

**Data assimilation into oceanic ecosystem models:
a critical review, some comparisons,
and recommendations**

Meric Srokosz

Room 254/43, Southampton Oceanography Centre (SOC)

Empress Dock, Southampton, SO14 3ZH, UK

Tel: +44-(0)1703-596414 (direct line); Fax: +44-(0)1703-596400

e-mail: M.Srokosz@soc.soton.ac.uk

(blank page)

Abstract

This report aims to critically review the work carried out to-date on the assimilation of data into ocean ecosystem and coupled bio-physical models. The review is prompted by the newly available ocean colour data from SeaWiFS and OCTS, and their potential use for data assimilation to improve both the models and their predictive capabilities. In addition, some initial results on the comparison of three data assimilation methods (adjoint, simulated annealing, Markov chain Monte Carlo) are summarised. The comparison has been carried out using a twin experiment approach with a simple predator-prey model, in order to focus on the assimilation techniques rather than the complexities of the model. From the review and comparison conclusions are drawn regarding the state of data assimilation into oceanic ecosystem models, and recommendations made about possible future work on this problem.

(blank page)

Contents

| | <i>page</i> |
|--|-------------|
| 1. Introduction | 7 |
| 2. Simple ecosystem models | 10 |
| 3. Coupled bio-physical models | 15 |
| 4. Some comparisons | 19 |
| 5. Discussion, conclusions and recommendations | 24 |
| References | 27 |
| Appendix: a note on the adjoint method | 31 |
| Abbreviations and acronyms | 32 |

(blank page)

1. Introduction

For many and various purposes, perhaps most importantly understanding the role of the oceans in the carbon cycle and the influence of anthropogenic CO₂ emissions into the atmosphere on that cycle (Siegenthaler & Sarmiento, 1993), considerable interest exists at the present time in developing robust oceanic ecosystem models (Fasham, 1993). The aim being, by coupling these biological models to physical models of the ocean circulation, to provide a description, and possibly a prediction, of the contribution of the ocean to global biogeochemical cycles (such as the carbon cycle; Abbott, 1992; Sarmiento & Armstrong, 1997), and to understand the behaviour of the oceanic food chain (including fisheries; Hofmann et al., 1991).

In attempting to carry this out two problems become apparent. First, in contrast to the modelling of the atmosphere or ocean, where a basic description of the physics is provided by the Navier-Stokes equations of fluid dynamics (Gill, 1982), there is no basic set of equations that describe the ocean ecosystem. Therefore a heuristic approach is generally taken, where sufficient biological components are modelled to describe the problem of interest (Fasham, 1993). For example, a simple three component (also called a three compartment) oceanic ecosystem model might include phytoplankton, zooplankton and nutrients and has been used with a degree of success for some studies (Evans & Parslow, 1985). The second problem is related to the first, in that such heuristic models have many parameters that are known to varying degrees of accuracy (for example, the 7 compartment model of Fasham et al., 1990, has 27 parameters). This limits the potential applicability of the models, unless the parameters can be better determined. This provides an ideal application for data assimilation in its guise of fitting models to data (Thacker & Long, 1988), and this will be the main application of data assimilation discussed in this report.

Another purpose of data assimilation is that of improving the predictive power of models. This is a familiar application in the context of meteorological forecasting. It is

also beginning to be used more widely in the area of physical ocean modelling, but is in its infancy with regard to assimilation of data into biological models. Many techniques have been developed for the assimilation of data into models of atmosphere and ocean physics (see Ghil & Malanotte-Rizzoli, 1991; Malanotte-Rizzoli, 1996). The applicability of these techniques to oceanic ecosystem models is by no mean automatic, and this report will show that much work remains to be done in this area.

Thus far we have seen the utility of data assimilation for improving both the descriptive and predictive powers of models. However, for this to be possible it is necessary to have good measurements of the biology to assimilate into the ecosystem model. Such measurements exist at limited number of sites (such as Bermuda; see Michaels et al., 1994), or for particular experiments (such as the North Atlantic Bloom Experiment - NABE; see Ducklow & Harris, 1993), but these may not be representative of what is happening at the ocean basin or global scale. The recent availability of ocean colour data from OCTS (Ocean Colour and Temperature Sensor, on the Japanese satellite ADEOS, from August 1996 to June 1997) and from SeaWiFS (Sea-viewing Wide Field-of-view Sensor, July 1997 onwards; Hooker & Esaias, 1993) provide measurements that might help in studying these larger scales. This is the first time that such data have been available since the demise of the Coastal Zone Color Scanner (CZCS) in the mid-1980s. From the ocean colour measurements it is possible to derive chlorophyll, which in turn is representative of the phytoplankton present in the near surface layers of the ocean. The limitation is that only one component of the ecosystem is measured, but in combination with a number of *in situ* measurements that might be sufficient to improve the models, through data assimilation.

As a result of these two factors, the need to improve oceanic ecosystem models and the availability of global ocean colour data, a number of studies have been carried out on data assimilation into biological models. No consensus has yet emerged on what the best way to proceed might be, and this is an active area of research at the present time

(though the total number of people working on the problem world-wide appears to be small).

In view of the above, this report aims to:

- a) critically review work to-date on data assimilation into biological and coupled bio-physical models;
- b) describe some initial results from testing a number of assimilation techniques using a simple predator-prey model;
- c) make preliminary recommendations regarding future work on data assimilation into biological and coupled bio-physical models.

Issues that are not addressed include: the possible chaotic behaviour of simple ecosystem models (see, for example, Scheffer, 1991; Truscott & Brindley, 1994), and the derivation of accurate measurements of chlorophyll from satellite ocean colour data (see, for example, Abbott & Chelton, 1991). Both these can potentially influence the assimilation of data, but lie beyond the scope of this report. In addition, this report will not give detailed descriptions of the assimilation techniques as these can be found in the literature cited, or in reviews such as that of Ghil & Malanotte-Rizzoli (1991), or in standard texts such as Malanotte-Rizzoli (1996) and Sen & Stoffa (1995).

The outline of the report is as follows: the next two sections of the report describe, and critically review, the work on assimilation that has been carried out using simple ecosystem models (section 2) and more complex coupled bio-physical models (section 3). Section 4 then gives some results from comparisons of assimilation techniques using a very simple predator-prey model. Finally, in section 5, discussion, conclusions and recommendations are presented. Note that some initial ideas on this topic have previously appeared in Srokosz et al. (1996), and a good introduction to modelling oceanic ecosystems may be found in Fasham (1993).

2. Simple ecosystem models

Although the ultimate aim is to assimilate data in fully coupled bio-physical ocean models¹, it is easier to begin with simpler ecosystem models and to build up an understanding of the issues involved in assimilating data into them. Despite their simplicity, relative to the coupled models, these ecosystem models still give many insights into the ocean biology (Fasham, 1993). The focus of this section is on these models.

The form of data assimilation that has been applied in this context is that of fitting the models to data, with the aim of better determining the parameters of the models. Altering the parameters in the model (for example, the zooplankton grazing or mortality rates) does not violate the conservation of biomass in the model, but may change the rate of transfer of biomass from one compartment to another (for example, by zooplankton grazing the phytoplankton). Biologically this is an important constraint to satisfy, as the creation / destruction of biomass is not realistic (in the following section on assimilation into coupled bio-physical models we will see that some of the techniques used do not satisfy this condition).

Lawson et al. (1995) were the first people to apply the adjoint technique (see appendix) of Thacker & Long (1988) to the assimilation of data into a biological model. The model used was a simple predator-prey one, perhaps representative in the oceanic context of zooplankton-phytoplankton interactions. They used the standard twin experiment technique where the first model run represents the “real world” from which “data” on the predator and prey are extracted, while the second model run attempts to fit the model to that data starting with parameter values and initial conditions that differ from the real world case. The success of the assimilation is measured by its ability to recover the original real world initial conditions and parameter values. The technique requires an initial guess for the unknowns (initial conditions and parameters) that the assimilation is trying to recover. The results obtained suggest that this technique

¹ Or in the climate context, coupled atmosphere-ocean-terrestrial bio-physical models, such as those run

worked reasonably well, even when noise of up to 20% of the signal was added to the “data”.

However, a number of problems exist with the Lawson et al. (1995) study. First, the parameter values chosen do not seem very representative of the usual cases studied with predator-prey models (see Renshaw, 1993, and section 4 below). Second, one initial guess was used throughout the study (except in one instance, the noisy case; see Lawson et al., 1995, Table 2) and it is unclear whether (particularly in the noisy case) the assimilation might not recover the original values if a different initial guess were used (we will return to this point in section 4 below). Third, they used only data taken from the first half day of their 15 day simulation for the assimilation and did not investigate the effect using differing amounts of data, except for one situation. That situation was the one in which they assimilated only prey data, which might correspond in the oceanic case to only having satellite information on the phytoplankton. Depending on the strength of the predator-prey interactions they found that it was possible to recover the initial conditions and parameter values by only assimilating prey data, but the number of iterations required for convergence of the technique was an order of magnitude larger than for the case where both predator and prey data were assimilated.

Lawson et al. (1996) subsequently applied the same approach to a more complex 5 compartment ecosystem model (phytoplankton, zooplankton, nitrate, ammonium and detritus). The twin experiments carried out again look promising. The main results of this study were that: the frequency of data assimilation required to recover parameter values is dependent on the time scale of the biological processes involved; the sampling strategy affects the recovery of the parameters, in particular the availability of zooplankton data even at a reduced frequency relative to the phytoplankton and nutrient data improves the parameter recovery rate (fewer iterations required); the ability to recover information on episodic events (forced in the study by nutrient injection) is

dependent on the timing of the sampling relative to those events (which may have implications with regard to the data required for assimilation into a model that is trying to reproduce the spring bloom in the North Atlantic, for example). Strangely, they carry out their model runs over periods of order 50 days rather than attempting to reproduce an annual cycle (which is more typical of studies using such models), but no explanation is given for this. It is perhaps related to their comment that a large (unspecified) number of iterations were required to achieve convergence, so perhaps shorter model runs are the result of computational restrictions. They give no information on their choice of initial guess, so it is difficult to tell whether this is a significant factor in their results.

More recently Spitz et al. (1997) have applied the adjoint technique to the Fasham et al. (1990) model, assimilating BATS (Bermuda Atlantic Time-series Study) data. They are able to obtain estimates of all the parameters but conclude that the model is not appropriate for the annual cycle of the BATS ecosystem. This latter conclusion seems odd as the model has previously been tested against data from Bermuda (Fasham et al., 1993).

Fasham & Evans (1995) use a constrained nonlinear optimisation technique (based on the conjugate direction method) to fit the model of Fasham et al. (1990) to NABE data. Their first result, obtained by giving all the data equal weight, led to an overestimation of the zooplankton (by a factor of 2) in the fitted model. Giving more weight to the zooplankton data improved that fit at the expense of a considerable underestimation of primary production and a poorer fit to the bacteria data. Despite being able to adjust the many parameters of the model it was not possible to obtain a good fit to all the data. This may be due to either: the limitations of the model, for example the data suggest two phytoplankton blooms, possibly due to different phytoplankton groups whereas the model only has a single phytoplankton compartment; or to the limitations of the data,

reliable determination of some of the biological variables, such as microzooplankton biomass or bacterial production, at sea, being difficult.

One problem that is apparent with the work discussed so far is the need to specify an initial guess for the parameters that are to be determined through the assimilation of data, and the fact that the method may not converge to the best solution in global sense (that is, it may only determine a locally optimum solution, of which there may be more than one, and the one found may depend on the initial guess chosen). To avoid this problem Matear (1995; following Krüger, 1993) applied the technique of simulated annealing (SA) to fit three different models to data from Station P in the subarctic Pacific. SA should provide the globally optimised values of the parameters required. The models studied by Matear (1995) were a simple 3 compartment model (zooplankton, phytoplankton and nitrate), a 4 compartment model (where the zooplankton were separated into two size classes - microzooplankton and mesozooplankton), and the 7 compartment model of Fasham et al. (1990). Since Matear (1995) was attempting to fit the models to a complete annual cycle, he constrained them to be periodic (with an annual period). His results suggest that the data available at Station P (nitrate, phytoplankton, mesozooplankton, net phytoplankton productivity) are adequate to constrain 10 of the model parameters, which is less than the total number of parameters, meaning that solutions obtained are not unique. He also compared SA with the conjugate gradient method, which failed for the 4 and 7 compartment models. His conclusion was that SA provided an efficient and robust method for assimilating the data into the model. He also concluded that the simplest 3 compartment model was adequate to explain the observations at Station P, and that additional data (ammonium and bacteria concentrations, for example) are required to justify the use of the more complex 7 compartment model. Two observations need to be made about this study. First, the use of what appears to be a heavily weighted periodicity constraint is not really discussed, so its effect on the results obtained is unclear (we will return to this point in section 4). Second, the version of SA used by

Matear (1995) is not true SA, but uses a fast annealing schedule² which violates the theoretical basis of SA that guarantees a globally optimal solution (Sen & Stoffa, 1995). This is clearly a limitation on this technique.

Hurt & Armstrong (1996) have also used SA, but they fit a 4 compartment model (ammonium, nitrate, phytoplankton, recycling term), with 11 parameters, to BATS data. They find that their new model fits the data better than the more complex model of Fasham et al. (1990), despite the fact that it does not explicitly include zooplankton, which is surprising. They use a likelihood measure of fit, but note that, “this is not necessarily the “correct” likelihood model”. It is unclear what effect a different or “correct” choice of likelihood might have on their results.

Another problem with the techniques discussed so far is that of determining the accuracy of the parameters obtained through data assimilation into a model. To overcome this Harmon & Challenor (1997) have applied a technique known as the Markov chain Monte Carlo (MCMC) method to the problem. This method is based on using Bayes theorem, as means of incorporating prior knowledge about the parameters, and the Metropolis-Hasting algorithm (Metropolis et al., 1951; Clifford, 1994) to generate a Markov chain that has the same statistical properties as the posterior (after assimilation of data) distribution of parameters. From the Markov chain it is possible to calculate, for example, the means and standard deviations of the parameters, as well as other properties such as correlations between the estimates. They also applied the method to the Fasham et al. (1990) model using a twin experiment technique, including a case with 20% Gaussian noise added to the “data”. Using sensitivity analysis they determined the 10 parameters to which the model was most sensitive and attempted to recover the top 5 or all 10 of these using the MCMC technique. They found that recovery of information on the parameters depended on the length of the Markov chain generated and its degree of stationarity. For simply reproducing the observations relatively short

² Sometimes known as simulated quenching.

chains, of length $O(10^4)$, were adequate. However, to recover higher order properties, such as the variances of the parameter estimates, much longer chains were needed, of length $O(10^6)$, with a corresponding increase in computational time³. They concluded that the MCMC was robust, gave information on both the parameter estimates and their variances as required, but was computationally intensive. The final point is not necessarily a problem if the assimilation procedure is only to be carried out a limited number of times.

From the above it can be seen that four different techniques - adjoint, conjugate direction, SA and MCMC - have been used to assimilate data into ecosystem models with varying degrees of success. However, no detailed comparison of the methods has been carried out so it is difficult to assess the methods against each other. We will return to this issue in section 4 of report. Next we move on to consider what has been done in the area of data assimilation into coupled bio-physical models.

3. Coupled bio-physical models

The first person to assimilate data into a coupled bio-physical model appears to have been Ishizaka (1990c), for the Southeastern U.S. continental shelf area. The data assimilated came from CZCS and the assimilation was of the predictive type (rather than the model fitting / parameter determination type). The model was a 4 compartment ecosystem model (nutrient, phytoplankton, zooplankton, detritus), vertically integrated with horizontal advection and eddy diffusion, but with a semi-empirical upwelling / downwelling term added (Ishizaka, 1990b). The coupling to the physics is through the advective velocities, which Ishizaka (1990a, b) obtained from optimally interpolated circulation fields⁴. Ishizaka's assimilation procedure is that of direct insertion, where the phytoplankton values in the model were simply replaced by those estimated from the

³ Typically ~100 hours on Silicon Graphics R4000 machine. The Fasham et al. (1990) model takes ~0.25 seconds to run on this machine, but its has to be run for $\sim 10^6$ iterations.

⁴ Strictly this is not a coupled model as the physics is taken from observations (Ishizaka, 1990a). The advective velocities could be obtained from a model, of course, rather than from data.

CZCS ocean colour data. Three types of adjustment were applied to the other compartments (nutrient, zooplankton, detritus) to allow for conservation of biomass, and this seemed to make little difference to the results obtained. The model was run forward from the assimilation time and the results compared with CZCS data at later times (comparisons made over a period of a few days). Two cases were considered: one in which the biological interaction and upwelling / downwelling terms were switched off, so that the resulting biological distributions are due to purely the physical advective and diffusive processes, and one where the biological interactions are included. The results suggest that advection dominates in this case, with biological processes acting as a secondary factor. However, the impact of the data assimilation lasted only a few days indicating the need to assimilate data frequently (every 1-2 days) to keep the model updated and the errors in prediction small. Assimilation of the data had a positive effect overall, but did degrade some aspects of the model (estimates of nutrient fluxes as compared to *in situ* data). Ishizaka (1990c) also tested the impact of only assimilating data over part of the region modelled, thus simulating the effect of partially cloudy data. This too seemed to give improved results, though it did cause discontinuities at the boundaries between the areas where data were available for assimilation and where there were none. Direct insertion of data is known to cause problems in physical model, as the model adjusts by radiation of waves (Ghil & Malanotte-Rizzoli, 1991), but this did not appear to cause problems for the advective-diffusive biological model used by Ishizaka (1990c). One observation to be made about the Ishizaka (1990c) study is that the use of a vertically integrated model avoids the issue of how to relate the satellite surface observation of chlorophyll to the subsurface structure in the biology (there being none in this model, as there is no depth dependence).

Subsequently, Ishizaka (1993) has reported some preliminary results on the use of direct insertion with simpler ecosystem models. He concludes that, "...assimilation of phytoplankton data into a simple time-dependent model is not straightforward.

Differences in error sources, methods of assimilation, and the timing of the assimilation all result in different model solutions... Furthermore, much *more* complex ecosystem structure is preferable for the real simulation and *this* increases uncertainty of the effects of the data assimilation.”⁵

Sarmiento et al. (1993) compared their results for the seasonal chlorophyll distribution of the Atlantic Ocean, obtained by embedding the Fasham et al (1990) ecosystem model in a ocean general circulation model (OGCM), with data from CZCS. They found reasonable agreement overall, but some specific discrepancies. Armstrong et al. (1995) followed this up by attempting to assimilate CZCS data into the same model at frequencies of 1 and 5 day⁻¹. Comparisons improved overall compared to the unforced case of Sarmiento et al. (1993), except at high latitudes; with the more frequent forcing giving better results. They attributed the discrepancy at high latitudes as being related to the modelling of the zooplankton grazing of the phytoplankton (see their paper for details). They therefore incorporated multiple grazing chains in the model, giving it a total of 13 compartments - nitrate, ammonium, bacteria and dissolved organic nitrogen (DON), plus three compartments each for detritus, phytoplankton and zooplankton (see Figure 16-2 of Armstrong et al., 1995; n.b. the Fasham et al., 1990, has only one compartment for each of detritus, phytoplankton and zooplankton). Assimilation of CZCS data at the same frequencies as before led to much better agreement. The assimilation technique used was “nudging” (see Ghil & Malanotte-Rizzoli, 1991, for more information on this), applied to bring the model chlorophyll values in the upper layer of the model (top 10m) towards the surface observations from CZCS. This procedure requires the model to adjust the subsurface biological distributions, and violates the conservation of biomass, which is either added or removed from the model depending on whether the satellite-model difference in chlorophyll is positive or negative. Simplistically one might think that forcing the model values towards the observations would eventually result in complete agreement, but this is not the case due

⁵ I have added the italicised words as the original does not read quite right.

to deficiencies in the model physics and biology. This again is a case of using data assimilation to fit a model to data.

The previous two models discussed were 2-D (horizontal) and 3-D in space, the next one that will be discussed is 1-D (vertical) model of Prunet et al. (1996a, b). This is a 10 compartment ecosystem model coupled to a 1-D mixed layer (Prunet et al., 1996a) and the data assimilated are from Station Papa (Station P; 50°N, 145°W). This is a fairly complex model, so the details will not be discussed here. The difference from the other studies discussed so far is that Prunet et al. (1996b) assimilate data into both the biological and physical components of their model. A variational assimilation technique is used, involving the minimisation of a cost function. In their first paper (Prunet et al., 1996a) they consider assimilating only surface chlorophyll data into the model and find that this can only partially constrain the parameters of the model (in particular, five linear combinations of the 11 model parameters that affect the surface chlorophyll distribution; the ecosystem model has a total of 47 parameters). Furthermore the assimilation of surface chlorophyll is not sufficient to constrain the vertical structure of the chlorophyll. They state that this case is nearly optimal in the sense that the ocean physics at Station P are comparatively simple and good quality *in situ* data are available to estimate the physical forcing. Therefore further testing of the approach in different oceanic regimes (biogeochemical provinces) is required. In their second paper (Prunet et al., 1996b) use a simpler 4 compartment ecosystem model (nitrate, phytoplankton, zooplankton, detritus) coupled to the same mixed layer model as in their first paper (Prunet et al., 1996a). They find that surface chlorophyll assimilation is not sufficient to reproduce the seasonal cycle of surface chlorophyll, temperature and nitrate in a robust manner. By additionally assimilating surface nitrate and temperature the model adjustment is improved when tested against other independent (not assimilated) surface chlorophyll and nitrate data. Comparison between the two models suggests that choice of model structure can affect the results obtained (for example, for the proportion of primary production consumed by grazing).

The final piece of work to consider in this section is that carried out on a recent cruise in the North Atlantic (see Srokosz et al., 1997). During the cruise an attempt was made to assimilate the biological and physical data being collected, in near real time, into the Harvard Ocean Prediction System (HOPS; Lozano et al., 1996). HOPS consists of a limited area primitive equation model, coupled with a 5 compartment biological model (phytoplankton, zooplankton, nitrate, ammonium, detritus). The data were assimilated using an objective analysis and optimal interpolation procedure (Robinson, 1996; Lozano et al., 1996), similar to those commonly used in meteorology and physical oceanography (Ghil & Malanotte-Rizzoli, 1991). Some problems were experienced in assimilating the data, but initial predictions were made on the cruise and some progress was achieved (see Srokosz et al., 1997). Further work is being carried out in a hindcast study that is currently underway.

4. Some comparisons

As noted above, in section 2, no comparison has been carried out between the various assimilation techniques that have been proposed for use with ecosystem models. The purpose of this section is to give some results from a number of initial comparisons of three of the methods - adjoint, SA, MCMC. Since the studies discussed above use a variety of ecosystem models it is difficult to make direct comparisons between them, so for this reason a single model is used here. This means that any differences in the results will be due to the assimilation techniques applied, rather than due to changes in the model which may be confounded with those from the techniques being tested. In addition, the discussion of section 2 showed that application of some of the techniques to some ecosystem models was computationally expensive (see, for example, Lawson et al., 1996; Harmon & Challenor, 1997). Therefore, the model chosen for study here is the very simple predator-prey one originally used by Lawson et al. (1995).

The model is

$$\dot{x}(t) = x(a_1 + a_2x + a_3y)$$

$$\dot{y}(t) = y(a_4 + a_5y + a_6x)$$

with initial conditions

$$x(t_0) = x_0 \text{ and } y(t_0) = y_0$$

where x and y represent the prey and predator, respectively (see Renshaw, 1993). The parameters to be determined by fitting model (x, y) to $(\hat{x}_j, \hat{y}_j; j = 1, \dots, N)$ data are

$$\theta_i \ (i = 1, \dots, 8) = a_i \ (i = 1, \dots, 6), x_0, y_0$$

Two versions of the model will be studied, that originally considered by Lawson et al. (1995) in which the solution spirals in to a stable fixed point, and a cyclical (periodic) solution which might be taken as representative of an annual cycle. Note that more realistic oceanic ecosystem models may have limit cycle behaviour, in that their solutions settle down to the same annual cycle even if the initial conditions are perturbed. The Fasham et al. (1990) model appears to exhibit this behaviour. Here different initial conditions will lead to different cyclical behaviour. The values of the parameters for the two cases are:

| | a_1 | a_2 | a_3 | a_4 | a_5 | a_6 | x_0 | y_0 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| stable fixed point | 4 | -2 | -4 | -6 | 2 | 4 | 1 | 1 |
| periodic | 1.5 | 0 | -0.1 | -0.25 | 0 | 0.01 | 10 | 15 |

As noted earlier, the Lawson et al. (1995) choice of values is non-standard (see Renshaw, 1993), and in particular as $a_5 > 0$ this gives a quadratic (rapid) growth term for the predators. This appears to cause some problems when assimilating data in this case.

To carry out the comparisons the standard twin experiment approach is used. Data are generated from the model, noise is added, and the assimilation is used to fit the model

to the data, but starting with incorrect parameter values. The table below shows which cases have been studied to-date:

| | adjoint | SA | MCMC |
|---------------------------------|---------|----|------|
| stable fixed point | √ | √ | √ |
| periodic | TBD | √ | √ |
| periodic with cyclic constraint | TBD | √ | √ |

The fixed point case is that of Lawson et al. (1995). The periodic case is analogous to the annual cycle in the ocean. The periodic with cyclic constraint case is a variation on the periodic case, in an attempt to investigate the effect on the assimilation of a periodicity constraint as used by Matear (1995). Since the details of the comparisons are fairly involved only the results obtained so far will be summarised here.

Since the SA and MCMC methods are statistical in nature the results obtained should be independent of the starting values (initial guess, or prior probability) assigned to the parameters. This is not the case in the adjoint method. Both SA and MCMC work by exploring the parameter space of possible values, while the adjoint will head for the nearest local minimum of the cost function in the parameter space. Lawson et al. (1995) use a single initial guess in their paper, here for the adjoint case the initial guess is randomly perturbed and the assimilation carried out for 100 cases to test whether the method is robust.

Starting with the fixed point case of Lawson et al. (1995), of the methods tested, the adjoint is the fastest computationally, but often fails in to converge depending on the circumstances (absence / presence / degree of noise in data). This situation seems to arise due to the quadratic growth term ($a_5 > 0$) which causes the assimilation to “blow up” numerically for some choices of the initial guess for the parameters. It is not clear why this is so, but using the single Lawson et al. (1995) initial guess causes no problem. However, they give no discussion in their paper as to why that guess is the appropriate one to use, and in any application of the method to fit a model to real data

the choice will be to some degree arbitrary. This is therefore a potential problem with the adjoint approach. Lawson et al. (1995) only assimilated data from the first half day of their 15 day simulation, which works well if there is no noise. For noisy data better results are obtained (unsurprisingly) if more data are assimilated (say, from the first 1.5 days).

Applying the SA algorithm to the stable fixed point case gave much poorer results than those reported by Matear (1995). The algorithm was not able to recover all the parameters, for reasons that are unclear at this stage. In contrast, the MCMC method worked well but took many iterations $O(10^6)$ when assimilating half a day of data with noise. If more data were used (1.5 days) then the number of iterations required to recover the parameters was smaller $O(10^5)$. Additionally, the variances and covariances of the parameter estimates were obtained from MCMC.

For the periodic case, to-date, only the SA and MCMC methods have been tested. Note that here it is only necessary to recover 6 parameters $(a_1, a_3, a_4, a_6, x_0, y_0)$, rather than the 8 required for the previous case. For this case the SA worked well and had no difficulty in recovering the parameter values. Similarly, the MCMC also recovered the parameter values well, and in many fewer iterations $O(10^4)$ than for the fixed point case. This is probably a result of assimilating data throughout the periodic cycle, rather than just for some initial time period as was done for the fixed point case.

Finally, an attempt was made to test the usefulness of adding a periodicity constraint in the cyclic case, as Matear (1995) had in his study. This is straightforward to do for the SA, but not so for the MCMC. In the latter case it was necessary to make an *ad hoc* modification of the likelihood function used, which probably invalidates the assumptions on which the MCMC method is based. Using the same relative weighting in the assimilation of the periodicity constraint as Matear (1995) led to the SA converging to a set of parameter values that gave a periodic solution of twice the

frequency (half the period) of the correct solution. Reducing the weight of the periodicity constraint sufficiently eventually allowed the SA to recover the correct parameter values. For the MCMC, the constraint distorts the likelihood function and causes the MCMC to recover incorrect parameter values. For this cyclic case the periodicity constraint seems to degrade the performance of both SA and MCMC. Matear (1995) gives no reason for including a periodicity constraint in his work. Hurtt & Armstrong (1996), who also used SA, did not use such a constraint. This suggests that it might not be necessary and that Matear (1995) might have obtained equally good results without it.

Some tests have also been carried out on assimilating only prey data, equivalent to having only phytoplankton data in the oceanic case (say, from satellite ocean colour), and on assimilating prey data plus occasional predator data, equivalent to having some zooplankton data from a small number of ship observations. Depending on the amount of noise in the data, the degree of coupling between the predator and prey (determined by the values of the parameters a_3 and a_6) and the method of assimilation used, it is possible in some circumstances to recover the parameters using just prey data. Adding the occasional predator observation improves the recovery if the data are not too noisy and sufficient observations are used. This problem requires further work in order to quantify these various effects.

Overall, these initial comparisons suggest that the MCMC method is the most robust, but not the most efficient computationally. The adjoint is the most efficient computationally, even taking account of the fact that 100 randomly varied initial guesses were used in the tests (this means that 100 runs of the adjoint case are quicker than 1 run of the MCMC). However, the adjoint is far from robust. SA has the advantage that it is easy to implement numerically, compared to the adjoint and MCMC, but it did not work well for the fixed point case. The advantage of the MCMC is that not only does it

recover the parameter values, but it also provides information how well it has performed through the information that it provides on their variances and covariances.

5. Discussion, conclusions and recommendations

From the above review (sections 2 and 3) and the comparison of three methods (section 4) the following points emerge:

- as yet there is no accepted “best” method for assimilating data into ecosystem models, and it is not apparent which method might be best for a specific application.
- the MCMC method seems to be robust, though relatively computational expensive, when applied to ecosystem models to fit them to data and so obtain improved parameter estimates.
- successful data assimilation is dependent on the quality and quantity of data available, and perhaps more importantly the density of observations relative to the occurrence of significant biological events.
- under certain conditions it is possible to constrain some models with only information about phytoplankton. This means that ocean colour data on their own may prove adequate for this purpose.
- for consistent ecosystem modelling it is necessary to ensure that the assimilation of data into the model does not violate the principle of conservation of biomass. Not all the studies reviewed adhere to this principle, and some make *ad hoc* adjustments to ensure that it is satisfied.
- the only attempt to assimilate data for predictive purposes is that of Ishizaka (1990c).
- the only study of simultaneous assimilation of biological and physical data into a coupled model is that of Prunet et al. (1996b).
- the review and comparison has raised more questions than answers, and indicates that the subject of assimilation into ecosystem models is very much in its infancy.

Issues that have not been considered in the work carried out up to the present time, and should perhaps be studied, include:

○ twin experiment studies to-date have been done under the “perfect model” assumption; that is, the same model is fitted to the “data” as generated the data. This will, of course, give over optimistic results. In practice, any model of the real oceanic ecosystem will have deficiencies, so will not satisfy the “perfect model” assumption. This problem can be investigated by generating data from a complex model and fitting a degraded or simpler model to that data. For certain types of problem it may not be necessary to use the most complex model available, and this too could be investigated in this manner.

○ rather than converting ocean colour data to chlorophyll or phytoplankton biomass, it might be possible to assimilate the colour data directly. This would be analogous to the recent work in meteorology, where satellite measured radiances are assimilated into the meteorological model directly, rather than first being converted to atmospheric temperatures. The model would be used to predict the expected colour based on the model phytoplankton values, using a bio-optical algorithm. The mismatch between the observed and predicted colour would then be used to drive the assimilation.

○ using the adjoint method requires the construction of the adjoint code corresponding to the model (see Lawson et al., 1995; Thacker & Long, 1988). Given the heuristic nature of ecosystem models this might prove a tedious task. However, automatic methods for adjoint code generation now exist (Giering & Kaminski, 1995; developed for meteorological applications) and could be used in this context.

The next step in developing data assimilation into ecosystem models is not entirely clear, given the wide range of problems that the preceding review has highlighted in the work carried out to-date. Consequently, the recommendations made below are tentative and aim to build on the results that are more sure.

Recommendations for potential ways forward for studying the assimilation of data into ecosystem models are:

- given the computational efficiency of the adjoint method, this could be developed further. First, by applying the automatic adjoint generating techniques (Giering & Kaminski, 1995), and second, by attempting to make it more robust.
- given the robustness of the MCMC method, this too might be developed further by attempting to make it more computationally efficient. One way forward could be combine the adjoint and MCMC methods in a way that takes advantages of the strengths of each (there are some indications that this might be possible; P. Challenor, pers. comm.).
- given the apparent importance of the frequency and timing of observations to be assimilated into the models, further investigation of the observation strategy might be useful in determining the best way to assimilate data.
- given the over optimistic nature of the results from twin experiments in the “perfect” model case, assimilation of data into simpler or degraded version of the model could be studied. This would give insight into the problem of data assimilation in the real oceanic case, where any model used will not fully represent the ecosystem being studied.
- given the ultimate aim of data assimilation into coupled bio-physical models, a simple 1-D coupled mixed layer ecosystem model could be developed and the simultaneous assimilation of biological and physical data studied (building on the initial work of Prunet et al., 1996b).

Acknowledgements

I am grateful to Robin Harmon for his work on the comparison of models, which made it possible to write section 4 of this report.

References

- Abbott M.R. 1992 Report of the U.S. JGOFS workshop on modeling and data assimilation, U.S. JGOFS Planning Report No. 14, 28pp.
- Abbott M.R. & Chelton D.B. 1991 Advances in passive remote sensing of the ocean. *Rev. Geophys.*, Suppl. 571-589.
- Armstrong R.A., Sarmiento J.L. & Slater R.D. 1995 Monitoring ocean productivity by assimilating satellite chlorophyll in to ecosystem models. pp.379-390 in *Ecological Time Series* (eds. T.M. Powell & J.H. Steele), Chapman & Hall.
- Clifford P. 1994 *in discussion of* "Approximate Bayesian inference with the weighted likelihood bootstrap" by Newton M.A. & Raferty A.E., *J. Roy. Statist. Soc. B*, **56**, 34-35.
- Ducklow H.W. & Harris R.P. 1993 Introduction to the JGOFS North Atlantic Bloom Experiment, *Deep-Sea Res.*, **40**, 1-8.
- Evans G.T. & Parslow J.S. 1985 A model of annual plankton cycles, *Biol. Oceanogr.*, **3**, 327-347.
- Fasham M.J.R. 1993 Modelling the marine biota, pp. 457-504 in *The Global Carbon Cycle* (ed. M. Heimann), Springer-Verlag.
- 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. & Evans, G.T. 1995. The use of optimisation techniques to model marine ecosystem dynamics at the JGOFS station at 47N 20W. *Phil. Trans. R. Soc. Lond.* **B348**, 203-209.
- 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**, 379-415.
- Ghil M. & Malanotte-Rizzoli P. 1991 Data assimilation in meteorology and oceanography, *Adv. Geophys.*, **33**, 141-266.

- Giering R. & Kaminski T. 1995 Recipes for adjoint code construction. Technical report, Max-Planck-Institut für Meteorologie.
- Gill A.E. 1982 Atmosphere-Ocean Dynamics, Academic Press Inc., San Diego, 662pp.
- Harmon R.T. & Challenor P.G. 1997 A Markov Chain Monte Carlo method for estimation and assimilation into models. *Ecol. Mod.*, **101**, 41-59.
- Hofmann E., Osborn T., Powell T., Price J., Rothschild B. & Roughgarden J. 1991 Theory and modeling in GLOBEC: a first step, U.S. GLOBEC Report, 9pp.
- Hooker S.B. & Esaias W.E. 1993 An overview of the SeaWiFS project, *Eos, Trans. AGU*, **74**, 241 & 245-246.
- Hurtt G.C. & Armstrong R.A. 1996 A pelagic ecosystem model calibrated with BATS data. *Deep-Sea Res. II*, **43**, 653-683.
- Ishizaka J. 1990a Coupling of coastal zone color scanner data to a physical-biological model of the southeastern U.S. continental shelf ecosystem, 1. CZCS data description and Lagrangian particle tracing experiments, *J. Geophys. Res.*, **95**, 20176-20181.
- Ishizaka J. 1990b Coupling of coastal zone color scanner data to a physical-biological model of the southeastern U.S. continental shelf ecosystem, 2. An Eulerian model, *J. Geophys. Res.*, **95**, 20183-20199.
- Ishizaka J. 1990c Coupling of coastal zone color scanner data to a physical-biological model of the southeastern U.S. continental shelf ecosystem, 3. Nutrient and phytoplankton fluxes and CZCS data assimilation, *J. Geophys. Res.*, **95**, 20201-20212.
- Ishizaka J. 1993 Data assimilation for biogeochemical models, pp. 295-316 in *Towards a model of ocean biogeochemical models* (eds. G.T. Evans & M.J.R. Fasham), Springer-Verlag.
- Krüger J. 1993 Simulated Annealing: a tool for data assimilation into an almost steady model state, *J. Phys. Oceanogr.*, **23**, 679-688.

- Lawson, M.L., Hofmann, E.E. & Spitz, Y.H. 1996 Time series sampling and data assimilation in a simple marine ecosystem model. *Deep-Sea Res. II*, **43**, 625-651.
- Lawson, M.L., Spitz, Y.H., Hofmann, E.E. & Long, R.B. 1995. A data assimilation technique applied to a predator-prey model. *Bull. Math. Bio.*, **57**, 593-617.
- Lozano C.J., Robinson A.R., Arango H.G., Gangopadhyay A., Sloan Q., Haley P.J., Anderson L. & Leslie W. 1996 An interdisciplinary ocean prediction system: assimilation strategies and structured data models, pp.413-452 in *Modern approaches to data assimilation in ocean modeling* (ed. P. Malanotte-Rizzoli), Elsevier, 468pp.
- Malanotte-Rizzoli P. (ed.) 1996 *Modern approaches to data assimilation in ocean modeling*, Elsevier, 468pp.
- Matear R.J. 1995 Parameter optimisation and analysis of ecosystems models at station P using simulated annealing. *J. Marine Res.*, **53**, 571-607.
- Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. & Teller E. 1953 Equations of state calculations by fast computing machine, *J. Chem. Phys.*, **21**, 1087-1091.
- Michaels A.F., Knap A.H., Dow R.L., Gundersen K., Johnson R.J., Sorensen J. & Close A. 1994 Seasonal patterns of ocean biogeochemistry at the U.S. JGOFS Bermuda Atlantic Time-series Study site, *Deep-Sea Res. I*, **41**, 1013-1038.
- Prunet P., Minster J-F., Ruiz-Pino D. & Dadou I. 1996a Assimilation of surface data in a one dimensional physical-biogeochemical model of the surface ocean 1. Method and preliminary results. *Global Biogeochem. Cycles* **10**, 111-138.
- Prunet P., Minster J-F., Echevin V. & Dadou I. 1996b 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 pCO₂ data. *Global Biogeochem. Cycles* **10**, 139-158.
- Renshaw E. 1993 *Modelling biological populations in space and time*, Cambridge University Press, 403pp.

- Robinson A. R. 1996 Physical processes, field estimation and an approach to interdisciplinary ocean modeling, *Earth-Science Rev.*, **40**, 3-54.
- Sarmiento J.L. & Armstrong R.A. 1997 U.S. JGOFS Synthesis and modelling project implementation plan, U.S. JGOFS Report, 73pp.
- 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, *Global Biogeochem. Cycles*, **7**, 417-450.
- Scheffer M. 1991 Should we expect strange attractors behind plankton dynamics - and if so, should we bother?, *J. Plankton Res.*, **13**, 1291-1305.
- Sen M. & Stoffa P.L. 1995 Global optimisation methods in geophysical inversion, Elsevier, 281pp.
- Siegenthaler U. & Sarmiento J.L. 1993 Atmospheric carbon dioxide and the ocean, *Nature*, **365**, 119-125.
- Spitz, Y.H., Moisan, J.R., Abbott, M.R. & Richman, J.G. 1997 Data assimilation and a pelagic ecosystem model: Parameterization using time series observations. submitted to *J. Marine Systems*.
- Srokosz M.A., Challenor P.G., Fasham M.J.R. & Harmon R. 1996 Using SeaWiFS (ocean colour) data in biological ocean model validation and data assimilation, pp.13-16 in: Proc. of the first SeaWiFS Exploitation Initiative (SEI) team meeting (eds. G.F. Moore & S.B. Hooker), Southampton, 24th Jan. 1995, NASA SeaWiFS Tech. Rep. Series, vol. 33, NASA Tech. Memo 104566.
- Srokosz M.A. et al. 1997 Plankton patchiness studies by ship and satellite - P²S³, *RRS Discovery Cruise 227*, 15 April - 16 May 1997, SOC Cruise Report No. 12, 76pp.
- Thacker W.C. & Long R.B. 1988 Fitting dynamics to data, *J. Geophys. Res.*, **93**, 1227-1240.
- Truscott, J.E. & Brindley, J., 1994. Equilibria, stability and excitability in a general class of plankton population models. *Phil. Trans. R. Soc. Lond. A.* **342**, 703-718.

Appendix: a note on the adjoint method

The adjoint method is strictly speaking an efficient way of calculating the gradient of the cost function, with respect to the unknown parameters, rather than a means of determining those parameters directly (Thacker & Long, 1988). The mismatch between the observations and the model are used to form a measure of mismatch known as the cost function (a variety of cost function formulations are possible; see Ghil & Malanotte-Rizzoli, 1991). The cost function is then minimised⁶ (usually numerically) with respect to the parameters that need to be determined. Many minimisation routines work most efficiently when supplied with the gradient of the cost function. In principle, this can be determined by differentiating the function with respect to the parameters, but may, in practice, be difficult to calculate for a complicated model that is implemented numerically. The adjoint technique allows this cost function gradient to be determined in a computationally efficient manner. This information is then used in the minimisation. It is usual to refer to the whole process as the adjoint method (as is done in this report).

It is necessary to remember that the adjoint method consists of these two components: the adjoint gradient calculation, and the minimisation of the cost function. It is possible to use a variety of techniques for that minimisation. Lawson et al. (1995, 1996) used a steepest descents quasi-Newton method, while Thacker & Long (1988) used a conjugate gradient descent method. The choice of minimisation routine might also affect the results of the assimilation. Lawson et al. (1995) tested a number of methods and found that, “...these yielded no significantly better results”.

The final point to bear in mind when using the adjoint method is that the adjoint of the model and the adjoint of the numerical implementation of the model are not necessarily the same thing. It is important that the adjoint of the numerical model be used in carrying out the assimilation (Giering & Kaminski, 1995).

⁶ The procedure of fitting the model to the data is sometimes also called optimisation (optimising of the model fit, minimising of the model-data mismatch).

Abbreviations and acronyms

BATS - Bermuda Atlantic Time-series Study

CZCS - Coastal Zone Color Scanner

GCM - general circulation model (of either ocean or atmosphere physics)

GLOBEC - Global ocean ecosystem dynamics

HOPS - Harvard Ocean Prediction System

JGOFS - Joint Global Ocean Flux Study

MCMC - Markov chain Monte Carlo method

NABE - North Atlantic Bloom Experiment

OCTS - Ocean Colour and Temperature Sensor

OGCM - ocean GCM

SA - Simulated Annealing

SeaWiFS - Sea-viewing Wide Field-of-view Sensor

SST - sea surface temperature

TBD - to be done