A multivariate approach to model skill assessment
J. I. Allen, P. J. Somerfield

Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH UK.

Abstract
Although the purpose of models is to simplify complex reality to allow the investigation of patterns, processes and relationships, many ecosystem models retain high levels of complexity. The outputs from such models are highly multivariate. Taking the view that a perfect model simulation of a spatial domain over a determinate time period will reproduce observed variables from the same place over the same period perfectly, we demonstrate how appropriate multivariate methods may be used to elucidate patterns within observations and model outputs, to compare patterns between them, and to explore the nature and spatio-temporal distribution of model errors. Analyses based on observations collected from the southern North Sea in 1988-89 are compared to analyses based on an equivalent dataset extracted from the output of the POLCOMS-ERSEM model. A combination of PCA and nonparametric multivariate approaches are used to demonstrate that in broad terms the model performs well, simulating patterns in, and interrelationships between, a range of variables. Errors are greatest in late winter and early spring, and are associated with inaccurate estimation of the magnitude of primary production in coastal waters and the amount of suspended particulate matter in the water column.

Keywords: Multivariate analysis; model skill assessment; Principal Components Analysis (PCA); Multidimensional Scaling (MDS); POLCOM-ERSEM; North Sea.
1. Introduction

The primary role of models in ecosystem science is to provide a simplification of complex reality (Allen et al., 2007a). The continual development and enhancement of coupled hydrodynamic-ecosystem models has reached the point where the outputs of such models are themselves of a complexity and volume that they require simplification in order to be understood. Conventional model skill assessment (e.g. Arhionditsis and Brett, 2004 and reference within) is either highly subjective or relies on simple univariate assessments of goodness of fit between model outputs and observations e.g. correlations, root mean square errors, bias and cost functions. An alternative approach is to see whether a model can reproduce observed relationships between variables. This is a particularly powerful test when these relationships can be viewed as emergent properties of the model, i.e. properties that are not hardwired into the model code and emerge due to the combination of interactions between the processes described. For example, the ERSEM model can reproduce observed relationships between ecosystem properties and climate related variables such as indices describing the North Atlantic Oscillation (NAO) and the position of the North Wall of the Gulf Stream, even though the latter is not present in meteorological forcing (Taylor et al., 2002).

The skill assessment problem is multivariate by nature. If we wish to understand not only how the individual model variables fit the data but also the relationships between variables we need to invoke a multivariate approach which allows the simultaneous examination of the ways in which numerous variables vary in relation to each other in both space and time. Such approaches are commonly used in marine ecology to interpret complex data sets. Multivariate approaches have been used investigate the patterns or models of variability in marine ecosystem model outputs (e.g. Allen et al., 2002; Blackford et al., 2004, Schrum et al., 2006, Allen et al., 2007a, Allen and Clarke, 2007) but rarely used for skill assessment (but see Allen et al., 2007a).

If we have a set of observations available for model validation we can subject them to multivariate analysis. If we then reconstruct a data set from the model by taking the nearest equivalents in space and time, we can subject them to the same analysis and compare the results. By definition, if the observations are the truth then the perfect model should exactly reproduce the observed multivariate patterns.

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. 1 shows the
area sampled during the NSP, on which our analyses focus. We use data (available from British Oceanographic Data Centre), including temperature, salinity, chlorophyll-a, nitrate, phosphate, ammonia, silicate and suspended sediment, collected at ~120 CTD (conductivity, temperature and depth) and water-sampling stations during each of 14 monthly cruises between August 1988 and September 1989 (~1600 stations in total). We compare these with the outputs of the POLCOMS-ERSEM model, which is a state-of-the-art coupled 3D hydrodynamic-ecosystem model for shelf seas (Allen et al., 2001; Holt et al., 2005). Focusing on a subset of model outputs for which we have appropriate comparative observations, we apply a combination of multivariate techniques to explore relationships between model outputs, observations, data-model mismatch and the distribution of errors within the model. We have two goals, to demonstrate the use of multivariate analytical tools for model skill assessment and to gain insight into the sources and distribution of model errors.

2. Model, Data and Multivariate Methods

2.1 Model

The simulations used the POLCOMS-ERSEM medium resolution shelf seas ecosystem model (Allen et al 2001). The model includes the density evolving physics of POLCOMS (Holt and James, 2001) and a size-fractionated SPM sub-model (Holt and James, 1999), coupled with the state-of-the-art biogeochemical processes of ERSEM (Blackford et al., 2004; Baretta et al., 1995). It is a coupled 3D hydrodynamic and ecosystem model (Allen et al., 2001, Holt et al., 2004), set up on a 1/10° longitude by 1/15° latitude horizontal grid (~7km resolution) with 20 s-levels (Song & Haidvogel, 1992) in the vertical, and boundaries following the North-West European Continental Shelf break (approximately along the 200m 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. The simulations are of the period of the data rich NERC North Sea Project (1988-89) and are described in detail in Holt et al. (2005). The model is spun up using 1988 forcing then run forward for
Model validation is presented in Lewis et al. (2006) and Allen et al. (2007a, b).

In ERSEM the ecosystem is represented as a network of interacting physical, chemical and biological processes. A ‘functional group’ approach is used to describe the biological components consisting of primary producers, consumers and decomposers, which in turn are subdivided on the basis of trophic links and/or size.

The dynamics of each functional group are dictated by physiological (ingestion, respiration, excretion and egestion) and population (growth, mortality and migration) processes that are described by fluxes of nutrients between groups. The adaptation of phytoplankton to variability in the light field assumes that the optimal growth parameter in the photosynthesis-irradiance equation adjusts, enabling optimisation of growth rates in a variety of regimes (Ebenhöh et al., 1997). Phytoplankton biomass is described by carbon, nitrogen, phosphorous and in the case of diatoms silicon components, because variable C/N/P cell quotas are allowed. Nutrient uptake is decoupled from the carbon processes that are dependent on internal cell nutrient quotas. Gross primary production is a function of temperature, availability of and adaptation to light, silicate concentration in the case of diatoms and phytoplankton biomass. The required nutrient uptake is calculated as a combination of uptake commensurate with carbon productivity and uptake necessary to address any internal shortfall of nutrients. Actual uptake is constrained by a maximum uptake dependent on an affinity parameter and external nutrient concentrations. If there is a shortfall in the available nutrients which prevent the entire gross production being fixed the excess carbon is diverted to the DOC pool via nutrient-stressed lysis.

The phytoplankton community (Fig. 2) consists of four functional types: picophytoplankton (0.2-2.0 μm), phytoflagellates (2-20 μm), phototrophic dinoflagellates (20-200 μm) and diatoms (20-200 μm). Photosynthetically active light (PAR) at the sea surface is modelled from astronomical values that are corrected for cloud cover. Extinction of light through the water column is a function of particulate concentrations and phytoplankton biomass and is given a background extinction parameter. Mesozooplankton, microzooplankton and heterotrophic flagellates make up the consumers (Fig. 2). Grazing uptake is a function of maximal ingestion rate, temperature, food availability and the zooplankton biomass. Zooplankton mortality
occurs in the model at a constant rate and may be triggered by low oxygen concentration. Each zooplankton functional group is predated upon.

2.2 Data
In this work we use the following data from the North Sea project (Charnock et al., 1994): temperature (28595 measurements); salinity (28490); chlorophyll-a (24820); total sediment (23645); nitrate (4467); phosphate (4856); silicate (4818). Location-time combinations for which all variables were not sampled were omitted. Data were averaged to get mean water column values for each station on each sampling occasion, 1099 sets of observations overall.

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., 1993). To calibrate the CTD sensors, temperatures were obtained from reversing thermometers and salinity determinations were made. 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 up to 25-50%. Spectrophotometric chlorophyll and phaeopigment determinations were used to calibrate the CTD fluorometer. Nutrients (nitrate, nitrite, silicate, phosphate and ammonium) were determined from water bottle samples using a Chemlab autoanalyser and have errors of about 1% for PO₄ and NO₃, and 4% for SiO₄ over the scales of measurement. SPM was measured as the particulate fraction retained on a 47µm 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 for 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.

2.3 Multivariate Methods
The majority of the multivariate analyses used here are described in detail in Clarke (1993) and Clarke and Warwick (2001) and were carried out using Primer v6 (Clark and Gorley 2006).
Principal Component Analysis (PCA e.g. Chatfield & Collins, 1980) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components (PCs). Each principal component is a linear combination of the original variables. The first principal component (PC1) accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible, while being constrained to be orthogonal to all preceding components. We apply the Kaiser criterion (Kaiser, 1960) when choosing the eigenvectors for analysis, which states that we only interpret factors with Eigen values greater than 1. Essentially we are saying that unless a factor extracts at least as much information as the equivalent of one original variable, we drop it. Prior to this we normalised (convert to a mean of zero and a standard deviation of one) the variables to account for the fact that they are measured on different scales (covariance based PCA).

An ordination plot derived from PCA is a 2D projection. If we are interested in relationships in the full multivariate space we need to construct matrices that reflect these relationships. There are many measures which reflect the resemblance between vectors (Legendre and Legendre, 1997; Clarke et al., 2006). The appropriate resemblance measure underlying PCA is Euclidian distance (Clarke et al., 2006). Correlations between vectors may also be used as resemblance measures. The Pearson product-moment correlation coefficient ($r$) is a measure of the strength of the linear correlation between two variables $X$ and $Y$ measured on the same object, that is, a measure of the tendency of the variables to increase or decrease together. This makes it ideal for assessing the goodness of fit between modelled and observed variables where a linear relationship is expected. Spearman's rank correlation coefficient ($\rho$) is a non-parametric measure of correlation; it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike the Pearson product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is strictly linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level.
Nonmetric multidimensional scaling (MDS; Clarke, 1993; Kruskal and Wish, 1978) constructs a ‘map’ or configuration of samples in a specified number of dimensions (two in this study) which attempts to satisfy all the conditions imposed by the rank order of distances in a resemblance matrix, thus placing similar samples close together in the ordination and avoiding the potentially misleading projection step characteristic of other ordination methods. How well the ordination succeeds in capturing the full multivariate information in 2 dimensions is indicated by the stress value

\[ \text{Stress} = \sqrt{\frac{\sum_j \sum_k (d_{jk} - \hat{d}_{jk})^2}{\sum_j \sum_k d_{jk}^2}} \]

where \( \hat{d}_{jk} \) is the distance predicted from a fitted monotonic-increasing regression corresponding to distance \( d_{jk} \). A stress value <0.05 indicates an excellent representation with no prospect of misinterpretation, <0.1 indicates a good representation, and <0.2 a useable representation in which detail should not be relied upon (Clarke, 1993). Values >0.3 indicate that the algorithm has not managed adequately to represent the full multivariate information in 2 dimensions. As the technique aims primarily to retain the rank order in the resemblance matrix and it is only the relative distance apart of the symbols that matters, the scaling of the plots is somewhat arbitrary, MDS ordinations do not have axes, and they may be reflected or rotated at will.

While patterns in resemblance matrices may usefully be compared by comparing ordination plots derived from them, such plots do not retain the full multivariate information in the underlying resemblance matrices, and no significance may be attached to apparent similarities or differences between plots. The appropriate test is a Mantel test (Mantel, 1967) of no relationship between the matrices. A correlation between corresponding elements in each distance matrix is calculated, and the significance of the correlation is determined by a Monte Carlo permutation procedure. In this a distribution of values which the correlation may take if there is no relationship between the 2 matrices is constructed, by relabeling one matrix at random and recalculating the correlation a large number (here 1000) of times. The observed value is then compared to the permutation distribution to assess the probability that a value as large, or larger, as the observed one could have arisen by chance. A
permutation procedure is essential because the lack of independence between elements of a resemblance matrix makes the standard tables for testing $\rho$ invalid.

Allen et al. (2007a) used a self-organising map (SOM) as a clustering tool to define biogeomes (regions of self-consistent biogeochemical properties). The SOM is an unsupervised neural network which is particularly adept at pattern recognition and classification (Kohonen, 1995). Model output for the study region was classified into five regions. We use this classification to define coastal and offshore waters to aid our analysis.

3. Multivariate model verification

3.1 The relationship between model and data.

The first question is: what is the relationship between the modelled and observed data? We have seven variables sampled at 122 stations for each of 14 months, so to reduce the dimensionality of the data set and identify new meaningful underlying variables we use Principal Component Analysis (PCA).

In the case of the observations the first three eigenvectors meet the Kaiser criteria and account for 76% of the observed variability (Table 1). For the model the first three eigenvectors account for 80% of the modelled variability. The first principal component accounts for 41.4% of the variability and the scores for the individual variables (Table 1) show that it essentially reflects the seasonal variability of nutrients. The second principal component accounts for 20.5% of the observed variability, reflecting the seasonality of chlorophyll, temperature and salinity. The third principal component accounts for 14% of the variability and is dominated by a mixture of temperature and suspended sediment (Table 1).

Ordination plots derived from PCA illustrate the major modes of variability in both the observations and the model (Fig. 3.). Fig. 3a shows the plot of PC1 and PC2 derived from the observations. Fig. 3b illustrates the plot of PC1 and PC2 derived from the simulations. The model PCA points (Fig. 3b) are much more tightly clustered together than the observations (Fig. 3a), implying that the model doesn’t reproduce all the observed variability, but both patterns have some similarities. In both cases samples from the autumn and winter months are to the right of the diagram and the spring and winter months to the left. The outliers to the right on PC1 in both
cases are in winter and in the coastal zone. High positive values on PC2 are more apparent for the model than for the observations and are possibly associated with the Rhine and Elbe river plumes.

Clearly there is correspondence between the first two modelled and observed principal components but not the third. This is confirmed by plotting the two components against each other (Fig. 4). Pearson correlations indicate that modelled PC1 reproduces 43% of the observed PC1 and modelled PC2 reproduces 39% of the observed PC2. Modelled PC3 has no skill at all \( r^2 < 0.003 \), implying that model simulation of SPM is poor.

3.2 Objective assessment of similarity in pattern

While comparing ordinates is useful, actually what we are interested in is the relationship in the full multivariate space. For this we need to construct matrices that reflect these relationships, and an appropriate measure is Euclidian distance (this is the distance measure implicit in PCA). A non parametric Mantel test (Fig. 5) shows a high rank correlation between model and observation \( \rho = 0.639 \) which is highly significant, as the observed value lies well outside the distribution of values derived from 100 random permutations so we are unlikely to be mistaken in rejecting the null hypothesis of no relationship between the matrices. While we could write this as \( p < 0.001 \), any number of permutations is unlikely to produce a correlation value as high as the one we actually observed. The highly significant correlation implies that the model has skill in reproducing overall patterns in the observations.

3.3 Deconstructing the error structure.

We now ask the question: which variables are being well represented by the model? The simplest approach is just to look at the correlations between the observed and modelled variables (Table 2), a high correlation implying high skill. The model has some skill for all variables; temperature and salinity have the most, SPM and chlorophyll the least. We are also interested in the interrelationships between observation and model for a whole range of variables simultaneously and we can investigate this by looking the correlations between both modelled and observed variables (Table 3). However this results in \( n(n-1)/2 = 91 \) correlations and we would
like to simplify the problem. Therefore we use nonmetric multidimensional scaling (MDS) to give us a simple visual representation of the relationships.

The MDS plot of a resemblance matrix based on squared Pearson correlations is shown in Fig 6a allowing us to assess whether the model is quantitatively reproducing the data in space and time. It shows strong correlations ($r^2 > 0.6$) between modelled and observed temperature and salinity, modelled nutrients and observed nitrate and silicate. There are weaker correlations ($r^2 > 0.2$) between modelled and observed SPM, temperature and modelled nutrients, and observed nutrients and salinity. The latter two suggest that the observed relationship between nutrients and salinity is not reproduced by the model, perhaps implying that freshwater nutrient inputs are not correctly defined. Modelled and observed chlorophyll are uncorrelated.

If we input a rank Spearman rank correlation resemblance matrix to MDS (Fig. 6b) we can assess the relationships between trends in the variables. It shows that the model is much better at capturing trends than producing an exact like with like fit. The analysis clusters variables into three groups: temperature and chlorophyll, salinity, sediment and nutrients. This is evidenced by the correlations ($\rho^2 > 0.40$) between modelled and observed chlorophyll and modelled and observed nitrate and silicate. The fact that the Spearman correlation between modelled and observed chlorophyll is good ($\rho^2 > 0.40$) while the Pearson correlation is poor ($r^2 < 0.20$) implies that the model is capturing the qualitative trends in, but not the quantitative amounts of, chlorophyll.

Both quantitatively and qualitatively (Fig. 6a, b) observed phosphate is decoupled from observed nitrate and silicate, a feature the model fails to reproduce. Modelled nutrients are all tightly coupled together, implying that the nitrogen and phosphate cycles are insufficiently decoupled in the model. Phosphorus turnover is much faster than nitrogen cycling and we speculate that the relative timescales of recycling are incorrect and therefore we need a better understanding of benthic and pelagic (microbial loop and bacterial) nutrient recycling.

3.4 Spatio-temporal variability of misfit errors

The previous analysis identified how well the model simulates variables at the scale of the whole North Sea project. However what we need to know now is when and where are they good? Earlier PCAs (Fig. 3a, b) show that variability is highest in
winter and spring but does it mean that the model fit is better or worse? For each variable we calculated a simple error metric; the difference between observation and model (mismatch). These were normalised to a mean of zero and a standard deviation of one and PCA applied. The first 3 PCs describe 69.3% of the error space (Table 1). The fact that a substantial number of the misfits cluster around the origin implies that the model has skill. PC1 is associated with nutrients, PC2 with chlorophyll and salinity and PC3 with temperature and salinity (Table 1). Fig. 7a shows clear seasonal differences in errors associated with nutrients in winter and spring (PC1) and chlorophyll in spring (PC2). These outliers are primarily associated with the coastal zone (Fig. 7b) as defined by SOM analysis (Allen et al. 2007a). Essentially this says that the model performance is poor in the coastal zone in winter and spring. The errors in nutrients may be attributed to poorly resolved freshwater nutrient inputs; these regions are strongly influenced by the Elbe and Rhine. Alternative explanations may be errors in advective transport or the rate of benthic recycling. The errors in chlorophyll imply errors in the timing of the spring bloom. As these are well mixed, optically complex (Case II), waters it is reasonable to assume that poor resolution of the near shore light climate is an issue. Factors that need to be better understood include wind-wave resuspension of sediments and terrestrial sources of both SPM and coloured dissolved organic matter.

3.5 Do the misfits represent a poor model or poor observations?

MDS based on a Pearson correlation matrix shows relationships between model outputs, observational data and the misfits between them (Fig. 8). In addition to the basic patterns described in Fig. 6a it shows that for phosphate, sediment and chlorophyll the misfits are strongly correlated ($r^2 > 0.6$) with the corresponding model variable. Nitrate and silicate misfits are weakly correlated ($r^2 > 0.2$), and temperature and salinity misfits uncorrelated, with the corresponding model variables. The Pearson correlations between model and misfit ($\rho_m$) and observations and misfit ($\rho_o$) are given in Table 2 and illustrate strong negative correlations between misfit and model. This essentially says that the higher the model value of the variable the more likely it is to overestimate the observation.

In assessing the skill of our model we have used observations as a surrogate for the true state of the system and asked: How well does the model fit the data? Both model
predictions and the observations reside in a halo of uncertainty and the true state of the system is assumed to be unknown, but to lie within the observational uncertainty (Stow et al this volume). A model starts to have skill when the observational and predictive uncertainty halos overlap, in the ideal case the halos overlap completely. We make the distinction between the model having skill (i.e. it reproduces aspects of the observations) and predictive skill (i.e. simulations lie within the uncertainty halo of the measurements) This is illustrated in Figure 9a where the shaded area indicates the approximate region of predictive skill; where the model misfit is less that the halo of uncertainty of the observations. Figure 9b illustrates how this works in practice. If \( \rho \) between model and observations < 0.8 then the misfit dominates the observational errors implying that while the model has some skill (\( \rho > 0 \)), the model processes or parameters determining that variable need to be investigated. Only if \( r > 0.9 \) can we begin to attribute uncertainties to errors in the data which implies predictive skill; temperature is the only variable which falls into this category. Further investigation is required to determine whether this is a feature of this simulation, a feature of the model used or a generic response for all models.

4. Conclusions
Multivariate analysis allows us to explore complex relationships. We have demonstrated the application of a range of techniques and shown that we can reduce the dimensions of the problem and generate multivariate and univariate goodness of fit measures, in terms of both magnitude and trend. Additionally we can identify seasonal and regional variations in model performance. We demonstrate that further work is required to improve the accuracy of the modelling of chlorophyll and nutrients in coastal waters, particularly in late winter and spring, and our understanding of benthic pelagic coupling, nutrient cycling and the optical properties of coastal seas.

There are some caveats that need to be placed on this type of analysis. Care is required to choose the correct methods. The analysis relies on the observations being a reasonable indication of the true state of the system. This requires that the observations have an acceptable accuracy and precision and that they are representative at the time and space scales of the model. The final limitation is data availably; there are currently a limited number of coherent data sets with which we
can do this type of analysis. It is to be hoped that the current expansion of ocean and coastal observing systems, along with improvements in the quality of satellite data products, will help to address this situation.

Acknowledgements.

JIA and PJS were funded as a part of theme 9 of the NERC Oceans 2025 core strategic science program and thank all those involved in the ongoing development of POLCOMS-ERSEM, particularly J Holt, R Proctor (POL) and J Blackford (PML).

References


Table 1. Results from Principal Component Analyses of the observed and modelled variables and the misfit between them.

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th></th>
<th>Model</th>
<th></th>
<th>Misfit</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC1 PC2 PC3</td>
<td>PC1 PC2</td>
<td>PC1 PC2</td>
<td>PC1 PC2</td>
<td>PC1 PC2</td>
<td>PC1 PC2</td>
</tr>
<tr>
<td>Eigenvectors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHL</td>
<td>-0.025 0.664 -0.153</td>
<td>-0.119 -0.640 -0.491</td>
<td>-0.165 0.702</td>
<td>0.184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO₃</td>
<td>0.567 0.013 -0.069</td>
<td>0.510 -0.039 0.095</td>
<td>0.556 -0.052</td>
<td>0.153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SiO₄</td>
<td>0.544 0.002 -0.066</td>
<td>0.050 0.104 0.023</td>
<td>0.535 -0.150</td>
<td>0.083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO₄</td>
<td>0.393 -0.310 -0.081</td>
<td>0.489 -0.021 0.214</td>
<td>0.443 0.080</td>
<td>-0.274</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-0.133 0.490 0.270</td>
<td>-0.341 -0.297 0.480</td>
<td>0.073 -0.371</td>
<td>0.633</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>-0.441 -0.465 -0.047</td>
<td>-0.188 0.646 -0.466</td>
<td>-0.234 -0.508</td>
<td>-0.527</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPM</td>
<td>0.126 -0.082 0.941</td>
<td>0.292 -0.267 -0.506</td>
<td>0.347 0.281</td>
<td>-0.426</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>2.9 1.44 1.00</td>
<td>3.34 1.45 0.83</td>
<td>2.47 1.33</td>
<td>1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Variance</td>
<td>41.4 20.5 14.0</td>
<td>47.7 20.7 11.9</td>
<td>35.3 19.1</td>
<td>14.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>41.4 62.0 76.0</td>
<td>47.7 68.4 80.3</td>
<td>35.3 54.4</td>
<td>69.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 The correlations between normalised observed and modelled Pearson correlation resemblance matrices for each variable ($\rho_{fit}$) along with the correlations between normalised observation model misfit and modelled ($\rho_{m}$) and observed ($\rho_{o}$) Pearson correlation resemblance matrices for each variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Chl</th>
<th>NO$_3$</th>
<th>PO$_4$</th>
<th>SiO$_4$</th>
<th>T</th>
<th>S</th>
<th>SPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{fit}$</td>
<td>0.37</td>
<td>0.66</td>
<td>0.51</td>
<td>0.56</td>
<td>0.99</td>
<td>0.85</td>
<td>0.46</td>
</tr>
<tr>
<td>$\rho_{m}$</td>
<td>-0.96</td>
<td>-0.69</td>
<td>-0.91</td>
<td>-0.61</td>
<td>0.39</td>
<td>-0.36</td>
<td>-0.90</td>
</tr>
<tr>
<td>$\rho_{o}$</td>
<td>-0.01</td>
<td>0.08</td>
<td>0.09</td>
<td>0.31</td>
<td>0.52</td>
<td>0.18</td>
<td>-0.03</td>
</tr>
</tbody>
</table>
Table 3. Pearson correlation coefficient matrix for observed (O) and modelled (M) variables (subscript).

<table>
<thead>
<tr>
<th></th>
<th>Ochl</th>
<th>ONO3</th>
<th>OPO4</th>
<th>O_T</th>
<th>O_sp</th>
<th>Mchl</th>
<th>MNO3</th>
<th>MPO4</th>
<th>MT</th>
<th>Ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ochl</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ONO3</td>
<td>-0.10</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPO4</td>
<td>-0.26</td>
<td>0.53</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_T</td>
<td>0.11</td>
<td>-0.29</td>
<td>-0.14</td>
<td>-0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_sp</td>
<td>-0.31</td>
<td>-0.70</td>
<td>-0.63</td>
<td>-0.28</td>
<td>-0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mchl</td>
<td>0.37</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.16</td>
<td>0.12</td>
<td>-0.19</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNO3</td>
<td>-0.04</td>
<td>0.66</td>
<td>0.57</td>
<td>0.46</td>
<td>-0.54</td>
<td>-0.32</td>
<td>0.24</td>
<td>-0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPO4</td>
<td>-0.22</td>
<td>0.58</td>
<td>0.56</td>
<td>0.44</td>
<td>-0.51</td>
<td>-0.18</td>
<td>0.21</td>
<td>-0.28</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>0.11</td>
<td>-0.30</td>
<td>-0.16</td>
<td>-0.18</td>
<td>0.99</td>
<td>-0.05</td>
<td>0.12</td>
<td>-0.55</td>
<td>-0.53</td>
<td>-0.43</td>
</tr>
<tr>
<td>Ms</td>
<td>-0.33</td>
<td>-0.61</td>
<td>-0.51</td>
<td>-0.22</td>
<td>0.86</td>
<td>-0.13</td>
<td>-0.31</td>
<td>-0.39</td>
<td>-0.19</td>
<td>-0.35</td>
</tr>
<tr>
<td>M_sp</td>
<td>0.01</td>
<td>0.20</td>
<td>0.14</td>
<td>0.07</td>
<td>-0.14</td>
<td>0.17</td>
<td>0.46</td>
<td>0.11</td>
<td>0.40</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Map of the southern North Sea study region indicating the sample stations and key features of the system.

Figure 2. A schematic diagram of the pelagic components of the ERSEM model. Arrows indicate the direction of fluxes of elements.

Figure 3. PCA of samples based on a) observed and b) modelled chlorophyll, nitrate, silicate, phosphate, temperature, salinity and suspended sediment. The modelled values are the nearest equivalent to the observation in space and time. Data sets have been normalised prior to the analysis. PC1 and PC2 refer to the first and second principal components respectively.

Figure 4. Relationships between the observed and modelled principal components as a timeseries; a) first principal component PC1, b) second principal component PC2 and c) third principal component PC3. R = Pearson product moment correlation.

Figure 5. Histogram of the 100 permutations used to determine the significance of the RELATE test relationship between model and observation. The vertical line indicates the value of ρ.

Figure 6. MDS plot of a) Pearson correlation resemblance matrix and b) Spearman rank correlation resemblance matrix of both observed and model data indicating the relationships between patterns in modelled and observed variables. The variables included are chlorophyll (CHL), nitrate (N), phosphate (P), silicate (Si), salinity (S), temperature (T) and Suspended Particulate Matter (SPM). Contours defined by the square of the correlation coefficients indicate the strength of the relationships between variables.

Figure 7. PCA of samples based on data-model mismatch showing the spatial and temporal distribution of model errors. The modelled values are the nearest equivalent to the observation in space and time. Data sets have been normalised prior to the analysis. Labelled according to season (a) and region (b). The seasons are defined as follows, Su = summer (July, Aug, Sept), Au = autumn (Oct, Nov, Dec), W = winter (Jan, Feb, Mar) and Sp = spring (April, May, June). The regions are defined as offshore (0) and coastal (1), based on SOM classification (Allen et al 2007).

Figure 8. MDS plot based on Pearson correlations between observed and modelled variables and model-data error. The variables included are chlorophyll (CHL), nitrate (N), phosphate (P), silicate (Si), salinity (S), temperature (T) and Suspended Particulate Matter (SPM).
Figure 9. The relationship between the correlation of modelled and observed variables and the correlation of the modelled variables and the misfit. a) is a schematic diagram illustrating model skill, the shaded area indicates the approximate region where the model has predictive skill, i.e. when the misfit is smaller than the model error. B) shows the distribution of model variables in skill space. The variables included are chlorophyll (CHL), nitrate (N), phosphate (P), silicate (Si), salinity (S), temperature (T) and Suspended Particulate Matter (SPM).
Freshwater and nutrient inputs
Seasonally stratified
Coastal erosion
Frontal zone
Optically complex CASE II (SPM & CDOM)
Freshwater and nutrient inputs
$y = 0.7051x - 5E-15$

$R^2 = 0.4315$
y = 0.6299x - 3E-14

$R^2 = 0.3947$
$y = -0.044x - 2E-14$

$R^2 = 0.0023$
Normalise
Resemblance: Pearson correlation

Correlation

2D Stress: 0.15

TYPE
△ o
▽ m

STOP

N
P
S
Si
SPM
CHL
Normalise
Resemblance: Spearman rank correlation

2D Stress: 0.09

Correlation

- Green: 0.2
- Blue: 0.5

TYPE

- △ o
- ▽ m
Normalise
Resemblance: Pearson correlation

TYPE

- o
- m
- e

Correlation

- 0.2
- 0.5

2D Stress: 0.22
Model vs Observations

Model
Underestimates

Misfit > Obs error

Missfit < Obs error

Model
Overestimates

ρ Model vs Observations

ρ Model vs Misfit

ρ Model vs Observations

a)

b)