Available online at www.sciencedirect.com
ScienceDirect
Journal of Marine Systems 68 (2007) 381–404
JOURNAL OF MARINE SYSTEMS
www.elsevier.com/locate/jmarsys

Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 2. Chlorophyll-a, nutrients and SPM

J. Icarus Allen [a,*], Jason T. Holt [b], Jerry Blackford [a], Roger Proctor [b]

[a] *Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth, PL1 3DH, UK*
[b] *Proudman Oceanographic Laboratory, Joseph Proudman Building, 6 Brownlow Street, Liverpool, L3 5DA, UK*

Received 27 April 2006; received in revised form 18 December 2006; accepted 4 January 2007
Available online 12 January 2007

Abstract

Marine systems models are becoming increasingly complex and sophisticated, but far too little attention has been paid to model errors and the extent to which model outputs actually relate to ecosystem processes. Here we describe the application of summary error statistics to a complex 3D model (POLCOMS-ERSEM) run for the period 1988–1989 in the southern North Sea utilising information from the North Sea Project, which collected a wealth of observational data. We demonstrate that to understand model data misfit and the mechanisms creating errors, we need to use a hierarchy of techniques, including simple correlations, model bias, model efficiency, binary discriminator analysis and the distribution of model errors to assess model errors spatially and temporally. We also demonstrate that a linear cost function is an inappropriate measure of misfit. This analysis indicates that the model has some skill for all variables analysed. A summary plot of model performance indicates that model performance deteriorates as we move through the ecosystem from the physics, to the nutrients and plankton.

Keywords: ERSEM; POLCOMS; North Sea; Nutrients; Chlorophyll-a; Model validation

1. Introduction

For several decades, there has been increasing concern that anthropogenic influences are having a detrimental effect on the ecosystem of the North Sea. A persistent problem is the enrichment of water by nutrients causing an accelerated growth of algae and higher forms of plant life to produce an undesirable disturbance to the balance of organisms present in the water and to the quality of the water concerned. We need to quantify issues of scale and natural variability to understand and manage human impacts on ecosystems successfully (Hardman-Mountford et al., 2005). It is also essential to be able to separate anthropogenic impacts from natural fluctuations. Regional-scale marine ecosystem models are tools with which we can potentially quantify this range of variability, and its causes, thus underpinning a marine management.

Recent reviews of the current state of coupled hydrodynamic-ecosystem modelling of the North West European Shelf (Jones, 2002; Moll and Radach, 2003 and references within) describe a wide range of modelling approaches from simple NPZ models to complex ecosystem models, with some including

* Corresponding author. Tel.: +44 1752 633468; fax: +44 1752633101.
E-mail address: jia@pml.ac.uk (J.I. Allen).

0924-7963/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jmarsys.2007.01.005

explicitly resolved benthic processes. A more recent review of the validation of these model systems (Radach and Moll, 2006) concluded that many model systems are capable of reproducing observations of the state variables correctly to within an order of magnitude, but that most of the models still need to be evaluated more intensively before their predictive potential can be judged.

A systematic analysis of the performance of 153 biological models that include plankton demonstrated that the efforts over the last decade to increase the level of biological detail and spatial complexity, and to explore longer simulation periods, have not led to a systematic or demonstrable improvement in model performance (Arhonditsis and Brett, 2004). They found that only 47% of the models assessed had any validation and only 30% determined some measure of goodness of fit. It would seem to be a basic requirement that before any model can be used for either scientific or policy application with any confidence an assessment of their accuracy and predictive capability is required.

Many marine management procedures involve assessing whether or not thresholds have been exceeded, for example the OSPAR common comprehensive procedure for eutrophication has specific levels to indicate eutrophication risk (OSPAR, 2003). For example, 'elevated' levels of winter dissolved inorganic nitrogen and/or phosphate concentrations are defined as a concentration of 50% above a salinity related and/or region specific background concentration; while for chlorophyll-a, 'elevated' levels are defined as a concentration of 50% above a spatially defined (offshore)/historical background concentrations. Consequently, there is a clear imperative to understand how well management models can resolve such thresholds.

The POLCOMS-ERSEM model system is a state-of-the-art coupled 3D hydrodynamic-ecosystem model for shelf seas. For this study, it has been applied to the North West European Continental Shelf on a ~7 km grid. The model is currently being evaluated within an operational framework using operationally available high-resolution atmospheric and lateral boundary forcing, allowing hindcast and near-real time nowcast simulations to be performed (Siddorn et al., in press). It is our aim in this paper and Holt et al. (2005) to give a comprehensive overview of the uncertainties associated with as many aspects of the coupled model as possible from a single simulation period. Holt et al. (2005) describe the validation of the physics of the model along with the basic phytoplankton dynamics.

Focusing on a subset of model outputs for which we have appropriate comparative observations, we apply a combination of error statistics and correlations in order to explore relationships between model outputs and observations, and the distribution of errors within the model; it is not our purpose at this stage to concentrate on the causes of model errors and possible solutions. In this paper, we discuss model performance when simulating chlorophyll-a, nutrients, and suspended particulate matter (SPM). We have two goals, firstly to propose a set of metrics to benchmark model performance against which we can assess the success of future model developments and secondly to use decision theory (receiver operator characteristics) to determine the ability of the model to discriminate thresholds.

A third paper in this series describes validation of these simulations against continuous plankton recorder (CPR) survey data (Lewis et al., 2006).

2. Model description

2.1. Data sets — simulations and observations

The medium resolution continental shelf (MRCS) model is a hindcasting/forecasting system, developed by Proudman Oceanographic Laboratory, Plymouth Marine Laboratory and The Met Office. It is based on a coupled 3D hydrodynamic and ecosystem model (POLCOMS-ERSEM; Allen et al., 2001; Holt et al., 2004), set up on a 1/10° longitude by 1/15° latitude horizontal grid (~7 km resolution) with 20-s levels (Song and Haidvogel, 1994) in the vertical and boundaries following the North-West European Continental Shelf break (approximately along the 200 m isobath, except for the Norwegian Trench). Boundary forcing for temperature, salinity, currents and sea surface elevation is obtained from a 1/6° longitude by 1/9° latitude (~12 km) Atlantic Margin Model, which is nested in the Met Office's FOAM system (Bell et al., 2000). An averaged annual cycle is used for boundary conditions since the operational system has not simulated the period of interest here (discussed below). The model includes the density evolving physics of POLCOMS (Holt and James, 2001) and a size-fractionated SPM submodel (Holt and James, 1999), coupled with the state-of-the-art biogeochemical processes of ERSEM (Baretta-Bekker et al., 1998; Blackford et al., 2004); Fig. 1 is a schematic of the pelagic model. We use a generic parameter set which was devised by fitting to data at 6 diverse stations: well mixed and a stratified North Sea station, oligotrophic eastern and western Mediterranean sea stations and upwelling and oligotrophic tropical Arabian Sea sites;

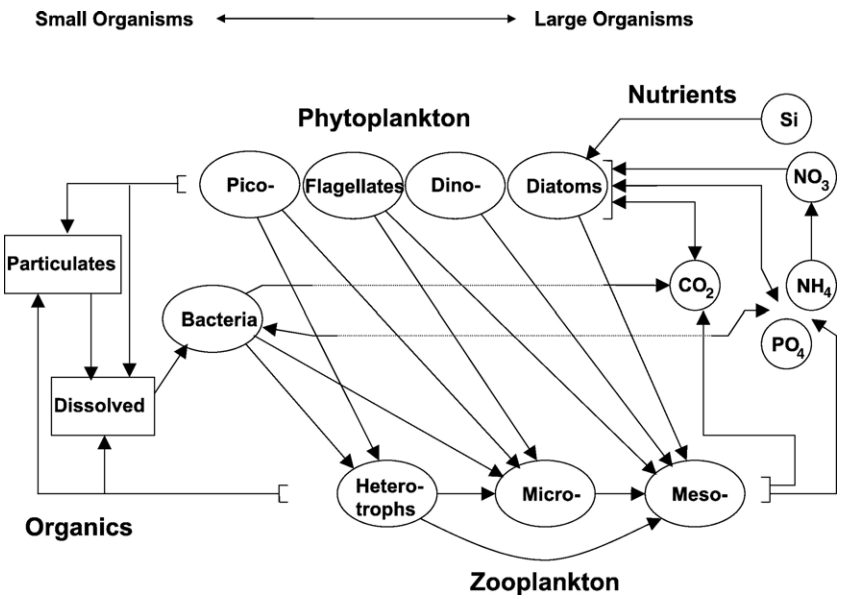

Fig. 1. A schematic diagram of the functional groups and linkages in the pelagic components of the ERSEM model.

this process and the resultant simulations are described by Blackford et al. (2004). The model is spun up using 1988 forcing then runs forward for 1988 and 1989, full details of the model experiment are given by Holt et al. (2005).

The North Sea community project (NSP, Charnock et al., 1994) collected a wealth of observational data from the southern North Sea in 1988 and 1989. Fig. 2 shows the model domain and the area sampled during the NSP, on which our analyses focus. Data (available from British Oceanographic Data Centre), including temperature, salinity, chlorophyll-a, nitrate, phosphate, ammonia, silicate and suspended sediment, were collected at ~120 CTD (conductivity, temperature and depth) and water-sampling stations during each of 16 monthly cruises between August 1988 and October 1989 (~1600 stations in total); these were preceded by a preliminary cruise in May 1988. Naturally, there are large variations in the quantity and quality of data for each variable, however, there are data to verify many variables in the ecosystem model, and all elements of the physics model apart from the turbulence variables. Holt et al. (2005) have previously reported a rigorous validation of the non-biological components of these simulations. They indicate that the model has a simulation skill for temperature, salinity, currents, tidal components, and potential energy anomaly, where they define skill as a quantifiable measure of agreement between model and observations.

In this work we use the following data sets for the analyses: temperature (28,595 measurements); salinity (28,490); chlorophyll-a (24,820); total sediment (23,645); oxygen (20,833); nitrate (4467); phosphate (4856); silicate (4818); ammonia (3532); 1% light depth (925). In the summary plot of model performance, we also consider dissolved oxygen, net primary production (Joint and Pomroy, 1993), semi-diurnal (M_2) tides, potential energy anomaly and daily mean residual velocities (Holt et al., 2005). The CTD profiles from the stations provide three dimensional distributions of conductivity (for salinity), temperature, depth, dissolved oxygen, transmittance (for suspended sediment concentration), fluorescence (for chlorophyll) and irradiance (from which the 1% light penetration depth was estimated). Water bottle samples were collected on almost all CTD casts, usually with bottles being fired at the bottom, middle and top of a cast. In order to calibrate the CTD sensors, temperatures were obtained from reversing thermometers and salinity determinations and dissolved oxygen measurements were made. Spectrophotometric chlorophyll and phaeopigment determinations were carried out, the chlorophyll values for a cruise were used to calibrate the CTD fluorometer. Sediment content was determined by filtration and the values used to calibrate the transmissometer. Nutrients (nitrate, nitrite, silicate, phosphate and ammonium) were determined from water bottle samples using an autoanalyser. Primary productivity was investigated on each survey cruise, the uptake of Carbon-14 being measured in an

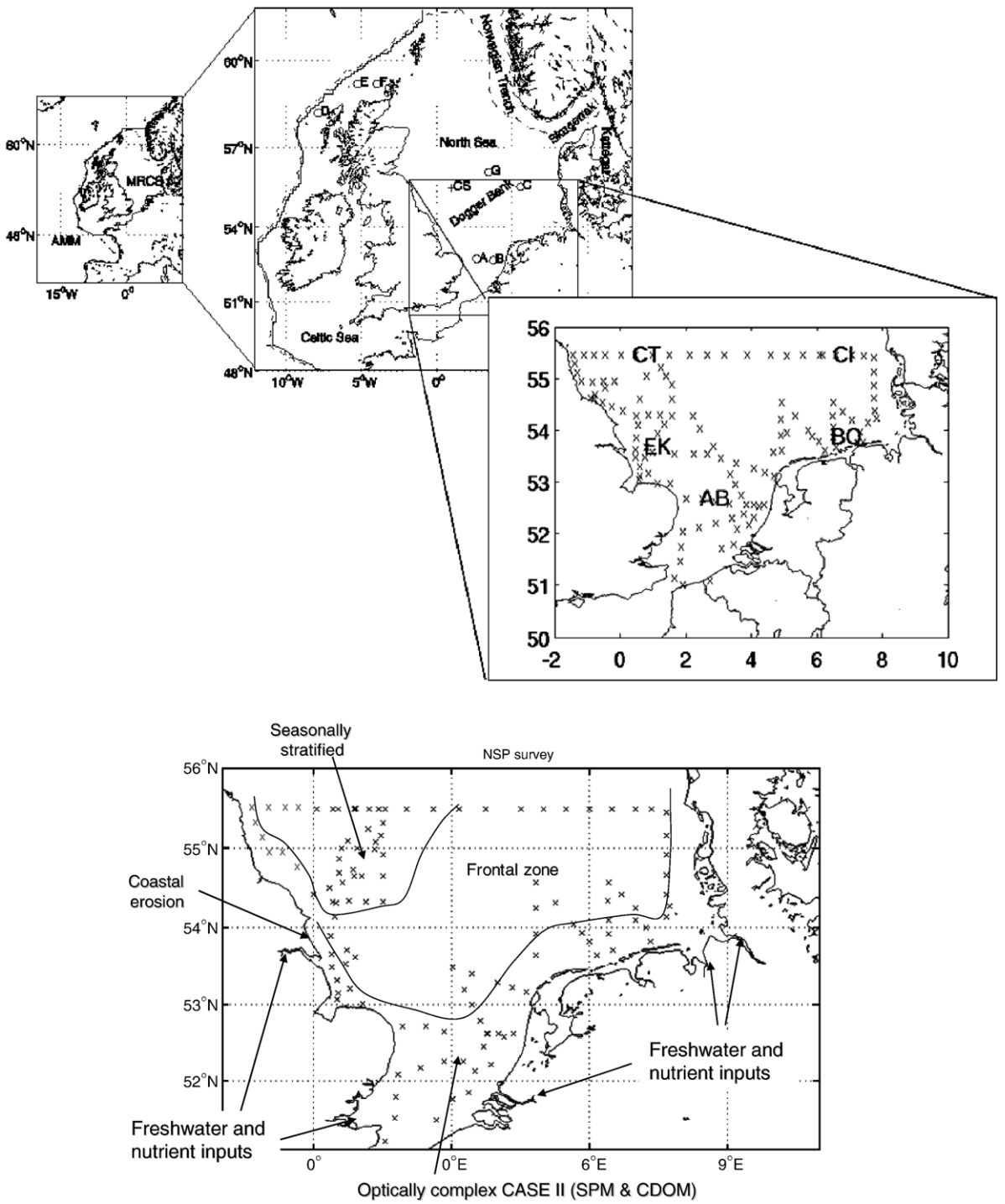

Fig. 2. The Medium Resolution Continental Shelf (MRCS) model domain. Also shown is the sub-domain of the MRCS analysed in this paper, the crosses indicate the NSP survey stations. The stations shown in Figs. 5–7 and the QSA are marked.

on-deck incubator. Water bottle samples from a pre-dawn CTD cast were taken from six depths (usually 1, 3, 7, 15, 20 and 30 m) instead of the more usual top, middle and bottom samples. Triplicate samples were then incubated at six simulated depths for a 24-hour period (Joint and Pomroy, 1993).

2.2. Model error calculations — measures of reliability?

An assessment of the confidence we can place on model results (known as model validation) must take into account the complex combination of model and observational uncertainties. Model errors derive from inaccuracies in process descriptions, parameterisation, initialisation and forcing functions. Errors in observations arise from basic measurement error, inappropriate scales of sample distribution (for example data that are over influenced by small-scale processes) or lack of replication in highly heterogenous systems and issues of methodology. A crucial issue is balancing precision (how well does the model fit each data point) with trend (i.e. how well the model reproduces the observed seasonal cycles). For example, even when the trend is well reproduced small differences in the timing of an event can lead to large errors in precision. The choice of error statistic (model efficiency, bias, cost function and so on) is crucial and a comprehensive validation process must consider several. In this paper we focus on precision and we consider the direct like-with-like comparison of model and data in space and time. We have deliberately chosen this rather unforgiving comparison to enable us to assess the model's short-term forecast potential. We have used the following eight criteria to assess model performance. In a companion paper (Lewis et al., 2006), we investigate the trends.

The *Nash Sutcliffe Model Efficiency* (Nash and Sutcliffe, 1970) of a model variable is a measure of the ratio of the model error to the variability of the data. It was developed to assess the performance of river catchment models, which exhibit a similar temporal variability to phytoplankton (rapid increases and decreases).

$$\mathrm{ME} = 1 - \frac{\sum_{n=1}^{N} (D_n - M_n)^2}{\sum_{n=1}^{N} (D_n - \bar{D})^2} \tag{1}$$

where D is the data, M the corresponding model estimate and the overbar indicates the mean of the data set for the chosen variable, N is the total number of model data matches and n is the nth comparison. The squaring of the error rewards a good fit and punishes a poor fit. Performance levels are categorised as follows >0.65 excellent, 0.65–0.5 very good, 0.5–0.2 good, <0.2 poor and are taken from Maréchal (2004).

The *percentage model bias* (the sum of model error normalized by the data) is given by

$$\mathrm{Pbias} = \frac{\sum_{n=1}^{N} (D_n - M_n)}{\sum_{n=1}^{N} D_n} * 100 \tag{2}$$

and gives measure of whether the model is systematically underestimating or overestimating the observations. The closer the value is to zero the better the model. Performance levels are categorised as follows |Pbias| <10 excellent, 10–20 very good, 20–40 good, >40 poor (Maréchal, 2004).

The *cost function* gives a non-dimensional value which is indicative of the "goodness of fit" between two sets of data; it quantifies the difference between model results and measurement data (see OSPAR Commission, 1998). The function is as follows:

$$\mathrm{CF} = \frac{1}{N} \sum_{n=1}^{N} \frac{|D_n - M_n|}{\sigma_D} \tag{3}$$

where σ_D is the standard deviation of the data. It is a measure of ratio of the model data misfit to a measure of the variance of the data; the closer the value is to zero the better the model. Performance criteria are generally scaled by numbers of standard deviations. Two sets of criteria have been used:

CF < 1 = very good, 1–2 = good, 2–5 = reasonable, >5 = poor; OSPAR Commission (1998).
CF < 1 = very good, 1–2 = good, 2–3 = reasonable, >3 = poor; Radach and Moll (2006).

The *skewness* of the error distribution characterizes the degree of asymmetry of a distribution around its mean.

$$\mathrm{Skew} = \frac{N}{(N-1)(N-2)} \sum_{n=1}^{N} \left(\frac{(D_n - M_n) - (\overline{D_n - M_n})}{\sigma_D} \right)^3 \tag{4}$$

Positive skewness indicates a distribution with an asymmetric tail extending toward large positive values, i.e. the model tends to make more underestimations. Negative skewness indicates a distribution within asymmetric tail extending toward values that are more negative (i.e. a greater number of large overestimations by the model). The standard error of skewness can be roughly estimated as $(6/N)^{0.5}$ (N is the number of data;

Tabachnick and Fidell, 1996) and skewness values of 2 standard errors or more (regardless of sign) can be taken to be substantially skewed. An order of magnitude analysis based on $N \sim 1000$ indicates that any of our error distributions with a skewness greater than ~ 0.15 is significantly skewed.

The *receiver operator characteristic* (ROC) curve is a graphical means of evaluating the predictive power of binary classification system as a discrimination threshold is varied; that is to say how useful a model is for a decision making process. It was devised during the Second World War as a means for radar operators to correctly identify hostile or friendly aircraft based on a radar signal, a situation where the incorrect identification of a hostile aircraft could be catastrophic. These techniques are now widely used in a number of fields, particularly medical research. Brown and Davis (2006) provide a detailed and accessible tutorial of the use of ROC curves and related metrics. We outline the metrics used in this paper below, following the nomenclature of Brown and Davis (2006).

The root is a simple yes/no decision, based on the comparison of two independent information sets (in our case observations and model) with respect to a threshold value. In a standard ROC analysis, the aim is to assess how well a test (model) can discriminate between two discrete observed outcomes (e.g. harmful algal bloom event or not; disease, no disease etc.). The decision process is illustrated by Fig. 3; there are four possible outcomes for each trial, either correctly positive (CP), correctly negative (CN), incorrectly positive (IP) and incorrectly negative (IN). We can use this approach to make an analysis of similarity of how well the model fits the data. The perfect model is one where all the points in a scatter diagram of model vs. data lie on the $x=y$ line (Fig. 3). If we set a threshold criteria (t) dividing the data into two sets and then compare it with the model using the same threshold (Fig. 3) we can assess model data similarity at that threshold, effectively assessing the model ability to discriminate that threshold. The perfect model will only give CP and CN outcomes; the more scatter there is in the model–data relationship the more IP and IN conditions will occur and the worse the model performance. By varying the threshold across the full range of observations, we obtain a non-parametric measure of the model's ability to simulate a given variable, which can be compared directly for other simulated variables.

The decision process can be further assessed by calculating the correct negative fraction (CNF) and the correct positive fraction (CPF).

$$\mathrm{CNF} = \frac{\mathrm{CN}}{\mathrm{CN} + \mathrm{IP}} \tag{5}$$

$$\mathrm{CPF} = \frac{\mathrm{CP}}{\mathrm{CP} + \mathrm{IN}} \tag{6}$$

CNF and CPF express the fraction of negative and positive events, which are correctly determined. These values are independent of the actual numbers of positive and negative events in the trials. The ROC curve is calculated by plotting CPF_i on the vertical axis and $1-\mathrm{CNF}_i$ on the horizontal axis for $i=1, k$ threshold values. A model ideal for decision making corresponds to a point in the top left hand corner of the ROC axis (i.e. CNF = 1 and CPF = 1). The top right (CPF = 1, CNF = 0) and bottom left (CPF = 0 and CNF = 1) correspond to the extremes of the decision process where every trial is always deemed either positive or negative. A completely random predictor

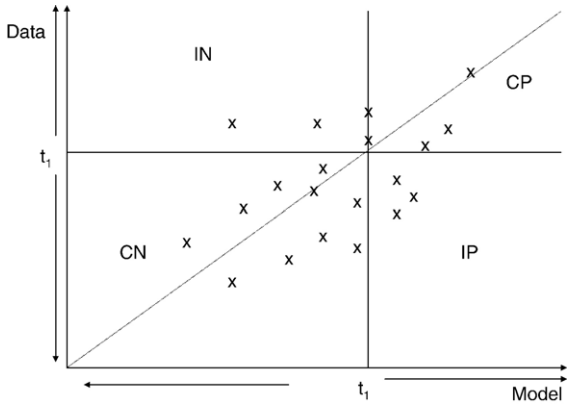

Fig. 3. Schematic diagram of the discrimination analysis.

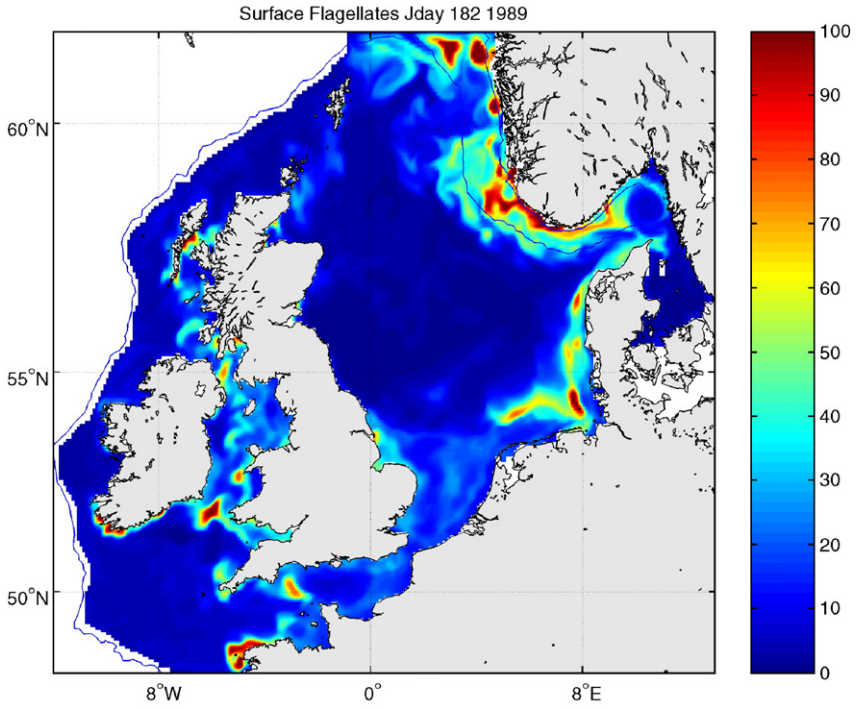

Fig. 4. Modelled spatial distribution of flagellates, 1st of July 1989.

gives a straight line at an angle of 45° from the horizontal. This is because as the threshold rises equal numbers of true and false positives occur. Results below this line suggest that the model gives consistently incorrect results.

Decisions based on CPF and CPN are estimators of probabilities of decisions contingent on events: if a positive event has occurred what is the probability I will make the correct decision. While these probabilities are useful they do not address the fundamental question, if I make a positive decision what's the probability that the decision is correct. The positive predictive value (PPV) and negative predictive value (NPV) can be expressed as (see Brown and Davis, 2006 for the theoretical background and derivation).

$$\mathrm{PPV} = \frac{\mathrm{CP}}{\mathrm{CP} + \mathrm{IP}} \tag{7}$$

$$\mathrm{NPV} = \frac{\mathrm{CN}}{\mathrm{CN} + \mathrm{IN}} \tag{8}$$

Values of PPV and NPV can range between 0 and 1, reflecting the intrinsic power of the decision; high values indicating a decision can be trusted, low values suggesting the decision should be regarded with suspicion.

The *median error* is defined as the 50th percentile of the error distribution.

The *ratio of the standard deviations of the data to model (RSD)* is given by: $\frac{\sigma_D}{\sigma_m}$ where σ is the standard deviation.

The *correlation coefficient (R)*is defined by

$$R = \frac{\sum_{n=1}^{N} \left(D_n - \bar{D}_n\right)\left(M_n - \bar{M}_n\right)}{\sqrt{\sum_{n=1}^{N} \left(D_n - \bar{D}_n\right)^2 \sum \left(M_n - \bar{M}_n\right)^2}}$$

It expresses the quality of a least squares fitting between two model and data ($R=0$ no relationship, $R=1$ perfect fit). The *square of the correlation coefficient (R^2)* expresses the percentage of the variability in data that can be accounted for by the model.

Finally, in addition to these objective criteria evidently many scientists base at least an initial, and sometimes their complete analyses of model outputs on a subjective visual comparison of plots of model against data. In order to begin to investigate how scientists

subjectively assess model outputs, and how this relates to the objective criteria described, we have conducted the following *Quantitative Subjective Analysis* experiment. A group of 16 scientists (both modellers and experimentalists) were asked to rank 15 model data comparisons visually on an arbitrary scale of 0 to 5 (0 poor, 5 excellent), using their own skill and judgement (i.e. they set the criteria). These scores have been collated to obtain a mean and standard deviation score for each graph. Additionally individual scores were correlated with the ME and Pbias associated with each model data comparison to assess the range of individual performance.

3. Results

As an example of the level of spatial detail the model produces, Fig. 4 shows the sea surface distribution of flagellate biomass on the 1st of July 1989. It demonstrates how the plankton distribution is constrained by the physical structure of the water column, in particular low biomass in the stratified central/northern North and Celtic Seas, and enhanced production along the fronts. Further examples of spatial output are published by Allen et al. (2001), Holt et al. (2004) and Holt et al. (2005).

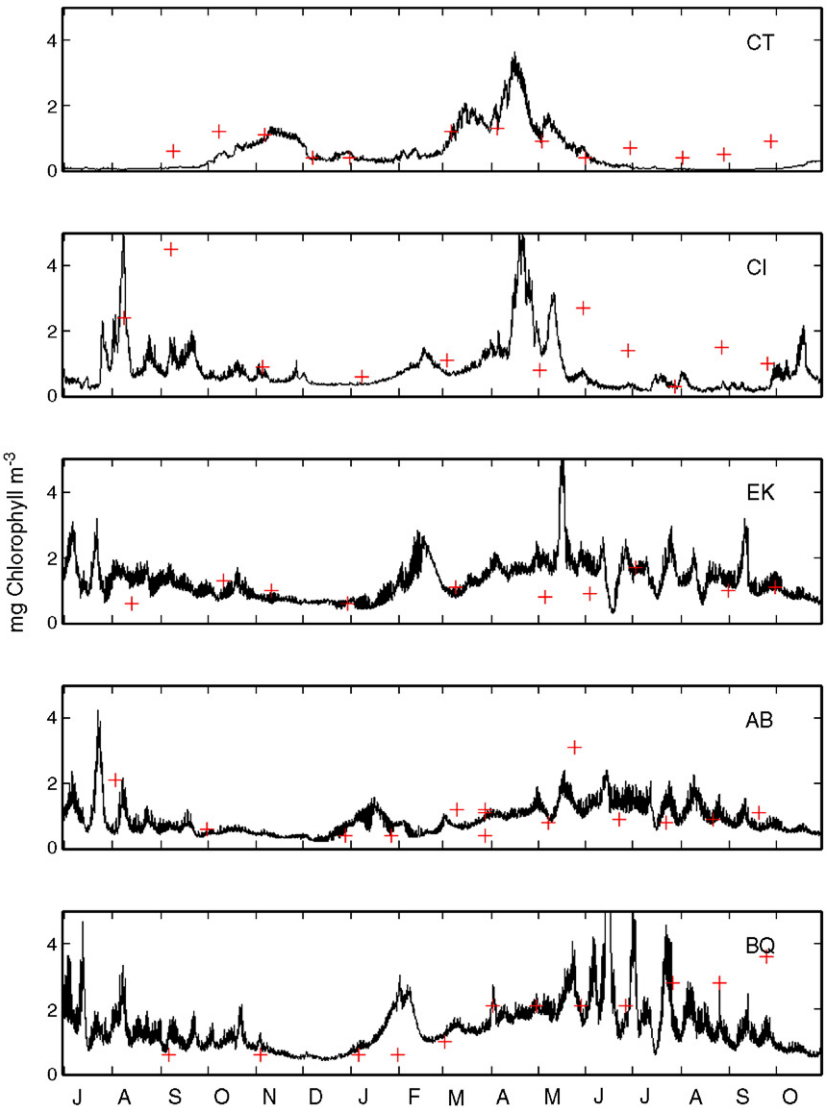

Fig. 5. Comparison of model (solid line) surface chlorophyll with data (cross) at stations CT CI, EK, AB and BQ. The mean and standard deviation of the Quantitative Subjective Analysis (QSA) score are shown.

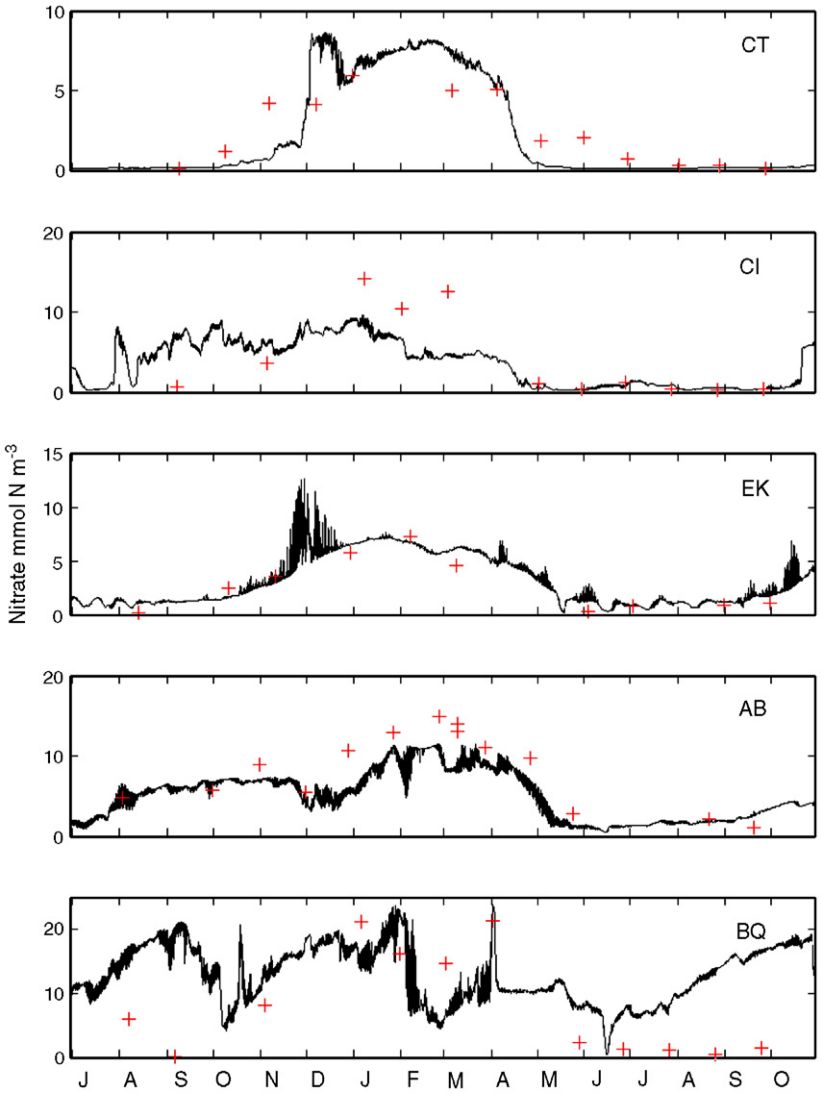

Fig. 6. Comparison of model (solid line) surface nitrate with data (cross) at stations CT, CI, EK, AB and BQ. The mean and standard deviation of the Quantitative Subjective Analysis (QSA) score are shown.

The seasonal cycles of chlorophyll, nitrate and phosphate at five stations in different parts of the southern North Sea are shown in Figs. 5–7. These stations were chosen because they are representative of the range of conditions found in the southern North Sea based on multivariate analysis (Allen et al., 2007). A visual assessment of data presented in this way relies on the subjective judgement of the evaluator as to whether the model performance is adequate or not. Simple changes to the presentation (e.g. scale choice, symbol size, line thickness) can influence the judgement of how good a simulation is. These results give a general indication that the model captures the general seasonal trends in chlorophyll, nitrate and phosphate, except at station BQ where the model fails to reproduce the observed nutrient draw down. In total, we have similar plots for 122 stations and 9 variables; consequently, it is difficult, if not impossible, to get a good feeling for the performance of the model based on a visual assessment, nor is it a suitable method of comparing the performance of model versions.

3.1. Summary statistics

A summary of basic model data fit metrics for the whole of the North Sea Project (space and time) for

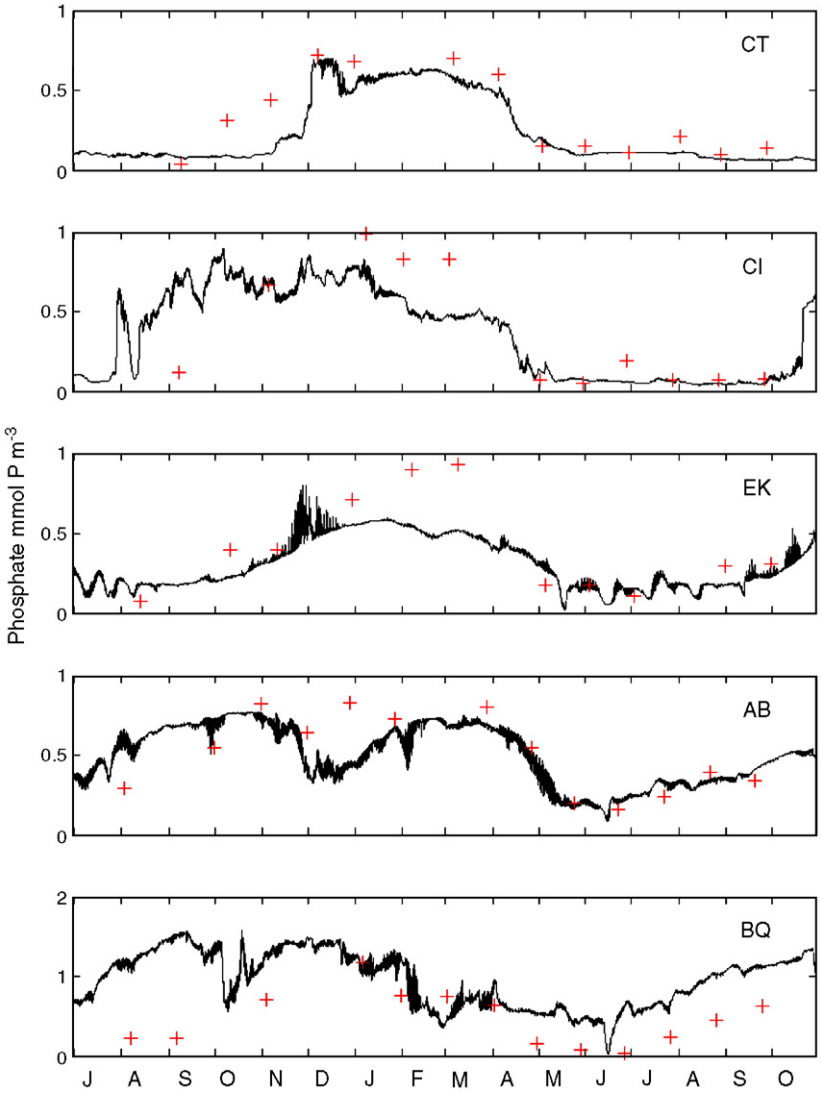

Fig. 7. Comparison of model (solid line) surface phosphate with data (cross) at stations CT, CI, EK, AB and BQ. The mean and standard deviation of the Quantitative Subjective Analysis (QSA) score are shown.

temperature, salinity, chlorophyll-a, nitrate, phosphate, oxygen, ammonia, silicate and total suspended matter (Fig. 8) indicates substantial differences between variables. The square of correlation coefficient R^2 (Fig. 8a) shows how much of the variability of the data can be reproduced by the model with the correct spatial and temporal distribution. R^2 is high (>0.65) for temperature and salinity, mid range ($0.35<>0.65$) for nitrate, oxygen and phosphate and low (<0.3) for the rest. However, all of the correlations are significant at a 95% confidence level. The distributions of model efficiency (Fig. 8b) and model bias (Fig. 8c) confirm this, the correlation coefficient reflecting the highest model efficiencies and the smallest bias. The results of the cost function (Fig. 8d) are ambiguous, having no clear correspondence with the other metrics; by the classification of OSPAR (1998) or Radach and Moll (2006), all the variables evaluated are good or very good.

3.2. Temporal propagation of errors

To investigate temporal changes in the error distribution we plot the median error value for each season (the median error) against the skew of the error (data minus model) (Fig. 9). We choose this analysis to assess whether errors are being propagated over time; increasing skew and median error is indicative of

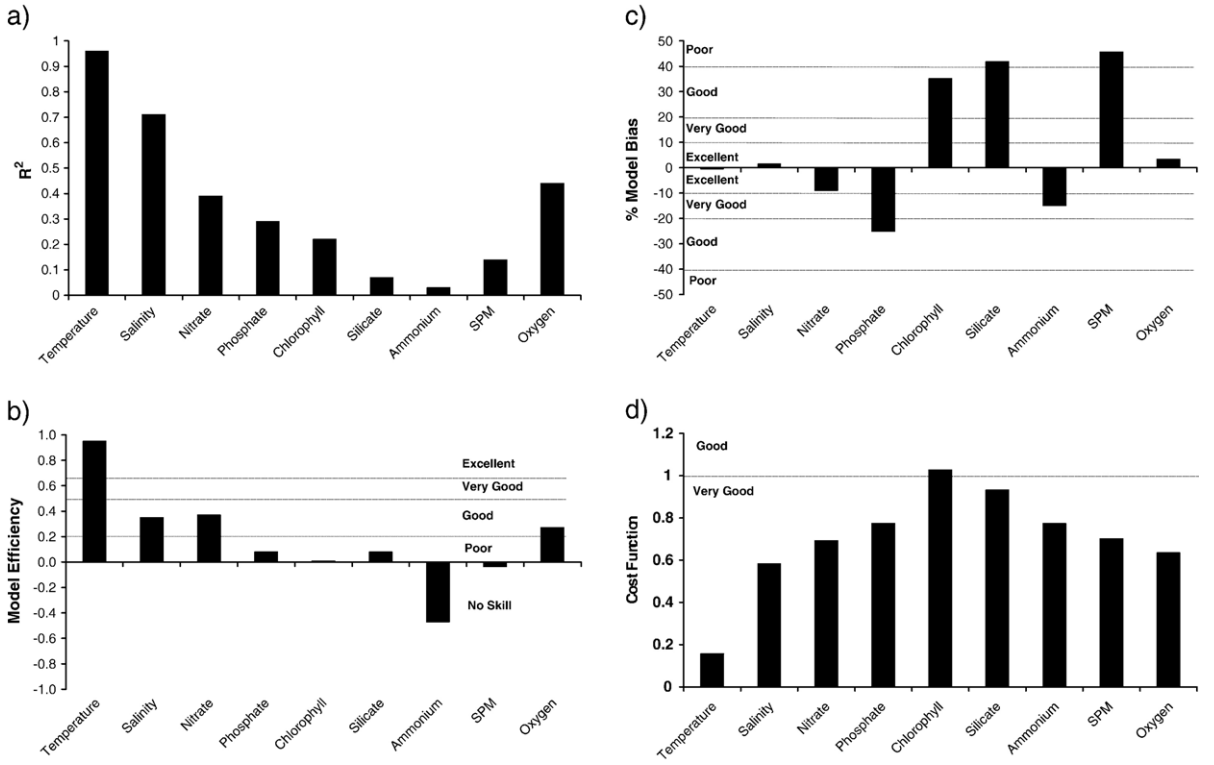

Fig. 8. Model performance summary statistics for the whole North Sea data set, (a) the square of correlation coefficient, (b) model efficiency, (c) model bias, (d) cost function.

increasing error. Temperature (Fig. 9a) shows distinct patterns; median errors are positive (underestimation) while skew is close to zero (±0.25) in winter and spring. In summer and autumn, the median error implies both a general overestimation of temperature by the model but an asymmetric distribution (positive skew) indicating that the tail of the distribution is skewed towards underestimation. The repeating pattern implies no long-term error propagation. Salinity (Fig. 9b) has the same median error (~0.5 psu, overestimation) throughout the simulation, but the skewness in the error distribution shifts from negative in winter and spring to positive in summer and autumn, possibly reflecting uncertainties in freshwater inputs and salinity stratification; too high in winter and spring, too low in summer and autumn. Nitrate (Fig. 9c) has positive skewness in autumn 88, winter 89 and spring 89 (underestimation); and negative median error, and skewness in the summer and autumn 89 (overestimation). Phosphate (Fig. 9d) errors and skew are close to zero in autumn 88 and winter 89, in spring and summer the median error indicates consistent overestimation, with an asymmetric tail biased towards underestimation in autumn 89. Chlorophyll-a (Fig. 9e) shows positive median error and skewness in all seasons except winter, implying consistent underestimation of plankton biomass. The skew and median error values are much higher in 1989 than in 1988 suggesting cumulative error propagation or more comprehensive data coverage in 1989. The skew and median errors for silicate (Fig. 9f) are always positive most notably in autumn and winter, again demonstrating consistent underestimation by the model and implying errors in either the recycling of silicate or the behaviour of diatoms in the model. Comparison of this simulation with continuous plankton recorder data (Lewis et al., 2006) indicates that the modelled diatom bloom occurs a month early. Silicate errors are larger in 1989 implying error significant propagation. Ammonia (Fig. 9g) simulations are characterized by highly positively skewed distributions. The median errors starts as positive (underestimation) in autumn 1988, and then shifts to being negative (overestimation) from winter 1989 onwards. SPM (Fig. 9h) shows large median errors and skew in the winter and spring, implying consistent underestimation during these seasons. Nutrients have the largest median errors in the winter, while chlorophyll has the largest errors in the summer; unsurprisingly nutrient

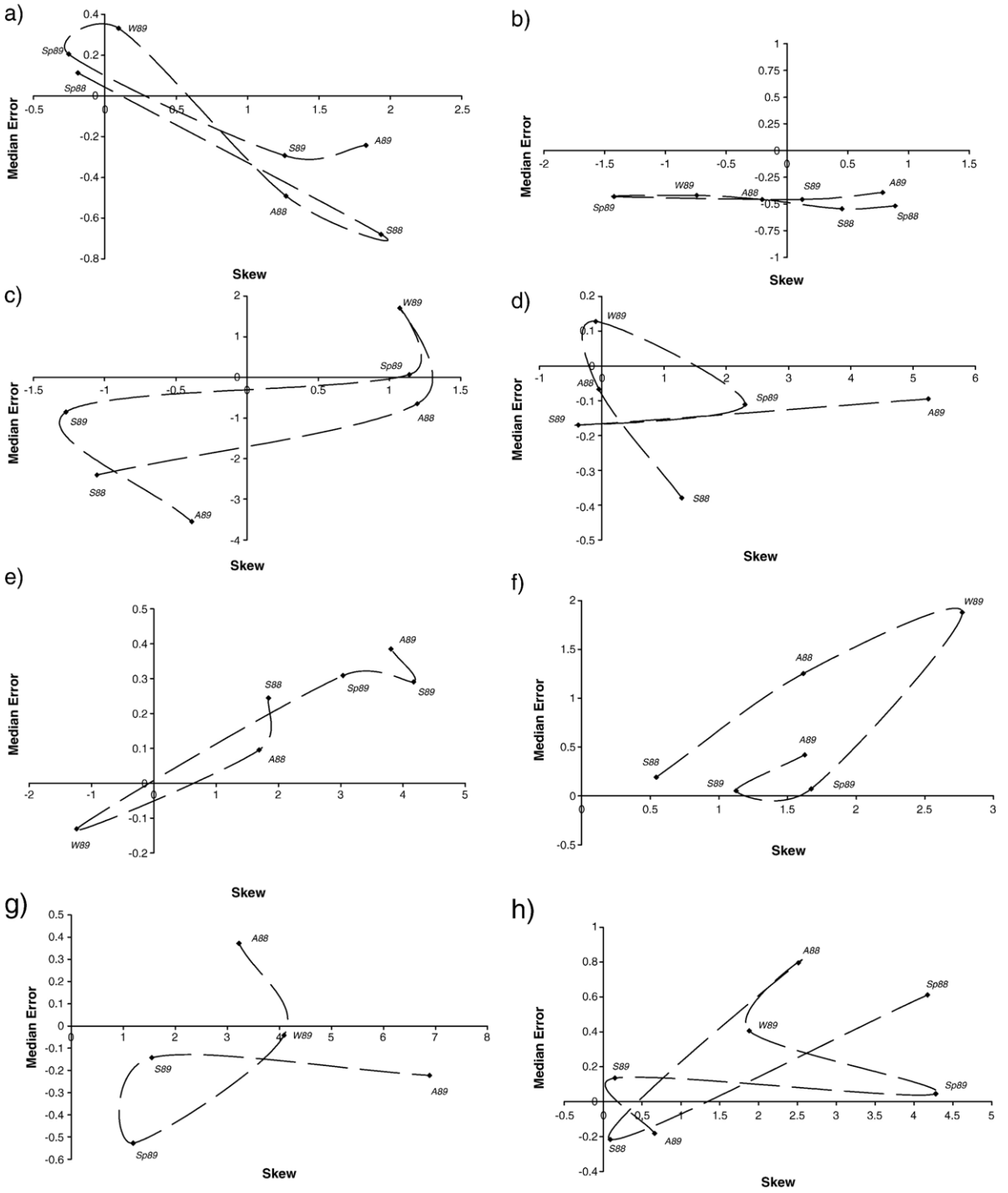

Fig. 9. Seasonal mean median model error vs. the skewness of the model error distribution for that season, (a) temperature, (b) salinity, (c) nitrate, (d) phosphate, (e) chlorophyll, (f) silicate, (g) ammonia, (h) suspended sediment. The dotted lines indicate the temporal progression. Seasons are defined as: winter (Jan–Mar), spring (Apr–Jun), summer (July–Sept), autumn (Oct–Dec). The error is calculated as model data, hence positive skew indicates an error distribution biased towards model underestimation, and negative skew vice versa. All of the skew values show a significant deviation from a normal distribution.

errors are inversely correlated to median errors of chlorophyll (R^2; NO_3=0.60; PO_4=0.41, SiO_4=0.84). The skew of silicate is also inversely correlated to the skew of chlorophyll (R^2; SiO_4=0.38), again pointing to the importance of diatoms in determining the simulation quality.

3.3. Spatial variability of errors

Tables 1 and 2 summarise the model efficiency and model bias for each variable calculated on a station-by-station basis over the period of the NSP survey and clearly show an ability to reproduce the observed variations in temperature, nitrate and phosphate. The model efficiency for salinity on a station-by-station basis is poor; this is because the variability in the observation at each station is much smaller than of the whole domain (the spatial variability is much greater than the temporal variability). The quality of simulation of chlorophyll-a, silicate ammonia and sediment is low with more than 80% of the stations classified as poor. The model bias shows that temperature and salinity simulations exhibit very low bias (99% stations good or better). The largest systematic biases are in SPM and silicate (57% and 70.5%). For the remaining variables between 30 and 50% of the stations are classified as poor.

To understand the spatial variability of the errors we have produced bubble plots of Bias and ME for each station (Figs. 10 and 11). For ME (Fig. 10), the temperature is simulated well except in the seasonally stratified NW of the domain, where errors occur due to inaccuracies in modelling mixing in strongly stratified water. Similar errors are found for nitrate, phosphate, chlorophyll, SPM and salinity, implying that errors in the thermal structure of the model may be propagated through biogeochemical variables. The other region of clear differences is the Case II waters (Fig. 2) where nitrate and (particularly) phosphate are poorly simulated, and silicate has some skill. The ME for both chlorophyll and SPM are mostly less than 0 indicating no model skill, however the simulations are better in the coastal zone than in the offshore regions. Ammonia has similar skill levels but displays the opposite pattern (better offshore, worse inshore). The simulation of salinity only shows skill in a few stations which are away from both regions influenced by stratification and major freshwater inputs.

Table 1
Percentage of NSP stations in each skill class as classified by model efficiency

% Stations	Excellent >0.65	Very good 0.65–0.5	Good 0.5–0.2	Poor <0.2
Temperature	81	3	4	12
Salinity	0	0	4	96
Chlorophyll	2	0	2	96
Nitrate	16	21	27	36
Phosphate	10	13	21	56
Silicate	3	6	9	82
Ammonium	0	0	2	98
Sediment	0	0	0	100

Table 2
Percentage of NSP stations in each skill class as classified by model bias

% Stations	Excellent <10	Very good 10–20	Good 20–40	Poor >40
Temperature	85	10	4	1
Salinity	100	0	0	0
Chlorophyll	10	22	32	38
Nitrate	17	12	37	34
Phosphate	8	11	34	47
Silicate	5	2.5	22	70.5
Ammonium	7.5	16	27.5	49
Sediment	11	13	19	57

The plots of simulation bias (Fig. 11) also demonstrate some clear patterns, particularly along the continental European coast, and the stratified NW of the region. Nitrate and phosphate are overestimated inshore and underestimated offshore, the reverse being true for chlorophyll, SPM and ammonia. The levels of bias for temperature and salinity are very low, while across the whole region silicate is systematically underestimated.

3.4. Overall model performance

Plotting R^2 against RSD (ratio of standard deviations, σ_D/σ_m) gives a simple representation of model performance (the closer R^2 and RSD are to 1 the better the fit; (Fig. 12)). This is a simplified version of the Taylor diagram, often used to assess climate model performance (Taylor, 2001). It clearly shows that the model has some degree of skill (in descending order of correlation coefficient) for temperature, tidal (M_2) elevations and currents, potential energy anomaly, salinity, oxygen, nitrate, phosphate and 1% light depth. The model clearly underestimates the variance of chlorophyll and silicate, both of these are systematically underestimated in the coastal regions of the model. We hypothesise that this is due to poor representation of the optical environment in the model and possible underestimation of the freshwater silicate loads. A bias plot for the 1% light depth (Fig. 13) shows

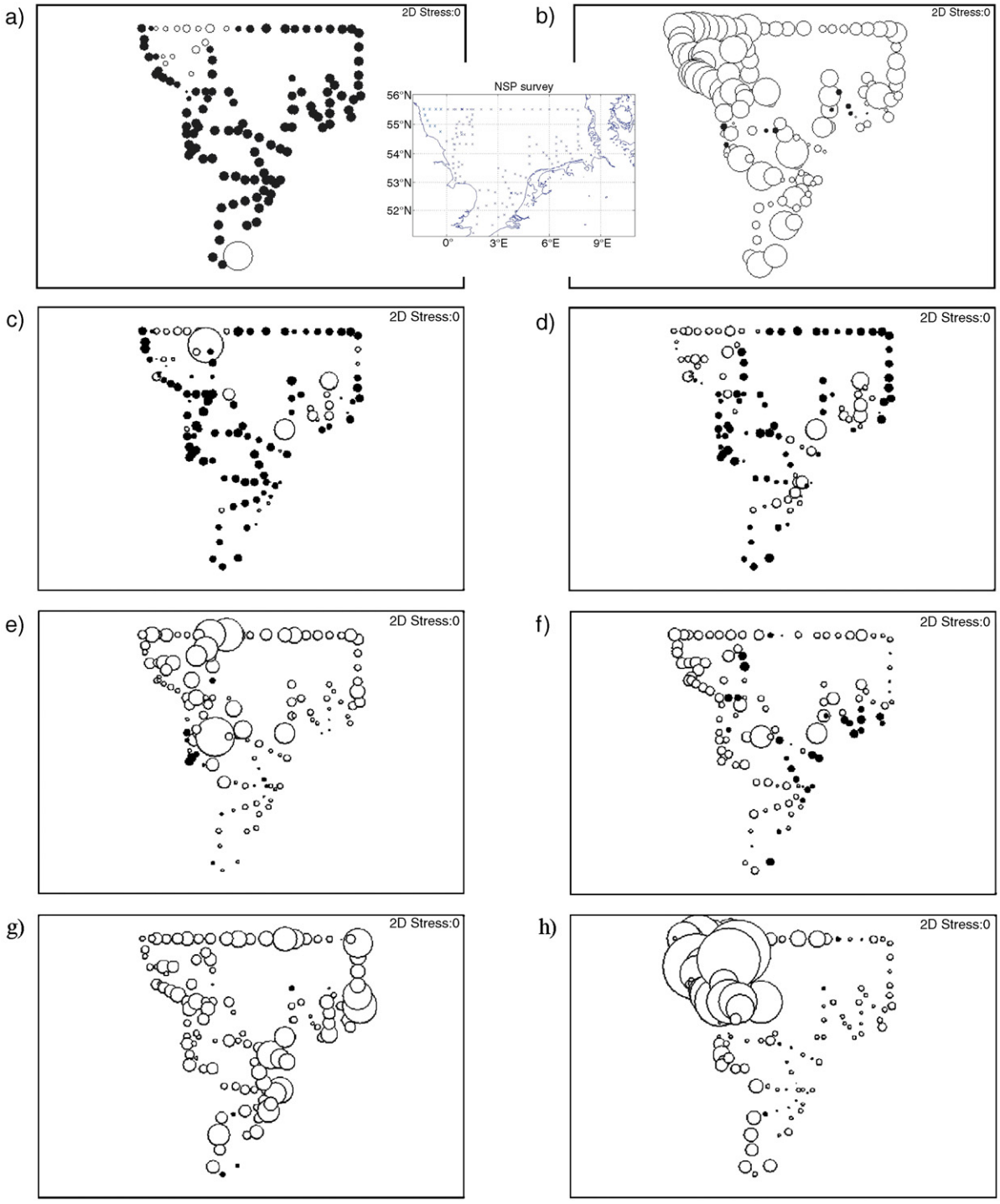

Fig. 10. Spatial distribution of model efficiency for each NSP station for the whole period of the survey, (a) temperature, (b) salinity, (c) nitrate, (d) phosphate, (e) chlorophyll, (f) silicate, (g) ammonia, (h) suspended sediment. Solid bubbles indicate ME positive, clear bubbles indicate ME negative. The size of the bubble indicates the ME value (i.e. large solid bubbles indicate the model performs well (max value = 1 the perfect model), large clear bubbles indicate poor performance). The *Bubble plots* were created using the PRIMER software (Plymouth Routines In Multivatiate Research v6, Clarke and Gorley, 2006). Non-metric multi-dimensional scaling ordination (MDS), derived from normalized Euclidean-distance matrices, are used to visualise spatial relationships between the stations to illustrate the spatial distribution of the bubbles.

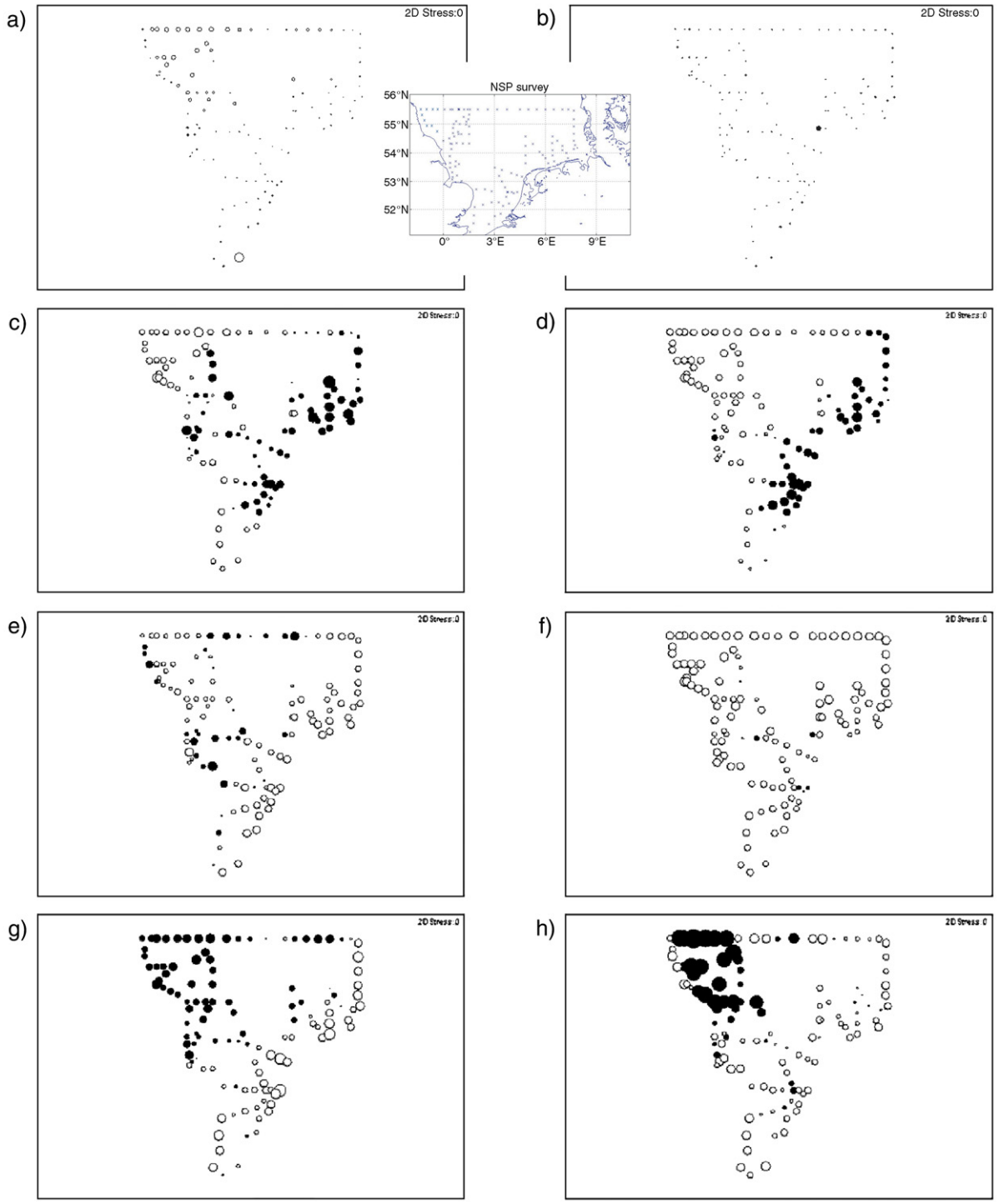

Fig. 11. Spatial distribution of model bias for each NSP station for the whole period of the survey, (a) temperature, (b) salinity, (c) nitrate, (d) phosphate, (e) chlorophyll, (f) silicate, (g) ammonia, (h) suspended sediment. Solid bubbles (see right) indicate negative bias (i.e. model overestimates), clear bubbles (see right) indicate positive bias (i.e. model underestimates). The size of the bubble indicates the size of the bias, the smaller the bubble the closer the bias is to zero.

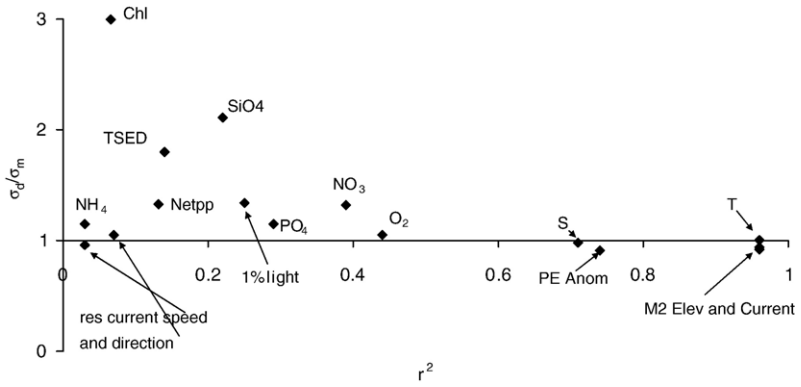

Fig. 12. A simplified (non-polar coordinate) Taylor diagram, giving an overall summary of model performance. It is a plot of the ratio of the standard deviations of data to model (*y* axis) against the square of the correlation coefficient between model and data (*x* axis).

that the model systematically overestimates the light penetration over the southern North Sea particularly in the western and southern regions. The freshwater nutrient inputs are monthly mean values used to drive the ERSEM box model (Patsch and Radach, 1997) and consequently underestimate the daily variability of the nutrient loads. The model clearly has no skill for ammonia suggesting that the parameterisations of its biological production, nitrification and denitrification processes require further work. The lateral boundary conditions exclude any residual currents or elevation, for example from a north–south density gradient. This is likely to result in an underestimate of the transport into the North Sea, as shown by Holt et al. (2005). How this might influence the ecosystem model is the subject of ongoing work.

3.5. Discriminating thresholds

Employing the ROC technique (Fig. 14) indicates that the model has some predictive skill for all variables analysed. Unsurprisingly temperature (Fig. 14a) has the most skill, followed by salinity, nitrate, silicate and phosphate (Fig. 14b, c, d and f; the curves are all above and to the left of the random 45° line). Salinity (Fig. 14b) shows high specificity but low sensitivity indicating a reduced predictive skill at small values. Chlorophyll-a demonstrates some skill at low concentrations (curve is above but close to the random line) but when the discrimination threshold is above 5 mg Chl m^{-3} (Fig. 14e) the model is random, suggesting no predictive skill during bloom periods. The ROC curves for ammonia and SPM (Fig. 14g, h)

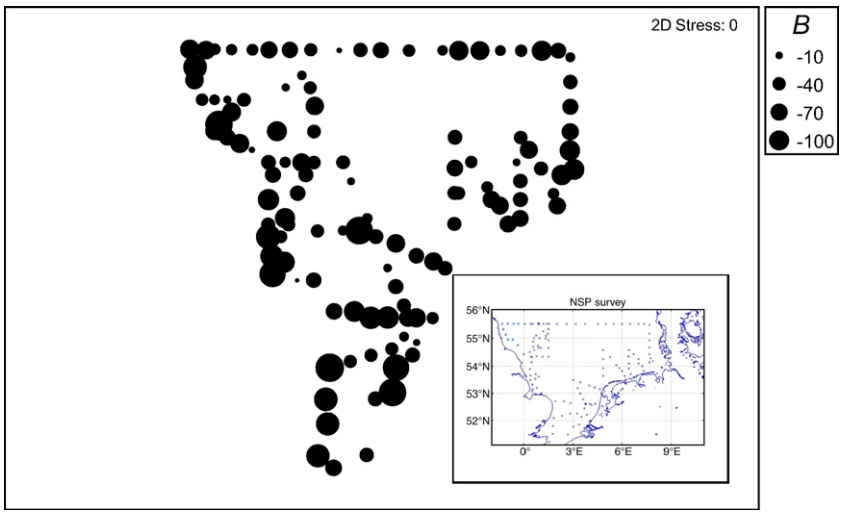

Fig. 13. Spatial distribution of model bias of 1% light for each NSP station for the whole period of the survey. Solid bubbles indicate negative bias (i.e. model overestimates). The size of the bubble indicates the size of the bias.

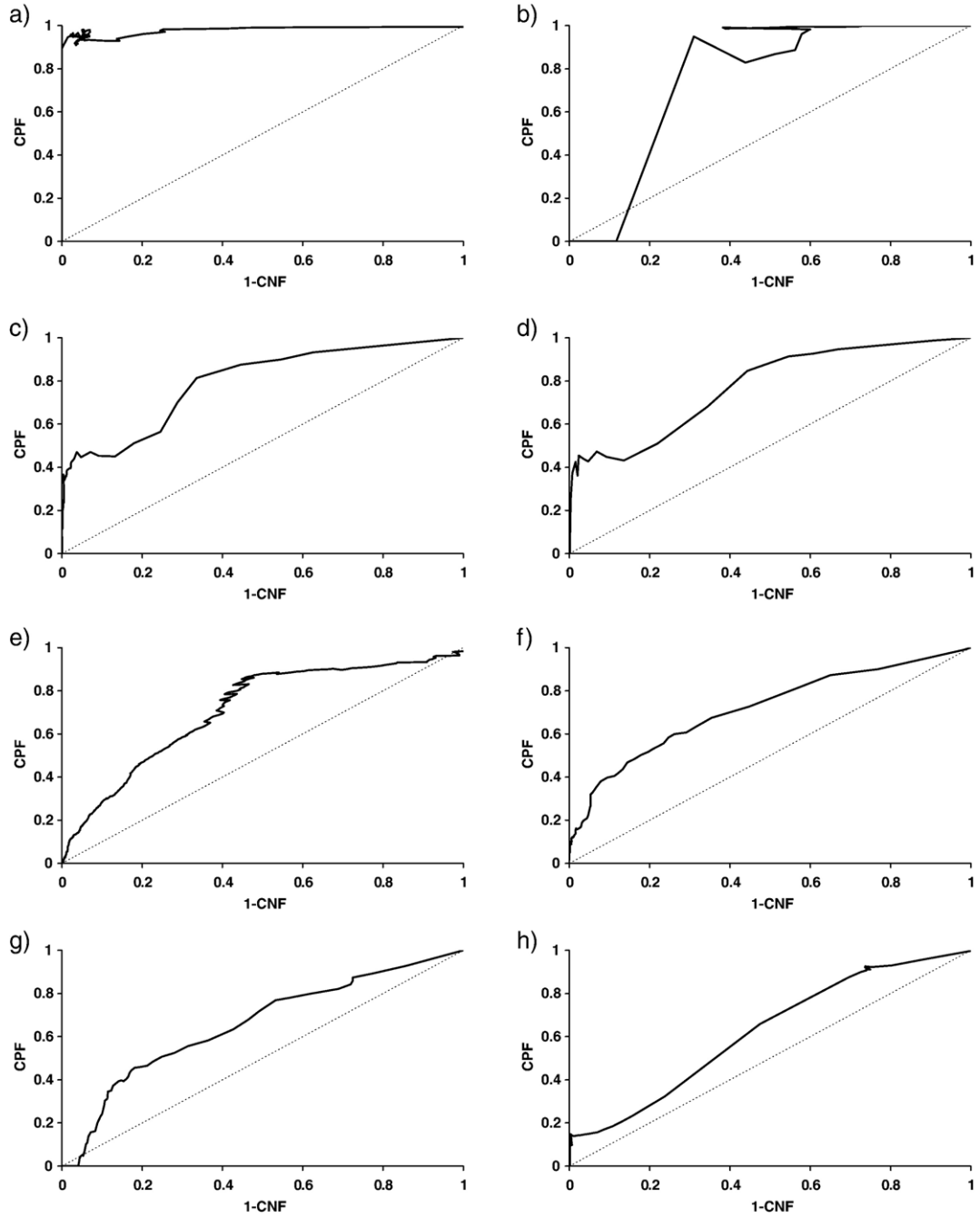

Fig. 14. Receiver operator characteristic (ROC) plots of model performance, (a) temperature, (b) salinity, (c) nitrate, (d) phosphate, (e) chlorophyll, (f) silicate, (g) ammonia, (h) suspended sediment. The sensitivity is the probability that case X classified correctly as above the threshold and the specificity (Sp) is the probability that X classified correctly as below the threshold. Dots indicate threshold point, calculated lowest threshold is top right, highest bottom left.

lie closest to, but just above the random line indicating a lack of predictive skill.

Fig. 15 shows the probabilities that a positive or negative decision is correct at a particular threshold for each of the variables considered. Temperature (Fig. 15a) is clearly the most reliable variable, with a greater than 90% probability that both positive and negative decisions are correct over the range 8–16 °C.

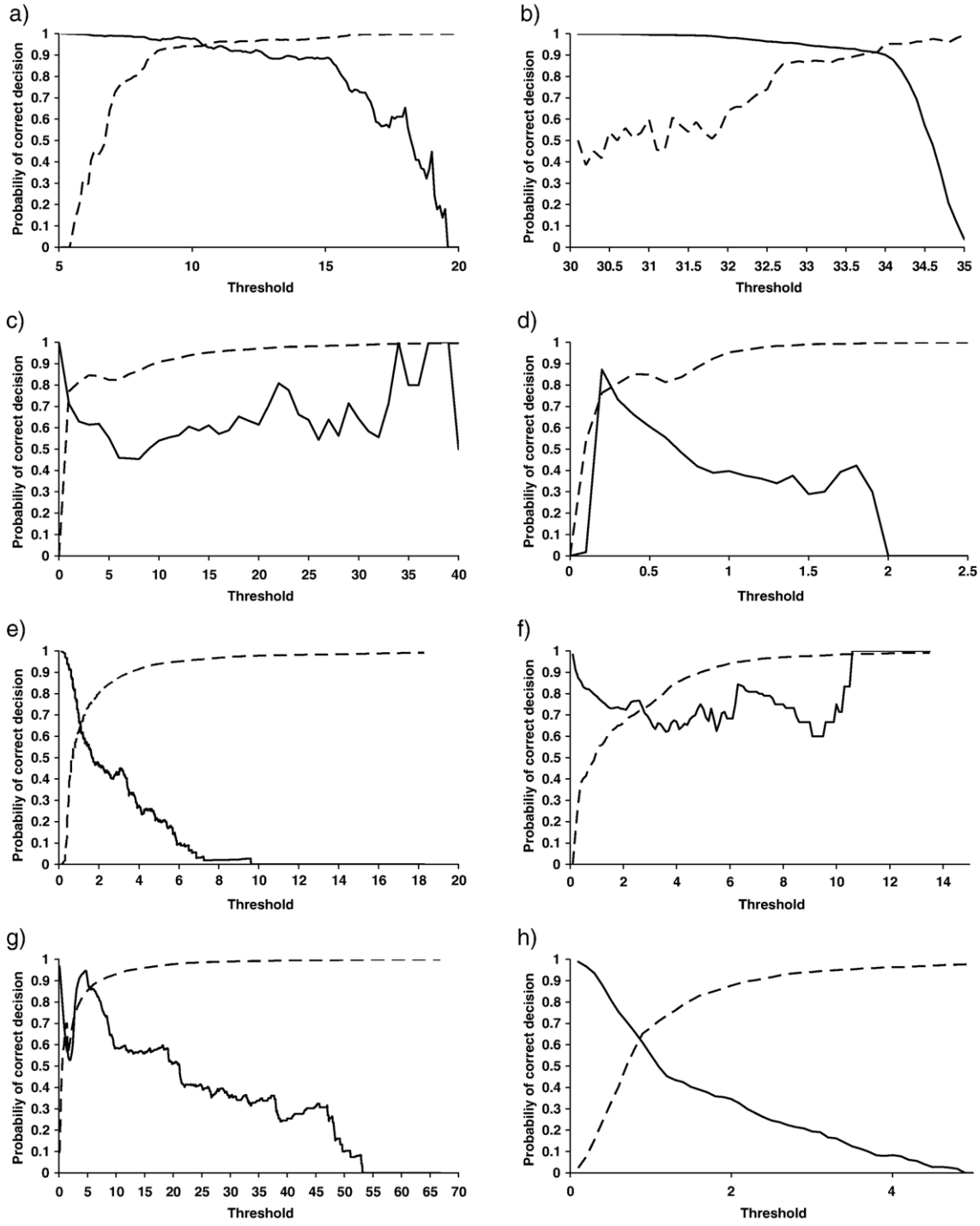

Fig. 15. Probability that a positive or negative decision is correct. As the discrimination threshold is varied, (a) temperature, (b) salinity, (c) nitrate, (d) phosphate, (e) chlorophyll, (f) silicate, (g) ammonia, (h) suspended sediment. Positive predictive value=solid line, negative predictive value=dashed line.

For salinity (Fig. 15b) the model can discriminate positive decisions over a wide range of thresholds (up to 34.0 psu) and negative decisions above this threshold. For nitrate, phosphate, chlorophyll silicate and suspended sediment (Fig. 15c–g) the negative predictive values are in excess of 0.9 over substantial ranges

Table 3
Average QSA scores for Figs. 5–7 along with ME and bias

	CT	CI	EK	AB	BQ
Chlorophyll					
QSA	2.74	2.22	3.91	3.45	2.33
S QSA	0.53	0.64	0.35	0.75	1.03
ME	0.59	0.68	0.62	0.74	−0.19
Bias	1.86	8.32	20.28	28.04	80.0
Nitrate					
QSA	2.93	2.25	2.44	3.12	3.00
S QSA	1.01	0.71	1.12	0.64	0.76
ME	0.69	0.18	0.51	0.59	0.01
Bias	29.2	39.24	−20.4	14.1	47.8
Phosphate					
QSA	3.13	2.38	2.25	3.18	1.5
S QSA	0.64	0.51	0.38	0.37	0.76
ME	0.82	0.91	0.72	0.33	−2.7
Bias	5.2	2.3	3.2	−7.9	125.5

of the data range. The ability to discriminate a positive event is poor; nitrate and silicate are best (~60–70% chance the decision is correct), chlorophyll, phosphate and suspended sediment having very little skill. The discriminatory skill for ammonia (Fig. 15h) is very poor.

3.6. Quantiative subjective analysis

The mean and standard deviations of the QSA scores for each model data comparison included in the test (Figs. 5–7) are shown in Table 3 along with the ME and bias for each graph. They indicate a wide range of assessment, the lowest scores are associated with stations CI and BQ, the highest scores with station AB, implying a consensus that AB is the best simulation and CI/BQ are the weakest. However these mean values show little correspondence with the ME and Pbias calculated for each of the plots. If we examine the relationships between individual QSA scores and goodness of fit criteria we can see a clear relationship (Fig. 16) which indicates the range of abilities of individuals to discriminate goodness of fit; some individuals apparently having no ability to discriminate goodness of fit visually.

4. Discussion

The purpose of quantitative modelling in science is to gain understanding of the natural world (Oreskes, 2003). In marine science, such models have two primary purposes. The first is a heuristic role, whereby models are used to corroborate a hypothesis, illuminate areas which require further study and identify where extra data are required. The second is as predictive tools, which can be used to aid management and assess the impact of man on the environment.

4.1. Assessing model performance

In this study, we have deliberately chosen to be unforgiving by making a direct model–data comparison in space and time, with no tolerance for errors in, for example, the time of a bloom or the depth of the thermocline, so we are effectively assessing the ability of the model to reproduce the short-term variability of the observations. This is appropriate as these models are being evaluated elsewhere (e.g. Siddorn et al., 2006) for their short-term operational forecast potential by the UK National Centre for Ocean Forecasting (www.ncof.gov.uk). Interannual simulations (e.g. Taylor et al., 2002) of this model, albeit in different physical environments, indicate a reasonable ability to resolve interannual

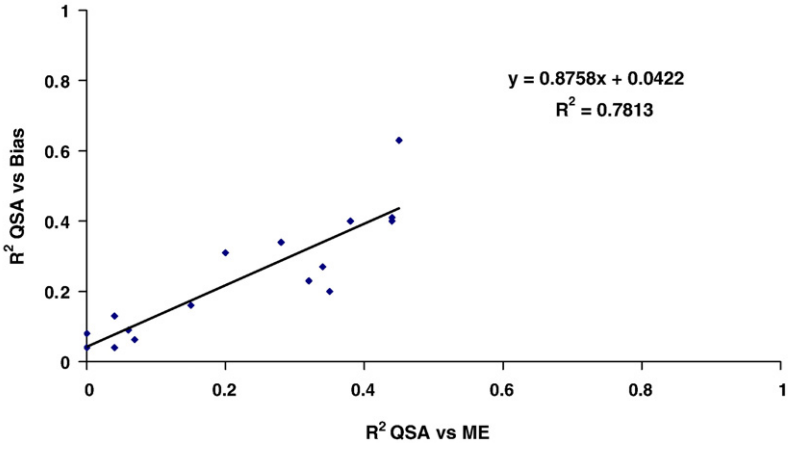

Fig. 16. Plot of the correlation between the QSA score and model efficiency against the correlation between QSA score and percentage model bias.

variability through the model ecosystem in response to changes in meteorological forcing.

The model does well in terms of hindcast performance for temperature and has some degree of skill for nitrate and phosphate in parts of the Southern North Sea as evidenced by the bubble plots (Figs. 10 and 11). Our analysis indicates that the model is incapable of reproducing the observed short-term variability in chlorophyll, silicate, ammonia and suspended sediment, although further analysis (Allen et al., 2007) indicates that it is capable of simulating bulk seasonal chlorophyll concentration. There is also evidence that the error in chlorophyll, silicate and ammonia, particularly the skew of the error distribution, increases as the simulation progresses (Fig. 8), indicating that data assimilation might be appropriate to address this.

Errors in plankton growth may arise from a combination of light, turbulence and nutrient availability. One obvious reason for the inability of the model to reproduce short-term small-scale dynamics is the poor resolution of the cloud cover used (based on daily satellite observations for the whole region) which has no spatial variation over the Southern North Sea and only limited temporal variation. The southern North Sea is a highly optically complex region and the model clearly fails to represent this (Fig. 13). The available light for photosynthesis is reduced by absorption by coloured dissolved organic matter (CDOM) and inorganic sediment. The model does not describe CDOM and the substantial land derived sources of CDOM into the region remain unquantified. The SPM sources are derived from NSP data and multiple linear regression (Holt and James, 1999b), and were then reduced 3.8 M ton/y in an attempt to improve the chlorophyll simulation. However, there are large uncertainties in land derived sediment inputs, both freshwater and from episodic coastal erosion particularly from the east coast of England (Holderness); the model ability to simulate 1% light depth is poor in this region (Fig. 13.). The overestimation of light levels in coastal waters leads to excessive winter diatom production, depleting silicate levels, and allowing zooplankton biomass to be maintained over the winter (as indicated by comparison with CRP data, tows HE and LG, Lewis et al., 2006). The combination of lower silicate concentrations (Fig. 11f, Allen et al., 2007) and enhanced grazing pressure prevents a sharp spring diatom bloom (as indicated by comparison with CRP data, tows HE and LG, Lewis et al., 2006) and substantial nitrate and phosphate draw down, hence the overestimation of nitrate and phosphate levels during the summer (Figs. 6 and 7; Allen et al., 2007). During 1988/89, there were no ocean colour satellite missions; this prevents us from attempting to correct the in-situ light field with satellite derived absorption fields. Similarly, uncertainty in the land derived inputs of nutrients, and the absence of atmospheric sources, may also impact on the plankton dynamics, although the variable C:N and C:P ratios in ERSEM phytoplankton allow them to buffer these changes (Allen, 1997).

In the central and northern North Sea, where the optics are less complex, we hypothesise that the accuracy of the turbulence model is the dominant factor. The spring bloom in 1989 at station CS in our model starts approximately a month early, when compared with in-situ fluorometry and CPR data (Holt et al., 2005; Lewis et al., 2006) and is dominated by diatoms. Huisman et al. (1999) introduced the notion of critical turbulence thresholds for phytoplankton growth; i.e. phytoplankton maintains a population as long as growth rates in the euphotic zone exceed rates of vertical transport downwards. Consequently, the simulation of vertical turbulent transport (as distinct from turbulent effects on the density profile) is critical to the simulation of the spring bloom. The consistent early modelled spring diatom bloom in the central and northern North Sea (e.g. Holt et al., 2005; Lewis et al., 2006) may imply that vertical mixing is too weak in these regions; this effect has been confirmed in preliminary experiments with a different turbulence model (the k-ε in GOTM based on Canuto et al., 2001). It may also imply errors in the parameterisation of diatoms. Undoubtedly we could reparameterise the diatom submodel to push back the timing of the spring bloom, but we are reluctant to make biological parameter changes to compensate for errors in physics.

There is an additional top–down control on phytoplankton growth in the North Sea, which remains largely unquantified. Observed grazing rates for UK coastal waters (e.g. Burkill et al., 1987) indicate that up to 60% of the phytoplankton standing stock are grazed out by microzooplankton. Clearly, it is important to make the distinction between errors in model processes, and those derived from external forcing.

4.2. Uncertainties in both the data and the model

The analysis presented in this paper is only possible because of the existence of a large self-consistent data set. Unfortunately, such data sets are rare, although the increasing number of in-situ mooring systems (e.g. in coastal-observatories) and underway data collection (including ships of opportunity) is beginning to improve the situation.

Establishing natural levels of variability within and between marine ecosystems is a prerequisite to rigorous model validation, and requires data collection that takes account the important (time and space) scales of variability. In assessing simulation performance, we need to consider both the quality of the validation data and the model forcing functions (river inputs, atmospheric forcing and open boundaries) all of which contain errors. So far, we have made a like-with-like comparison of model and measurements assuming that the measurements are accurate. Obviously, this is not always the case as the accuracy of measurements of some parameters is highly variable and the less certain we are of the measurement the harder it is to be certain the model is at fault.

Temperature and salinity from CTD are highly accurate measurements (errors of the order 0.0005 °C in temperature, 0.01 psu in salinity; Lowry et al., 1992). The nutrient measurements were made with a Chemlab autoanalyser and are very accurate, with errors of the order 1% for PO_4 and NO_3, and 4% SiO_4 over the scales of measurement (D. Hydes, pers com). In these cases we can place more trust in the error statistics. Values of ammonia below 0.5–1.0 mmol m^{-3} may not be very reliable (D. Hydes pers com) implying that we can place less trust in the error estimates. The chlorophyll-a data used here were measured using a CTD fluorometer calibrated to in-situ chlorophyll-a. In-situ chlorophyll was measured using the method of Strickland and Parsons (1972) who state that at 0.5 mg Chl m^{-3}, the measurement error is $\pm \frac{0.26}{\sqrt{n}}$ where n is the number of replicates. This implies that at low chlorophyll concentrations the errors in the chlorophyll are of the order 25–50%. When the measurement errors are coupled with calibration errors for the fluorometer, we can only conclude that the errors statistics calculated here should be treated with caution. SPM is measured as the particulate fraction after seawater is passed through a 47 μ filter and includes both the inorganic and organic fractions, this was then used to calibrate the CTD transmissometer. The uncertainty in such measurements can be large, particularly near bed SPM concentrations and at slack water (Jago and Bull, 2000). In an error quantification exercise Jago and Bull (2000) estimate the errors in transmissometer derived sediment fluxes to be of the order of 20%, when compared with the equivalent gravimetric flux. This implies that the errors statistics for SPM may not be very reliable. Primary production measured using the C_{14} technique is also highly variable with errors up to an order of magnitude, indicating that the error statistics for primary production should be viewed with some scepticism.

To interpret our results we also need estimates of the model errors. These have three components. The formal model error, defined as the divergence between the true mathematical solution to an equation and a numerical solution is not generally accessible for the full model but can be investigated through convergence experiments. The second is the propagation of errors in response to uncertainties in initial conditions, forcing functions and parameters. We can get an understanding of model variance in response to such perturbations by undertaking an ensemble sensitivity analysis. However, the computational cost of such an analysis means that it is currently not possible for models of this scale. Finally there is the uncertainty associated with the approximations needed to formulate the model, for example how to partition the ecosystem into functional groups.

In these simulations, an averaged annual cycle is used for the open ocean physical boundary conditions and a zero net flux boundary condition used for the ecosystem variables. We can see from the skew analysis that there is systematic propagation of errors which one can speculate may result from the propagation of boundary errors through the model domain. We know that the Atlantic Ocean is a major source of nutrients to the southern North Sea (e.g. Howarth et al., 1993). For example Seitzinger and Giblin (1996) estimated net flux of nitrogen from the North Atlantic onto the NW European shelf to be 2.44 Mt y^{-1}, the amount required in excess of river and atmosphere contributions to balance the loss of nitrate caused by denitrification; so we might expect the model to be sensitive to such effects. The acquisition of better boundary conditions is ongoing, and the impacts on the model system are being assessed. The benchmarking of model performance (discussed below) is crucial to the assessment of changes in the model.

4.3. Choice of metric

Previous model validation exercises in this region (e.g. Moll, 2000; Radach and Moll, 2006) have focused on the use the OSPAR recommended cost function, (OSPAR, 1998). Cost functions are a measure of model data mismatch and are primarily used in data assimilation, usually taking the form of the difference between model and observation, scaled by some measure of the variance of the data; i.e. if the cost function is less than 1 the model data mismatch is less than the variance of the data. They can be both univariate and multivariate. In data assimilation, the aim is to determine the set of parameters or initial conditions that minimise the cost function, thus drawing the simulation towards the data. Our results clearly imply that, used on its on own for classification, the OSPAR cost function is flawed: all the model variables considered would be classed as very

good, while the other metrics used (e.g. ME, Pbias, R^2) clearly indicate that this is not always the case. This is most notable for SPM and ammonia, which the ROC test demonstrates to be close to random. There is a basic problem with the mathematical structure of this function. Because it is linear, it neither rewards goodness of fit nor punishes poor fit. A cost function including a power term in the model data mismatch (e.g. that used in Holt et al., 2005), data variance rewards goodness of fit and punishes misfit, and we recommend the use of such metrics.

In order to assess whether changes to our model system, e.g. parameter changes, new variables or changes to forcing functions, are effective we need to be able to compare to a reference or 'benchmark' simulation. To systematically benchmark model performance a hierarchy of tests is required. As a starting point, we suggest the following.

1. The ROC, a simple binary discriminator with variable threshold, allows a basic assessment of whether the model has any skill, or is just a random number generator.
2. The simplified Taylor plots provide an overview of model performance and can be used as a benchmark to assess the success of model developments. Crucially this allows us to see whether the model is reproducing the observed variance. We should however caution that these metrics are a summary and are not a substitute for the detailed analysis of model behaviour required to gain insight into process descriptions.
3. The combination of model efficiency and bias is more informative than using a simple cost function. A bare minimum performance level should be that the ME is greater than zero, i.e. model errors are less than the variability of the data (Allen et al., 2007). Understanding model bias is crucial when models are to be used to define indicators for environmental management, and a bias of less than 40% is the bare minimum acceptable performance criteria (Allen et al., 2007). Ultimately, we need to be far more stringent.
4. Temporal analysis of error propagation allows the identification of poorly described processes and potential target variables for data assimilation.
5. Spatio-temporal analysis of variability in errors allows the diagnosis of model errors and defines both regions and processes to work on when combined with cluster type analysis that identifies biogeomes (these are regions of self-consistent biogeochemical properties, which have not been considered here but are described by Allen et al., 2007).

The QSA points to the range of assessments that individuals will make about the quality of model output compared with data. Subjective goodness of fit is one of the criteria used by many analysts but this begs the rhetorical question: should we dismiss those analysts who use other criteria? Clearly, there is a role for experience, expert knowledge of the system and intuition when benchmarking model results. Consequently, we should add the following caveat; when applying any statistical tests to the model–data comparisons we should not lose sight of common sense, the basic assessment of whether or not the results exhibit plausible behaviour is still highly relevant.

4.4. Discrimination thresholds

Discrimination analysis allows us to assess model performance in a way which is potentially highly relevant for environmental management, where many decisions are based on thresholds. Once thresholds are defined, we can use the discrimination analysis to determine the probability that a predicted elevated level in the model is correct. If for example the threshold for elevated nitrate (DIN) is 15 mmol m^{-3} then PPV = 0.61 and PPN = 0.95, so if the model predicts elevated nitrate, there is a 61% chance it is correct; there is also a 95% chance that a prediction below the threshold is correct. For chlorophyll, if we set a threshold of 5 mg Chl m^{-3} PPV = 0.21 and PPN = 0.93 so there is a 21% chance that if the model predicts elevated concentration it is correct and there is also a 93% chance that a prediction below the threshold is correct. While we cannot be sure with any great confidence that the model can reliably discriminate elevated nutrient or chlorophyll concentrations, we can be more certain of model indications of non-elevated concentrations.

5. Conclusions

All models are open systems and by definition are simplifications that do not completely encompass the natural system (Oreskes, 2003). In the case of marine ecosystem models they are open with respect to the assumptions made about the complexity of the system, the empirical adequacy of the equations and how well the model variables represent elements of the system. Consequently, models can be only confirmed by the demonstration of agreement between observation and prediction, but confirmation is inherently partial (Oreskes et al., 1994). This partial confirmation is however immensely useful in allowing us to test our understanding of the system being modelled.

The analyses presented here are among the first steps towards understanding model uncertainty, identifying dysfunctional process descriptions and hence improving the articulation of detail in ecosystem models. If we are to make reliable simulations, we need to be able to quantify and understand how model errors propagate. This also requires an appreciation of the inherent errors in the data. A sceptical reader may argue at this point that the process errors we have identified by this analysis could have been equally well determined from a combination of conventional plots, visual analysis and the combination of the modeller's knowledge and intuition, which is ultimately dependent on the skill of the modeller rather than the model. This is to some extent true and the metrics used tend to confirm what we already knew from previous studies. However, the QSA (while not rigorous) clearly indicates the wide range of outcomes subjective analysis can result in. The object metrics do however give us clear confidence intervals, which can be taken to policy makers and are the first step towards a quantitative risk assessment.

Model development is an iterative process: only by a quantitative benchmarking of model uncertainty can we reduce subjectivity when assessing changes to a model. This is important whether the models are used in heuristic or forecast mode. However, we must be cautious, as Flynn (2005) points out 'just because a model gives a fit to a particular data set, it does not guarantee the structure is not dysfunctional'. Consequently, there is a balance to be achieved between statistical fit and intuitive understanding of system function. We should also bear in mind that a healthy dose of scepticism is always useful when interpreting models.

Acknowledgements

This work was partly funded by the EC MERSEA Integrated project (Co No AIP3-CT-2003-502885) and partly by the NERC core strategic research programs of the Plymouth Marine Laboratory (JIA, JCB) and the Proudman Oceanographic Laboratory (JTH, RP). We would also like to thank the participants in the QSA experiment. Finally we thank the referees for their challenging but constructive comments.

References

Allen, J.I., 1997. A modelling study of ecosystem dynamics and nutrient cycling in the Humber Plume, UK. J. Sea Res. 38, 333–359.

Allen, J.I., Blackford, J.C., Holt, J., Proctor, R., Ashworth, M., Siddorn, J., 2001. A highly spatially resolved ecosystem model for the North West European Continental Shelf. Sarsia 86, 423–440.

Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic ecosystem models. J. Mar. Syst. 64, 3–14.

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

Baretta-Bekker, J.G., Baretta, J.W., Hansen, A.S., Riemann, B., 1998. An improved model of carbon and nutrient dynamics in the microbial food web in marine enclosures. Aquat. Microb. Ecol. 14 (1), 91–108.

Bell, M.J., Forbes, R.M., Hines, A., 2000. Assessment of the FOAM global data assimilation system for real time operational forecasting. J. Mar. Syst. 25, 1–22.

Blackford, J.C., Allen, J.I., Gilbert, F.J., 2004. Ecosystem dynamics at six contrasting sites: a generic model study. J. Mar. Syst. 52, 191–215.

Brown, C., Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: a tutorial. Chemometr. Intell. Lab. Syst. 80, 24–38.

Burkill, P.H., Mantoura, R.F.C., Llewllyn, C.A., Owens, N.J.P., 1987. Microzooplankton grazing and selectivity of phytoplankton in coastal waters. Mar. Biol. 93, 581–590.

Canuto, V.M., Howard, A., Cheng, Y., Dubovikov, M.S., 2001. Ocean turbulence. Part I: one-point closure model—momentum and heat vertical diffusivities. J. Phys. Oceangr. 31 (6), 1413–1426.

Charnock, H., Dyer, K., Huthnance, J.M., Liss, P.S., Simpson, J.H., Tett, P.B. (Eds.), 1994. Understanding the North Sea System. Royal Society of London. 222 pp.

Clarke, K.R., Gorley, R.N., 2006. PRIMER v6: User Manual/Tutorial. PRIMER-E Ltd. Plymouth.

Flynn, K.J., 2005. Castles built of sand: dysfunctionality in plankton models and inadequacy of dialogue between biologists and modellers. J. Plankton Res. 27, 1205–1210.

Hardman-Mountford, N., Allen, J.I., Frost, M.T., Hawkins, S.J., Kendall, M.A., Mieszkowska, N., Richardson, K., Somerfield, P.J., 2005. Diagnostic monitoring of a changing environment: an alternative UK perspective. Mar. Pollut. Bull. 50, 1463–1471.

Holt, J.T., James, I.D., 1999. A simulation of the southern North Sea in comparison with measurements from the North Sea Project. Part 2: suspended particulate matter. Cont. Shelf Res. 1, 1617–1642.

Holt, J.T., James, I.D., 2001. An *s*-coordinate model of the northwest European Continental Shelf. Part 1. Model description and density structure. J. Geophys. Res. 106 (C7), 14015–14034.

Holt, J.T., Proctor, R., Blackford, J.C., Allen, J.I., Ashworth, M., 2004. Advective controls on primary production in the stratified western Irish Sea: an eddy-resolving model study. J. Geophys. Res. 109 (C05024). doi:10.1029/2003JC001951.

Holt, J.T., Allen, J.I., Proctor, R., Gilbert, F., 2005. Error quantification of a coupled high-resolution coupled hydrodynamic-ecosystem coastal ocean model: Part 1. Model overview and assessment of the hydrodynamics. J. Mar. Syst. 57, 167–188.

Howarth, M.J., Dyer, K.R., Joint, I.R., Hydes, D.J., Purdie, D.A., Prieur, L., Jones, J.E., Lowry, R.K., Moffat, T.J., Pomroy, A.J., Proctor, R., 1993. Seasonal cycles and their spatial variability. Philos. Trans. R. Soc. Lond. Ser. A: Math. Phys. Eng. Sci. 343, 383–403.

Huisman, J., van Oostveen, P., Weissing, F.J., 1999. Critical depth and turbulence: two different mechanisms for the development of phytoplankton blooms. Limnol. Oceangr. 44, 1781–1787.

Jago, C.F., Bull, C.F.J., 2000. Quantification of errors in transmissometer-derived concentration of suspended particulate matter in the coastal zone: implications for flux determinations. Mar. Geol. 169, 273–286.

Joint, I.R., Pomroy, A.J., 1993. Phytoplankton biomass and production in the southern North Sea. Mar. Ecol. Prog. Ser. 99, 169–182.

Jones, J.E., 2002. Coastal and shelf seas modelling, in the European context. Oceanogr. Mar. Biol. 40, 37–141 (An annual review).

Lewis, K., Allen, J.I., Richardson, A.J., Holt, J.T., 2006. Error quantification of a high-resolution coupled hydrodynamic ecosystem coastal ocean model: Part 3. Validation with CPR data. J. Mar. Syst. 63, 209–224.

Lowry, R., Crammer, K., Rickards, L., 1992. North Sea Project CD ROM and User Guide. British Oceanographic Data Centre. Natural Environmental Research Council, Swindon UK.

Maréchal, D., 2004. A soil-based approach to rainfall-runoff modelling in ungauged catchments for England and Wales. PhD Thesis, Cranfield University 157pp.

Moll, A., 2000. Assessment of three-dimensional physical–biological ECOHAM1 simulations by quantified validation for the North Sea with ICES and ERSEM data. ICES J. Mar.Sci. 57, 1060–1068.

Moll, A., Radach, G., 2003. Review of three-dimensional ecological modelling related to the North Sea shelf system—Part 1: models and their results. Prog. Oceanog. 57, 175–217.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I—a discussion of principles. J. Hydrol. 10 (3), 282–290.

Oreskes, N., 2003. The role of quantitative models in science. In: Canham, C.D., Cole, J.J., Lauenroth, Wi.K. (Eds.), Models inz Ecosystem Science. Princeton University Press, Princeton, pp. 13–31.

Oreskes, N., Sharder-Frechette, K., Belits, K., 1994. Verification, validation and confirmation of numerical models in earth sciences. Science 263, 641–646.

OSPAR Commission, 1998. Report of the Modelling Workshop on Eutrophication Issues. 5–8 November 1996. Den Haag, The Netherlands. OSPAR Report, 86 pp.

OSPAR Commission 2003 OSPAR Integrated Report 2003 on the Eutrophication Status of the OSPAR Maritime Area Based Upon the First Application of the Comprehensive Procedure. 59 pp.

Patsch, P., Radach, G., 1997. Long term simulations of the eutrophication of the North Sea: temporal development of nutrients, chlorophyll and primary production in comparison to observations. J. Sea Res. 38, 275–310.

Radach, G., Moll, A., 2006. Review of three-dimensional ecological modelling related to the North Sea shelf system. Part II: model validation and data needs. Oceanog. Mar. Biol. 44, 1–60 (An Annual Review).

Seitzinger, S.P., Giblin, A.E., 1996. Estimating denitrification in the North Atlantic continental shelf sediments. Biogeochemistry 35, 235–260.

Siddorn, J.R., Allen, J.I., Blackford, J., Gilbert, F., Holt, J.T., Proctor, R., Holt, M., Osbourne, J., 2006. Modelling the hydrodynamics and ecosystem of the North-West European continental shelf for operational oceanography. J. Mar. Syst. doi:10.1016/j.jmarsys.2006.01.018.

Song, Y., Haidvogel, D., 1994. A semi-implicit ocean circulation model using a generalized topography-following coordinate system. J. Comp. Phys. 115, 228–244.

Strickland, J.D.H., Parsons, T.R., 1972. A practical handbook of seawater analysis, 2nd edn. Bulletin of the Fisheries Research Board of Canada, vol. 167.

Tabachnick, B.G., Fidell, L.S., 1996. Using Multivariate Statistics, 3rd ed. Harper Collins, New York.

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

Taylor, A.H., Allen, J.I., Clarke, P.A., 2002. Extraction of a weak climatic signal by an ecosystem. Nature 416, 629–631.