Split-domain calibration of an ecosystem model using satellite ocean colour data

John C. P. Hemmings *, Meric A. Srokosz, Peter Challenor, Michael J. R. Fasham

Southampton Oceanography Centre, European Way, Southampton SO14 3ZH, UK

Abstract

The application of satellite ocean colour data to the calibration of plankton ecosystem models for large geographic domains, over which their ideal parameters cannot be assumed to be invariant, is investigated. A method is presented for seeking the number and geographic scope of parameter sets which allows the best fit to validation data to be achieved. These are independent data not used in the parameter estimation process. The goodness-of-fit of the optimally calibrated model to the validation data is an objective measure of merit for the model, together with its external forcing data. Importantly, this is a statistic which can be used for comparative evaluation of different models. The method makes use of observations from multiple locations, referred to as stations, distributed across the geographic domain. It relies on a technique for finding groups of stations which can be aggregated for parameter estimation purposes with minimal increase in the resulting misfit between model and observations.

The results of testing this split-domain calibration method for a simple zero-dimensional model, using observations from 30 stations in the North Atlantic, are presented. The stations are divided into separate calibration and validation sets. One year of ocean colour data from each station were used in conjunction with a climatological estimate of the station's annual nitrate maximum. The results demonstrate the practical utility of the method and imply that an optimal fit of the model to the validation data would be given by two parameter sets. The corresponding division of the North Atlantic domain into two provinces allows a misfit-based cost to be achieved which is 25% lower than that for the single parameter set obtained using all of the calibration stations. In general, parameters are poorly constrained, contributing to a high degree of uncertainty in model output for unobserved variables. This suggests that limited progress towards a definitive model calibration can be made without including other types of observations.

Key words:
ecosystem modelling, parameter estimation, data assimilation, biogeochemical provinces

1 Introduction

An ability to predict the response of the pelagic ecosystem to physical changes in the environment is a prerequisite for quantifying potentially important climate feedbacks involving the marine carbon cycle. Progress in this area requires the development of ecosystem models which can be used to extrapolate reliable biological predictions from external forcing data describing physical variability. These models can be coupled with basin-scale or global-scale general circulation models. Useful preliminary results have already been obtained in this way with some specific ecosystem models (e.g. Sarmiento et al., 1993; Oschlies and Garçon, 1998; Oschlies et al., 2000; Sarmiento et al., 2000; Gregg, 2001; Oschlies, 2001; Palmer and Totterdell, 2001; Christian et al., 2002). However, a wide range of candidate ecosystem models are available, varying in complexity and employing many different structures and functional forms. Evaluating their relative merits in an objective way poses a major challenge.

All of the candidate models contain a large number of parameters, many of which are poorly known or represent quantities which, in nature, are highly variable in time and space and across taxa. Parameter uncertainty acts as a barrier to the evaluation of one model against another. The models need first to be calibrated against observational data. This has been done for a number of models using time series data collected at various sites. The Bermuda Atlantic Time-series Study (BATS 32°N 64°W) site has received particular attention (Hurtt and Armstrong, 1996, 1999; Spitz et al., 1998; Fennel et al., 2001; Spitz et al., 2001; Schartau et al., 2001). Calibrations have also been performed for Ocean Weather Station Papa (50°N 145°W) in the Pacific (Matear, 1995; Prunet et al., 1996a,b), the North Atlantic Bloom Experiment site at 47°N, 20°W (Fasham and Evans, 1995; Evans, 1999) and at the Tropical Atmosphere Ocean mooring at 140°W in the equatorial Pacific (Friedrichs, 2002). Hurtt and Armstrong (1999) calibrated a model using BATS data and observations from Ocean Weather Station India (OWSI 59°N 20°W) simultaneously. These studies are all based on *in situ* data although satellite ocean colour measurements, now routinely available from SeaWiFS (McClain et al., 1998) and other sensors, are now beginning to be used to improve the coverage of the annual cycle (Friedrichs, 2002).

Model parameters are estimated by first defining either a cost function or a likelihood function, based on the misfit between the model output and observations. This is a function of the model's parameter set, defined by a vector in the model's parameter space. An optimization technique is then applied to search the parameter space for values which minimize the cost or maximize

* Corresponding author. Fax: +44 (0)23 8059 6400
Email address: J.Hemmings@soc.soton.ac.uk (John C. P. Hemmings).

the likelihood. In a number of cases the calibration studies are supported by identical twin experiments in which simulated observations, sometimes including stochastic errors, are assimilated in an attempt to recover the parameter set of the model from which they were generated (Spitz et al., 1998; Fennel et al., 2001; Spitz et al., 2001; Schartau et al., 2001; Friedrichs, 2001). These experiments provide important information on the robustness of the optimization methods and the ability of different numbers and types of observations to constrain parameter values independently, under the assumption that the model is a perfect description of reality. However, as pointed out by Schartau et al. (2001) inferences from twin experiments are only strictly applicable to optimization results for real observations if the reference parameter values used to generate the simulated observations are close to the optimum values sought for describing the real data.

While an important goal in climate modelling is to develop models which can be applied globally, virtually all of the calibrations performed so far are based on observations at single locations. The geographic scope over which they can be usefully applied has yet to be properly assessed, although Hurtt and Armstrong (1999) did examine the performance of their model, calibrated using Atlantic observations, at the Hawaii Ocean Time-series site in the Pacific. Formal validation of model output against independent data not used in the parameter estimation process, whether this be local data for a different time period or data from another location, is important for quantifying the predictive skill of a model. The calibrated models of Prunet et al. (1996a,b) and Friedrichs (2002) were tested against independent local data. However, in general, this aspect of model evaluation has so far received relatively little attention. Satellite ocean colour measurements provide estimates of surface chlorophyll concentration over multiple annual cycles throughout most of the world ocean. They are therefore an invaluable resource for calibrating and validating models and for analysing the geographic scope of models which have been calibrated for specific locations.

Although ocean colour measurements give information relating to phytoplankton components of the ecosystem only, their use in model calibration in conjunction with other observations should lead to more widely applicable parameter vectors. The exploitation of these data for such purposes is just beginning. Losa et al. (submitted) have recently used ocean colour data to examine the variability in model parameters arising from local calibrations over the North Atlantic basin, while Hemmings et al. (2003) attempted to retrieve parameter vectors applicable over a wide range of different environmental conditions by fitting a model simultaneously at multiple stations. The present study builds on the latter work, seeking the most effective way of using ocean colour data in model calibration and validation.

When calibrating a model using data from multiple locations simultaneously

it is implicitly assumed that the same parameter vector is appropriate for all locations and that spatial variability is purely a result of the different environmental conditions described by the model's external forcing data. However, as demonstrated for the North Atlantic (Hemmings et al., 2003), assuming *a priori* that a single parameter vector is appropriate for a large domain can lead to a sub-optimal calibration. Differences may exist between ecosystems in different regions which are independent of factors represented in the forcing data, but can be reproduced by different parameter vectors. The most obvious example is differences in the taxonomic composition of the plankton. It is therefore desirable to investigate whether better results can be obtained by splitting the geographic domain into provinces and calibrating the model separately for each. This immediately introduces the problem of how best to split the domain. One possible approach is to use the biogeochemical provinces defined by Longhurst (1998), which are based on observed spatial variation in the characteristics of the annual phytoplankton cycle, combined with our knowledge of the relevant physical features of ocean regions. However, the extent to which the observed differences in the annual cycle imply significant differences in the ecosystem response to physical forcing, as opposed to simply reflecting the variability in that forcing, is unclear. We present here a more flexible approach, which does not require any prior assumptions about the geographic scope of individual parameter vectors. The utility of the method is tested by applying it to a simple candidate model.

A major problem in model calibration is the difficulty of locating the global minimum of a cost function in parameter space. The effects of different parameters on model output are often correlated, potentially making the inverse problem underdetermined, even in identical twin experiments where the model is perfect and there is no observational error. In real applications, model inadequacy, observation error and poor data availability compound the problem. Observational constraints on parameter values are normally weak and the existence of a single global minimum which is significant in the presence of observational error is unlikely. Fortunately though, to objectively choose between models we do not actually need to locate the cost function minimum for each in the parameter space, provided we can estimate its value. Therefore, in contrast with previous work, the emphasis here is on estimating the minimum cost obtainable for the given model, together with the associated posterior parameter probability distribution(s), rather than on finding a single optimal parameter vector for each province. While it is unnecessary for the parameters to be independently constrained by the observations to estimate the minimum cost, the extent to which they can be constrained does have important implications for the uncertainty associated with model predictions. The uncertainty in model output associated with the posterior parameter distributions is therefore examined.

The calibration method is first described generically in Section 2. Section 3

then describes how the method was tested by using it to evaluate the candidate model for the North Atlantic. The results of the test are presented in Section 4 and the utility of the method is discussed in Section 5.

2 The calibration method

2.1 Overview

The calibration procedure involves minimizing a cost based on the misfit between model output and observations at multiple locations in the model domain, referred to as stations. These stations are divided into two sets. One is a calibration set, from which stations are combined in different calibration groups to obtain parameter vector estimates. The other is a validation set. Data from these stations are not used in the parameter estimation process. Instead they are used to evaluate alternative model calibrations, each specified either by one parameter vector for the whole domain or a number of provincial parameter vectors which together cover the domain. These independent validation data are vital for assessing the generality of the calibrated model and thereby its value for prediction.

The division of stations between the calibration and validation sets is done in such a way that the data in each set are statistically similar, constituting independent samples from the same population. In order that calibrations involving multiple parameter vectors can be tested, it is important that each set provides similar coverage of any given geographic region. We expect similar results whichever set is chosen as the calibration set. This assumed robustness to sampling error is supported by the results of Hemmings et al. (2003). They showed that compatible parameter estimates could be obtained by fitting a model to 3 independent basin-wide calibration sets, provided each set was similarly distributed over the range of different environmental conditions present.

The split-domain calibration method is applied to a given model with the aim of finding the number and geographic scope of parameter vectors which allow the lowest possible cost of the calibrated model, with respect to the stations in the validation set, to be obtained. This validation cost can be seen as an objective 'measure of merit' for the model. Because it reflects the model's behaviour with optimal parameter vectors, rather than with arbitrary parameters or parameters based on calibrations for arbitrary provinces, it can be used for comparative evaluation of different models. The use of independent validation data means that models of different complexity and different numbers of free parameters can be directly compared.

The algorithm for finding the optimal calibration is to first seek the best single-parameter vector calibration for the domain and then investigate whether a better calibration can be obtained by splitting the domain into two geographic provinces. If one or more ways of splitting the domain are found which lead to improved calibrations, then the best domain division is accepted and the algorithm is recursively applied to each province until either there is no decrease in the validation cost or no practical geographic split is possible.

It is not assumed that the best single-parameter vector calibration for a domain is the one that uses all the calibration stations. There may be good reason to exclude atypical stations to prevent them adversely affecting the calibration result. Such atypical stations might exist due to real but local effects, poor forcing data or differences in the type or number of observations available. Instead a novel method of aggregating stations into 'natural' calibration groups is employed. This allows groups comprising different numbers of stations to be identified, each of which is the group of a particular size best satisfied by a single parameter vector. No account is taken of the station positions prior to aggregation, but the geographic distribution of stations in each of the emerging groups is examined. Large groups with good coverage of the domain provide alternative calibration groups for the undivided domain. In addition, any of the smaller groups in which stations are geographically clustered indicates a natural province within the domain, on the basis of which a trial split-domain calibration can be performed.

The calibration algorithm is implemented by a hierarchy of procedures which are described in the following sub-sections, starting from the lowest level. The most fundamental of these, the 'parameter optimization procedure' is invoked by the 'station aggregation procedure' which is in turn invoked by two higher-level procedures to identify optimal provinces and calibration groups. Finally Section 2.5 describes important modifications to the basic method which allow observation error to be taken into account. For reference, a list of standard terms and symbols used in the description is given in Table 1.

2.2 Parameter optimization

The parameter optimization procedure searches for parameter vectors which minimize a cost based on the model misfit to observations. The variation of the misfit cost in parameter space is defined by a cost function $J(\vec{p})$, which is some function of the misfit between the output of the model with parameter vector $\vec{p}$ and observations at one or more stations. The actual cost function used is application dependent but is subject to the requirement, imposed by the station aggregation procedure, that it can be evaluated at each station independently. The aggregation procedure allows for variability in cost be-

Table 1
Standard notation for the split-domain calibration method

Item	Symbol or function	Equation
Domains		
Primary province	A	
Complementary province	B	
Potential primary province	$Apot$	
Potential complementary province	$Bpot$	
Station sets		
Calibration set	$\boldsymbol{D}$	
Validation set	$\boldsymbol{V}$	
Group of size n	$\boldsymbol{H}^n$	
True optimal group of size n	$\boldsymbol{H}^n_{\mathrm{OPT}}$	
Size n group best satisfied by parameter vector $\vec{p}$	$\boldsymbol{H}^n_{\mathrm{BEST}}(\vec{p})$	
Size n aggregation group	$\boldsymbol{G}^n$	
Calibration group	$\boldsymbol{C}$	
Parameter vectors		
Locally optimal parameter vectors found for group $\boldsymbol{H}$	$\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{H})$	
Optimal parameter vector found for group $\boldsymbol{H}$	$\vec{p}_{\mathrm{BEST}}(\boldsymbol{H})$	
Parameter vector search set	$\boldsymbol{P}$	
Costs		
Misfit for parameter vector $\vec{p}$ to observation from set $\boldsymbol{X}$	$M(\vec{p}, \boldsymbol{X})$	4
Cost for parameter vector $\vec{p}$ and group $\boldsymbol{H}$	$J(\vec{p}, \boldsymbol{H})$	
Cost function minimum found for group $\boldsymbol{H}$	$J_{\mathrm{BEST}}(\boldsymbol{H})$	
Baseline cost for station s	$J_{\mathrm{BEST}}(s)$	
Cost deviation for parameter vector $\vec{p}$ at station s (no observation error)	$\Delta J(\vec{p}, s)$	1
Group max. cost deviation for vector $\vec{p}$ over group $\boldsymbol{H}$ (no observation error)	$\Delta J_{\mathrm{MAX}}(\vec{p}, \boldsymbol{H})$	2
Aggregation penalty for group $\boldsymbol{H}$ (no observation error)	$\Delta J_{\mathrm{MAX}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}), \boldsymbol{H}\}$	2
Split-domain validation cost for provincial parameter vectors $\vec{p}_A$ and $\vec{p}_B$	$J^{\mathrm{SPLIT}}(\vec{p}_A, \boldsymbol{V}_A, \vec{p}_B, \boldsymbol{V}_B)$	
Misfit probability distribution	$\boldsymbol{M}(\vec{p}, \boldsymbol{X})$	6
Cost probability distribution	$\boldsymbol{J}(\vec{p}, \boldsymbol{H})$	
Cost probability distribution estimate	$\hat{\boldsymbol{J}}(\vec{p}, \boldsymbol{H})$	
Baseline cost distribution estimate for station s	$\hat{\boldsymbol{J}}(\vec{p}_{\mathrm{BEST}}(s), s)$	
Cost deviation (observation error present)	$U(\vec{p}, \boldsymbol{H})$	
Group maximum cost deviation (observation error present)	$U_{\mathrm{MAX}}(\vec{p}, \boldsymbol{H})$	7
Aggregation penalty (observation error present)	$U_{\mathrm{MAX}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}), \boldsymbol{H}\}$	7
Split-domain validation cost distribution estimate	$\hat{\boldsymbol{J}}^{\mathrm{SPLIT}}(\vec{p}_A, \boldsymbol{V}_A, \vec{p}_B, \boldsymbol{V}_B)$	

tween stations arising from factors other than the suitability of the parameter vector. However, cost function design choices which increase this variability unnecessarily should be avoided. In particular, a cost function based on a mean observation misfit is preferable to one based on a total misfit because of its reduced sensitivity to differences in the number of observations at each sta-

tion. In other work, the cost function has sometimes included a penalty term based on parameter deviations from their *a priori* estimated values (Fasham and Evans, 1995; Matear, 1995; Schartau et al., 2001). Such additional terms should ideally be made independent of model design so that different models can be compared.

Given a group of one or more stations $\boldsymbol{H}$, the parameter optimization procedure explores the group's cost function $J(\vec{p}, \boldsymbol{H})$ and provides an estimate $J_{\mathrm{BEST}}(\boldsymbol{H})$ of its minimum value. Cost function minima are located in parameter space using an optimizing routine. Powell's conjugate direction set method (Press et al., 1992) was used in the present study to search a finite parameter space, the bounds for each parameter being prescribed in the model definition. The cost function minimum found by a single application of the optimizer is dependent on its starting point in parameter space and may only be a local minimum. The global minimum can be estimated with some degree of confidence, albeit difficult to quantify, by running an ensemble of optimizations with different initial parameter vectors and selecting the smallest of all the minima found. The positions in parameter space of all the minima found define a set of locally optimal parameter vectors $\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{H})$. The 'best' parameter vector $\vec{p}_{\mathrm{BEST}}(\boldsymbol{H})$ is defined as the parameter vector in $\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{H})$ associated with the lowest minimum $J_{\mathrm{BEST}}(\boldsymbol{H})$ such that $J_{\mathrm{BEST}}(\boldsymbol{H}) = J(\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}), \boldsymbol{H})$. The ensemble approach has been used, in combination with the variational adjoint optimization method, by Friedrichs (2002) and by Schartau et al. (2001).

In the present study, the initial parameter vectors were drawn from a prior, joint normal probability distribution, scaled asymmetrically about *a priori* expected values such that displacements of 3 standard deviations correspond to the prescribed bounds. Parameter covariances in the prior distribution are zero. Choice of ensemble size is necessarily a compromise between coverage of the parameter space and computational load. A size of 100 was chosen here, compared with the 50 initial points used by Friedrichs (2002) and 600 initial points used by Schartau et al. (2001). Efficient coverage of the multivariate space was achieved using Latin hypercube sampling (McKay et al., 1979). This technique involves forming a grid which divides the prior distribution along each dimension into a number of intervals of equal probability, the number being equal to the sample size. The sample is then drawn randomly from the grid boxes in such a way that each interval in each dimension is sampled once only, ensuring that all intervals in all dimensions are represented. In the absence of a parameter penalty term in the cost function, the parameter bounds were applied by use of a mapping function which distorts the infinite space seen by the optimizer, such that displacements made when approaching the bounds translate to infinitesimally small steps in the actual parameter space. The mapping function employed takes the same form as the parameter penalty term of Fasham and Evans (1995).

In a broad sense, the station aggregation procedure might be classified as a form of cluster analysis. Its role is to identify a series of station groups, comprising different numbers of stations from the calibration set, each of which is the group of a particular size best satisfied by its optimal parameter vector. Combining stations for parameter optimization purposes improves the generality of the model and, potentially, its ability to describe independent data. However, the improvement normally comes at the expense of degrading the fit to the calibration data. This is because cost function minima for individual stations in a calibration group are unlikely to be coincident in the parameter space and, if they are not, the group's cost function minimum will be greater than the minima for each of the individual stations. This will be true for any sensible cost function formula by which the misfits for different stations are combined. The principle behind the station aggregation procedure is to group stations in such a way as to keep a carefully chosen measure of the cost penalty incurred as small as possible. We refer to this measure as the group's aggregation penalty.

The group of size n best satisfied by any given parameter vector $\vec{p}$ must be formed from the n stations which are individually best satisfied. Stations must therefore be compared on the basis of how well their data are satisfied by $\vec{p}$. The degree to which a station s is satisfied by the parameter vector $\vec{p}$ is quantified by the cost deviation

$$\Delta J(\vec{p}, s) = J(\vec{p}, s) - J_{\mathrm{BEST}}(s), \tag{1}$$

which is the difference between the misfit cost for the model with parameter vector $\vec{p}$ at station s and a baseline cost determined by optimizing the model for station s only. The cost deviation from the baseline is zero for $\vec{p} = \vec{p}_{\mathrm{BEST}}(s)$ and increases as $\vec{p}$ becomes less suitable for that station.

The aggregation penalty for a group of n stations $\boldsymbol{H}^n = \{s_1, ..., s_n\}$ is defined as the group maximum cost deviation for the group's optimal parameter vector $\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n)$, given by

$$\Delta J_{\mathrm{MAX}} \left\{ \vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n), \boldsymbol{H}^n \right\} = \max_{i=1}^{n} \left\{ \Delta J(\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n), s_i) \right\}. \tag{2}$$

For a given parameter vector $\vec{p}$, the group maximum cost deviation $\Delta J_{\mathrm{MAX}}(\vec{p}, \boldsymbol{H}^n)$ is minimized for a calibration set $\boldsymbol{D}$ simply by selecting the n stations from $\boldsymbol{D}$ with the lowest individual cost deviations. This forms the optimal group for parameter vector $\vec{p}$, denoted $\boldsymbol{H}^n_{\mathrm{BEST}}(\vec{p})$. Finding the group of size n which minimizes the aggregation penalty, denoted $\boldsymbol{H}^n_{\mathrm{OPT}}$, is less straightforward be-

cause the group's optimal parameter vector $\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n_{\mathrm{OPT}})$ is not known in advance.

Ideally the parameter optimization procedure would be applied to every possible group of the required size to allow direct comparison of the aggregation penalties for all groups. However, this approach quickly becomes infeasible as the number of stations in the calibration set increases. To select the best group of size n from a calibration set of size N, the number of parameter optimizations required would be

$$ {}_N C_n = \frac{N!}{n!(N-n)!}. \tag{3}$$

Instead, the group selected by the aggregation procedure is that which gives the minimum value of the group maximum cost deviation over a finite set of promising parameter vectors, referred to as the search set $\boldsymbol{P}$. The set is assumed to contain at least one parameter vector $\vec{p}$ for which the optimum group's group maximum cost deviation is close to its aggregation penalty. i.e. $\Delta J_{\mathrm{MAX}}(\vec{p}, \boldsymbol{H}^n_{\mathrm{OPT}}) \approx \Delta J_{\mathrm{MAX}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n_{\mathrm{OPT}}), \boldsymbol{H}^n_{\mathrm{OPT}}\}$. The selected group is referred to as the size n aggregation group.

Station aggregation is performed by the following stepwise procedure, in which the parameter vector search set for each step is obtained by optimizing for a smaller group of stations already aggregated. The procedure identifies a series of aggregation groups $\boldsymbol{G}^2, ..., \boldsymbol{G}^{N-1}$.

$\boldsymbol{P} = \boldsymbol{P}_{\mathrm{init}}$, $n = 1$
While n less than size of $\boldsymbol{D}$
    For each $\vec{p} \in \boldsymbol{P}$
        For each $s \in \boldsymbol{D}$
            Evaluate $\Delta J(\vec{p}, s)$
        Form $\boldsymbol{H}^{n+1}_{\mathrm{BEST}}(\vec{p})$
        Evaluate $\Delta J_{\mathrm{MAX}}\{\vec{p}, \boldsymbol{H}^{n+1}_{\mathrm{BEST}}(\vec{p})\}$
    $\boldsymbol{G}^{n+1} = \boldsymbol{H}^{n+1}_{\mathrm{BEST}}$ with lowest ΔJ_{MAX}
    $n = n + 1$
    $\boldsymbol{P} = \boldsymbol{P}_{\mathrm{good}}(\boldsymbol{G}^n)$

At the nth step, a new aggregation group $\boldsymbol{G}^{n+1}$ is identified by selecting, from all possible groups of size $n + 1$ in the calibration set $\boldsymbol{D}$, the group with the lowest group maximum cost deviation for a parameter vector in the search set $\boldsymbol{P}$. The new group $\boldsymbol{G}^{n+1}$ has one more station than the previous aggregation group $\boldsymbol{G}^n$. It does not have to include all of the stations in $\boldsymbol{G}^n$. Such a restriction is unnecessary and could cause the selection of a non-optimal group. The model is then optimized for the new group to obtain a new set of parameter vectors and the process is repeated until all but one of the stations

in the calibration set are included.

The behaviour of the station aggregation procedure at the nth step is illustrated in Fig. 1 for a simple case where the model has only one parameter. The figure shows how the procedure chooses between two groups $\boldsymbol{H}_1^{n+1}$ and $\boldsymbol{H}_2^{n+1}$. The group which we aim to select is that with the lowest aggregation penalty, defined with reference to the group's cost function minimum. Comparison of the aggregation penalties shows that $\boldsymbol{H}_1^{n+1}$ is the true optimal group. However, in practice the aggregation penalties are unknown because minimization of the cost function for all possible groups is too expensive. In the absence of information regarding the location of each group's cost function minimum, the group maximum cost deviation, designed for selecting between groups, has the additional function of acting as a proxy for the cost function in parameter space. While its variation in parameter space is broadly similar to that of the cost function, there are differences because it is sensitive to different model output: the group maximum cost deviation is sensitive to output at the station least favoured for inclusion in the group, whereas the cost function might be particularly sensitive to output at another station, such as the station with the most reliable observations. The implications of this are discussed in Appendix A.

The main limitation is that it is only possible to evaluate the group maximum cost deviation at a relatively small number of points in the parameter space. The set of sample points in parameter space, the search set $\boldsymbol{P}$, is determined with reference to the cost function for the optimal group of size n, $\boldsymbol{G}^n$. The underlying assumption is that the set $\boldsymbol{P}_{\text{good}}(\boldsymbol{G}^n)$ is, by virtue of its pre-conditioning, sufficiently representative of the most promising areas of parameter space for a similar but slightly larger group to allow the best such group to be selected. For the example in Fig. 1, there are three parameter values in $\boldsymbol{P}$ within the region of interest, p_1, p_2 and p_3. The lowest point sampled on either of the group maximum cost deviation curves lies on that for group $\boldsymbol{H}_1^{n+1}$ at p_2. The selection is therefore successful. Examples are given in Appendix A where this is not the case, to illustrate the limitations of the method.

To start the procedure, rather than choosing an arbitrary station s and restricting the initial set of parameter vectors $\boldsymbol{P}_{\text{init}}$ to $\boldsymbol{P}_{\text{good}}(s)$, we aim for a more robust result by pooling the sets of parameter vectors obtained by optimizing for all stations individually. This increases the initial number of trial parameter vectors by a factor equal to the number of stations in the calibration set.

The effectiveness of the aggregation penalty as a criteria for evaluating station groups relies on the assumption that variation of the cost deviation between stations for a given parameter vector is dominated by variation in the suitabil-

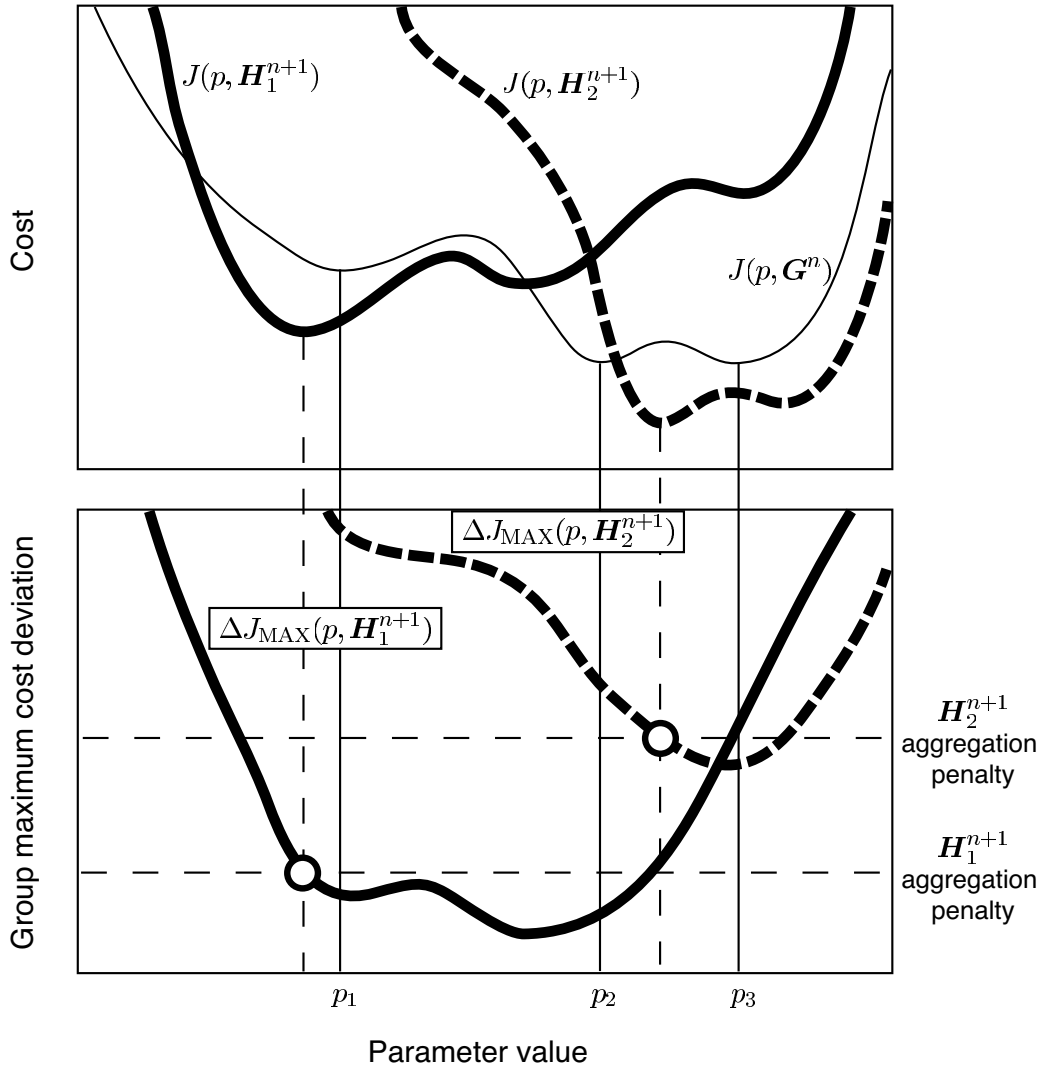

Fig. 1. Behaviour of the station aggregation procedure at step n for a single parameter model. The group with the lower aggregation penalty $\boldsymbol{H}_1^{n+1}$ is successfully identified as a result of sampling the cost deviation curves for the two groups at p_1, p_2 and p_3.

ity of that parameter vector. However, other potential sources of variation in cost exist. The limitations of the model structure and/or the functional forms it uses to describe biogeochemical processes may make the model itself less compatible with observations at one station than another, irrespective of how optimal its parameter vector is. Likewise, the quality of the external forcing data may vary between stations. Also, because the model is imperfect, we should expect a tendency for the mean misfit to increase with the number and type of independent observations, so the availability of observations at each station acts as another source of variation. Using the deviation from the station's baseline cost, in place of the absolute cost, removes much of the variance associated with these factors but some residual variance may remain. A more robust measure of the cost deviation, which takes into account observation

error, is introduced in Section 2.5.

2.4 Identification of provinces and calibration groups

The calibration algorithm is implemented by two fundamental procedures, both of which invoke the station aggregation procedure. The first of these is the 'whole-domain calibration procedure', which seeks the optimum group of stations for a single-parameter vector calibration of a given geographic domain. The second is the 'split-domain calibration procedure', which seeks optimal station groups for a two-parameter vector calibration of the domain. Any aggregation group which shows good geographic coverage of some province within the domain can serve as a starting point for this procedure. Such groups are referred to as province indicator groups. Each aggregation group of suitable size identified during the application of the whole-domain calibration procedure to the domain is a potential candidate. These candidate groups are assessed with regard to the geographic distribution of their member stations, ignoring any with less than 3 stations or any which leave less than 3 stations remaining in the domain's calibration set. A single province can include multiple regions of the domain, each represented by at least 3 stations, allowing for the possibility that geographically separated regions might share similar ecological responses to environmental forcing.

For any given domain, one application of the whole-domain calibration procedure is required, followed by zero or more applications of the split-domain calibration procedure, depending on the number of province indicator groups found. The split-domain calibrations are evaluated against each other and against the whole-domain calibration by comparing the costs of the calibrated model with respect to the domain's validation data. If the best calibration is a split-domain calibration then the associated provinces define two new domains to which the calibration algorithm is applied recursively.

2.4.1 Whole-domain calibration procedure

In the whole-domain calibration procedure, a trial optimization is first done using all stations in the domain's calibration set $\boldsymbol{D}$. The best parameter vector thus obtained $\vec{p}_{\mathrm{BEST}}(\boldsymbol{D})$ is evaluated by running the model at all stations in the validation set $\boldsymbol{V}$ to obtain the validation cost $J\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{D}), \boldsymbol{V}\}$. The station aggregation procedure is then applied to the stations in $\boldsymbol{D}$, giving a series of alternative calibration groups $\boldsymbol{G}^i$, of increasing size i, The validation cost $J\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{G}^i), \boldsymbol{V}\}$ is determined for each of these groups. All of the validation costs are compared and the group with the lowest cost becomes the domain's final calibration group $\boldsymbol{C}$. This is either the full calibration set ($\boldsymbol{C} = \boldsymbol{D}$) or one

of the aggregation groups ($\boldsymbol{C} = \boldsymbol{G}^i$ for some i). The validation cost for the whole-domain calibration is $J\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}), \boldsymbol{V}\}$. The set of province indicator groups is formed from all suitable aggregation groups.

2.4.2 Split-domain calibration procedure

The split-domain calibration procedure splits the domain into two provinces using, as a starting point, a specific province indicator group $\boldsymbol{G}_A$ ($\boldsymbol{G}_A = \boldsymbol{G}^i$ for some i). The associated province is referred to as the primary province A. The remainder of the domain is referred to as the complementary province B. Any single stations or pairs of stations which are geographically isolated by the province indicator group $\boldsymbol{G}_A$ cannot be considered indicative of a separate region. They are therefore considered to be atypical stations in province A rather than province B stations.

Geographic borders are established between the two provinces, allowing the validation set $\boldsymbol{V}$ to be split into separate sub-sets $\boldsymbol{V}_A$ and $\boldsymbol{V}_B$ for each province. The full calibration set $\boldsymbol{D}$ is likewise split into sub-sets $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$. In the present study, lines of latitude were used where possible and, where calibration stations at the same latitude fell in different provinces, intervening validation stations were assigned such that there was at least one validation station at the same latitude in each province if possible.

Calibration groups $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$ are identified by applying the whole-domain calibration procedure to each of the provincial calibration sets $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$. For province B, this involves another application of the station aggregation procedure. For province A, the previous aggregation results can be used. However, if there are stations in the calibration set $\boldsymbol{D}_A$ which are not in the group $\boldsymbol{G}_A$, the aggregation procedure must be repeated, starting from $\boldsymbol{G}_A$, to check whether the series of aggregation groups can be extended within province A. This may now be possible as a consequence of the exclusion of stations outside A from the calibration set. The validation cost for the split-domain calibration

$$J^{\mathrm{SPLIT}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_A), \boldsymbol{V}_A, \vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_B), \boldsymbol{V}_B\}$$

is determined by running the model with the best parameter vector from the appropriate calibration at each validation station in $\boldsymbol{V}$. That is $\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_A)$ in $\boldsymbol{V}_A$ and $\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_B)$ in $\boldsymbol{V}_B$.

Estimates of observation error provide valuable information for any data assimilation scheme, their principal use being to weight the model misfit to individual observations such that higher precision observations have a greater influence on the model. Taking observation error into account also allows statistical significance to be associated with differences in the misfit cost between different model integrations. This allows the cost deviation for a station to be redefined in terms of a statistic for significance testing. It also allows the significance of differences in validation cost to be taken into account when comparing alternative model calibrations for a domain.

2.5.1 Model misfit

The misfit is defined here in terms of the squared deviation of the model from the observation, weighted according to the observation variance. While this is a useful definition which is consistent with previous work, other definitions of model misfit could be used. The misfit of the model with parameter vector $\vec{p}$ with respect to the ith observation from the set $\boldsymbol{X}$ at the jth station is given by

$$M_{ij}(\vec{p}, \boldsymbol{X}) = \frac{\{x_{ij\mathrm{MODEL}}(\vec{p}) - \overline{x_{ij\mathrm{OBS}}}\}^2}{\mathrm{var}(x_{ij\mathrm{OBS}})} \tag{4}$$

where the subscripts 'MODEL' and 'OBS' denote model predictions and observed values of x ($x \in \boldsymbol{X}$) respectively. Each observation $\overline{x_{ij\mathrm{OBS}}}$ is the expected value of $\boldsymbol{X}$, for a particular time and location, derived from observational data. This is the estimated mean of a population of possible values affected by both measurement and sampling error. The population variance quantifies the observation error due to the combined effect of these sources. Division by an estimate of the variance $\mathrm{var}(x_{ij\mathrm{OBS}})$ non-dimensionalizes the misfit so that a misfit of unity or less implies a model deviation not exceeding the observation error.

2.5.2 Cost function probability distribution

The cost function used in the parameter optimization procedure is a function of the model misfit to the expected observation values M_{ij}. (A specific example is given in Section 3.2). It returns a single value for a particular parameter vector and station group. To allow for observation error in cost comparisons, it is necessary to replace this value with an estimate of the cost function's probability distribution, as a function of the probability distribution of the

observations. Observational uncertainty is represented by a joint probability distribution in a multi-dimensional space having one dimension for each available observation. For any group of stations, the cost distribution is defined by the application of the cost function to the appropriate lower-dimensional subset of this observation distribution.

To derive an estimate $\hat{\boldsymbol{J}}$ of a particular cost distribution $\boldsymbol{J}$, the cost function is applied to a sample of observation vectors drawn from a probability distribution $\boldsymbol{\Omega}$: a model of the observation probability distribution. Its multivariate mean is the vector of expected observation values $\overline{x_{ij\mathrm{OBS}}}$, while its variance structure is chosen to be consistent with the estimated observation error. The observation error at each time and location is assumed to have a normal distribution with zero mean and variance equal to the estimate $\mathrm{var}(x_{ij\mathrm{OBS}})$. Error covariances between stations are assumed to be zero. This is implicit in the method because it must be possible to evaluate costs at stations individually. In the present work, the errors are also assumed to be temporally independent so that all error covariances are zero. The univariate observation distribution for the ith observation from the data set $\boldsymbol{X}$ at the jth station is therefore given by

$$\boldsymbol{\Omega}_{ij}(\boldsymbol{X}) = \overline{x_{ij\mathrm{OBS}}} + \boldsymbol{E}\sqrt{\mathrm{var}(x_{ij\mathrm{OBS}})} \tag{5}$$

where $\boldsymbol{E}$ is the normal distribution with zero mean and unit variance. $\boldsymbol{\Omega}_{ij}$ is substituted for the expected observation value in the misfit expression from Eq. (4) to give a misfit probability distribution

$$\boldsymbol{M}_{ij}(\vec{p}, \boldsymbol{X}) = \frac{\{x_{ij\mathrm{MODEL}}(\vec{p}) - \boldsymbol{\Omega}_{ij}(\boldsymbol{X})\}^2}{\mathrm{var}(x_{ij\mathrm{OBS}})} \tag{6}$$

This is substituted for the misfit M_{ij} in the cost function to define the cost probability distribution $\boldsymbol{J}$.

In the present study, the sample of observation vectors from $\boldsymbol{\Omega}$ for the required cost distribution estimates was drawn using Latin hypercube sampling, with a sample size of 100. In general, cost distributions are not well approximated by a normal distribution and cost distribution estimates should therefore be compared using a non-parametric test. The robust rank-order test (Siegel and Castellan, 1988) was used here. The test provides a statistic from which the statistical significance of the differences between the distribution estimates is deduced. The lowest of two cost distributions is taken to be that with the lowest median. The median is used in preference to the mean because of its reduced sensitivity to the shape of the probability distribution.

2.5.3 Modification of the station aggregation procedure

The new cost deviation for a parameter vector $\vec{p}$ at station s, taking into account observation error, is the test statistic $U(\vec{p}, s)$ derived by comparing the cost distribution estimate $\hat{\boldsymbol{J}}(\vec{p}, s)$ with the station's baseline cost distribution estimate $\hat{\boldsymbol{J}}(\vec{p}_{\mathrm{BEST}}(s), s)$. Substituting for the cost deviation ΔJ in Eq. 2 gives the new aggregation penalty for a group $\boldsymbol{H}^n$

$$U_{\mathrm{MAX}}\left\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n), \boldsymbol{H}^n\right\} = \max_{i=1}^{n}\left\{U(\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}^n), s_i)\right\}. \tag{7}$$

The station aggregation procedure, defined in Section 2.3, is modified by replacing the old cost deviation $\Delta J(\vec{p}, s)$ with $U(\vec{p}, s)$ and the old group maximum cost deviation $\Delta J_{\mathrm{MAX}}\{\vec{p}, \boldsymbol{H}^{n+1}_{\mathrm{BEST}}(\vec{p})\}$ with $U_{\mathrm{MAX}}\{\vec{p}, \boldsymbol{H}^{n+1}_{\mathrm{BEST}}(\vec{p})\}$. Unfortunately, a complication arises because infinite values of the cost deviation $U(\vec{p}, s)$ occur if there is no overlap between the two cost distribution estimates. In that case, the parameter vector $\vec{p}$ has no measurable merit with respect to the station s data and the station is considered to be 'not satisfied' by the parameter vector. The existence of such stations introduces an additional termination condition for the station aggregation procedure. It may now terminate, at any step n, if no group of the required size $n+1$ is satisfied by any of the available parameter vectors in the current search set $\boldsymbol{P}$. That is if, for all $\vec{p} \in \boldsymbol{P}$, there are less than $n+1$ stations s ($s \in \boldsymbol{D}$) with finite $U(\vec{p}, s)$, so that the group $\boldsymbol{H}^{n+1}_{\mathrm{BEST}}(\vec{p})$ cannot be formed. Early termination of the station aggregation procedure saves on computation time and could be introduced with the original cost deviation definition for this purpose, by setting an upper limit to ΔJ. However, it does introduce a complication which must be allowed for in the split-domain calibration procedure as described in the next section.

2.5.4 Modification of the calibration algorithm

In the calibration algorithm, shown in its final form in Fig. 2, comparisons between alternative model calibrations are now made with reference to the median values of their validation cost distribution estimates and the split-domain calibration procedure is modified to allow for provinces with incomplete station aggregation.

Given a set of calibration stations $\boldsymbol{D}$ and a corresponding set of validation stations $\boldsymbol{V}$, a whole-domain calibration is first carried out to obtain a calibration group $\boldsymbol{C}$ with validation cost distribution estimate $\hat{\boldsymbol{J}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}), \boldsymbol{V}\}$. In the new whole-domain calibration procedure, the group $\boldsymbol{G}^i$ with the lowest median validation cost $\hat{\boldsymbol{J}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{G}^i), \boldsymbol{V}\}$ is favoured over the full calibration set $\boldsymbol{D}$ only if its validation cost is significantly lower than $\hat{\boldsymbol{J}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{D}), \boldsymbol{V}\}$, at a probability of 95%. Split-domain calibrations are then performed for each

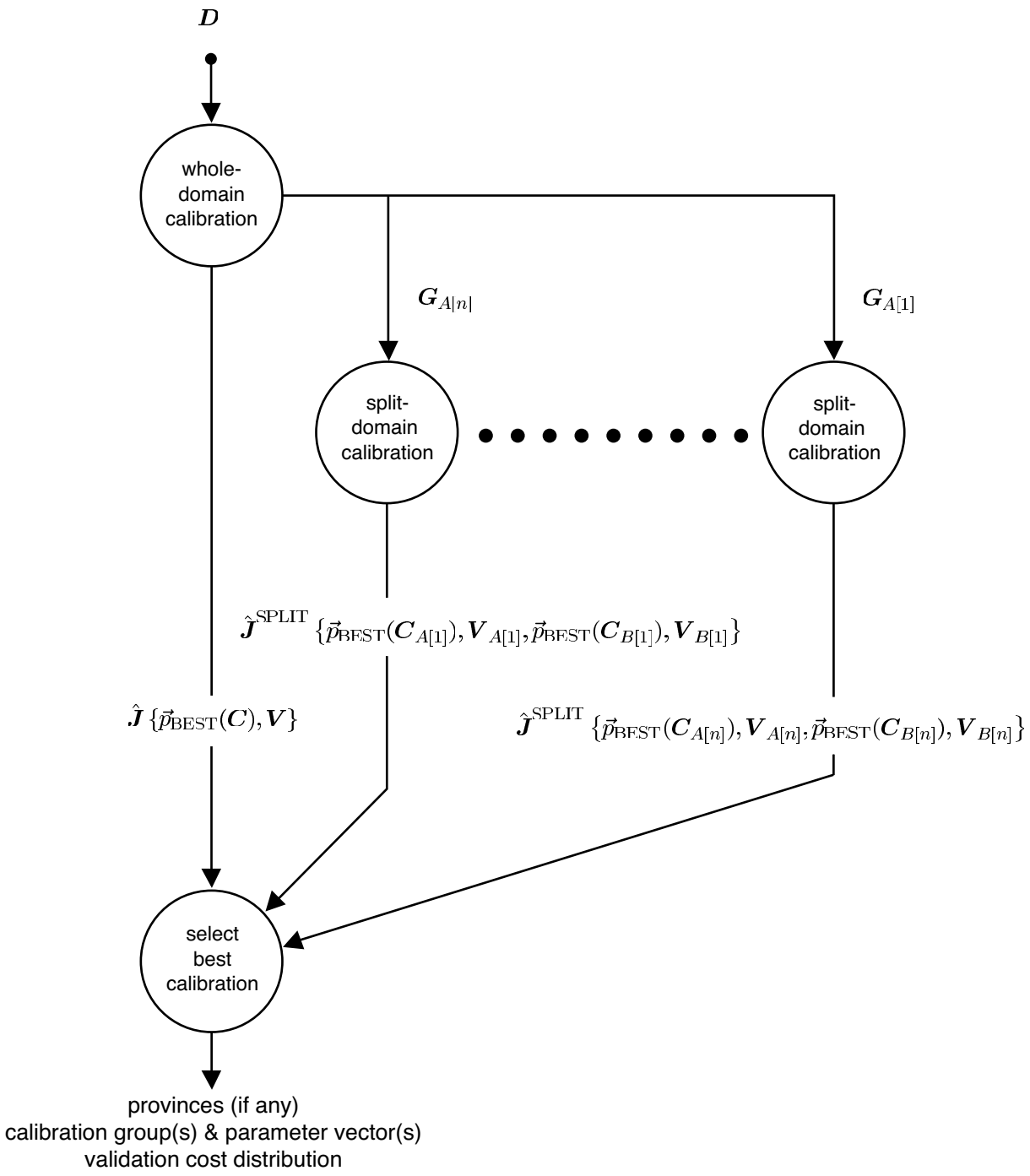

Fig. 2. The calibration algorithm used to identify the optimum calibration for a given domain having the set of calibration stations $\boldsymbol{D}$. (See text for details.)

province indicator group identified. The jth such group is denoted $\boldsymbol{G}_{A[j]}$ and the calibration groups obtained are $\boldsymbol{C}_{A[j]}$ and $\boldsymbol{C}_{B[j]}$. The validation sets for each province in the jth split-domain calibration are $\boldsymbol{V}_{A[j]}$ and $\boldsymbol{V}_{B[j]}$. The

split-domain cost distribution estimate

$$\hat{\boldsymbol{J}}^{\mathrm{SPLIT}}\left\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_{A[j]}), \boldsymbol{V}_{A[j]}, \vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_{B[j]}), \boldsymbol{V}_{B[j]}\right\}$$

with the lowest median is identified and compared with the whole-domain cost distribution estimate. If it is significantly lower, at a probability of 95%, then the split-domain calibration is accepted. Otherwise the whole-domain calibration is accepted.

The requirement for modifying the split-domain calibration procedure arises because, for symmetry, the geographic border or borders between the two provinces A and B should be drawn between two aggregation groups. We cannot now assume that all stations in the complementary province B will aggregate. If they do not, there may be a gap between the aggregation groups and an adjustment may need to be made to the border so that it bisects this unrepresented area or otherwise divides it in a sensible way.

The modified split-domain calibration procedure is shown in Figure 3. Prior to any border adjustment, the provinces are now referred to as the potential primary province *Apot* and the potential complementary province *Bpot*. The station aggregation procedure is applied to the set of calibration stations within *Bpot*, denoted $\boldsymbol{D}_{Bpot}$. Unless forced to terminate early, the aggregation procedure normally terminates when the group $\boldsymbol{G}^{N-1}$ is found, where N is the size of the calibration set. However, in this case, there is a requirement to know whether the full calibration set $\boldsymbol{D}_{Bpot}$ is also an aggregation group. The procedure is therefore extended to test whether the final station would aggregate. That is, whether it is satisfied by a parameter vector in the search set $\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{G}^{N-1})$. The largest aggregation group obtained $\boldsymbol{G}_B$ is used, together with $\boldsymbol{G}_A$, to define the border or borders between the final provinces A and B. The whole-domain calibration procedure is then applied to A and B as before. For province B, this now only involves determining the validation costs. For province A, extension of the series of aggregation groups beyond $\boldsymbol{G}_A$ may also be required.

3 Testing the method

The split-domain calibration method described in the previous section was applied to evaluate a simple ecosystem model, constructed with sufficient realism to demonstrate the method's practical utility. The present study is restricted to the North Atlantic but, to cover a wide range of physical and nutrient regimes, the chosen stations are widely distributed over the basin on a 5 degree grid, from 27.5°N to 67.5°N. Stations on this grid were screened to remove

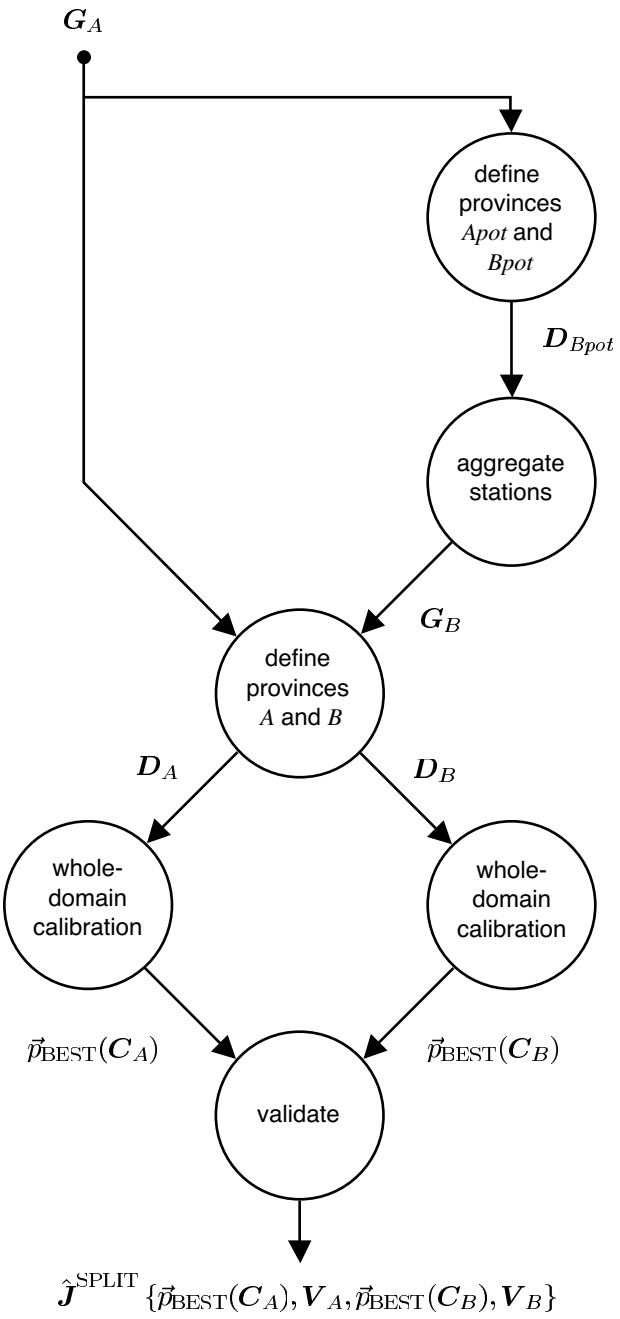

Fig. 3. The procedure for obtaining the optimum split-domain calibration based on a given province indicator group $\boldsymbol{G}_A$. (See text for details.)

any at which there were obvious problems with the model's external forcing data. Only about half were used. These 30 remaining stations are divided into calibration and validation sets of 15 stations each.

The normal application of the calibration method involves continuing the station aggregation procedure at step 2 with the most promising pair of stations identified in step 1. However, other pairs can have group maximum cost deviation values which are only slightly higher. The actual pair of stations selected

depends on the finite sample of parameter vectors available from the individual station optimizations and alternative samples could result in the selection of different pairs. To test the robustness of the results to the selection of the initial pair, additional aggregation experiments were performed with alternative pairs.

3.1 Observations

The observed variables used in this test are from two data sets: $\log \boldsymbol{\mathcal{C}}$, where $\boldsymbol{\mathcal{C}}$ is the set of chlorophyll *a* concentrations, and the set of annual nitrate concentration maxima $\boldsymbol{\mathcal{N}}$. Although the surface layer chlorophyll data derived from satellite ocean colour are invaluable because of their good spatial and temporal coverage, they provide information relating to one ecosystem component only and this is recognized as a major limitation. Climatological estimates of the annual maximum nitrate concentration in the upper mixed layer therefore provided a useful additional constraint. Seasonal estimates of mixed layer nitrate concentration are available but were not used in the present study. This is partly because they are much less reliable than the annual maxima, being especially prone to sampling error associated with interannual variability, but also because their comparison with the output available from the candidate model is not straightforward. The latter problem is explained in Section 3.3. Observational nitrate estimates are the only relevant *in situ* data presently available for the whole basin. Other *in situ* data are only available at isolated locations and were not used.

The resolution of the data should reflect the application for which the model is to be evaluated. We envisage a hypothetical target application requiring field predictions down to time scales of the order of 1 day and space scales of 200 km. The observed mean and variance estimates of $\log \boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{N}}$ required for the cost function are therefore those describing data distributions over a 200 km length scale. This is considered to be sufficiently large to prevent major problems with aliasing due to mesoscale eddy activity. Any observation within 100 km of a station is treated as a possible realization of the true value at the station location. Chlorophyll data from a specific year were used, in preference to climatological values, to avoid the potentially serious loss of information which can result from combining multiple annual cycles in which phytoplankton blooms are out of phase.

All chlorophyll *a* estimates within 100 km of the nominal station positions were extracted from daily SeaWiFS 9 km Standard Mapped Image data for 1998, this being the first complete calendar year of data available. $\log \boldsymbol{\mathcal{C}}$ was calculated for each valid pixel and its mean and variance were estimated from the sample of all valid pixels within 100 km radius of the station position,

pixels being weighted by area to allow for meridional variation. The variance quantifies the observation error, which is a combination of the measurement error associated with the SeaWiFS Chlorophyll estimates (±35%) and sampling error associated with mesoscale variability. To avoid bias due to poor coverage, samples of less than 10 pixels and samples with a standard deviation in meridional or zonal position of less than 30 km were not used. A sample giving complete coverage has a standard deviation of 50 km. The log transformation is required to give a pseudo-normal distribution. This was tested by examining combined probability density functions for $\boldsymbol{\mathcal{C}}$ and $\log \boldsymbol{\mathcal{C}}$ for observations at all times and all stations. It is also supported by theoretical considerations and other empirical data (Campbell, 1995).

The nitrate maximum normally occurs in late winter as a result of deep winter-time mixing. Observed values of $\boldsymbol{\mathcal{N}}$ were estimated, following the method of Glover and Brewer (1988), by interpolating vertical nitrate profiles, extracted from World Ocean Atlas (WOA) 1° analyzed annual mean fields (Conkright et al., 1998), to the average depth of the mixed layer over the period February-April. Where this depth is greater than 500 m, the concentration at 500 m was used. The average mixed layer depth was estimated from averaging monthly data on a 1° grid. The mixed layer depth estimates of Levitus et al. (1982) based on a density difference criterion ($\Delta\sigma_t = 0.125$) were used as these are readily available. The processing differs slightly from that of Glover and Brewer (1988) who averaged winter-time hydrographic profiles (Levitus et al., 1982) over the 3 month period before calculating mixed layer depth using a variable σ_t criterion. Their criterion was equivalent to a 0.5°C temperature criterion in the absence of salinity stratification. Despite these differences in processing, the resulting winter-time mixed layer depth fields are very similar, with the exception of a few northern regions above 55°N where the use of the fixed σ_t criterion gives some exceptionally high values. The cut off at 500 m means that this has little effect on the nitrate maximum estimates. Nitrate values obtained where the winter-time mixed layer depth is less than 100 m are considered unreliable because of the likelihood of strong seasonal bias in the nitrate concentrations observed above this depth. These were therefore omitted.

The estimated annual nitrate maximum field is shown in Fig. 4, together with the calibration and validation stations. For ease of reference, stations are numbered according to their grid position in the format YYXX. They are assigned alternately to calibration and validation sets in numerical order. To obtain the estimates of the 200 km scale mean for the cost function, the winter-time nitrate field was averaged over 3° boxes centred on station locations. This was to reduce contamination by smaller scale features which, given the sparse nature of the original nitrate sampling and the sensitivity of the method to error in mixed layer depth, are more likely to be due to estimation errors than real spatial patterns. Any 3° means based on less than 5 values (56% coverage)

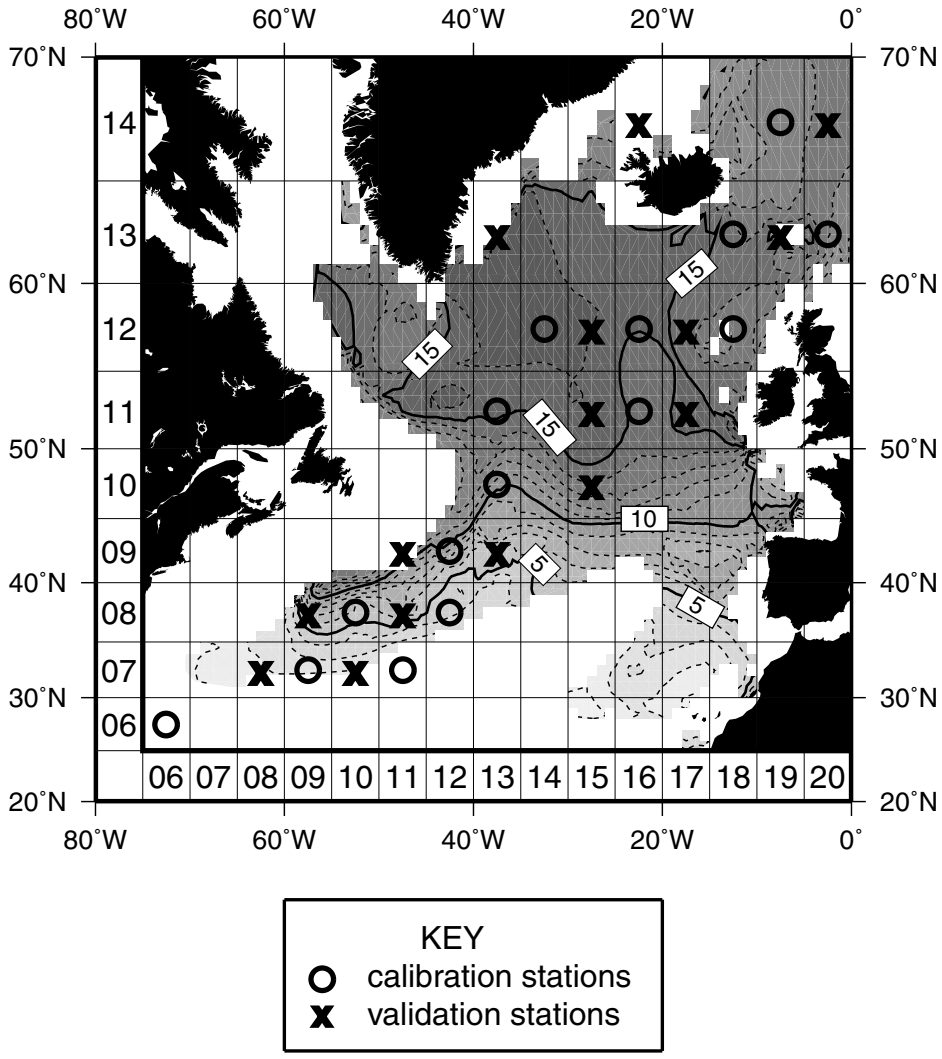

Fig. 4. Estimated annual maximum nitrate concentration in the mixed layer (mmol m^{-3}) and distribution of calibration and validation stations. Each station is identified by a four digit station number of the form YYXX, formed by concatenating the 2 digit meridional and zonal position numbers shown on the grid.

were discarded. In the absence of information about the variance of the nitrate maximum at the 200 km scale it's standard deviation was somewhat arbitrarily set to 1 mmol m^{-3} for all stations.

Five of the stations have no nitrate observation. These are 0606 (27.5°N 72.5°W), 0711 (32.5°N 47.5°W), 0911 (42.5°N 47.5°W), 1313 (62.5°N 37.5°W) and 1416 (67.5°N 22.5°W). Two of these stations are in the calibration set. Including these stations in the analysis was experimental, allowing the effects of missing data to be examined.

3.2 Cost function

The cost function returns the sum of the misfit costs for each of the observed variables

$$J = J_{\text{chl}} + J_{\text{nit}} \tag{8}$$

The chlorophyll misfit cost for parameter vector $\vec{p}$ and group $\boldsymbol{H}$ is the mean misfit for $\log \boldsymbol{\mathcal{C}}$ over all stations and observation times. So

$$J_{\text{chl}}(\vec{p}, \boldsymbol{H}) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n_j} M_{ij}(\vec{p}, \log \boldsymbol{\mathcal{C}})}{\sum_{j=1}^{n} n_j} \tag{9}$$

where n is the number of stations in $\boldsymbol{H}$ and n_j is the number of chlorophyll observations for the jth station, which varies largely due to cloud cover. The nitrate misfit cost is the mean misfit for $\boldsymbol{\mathcal{N}}$ over all stations

$$J_{\text{nit}}(\vec{p}, \boldsymbol{H}) = \frac{1}{n} \sum_{j=1}^{n} M_j(\vec{p}, \boldsymbol{\mathcal{N}}) \tag{10}$$

For the purposes of constructing the simulated observation probability distribution $\boldsymbol{\Omega}$, $\log \boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{N}}$ were treated as a single data set. Each observation vector in the sample drawn from $\boldsymbol{\Omega}$ for the cost distribution estimates comprises N nitrate values and $\sum_{j=1}^{N} n_j$ chlorophyll values, where N is the total number of stations in the calibration and validation sets. i.e. $N = 30$.

3.3 Model

The specific model used to test the method is of relatively minor importance as the aim of the study is to find an effective method for evaluating any candidate model. However, a certain level of realism is required to properly demonstrate the method's utility for the intended application. A phytoplankton-zooplankton-nutrient (PZN) model of nitrogen flow with 12 free parameters, based on the 7 compartment model of Fasham et al. (1990), was chosen largely for its simplicity. The model equations are given in Appendix B. It is run in a zero-dimensional context to describe the plankton ecosystem in the upper ocean mixed layer. This is in keeping with the previous work, excepting that of Prunet et al. (1996a,b) who used a 1-dimensional model. Keeping the spatial dimension of the model low is important because of the longer integration time for higher dimensional models, combined with the iterative nature of the optimization process which requires a large number of integrations.

The model's external forcing data includes spatially varying annual cycles of day length, photosynthetically available radiation (PAR), mixed layer depth and phytoplankton maximum growth rate modelled as a function of temperature. These are augmented by spatially varying annual mean nitrate profiles, which define the sub-surface nitrate at the base of the mixed layer. While there are no explicit horizontal fluxes, keeping the sub-surface nitrate profiles constant implicitly includes the effect of horizontal nitrate fluxes on time scales longer than a year.

Year-specific satellite data for 1998 were used to define the PAR and temperature cycles but were not available for mixed layer depth and nitrate. Climatological mixed layer depth and nitrate fields were therefore substituted. A consequential limitation of the model is that the temporal resolution of its mixed layer depth forcing data is inconsistent with the requirements of the hypothetical target application for which it is being evaluated. Mixed layer depth data from a climatological integration of a 3-dimensional general circulation model (see Appendix B) were used in preference to observational estimates. Use of model data is prudent because the observational estimates available do not sufficiently resolve the rapid shoaling of the mixed layer in the spring. To minimize problems due to incorrect timing of spring shoaling in the simulation, surface layer temperature cycles from the simulation were screened against the 1998 temperature observations and stations where the model and observations appeared to be out of phase were excluded. For the remaining stations, the mean winter-time (February-April) mixed layer depth from the simulation is compared with the observational estimate used to determine the annual nitrate maximum (Fig. 5). With the exception of some of the northern values, where the observational estimate is high as a result of the fixed σ_t criterion, there is generally good agreement between the two estimates, although the simulated values do show a high bias. Full details of the model's external forcing data are given in Appendix B.

The outputs required for the cost function are not explicitly modelled as state variables. Model chlorophyll concentration is derived from phytoplankton concentration using a spatially and temporally constant chlorophyll to nitrogen ratio which is one of the model's free parameters. The annual nitrate maximum is assumed to be equal to the annual nutrient maximum. In addition, the nutrient concentration immediately below the mixed layer is assumed to be equal to the nitrate concentration at the same depth, thereby allowing the amount of nutrient entrained into the mixed layer from below to be determined from the forcing data. In both cases, the concentrations are relatively high. The assumption would be more difficult to justify for surface nitrate observations in spring and summer when ammonium and other labile forms of dissolved nitrogen become important as nitrate is depleted.

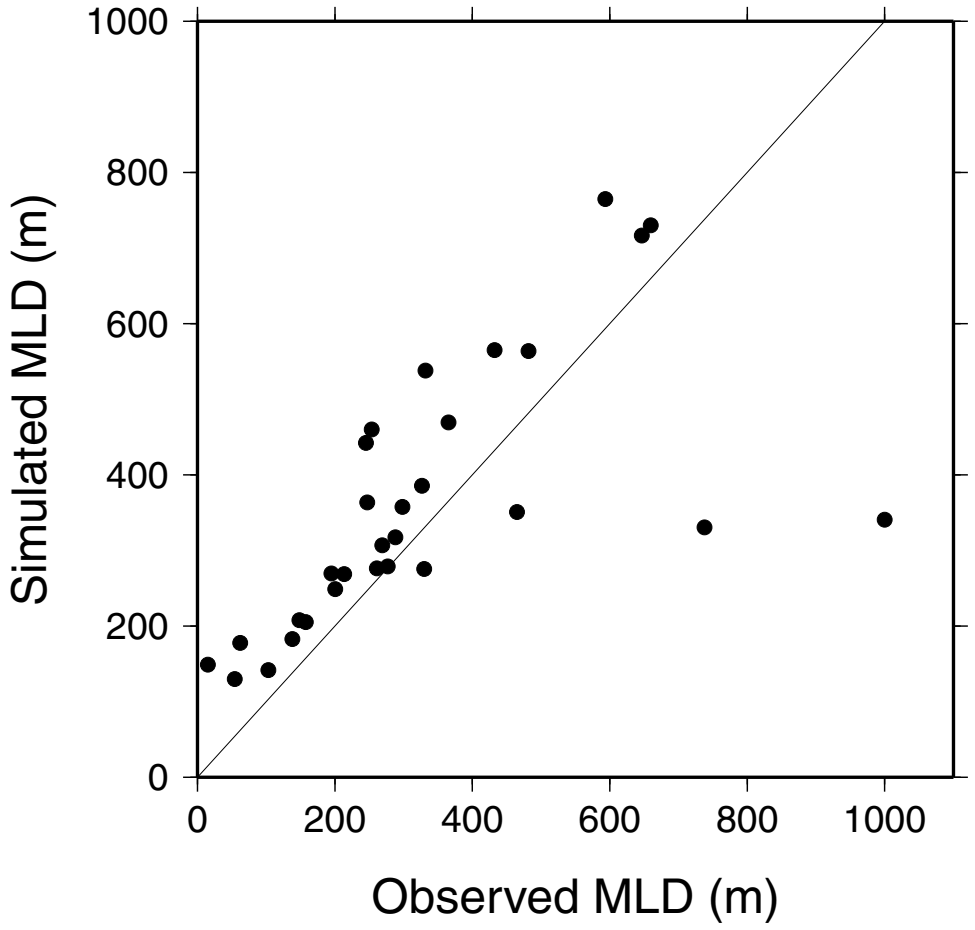

Fig. 5. Mean simulated mixed layer depth for the period February-April from the general circulation model compared with observational estimates for the same period. The latter are derived from the 1° monthly data of Levitus et al. (1982), based on a density difference criterion $\Delta\sigma_t = 0.125$. The 1:1 line is shown for reference.

4 Results

The results of applying the split-domain calibration method as defined in Section 2 to the PZN model are described here. The results of the robustness experiment are given in Appendix C.

4.1 Calibration domains

The results of applying the calibration algorithm to the North Atlantic domain are summarized in Table 2. Results are tabulated for the whole-domain calibration and each of the split-domain calibrations. The median validation

Table 2
Summary of calibration results for the North Atlantic domain

	Median validation cost	r.m.s. residuals $r(\log \mathcal{C})$	r.m.s. residuals $r(\mathcal{N})$	Domain or province latitude range (°N)	No. of stations in calibration set	Maximum coverage by station aggregation	Calibration group coverage (if different)
Calibration 1A	13.13	2.88	1.70	25-70	15	9(60%)	15(100%)
Calibration 2A	9.84*	2.37	1.51	25-45 45-70	6 9	5(83%) 9(100%)	4(67%) –
Calibration 2B	10.66*	2.40	1.69	25-40 40-70	5 10	3(60%) 10(100%)	– –
Calibration 2C	11.69*	2.72	1.52	25-70 50-60	11 4	9(82%) 4(100%)	– –

Validation costs for split-domain calibrations which are significantly lower (at 95%) than that for the whole-domain calibration (Calibration 1A) are marked *. Coverage of the domain or province is expressed in terms of the number of calibration stations and, in brackets, the proportion of the domain or province this represents.

costs are the medians of the distributions

$$\hat{\boldsymbol{J}}\left\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}), \boldsymbol{V}\right\}$$

and

$$\hat{\boldsymbol{J}}^{\mathrm{SPLIT}}\left\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_{A[i]}), \boldsymbol{V}_{A[i]}, \vec{p}_{\mathrm{BEST}}(\boldsymbol{C}_{B[i]}), \boldsymbol{V}_{B[i]}\right\}$$

for the whole-domain calibration and the ith split-domain calibration respectively. The root mean square residuals for the chlorophyll and nitrate validation data are also shown. These are the residuals with respect to the observation means, given by:

$$r(\log \boldsymbol{\mathcal{C}}) = \sqrt{\frac{\sum_{j=1}^{n}\sum_{i=1}^{n_j} M_{ij}(\vec{p}_j, \log \boldsymbol{\mathcal{C}})}{\sum_{j=1}^{n} n_j}} \tag{11}$$

and

$$r(\boldsymbol{\mathcal{N}}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} M_j(\vec{p}_j, \boldsymbol{\mathcal{N}})} \tag{12}$$

where n is the number of stations in the validation set $\boldsymbol{V}$, n_j is the number of chlorophyll observations for the jth station and $\vec{p}_j$ is the parameter vector

applicable to the jth station. The residual values are dimensionless, each referring to a number of standard deviations of the error distribution estimate for the observation. However, for nitrate, a constant observation error estimate with a standard deviation of 1 mmol m^{-3} is used throughout the domain, so the values are equivalent to concentrations expressed in mmol m^{-3}.

Latitude ranges are given in the table for each domain or province. The provincial latitude ranges are for identification purposes and do not constitute complete province descriptions. The actual geographical extent of each province is shown in Fig. 6. The maximum coverage of the domain or province by station aggregation indicates how representative the largest aggregation group from the applicable whole-domain calibration is of the full calibration set. A coverage of 100% implies that the largest aggregation group is effectively the full calibration set. This means that the station aggregation procedure completed successfully and that the single remaining station is satisfied by a parameter vector in the final search set. The omission of a small proportion of the calibration set's stations is acceptable: the omitted stations are considered atypical. However, in some cases the coverage is as low as 60%, calling into question the applicability of the aggregation result to the whole domain or province. The calibration group referred to in the table is usually the largest aggregation group. However, as indicated in Section 2.4.1, it can be one of the smaller groups if the optimal parameter vector for that group gives a lower validation cost. It can also be the full calibration set, irrespective of whether this qualifies as an aggregation group. Where the calibration group is not the largest aggregation group, its coverage is given as a separate entry in the table.

The whole-domain calibration result for the North Atlantic domain is referred to as Calibration 1A, where 1 denotes the number of parameter vectors required to cover the domain. This is based on the full 15 station calibration set. When the station aggregation procedure was applied, the maximum number of stations aggregated was only 9 out of 15. The calibration based on the 9 station group gives a median validation cost of 14.34, which is lower than those for the smaller groups but higher than that for the full calibration set.

Three different ways of splitting the domain into two were identified by choosing province indicator groups of different sizes. The resulting provinces are shown in Fig. 6, in order of the median validation costs of the associated calibrations. All three of these divisions produce calibrations which are significantly better (at 95%) than Calibration 1A. The best of these split-domain calibrations, Calibration 2A, has a median validation cost 25% less. Application of the calibration algorithm to each of the southern and northern provinces identified in Calibration 2A did not reveal any sensible geographic sub-divisions of these provinces as no province indicator groups were found. Calibration 2A is therefore the accepted calibration. It is encouraging to note that the two calibration stations without nitrate observations, Station 0606 (27.5°N 72.5°W)

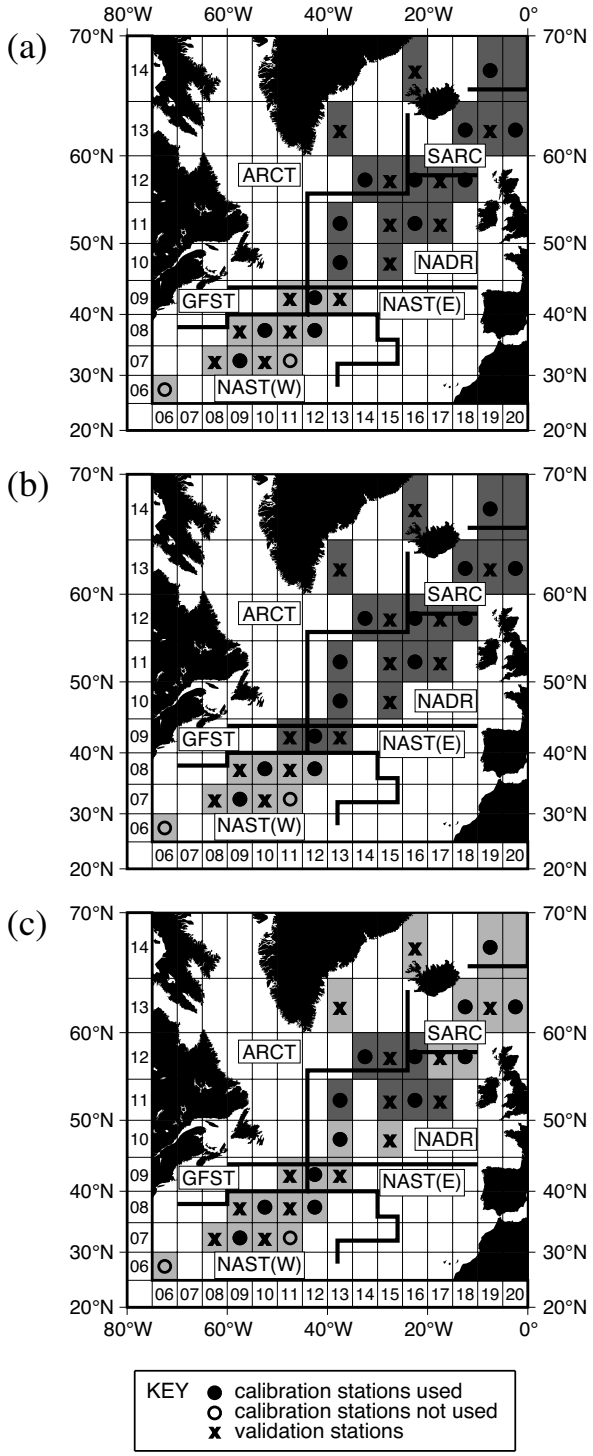

Fig. 6. Province extents for (a) Calibration 2A, (b) Calibration 2B and (c) Calibration 2C. The two provinces for each calibration are shown by light and dark shading. Biogeochemical provinces defined by Longhurst (1998) at 2° resolution are shown for reference. ARCT: Atlantic Arctic Province; SARC: Atlantic Subarctic Province; NADR: North Atlantic Drift Province; GFST: Gulf Stream Province; NAST: North Atlantic Subtropical Gyral Province.

and Station 0711 (32.5°N 47.5°W), are excluded from the calibration groups in all of the split-domain calibration results. This demonstrates that the method is able to allow for atypical stations.

Shown for reference in Fig. 6 are the divisions between the oceanic biogeochemical provinces defined by Longhurst (1998). In Calibration 2A, the division between the provinces is coincident with the boundary between the North Atlantic Drift Province and the North Atlantic Subtropical Gyral Province. This is consistent with the idea that model parameters should at least be invariant within the pre-defined biogeochemical provinces. The second best split-domain calibration, Calibration 2B, is very similar to the first and although the division is slightly further south it too coincides with pre-defined biogeochemical province boundaries. In the last split-domain calibration, Calibration 2C, one of the provinces is geographically divided such that northern and southern regions share the same parameter vector, while a mid-latitude region forms the alternate province. There is no clear relationship between these provinces and the pre-defined biogeochemical provinces.

Although there are obvious advantages in combining stations when calibrating a model for application to a wide area, the benefits are less clear for local studies in which a model is required to produce time-series estimates for a single location. In that case, we might expect local calibrations to have greater relevance. It is therefore informative to compare the calibrated model's goodness-of-fit at the validation stations with that obtained by extrapolating individual station calibrations locally.

Local calibrations were tested at each of the 15 validation stations in the domain by applying the optimal parameter vector for the nearest calibration station at the same latitude. Where stations were equally close geographically, that with the closest nitrate observation value was chosen. Stations with no nitrate data were ignored. The median validation cost for the North Atlantic domain based on these local, single-station calibrations is 12.92. While the median cost for the whole-domain calibration (Calibration 1A) is higher than this, the costs for all 3 of the split-domain calibrations (2A, 2B and 2C) are significantly lower (at 95%), with reductions of up to 24% (Calibration 2A).

The cost distributions for Calibration 2A were also compared with the local calibration costs at each validation station individually. Calibration 2A gives significantly lower costs at 8 of the stations (at 95%), while giving significantly higher costs at only 5 stations. The lower local calibration costs at the latter stations suggests the existence of regions, smaller than the Calibration 2A provinces, over which multiple parameter vectors might be required to best reproduce the spatial variability in the validation data. However, they could also be a consequence of the validation data not providing a truly independent test. Although it seems reasonable to assume that observation error is

uncorrelated between adjacent stations, so that the observations at each can be treated as independent samples of the annual chlorophyll cycle, the cycles themselves tend to be correlated between stations. The power of the validation data to test the generality of the calibrated model is to some extent compromised by this. In general, correlation between observations at validation and calibration stations will tend to introduce an unwanted bias towards the selection of smaller provinces, since a combination of the model's descriptive and predictive abilities are tested, rather than purely its predictive skill. It is therefore not possible to determine whether the favourable local parameter vectors reflect true local biogeochemical characteristics or are simply compensating for deficiencies in the model. In general, the single-station calibrations seem unreliable and the results presented here clearly demonstrate the advantage of combining multiple stations for local as well as basin-scale applications.

4.2 Spatial distribution of model error

The spatial distribution of the model misfits are summarized by maps of the station r.m.s. model residual for chlorophyll, defined at the jth station by

$$r_j(\log \boldsymbol{\mathcal{C}}) = \sqrt{\frac{1}{n_j} \sum_{i=1}^{n_j} M_{ij}(\vec{p}_j, \log \boldsymbol{\mathcal{C}})} \tag{13}$$

and the nitrate residual

$$\varepsilon_j(\boldsymbol{\mathcal{N}}) = \nu_{j\mathrm{MODEL}}(\vec{p}_j) - \nu_{j\mathrm{OBS}} \tag{14}$$

where $\nu \in \boldsymbol{\mathcal{N}}$. These are presented here (Fig. 7 and 8) for the whole-domain calibration (Calibration 1A) and for each of the split-domain calibrations (Calibrations 2A, 2B and 2C).

The chlorophyll error map for the whole-domain calibration (Fig. 7a) shows an approximate r.m.s. residual of between 2 and 4 observational standard deviations, representing a rather poor fit throughout the domain. The corresponding map for nitrate (Fig. 8a) shows that the winter-time nitrate maximum is underestimated everywhere by the model with the exception of Station 1319 (62.5°N 7.5°W). The magnitude of the error is 2 mmol m^{-3} or less almost everywhere though, which is relatively small compared with the chlorophyll error, based on the assumption implicit in the cost function of a 1 mmol m^{-3} error in the observational nitrate estimate.

In contrast with the whole-domain calibration, the split-domain calibrations all show some stations with chlorophyll r.m.s. errors of less than 2 standard

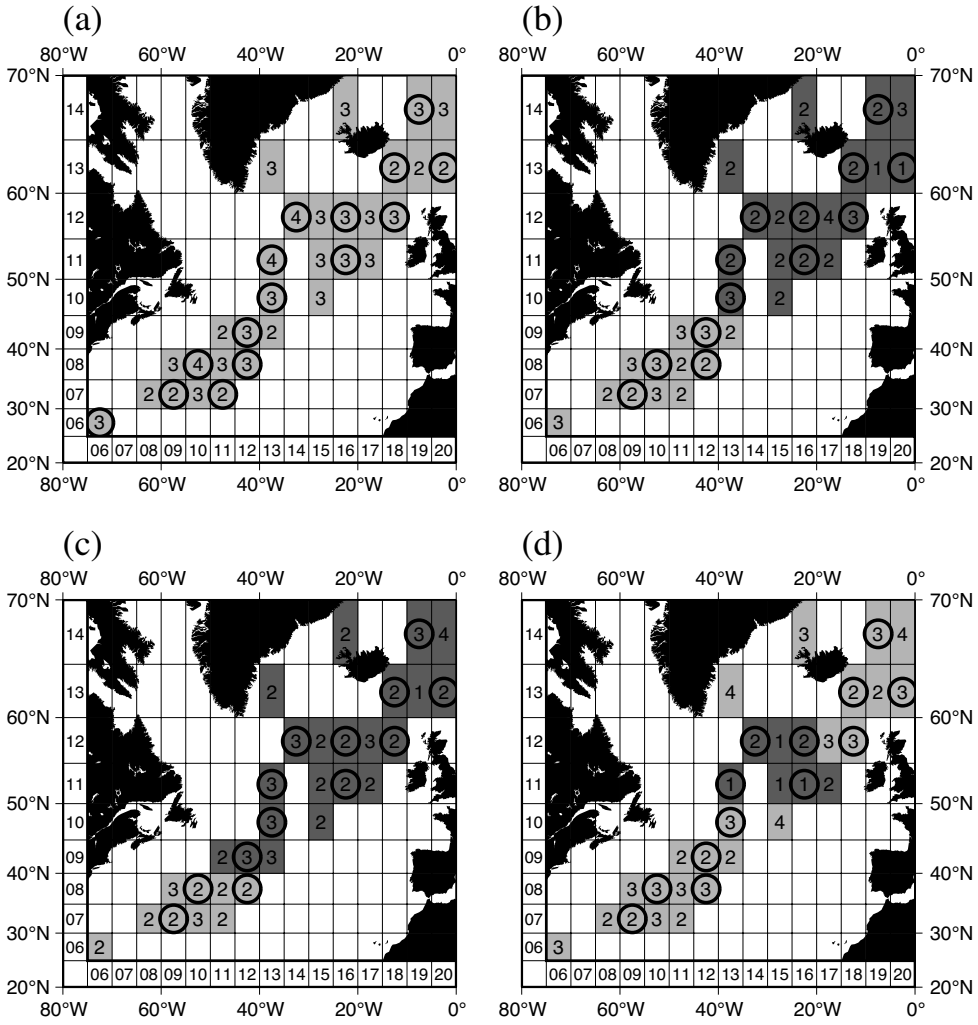

Fig. 7. Approximate station r.m.s. residual for chlorophyll $r_j(\log \boldsymbol{C})$, in number of standard deviations, for (a) Calibration 1A, (b) Calibration 2A, (c) Calibration 2B and (d) Calibration 2C. Circled stations are the calibration stations used. The extent of each province is indicated by the shading.

deviations (Fig. 7b-d). In Calibration 2A, there is some tendency for chlorophyll errors to be lower in the northern province than in the south. Almost everywhere in this northern province the fit to chlorophyll is better than in Calibration 1A, while the nitrate errors (Fig. 8b) are very similar.

The association between the chlorophyll errors and the different provinces in Calibration 2A is interesting. A clear association is also evident in Calibration 2C, where lower errors (Fig. 7d) are associated with the smaller mid-latitude domain (Fig. 6c). Association of lower errors with a particular province suggests that the model may be better suited to that province. With regard to Calibration 2A, it is possible that the relationship between the observed vari-

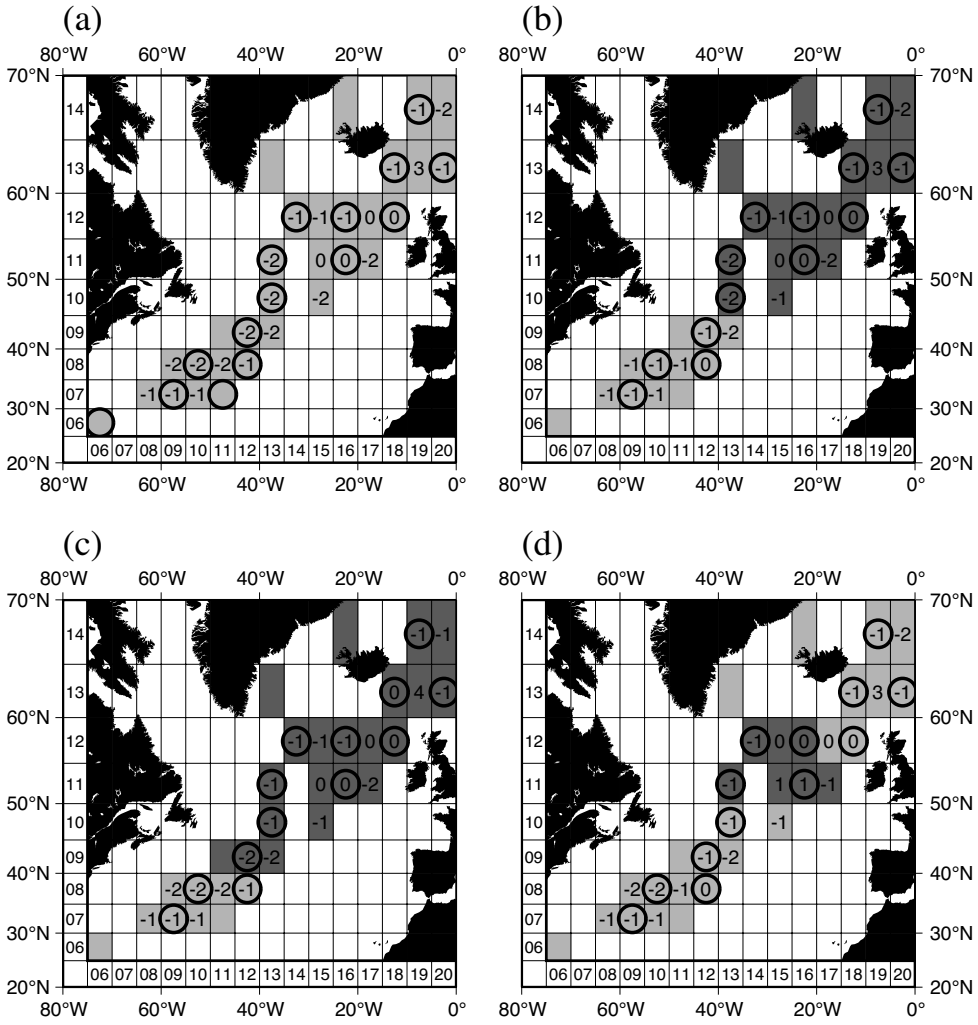

Fig. 8. Approximate nitrate residual $\varepsilon_j(\boldsymbol{\mathcal{N}})$ (mmol m^{-3}) for (a) Calibration 1A, (b) Calibration 2A, (c) Calibration 2B and (d) Calibration 2C. Circled stations are the calibration stations used. The extent of each province is indicated by the shading.

ability in chlorophyll and the environmental variability represented by the forcing data is stronger in the north than in the south, making the model better suited to the northern province. Certainly the fact that further divisions of the relatively large northern province do not improve the calibration implies that the forcing data provide at least part of the required spatial variability. This result is obtained despite the potential bias towards small province selection due to unwanted correlation between validation and calibration data. Because the northern province contains a much larger set of calibration data than the southern province, covering a wider range of environmental conditions, the northern province parameter estimates should be statistically more robust and more generic. Importantly, as discussed by Fennel et al. (2001),

the presence of larger blooms and more pronounced seasonal variation allow the functions describing the processes in the model to be sensibly constrained over a greater dynamic range.

4.3 Parameters

Unless parameters are well constrained by the observations, cost values which are only slightly higher than the lowest found J_{BEST} can occur over large areas of the parameter space. In many cases, an optimization result is therefore better represented in the form of a joint probability distribution for the optimal parameter values, rather than by a single vector. This posterior parameter distribution contains information about the degree of parameter constraint achieved as well as the correlation between different parameters. An estimate of the posterior parameter distribution can be derived from $\boldsymbol{P}_{\mathrm{good}}$ by removing outliers associated with unacceptably high costs, as done by Schartau et al. (2001). However, the definition of an unacceptably high cost is somewhat arbitrary. Accepting a wide range of costs increases the likelihood of including sub-optimal parameter vectors, while restricting the range reduces the size of the sample so that it becomes less representative of the distribution in parameter space of possible global minima. Schartau et al. (2001) defined costs within 25% of J_{BEST} as acceptable. A statistical approach is used here, which takes into account the observation error by making use of cost probability distribution estimates determined for each parameter vector, in place of single cost values. An unacceptably high cost distribution is defined as one which differs significantly from that with the lowest median at a chosen level of probability.

The accepted calibration result, Calibration 2A, implies that the model should have two regional parameter vectors. Estimates of the two posterior parameter joint probability distributions for Calibration 2A were determined from the parameter optimization results for the two calibration groups: for each province, a low cost subset of parameter vectors was selected from the set $\boldsymbol{P}_{\mathrm{good}}$ based on 100 different starting points in parameter space. The parameter vectors excluded were those with calibration cost distributions greater than that that with the lowest median at a significance level of 99.99%. The joint parameter distribution represented by the remaining subset shows the area of parameter space to which the model is constrained with a probability of 99.99% by the observations. The rejection level was chosen pragmatically so that the samples retained for both provinces were large enough to provide useful estimates. To investigate the constraint achieved at a lower probability level would require a larger ensemble size. The median calibration costs associated with the southern province parameter distribution, represented by a sample of 11 parameter vectors, range from 8.54 to 9.29. The corresponding range for the northern

province sample of 13 parameter vectors is 7.78 to 8.25. The upper costs for southern and northern provinces are 9% and 6% higher than the lowest costs respectively. These cost ranges are small compared with the 25% cost difference considered acceptable by Schartau et al. (2001). However, relatively small cost differences are highly significant here as a consequence of averaging over a large number of observations.

The univariate posterior parameter distribution estimates for the accepted calibration are shown in Fig. 9. These are the projections of the joint probability distribution estimates onto the parameter axes. In the case of both southern and northern calibrations, most parameters appear poorly constrained. There is evidence (not presented here) to suggest that some parameters may be much better constrained by the observations at a probability of 95%. However, this is based on sample sizes of just 3 for each province and cannot therefore be considered very reliable.

In some cases, parameters can be difficult to constrain because their optimal values are not independent. This is reflected by non-zero covariances in the posterior probability distributions. Parameter dependencies were investigated by calculating the Pearson correlation coefficients for all possible parameter pairs in each sample. In contrast with techniques used by other workers (Matear, 1995; Fennel et al., 2001), which are based on analysis of the Hessian matrix of the cost function at its minimum, our statistical approach is global with respect to the parameter space, allowing for the existence of multiple minima within the acceptable cost range.

Large positive and negative correlation coefficients were found for a number of the parameter pairs, the most notable of which are consistent between the independent results for southern and northern provinces. These are the two largest negative correlations and the largest positive correlation found in each case. The correlated parameters are the zooplankton ingestion half-saturation constant k_{G} and the chlorophyll to nitrogen ratio χ (correlation coefficients of -0.76 and -0.74 for the southern and northern provinces respectively), the zooplankton excretion rate μ and the chlorophyll to nitrogen ratio (-0.70 and -0.63) and the zooplankton excretion rate and the cross-pycnocline mixing rate m (+0.84 and +0.63). It is difficult to see any clear reasons for these parameter pairs to be correlated in reality and the relationships may be a consequence of unrealistic constraints imposed by the model design and/or forcing data.

Despite the uncertainty in parameter values, the posterior parameter distributions show some interesting patterns. Estimates of phytoplankton specific mortality ϕ_{P} are consistently low, suggesting that phytoplankton mortality is not an important process in the model. With the exception of some southern domain results, the initial slope of the photosynthesis versus irradiance (P-I) curve α tends to be high, indicating a weaker light limitation effect than ex-

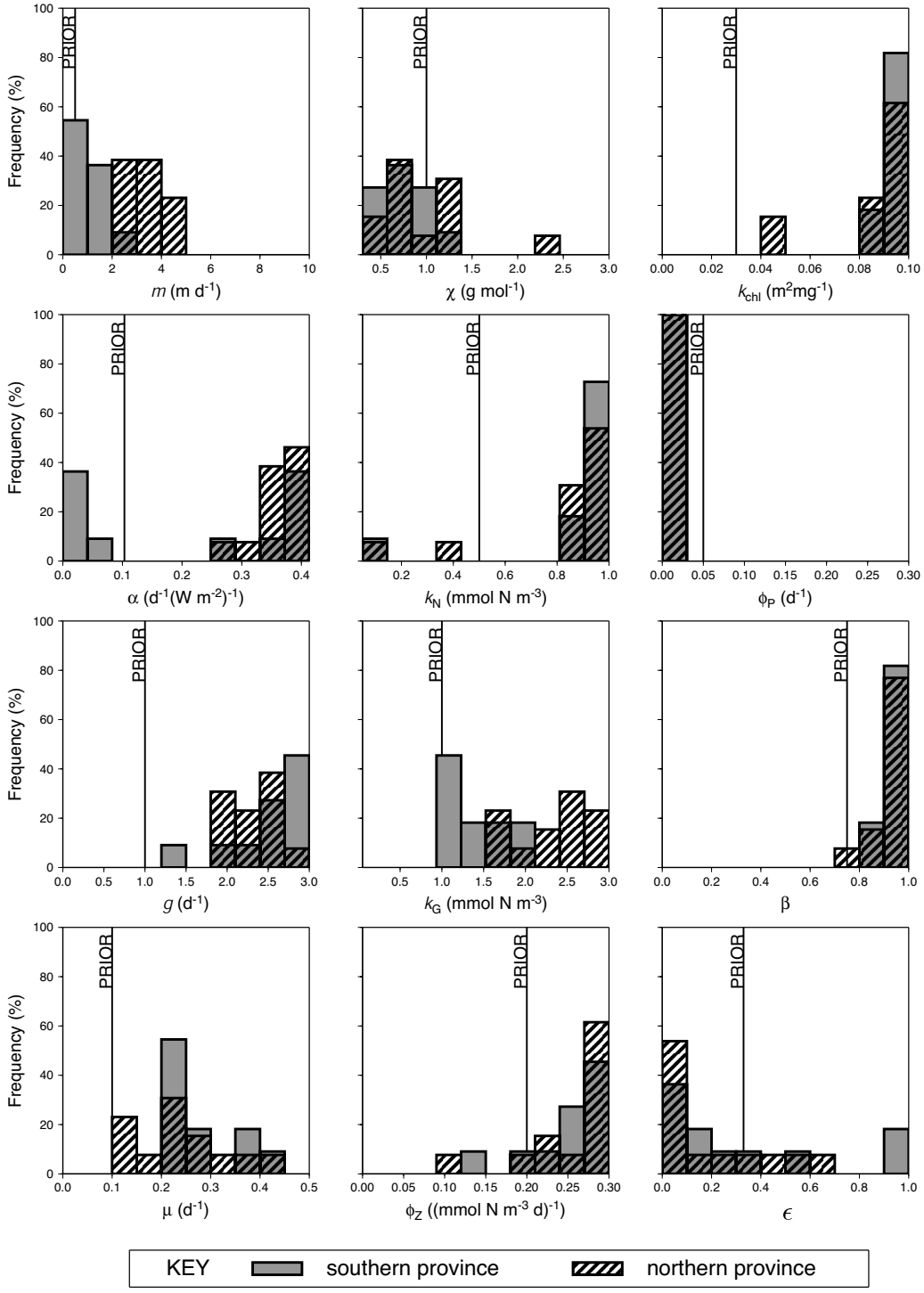

Fig. 9. Univariate posterior parameter distribution estimates for the accepted calibration (Calibration 2A), showing prior 'expected' values (see Table B.1). In each case the *abcissa* corresponds to the prescribed range over which the parameter is allowed to vary. Parameter values are interpreted with reference to the model equations given in Appendix B.

pected. The weak light limitation is compensated for at high phytoplankton concentrations by exceptionally high values of the chlorophyll light attenuation coefficient k_{chl}. The half-saturation coefficient for nutrient k_{N} tends to be higher than expected, implying that nutrient limitation is still important at relatively high nutrient concentrations. The zooplankton maximum ingestion rate g, assimilation efficiency β, excretion rate μ and mortality parameter ϕ_{Z} all tend to be high, producing a faster turnover of material by the zooplankton than the prior parameter values. Grazing pressure on the phytoplankton is high, but the high values for the zooplankton ingestion half-saturation constant k_{G} compensate for this by reducing the grazing at low phytoplankton concentrations.

We should note here that the prescribed bounds do in some cases allow parameters to vary over rather wider ranges than those that might be consistent with a perfect model structure. This can help to expose weaknesses of the model. The high values of the P-I slope α, for example, seem unrealistic. Most of the values obtained, often greater than 0.3 d^{-1} $(\mathrm{W\ m}^{-2})^{-1}$, are certainly not consistent with seasonal averages for the provinces of the Polar and Westerlies biomes presented by Sathyendranath et al. (1995), although the comparison is not straightforward. Their values, derived from *in situ* observations, are based on growth expressed in units of carbon per unit chlorophyll and, in addition, are only directly comparable with the model values in the absence of nutrient limitation. The values remaining after removal of the low summer and autumn values for the Westerlies provinces, where nutrient depletion is an important factor, vary over a very small range from 0.09 to 0.11 mg C $(\mathrm{mg\ Chl})^{-1}$ h^{-1} $(\mathrm{W\ m}^{-2})^{-1}$. When combined with the range of chlorophyll to nitrogen ratios suggested by our optimization results (about 0.5 to 1.5 g mol^{-1}) and the Redfield carbon to nitrogen ratio (6.625) this gives a range of values from 0.014 to 0.050 d^{-1} $(\mathrm{W\ m}^{-2})^{-1}$. Fig. 9 shows only a relatively small proportion of the parameter values within this range, all of which are for the southern province.

The most obvious differences in parameter distributions between the southern and northern provinces are a tendency in the north for higher cross-pycnocline mixing rates m, higher P-I slopes α and higher zooplankton ingestion half-saturation constants k_{G}. In the case of the half-saturation constant, the pattern is consistent with that found by Losa et al. (submitted) for a similar model calibrated locally. However their results for the P-I slope indicate a reverse pattern, with the highest values of the P-I slope occurring in the south. Higher mixing rates in the north, where the stratification tends to be weaker, are consistent with expectations.

Fig. 10 shows the modelled annual cycle of chlorophyll for the accepted calibration at each of the validation stations. To demonstrate the uncertainty associated with the lack of parameter constraint, the annual cycles for the best parameter vector are shown together with those for all other parameter vectors in the posterior distribution. Examples of the corresponding cycles of the state variables are shown in Fig. 11 for a low latitude station close to the BATS site (Station 0708 at 32.5°N 62.5°W) and a high latitude station close to OWSI (Station 1217 at 57.5°N 17.5°W). The variance in the posterior parameter distributions causes only a small amount of variability in the observed variables (Fig. 10 and Fig. 11c) but there is much greater uncertainty in phytoplankton and zooplankton biomass and in summer nutrient levels (Fig. 11).

A further concern is that the model does not appear to represent the temporal variability in chlorophyll very well. It captures almost none of the seasonal evolution in the south, where the observations show a clear bloom in spring (with the exception of Station 0708 and Station 0710) followed by a steady decline in summer and a rise again in the autumn. The model does show a similar bloom response at about the right time but this takes the form of a damped oscillation about a steady state rather than following the subsequent variation in the observations. In the north, the model does capture the elevated chlorophyll levels in spring and summer but fails to capture the strong variation at time-scales of weeks to months which is superimposed on top of this at some stations; most notably the spring and autumn bloom pattern at Station 1217.

Examination of the chlorophyll output at the calibration stations (Fig. 12) shows immediately that the problem is not just a symptom of applying the calibrated model to independent data. In addition, the individual station results in this figure show that the problem cannot be wholly attributed to the parameter compromises made during station aggregation. The shorter timescale features are poorly represented by the locally optimized model for at least 6 of the 15 stations (stations 0709, 0810, 0912, 1013, 1216 and 1218). The problem therefore appears to be inherent in the model and/or the forcing data. With the exception of Station 1216, the output at all of these 6 stations shows very strong oscillations. There are particular problems with the southern province calibration. In fact, when the two stations without nitrate observations (Stations 0606 and 0711) are excluded, the model only shows an annual cycle consistent with observations at one of the four remaining southern stations (Station 0812, 37.5°N 42.5°W). At the two stations without nitrate observations, the locally optimized model actually reproduces the chlorophyll record extremely well. However, this contrast with the results for

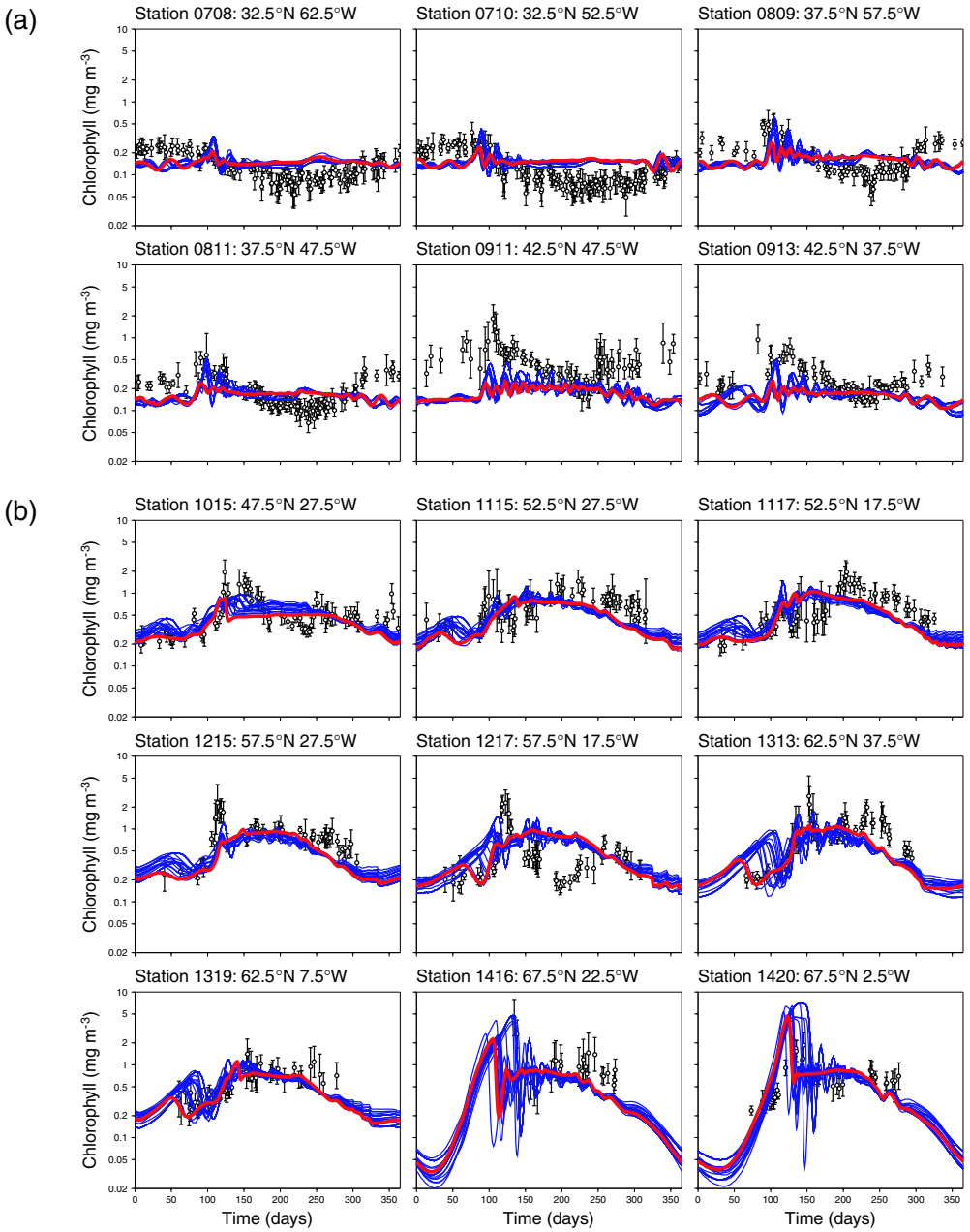

Fig. 10. Model annual cycles of chlorophyll at each of the validation stations in (a) the accepted southern province and (b) the accepted northern province. Chlorophyll observations are shown for reference with error bars at 1 standard deviation. At each station, the model output for each of the parameter vectors in the posterior distribution arising from the accepted calibration (Calibration 2A) is shown. The output for the parameter vector $\vec{p}_{\mathrm{BEST}}$ is highlighted.

the other stations seems suspicious, serving only to underline the importance of the additional constraint provided by the nitrate observations.

In general, the model appears to be smoothing out blooms, perhaps because

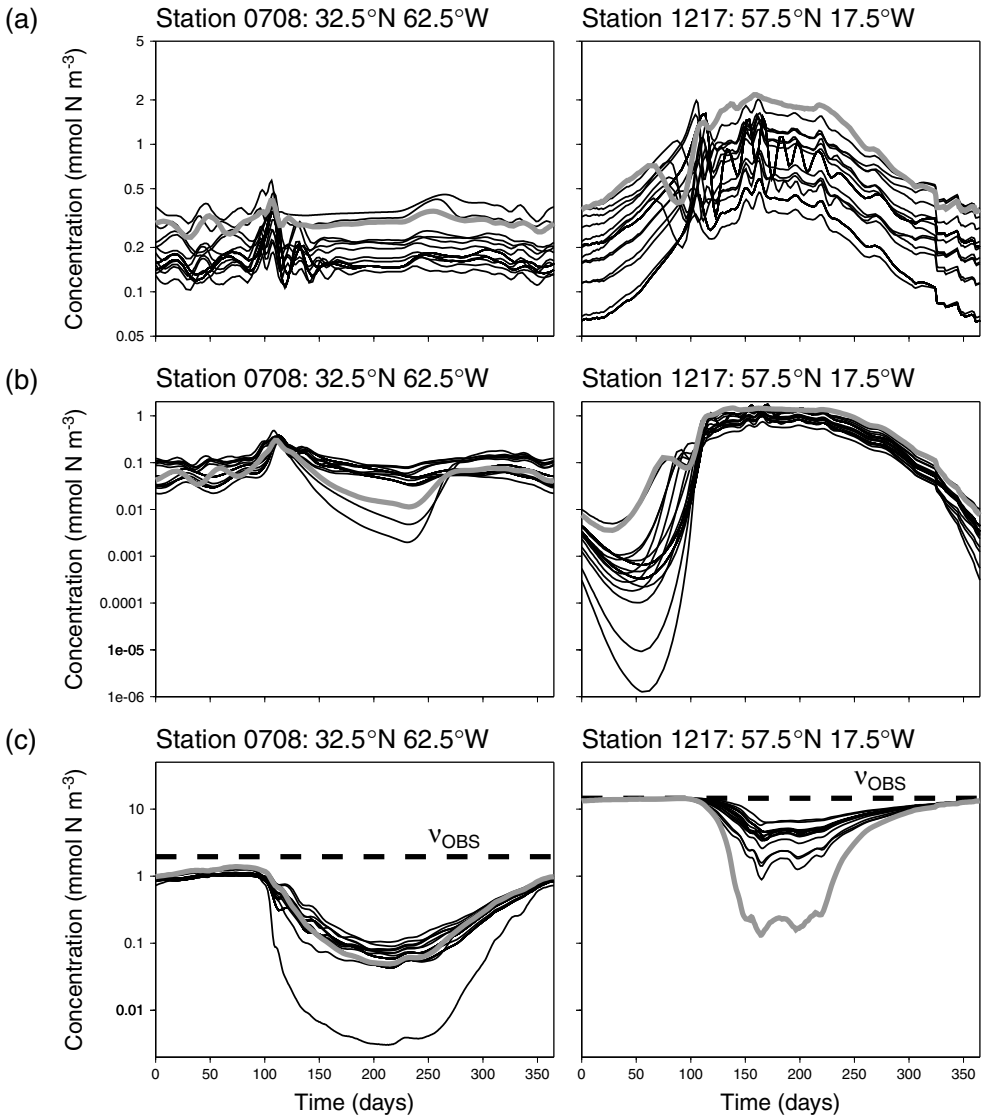

Fig. 11. Model annual cycles of (a) phytoplankton, (b) zooplankton and (c) nutrient at a southern province validation station and a northern province validation station. The observed annual nitrate maximum ν_{OBS} is shown for reference in (c). At each station, the model output for each of the parameter vectors in the posterior distribution arising from the accepted calibration (Calibration 2A) is shown. The output for the parameter vector $\vec{p}_{\mathrm{BEST}}$ is highlighted.

of deficiencies in the forcing data which prevent it reproducing the temporal structure in detail. In particular, mixed layer depth is taken from the output of a climatologically forced general circulation model, while a significant part of the observed variability in chlorophyll may be due to year-specific variability in mixed layer depth, including weather events on time scales of days to weeks. The problem might be alleviated in future work by using an independently validated mixed layer depth field from a general circulation model run with

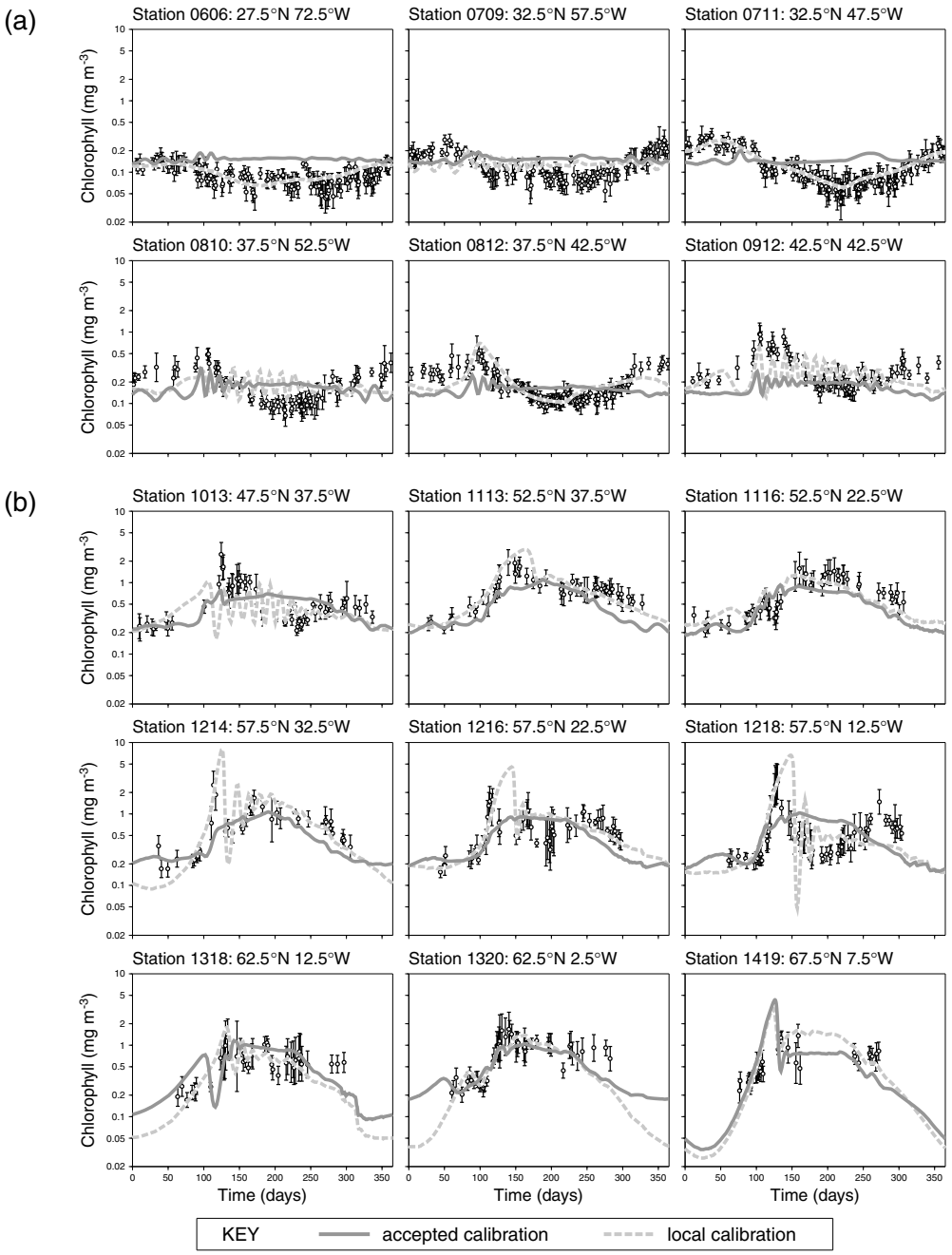

Fig. 12. Model annual cycles of chlorophyll at each of the calibration stations in (a) the accepted southern province and (b) the accepted northern province. Chlorophyll observations are shown for reference with error bars at 1 standard deviation. At each station, the model output given by the parameter vector $\vec{p}_{\mathrm{BEST}}$ from the accepted calibration (Calibration 2A) is shown, together with that given by $\vec{p}_{\mathrm{BEST}}$ for the local, single station calibration.

year-specific meteorological forcing. However, there may of course be limitations of the ecosystem model itself which prevent it responding realistically to the forcing data provided.

The inability of the model to reproduce the temporal structure leads to the smoother fit which explains a number of features of the parameter distributions: the high chlorophyll attenuation coefficients k_{chl} and high grazing rates g reduce the peak phytoplankton concentrations, as does the high cross-pycnocline mixing rate m. However, these values do not prevent the initial spring increase in phytoplankton because of the effect of other parameters: the high values for the zooplankton ingestion half-saturation constant k_{G} reduce the grazing pressure when phytoplankton concentrations are low and the P-I slope α, having high values in the north, reduces light limitation in regions where low surface PAR and deep mixed layers might otherwise prevent early accumulation of biomass. The north-south differences in these two parameters may be a consequence of a requirement for increased damping in the north to cope with the stronger physical forcing, the requirement being met by high mixing rates. Certainly, given the rather unsatisfactory performance of the model, great care should be taken to avoid over-interpretation of its posterior parameter distributions. Likewise, limited ecological significance should be ascribed to the provinces identified.

5 Discussion

The split-domain calibration method has been shown to be a practical method for estimating the best model fit to validation data over a large domain, while avoiding prior assumptions about the geographic scope of individual parameter vectors. Such assumptions are undesirable because they could lead to sub-optimal calibrations. The method allows us to obtain a measure of merit for a given model, together with its forcing data, which can be used to compare different models or different forcing data sets. For the present model and forcing data, this measure, the median validation cost for the best calibration found, is 25% lower than the equivalent cost for the calibration based on the full set of 15 stations. It is also 24% lower than the cost obtained when the model was calibrated locally, using individual stations. These results clearly demonstrate the utility of the method. However, it is recognised that the possible existence of a better solution to the calibration problem cannot be disproved.

Further developments of the method could be considered which would increase confidence in the result, albeit at the expense of a higher computational load. The simplest of these would be to increase the ensemble size in the parameter optimization procedure to allow the station aggregation procedure to explore

more of the parameter space. Another development to consider relates to how the hierarchy of calibration solutions with different numbers of provinces is explored. In the method described here, a split-domain calibration result which does not produce an improvement over the corresponding whole-domain calibration is rejected without exploring the consequences of possible domain divisions further down in the hierarchy. A more thorough approach would be to explore all possible branches. In that case the recursion would be terminated only when no province indicator groups were found.

For the present model, the best representation of the observed spatial variability across the basin is achieved by introducing spatial variation in parameters as well as forcing data. For the purposes of making predictions of biological responses to physical change, a more useful model would be one with a single parameter vector in which the spatial variability was determined purely by the forcing. The work of Hurtt and Armstrong (1999) suggests this might be feasible for the North Atlantic. Their model was able to fit data from BATS and OWSI simultaneously when they included multiple size classes of phytoplankton and detritus and a variable chlorophyll to nitrogen ratio for the phytoplankton. The latter was modelled as a function of the model's state variables and its forcing data. Advantages might therefore be gained from replacing the chlorophyll to nitrogen ratio in the present model by a similar function. Other parameters might also be replaced by functions of internal and/or external variables. The cross-pycnocline mixing rate, for example, could perhaps be expressed as some empirical function of the available forcing data. The possibility should be explored for the remaining biological parameters as well. However, as pointed out by Longhurst (1998), we should not necessarily expect to find ecological continuity in the plankton response to environmental forcing, because biological responses are often species dependent and are further complicated by species succession. The use of different model parameter vectors, with some form of smooth transition over domain boundaries, might therefore be the only sensible way of representing some regional variations.

The high levels of uncertainty in the aspects of model output which are not directly constrained by the observations demonstrate clearly the potential value of augmenting the data used in this study with other types of observations. Improving the forcing data and/or the model might lead to better constrained parameter vectors, but it seems unlikely that it would be possible to derive a single parameter vector which can be used with confidence in an application without taking into account other observations. Whether the additional data are observations of different variables or at different times of year, careful consideration must be given to their spatial distribution because differences in the type of constraints imposed at different stations can impact on the way a domain is divided. The potential for this is shown by the fact that the station aggregation procedure distinguishes between stations with and without a nitrate observation. In this study, because there were only two stations without

nitrate, these were treated as atypical and had no impact on the result. In general though, it is clear that the spatial distribution of different observation types needs to be as uniform as possible.

In the absence of other types of observations, parameter constraints can be improved by reducing the number of free parameters or including parameter penalties or other criteria in the cost function. However, unless such constraints can be properly justified, the associated reduction in uncertainty is misleading. While a model with fewer free parameters might be easier to constrain with the available observations, the advantage is gained at the expense of introducing arbitrary constraints in the form of fixed parameter values. The same problem applies to the issue of model complexity in general. The model tested here has many built-in constraints which simplify its design but are difficult to justify on theoretical grounds. More complex models generally impose fewer arbitrary constraints but, as a consequence, their parameters are more difficult to constrain by data assimilation. Ignoring the practical limits to model complexity, it can be argued that an ideal model is one constrained only by observations and well-established theory. If we move closer to such a model we expect the uncertainty in model output to increase, even allowing for improvements in the observation set. Although this presents practical problems, it does better represent the limits to our understanding of the real system. For some applications it may be desirable to accept the uncertainty and monitor its effects, rather than try to remove it artificially.

The need for stations to be treated independently in the split-domain calibration method, combined with the iterative nature of the optimization techniques, restrict the practical application of the method to models with low spatial dimension. Evans (1999) showed that the potential errors in parameter estimates associated with using a zero-dimensional model, with a homogenous surface layer and zero biomass concentrations below, are serious. Consideration should therefore be given to resolving the water column more explicitly, within the constraints of available computing power. One approach is that of Friedrichs (2002) who fixed the form of the vertical profile for each of the state variables, allowing only its magnitude to vary. Alternatively, given sufficient computing power, a 1-dimensional model might be employed as in the work of Prunet et al. (1996a,b). The issue of horizontal fluxes, especially advection, must also be addressed. A useful framework for this which avoids the need for repeated runs of an expensive 3-dimensional ecosystem model was presented by Gunson et al. (1999). Their approach involves the application of an ecosystem model to Lagrangian water columns following surface water trajectories extracted from a general circulation model. The length scale of such trajectories, over the integration time of the model, would of course impose limitations on the resolution to which province boundaries could be defined.

The calibration method demonstrated here for a marine ecosystem model can

be applied to other fields of research where model run-time is not prohibitive. It has potential value in any situation where it cannot be assumed that a single parameter vector is appropriate for the whole model domain. In the example here observation records are associated with geographic locations, but they could be associated with points in any ordinal space.

Acknowledgements

The authors would like to thank the SeaWiFS Project (Code 970.2) and the Distributed Active Archive Center (Code 902) at the Goddard Space Flight Center, Greenbelt, MD 20771, for the production and distribution of data respectively. these activities are sponsored by NASA's Mission to Planet Earth Program. Thanks are due to the National Center for Atmospheric Research for data distribution and to Yanli Jia for providing the simulated mixed layer depth data. The authors would also like to thank Marjorie Friedrichs and one anonymous referee for their helpful comments and suggestions on the the manuscript. This research was supported under the NERC Data Assimilation Thematic Programme award number NER/T/S/1999/00104.

A Limitations of the station aggregation procedure

The station aggregation procedure seeks groups which have the minimum aggregation penalty for their size. There are two limitations which can prevent it from consistently identifying the correct group. The principal limitation arises from inadequate sampling of the parameter space when evaluating the group maximum cost deviation for the candidate groups. The other, secondary limitation is a consequence of the differences between the variation of the cost function and the group maximum cost deviation in parameter space.

The area of parameter space sampled, the search set $\boldsymbol{P}$, depends on the analysis of the cost function for a smaller aggregation group $\boldsymbol{G}^n$, performed by the parameter optimization procedure ($\boldsymbol{P} = \boldsymbol{P}_{\text{good}}(\boldsymbol{G}^n)$). The sampling problem can be addressed in part by increasing the ensemble size used in this analysis, thus increasing the probability that all relevant minima are found. However, the procedure still relies on there being a broad similarity between the smaller group's cost function and the variation in parameter space of the cost deviation for the optimum group sought. If this is not the case then the wrong group may be selected. This is illustrated in Fig. A.1, where the location of the minimum in the cost function for $\boldsymbol{G}^n$ is much closer to that of group maximum cost deviation curve for the non-optimal group $\boldsymbol{H}_2^{n+1}$ than that for the optimal group $\boldsymbol{H}_1^{n+1}$. The lowest group maximum cost deviation sampled is

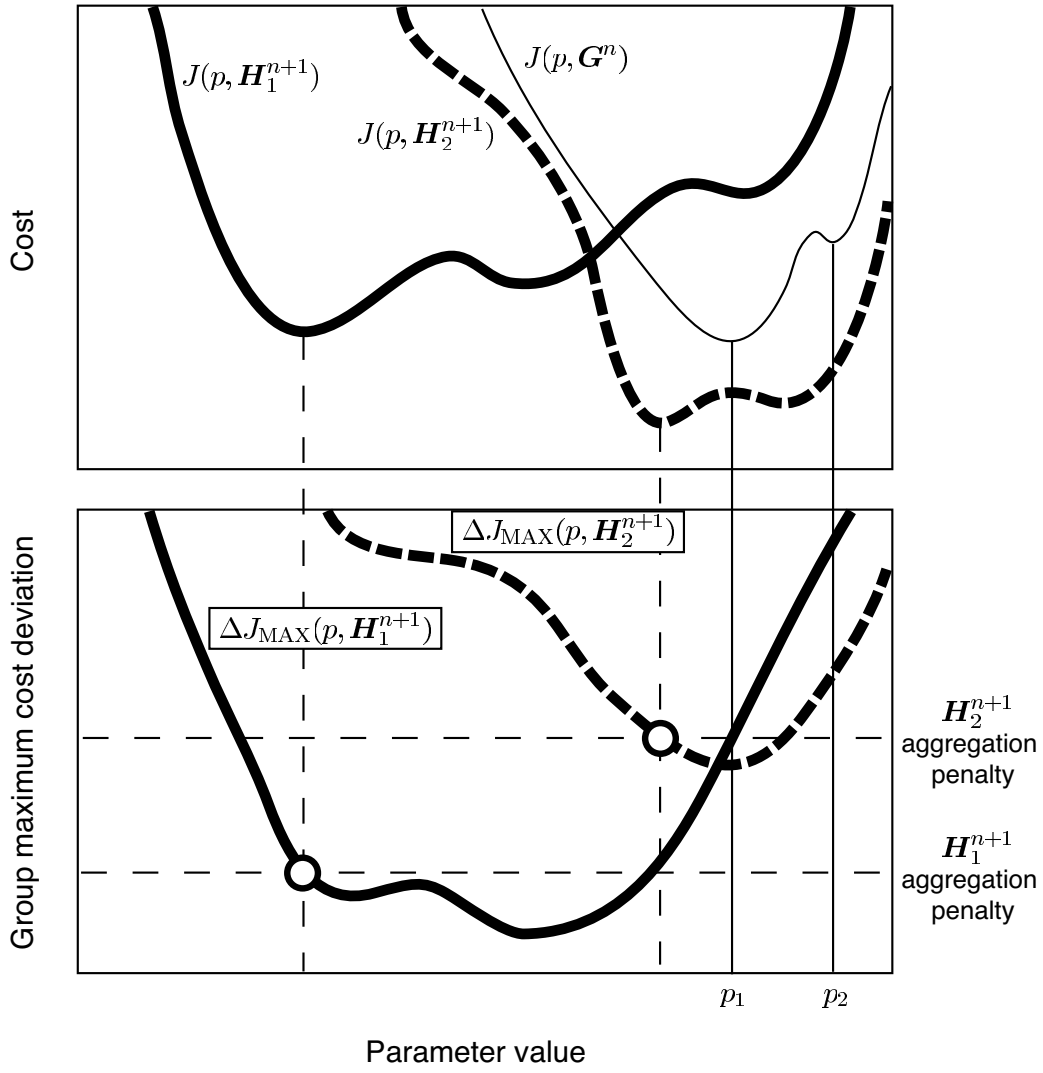

Fig. A.1. Failure of the station aggregation procedure to select the optimal group $\boldsymbol{H}_1^{n+1}$ due to poor representation of the parameter space in the region of its lowest group maximum cost deviation. The lowest group maximum cost deviation found is that for $\boldsymbol{H}_2^{n+1}$ at p_1.

on the curve for $\boldsymbol{H}_2^{n+1}$ at p_1. In general, there is a bias towards selection of groups which are similar to those already aggregated. This is likely to be more of a problem when groups are small due to the large relative increase in size at one step of the aggregation procedure and the minimal pre-conditioning of the search set $\boldsymbol{P}$.

The secondary limitation means that perfect representation of the parameter space, such that the set of sample points includes the optimal parameter vector for the group sought (i.e. $\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}_{\mathrm{OPT}}^{n+1}) \in \boldsymbol{P}$), does not guarantee success. This is because the lowest value of the group maximum cost deviation for a group can be smaller than its aggregation penalty. This is a potential problem for groups with similar aggregation penalties as shown in Fig. A.2. Here the

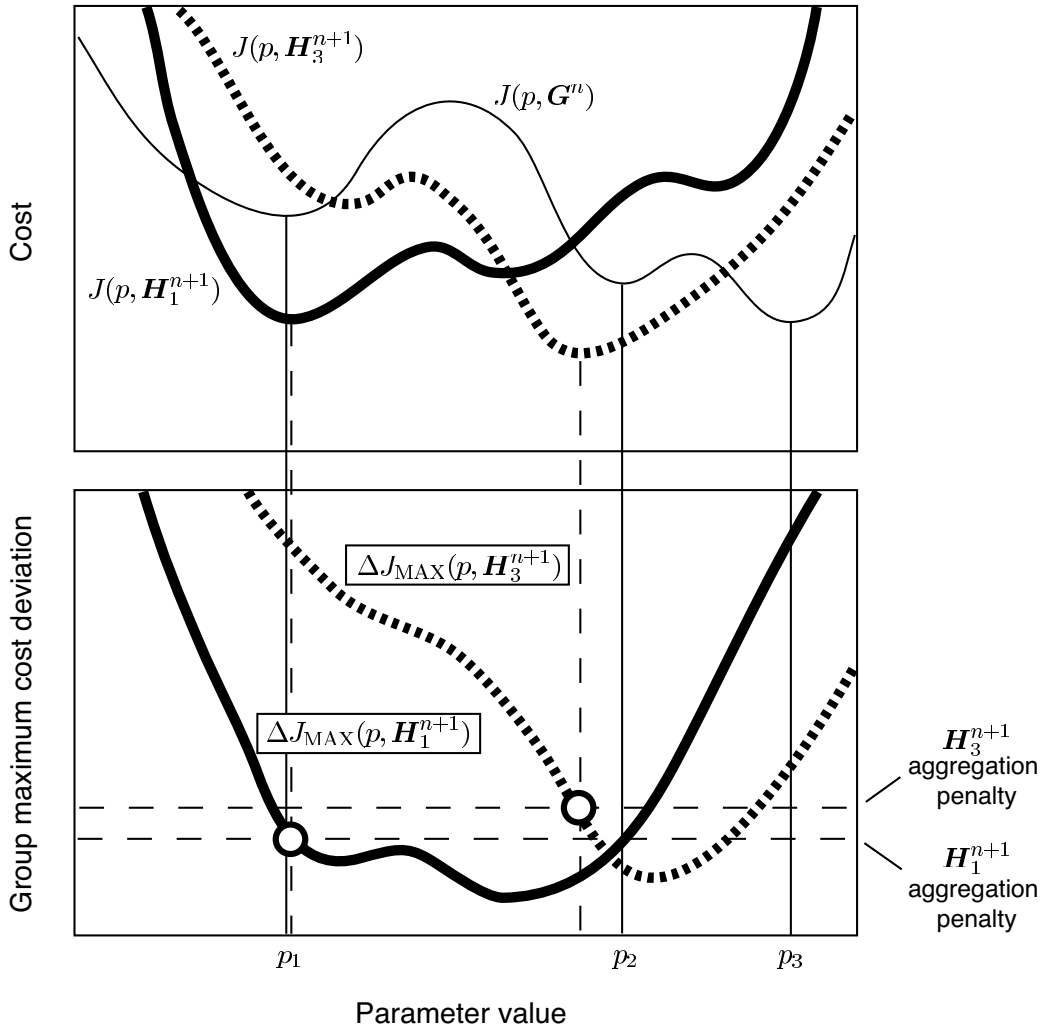

Fig. A.2. Failure of the station aggregation procedure to select the optimal group $\boldsymbol{H}_1^{n+1}$ due to the presence of a lower group maximum cost deviation for a non-optimal group $\boldsymbol{H}_3^{n+1}$ at p_2 in the search space defined by p_1, p_2 and p_3.

non-optimal group $\boldsymbol{H}_3^{n+1}$ has a group maximum cost deviation curve which dips below the aggregation penalty for the optimal group $\boldsymbol{H}_1^{n+1}$. In this example, the sampling is such that the wrong group is selected. The lowest group maximum cost deviation is on the $\boldsymbol{H}_3^{n+1}$ curve at p_2.

Avoidance of this problem relies firstly on the correlation between the aggregation penalty and the minimum value of the group maximum cost deviation over different groups and secondly on the greater likelihood of the correct group being sampled in the region of its minimum as a result of the pre-conditioning. The correct group $\boldsymbol{H}_{\mathrm{OPT}}^{n+1}$ is selected if at least one of the parameter vectors in $\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{G}^n)$ gives a group maximum cost deviation for $\boldsymbol{H}_{\mathrm{OPT}}^{n+1}$ sufficiently close to or less than its aggregation penalty $\Delta J_{\mathrm{MAX}}\{\vec{p}_{\mathrm{BEST}}(\boldsymbol{H}_{\mathrm{OPT}}^{n+1}), \boldsymbol{H}_{\mathrm{OPT}}^{n+1}\}$ to be lower than the group maximum cost deviation $\Delta J_{\mathrm{MAX}}\{\vec{p}, \boldsymbol{H}_{\mathrm{BEST}}^{n+1}(\vec{p})\}$ for competing groups, for all $\vec{p}$ in $\boldsymbol{P}_{\mathrm{good}}(\boldsymbol{G}^n)$.

B Test model definition

B.1 External forcing data

The ecosystem model's external forcing data consists of spatially varying annual cycles of day length $d(t)$, daily mean photosynthetically available radiation (PAR) $\overline{I}(t)$, mixed layer depth $M(t)$ and phytoplankton maximum growth rate $V_{\mathrm{P}}(t)$ and spatially varying annual mean vertical nitrate profiles $N_{\mathrm{s}}(z)$. The maximum growth rate is modelled as a function of water temperature $T(t)$ using the empirical relationship derived by Eppley (1972):

$$V_{\mathrm{P}} = 0.6(1.066^{T}) \tag{B.1}$$

Observed sea surface temperature (SST) for the same year as the chlorophyll observations, 1998, was used to derive $T(t)$. The SST measurements are 8 day mean Advanced Very High-Resolution Radiometer (AVHRR) data at a resolution of 18 km, averaged over a circle of 100 km radius about each station to match the length scale of the chlorophyll observations.

The PAR forcing $\overline{I}(t)$ was derived from SeaWiFS 8 day mean PAR Standard Mapped Image data for 1998 at a resolution of 9 km. This is an estimate of the downwelling irradiance reaching the sea surface. Again the data are averaged over a 100 km radius about each station. For times at which observed PAR is available that value was used. To define PAR at other times a simple transmission model of the form introduced by Evans and Parslow (1985) was employed. This is

$$\overline{I} = a(1 - bC)\overline{I_S} \tag{B.2}$$

where $\overline{I_S}$ is the daily mean solar radiation at the top of the atmosphere, integrated over all wavelengths, as a function of time and latitude. This is determined using standard formulae (e.g. Brock, 1981). The variable C is the fractional cloudiness and a and b are constants. An 'effective' cloudiness time series was derived which satisfies the transmission model at the times when PAR is defined by the observations. This was then linearly interpolated between observation times. Where gaps between observations are more than 8 days the missing data was first filled in by linearly interpolating a 3 point running mean of the effective cloudiness. The values of a and b, 0.38 and 0.45 respectively, were determined by fitting the model to observational data from 36 stations distributed across the North Atlantic. The PAR data used were the 1998 SeaWiFS values. The cloudiness data were extracted from climatological fields produced by Bishop et al. (1994) using International Satellite Cloud Climatology Project (ISSCP) data (Rossow and Schiffer, 1991).

Mixed layer depth time series $M(t)$ were extracted from the output of a climatologically forced general circulation model. The general circulation model was an implementation of the Miami Isopycnic Co-ordinate Ocean Model (MICOM) (Bleck et al., 1992) for the North Atlantic, described by Jia (2000). The time series were extracted from the final year of a 16 year integration of the coarse resolution (4/3°) version. The mixed layer in this model is of the vertically homogenous Kraus-Turner formulation (Kraus and Turner, 1967). Mixed layer model parameters for this particular run (Jia, personal communication 1997) had been tuned to data collected during a subduction experiment conducted in an area of the North Atlantic from 18-33°N and 22-34°W (Moyer and Weller, 1997). The distribution of the layer densities differs slightly from those in Jia (2000), giving higher resolution over part of the lower density range. MICOM output data were averaged over 5° boxes centred on station locations to remove aliasing associated with the discrete density layers.

The annual mean nitrate profiles were determined by fitting profiles of the form

$$N_{\mathrm{s}} = a_{\mathrm{N}} \ln(b_{\mathrm{N}} z + 1) \tag{B.3}$$

to World Ocean Atlas annual mean 1° analyzed nitrate data (Conkright et al., 1998) at each station. Only data between depths of 100 m and the mixed layer depth forcing maximum were used. Data above 100 m are considered unreliable due to seasonal variations in the observations. Extrapolation of the observed profile into shallower depths is justified to some extent because water below 100 m shows a partial signature of near surface nitrate depletion, becoming weaker with depth, as a result of vertical mixing. Although each model profile is forced through the origin, it is assumed that entrainment of nitrate in the model is suppressed when the sub-surface nitrate concentration is less than that in the mixed layer, so in areas where nitrate is not used up in the summer the upper part of the profile has little influence on the model results.

B.2 Ecosystem model equations

The model's state variables are the concentrations of nitrogen in phytoplankton (P), zooplankton (Z) and nutrient (N) pools. The nitrate and ammonium pools modelled separately by Fasham et al. (1990) are combined into a single pool and the detritus pool and the pools associated with the microbial loop are absent. Detrital material such as dead plankton or faecal pellets is immediately exported from the system as it is produced. New nitrogen enters the system from below the mixed layer as a result of nutrient entrainment during mixed layer deepening and by diffusive mixing across the mixed layer base, parameterized by a constant mixing rate m. Plankton concentrations are zero

below the mixed layer.

The rate of change of the phytoplankton concentration is given by

$$\frac{\mathrm{d}P}{\mathrm{d}t} = P\overline{J}Q - G_{\mathrm{P}} - \phi_{\mathrm{P}}P - \frac{(m + h^{+})P}{M} \tag{B.4}$$

where $h^{+} = \max(\mathrm{d}M/\mathrm{d}t, 0)$. The factor $\overline{J}$ is the daily mean light limited specific growth rate which is a function of PAR, mixed layer depth and the self-shading effect of phytoplankton biomass. Q is a nutrient limitation factor given by

$$Q = \frac{N}{k_{\mathrm{N}} + N} \tag{B.5}$$

where k_{N} is the Michaelis-Menten half-saturation constant for nutrient. The loss term G_{P} is the zooplankton grazing rate given by

$$G_{\mathrm{P}} = \frac{gZP}{k_{\mathrm{G}} + P} \tag{B.6}$$

where g is the maximum ingestion rate and k_{G} is the half-saturation constant for zooplankton ingestion. The remaining loss terms are exports from the system. These are the phytoplankton mortality, parameterized by a constant specific mortality rate ϕ_{P}, and the physical flux due to dilution as a result of vertical mixing processes.

Following Fasham et al. (1990), the light limited growth rate J at a given depth and time is defined by

$$J = \frac{V_{\mathrm{P}}\alpha I_{\mathrm{z}}}{\sqrt{V_{\mathrm{P}}^{2} + \alpha^{2} I_{\mathrm{z}}^{2}}} \tag{B.7}$$

where α is the initial slope of the photosynthesis versus irradiance (P-I) curve and I_{z} is the underwater light field. The light field is modelled in terms of the PAR directly below the sea surface I_0, the attenuation of PAR due to water k_{w} ($0.04\ \mathrm{m}^{-1}$) and the specific attenuation of PAR due to chlorophyll k_{chl}, the model taking the simple Beer's law form

$$I_{\mathrm{z}} = I_0 \exp\{-(k_{\mathrm{w}} + k_{\mathrm{chl}}\chi P)z\} \tag{B.8}$$

For the purposes of integrating over the day, the time since sunrise t_{D} is treated independently of the time of year t. The variation of I_0 with time of day is

modelled as a triangular function

$$I_0 = 2\frac{\overline{I}}{d} f_{\mathrm{D}}(t_{\mathrm{D}}) \tag{B.9}$$

where f_{D} increases linearly from 0 to 1 between $t_{\mathrm{D}} = 0$ and $t_{\mathrm{D}} = d/2$ and decreases linearly from 1 to 0 between $t_{\mathrm{D}} = d/2$ and $t_{\mathrm{D}} = d$. This ensures that the daily mean light limited growth rate in the mixed layer

$$\overline{J} = \frac{1}{M} \int_0^d \int_0^M J \, \mathrm{d}z \, \mathrm{d}t_{\mathrm{D}} \tag{B.10}$$

has an analytical solution (Evans and Parslow, 1985).

The zooplankton equation is

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \beta G_{\mathrm{P}} - \mu Z - \phi_{\mathrm{Z}} Z^2 - \frac{(m + h^+)Z}{M} \tag{B.11}$$

where β is the assimilation efficiency, μ is the zooplankton specific excretion rate and ϕ_{Z} is the zooplankton specific mortality parameter. The zooplankton excretion term represents a nitrogen flow from the zooplankton to nutrient pools. Nitrogen associated with zooplankton mortality is exported. Zooplankton faecal material (the fraction not assimilated) is divided between labile material, which is transferred to the nutrient pool, and refractory material, which is exported.

The nutrient equation is

$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}t} = & -P\overline{J}Q + \mu Z + (1 - \epsilon)(1 - \beta)G_{\mathrm{P}} \\ & + \frac{m + h^+}{M} \max\{N_{\mathrm{s}}(M) - N, 0\} \end{aligned} \tag{B.12}$$

where ϵ is the exported fraction of zooplankton faecal material and N_{s} is the nutrient concentration immediately below the base of the mixed layer. Because the nutrient profile does not vary temporally, sub-surface nutrient concentrations can be less than those in the mixed layer when the mixed layer first shoals in the spring. This can cause negative fluxes which are simply an artefact of the model. Such fluxes are suppressed, as indicated in Eq. (B.12), on the assumption that the concentration below the mixed layer is always in reality at least as high as that within the mixed layer.

Table B.1
Model parameters

Parameter	Symbol	Unit	Prior value	Lower bound	Upper bound
cross-pycnocline mixing rate	m	m d^{-1}	0.5	0	10
phytoplankton chlorophyll:N ratio	χ	g mol^{-1}	1	0.3	3
chlorophyll light attenuation coefficient	k_{chl}	m^2 mg^{-1}	0.03	0	0.1
initial slope of P-I curve	α	d^{-1} (W m^{-2})$^{-1}$	0.1	0	0.41
nutrient uptake half-saturation constant	k_N	mmol N m^{-3}	0.5	0.05	1
phytoplankton mortality rate	ϕ_P	d^{-1}	0.05	0	0.3
zooplankton maximum ingestion rate	g	d^{-1}	1	0	3
zooplankton ingestion half-saturation constant	k_G	mmol N m^{-3}	1	0.05	3
zooplankton assimilation efficiency	β		0.75	0	1
zooplankton excretion rate	μ	d^{-1}	0.1	0	0.5
zooplankton mortality parameter	ϕ_Z	(mmol N m^{-3} d)$^{-1}$	0.2	0	0.3
export fraction of zooplankton faeces	ϵ		0.33	0	1

The prior 'expected' values and prescribed ranges used for the 12 free parameters are given Table B.1. Initial concentrations are $N = 1$ mmol N m^{-3}, $P = 0.02$ mmol N m^{-3} and $Z = 0.002$ mmol N m^{-3}. All predicted values used in the cost function are taken from the second year of the model integration. The sensitivity of this model output to initial concentrations is considered to be negligible in the context of this study.

C Robustness to choice of initial station pair

A series of experiments were performed, testing the robustness of the results to differences in the initial station pair during station aggregation. Alternative pairs were selected from the set of station pairs best satisfied by parameter vectors in the initial search set. These are the pairs $\boldsymbol{H}^2_{\mathrm{BEST}}(\vec{p})$ for all $\vec{p}$ in the initial search set which give finite cost deviations for more than one station. In each experiment, only independent alternative station pairs were selected as these were considered most likely to produce different aggregation results.

In the present study, three applications of the calibration algorithm were required: one to the North Atlantic domain in its entirety and one to each of the provinces in the accepted split-domain calibration. Normally, each application of the calibration algorithm requires one application of the station aggregation procedure to the whole domain and another to each of the potential complementary provinces identified. The latter applications occur as the split-domain calibration procedure is invoked for each province indicator group. The robustness experiments involved two extra applications of the aggregation procedure, from step 2 onwards, in all instances where it was invoked, except in cases where there were less than 3 independent pairs available. The independent station pairs with the 2nd and 3rd lowest group maximum cost

Table C.1
Summary of calibration results for the North Atlantic domain

	Median validation cost	r.m.s. residuals $r(\log \mathcal{C})$	r.m.s. residuals $r(\mathcal{N})$	Domain or province latitude range (°N)	No. of stations in calibration set	Maximum coverage by station aggregation	Calibration group coverage (if different)
Calibration 1A	13.13	2.88	1.70	25-70	15	9(60%) or 11(73%)	15(100%)
Calibration 1B	12.77	2.80	1.72	25-70	15	13(87%)	–
Calibration 2A	9.84*	2.37	1.51	25-45 45-70	6 9	5(83%) 9(100%)	4(67%) –
Calibration 2B	10.66*	2.40	1.69	25-40 40-70	5 10	3(60%) 10(100%)	– –
Calibration 2C	11.69*	2.72	1.52	25-70 50-60	11 4	9(82%) 4(100%)	– –
Calibration 2D	12.14*	2.82	1.48	30-70 45-60	9 6	7(78%) 6(100%)	– –
Calibration 2E	12.96	2.89	1.61	25-60 30-70	6 9	5(83%) 8(89%)	– –
Calibration 2F	11.98*	2.74	1.55	25-70 45-60	11 4	9(82%) 4(100%)	– –
Calibration 2G	11.51*	2.67	1.55	25-45 30-70	4 11	4(100%) 10(91%)	3(75%) –
Calibration 2H	10.97*	2.34	1.91	30-70 45-60	9 6	7(78%) 6(100%)	3(33%) –

The validation cost for Calibration 1B is significantly lower (at 95%) than the full 15 station calibration (Calibration 1A). Validation costs for split-domain calibrations which are significantly lower (at 95%) than that for the best whole-domain calibration (Calibration 1B) are marked *. Coverage of the domain or province is expressed in terms of the number of calibration stations and, in brackets, the proportion of the domain or province this represents. The two alternative values for the maximum coverage by station aggregation for Calibration 1A are associated with different initial station pairs which produce the same final calibration result.

deviations were chosen.

The full set of whole-domain and split-domain calibration results obtained by applying the calibration algorithm to the North Atlantic domain is summarized in Table C.1. This includes the results for all of the alternative initial station pairs; those for the station pairs with the lowest group maximum cost

Table C.2
Robustness experiments for the North Atlantic domain

Rank	Station pair	Agg. penalty	Whole-domain calibration	Size of province indicator group $\boldsymbol{G}_A$	Size of calibration set $\boldsymbol{D}_{Bpot}$	Rank in $\boldsymbol{D}_{Bpot}$	Staion pair	Agg. penalty	Split-domain calibration
1	0711, 0812	62.5	1A	3	10	1	1113, 1419	6.1	2B
						2	1013, 1218	10.5	2B
						3	1116, 1318	5.2	2B
				4	9	1	} as above		2A
						2			2A
						3			2A
				9	4	1	1113, 1116	6.6	2C
						2	1214, 1216	9.5	2C
2	1113,1419	6.1	1B	8	7	1	0711, 0812	62.5	2A
						2	0709, 0810	7.0	2A
						3	0912, 1013	15.1	2A
3	1013, 1218	10.5	1A	4	11	1	0711, 0812	62.5	2D
						2	1113, 1419	6.1	2F
						3	0709, 1318	10.0	2H
				7	8	1	0711, 0812	62.5	2A
						2	1113, 1214	7.3	2A
						3	0709, 0810	7.0	2A
				8	7	1	1113, 1214	7.3	2E
						2	0709, 0812	7.0	2G
4	1116,1318	5.2	1B	–	–	–	–	–	–
5	0709, 0810	7.0	1B	–	–	–	–	–	–

Province indicator groups were not analyzed for the 4th and 5th station pairs.

deviations, shown in Table 2, are duplicated here for completeness. Table C.2 shows the experiments from which the calibration results were obtained. Each application of the station aggregation procedure was repeated from step 2 with the alternative initial station pairs. These are ranked in increasing order of their group maximum cost deviations $U_{\mathrm{MAX}}(\vec{p}, \boldsymbol{H}^2_{\mathrm{BEST}}(\vec{p}))$. Inspection of the pairs' aggregation penalties, determined after they were selected, shows that ranking by aggregation penalties would be different. If the rank is intended to reflect the degree to which the station pairs are satisfied by their own optimal parameter vectors, then this discrepancy implies a ranking error. Ranking error should be less likely for groups of more than two stations

because these larger groups are selected using parameter vector search sets with more pre-conditioning. However, an analysis of aggregation penalties for alternative groups of larger size is beyond the scope of this paper.

For each alternative station pair, the result of the applying the whole-domain calibration procedure is tabulated, followed by the results of applying the split-domain calibration procedure for each of the emerging province indicator groups, using each alternative station pair in the potential complementary province. The province indicator groups are distinguished by the number of stations aggregated. The size of the potential complementary province for each is then given in terms of the number of stations in its calibration set, to which the station aggregation procedure is applied.

Each of the 3 alternative station pairs in the whole-domain aggregation for the North Atlantic domain produced different results in terms of province indicator groups, giving a total of 7 such groups. When only the highest ranked pair in the potential complementary province's calibration set was chosen, this led to 5 different split-domain calibration results, of which 2 were not identified in the standard application of the method (Calibration 2D and Calibration 2E). However, for all 3 North Atlantic domain pairs, the split-domain calibration for at least one of the province indicator groups produced Calibration 2A and none of the other results improved on this. Calibration 2A therefore appears to be a robust result.

The use of alternative station pairs in the potential complementary province aggregation affected the results in only 2 out of the 8 split-domain calibrations. This did produce 3 additional calibration results, but again all have higher costs than Calibration 2A. One of these results, Calibration 2H, is the same as Calibration 2D in terms of the geographical extent of its provinces but has a different calibration group. The new group gives a lower validation cost despite its low coverage of 33%.

Although the accepted calibration remained the same throughout, the experiments did reveal an improved whole-domain calibration for the North Atlantic domain. This calibration, referred to as Calibration 1B (Table C.1), included 13 out of the 15 calibration stations. The excluded stations are Station 0606 (27.5°N 72.5°W) and Station 0711 (32.5°N 47.5°W): the two stations for which no winter-time nitrate estimates are available. The median cost for Calibration 2A is 23% lower than that for Calibration 1B, compared with the 25% improvement it represents over Calibration 1A, the result based on the full calibration set. To follow up this result and explore the robustness of the whole-domain calibration more fully, two more independent station pairs were tried: those with the 4th and 5th lowest group maximum cost deviations. 3 out of the 5 alternative station pairs produced Calibration 1B. These were also the 3 pairs with the lowest aggregation penalties. The other 2 pairs produced

Table C.3
Summary of calibration results for the accepted northern province

	Median validation cost	r.m.s. residuals $r(\log \mathcal{C})$	r.m.s. residuals $r(\mathcal{N})$	Domain or province latitude range (°N)	No. of stations in calibration set	Maximum coverage by station aggregation	Calibration group coverage (if different)
Calibration 2A	9.70	2.16	1.74	45-70	9	9(100%)	–
Calibration 3A	10.36	2.31	1.76	45-60 50-70	4 5	4(100%) 5(100%)	– –
Calibration 3B	9.46	2.16	1.72	45-65 50-70	6 3	6(100%) 3(100%)	– –
Calibration 3C	9.49	2.21	1.62	45-60 50-70	3 6	3(100%) 6(100%)	– –

None of the split-domain calibrations have validation costs significantly lower (at 95%) than that for the whole-domain calibration (Calibration 2A). Coverage of the domain or province is expressed in terms of the number of calibration stations and, in brackets, the proportion of the domain or province this represents.

Table C.4
Robustness experiments for the accepted northern province

Rank	Station pair	Agg. penalty	Whole-domain calibration	Size of province indicator group $\boldsymbol{G}_A$	Size of calibration set $\boldsymbol{D}_{Bpot}$	Rank in $\boldsymbol{D}_{Bpot}$	Staion pair	Agg. penalty	Split-domain calibration
1	1113, 1419	6.1	2A	no groups found	–	–	–	–	–
2	1013, 1218	10.5	2A	4	5	1 2	1113, 1419 1318, 1320	6.1 4.7	3A 3A
				6	3	1	1113, 1419	6.1	3B
3	1116, 1318	5.2	2A	4	4	1 2	1113, 1419 1013, 1214	6.1 10.5	3B 3B
				5	3	1	1113, 1214	7.3	3C

results which did not significantly improve on that for the full calibration set. In both cases, the maximum coverage of the domain by station aggregation was rather low (60% and 73%).

Application of the calibration algorithm to the accepted northern province,

with 3 alternative station pairs, also produced different results in terms of potential calibration groups for each pair (Tables C.3 and C.4). Whereas the best station pair had not produced any province indicator groups, the extra pairs gave a total of 4 between them, leading to 3 different split-domain calibration results. However, none of these significantly improved on Calibration 2A. The same 3 parameter vector result, Calibration 3B, was obtained in both of the experiments where one or more province indicator groups were found. This is the lowest cost 3 parameter vector calibration. Alternative station pairs were tried in 2 of the 4 potential complementary province aggregations with no effect on the results. The calibration algorithm was applied to the accepted southern province with 2 alternative station pairs. Again, use of the additional pair did not affect the southern province result, which was that no province indicator groups for sub-provinces were identified.

In conclusion, the results appear fairly robust to the choice of initial station pair in the station aggregation procedure, although they may be less so for larger domains. The best calibration result obtained was not dependent on the choice of the initial pair. However, the best whole-domain calibration result was: an improved calibration result was obtained when lower ranked initial station pairs were used. This may simply be due to the poor coverage achieved by aggregation when starting from the highest ranked pair. The results suggest that the overhead of investigating alternative pairs may be justified, particularly when coverage is poor. A practical approach might be to pool sets of parameter vectors optimized for multiple pairs, for the purposes of seeking the best 3 station group, in the same way that individual station optimization results were pooled to get the initial search set for seeking the best station pair.

References

Bishop, J.K.B., McLaren, J., Garraffo, Z., Rossow, W.B., 1994. Documentation and description of surface solar irradiance data sets produced for SeaWiFS. A draft document dated 10/30/94. Lamont Doherty Earth Observatory, Columbia University.

Bleck, R., Rooth, C., Hu, D.M., Smith, L.T., 1992. Salinity-driven thermocline transients in a wind-forced and thermohaline-forced isopycnic coordinate model of the North Atlantic. Journal of Physical Oceanography 22, 1486-1505.

Brock, T.D., 1981. Calculating solar radiation for ecological studies. Ecological Modelling 14, 1-19.

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

Christian, J.R., Verschell, M.A., Murtugudde, R., Busalacchi, A.J., McClain,

C.R., 2002. Biogeochemical modelling of the tropical Pacific Ocean. I: seasonality and interannual variability. Deep-Sea Research II 49, 509-543.

Conkright, M., O'Brien, T., Levitus, S., Boyer, T.P., Antonov, J., Stephens, C., 1998. World Ocean Atlas 1998 Vol 10: Nutrients and Chlorophyll of the Atlantic Ocean. NOAA Atlas NESDIS 36. U.S. Govt. Printing Office, Washington.

Eppley, R.W., 1972. Temperature and phytoplankton growth in the sea. Fishery Bulletin 70, 1063-1085.

Evans, G.T., 1999. The role of local models and data sets in the Joint Global Ocean Flux Study. Deep-Sea Research I 46, 1369-1389.

Evans, G.T., Parslow, J.S., 1985. A model of annual plankton cycles. Biological Oceanography 3, 327-347.

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

Fasham, M.J.R., Evans, G.T., 1995. The use of optimization techniques to model marine ecosystem dynamics at the JGOFS station at 47°N 20°W. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 348, 203-209.

Fennel, K., Losch, M., Schröter, J., Wenzel, M., 2000. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. Journal of Marine Systems 28, 45-63.

Friedrichs, M.A.M., 2001. A data assimilative marine ecosystem model of the central equatorial Pacific: Numerical twin experiments. Journal of Marine Research 59, 859-894.

Friedrichs, M.A.M., 2002. Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the central equatorial Pacific Ocean. Deep-Sea Research II 49, 289-319.

Glover, D.M., Brewer, P.G., 1988. Estimates of winter-time mixed layer nutrient concentrations in the North Atlantic. Deep-Sea Research 35, 1525-1546.

Gregg, W.W., 2001. Tracking the SeaWiFS record with a coupled physical/biogeochemical/radiative model of the global oceans. Deep-Sea Research II 49, 81-105.

Gunson, J., Oschlies, A., Garçon, V., 1999. Sensitivity of ecosystem parameters to simulated satellite ocean color data using a coupled physical-biological model of the North Atlantic. Journal of Marine Research 57, 613-639.

Hemmings, J.C.P., Srokosz, M.A., Challenor, P., Fasham, M.J.R., 2003. Assimilating satellite ocean colour observations into oceanic ecosystem models. Philosophical Transactions of the Royal Society of London. Series A, Mathematical, Physical and Engineering Sciences 361, 33-39.

Hurtt, G.C., Armstrong, R.A., 1996. A pelagic ecosystem model calibrated with BATS data. Deep-Sea Research II 43, 653-683.

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

Jia, Y.L., 2000. Formation of an Azores Current due to Mediterranean overflow

in a modeling study of the North Atlantic. Journal of Physical Oceanography 30, 2342-2358.

Kraus, E.B., Turner, J.S., 1967. A one-dimensional model of the seasonal thermocline. II: the general theory and its consequences. 19, 98-105.

Levitus, S., 1982. Climatological Atlas of the World Ocean. NOAA Professional Paper 13. U.S. Govt. Printing Office, Washington.

Longhurst, A., 1998. Ecological Geography of the Sea. Academic Press, San-Diego.

Losa, S.N., Kivman, G.A., Ryabchenko, V.A., submitted. Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data? Journal of Marine Systems.

Matear, R.J., 1995. Parameter optimization and analysis of ecosystem models using simulated annealing: A case study at Station P. Journal of Marine Research 53, 571-607.

McClain, C.R., Cleave, M.L., Feldman, G.C., Gregg, W.W., Hooker, S.B., Kuring, N., 1998. Science quality SeaWiFS data for global biogeochemical research. Sea Technology 39, 10-16.

McKay, M.D., Conover, W.J., Beckman, R.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239-245.

Moyer, K.A., Weller, R.A., 1997. Observations of surface forcing from the subduction experiment: A comparison with global model products and climatological datasets. Journal of Climate 10, 2725-2742.

Oschlies, A., 2001. Model-derived estimates of new production: new results point towards lower values. Deep-Sea Research II 48, 2173-2197.

Oschlies, A., Garçon, V., 1998. Eddy-enduced 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. Global Biogeochemical Cycles 14, 499-523.

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

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992. Numerical Recipes in C: the Art of Scientific Computing. Cambridge University Press, Cambridge.

Prunet, P., Minster, J.F., Echevin, V., Dadou, I., 1996a. 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. Global Biogeochemical Cycles 10, 139-158.

Prunet, P., Minster, J.F., Ruiz-Pino, D., Dadou, I., 1996b. Assimilation of surface data in a one-dimensional physical-biogeochemical model of the surface ocean. 1. Method and preliminary results. Global Biogeochemical Cycles 10, 111-138.

Rossow, W.B., Schiffer, R.A., 1991. ISCCP cloud data products. Bulletin of

the American Meteorological Society 72, 2-20.

Sarmiento, J.L., Monfray, P., Maier-Reimer, E., Aumont, O., Murnane, R.J., Orr, J.C., 2000. Sea-air CO_2 fluxes and carbon transport: a comparison of three ocean general circulation models. Global Biogeochemical Cycles 14, 1267-1281.

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 Biogeochemical Cycles 7, 417-450.

Sathyendranath, S., Longhurst, A., Caverhill, C.M., Platt, T., 1995. Regionally and seasonally differentiated primary production in the North Atlantic. Deep-Sea Research I 42, 1773-1802.

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

Siegel, S., Castellan, N.J., 1988. Nonparametric Statistics for the Behavioural Sciences. McGraw Hill, New York.

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 Research II 48, 1733-1768.

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. Journal of Marine Systems 16, 51-68.