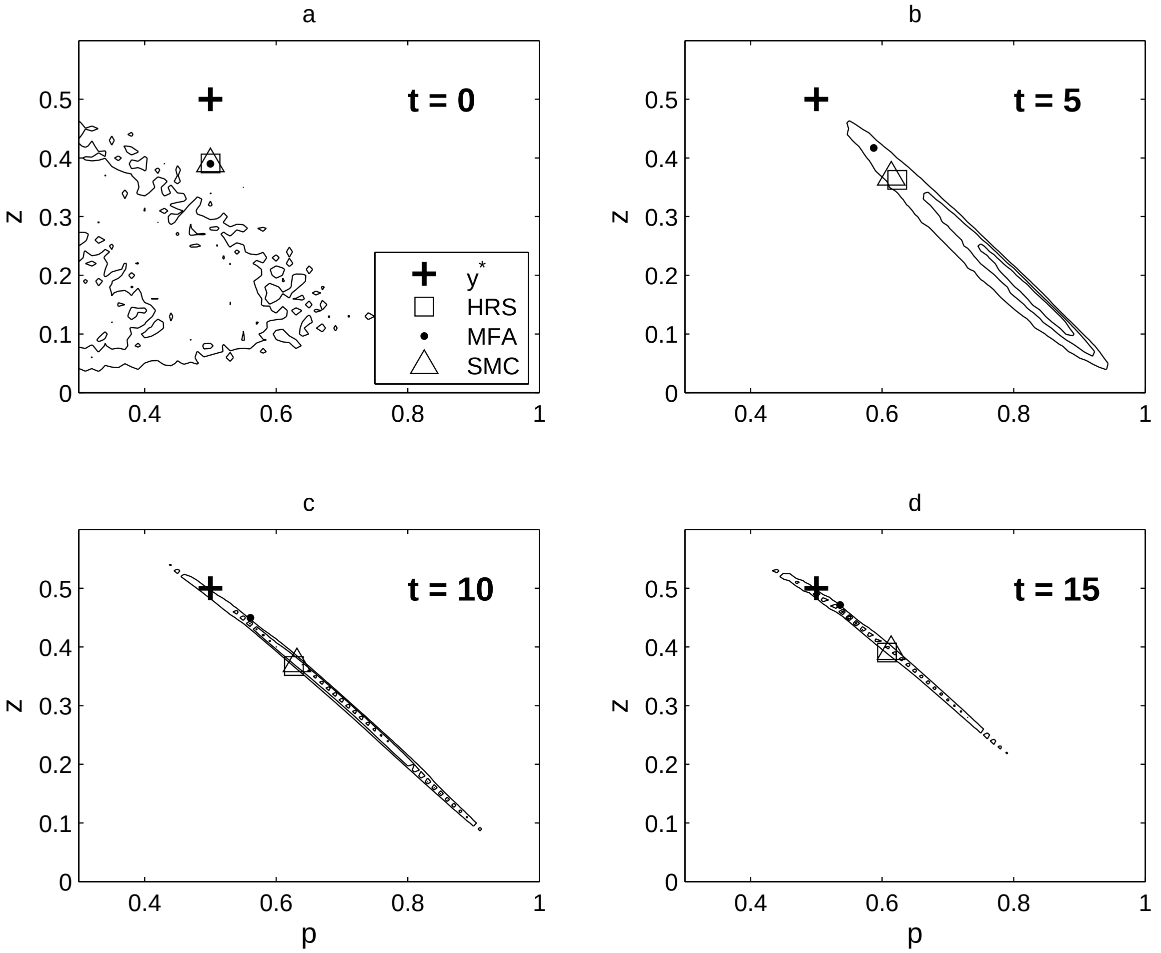

Figure 4.2: Phase space phytoplankton-zooplankton distributions for Example 1a (weak nonlinearity, no mixing) at t = 0 (a), t = 5 (b), t = 10 (c) and t = 15 (d). Symbols mark the position of the equilibrium (cross), the mean of the High Resolution Simulation (square), the mean predicted by the Mean Field Approximation (dot), and the mean predicted by the Spatial Moment Closure model (triangle). Units are non-dimensional (see section 4.3.1)

have Δ_Z^{MFA} =9% vs. Δ_Z^{SMC} <1%. The % error in the total primary production over the interval of PP-transience is similar ($\delta_{PT}^{MFA} = +13\%$ vs. $\delta_{PT}^{SMC} = -1\%$). In plankton data, such errors could easily be undetectable. Also, mixing has not yet been included. This is a crucial ocean process expected to suppress discrepancies between the HRS mean fields and the MFA.

Nevertheless, it is instructive to investigate the controls on $\Delta^{MFA/SMC}$ in this simple model with no mixing. To this end, the above experiment is repeated for 500 different values of (α, β) randomly distributed (uniformly and independently) over plausible ranges $(0.01 < \alpha < 2, 0.01 < \beta < 2)$ with the restriction $\alpha > \beta$ to provide a positive and stable equilibrium. Results are discarded of runs for which the P^{SMC} or Z^{SMC} exceeded 100 units (20 cases) and for which the nonlinear transience time exceeded the run time of 100 units (35 cases). Figure 4.3 shows the resulting $\Delta_P^{MFA/SMC}$ plotted against the linear transience time τ_{lin} (see section 4.2.3), excluding 4 cases with $\Delta_P^{SMC} > 0.5$ by choice of axis limits. There is a clear positive correlation between $\Delta_P^{MFA,SMC}$ and τ_{lin} ($r^2 = 0.76$ and 0.41, $p < 0.001$ for both). The relationship is remarkably clean though nonlinear up to $\tau_{lin} \approx 6$, becoming noisy for higher τ_{lin} as the performance of both models becomes erratic. This supports the intuitive view that for more oscillatory transience with longer τ_{lin}, the phase space distribution develops more complex 'boomerang-shaped' and skewed distributions which cause stronger MFA discrepancies and pose difficulties for the SMC, which assumes a compact normal distribution in log space. Even so, the SMC model is clearly quite robust and yields improved accuracy for almost all sensible parameter values.

4.3.2 Example 1b: Weak nonlinearity with realistic mixing rates ($\lambda > 0$)

The fine-scale population dynamics in the HRS (4.17) are assumed to be accurate at the scale of laboratory and mesoscale experiments (1–10m in the ocean surface mixed layer). The Okubo relation: $\lambda_a \approx 120L^{-0.85}$ day^{-1} (plus or minus an order of magnitude) implies a minimal mixing rate of $\lambda_a = 1.7$ day^{-1} at the scale $L = 10$ m. Assuming that (p, z) may be treated as passive tracers as regards their eddy-mixing timescales (a questionable assumption — see section 4.4), and assuming a maximum phytoplankton growth rate of $a = 1$ day^{-1}, a minimum non-dimensional mixing rate ('diffusive Damköhler number') is given by $\lambda^{min} = \lambda_a^{min}/a = 1.7$. Repeating the first experiment of section 4.3.1 using $\lambda_p = \lambda_z = 1.7$ yields negligible discrepancy between the HRS/MFA/SMC. Hence in this case, the fine-scale dynamics may be safely extrapolated from $L = 10$ m to an ocean model grid cell of scale $D = 512L \approx 5$km, as might be used in a standard regional ocean model or a very high resolution General Circulation Model (GCM) for the global ocean.

By a similar argument, a minimal dimensionless mixing rate for $L = 100$ m is $\lambda_p = \lambda_z = 0.24$, which yields $(\Delta_P^{MFA}, \Delta_P^{SMC}) = (3,\ 1)\%$ for an ocean grid cell of scale $D \approx 50$km, as in a moderate-resolution GCM. For the largest relevant fine scale $L = 1$km, using $\lambda_p = \lambda_z = 0.03$, $(\Delta_P^{MFA}, \Delta_P^{SMC}) = (6,\ 1)\%$ for $D \approx 500$km. Hence, if 10% error is taken as a threshold of importance, the results suggest that, for this model formulation, the error involved in extrapolating transient behaviour from laboratory to ocean model grid cell scale is negligible for all practical grid cell scales.

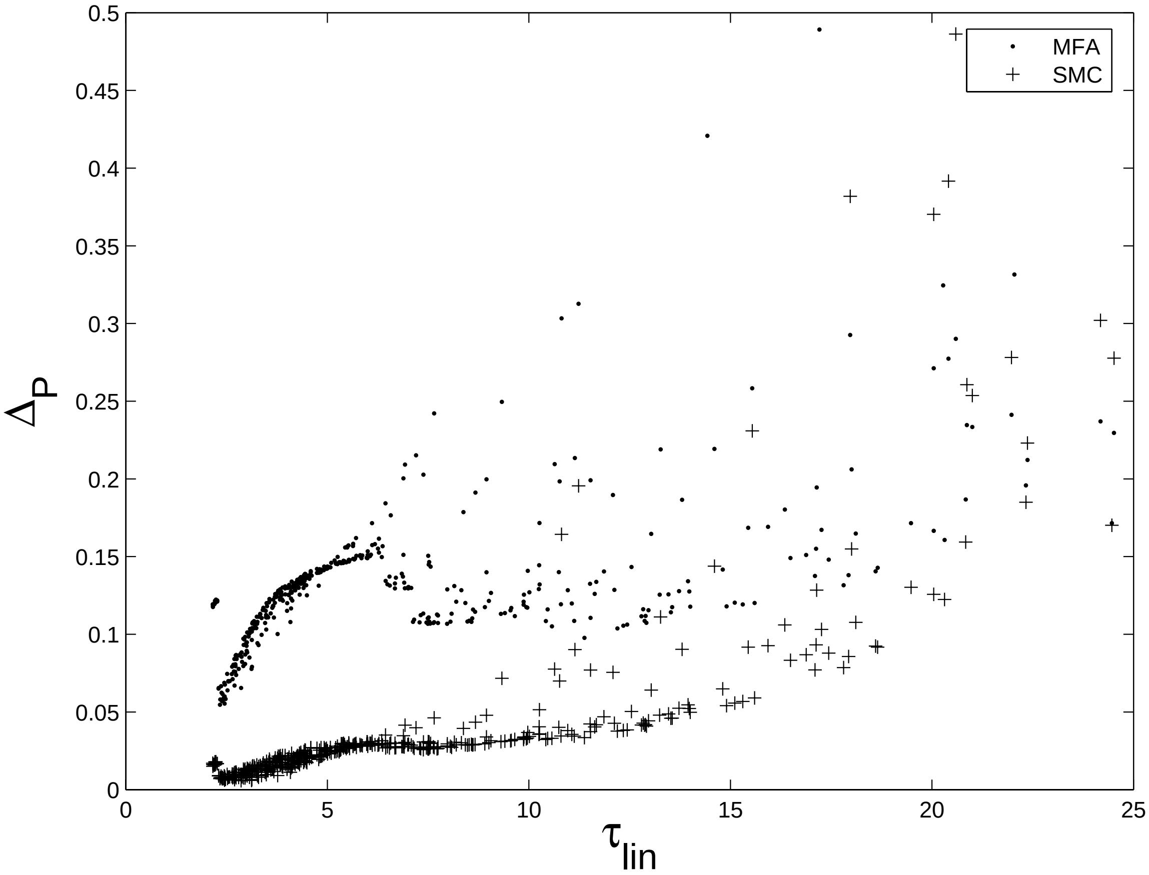

Figure 4.3: Average (rms) errors in the phytoplankton mean field over transient interval, Δ_P, vs. linear transience time, τ_{lin}, for the Mean Field Approximation (dots) and Spatial Moment Closure (crosses) model, using a weakly nonlinear model with 500 sets of parameter values randomly-chosen from plausible ranges. Mixing rate $\lambda = 0$. Units are non-dimensional (see section 4.3.1).

4.3.3 Example 2a: Moderate nonlinearity with no mixing ($\lambda = 0$)

Moderately nonlinear dynamics, in the sense of discrepancy between the HRS mean field dynamics and the MFA, result from modifying the formulation in (4.16) to account for grazer saturation. Following B03 and numerous other studies (Matthews and Brindley, 1997; Neufeld et al., 2002; Steele and Henderson, 1981; Truscott and Brindley, 1994), the following simple plankton model is used:

$$\begin{aligned} \dot{p} &= ap(1-bp) - \frac{cp^2 z}{k^2+p^2} \\ \dot{z} &= \frac{rcp^2 z}{k^2+p^2} - dz \end{aligned} \tag{4.18}$$

where the new parameter k is the half-saturation constant for zooplankton grazing. Note also the change in units of c from (pz-unit^{-1} day^{-1}) in (4.16) to (day^{-1}) in (4.18). The Holling III functional response implies a slow (quadratic) increase in grazing pressure at low phytoplankton concentrations ($p < k$) as well as saturation at high concentrations ($p > k$). The former is consistent with the presence of an alternative food source for the zooplankton, and a minimal level of zooplankton behavioural complexity (see B03 for more detailed justification). Dynamically, the inclusion of saturation effects raises the 'degree' of nonlinearity above the quadratic nonlinearity of the Lotka-Volterra dynamics in (4.16). Applying the rescalings: $t = t'/a$, $p = p'/b$, $z = az'/c$ (note that z remains dimensional) yields the following scaled system:

$$\begin{aligned} \dot{p} &= p(1-p) - \frac{p^2 z}{\nu^2+p^2} \\ \dot{z} &= \frac{\alpha p^2 z}{\nu^2+p^2} - \beta z \end{aligned} \tag{4.19}$$

where primes have been dropped, $\alpha = rc/a$, $\beta = d/a$ and $\nu = kb$. The system (4.19) has a single coexistence equilibrium at ($p^* = (\beta\nu^2/(\alpha-\beta))^{1/2}$, $z^* = (\alpha/\beta)p^*(1-p^*)$), which exists in the positive quadrant if $0 < p^* < 1$ and is stable if $K = 1 - 2p^* - 2(\nu\beta/\alpha)^2 z^*/(p^*)^3 < 0$ (see section 4.2.3). Plausible parameter values ($\alpha = 1$, $\beta = 0.1$, $\nu = 0.1$) are used as an example — hence the zooplankton grow as fast as the phytoplankton but die much slower, and grazing half-saturates well below the phytoplankton carrying capacity. The resulting equilibrium is a stable spiral.

The initial pz distribution is as in Example 1: $(p(0), z(0))$ lognormally distributed with mean values ($P(0) = p^*, Z(0) = 0.78z^*$) and 100% variance with no covariance or lag-covariance ($C_{pp}(0,0) = P(0)^2, C_{pz}(0,0) = 0, C_{zz}(0,0) = Z(0)^2$). The individual pz trajectories subsequently decay towards the homogeneous equilibrium ($p(x,y) = p^*, z(x,y) = z^*$), within the 20 units of integration time. As before, the MFA is given by replacing (p,z) by (P,Z) in (4.19), whilst the SMC model is given by the 5 ODEs (4.14 and 4.3.1), substituting ($\tilde{f} = (1 - e^{\tilde{P}}), \tilde{g} = e^{\tilde{P}}/(\nu^2 + e^{2\tilde{P}})$), $(d,e) = (\beta, \alpha)$, ($\lambda_p = 0, \lambda_z = 0$), and using the formulae (4.7 and 4.8) to covert moments to/from log space.

The results are shown in Figure 4.4. Compared to the weakly nonlinear example (Figure

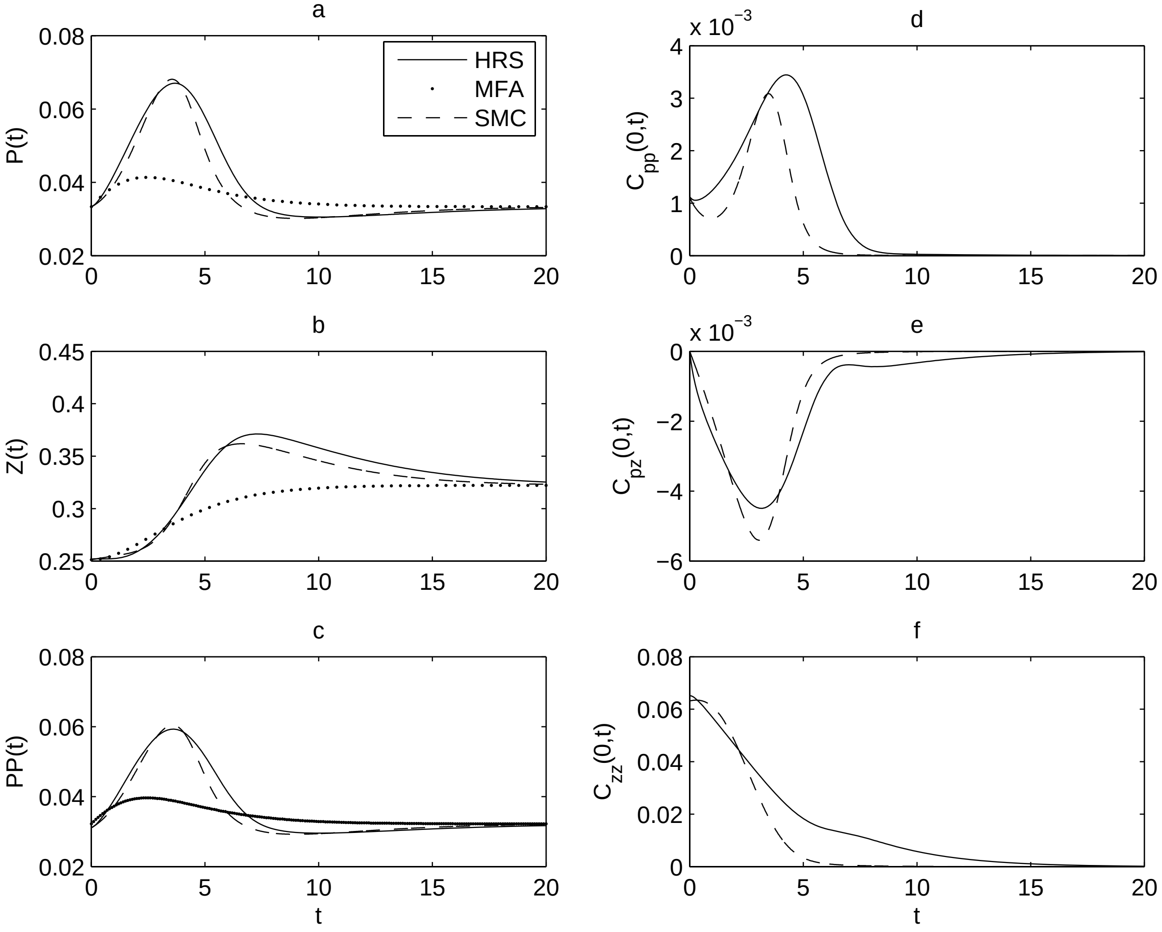

Figure 4.4: Time Series for Example 2a (moderate nonlinearity, no mixing), showing High Resolution Simulation (HRS), Mean Field Approximation (MFA) and Spatial Moment Closure (SMC) output for phytoplankton and zooplankton spatial mean fields (a, b), mean primary productivity (c) and spatial covariances (d–f). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

4.1), the spatial heterogeneity amplifies the mean phytoplankton bloom to a much greater extent (Fig. 4.4a), and the zooplankton mean field undergoes a larger, this time positive, deviation from the MFA (Fig. 4.4b). In this case, the mean primary productivity is strongly enhanced relative to the MFA (Fig. 4.4c), the increase in P outweighing the deficit due to C_{pp} (Fig. 4.4d).

The evolution of the phase space distribution (Figure 4.5) is qualitatively similar to Example 1 (Figure 4.2), except that here the initial asymmetric contraction is less severe (Fig. 4.5b), corresponding to weaker negative covariance (Fig. 4.4e) and the distribution initially expands in the p direction (Fig. 4.5b), as indicated by the variance time series (Fig. 4.4d). As in Example 1, the distribution remains strongly skewed, as indicated by the separation of means from modal peaks (Figs. 4.5b–d), but here the skew is in the general direction of approach to the spiral equilibrium, rather than along a slow manifold (Figs. 4.2b–d).

All features of the domain-scale HRS dynamics are predicted and explained by the SMC model, albeit less accurately than in Example 1. Again, it suffices to refer to (4.11 and 4.12) for qualitative explanations rather than the more accurate but less tractable system (4.14)

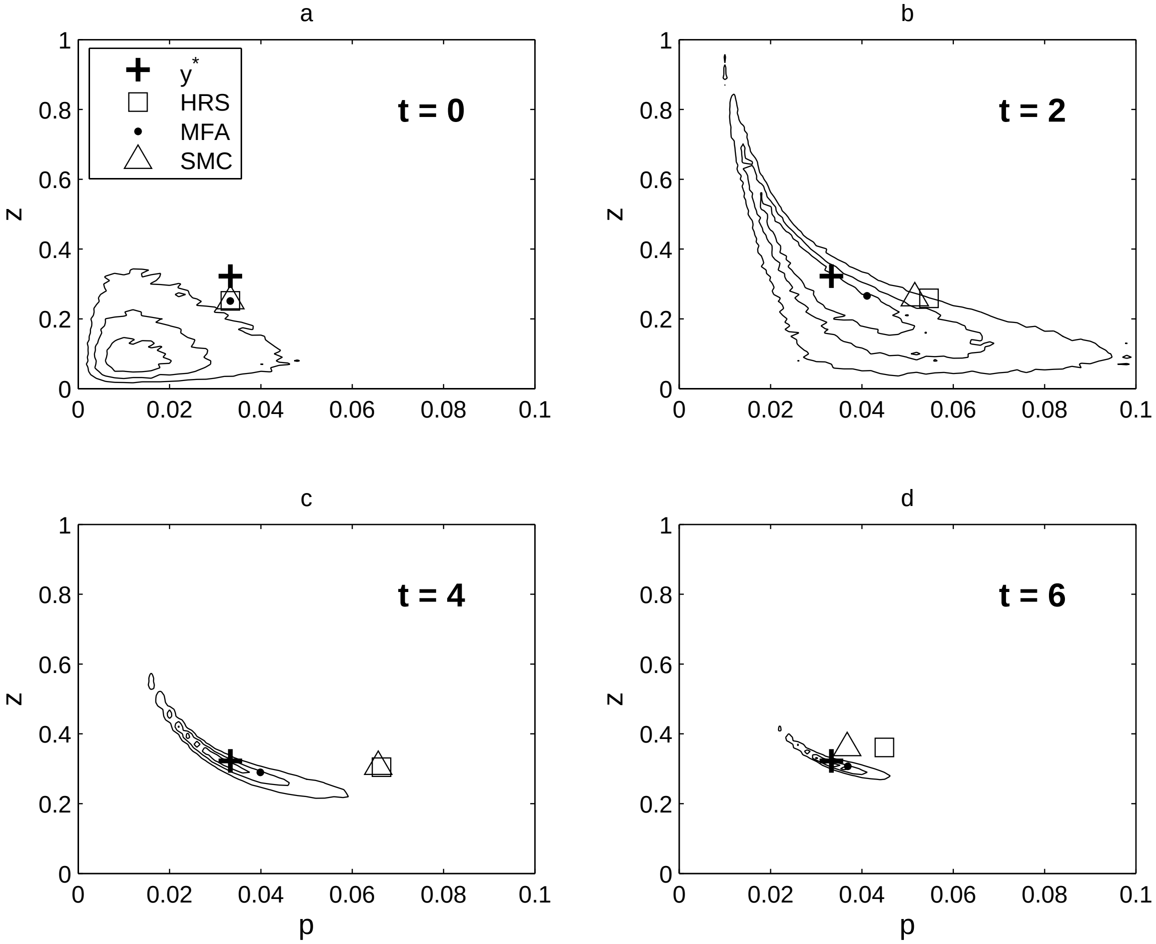

Figure 4.5: Phase space phytoplankton-zooplankton distributions for Example 2a (moderate nonlinearity, no mixing) at t = 0 (a), t = 2 (b), t = 4 (c) and t = 6 (d). Symbols mark the position of the equilibrium (cross), the mean of the High Resolution Simulation (square), the mean predicted by the Mean Field Approximation (dot), and the mean predicted by the Spatial Moment Closure model (triangle). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

and 4.3.1) which was actually employed. As before, grazing acts on the initial heterogeneity to decouple the predator and prey populations (imbalance of C_{pp} and C_{zz} terms in $\dot{C}_{pz}$ leads to negative C_{pz}) which in turn suppresses the domain-scale grazing losses and zooplankton growth (C_{pz} terms in $\dot{P}$ and $\dot{Z}$). However, with the sigmoidal functional response, the zooplankton reduce their grazing activity where phytoplankton are depleted ($g''(p)$ positive for low p), which tends to exacerbate the predator-prey decoupling, allowing an even larger domain-scale bloom. Meanwhile, the spatial variance in phytoplankton grows due to local predator release (C_{pz} term in $\dot{C}_{pp}$). Because in this case the p and z maximum growth rates are equal ($\alpha = 1$), the zooplankton are able to rapidly respond to the blooming crop. Also, the sigmoidal grazing response allows the zooplankton to better exploit relatively phytoplankton-rich areas, thus enhancing their domain-scale grazing and growth (g'' terms in $\dot{P}$ and $\dot{Z}$). The bloom is further limited by the enhancement of logistic self-limitation due to spatial variance in phytoplankton (f' term in $\dot{P}$). Hence the system is restored rather more rapidly to the homogeneous equilibrium (cf. Fig. 4.1).

The SMC reduces the rms error in the MFA by 24% of the equilibrium values for P (Δ_P^{MFA} =35%, Δ_P^{SMC} =11%), by 8% of the equilibrium values for Z (Δ_Z^{MFA} =11%, Δ_Z^{SMC} =3%), by 20% of the equilibrium values for PP (Δ_{PP}^{MFA} =28%, Δ_{PP}^{SMC} =8%), and reduces the % error in total production by a factor of 3 (δ_{PT}^{MFA} =-9%, δ_{PT}^{SMC} =-3%) (see section 4.3.1 for definitions of these measures). Such errors could well be detectable in data.

Again, before investigating mixing effects, the controls on $\Delta^{MFA/SMC}$ are investigated. The above experiment is repeated for 500 different values of (α, β, ν), chosen as uniform and independent random variables in 'plausible' ranges $0.01 < \alpha < 2$, $0.01 < \beta < 2$ and $0.01 < \nu < 2$, integrating for 100 time units. Again, results are discarded for runs where P^{SMC} or Z^{SMC} exceeds 100 units (14 cases) and for which the nonlinear transience time exceeds the run time of 100 units (37 cases). The resulting rms errors ($\Delta_P^{MFA}, \Delta_P^{SMC}$) were plotted against different parameters. No significant correlation was observed with linear transience time τ_{lin} (see section 4.2.3) or α or β. However, a significant inverse correlation was observed with ν (Figure 4.6 — excluding 22 cases by choice of axis limits) ($r^2 = 0.66$, 0.13, $p < 0.001$, 0.01). This supports the intuitive view that decreasing ν increases the 'nonlinearity' of (4.19), with the system tending towards a third-degree nonlinear system for large ν. Also, Figure 4.6 clearly demonstrates the robustness of the SMC model, which significantly outperforms the MFA for almost all sensible parameter values.

4.3.4 Example 2b: Moderate nonlinearity with realistic mixing rates ($\lambda > 0$)

To explore the effect of mixing, the scaled mixing rate λ was varied over 2.5 orders of magnitude from 0.01 to 3 (see Figure 4.7), assuming equal diffusion rates of p and z ($\lambda_p = \lambda_z = \lambda$). Recall that λ is the ratio of the dimensional mixing rate $\hat{\lambda}$ to the maximum phytoplankton growth rate a. Hence, assuming $a \approx 1$ day^{-1}, a range: $\hat{\lambda} = 0.01$–3 day^{-1} is explored. The Okubo relation: $\lambda_a \approx 120L^{-0.85}$ day^{-1} (plus or minus an order of magnitude) then implies that the range of λ is relevant to grid cells in our model of scale L = 100m–10km (plus or minus an order of magnitude). The domain scale $D = 512L$ in the HRS is assumed to be the grid cell scale of a practical ocean model. Therefore, the slowest mixing results pertain to poorly-mixed grid cells of scale 500km (representative of a coarse-resolution GCM),

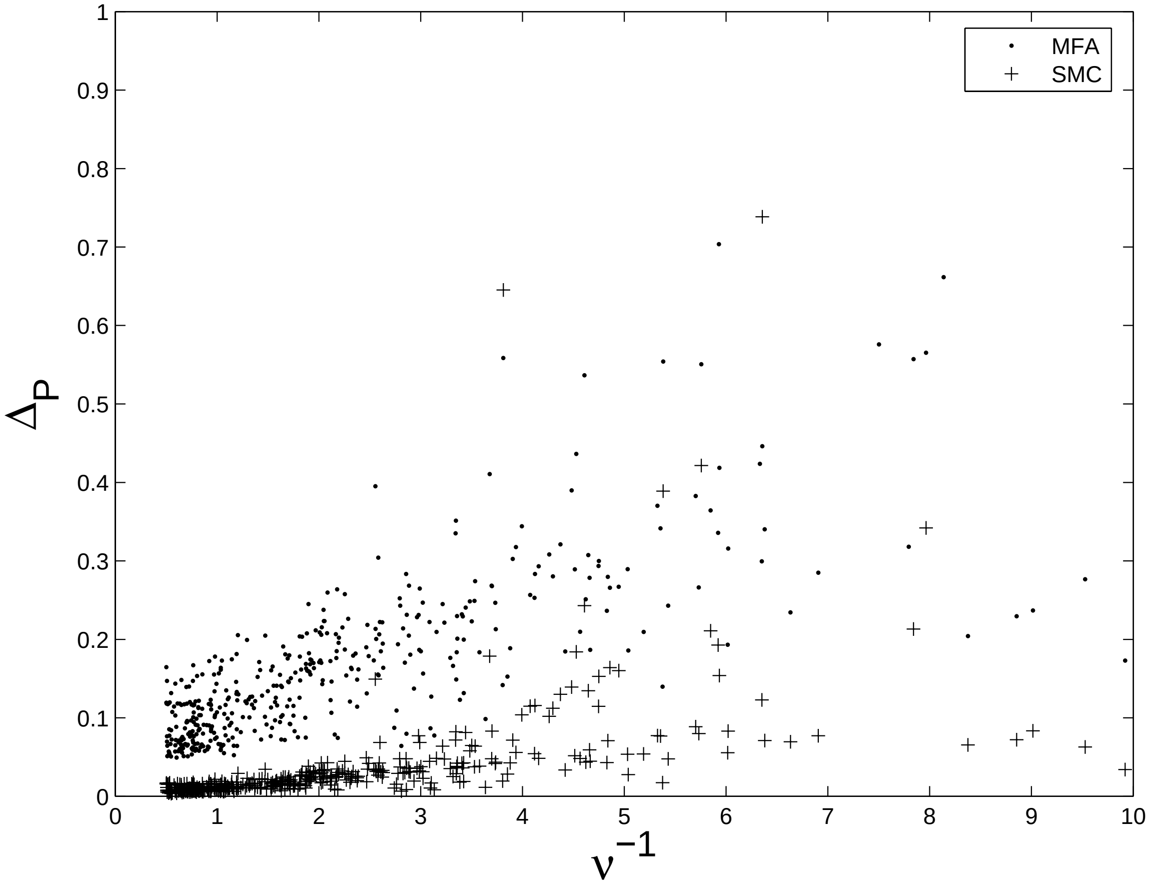

Figure 4.6: Average (rms) errors in the phytoplankton mean field over transient interval Δ_P vs. ν^{-1} for the Mean Field Approximation (dots) and Spatial Moment Closure (crosses) model, using a moderately nonlinear model with 500 sets of parameter values randomly-chosen from plausible ranges. Mixing rate $\lambda = 0$. Units are non-dimensional (see section 4.3.3).

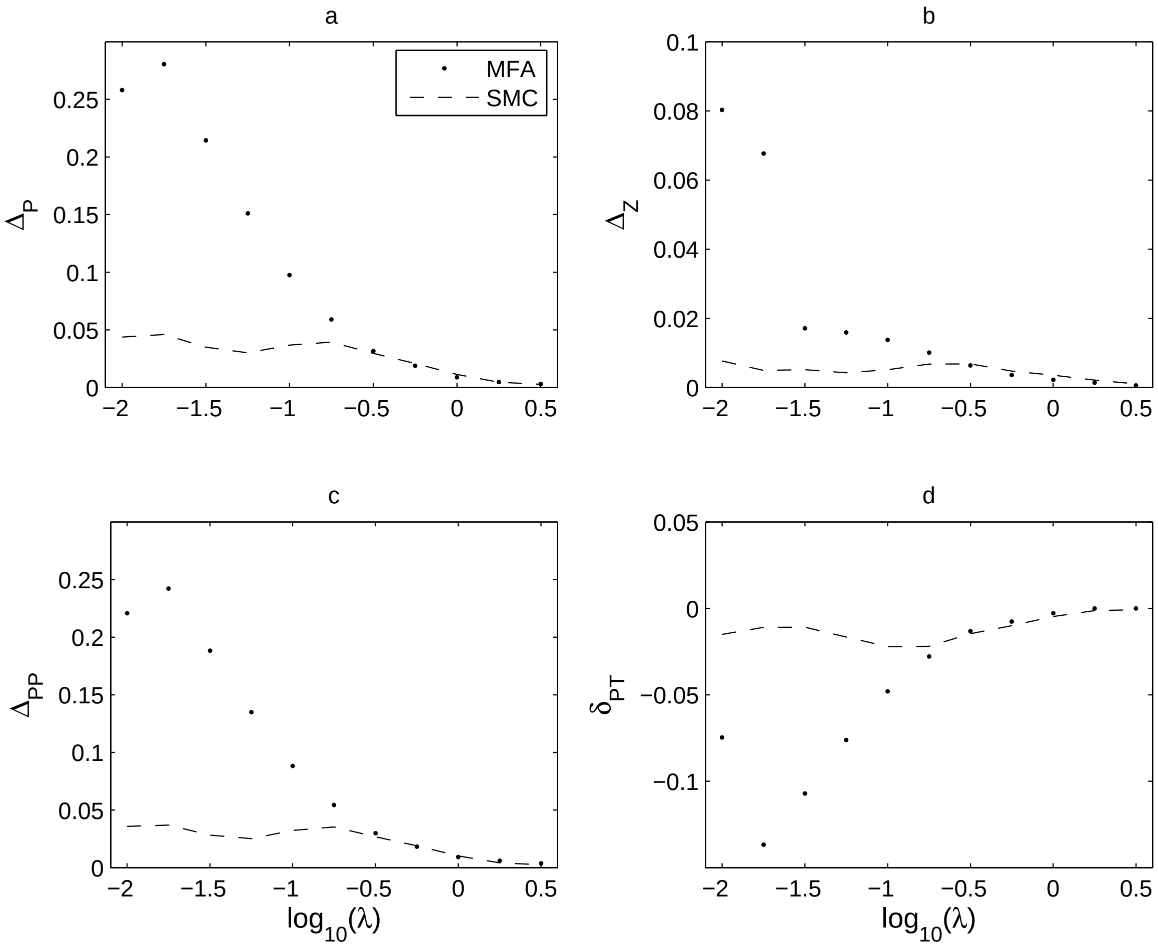

Figure 4.7: Average (rms) errors for Example 2b (moderate nonlinearity with mixing) vs. $\log_{10}\lambda$ where λ is the ratio of the mixing rate to the maximum phytoplankton growth rate, showing rms errors in the phytoplankton and zooplankton mean fields (a, b) and mean primary productivity (c), and fractional error in the total transient production (d). Results are shown for the Mean Field Approximation (dots) and Spatial Moment Closure model (dashes). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

whilst the fastest mixing results pertain to poorly-mixed grid cells of scale 5km (typical of a regional/coastal ocean model).

Figures 4.7a–d show that mixing generally reduces the error of the MFA. Errors in P (Fig. 4.7a) are largest, ranging from around 25% of the equilibrium value for $\lambda \approx 0.01$ to less than 1% for $\lambda \approx 3$, and becoming significant ($> 10\%$) at rates lower than $\lambda \approx 0.1$. This threshold corresponds to $L \approx 4$km, using the Okubo relation, hence a domain scale of 2000km. Given the order of magnitude uncertainty in the Okubo relation (and its inverse), this suggests that significant error in mean phytoplankton fields may be incurred by ocean model grid cells of scale 200km or more, when the dynamics extrapolated from smaller scales are moderately nonlinear.

Similar errors and threshold mixing rates are obtained if primary productivity is the quantity of interest (Fig. 4.7c). If only the % error in total primary production during the transient interval is important, this example suggests that lower mixing rates ($\lambda \approx 10^{-1.5} \approx 0.03$) can be tolerated (Fig. 4.7d). Curiously, the % errors in zooplankton mean fields are much smaller than those of the phytoplankton (Fig. 4.7b), even though, in this example, prey and predator populations have equal maximum growth rates.

The SMC model does a consistently good job of correcting these errors. SMC accuracy does decrease with decreased mixing, but never gets worse than 5%. Importantly, the SMC maintains significant improvement over the MFA (factor of 2 or more error reduction) for mixing rates as high as $\lambda = 0.1$. This shows that the simplistic treatment of mixing in the SMC model (neglecting interactions with lag covariances) is nevertheless a useful approximation for moderate levels of biological nonlinearity.

4.3.5 Example 3a: Strong nonlinearity with no mixing ($\lambda = 0$)

Strong nonlinearity is invoked by modifying the parameter values in (4.18) to those used in B03, where (4.18) produced 'excitable' behaviour. This means that, for some initial conditions, the transient plankton trajectories undergo large bloom-like excursions before the phytoplankton are grazed down and the system enters a 'refractory' phase, whereby the zooplankton die off and the phytoplankton slowly recover. In the scaled system (4.19), the B03 parameter values are ($\alpha = 0.05/0.43$, $\beta = 0.05 \times 0.34/0.43$, $\nu = 0.053$). Furthermore, the initial pz distribution of B03 is used: a uniform 'top hat' in pz space centred on $(P(0) = p^*, Z(0) = 0.78z^*)$ with initial value ranges $((1 - 0.657)P(0) < p < (1 + 0.657)P(0)$, $(1 - 0.277)Z(0) < z < (1 + 0.277)Z(0))$, and as before, no initial covariance or lag covariance. This in fact implies initial coefficients of variation of 0.379 and 0.160 for p and z respectively, which are within observed ranges.

The results are shown in Figure 4.8. As found in B03, the P and Z peaks in the HRS are greatly enhanced relative to the MFA (factors of 4 and 2 larger respectively — see Figs. 4.8a,b). The mean primary productivity is also enhanced (Fig. 4.8c) but only by a factor of 2, since the greatly increased P is compensated by both logistic self-limitation and the strong peak in C_{pp} (see Fig. 4.8d and recall: $PP = P(1 - P) - C_{pp}$). The slightly broader peak in P with respect to C_{pp} results in the double-peak in PP (Fig. 4.8c). The peaks in C_{pz} and C_{zz} are roughly an order of magnitude smaller than in C_{pp} (Figs. 4.8e,f).

The phase space evolution is dominated by this p variability (Figure 4.9). In fact, the distribution becomes very strongly bimodal (Fig. 4.9b — note the change in scale of the p axis), with such sharp peaks that we had to take logarithms of our 2D histogram frequencies in order to obtain visible contours. This is a consequence of the excitability threshold — over a very small range of $z(0)$, the peak values reached by the individual subsystems increases from a modest level (attained by the MFA) to a level almost an order of magnitude greater (see Richards and Brentnall, 2006). The bimodality persists as the distribution 'rotates' anticlockwise and contracts towards the equilibrium, meanwhile developing skew in the average direction of approach to equilibrium (see Figs. 4.9c,d).

The first ISR model tried was the dynamical second moment closure model in normal space (4.11 and 4.12) with the appropriate functions and constants. Unfortunately, for these parameter values and initial conditions, it blows up! P goes negative and C_{pp} increases without bound. This was always a danger with (4.11), because the fC_{pp} term in $\dot{P}$ can drive the mean phytoplankton field below zero if C_{pp} fails to decrease rapidly enough, most likely due to inaccuracy in the (zeroth order) covariance dynamics (4.12). Therefore, alternatives are sought. A casual comparison of Figs. 4.8a,b and Figs. 4.8d–f suggests an 'algebraic'

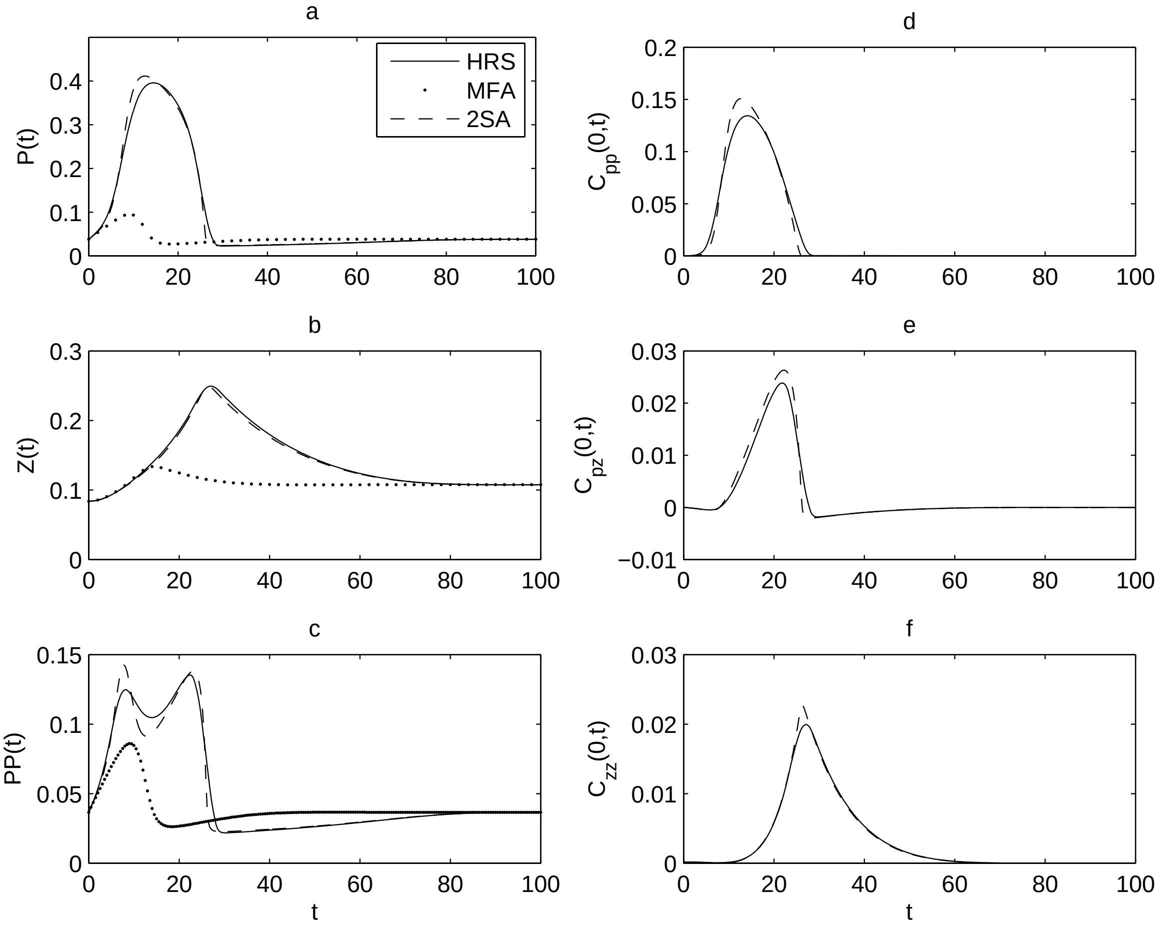

Figure 4.8: Time series for Example 3a (strong nonlinearity, no mixing), showing High Resolution Simulation (HRS), Mean Field Approximation (MFA) and 2-Spike Approximation (2SA) output for phytoplankton and zooplankton spatial mean fields (a, b), mean primary productivity (c) and spatial covariances (d–f). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

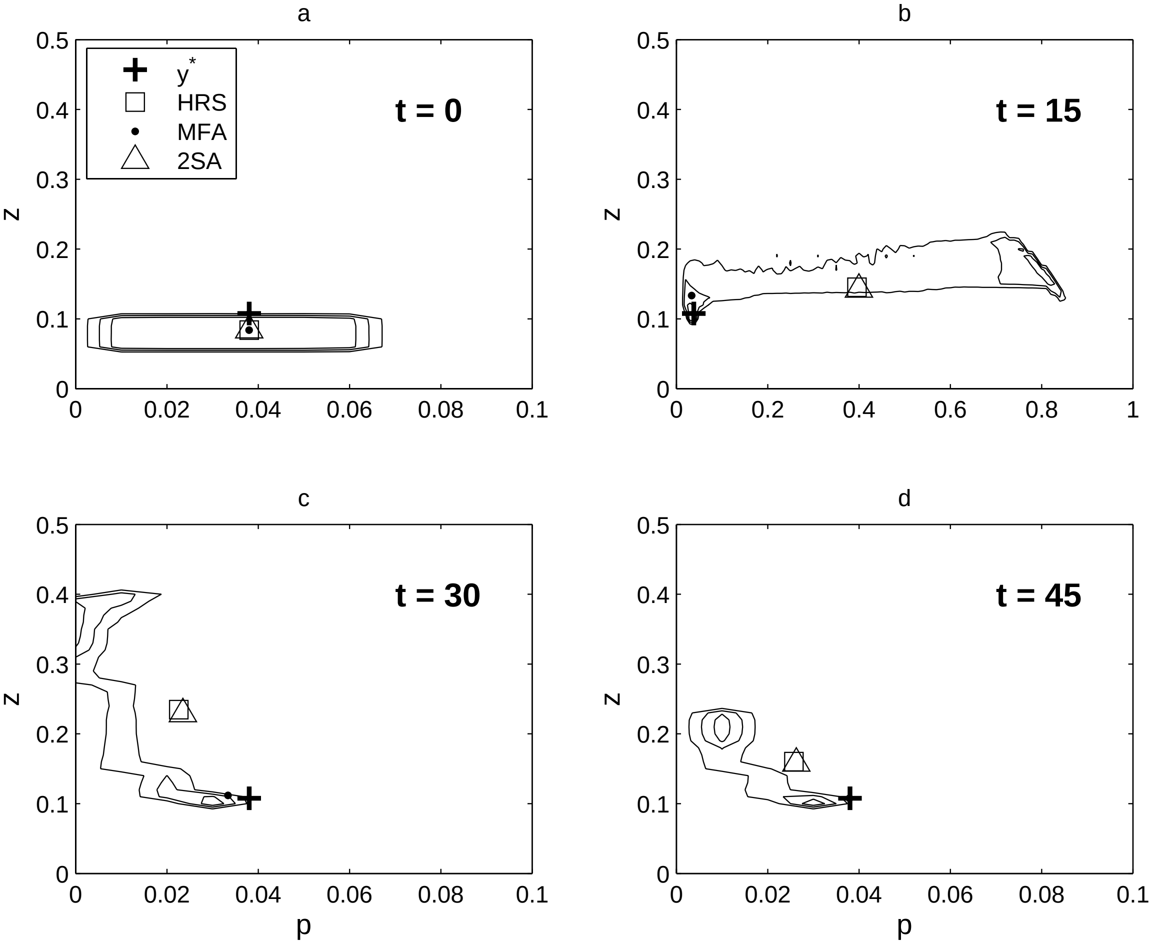

Figure 4.9: Phase Space phytoplankton-zooplankton distributions for Example 3a (strong nonlinearity, no mixing) at t = 0 (a), t = 15 (b), t = 30 (c) and t = 45 (d). Symbols mark the position of the equilibrium (cross), the mean of the High Resolution Simulation (square), the mean predicted by the Mean Field Approximation (dot), and the mean predicted by the 2-Spike Approximation (triangle). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

closure, exploiting an approximately quadratic relation between covariances and mean fields. By this method the following model is fitted to HRS output:

$$C_{ij}(0,t) = \gamma_{ij}(Y_i - y_i^*)(Y_j - y_j^*) \tag{4.20}$$

where the γ_{ij} are time-independent regression constants and the y^* are the equilibrium values. The result is ($\gamma_{pp} = 1.09$, $\gamma_{pz} = 0.90$, $\gamma_{zz} = 1.00$) with r^2 values (0.9916, 0.9973, 0.9995). Using (4.20) in combination with (4.11) did yield a stable solution, reducing Δ_P by a factor 2–3. However, this model was unsatisfactory because for $\lambda > 0$ (next section) it required re-fitting all 3 constants (or at least one) to HRS output, as well the use of an ad hoc decay constant to maintain stability, and even then showed dramatically poor performance for certain high mixing rates.

Clearly, a different approach is required. The phase space view (Fig. 4.9) suggests strong bimodality of the phase space distribution under the dynamical flow, as does the remarkably accurate approximation (4.20). To see the latter, consider a bimodal distribution consisting of two delta-function peaks (spikes), with an 'unexciting' fraction q at y_1 and a 'exciting' fraction $(1-q)$ at y_2. The variance of this distribution may be expressed as $C_{yy} = q/(1-q)(Y - y_1)^2$ where Y is the mean. Since in this example the average of the unexciting trajectories is always close to equilibrium (see Fig. 4.9), the approximation (4.20) is accurate. The implication is that the evolution of the phase space distribution might be well approximated by a '2-Spike Approximation' (2SA — see section 4.2.2). Under this approximation, which neglects all lag covariances, we have:

$$\begin{aligned}
\dot{p}_1 &= p_1(1-p_1) - \frac{p_1^2 z_1}{\nu^2 + p_1^2} + 2^D(1-q)\lambda_p(p_2 - p_1) \\
\dot{z}_1 &= \frac{\alpha p_1^2 z_1}{\nu^2 + p_1^2} - \beta z_1 + 2^D(1-q)\lambda_z(z_2 - z_1) \\
\dot{p}_2 &= p_2(1-p_2) - \frac{p_2^2 z_2}{\nu^2 + p_2^2} + 2^D q\lambda_p(p_1 - p_2) \\
\dot{z}_2 &= \frac{\alpha p_2^2 z_2}{\nu^2 + p_2^2} - \beta z_2 + 2^D q\lambda_z(z_1 - z_2)
\end{aligned} \tag{4.21}$$

where subscripts (1, 2) denote unexciting and exciting spikes respectively. Initialising (4.21) presents some difficulties, because the 2SA does not apply for small t (the HRS initial condition is a 'top hat' in phase space). Moreover, all first and second moments (P, Z, C_{pp}, C_{pz}, C_{zz}) of the HRS initial condition cannot be reproduced with only two spikes (four spikes would do it, but at the price of more complexity). However, the z initial condition is known to be a far more sensitive control on the excitability of (4.19). Therefore, the initial variance $C_{zz}(0,0)$ only is reproduced, neglecting the initial variance $C_{pp}(0,0)$. Hence, using the above formula, the unexciting fraction q is initialised at ($p_1(0) = P(0)$, $z_1(0) = Z(0) + ((1-q)C_{zz}(0,0)/q)^{1/2}$) and the exciting fraction is initialised at ($p_2(0) = P(0)$, $z_2(0) = (Z(0) - qz_1(0))/(1-q)$). The fraction q of unexciting subsystems is determined by regressing HRS output for C_{pp} on $(P - p_1)^2$, yielding $q \approx 0.52$. This fit is even better than for (4.20) (r^2 = 0.9946) and note that it only has to be done once for all λ. After running (4.21), $y_i^{(1,2)}(t)$ is used to determine mean fields $Y_i = qy_i^{(1)} + (1-q)y_i^{(2)}$ and local covariances $C_{ij} = q/(1-q)(Y_i - y_i^{(1)})(Y_j - y_j^{(1)})$, where subscripts label ecosystem components and superscripts label spikes.

The performance of the 2SA is in this case quite spectacular (Fig. 4.8). The approximation reduces the errors in P and Z by factors 11 and 21 respectively relative to the MFA ($\Delta_P^{MFA} = 371\%$, $\Delta_P^{2SA} = 33\%$, $\Delta_Z^{MFA} = 58\%$, $\Delta_Z^{2SA} = 3\%$), therefore fully justifying the approximation of the phase space distribution as two spikes. The local covariances and consequently the primary productivity are also accurately predicted.

4.3.6 Example 3b: Strong nonlinearity with mixing ($\lambda > 0$)

To explore the effect of mixing, the scaled mixing rate λ is again varied over 2.5 orders of magnitude from 0.01 to 3 (see Figure 4.7), this time assuming unequal diffusion rates of p and z for parity with B03 ($\lambda_p = 2\lambda_z = \lambda$). As showed by B03, when the mixing rate is increased from low values, the discrepancy between HRS mean fields and the MFA due to bloom enhancement initially *in*creases (Figs. 4.10a,b). Maximal discrepancies occur for mixing rates approximately $10^{-1.3} = 0.05$ times the phytoplankton growth rate. For higher mixing rates, the MFA decreases monotonically but gradually, so that there is still significant error (> 10%) even for rates of order the phytoplankton growth rate. Assuming a typical phytoplankton growth rate of 1 day^{-1}, this implies a threshold HRS grid cell scale of $L = (1/120)^{-1/0.85} \approx 280$ m using the Okubo relation. Given $D = 512L$ and the order-of-magnitude uncertainty in the Okubo relation, this implies that ocean model grid cells of size 15 km or more could incur significant error, if nonlinearity is this strong.

Errors in mean primary productivity Δ_{PP}^{MFA} follow a similar pattern to errors in mean phytoplankton Δ_P^{MFA}, since the effect of underestimating mean fields outweighs the effect of underestimating phytoplankton variance, although the opposition moderates the size of the % errors (Fig. 4.10c). Consequently, the MFA underestimates the total production during the transient interval by up to about 40% (Fig. 4.10d).

The 2SA model (4.21) accommodates the effect of mixing via an asymmetric interaction between the unexciting and exciting spikes (see section 4.2.2). As Figure 4.10 shows, the 2SA does a good job of reducing the errors by factors 4–5 at low mixing rates, up to $\lambda \approx 10^{-1.2} \approx 0.06$ times the phytoplankton growth rate. At higher rates, the 2SA model overestimates the effect of mixing, bringing it into agreement with the MFA at $\lambda \approx 10^{-0.75} \approx 0.18$. Hence the HRS becomes poorly approximated by the 2SA for $\lambda > 0.06$. Looking at the phase space distributions for $\lambda = 0.1$ (Fig. 4.11), they are in fact still bimodal but now, under the cohesive influence of mixing, they have acquired substantial ‘body’ (cf. Figs. 4.11b and 4.9b).

As might be anticipated, the dynamical SMC model starts to work where the 2SA starts to fail (see Fig. 4.10). For mixing rates higher than $\lambda \approx 0.055$, the SMC ceases to crash, and in fact slightly outperforms the 2SA at $\lambda \approx 0.06$. However, with increasing λ, it prematurely converges on the MFA even faster than the 2SA. Hence the increase in 2SA/SMC error at $\lambda \approx 0.1$ is likely a result of neglected lag covariances, rather than a breakdown of bimodality. A view of the spatial distributions of p and z at $t = 15$, for $\lambda = 0.1$, is shown in Figure 4.12. Patches are clearly still small relative to the domain scale at this mixing rate (their spatial extent increases as $\lambda^{1/2}$ — see B03), but they are large enough to violate the assumption of zero covariance at one unit lag (see Fig. 4.13). Consequently, neither approximation improves on the MFA for $\lambda > 0.18$, which, for phytoplankton growth rates of order 1 day^{-1}, corresponds

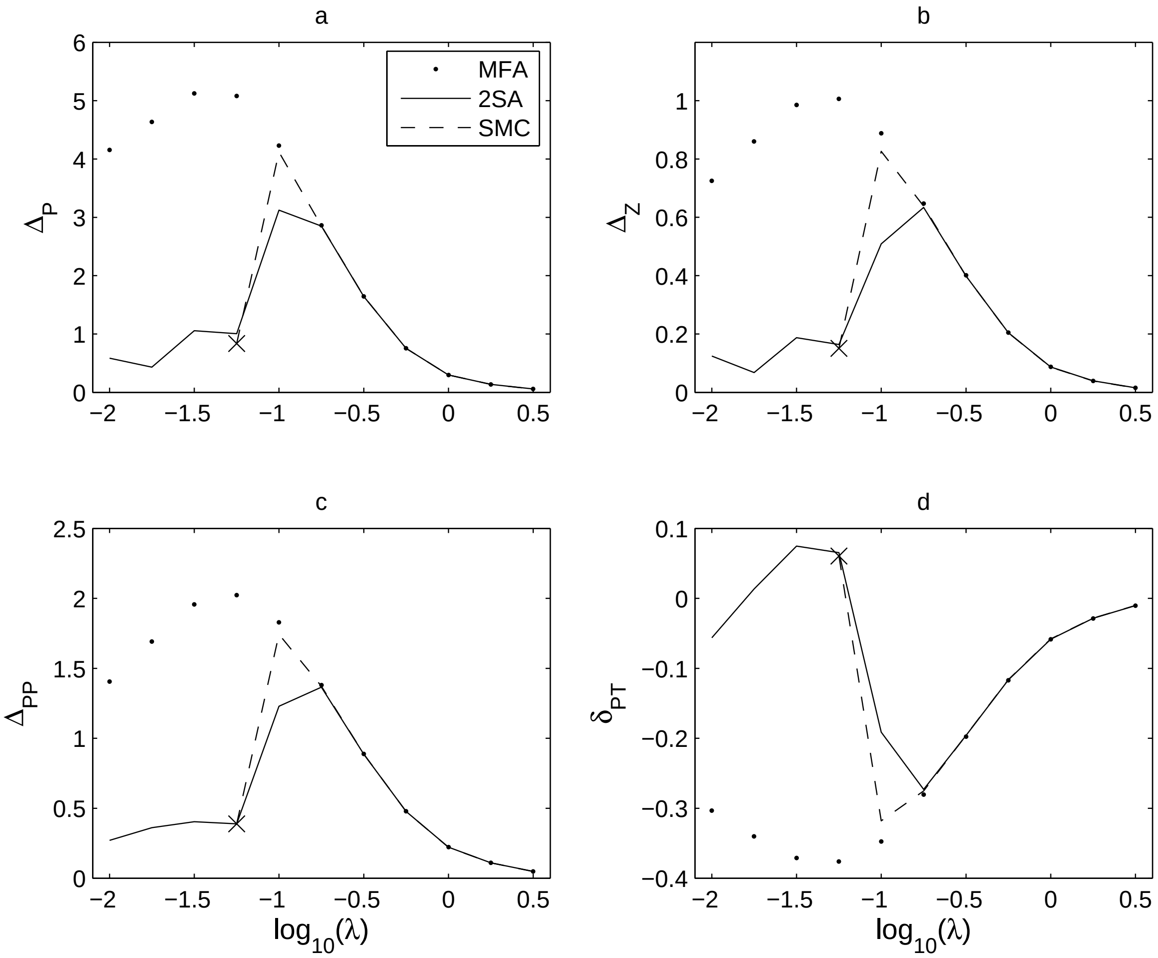

Figure 4.10: Average (rms) errors for Example 3b (strong nonlinearity with mixing) vs. $\log_{10} \lambda$ where λ is ratio of the mixing rate to the maximum phytoplankton growth rate, showing rms errors in the phytoplankton and zooplankton mean fields (a, b) and mean primary productivity (c), and fractional error in the total transient production (d). Results are shown for the Mean Field Approximation (dots), 2-Spike Approximation (solid) and Spatial Moment Closure model (dashes). The 'X' marks the lowest mixing rate for which the Spatial Moment Closure model was stable. Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

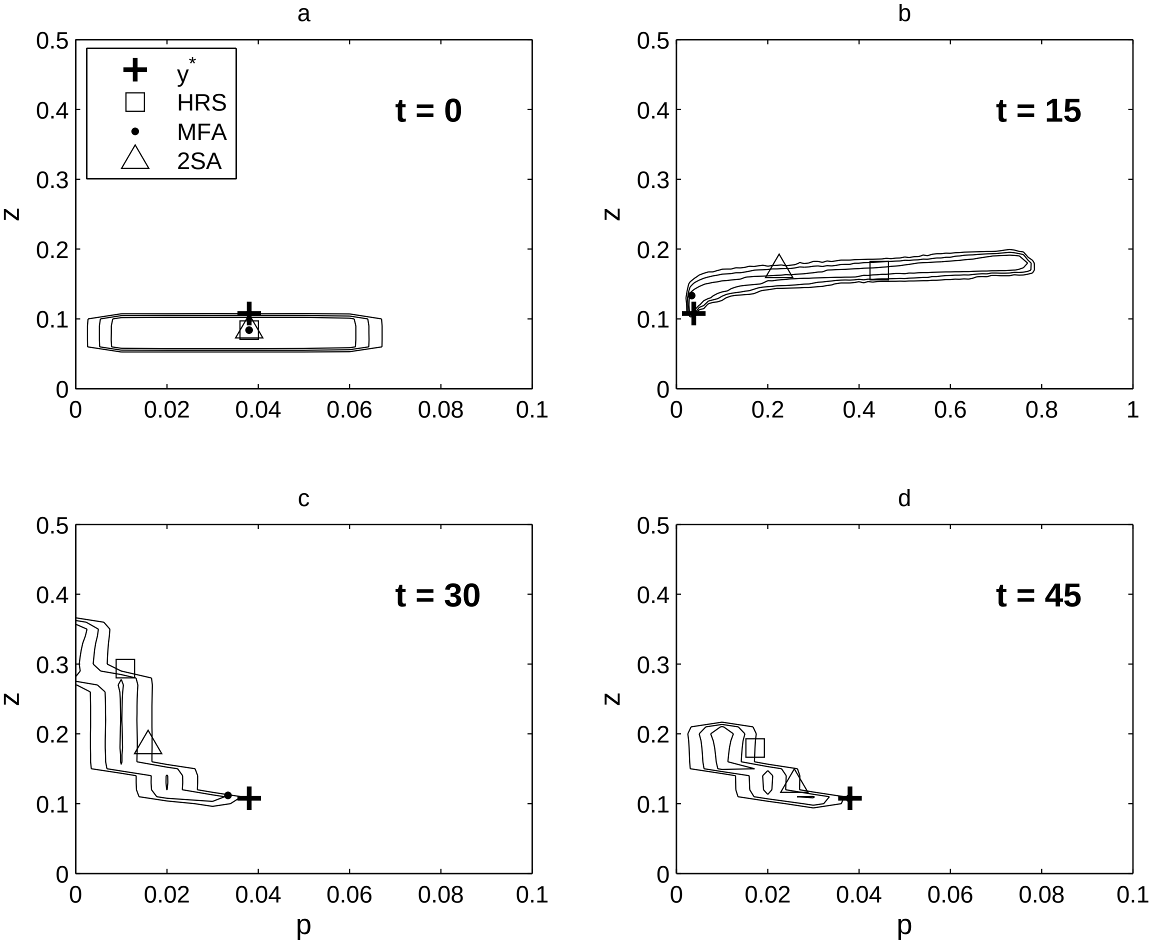

Figure 4.11: Phase Space phytoplankton-zooplankton distributions for Example 3b (strong nonlinearity with mixing rate $\lambda = 0.1$) at t = 0 (a), t = 15 (b), t = 30 (c) and t = 45 (d). Symbols mark the position of the equilibrium (cross), the mean of the High Resolution Simulation (square), the mean predicted by the Mean Field Approximation (dot), and the mean predicted by the 2-Spike Approximation (triangle). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

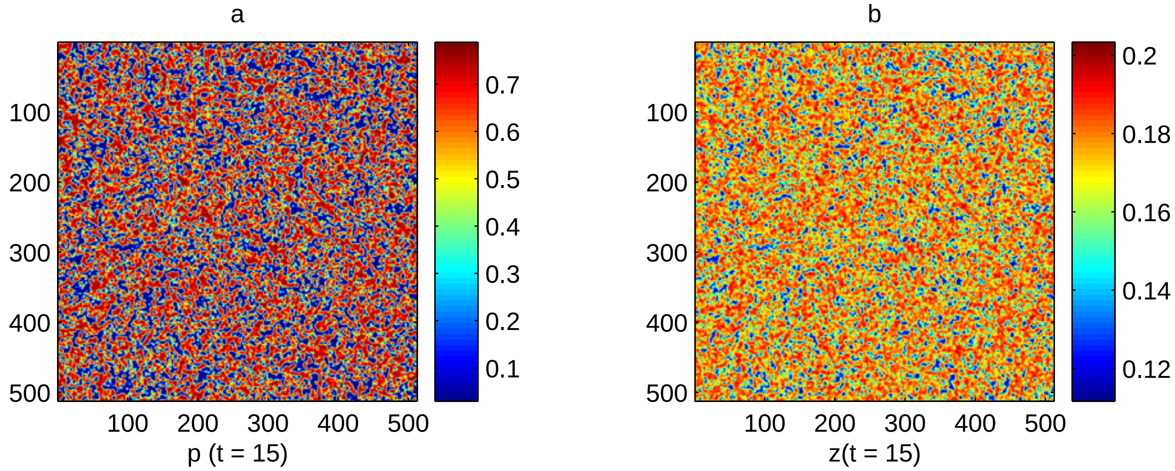

Figure 4.12: Snapshots of the reaction-diffusion model output for Example 3b (strong non-linearity with mixing rate $\lambda = 0.1$) at time $t = 15$, showing concentrations of phytoplankton (a) and zooplankton (b). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

to ocean model grid cells of scale 100–10000 km.

4.4 Discussion

The results show that a judicious choice of ISR model can significantly improve upon the MFA in approximating the domain-scale transient behaviour of simplified, two-component, reaction-diffusion simulation models. As marine ecosystem modellers, we want to know roughly how beneficial ISR is likely to be for practical plankton models. This preliminary study is much too simplified to answer this question. This section discusses these simplifications and how they may be relaxed in future extensions of this work.

Perhaps the most fundamental shortfalls of this study, as regards realism, are the assumptions of transient dynamics and transient spatial variability. Temporal and spatial variability are not, generally speaking, transient phenomena in marine ecosystems. A realistic simulation should maintain a periodic seasonal cycle and a stable heterogeneous state, ideally producing spatial power spectra which are consistent with observations over a yearly cycle. A periodic seasonal cycle is achieved by including spatially-homogeneous time dependence in the simulation model parameters (e.g. light and stratification dependence in the phytoplankton growth rate), and this time dependence is simply carried through into the ISR models. However, it is not clear that the benefits of our SMC model with transient dynamics will carry through to forced limit-cycle dynamics, where the phase space distribution may contort as it ‘rounds corners’ such that it is poorly-approximated by a compact (log)normal distribution (as in the weakly transient runs of Figure 4.3). Alternative or modified ISR models may therefore be required, perhaps exploiting approximate variability in phase (e.g. spikes along a common trajectory, or moment closure with respect to phase).

Numerous mechanisms have been proposed for stable heterogeneity or ‘patchiness’ in plankton ecosystems (e.g. Martin, 2003 for a recent review of meso-/submesoscale mechanisms). Suffice to say here that homogeneous plankton concentrations are very seldom ob-

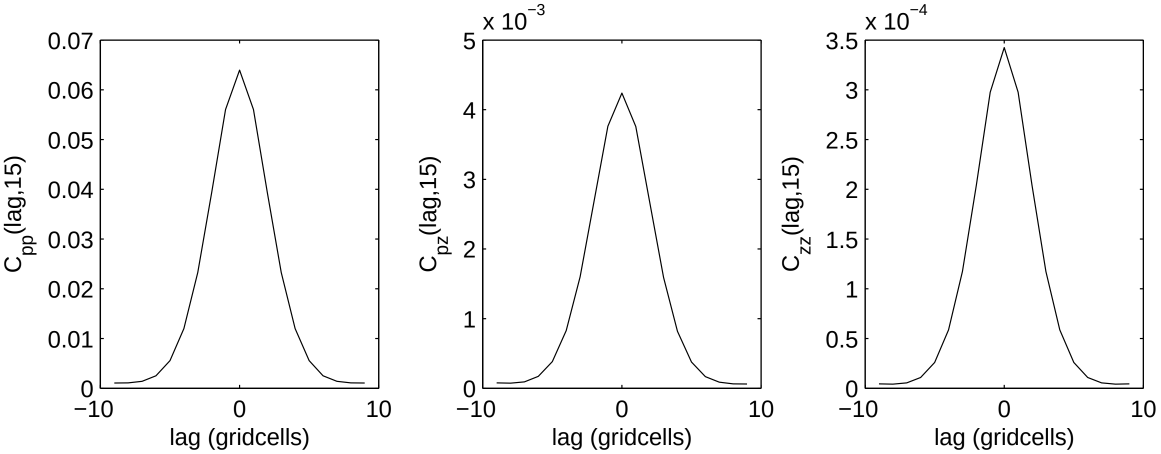

Figure 4.13: Snapshots of the lag covariance functions for the reaction-diffusion model output in Example 3b (strong nonlinearity with mixing rate $\lambda = 0.1$) at time $t = 15$, showing the phytoplankton lag covariance C_{pp}(lag,15) (a), the phytoplankton-zooplankton lag covariance C_{pz}(lag,15) (b) and the zooplankton lag covariance C_{zz}(lag,15) (c). Phytoplankton units are non-dimensional. Zooplankton concentration is scaled by a constant with units (plankton concentration unit)$^{-2}$. Time units are non-dimensional (see section 4.3.3).

served over scales of model grid cells in the ocean, and that seasonal forcing by stratification and light levels is likely to provide ample spatial variability to destabilise the homogeneous state. The ISR methods presented in this paper are easily extended to include sources of spatial variability. A spatially-variable parameter in a system of m components is simply ‘promoted to the status of a variable’, and the same ISR methods are then applied to the $(m+1)$-component system. As a trivial example, consider an exponentially-growing phytoplankton crop with a growth rate that varies randomly in the horizontal, with no mixing. At the fine scale:

$$\begin{aligned} \dot{p} &= ap \\ \dot{a} &= 0 \end{aligned} \tag{4.22}$$

The SMC model for the domain scale, neglecting third central moments, is then given by:

$$\begin{aligned} \dot{P} &= AP + C_{pa} \\ \dot{A} &= 0 \\ \dot{C}_{pp} &= 2AC_{pp} + 2PC_{pa} \\ \dot{C}_{pa} &= AC_{pa} + PC_{aa} \\ \dot{C}_{aa} &= 0 \end{aligned} \tag{4.23}$$

Note how even linear growth models become effectively nonlinear when spatial variability is included in the parameters. In fact, the dynamics of chlorophyll spatial variance as a function of covariances with light and nutrient parameters have previously been considered in Smith et al. (1996), albeit not in the context of an ISR model. Alternatively, the data may support strong bimodality of the instantaneous distribution in pa space, suggesting use of the following 2SA to derive domain-scale behaviour:

$$\begin{aligned}
\dot{p}_1 &= a_1 p_1 \\
\dot{a}_1 &= 0 \\
\dot{p}_2 &= a_2 p_2 \\
\dot{a}_2 &= 0
\end{aligned} \tag{4.24}$$

Perhaps a greater challenge is how to extend the ISR models to account for non-zero lag covariances, without invoking a high-resolution grid in lag space. This was attempted by parameterising the time-dependent ratio of one-unit to zero lag covariances $\eta_{ij}(t) \equiv C_{ij}(1,t)/C_{ij}(0,t)$, allowing closure of (4.2 and 4.3). This ratio is related to the width $\sigma_{ij}(t)$ of a Gaussian lag-covariance function (a decorrelation length) by $\eta_{ij} = \exp(-1/(2\sigma_{ij}^2))$ (space being scaled by the high-resolution grid spacing L). For the transient simulations, a robust way to do this was not found. However, in the more realistic situation where mixing is balanced by spatial variability in the forcing, spatial decorrelation lengths and hence η_{ij} may in fact be more amenable to parameterisation, perhaps based on spatial power spectra in simulations or real data fits, or perhaps even theoretical predictions based on steady or quasi-steady state assumptions.

The most conspicuous physical simplification in the simulations is the neglect of grid-scale advection, or stirring (the Fickian mixing accounting for *sub* grid-scale advection). The authors of B03 have in fact already examined how this process affects their previous findings (Richards and Brentnall, 2006). Their results seem to support the intuitive expectation that grid-scale stirring should enhance the effective mixing rate of the grid cells, by stretching incipient patches into filaments upon which mixing acts more efficiently. This would imply that the simulations in this study underestimate mixing over the model domain and hence within practical ocean model grid cells. On the other hand, sub grid-scale advection was parameterised via an empirical (Okubo) relation for *passive* tracers (based on dye experiments). Pasquero (2005) showed that the effective mixing rate, or rate of rms displacement, of a reacting tracer in a turbulent fluid is better modelled as the mixing rate at a reaction timescale after the tracer is released. This rate is *less* than the asymptotically constant (Brownian motion) rate associated with a passive tracer, implying that the passive scalar parameterisation of sub grid-scale mixing overestimates mixing of plankton due to turbulent advection at these scales. It is not obvious that neglecting these two opposing effects should bias the results beyond the order-of-magnitude uncertainty inherent in the Okubo relation.

In terms of biological complexity, two-component (pz) systems are very rarely considered adequate for realistic plankton ecosystem simulations (four-component $npzd$ models are generally considered minimal). ISR model complexity can be predicted in terms of the number of dynamical variables. The nSA model has nm components, and therefore its complexity increases only linearly with m. However, the SMC model with all first and second moments dynamically resolved has $m(m+3)/2$ components, so its complexity increases quadratically. This could make the SMC approach impractical for highly complex systems, especially if the larger number of variables increases the chance of dynamically aberrant behaviour (see section 4.3.5). Another effect of having $m > 2$ is a greater potential for very strongly nonlinear — even chaotic — dynamical behaviour at the subsystem level. This is liable to increase the errors in both MFA and ISR models, although complexity of individual subsystem trajectories does not necessarily imply a complex evolution of the phase space distribution, especially if stochasticity is also included (see below).

As argued in section 4.1, there are reasonable grounds for assuming deterministic population dynamics for lower trophic-level plankton at averaging scales of perhaps 0.1m and larger. However, it may yet be desirable to include stochasticity as a representation of fine-scale environmental or demographic fluctuations, in which case the SMC method may be generalised (Rodriguez and Tuckwell, 1996). Rather than approximating the Louiville dynamics, as in this study, the SMC would then approximate the Fokker-Planck dynamics. Stochasticity may 'smear out' the phase space distribution enough to prohibit successful nSA models. Conversely, it could play to the advantage of the SMC model by improving smoothness and (log)normality of the phase space distribution (cf. Example 3a).

Finally, note that the particular distributional or moment closure assumptions underlying the ISR models in this study are not essential. The aim here was merely to show that successful ISR approximations are *possible* for plankton HRSs via certain general methods. For example, the neglect of third and higher central moments in the SMC model may lead to predictions of *negative* third non-central moments, and for this reason might be exchanged for an 'asymmetric' or 'higher power' closure assumption. On the other hand, this 'reasonable first guess' assumption at least has a 'physical' interpretation in terms of compactness and normality of the phase space distribution, so that it may be suggested by phase space data. It is difficult to say what a distribution which obeys, for example, the Kirkwood superposition approximation should look like in phase space (Murrell et al., 2004).

4.5 Conclusions

The process of turbulent mixing in plankton ecosystems may allow extrapolation of the population dynamics over a range of spatial scales. If a region of ocean is sufficiently well mixed, the 'mean field approximation', whereby population dynamics valid at the laboratory scale are applied to regional mean fields, is accurate. However, if the mixing rate is slow enough, the fine-scale spatial variability is large enough, and the nonlinearity in the fine-scale plankton dynamics is strong enough, the regional mean fields can exhibit an 'emergent' behaviour which is very poorly predicted by the mean field approximation. Brentnall et al. (2003) gave a simple example of a transient, spatially-variable, two-component reaction-diffusion system where low diffusion rates produced mean field blooms a factor of 4-5 larger than those predicted by the mean field approximation. This discrepancy was investigated here for similar transient, two-component, 2D reaction-diffusion systems with different plankton model formulations and parameter values (resulting in different 'strengths' of biological nonlinearity) and different rates of sub grid-scale mixing ('eddy-diffusion'). Significant failure of the mean field approximation, due to 'biological Reynolds fluxes', was quite generic, although usually less extreme than in Brentnall et al. (2003), given realistic initial spatial variances and sub grid-scale mixing rates applicable to practical ocean models.

Spatially-implicit models were investigated to better approximate the coarse-scale dynamics by making statistical assumptions about the phase space distribution of fine-scale sub-populations. The main interest was is in spatial moment closure models, whereby the mean field approximation is extended to include corrections due to higher order moments, coupled to the mean fields by dynamical nonlinearity. A closed system was obtained by neglecting the contributions of lag covariances as well as third and higher central moments, effectively assuming a compact normal distribution in phase space. A similar approximation was derived for lognormal statistics. These approximations accurately predicted phytoplankton/zooplankton mean fields and spatial covariances, as well as bulk primary productivity, assuming weak-to-moderate strengths of biological nonlinearity, and over the full range of sub grid-scale turbulent mixing rates relevant to ocean models. However, they were less accurate in cases of longer transience and strong nonlinearity, when they may become unstable. An example of the latter is the excitable system of Brentnall et al. (2003) at low mixing rates. Here, an alternative 'two-spike approximation', assuming a strongly-peaked bimodal distribution comprising 'exciting' and 'unexciting' fractions, was found to be accurate. At high mixing rates, both spatially-implicit models as well as the mean field approximation were inaccurate in this example, mainly due to the neglect of lag covariances.

Inferences regarding the likely benefits of these methods for realistic plankton models are limited by the simplifications employed in this study, notably: the use of transient dynamics rather than seasonally-forced behaviour, the neglect of sources of spatial variability which might stabilise decorrelation lengths, the neglect of advection, and the restriction to only two reacting components. The methods described here may, however, be extended to accommodate these factors in future investigations.

Appendix 4A: Spatial Moment Dynamics in Discrete Space

Consider a 1D chain of N identical, subsystems each with m components interacting via Fickian diffusive mixing with their nearest neighbours. The dynamics of the i^{th} component of the n^{th} subsystem obeys:

$$\dot{y}_i^{(n)} = F_i(y^{(n)}) + \lambda_i(y_i^{(n-1)} + y_i^{(n+1)} - 2y_i^{(n)}) \tag{4A.1}$$

where F_i is some generally nonlinear function of y describing the biological interactions. It is assumed that (4A.1) applies to every subsystem including those at the boundaries (hence assuming periodic boundary conditions) or that N is large enough and the boundary fluxes are small enough that the effect of boundary populations on the mean dynamics can be neglected. $F_i(y^{(n)})$ can be expressed as an m-dimensional Taylor expansion about the spatial mean field $Y \equiv N^{-1}\sum_{n=1}^{N} y^{(n)}$ and the i^{th} component of the n^{th} subsystem can be expressed as a sum of the mean field Y_i plus some (not necessarily small) deviation $\delta y_i^{(n)}$:

$$\dot{Y}_i + \delta\dot{y}_i^{(n)} = F_i(Y) + \sum_{r=1}^{\infty} \frac{1}{r!}\frac{\partial^{(r)} F_i(Y)}{\partial y_{s_1}\ldots\partial y_{s_r}}\delta y_{s_1}\ldots\delta y_{s_r} + \lambda_i(\delta y_i^{(n-1)} + \delta y_i^{(n+1)} - 2\delta y_i^{(n)}) \tag{4A.2}$$

using the convention that identical indices are implicitly summed over. Averaging over space has no effect on terms involving the spatial mean field, eliminates terms involving a single power of δy, and converts the r^{th} powers of δy into r^{th}-order central moments $M^{(r)}$. Hence:

$$\dot{Y}_i = F_i(Y) + \sum_{r=2}^{\infty} \frac{1}{r!} \frac{\partial^{(r)} F_i(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} M^{(r)}_{s_1 \dots s_r}(\underline{0}^{(r-1)}, t) \qquad (4A.3)$$

using $\underline{0}^{(r-1)}$ to denote a vector of $(r-1)$ zeros describing the lag coordinates of the r^{th}-order moment, denoted $M^{(r)}$. Note that the physical interaction terms have no direct effect on the mean field, because linear diffusive mixing is assumed. So far, no approximation has been made. The dynamics of a system of N sub-populations are being transformed to an infinite system or 'hierarchy' of spatial moments. A 'closure assumption' is required to truncate this hierarchy. For instance, assuming that the net effect of the terms involving third and higher moments on the mean field dynamics is negligible yields equation (4.2):

$$\dot{Y}_i \approx F_i(Y) + \frac{1}{2} \frac{\partial^2 F_i}{\partial y_j \partial y_k} C_{jk}(0, t)$$

To derive the dynamics of the higher moments, $\dot{\delta y}_i^{(n)}$ is obtained by subtracting (4A.3) from (4A.2):

$$\dot{\delta y}_i^{(n)} = \sum_{r=1}^{\infty} \frac{1}{r!} \frac{\partial^{(r)} F_i(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} (\delta y_{s_1} \dots \delta y_{s_r} - M^{(r)}_{s_1 \dots s_r}(\underline{0}^{(r-1)}, t)) + \lambda_i (\delta y_i^{(n-1)} + \delta y_i^{(n+1)} - 2\delta y_i^{(n)}) \qquad (4A.4)$$

Consider, for example, the zero-lag covariances: $C_{kj}(0,t) \equiv N^{-1} \sum_{n=1}^{N} \delta y_k^{(n)} \delta y_j^{(n)}$. These obey $\dot{C}_{kj}(0,t) = N^{-1} \sum_{n=1}^{N} (\dot{\delta y}_k^{(n)} \delta y_j^{(n)} + \delta y_k^{(n)} \dot{\delta y}_j^{(n)})$. Therefore, to obtain $\dot{C}_{ij}(0,t)$, we multiply (4A.4) by δy_j, average over space (eliminating the already-present moment terms), and add the same again swapping i and j, yielding:

$$\begin{aligned} \dot{C}_{ij}(0,t) &= \sum_{r=1}^{\infty} \frac{1}{r!} \left(\frac{\partial^{(r)} F_i(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} M^{(r+1)}_{s_1 \dots s_r j}(\underline{0}^{(r)}, t) + \frac{\partial^{(r)} F_j(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} M^{(r+1)}_{s_1 \dots s_r i}(\underline{0}^{(r)}, t) \right) \\ &- 2(\lambda_i + \lambda_j)(C_{ij}(0,t) - C_{ij}(1,t)) \qquad (4A.5) \end{aligned}$$

using the symmetries $C_{ij} = C_{ji}$ and $C_{ij}(1,t) = C_{ij}(-1,t)$ to simplify the last term. The generalisation to a D-dimensional grid consists of replacing the factor of 2 in the mixing term with a factor 2^D. Note the appearance of the spatial covariances at one unit spatial lag $C_{ij}(1,t) \equiv N^{-1} \sum_{n=1}^{N} \delta y_k^{(n)} \delta y_j^{(n+1)}$, where the spatial unit here is the separation of adjacent populations. Whilst the nonlinear biology couples zero-lag moments of different order, the linear mixing couples same-order moments of different spatial lag. The assumption of local biological interactions causes the biological dynamics in the moment equations to be uniform in lag space. This implies a set of discrete reaction-diffusion equations for the covariance dynamics in lag space with a uniform reaction term:

$$\begin{aligned} \dot{C}_{ij}(1,t) &= \sum_{r=1}^{\infty} \frac{1}{r!} \left(\frac{\partial^{(r)} F_i(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} M^{(r+1)}_{s_1 \dots s_r j}(\underline{0}^{(r-1)}, 1, t) + \frac{\partial^{(r)} F_j(Y)}{\partial y_{s_1} \dots \partial y_{s_r}} M^{(r+1)}_{s_1 \dots s_r i}(\underline{0}^{(r-1)}, 1, t) \right) \\ &+ (\lambda_i + \lambda_j)(C_{ij}(0,t) + C_{ij}(2,t) - 2C_{ij}(1,t)) \qquad (4A.6) \end{aligned}$$

If the net contribution of terms involving third and higher order moments is neglected in (4A.5), we obtain (4.3):

$$\dot{C}_{ij}(0,t) \approx \frac{\partial F_i}{\partial y_k} C_{kj}(0,t) + \frac{\partial F_j}{\partial y_k} C_{ki}(0,t) - 2(\lambda_i + \lambda_j)(C_{ij}(0,t) - C_{ij}(1,t))$$

Appendix 4B: Spatial Moment Dynamics in Continuous Space

The spatially-continuous version of (4A.1) is:

$$\dot{y}_i = F_i(y) + \kappa_i \frac{\partial^2 y_i}{\partial x^2} \tag{4B.1}$$

where $\kappa_i = L^2 \lambda_i$ is the eddy-diffusivity. This leads to the same equation (4A.3):

$$\dot{Y}_i = F_i(Y) + \overline{\delta F_i} \tag{4B.2}$$

where the bar denotes an average over space (x), $Y \equiv \overline{y}$, and $\delta F \equiv F(y) - F(Y)$ denotes the dynamical discrepancy from the MFA, expressed as a Taylor series in (4A.2). Hence the equivalent of (4A.4) is:

$$\dot{\delta y_i} = F_i(y) - F_i(Y) - \overline{\delta F_i} + \kappa_i \frac{\partial^2 y_i}{\partial x^2} \tag{4B.3}$$

Now defining the lag covariances $C_{ij}(l,t) \equiv \overline{y_i(x) y_j(x+l)}$ leads to:

$$\dot{C}_{ij}(l,t) = B(l,t) + \kappa_j \overline{y_i(x,t) \frac{\partial^2 y_j(x+l,t)}{\partial x^2}} + \kappa_i \overline{y_j(x+l,t) \frac{\partial^2 y_i(x,t)}{\partial x^2}} \tag{4B.4}$$

writing the sum of terms due to the biological dynamics, given in (4A.5), as $B(l,t)$. Integrating the second and third terms once by parts, and using the periodic boundary conditions yields:

$$\dot{C}_{ij}(l,t) = B(l,t) - (\kappa_i + \kappa_j) \overline{\frac{\partial y_i(x,t)}{\partial x} \frac{\partial y_j(x+l,t)}{\partial x}} \tag{4B.5}$$

Note that for $i = j$ and $l = 0$, the mixing term simplifies to $-2\kappa_i \overline{(\partial y_i / \partial x)^2}$, which is analogous to the 'molecular smearing' term for scalar variance (Tennekes and Lumley, 1972). Integrating the mixing term again by parts, changing the sign and shifting both derivatives onto $y_j(x+l,t)$, and swapping x-derivatives for l-derivatives, yields:

$$\dot{C}_{ij}(l,t) = B(l,t) + (\kappa_i + \kappa_j) \frac{\partial^2 C_{ij}(l,t)}{\partial l^2} \tag{4B.6}$$

Hence, for a passive scalar ($F = B = 0$) with delta function initial lag-covariance density $C_{ij}(l,0) = C^0_{ij} \delta(0)$, the solution of (4B.6) is a Gaussian with a linearly increasing variance in lag space:

$$\frac{C_{ij}(l,t)}{C^0_{ij}} = \frac{1}{\sqrt{4\pi(\kappa_i + \kappa_j)t}} \exp\left(\frac{-l^2}{4(\kappa_i + \kappa_j)t} \right) \tag{4B.7}$$

The solution in D dimensions is just a product of these solutions in each dimension l_x, l_y etc.

Chapter 5

Conclusions

5.1 Overview

The overall aim of this thesis was to investigate new (within plankton ecology) statistical methods for mitigating the effects of unpredictable spatial variability on plankton modelling. Clearly, within this broad remit, the investigations could not be exhaustive. Nevertheless, as discussed below, they have exposed some specific dangers and revealed some promising new solutions. As a necessary complement to this work, more rigorous methods were developed for testing and inter-comparing plankton models and prediction methods (such as those proposed herein) with real data. Although highly simplified plankton models were used, the methods are general and may, with suitable adaptation, be of benefit in a wide range of plankton models, especially as computational resources continue to grow. The most general conclusion then, is that plankton modellers can and should exploit statistical methods to account for limitations on their ability to predict spatial variability.

In the following, basic findings are first summarised, followed by their primary limitations and importance in a wider context, referring the reader to individual chapters for more detailed discussion. Promising avenues for extensions of this work are also suggested.

5.2 Basic Findings

- If unpredictable spatial variability may be approximated by time lags (as appears to be sometimes the case), this property may be exploited in a 'Lagrangian' model-data fit to recover biological parameter estimates and a dynamical trajectory applicable to a single, 'typical' fine-scale parcel of fluid. The resulting 'variable lag fit' greatly increases the effective resolution of the model without suffering from unpredictable model inputs.
- An alternative approach of fitting 'coarse-scale' biological parameters with weak prior constraints, using a simple bulk plankton model and data from Georges Bank, fails to produce more skillful predictions than a strictly empirical model, at least if 100km-scale boundary fluxes are neglected. This is demonstrated using a new skill assessment method which distinguishes calibration from validation data sets and also replicates them over a simulated ensemble of data sets generated by a fitted statistical model.
- Unpredictable spatial variability can have serious impacts on grid cell-scale plankton dynamics via 'biological Reynolds fluxes' due to nonlinearity in the sub grid-scale biolog-

ical dynamics. More accurate coarse-resolution plankton models can be formulated by parameterising these fluxes, using ‘spatial moment closure’ and other methods which model the evolution of the distribution of sub grid-scale plankton concentrations in phase space.

5.3 Limitations and Wider Importance

By assuming fine-scale spatial variability in the form of time lags, Chapter 2 showed that a ‘variable lag fit’ method mitigates the problems of unpredictable model inputs and spatial undersampling in deriving fine-scale biological parameter estimates from ocean data. However, it is unclear how applicable the assumption of time lag variability or ‘phase relation’ might be to real plankton ecosystems. Comparing the time series vs. phase space data in Figure 2.1 does suggest intuitively that time lags are important for that particular survey area in the North Atlantic, but their statistical significance in these data has not been quantified (this would require use of the methods developed in Chapter 2, as discussed below). Moreover, the synoptic output of the model used in Srokosz et al. (2003) does not suggest any preferential alignment of the data spread along the phase space trajectory. If, as this model output suggests, the alignment in real data is just a consequence of the sampling asynopticity, then the variable lag fit will be no better (in fact worse) than the standard method.

However, even if time lags are a poor model for fine-scale spatial variability, the lag-fitting method may yet, with suitable adaptation, be useful in other ways. First, the variable lag fit can be regarded as an example of a general strategy to simplify and perhaps better model the statistics of errors in plankton model dynamics and initial/boundary conditions. Previous studies have attempted to fit these errors separately, but this generally entails neglect of one or more error sources and/or strong, somewhat *ad hoc* assumptions about error statistics, often motivated by computational concerns and smoothness of solution (e.g. Robinson, 1996; Anderson et al., 2000; Losa et al., 2003, 2004; Li et al., 2006; Smith et al., submitted; Nerger and Gregg, 2008; Gregg, 2008). It may prove easier to obtain empirical estimates of error statistics for lags/phase errors or other ‘distortions’ which exploit low-dimensional patterns in model-data misfit. Atmospheric weather forecasters have begun to investigate this as a possible way of incorporating into model-fitting/skill-assessment the pattern similarity evaluation process which occurs automatically in the brains of subjective human experts, when the information is presented in the right way (Hoffman et al., 1995). Another way of making model-data discrepancy metrics ‘suitably forgiving’ might be to formulate cost functions involving Fourier spectral densities or other spatio-temporal statistics such as spatial covariances (see below) or principal components. Although such approaches do not appear to have found utility in atmospheric weather forecasting or physical oceanography, they may yet be useful to plankton modelling seeking to constrain poorly-known biological parameters.

Chapter 3 found that a ‘weak prior’ bulk plankton process model generally could not outperform a strictly empirical model in forecast mode. However, the weak prior bulk modelling approach could not be refuted in view of the limitations of the study, including the potential importance of neglected net plankton boundary fluxes. If 100km-scale net boundary fluxes of phytoplankton cannot be neglected, estimates must be provided by observations or a larger-scale model (i.e. ‘nesting’ or ‘downscaling’). In lieu of such information, a better test site for the method might have been in the open ocean where net boundary fluxes may be less important, rather than an exceptionally productive shelf sea area like Georges Bank.

Nevertheless, the study demonstrated some important characteristics of the weak prior method which are likely to be robust to such limitations. The ensemble-generated parameter uncertainty estimates were about 50–100% in all parameters, much larger than generated by the inverse Hessian method (still popular for its computational efficiency despite inaccuracy with poorly constrained, nonlinear model fits, as demonstrated by Vallino (2000) and Schartau (2001)), and much more accurate on the basis of twin tests. Two of the seven fitted parameters were unconstrainable in the sense of being highly correlated (Matear, 1995), but probably because of the lack of microzooplankton data. Hence, three half-years of chlorophyll data were insufficient to properly constrain the seven parameters of a weak-prior bulk model, partly due to the weak prior constraints, and partly due to uncertainty in the nutrient and mesozooplankton forcings (even at the 100km scale).

Chapter 3 also demonstrated a systematic approach to defining and then estimating predictive skill, using both data 'division' (or 'cross-validation') and data 'simulation' (or 'parametric bootstrap') techniques. The first step of mathematically defining the predictive objectives via a 'skill function' guided subsequent calibration and skill assessment methods. Current debate on 'correct' choices of cost function (e.g. Evans, 2003) might be better informed if practitioners specified the sense in which they want their estimates/predictions to be 'accurate'. Similarly, the choice of appropriate skill metrics is also informed by the skill function and its specific inter-/extrapolative demands. It appears relatively easy to ensure that skill metrics are independent i.e. not biased by covariances between calibration and validation data errors. Much more difficult is to devise skill tests which are representative of the inter-/extrapolative demands of the skill function (a particular challenge for climate modellers at present). Accurate data simulation models extend our ability to assess model skill, allowing us to distinguish bias and variance in the model's predictive error distribution at various levels of extrapolation, and then act accordingly to improve the model formulation and/or estimation method. Also, they provide independent testbeds in which the skill of *all* models and prediction methods may be objectively measured and inter-compared. At present, there is an urgent need to prove objectively that mechanistic plankton models can provide better extrapolative skill than strictly empirical models (e.g. a Gaussian process or spatiotemporal 'objective analysis').

The simulated annealing optimisation method used in Chapters 2 and 3 may be too computationally expensive for use outside 0D models at present — O(10^6) iterations were used to fit a single data set. However, the 'genetic' algorithm global optimisation method has been successfully implemented with 1D (Schartau and Oschlies, 2003) and 3D (Huret et al., 2007) models. Moreover, there was surely useful information in the O(10^6) 'suboptimal' runs discarded by our method. This might be utilised in a Monte Carlo integrations involving the posterior distribution, allowing 'posterior mean' estimates of parameters/trajectories which may be more robust than the 'posterior mode' estimates employed in this thesis (Harmon and Challenor, 1997). Furthermore, such integrations may provide accurate Bayesian estimates of parameter and prediction statistics (over a 'true' ensemble) without requiring a simulated ensemble of data fits. However, such a shortcut stands in need of validation by realistic twin tests, accounting for misspecification of the dynamical model and measurement error hypothesis. The suboptimal runs may also inform a multi-dimensional, nonlocal model 'sensitivity' (cf. Schartau, 2001), perhaps allowing automatic fixing of insensitive parameters over the course of the parameter search.

Chapter 4 investigated an intuitively more attractive alternative to the weak prior approach (sometimes referred to pejoratively as 'data-fitting') which less invokes the unpredictable model input problem encountered by high explicit spatial resolution. This alternative is statistical implicit spatial resolution: parameterising the mean effects of unresolved variability on coarse-scale dynamics. In principle, this approach would allow the results of multi-species laboratory and 10m-scale mesocosm experiments (e.g. Vallino, 2000) to be imposed as *strong* priors on the ocean data fit, or indeed the experimental and ocean data might be fitted simultaneously, in both cases mitigating the underconstraint problem in plankton model fitting. Furthermore, if the parameterisation is sufficiently mechanistic, it may extrapolate better beyond fitted data than a 'weak prior' model where coarse spatial resolution is compensated by adjustment of biological parameters.

In particular, the effects of hitherto-unparameterised 'biological Reynolds fluxes' on grid cell-scale plankton ecosystem dynamics were examined and parameterised using spatial moment closure and '*n*-spike approximation' methods. The simulation study of Chapter 4 was highly simplified, most notably limited by the neglect of grid-scale advection and restriction to transient dynamics. It was not, therefore, a credible means to quantify the significance of biological Reynolds fluxes, or the expediency of the parameterisations, in real plankton ecosystems. Nevertheless, the fact that the parameterisation methods were robustly effective in this artificial scenario is encouraging.

In practice, of course, a spatially-implicit model will still need to be fitted to ocean data to account for the integration over species and physiological adaptations, and yes, the parameterisations will require more parameters to be fitted. Is this approach therefore subject to the criticism made in the Introduction, of approaches which assume that parameter-hungry methods successful in physical model fitting will work for plankton ecology, and which are in fact likely to exacerbate the underconstraint problem? There are two key differences. First, the dynamics of the coarse-scale mean quantities, derived by the moment expansion, and by assumption the focus of predictive effort, guide the choice of a parsimonious set of degrees of freedom for the model from the hierarchy of moment dynamics. Hence the model can be formulated to best fulfil its predictive purpose, rather than to produce output which 'looks more realistic'.

Second, by fitting spatial moment dynamics, more degrees of freedom in the data can be fitted by the *same* biological parameters, without invoking the problem of deterministic unpredictability at the small scale. Fitting higher spatial moments would require accounting for the non-independence of mean field and spatial covariance data, the use (or estimation) of some extra erroneous model inputs, and possibly estimation of a few 'closure parameters'. Nevertheless, it seems possible that the net effect could be increased constraint on biological parameters. The data set required by such a model should include high-resolution sampling in at least one 'typical' subregion of the sampled region, such that sub grid scale covariance dynamics might be extrapolated from subregion to region. Fitting moment dynamics could also be implemented in a high explicit resolution model, as an extension of the mean field fit. This may, however, raise issues of computational practicality, and fail to accurately constrain ‘surplus’ degrees of freedom in the higher explicit resolution model.

In summary, plankton modellers can only really hope to predict the *statistical* effects of fine-scale spatial variability over extended timescales. Existing methods of formulating, calibrating and validating plankton models do not sufficiently acknowledge this fact. It should therefore be no surprise that new statistical methods have much potential to improve this situation by accounting for unpredictable spatial variability in plankton modelling. The methods investigated herein reconcile current enthusiasm for deterministic data assimilation and prediction with an older remit of achieving ‘statistical’ agreement with data. They also demonstrate the utility of considering plankton ecosystems as dynamical systems in ‘phase space’. However, much more testing and development is required to bring proven benefits in general applications. In addition, use of these methods will require feedback to observationalists on how to design better surveys to meet the new modelling demands. Hopefully, this work will encourage plankton ecologists of all kinds to think more about plankton population behaviour on different spatial scales of averaging, and to discuss with balanced, non-evangelical statisticians how best to sample and model these populations. Interdisciplinary scientists should be inspired to work on this problem, of how the behaviour of microscopic life in a turbulent fluid environment regulates the behaviour of a planet’s climate, and of ecosystem behaviour on all spatial scales in between.

Abraham, E.R., 1998. The generation of plankton patchiness by turbulent stirring. Nature. 391, 577-580.

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N. and Csaki, F.(Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest. pp. 267–281.

Anderson, T.R., 2005. Plankton functional type modelling: running before we can walk? J. Plank. Res. 27, 1073-1081.

Anderson, T.R., 2006. Confronting complexity: reply to Le Quéré and Flynn. J. Plank. Res. 28(9), 877-878.

Anderson, L.A., Robinson, A.R., Lozano, C.J., 2000. Physical and biological modeling in the Gulf Stream region: I. Data assimilation methodology. Deep-Sea Res. I. 47, 1787–1827.

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modelling. Mar. Ecol. Prog. Ser. 271, 13-26.

Armi, L.A., Flament, P., 1985. Cautionary remarks on the spectral interpretation of turbulent flows. J. Geophys. Res. 90, 11779–11782.

Azam, F., Fenchel, T., Field, J.G., Gray, J.S., Meyer-Reil, L.A., Thingstad, F., 1983. The ecological role of water-column microbes in the sea. Mar. Ecol. Prog. Ser. 10, 257–263.

Baird, M.E., Emsley S.M., 1999. Towards a mechanistic model of plankton population dynamics. J. Plankton Res. 21, 85–126.

Batchelder, H.P., Edwards, C.A., Powell, T.M., 2002. Individual-based models of copepod populations in coastal upwelling regions: implications of physiologically and environmentally influenced diel vertical migration on demographic success and nearshore retention. Prog. Oceanogr., 53, 307–333.

Bees, M.A., 1998. Planktonic Communities and Chaotic Advection in Dynamical Models of Langmuir Circulation. App. Sci. Res. 59, 141–158.

Bennett, A.F., 2002. Inverse modelling of the ocean and atmosphere, Cambridge University Press, Cambridge.

Birch, D.A., Young, W.R., 2006. A master equation for a spatial population model with pair interactions. Theor. Pop. Biol. 70(1), 26–42.

Bolker, B.M., Pacala, S.W., 1997. Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. Theor. Pop. Biol. 52, 179–197., doi:10.1006/tpbi.1997.1331

Brentnall, S.J., Richards, K.J., Brindley, J., Murphy, E., 2003. Plankton patchiness and its effect on larger-scale productivity. J. Plankton Res. 25, 121–140.

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

Campbell, J. W., 1995. The lognormal distribution as a model for bio-optical variability in the sea. Journal of Geophysical Research. 100(C7), 13237–13254.

Carniel, S., Vichi, M., Sclavo, M., 2007. Sensitivity of a coupled physical-biological model to turbulence: high-frequency simulations in a northern Adriatic station. Chemistry and Ecology. 23(2), 157–175.

Charria, G., Dadou, I., Cipollini, P., Drévillon, M., De Mey, P., Garçon, V., 2006. Understanding the influence of Rossby waves on surface chlorophyll concentrations in the North

Atlantic Ocean. J. Mar. Res. 64, 43–71.

Cipollini, P., Cromwell, D., Challenor, P.G., Raffaglio, S., 2001. Rossby waves detected in global ocean colour data. Geophys. Res. Lett. 24, 889–892.

Cowles T.J., Desiderio, R.A., Carr, M., 1998. Small-Scale Planktonic Structure: Persistence and Trophic Consequences. Oceanography. 11(1), 4–9.

Dadou, I., Evans, G.E., Garçon, V., 2004. Using JGOFS *in situ* and ocean color data to compare biogeochemical models and estimate their parameters in the subtropical North Atlantic Ocean. J. Mar. Res. 62, 565–594.

Dandonneau, Y., Vega, A., Loisel, H., du Penhoat, Y., Menkes, C., 2003. Oceanic Rossby Waves Acting As a "Hay Rake" for Ecosystem Floating By-Products. Science. 302, 1548–1551.

Davis C.S., Gallager, S.M., Solow, A.R., 1992. Microaggregations of Oceanic Plankton Observed by Towed Video Microscopy. Science. 257, 230–232.

Denman, K., Okubo, A., Platt, T., 1977. The chlorophyll fluctuation spectrum in the sea. Limn. and Oceanog. 22(6), 1033–1038.

Denman, K.L., 2003. Modelling planktonic ecosystems: parameterizing complexity. Prog. Oceanog. 57, 429–452.

Denman, K.L., Brasseur, G., Chidthaisong, P., Ciais, P., Cox, P.M., Dickinson, R.E., Hauglustaine, D., Heinze, C., Holland, E., Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P.L., Wofsy, S.C., Zhang, X., 2007. Couplings Between Changes in the Climate System and Biogeochemistry. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L.(Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Derenbach, J.B., Astheimer, H., Hansen, H.P., Leach, H., 1979. Vertical Microscale Distribution of Phytoplankton in Relation to the Thermocline. Mar. Ecol. Prog. Ser. 1, 187–193.

Dieckmann, U., Law, R., Metz, J.A.J, 2000. The Geometry of Ecological Interactions: Simplifying Spatial Complexity. Cambridge University Press, Cambridge. 564 pp.

Di Lorenzo, E., Moore, A.M., Arango, H.G., Cornuelle, B.D., Miller, A.J., Powell, B., Chua, B.S., Bennet, A.F., 2007. Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): Development and application for a baroclinic coastal upwelling system. Ocean. Mod. 16, 160–187.

Dombrowsky, E., De Mey, P., 1992. Continuous Assimilation in an Open Domain of the Northeast Atlantic 1. Methodology and Application to AthenA-88. J. Geophys. Res. 97(C6), 9719–9731.

Doubell, M.J., Seuront, L., Seymour, J.R., Patten, N.L., Mitchell, J.G., 2006. High-Resolution Flourometer for Mapping Microscale Phytoplankton Distributions. App. and Env. Microbiol. 72(6), 4475–4478. doi:10.1128/AEM.02959-05.

Edwards, A. M. and J. Brindley. 1996. Oscillatory behavior in a three-component plankton population model. Dynamics and Stability of Systems, *11*, No. 4, 347–370.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Evans, G.T., 1999. The role of local models and data sets in the Joint Global Ocean Flux Study. Deep-Sea Res. I. 46, 1369–1389., doi:10.1016/S0967-0637(99)00010-2

Evans, G.T., 2003. Defining misfit between biogeochemical models and data sets. J. Mar. Sys. 40–41, 49–54.

Evans, G.T. and Parslow, J.S., 1985. A Model of Annual Plankton Cycles. Biol. Oceanog. (3) 3, 327–347.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367.

Falkowski, P.G., Ziemann, D., Kolber, Z., Bienfang, P.K., 1991. Role of eddy pumping in enhancing primary production in the ocean. Nature. 352, 55–58.

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

Fasham, M.J.R., Sarmiento, J.L., Slater, R.D., Ducklow, H.W., Williams, R., 1993. Ecosystem Behaviour at Bermuda station ”S” and Ocean Weather Station ”India”: a General Circulation Model and Observational Analysis. Global Biogeochem. Cycles. 7(2), 379–415.

Fasham, M.J.R., Evans, G.T., 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., Losch, M., Schroter, J., Wenzel, M., 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. J. Mar. Sys. 28, 45–63., doi:10.1016/S0924-7963(00)00083-X

Ferguson, N.M., Donnelly, C.A., Anderson, R.M., 2001. The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions. Science, 292, 1155–1160.

Fernández, E., Pingree, R.D., 1996. Coupling between physical and biological fields in the North Atlantic subtropical front southeast of the Azores. Deep Sea Res. I. 43(9), 1369–1393.

Filipe, J.A.N., Gibson, G.J., 2001. Comparing Approximations to Spatio-temporal Models for Epidemics and Local Spread. Bull. Math. Biol., 63, 603–624, doi:10.1006/bulm.2001.0234.

Filipe, J.A.N., Maule, M.M., 2003. Analytical methods for predicting the behaviour of population models with general spatial interactions. Math. Biosci., 183, 15–35, doi:10.1016/S0025-5564(02)00224-9.

Flierl, G., McGillicuddy, D.J., 2002. Mesoscale and Submesoscale Physical-Biological Interactions. Chapter 4 in *The Sea*, Volume 12, edited by Robinson, A.R., McCarthy, J.J., Rothschild, B.J., John Wiley & Sons, Inc., New York.

Flierl, G., Grünbaum, D., Levin, S., Olson, D., 1999. From Individuals to Aggregations: the Interplay between Behaviour and Physics. J. Theor. Biol. 196, 397–454, doi:10.1016/j.jtbi.1998.0842

Flynn, K.J., 2001. A mechanistic model for describing dynamic multi-nutrient, light, temperature interactions in phytoplankton. J. Plank. Res. 23, 977–997.

Flynn, K.J., 2005. HORIZONS. Castles built on sand:dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modellers. J. Plank. Res. 27(12), 1205-1210.

Follows, M.J., Dutkiewicz, S., Grant, S., Chisholm, S.W., 2007. Emergent Biogeography of Microbial Communities in a Model Ocean. Science, 315, 1843–1846.

Folt, C.L., Burns, C.W., 1999. Biological drivers of zooplankton patchiness. TREE.

14(8), 300–305.

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.

Franks, P.J.S., 1992a. Sink or swim: accumulation of biomass at fronts. Mar. Ecol. Prog. Ser. 82, 1–12.

Franks, P.J.S., 1992b. Phytoplankton Blooms at Fronts: Patterns, Scales, and Physical Forcing Mechanisms. Rev. Aquatic. Sci. 6(2), 121–137.

Franks, P.J.S., 1995. Thin layers of phytoplankton: a model of formation by near-inertial wave shear. Deep-Sea Res. I. 42, 75–91.

Franks, P.J.S. and Chen, C., 1996. Plankton production in tidal fronts: A model of Georges Bank in the summer. J. Mar. Res. 54, 631–651.

Franks, P.J.S., Walstad, L.J., 1997. Phytoplankton patches at fronts: A model of formation and response to wind events. J. Mar. Res. 55, 1–29.

Franks, P.J.S. and Chen, C., 2001. A 3-D prognostic numerical model study of the Georges Bank ecosystem. Part II: biological-physical model. Deep-Sea Res. II. 48, 457–482.

Franks, P.J.S., Jaffe, J.S., 2001. Microscale distributions of phytoplankton: initial results from a two-dimensional imaging flourometer, OSST. Mar. Ecol. Prog. Ser. 220, 59–72.

Franks, P.J.S., 2005. Plankton patchiness, turbulent transport and spatial spectra. Mar. Ecol. Prog. Ser. 294, 295–309.

Friedrichs, M.A.M., 2001. A data assimilative marine ecosystem model of the central equatorial Pacific: Numerical twin experiments. J. Mar. Res. 59, 859–894.

Friedrichs, M.A.M., Hood, R.R., Wiggert J.D., 2006. Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data. Deep-Sea Res. II. 53, 576–600., doi:10.1016/j.dsr2.2006.01.026

Friedrichs, M.A.M., Dusenberry, J.A., Anderson, L.A., Armstrong, R., Chai, F., Christian, J.R., Doney, S.C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D., Moore, J.K., Schartau, M., Spitz, Y.H., Wiggert., J.D., 2007. Assessment of skill and portability in regional marine biogeochemical models: the role of multiple planktonic groups. J. Geophys. Res., in press.

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.

Fussman, G.F., Ellner, S.P., Shertzer, K.W., Hairston, N.G. Jr., 2000. Crossing the Hopf Bifurcation in a Live Predator-Prey System. Science, 290, 1358–1360.

Garcia-Gorriz, E., Hoepffner, N., Ouberdous, M., 2003. Assimilation of SeaWiFS data in a coupled physical-biological mode of the Adriatic Sea. J. Mar. Sys. 40–41, 233–252.

Garçon, V.C., Oschlies, A., Doney, S., McGillicuddy, D., Waniek, J., 2001. The role of mesoscale variability on plankton dynamics in the North Atlantic. Deep-Sea Res. II. 48, 2199–2226.

Gent, P.R., McWilliams, J.C., 1990. Isopycnal mixing in ocean circulation models. J. Phys. Oceanogr. 20, 150–155.

Gent, P.R., Willebrand, J., McDougall, T.J., McWilliams, J.C., 1995. Parameterizing Eddy-Induced Tracer Transports in Ocean Circulation Models. J. Phys. Oceanogr. 25, 463–474.

Gregg, W.W., 2008. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. J. Mar. Sys. 69, 205–225.

Guirey, E.J., Bees, M.A., Martin, A.P., Srokosz, M.A., Fasham, M.J.R., 2007. Emergent features due to grid-cell biology: synchronisation in biophysical models. Bull. Math. Biol. 69(4), 1401–1422.

Harmon, R.T., Challenor, P.G., 1997. A Markov Chain Monte Carlo method for estimation and assimilation into models. Ecol. Mod. 101, 41–49., doi:10.1103/PhysRevE.70.016216

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer-Verlag, New York.

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

Hernández-García, E., López, C., 2004. Clustering, advection, and patterns in a model of population dynamics with neighborhood-dependent rates. Phys. Rev. E., 70, 016216, doi:10.1016/S0924-7963(00)00083-X

Hillary, R.M., Bees, M.A., 2004. Synchrony & Chaos in Patchy Ecosystems. Bull. Math. Biol. 66, 1909–1931.

Hoffman, R.N., Liu, Z., Louis, J-F., Grassotti, C., 1995. Distortion Representation of Forecast Errors. Mon. Weather Rev. 123, 2758–2770.

Hsieh, C., Glaser, S.M., Lucas, A.J., Sugihara, G., 2005. Distinguishing random environmental fluctuations from ecological catastrophes for the North Pacific Ocea. Nature. 435, 336–340.

Huret, M., Gohin, F., Delmas, D., Lunven, M., Garçon, V., 2007. Use of SeaWiFS data for light availability and parameter estimation of a phytoplankton production model of the Bay of Biscay. J. Mar. Sys. 65, 509–531.

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

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

Hurvich, C.M., Tsai, C-L., 1989. Regression and time series model selection in small samples. Biometrika, 76, 297–307.

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

Ji, R., Chen, C., Franks, P.J.S., Townsend, D.W., Durbin, E.G., Beardsley, R.C., Lough, R.G., Houghton, R.W., 2006. Spring bloom and associated lower trophic level food web dynamics on Georges Bank: from observation to model. Submitted to Deep-Sea Res. II. 53, 26562683.

Ji, R., Davis, C., Chen, C., Beardsley, R., 2007. Influence of local and external processes on the annual nitrogen cycle and primary productivity on Georges Bank: A 3-D biological-physical modeling study. J. Mar. Sys. submitted.

Keeling, M.J., 2000a. Metapopulation moments: coupling, stochasticity and persistence. J. Animal Ecol. 69, 725–736.

Keeling, M.J., 2000b. Multiplicative Moments and Measures of Persistence in Ecology. J. Theor. Biol. 205, 269–281.

Keeling, M.J., Wilson, H.B., Pacala, S.W., 2002. Deterministic Limits to Stochastic Spatial Models of Natural Enemies. Am. Nat. 159, 57–80.

Kierstead, H., Slobodkin, L.B., 1953. The size of water masses containing plankton blooms. J. Mar. Res. 12, 141–147.

Killworth, P.D., Cipollini, P., Mete Uz, B., Blundell, J.R., 2004. Physical and biological mechanisms for planetary waves observed in satellite-derived chlorophyll. J. Geophys. Res. 109, C07002, doi:10.1029/2003JC001768.

Kolmogorov, A.N., 1941. The local structure of turbulence in incompressible fluid for very large Reynolds numbers. Dokl. Akad. Nauk SSSR. 30, 301–305. (In Russian). (Trans. Proc. Roy. Soc. Lond. A. 434, 9–13, (1991)).

Kuroda, H., Kishi, M.J., 2004. A data assimilation technique applied to estimate parameters for the NEMURO marine ecosystem model. Ecological Modelling. 172, 69–85.

Law, R., Dieckmann, U., 2000a. Moment Approximations of Individual-based Models. In: The Geometry of Ecological Interactions: Simplifying Spatial Complexity. Cambridge University Press, Cambridge.

Law, R., Dieckmann, U., 2000b. A Dynamical System for Neighborhoods in Plant Communities. Ecol., 81(8), 2137–2148.

Law, R., Murrell, D.J., Dieckmann, U., 2003. Population Growth in Space and Time: Spatial Logistic Equations. Ecol., 84(8), 252–262.

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

Leonard, T., Hsu, J.S.J., 1999. Bayesian Methods. Cambridge University Press, 333pp.

Levasseur, A., Shi, L., Wells, N.C., Purdie, D.A., Kelly-Gerreyn, B.A., 2007. A three-dimensional hydrodynamic model of estuarine circulation with an application to Southampton Water, UK. Est. Coast. Shelf Sci. 73, 753–767.

Lévy, M., Klein, P., Treguier, A., 2001. Impact of sub-mesoscale physics on production and subduction of phytoplankton in an oligotrophic regime. J. Mar. Res. 59, 535–565.

Lévy, M., 2006. Modulation de la production biologique par la turbulence océanique de mésoéchelle, Habilitation à diriger des Recherches, Univ. Paris 6, soutenue le 6 juin 2006.

Lewis, M.A., Pacala, S., 2000. Modeling and analysis of stochastic invasion processes. J. Math. Biol. 41, 387–429, doi:10.1007/s002850000050.

Li, X., McGillicuddy D.J., Jr., Durbin, E.G., Wiebe, P.H. Biological control of the vernal population increase of Calanus Finmarchicus on Georges Bank. Deep-Sea Res. II. 53, 2632–2655., doi:10.1016/j.dsr2.2006.08.011.

Lochte, K., Pfannkuche, O., 1987. Cyclonic cold-core eddy in the eastern North Atlantic II. Nutrients, phytoplankton and bacterioplankton. Mar. Ecol. Progr. Ser. 39, 153–164.

Losa, S.N., Kivman, G.A., Schröter, J., Wenzel M., 2003. Sequential weak constraint parameter estimation in an ecosystem model. J. Mar. Sys. 43, 31–49., doi:10.1016/j.jmarsys.2003.06.001.

Losa, S.N., Kivman, G.A., Ryabchencko, V.A., 2004. Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data? J. Mar. Sys. 45, 1–20., doi:10.1016/j.jmarsys.2003.08.005.

Lynch, D.R., McGillicuddy D.J., Jr., Werner, F.E., 2007. Skill Vocabulary - a Starting Point. J. Mar. Sys., this volume.

Mahadevan, A., Archer, D., 2000. Modeling the impact of fronts and mesoscale circulation on the nutrient supply and biogeochemistry of the upper ocean. J. Geophys. Res. 105,(C1), 1209–1225.

Martin, A.P, Richards, K.J., 2001. Mechanisms for vertical nutrient transport within a North Atlantic mesoscale eddy. Deep-Sea Res. II. 48, 757–773.

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

Martin, A.P., 2004. A Malthusian curb on spatial structure in microorganism populations. J. Theor. Biol. 230, 343–349., doi:10.1016/j.jtbi.2004.05.017.

Martin, A.P, Zubkov, M.V., Burkhill, P.H., Holland, R.J., 2005. Extreme spatial variability in marine picoplankton and its consequences for interpreting Eulerian time-series. Biol. Lett. 1, 366–369.

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.

Matthews, L., Brindley, J., 1997. Patchiness in plankton populations. Dyn. Stab. Sys. 12, 39–59.

McCauley, E., Nisbet, R.M., Murdoch, W.W., de Roos, A.M., Gurney, W.S.C., 1999. Large-amplitude cycles of *Daphnia* and its algal prey in enriched environments. Nature, 402, 653–656.

McGillicuddy Jr., D.J., Robinson, A.R., 1997. Eddy-induced nutrient supply and new production in the Sargasso Sea. Deep-Sea Res. I. 44(8), 1427–1450.

McGillicuddy Jr., D.J., Lynch, D.R., Moore, A.M., Gentleman, W.C., Davis, C.S., Meise, C.J., 1998. An adjoint data assimilation approach to diagnosis of physical and biological controls on *Pseudocalanus* spp. in the Gulf of Maine–Georges Bank region. Fish. Oceanogr. 7:3/4, 205–218.

McGillicuddy Jr., D.J., Lynch, D.R., Wiebe, P., Runge, J., Durbin, E.G., Gentleman W.C., Davis, C.S., 2001. Evaluating the synopticity of the US GLOBEC Georges Bank broad-scale sampling pattern with observational system simulation experiments. Deep-Sea Res. II, 48, 483499.

McGillicuddy Jr., D.J., Anderson, L.A., Doney, S.C., Maltrud, M.E., 2003. Eddy-driven sources and sinks of nutrients in the upper ocean: Results from a 0.1^o resolution model of the North Atlantic. Glob. Biogeo. Cycles. 17(2), 1035, doi:10.1029/2002GB001987.

McGillicuddy Jr., D.J., Anderson, L.A., Bates, N.R., Bibby T., Buesseler, K.O., Carlson, C.A., Davis, C.S., Ewart, C., Falkowski, P.G., Gooldthwaite, S.A., Hansell, D.A., Jenkins, W.J., Johnson, R., Kosnyrev, V.K., Ledwell, J.R., Li, Q.P., Siegel, D.A., Steinberg, D.K., 2007. Eddy/Wind Interactions Stimulate Extraordinary Mid-Ocean Plankton Blooms. Science. 316, 1021–1026.

McLeod, A.I., Quenneville, B., 2001. Mean likelihood estimators. Stat. and Comp. 11, 57–65.

Miller, A.J., Gabric, A.J., Moisan, J.R., Chai, F., Neilson, D.J., Pierce, D.W., Di Lorenzo, E., 2006. Global change and oceanic primary productivity: Effects of ocean-atmosphere-biological feedbacks. In: *Global Climate Change and Response of the Carbon Cycle in the Equatorial Pacific and Indian Oceans and Adjacent Land Masses* H. Kawahata and Y. Awaya, Eds., Elsevier Oceanography Series, 73, pp. 29–65.

Mitchell, J.G., Fuhrman, J.A., 1989. Centimeter scale vertical heterogeneity in bacteria and chlorophyll *a*. Mar. Ecol. Prog. Ser. 54, 141–148.

Murrell, D.J., Dieckmann U., Law, R., 2004. On moment closures for population dynamics in continuous space. J. Theor. Biol. 229, 421–432., doi:10.1016/j.jtbi.2004.04.013

Nerger, L., Gregg, W.W., 2008. Assimilation of SeaWiFS data into a global ocean-biogeochemical model using a local SEIK filter. J. Mar. Sys. 68, 237–254.

Neufeld, Z., Haynes, P., Garçn, V., Sudre, Joel, 2002. Ocean fertilization may initiate a large scale phytoplankton bloom. Geophys. Res. Lett. 29(11), 1534, doi: 10.1029/2001GL013677, 2002.

Neyman, J. and E. L. Scott. 1948. Consistent estimates based on partially consistent observations. Econometrica, *16*, No. 1. 1–32.

O'Hagan, A., Forster, J., 2004. Kendall's advanced theory of statistics (2nd edn), Vol. 2B: Bayesian Inference, Oxford University Press, New York, 480 pp.

Okubo, A., 1971. Oceanic diffusion diagrams. Deep-Sea Res. 18, 789–802.

Okubo, A., 1976. Remarks on the use of "diffusion diagrams" in modeling scale-dependent diffusion. Deep-Sea Res. 23, 1213–1214.

Okubo, A., 1978. Horizontal dispersion and critical scales for phytoplankton. In J.H., Steele (Ed.), *Spatial pattern in plankton communities* (pp. 21–42). Plenum.

Okubo, A., Levin, S., 1980. Diffusion and Ecological Problems: Modern Perspectives, second ed., Interdisciplinary Applied Mathematics, vol. 17, Springer, New York. 467 pp.

Oschlies, A., Garçon, V., 1998. Eddy-induced enhancement of primary production in a model of the North Atlantic Ocean. Nature. 394, 266–269.

Oschlies, A., Koeve, W., Garçon, V., 2000. An eddy-permitting coupled physical-biological model of the North Atlantic 2. Ecosystem dynamics and comparison with satellite and JGOFS local studies data. Glob. Biogeochem. Cycles. 14(1), 499–523.

Oschlies, A., 2002. Can eddies make ocean deserts bloom? Glob. Biogeochem. Cycles. 16(4), 1106, doi:10.1029/2001GB001830.

Oschlies, A., 2004. Feedbacks of biotically induced radiative heating on upper ocean heat budget, circulation, and biological production in a coupled ecosystem-circulation model. J. Geophys. Res. 109, C12031, doi:10.1029/2004JC002430.

Oschlies, A., Schartau, M., 2005. Basin-scale performance of a locally optimized marine ecosystem model. J. Mar. Res. 63, 335–358.

Ovaskeinen, O., Cornell, S., 2006. Space and stochasticity in population dynamics. PNAS, 103(34), 12781–12786.

Palmer, J.R., Totterdell, I.J., 2001. Production and export in a global ocean ecosystem model. Deep-Sea Res. I. 48, 1169–1198.

Pascual, M., Levin, S.A., 1999. Spatial Scaling in a Benthic Population Model with Density-Dependent Disturbance. Theor. Pop. Biol. 56, 106–122.

Pasquero, C., 2005. Differential eddy diffusion of biogeochemical tracers. Geophys. Res. Lett. 32(11), L171603, doi: 10.1029/2005GL023662, 2005.

Petrovskii, S.V., 1999a. On the diffusion of a plankton patch in a turbulent ocean. Oceanology. 39, 737–742.

Petrovskii, S.V., 1999b. On the plankton front waves accelerated by marine turbulence. J. Mar. Sys. 21, 179–188.

Platt, T., 1972. Local phytoplankton abundance and turbulence. Deep-Sea Res. 19, 183–187.

Platt T., Sathyendranath, S., Ravindran, P., 1990. Primary production by phytoplankton: analytic solutions for daily rates per unit area of water surface. Proc. R. Soc. Lond. 241, 101–111.

Popova, E.E., Srokosz, M.A., Smeed, D.A., 2002a. Real-time forecasting of biological and physical dynamics at the Iceland-Faeroes Front in June 2001. Geophys. Res. Lett. 29(4), 1055, 10.1029/2001GL013706, 2002.

Popova, E.E., Lozano, C.J., Srokosz, M.A., Fasham, M.J.R., Haley, P.J., Robinson, A.R., 2002b. Coupled 3D physical and biological modelling of the mesoscale variability in North-East Atlantic in spring 1997: biological processes. Deep-Sea Res. I. 49, 1741–1768.

Popova, E.E., Coward, A.C., Nurser, G.A., de Cuevas, B., Fasham, M.J.R., Anderson, T.R., 2006. Mechanisms controlling primary and new production in a global ecosystem model — Part I: Validation of the biological simulation. Ocean Sci. 2, 249–266.

Powell, T.M., Okubo, A., 1994. Turbulence, diffusion and patchiness in the sea. Phil. Trans. R. Soc. Lond. B. 343, 11–18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P, 1999. Numerical Recipes in Fortran 77, Second Edition, The Art of Scientific Computing. Cambridge University Press, Cambridge.

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 Biogeochemical Cycles, *10*, No. 1, 111–138.

Prunet, P., Minster, J-F, Echevin, V., 1996. Assimilation of surface data in a one-dimensional physical biogeochemical model of the surface ocean 2. Adjusting a simple trophic model to chlorophyll, temperature, nitrate, and pCO_2 data. Glob. Biogeochem. Cycles. 10(1), 139–158.

Renshaw, E., 1991. Modelling biological populations in space and time. Cambridge University Press, Cambridge. 403 pp.

Reynolds, O., 1895. On the dynamical theory of incompressible viscous fluids and the determination of the criterion. Philos. Trans. R. Soc. London Ser. A. 186, 123–164.

Richards, K.J., Brentnall, S.J., 2006. The impact of diffusion and stirring on the dynamics of interacting populations. J. Theor. Biol. 238, 340–347., doi:10.1016/j.jtbi.2005.05.029

Robinson, A.R., 1996. Physical processes, field estimation and an approach to interdisciplinary ocean modeling. Earth-Science Rev. 40, 3–54.

Rodriguez, R., Tuckwell, H.C., 1996. Statistical properties of stochastic nonlinear dynamical models of single spiking neurons and neural networks. Phys. Rev. E. 54(5), 5585–5590.

Rothstein, L.W., Cullen, J.J., Abbott, M., Chassignet, E.P., Denman, K., Doney, S.C., Ducklow, H., Fennel, K., Follows, M., Haidvogel, D., Hoffman, E., Karl., D.M., Kindle, J., Lima, I., Maltrud, M., McClain, C., McGillicuddy, D., Olascoaga, M.J., Spitz, Y., Wiggert, J., Yoder, J., 2006. Modeling Ocean Ecosystems. Oceanography. 19(1), 22–51.

Ryabchenko, V.A., Fasham, M.J.R., Kagan, B.A., Popova, E.E., 1997. What causes short-term oscillations in ecosystem models of the ocean mixed layer? J. Mar. Sys. 13, 33–50., doi:10.1016/S0924-7963(96)00110-8

Sander, J., 1998. Dynamics equations and turbulent closures in geophysics. Continuum Mech. Thermodyn. 10, 1–28.

Sarmiento, J.L., Slater, R.D., Fasham, M.J.R., Ducklow, H.W., Toggweiler, J.R., Evans, G.T., 1993. A seasonal three-dimensional ecosystem model of nitrogen cycling in the north Atlantic euphotic zone. Glob. Biogeochem. Cycles. 7(2), 417–450.

Sarmiento, J.L., Slater, R.D., Barber, R., Bopp, L., Doney, S.C., Hirst, A.C., Kleypas, J., Matear, R., Mikolajewicz, U., Monfray, P., Soldatov, V., Spall., S.A., Stouffer, R., 2004. Response of ocean ecosystems to climate warming. Glob. Biogeochem. Cycles. 18, GB3003, doi:10.1029/2003GB002134, 2004.

Schartau, M., 2001. Data-assimilation studies of marine, nitrogen based, ecosystem models in the North Atlantic Ocean. Doctoral thesis, Kiel, 2001.

Schartau, M., Oschlies, A., Willebrand, J., 2001. Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method. Deep-Sea Res. II. 48, 1769–1800.

Schartau, M. and Oschlies, A., 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 (6) 761–793.

Scheffer, M., Baveco, J.M., DeAngelis, D.L., Rose, K.A., van Ness, E.H., 1995. Super-individuals a simple solutioin for modelling large populations on an individual basis. Ecol. Modelling. 80, 161–170.

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

Sen, M., Stoffa, P.L., 1995. Global optimisation methods in geophysical inversion, Elsevier.

Six, K.D., Maier-Reimer, E., 1996. Effects of plankton dynamics on seasonal carbon fluxes in a ocean general circulation model. Glob. Biogeochem. Cycles. 10(4), 559–583.

Smith, C.L., Richards, K.J., Fasham, M.J.R., 1996. The impact of mesoscale eddies on plankton dynamics in the upper ocean. Deep-Sea Res. I,. 43(11–12), 1807–1832.

Smith, K.W., McGillicuddy, D.J., Lynch, D.R. Parameter estimation using an ensemble smoother: the effect of circulation in biological estimation. J. Mar. Sys., submitted.

Spall, S.A., Richards, K.J., 2000. A numerical model of mesoscale frontal instabilities and plankton dynamics — I. Model formulation and initial experiments. Deep-Sea Res. I. 47, 1261–1301.

Spitz, Y.H., Moisan, J.R., Abbott, M.R., Richman, J.G., 1998. Data assimilation and a pelagic ecosystem model: parameterization using time series observations. J. Mar. Sys. 16, 51–68., doi:10.1016/S0924-7963(97)00099-7

Spitz, Y.H., Moisan, J.R., Abbott, M.R., 2001. Configuring an ecosystem model using data from the Bermuda Atlantic Time Series (BATS). Deep-Sea Res. II. 48, 1733–1768.

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., Henderson, E.W., 1981. A simple plankton model. American Naturalist 117, 676–691.

Steele, J.H., 1998. Incorporating the microbial loop in a simple plankton model. Proc. R. Soc. Lond. B. 265, 1771–1777.

Steele, J.H., Collie, J.S., Bisagni, J.J., Gifford, D.J., Fogarty, M.J., Link, J.S., Sullivan, B.K., Sieracki, M.E., Beet, A.R., Mountain, D.G., Durbin, E.G., Palka, D., Stockhausen, W.T., 2007. Balancing end-to-end budgets of the Georges Bank ecosystem. Prog. in Oceanog. 74 (4) 423448. doi:10.1016/j.pocean.2007.05.003

Stock, C.A., McGillicuddy, D.J., Solow, A.R., Anderson, D.M., 2005. Evaluating hypotheses for the initiation and development of *Alexandrium fundyense* blooms in the western Gulf of Maine using a coupled physical-biological model. Deep-Sea Res. II. 52, 2715–2744.

Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. J. Roy. Stat. Soc. B. 36(2), 111–147.

Stuart, A., K. Ord and S. Arnold. 1999. Kendall's advanced theory of statistics (6th edn), Vol. 2A: Classical Inference and the Linear Model, Oxford University Press, New York, 885 pp.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106(D7), 7183–7192.

Tennekes, H., Lumley, J.L., 1972. A First Course in Turbulence. The MIT Press, Cambridge.

Thacker, W.C., 1989. The Role of the Hessian Matrix in Fitting Models to Measurements. J. Geophys. Res. 94(C5), 6177–6196.

Thiébaux, H.J., Pedder, M.A., 1987. Spatial Objective Analysis: with applications in atmospheric science. Academic Press Inc., London. 299 pp.

Tjiputra, J.F., Polzin, D., Winguth, A.M.E., 2007. Assimilation of seasonal chlorophyll and nutrient data into an adjoint three-dimensional ocean carbon cycle model: Sensitivity analysis and ecosystem parameter optimization. Glob. Biogeochem. Cycles. 21, GB1001,doi:10.1029/2006GB002745, 2007.

Townsend, D.W. and Thomas, M., 2002. Springtime nutrient and phytoplankton dynamics on Georges Bank. Mar. Ecol. Prog. Ser. 228, 57–74.

Truscott, J.E., Brindley, J., 1994. Ocean plankton populations as excitable media. Bull. Math. Biol. 56, 981–998.

Umlauf, L., Burchard, H., 2005. Second-order turbulence closure models for geophysical boundary layers. A review of recent work. Cont. Shelf Res. 25, 795–827., doi:10.1016/j.csr.2004.08.004

Uz, B.M., Yoder, J.A., Osychny, V., 2001. Pumping of nutrients to ocean surface waters by the action of propagating planetary waves. Nature, 409, 597–600.

Vainstein, M.H., Rubí, J.M., Vilar, J.M.G., 2007. Stochastic population dynamics in turbulent fields. Eur. Phys. J. Special Topics 146, 177-187, doi: 10.1140/epjst/e2007-00178-7.

Vallino, J.J., 2000. Improving marine ecosystem models: Use of data assimilation and mesocosm experiments. J. Mar. Res. 58, 117–164.

Woods, J.D., 1988. Mesoscale upwelling and primary production. In: Rothschild, B.J. (Ed.), *Towards a Theory on Biological-Physical Interactions in the World Ocean*, NATO ASI Series. Kluwer Academic Press, Dordrecht, pp. 7–38.

Woods, J.D., 2005. The Lagrangian Ensemble metamodel for simulating plankton ecosystems. Prog. Oceanogr. 67, 84–159.

Yamazaki, H., Mackas, D.L., Denman, K., 2002. Coupling small-scale physical processes with biology. Chapter 3 in *The Sea*, Volume 12, edited by Robinson A.R., McCarthy, J.J., Rothschild, B.J., New York.

Young G. A., 1994. Bootstrap — more than a stab in the dark?. Stat. Sci. 9, 382–395.