

Ocean Forecasting Systems – product evaluation and skill

Matthew Martin

Met Office, Fitzroy Road, Exeter, UK.

Abstract

The evaluation of output from ocean forecasting systems is important in order to inform users how much confidence can be placed in the products, and helps identify areas for improvement in the systems. An overview of the statistical methods which can be used to perform the evaluation is provided. Examples of some commonly used methods from various GODAE systems are given, including evaluation of large-scale model performance, the use of output from data assimilation systems, the use of independent data, and comparison of forecasts with analyses.

Introduction

The aim of ocean forecasting systems is to provide information about the past, present and future state of the ocean to a range of users. There are a wide range of applications including defence, ship routing, oil spill prediction, weather forecasting, climate monitoring and scientific research. In order for the outputs produced by ocean forecasting systems to be useful for these applications, the ability of the systems to represent the real world must be assessed. This will inform the users of where and when the products can be used, and with how much confidence. It also aids development of the forecasting systems themselves, highlighting areas where improvements can be made.

The state variables from ocean forecasting systems are the sea surface height and the three-dimensional temperature, salinity and currents. For those systems which include sea-ice models, the sea-ice concentration, velocity and thickness are also produced. Other diagnostic quantities such as mixed-layer depth and transports are also of interest to users of model output. The applications which use this ocean information cover a wide range of time and space scales from large-scale climate monitoring and seasonal forecasting applications, which look at evolution

over months with basin-wide and global coverage, through to the analysis and prediction of mixed-layer depth over a diurnal cycle. This range of spatial and temporal scales must therefore be taken into account when assessing the products.

The amount of information available from ocean forecasting systems is immense - the model state vector usually contains at least of the order of 10^7 variables at any particular time (and for some ocean forecasting systems significantly more than this). It is usually impossible for users of the output to access all of this data, and so some post-processing is often performed to synthesise this information. This may involve interpolation or averaging in space and time, and may also involve production of other diagnostic information which is of more relevance to a particular user. The impact of the post-processing on the accuracy of the data which are provided must be assessed, usually by assessing the post-processed fields directly.

In order to evaluate products from ocean forecasting systems and relate them to the real world, observations are required. These could take the form of climatologies, analyses of satellite data, or raw observed values. In all cases, the accuracy of the observations needs to be assessed and taken into account when performing the comparison between model and observations. It is also important to use observations which have been quality controlled – comparison with “bad” observations can lead to confusing results.

A number of aspects affect the quality of the output from ocean forecasting systems. The most obvious is the quality of the model used to produce the forecast, including its horizontal and vertical resolution, and the parametrisations which are used. The surface forcing fields used to drive the model (or in the case of a coupled model, the quality of the atmospheric model) also have a significant impact on the quality of the ocean forecast. For regional models, lateral boundaries can play a significant role. The data assimilation scheme used to initialise the model has a large impact on the accuracy of the forecast - the type and number of observations being used in the assimilation, the assimilation scheme itself, and the quality control of observations all have an impact on the accuracy of the analysis and the subsequent forecast.

A review of some statistical concepts which are required to assess model output is given in the next section. A summary of the main issues with the observations available for use in the evaluation of model products is then given, followed by some specific examples of product evaluation. An overall summary is then given.

Statistical concepts

A number of statistical measures are required to thoroughly assess the output of ocean forecasting systems. Three different concepts are described here, aimed at determining the *accuracy* of the analysis and forecast, the ability of the model to represent the *patterns* of the observations, sometimes termed association (Murphy *et al.* 1995), and the *skill* of the forecast. Representing this information in a con-

cise way can be done through some well-known summary diagrams which are briefly described.

Accuracy

We assume that there is a verification data set consisting of a set of N observations, y_i , $i = 1, \dots, N$, with mean value \bar{y} . The model values at the same time and location as these observation are denoted x_i , $i = 1, \dots, N$ and have mean value \bar{x} . The accuracy of the forecast is usually assessed using the root mean square differences between the model and observed values:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2}. \quad (1.1)$$

The mean difference is also a useful measure of the ability of the model to represent the mean observed state:

$$MD = \frac{1}{N} \sum_{i=1}^N (x_i - y_i). \quad (1.2)$$

Pattern

The ability of the model to reproduce the pattern in the observations can be measured using the correlation coefficient:

$$R = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}, \quad (1.3)$$

where σ_x and σ_y are the standard deviations of the model and observations respectively.

The correlation coefficient provides information about whether the patterns in the model are similar to the patterns of the observations, but not about the amplitude of variation in the two fields. It reaches a value of 1 when the two fields have the same centred pattern of variation, a value of -1 when the two fields vary in the opposite sense to each other, and a value of zero when no correlation exists be-

tween the two fields. The square of the correlation coefficient, R^2 , is also a useful quantity as it provides information on the fraction of the variance explained.

When the dominant source of variability in a field is a large scale signal, for instance the seasonal cycle, most ocean models would easily reproduce the signal, resulting in high values of R . However, ocean forecasting systems produce information at smaller temporal and spatial scales. To assess these, it is instructive to calculate the anomaly correlation coefficient:

$$ACC = \frac{\sum_{i=1}^N (x_i - C_i)(y_i - C_i)}{\sqrt{\sum_{i=1}^N (x_i - C_i)^2 \sum_{i=1}^N (y_i - C_i)^2}}, \quad (1.4)$$

which provides information about the ability of the model forecast to reproduce the observational information when the seasonally varying climate signal, denoted C , has been removed.

Skill

Determining the skill of a model forecast is dependent on the application and it is not possible to define one skill score that is universally appropriate. A number of scores have been suggested in the literature, some examples of which are given below.

The skill of a forecast can be defined as the accuracy of the forecast relative to the accuracy of a reference field such as a climatology or persistence (Murphy, 1995). A simple way of measuring this is given by:

$$SS1 = 1 - \frac{MSD}{MSD_{ref}}, \quad (1.5)$$

which measures the relative accuracy of the forecast to some reference, where MSD indicates Mean Square Difference (the square of equation 1.1) and the subscript *ref* indicates that the model value in equation 1.1 has been replaced by a climatology or persistence estimate. A value of 1 implies that the forecast has perfect skill while a value of zero implies no skill.

In the above skill score, no account is taken of correlations or bias. Taylor (2001) suggests the following score which is based on the correlation coefficient and the model and observed variances:

$$SS2 = \frac{4(1+R)}{(\hat{\sigma}_x + 1/\hat{\sigma}_x)^2(1+R_o)}, \quad (1.6)$$

where R_o is the maximum correlation attainable and $\hat{\sigma}_x = \sigma_x / \sigma_y$ is the normalised standard deviation.

Another skill score which uses the correlation, variances, and also includes the biases in model and observations, as suggested by Metzger *et al.* (2008), is given by:

$$SS3 = R^2 - [R - (\sigma_y / \sigma_x)]^2 - [(\bar{y} - \bar{x}) / \sigma_x]^2. \quad (1.7)$$

For probabilistic forecasting systems, a wide range of skill scores are often used such as the Brier skill score (Brier, 1950), or the Relative Operating Characteristics (ROC) score which is used to determine the relationship between the number of events which were correctly forecast to the number of false alarms. These skill scores are widely used in seasonal prediction systems and ensemble weather forecasting systems, but few short-range ocean forecasting systems currently produce ensemble forecasts. These skill scores are not covered in depth here - see Atger (1999) and references therein for further information.

Summary diagrams

In order to characterise the differences between the model and observations it is important to take into account the correspondence in both the patterns and the variances of the two fields. We define the centred pattern RMSD as:

$$CRMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N [(x_i - \bar{x}) - (y_i - \bar{y})]^2}. \quad (1.8)$$

Taylor (2001) noticed that a simple relationship exists between the correlation coefficient, the centred pattern RMS difference, and the variances of the fields in question. The relationship is given by:

$$CRMSD^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_yR, \quad (1.9)$$

which takes the same form as the law of cosines. This relationship can be used to plot the information about R, CRMSD and the variances in the model and obser-

vations as a point on a single diagram. In order to make it possible to compare fields with different units, the statistics can be non-dimensionalised by normalising each variable in equation (1.9) by the standard deviation in the observed field, which leaves the correlation coefficient unchanged. A schematic Taylor diagram is shown in Fig. 1. If the model exactly reproduced the observations, it would lie at the point indicated by the black circle. The distance between this black circle and the actual model point (the blue diamond in this example) represents the CRMSD and the dotted arcs on the diagram represent lines of constant CRMSD. The correlation coefficient is represented on the outer arc of the diagram with increasing correlation with the angle from the y-axis. The normalised standard deviation is represented as the distance to the origin, with a ratio of one denoted by the dashed arc (if the point is closer to the origin the model has lower variance than the observations). The power of the Taylor diagram lies in the ability to plot numerous model runs on a single diagram and to compare these various aspects of the models' performance.

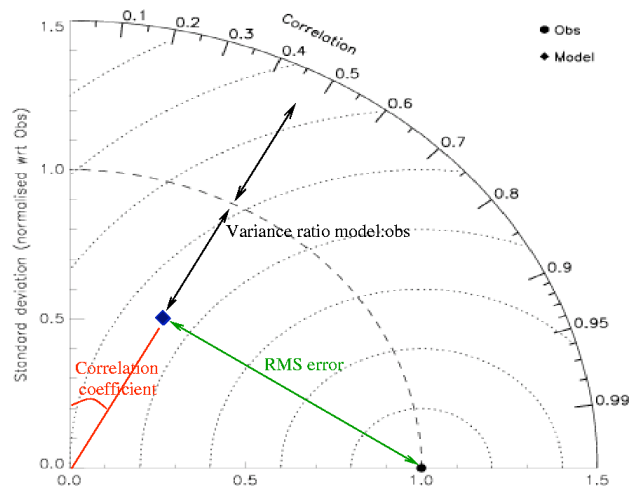


Fig. 1. Schematic description of a Taylor diagram.

One drawback of the Taylor diagram is that the mean error of the models is not accounted for. The so-called Target diagram (Jolliff et al., 2009) can be used to represent complementary information about the statistical performance of models. In this case, the relationship between the total mean square difference and the unbiased MSD and bias, $RMSD^2 = MD^2 + CRMSD^2$, is plotted on a diagram

where the x-axis represents *CRMSD* and the y-axis represents the bias. Since *CRMSD* is a positive quantity by definition, the negative x-axis can be utilised to include information about the standard deviation difference by multiplying the *CRMSD* by the sign of the standard deviation difference.

Observations

Various observation types are available for use in validating and verifying ocean forecasting systems and are detailed elsewhere in this summer school. Some general points about the use of these data in evaluating model output are outlined here.

For satellite data, a number of levels of processing are performed to produce observations of the quantities which are output by ocean models. For instance, sea surface temperature (SST) data undergoes various levels of data processing from the level 1 brightness temperatures measured by the satellites, through the level 2 conversion to sea surface temperature at the native resolution, the level 3 re-gridding of the data, through to the level 4 objective analyses. Each level of processing affects the accuracy and representativeness of the data so it is important to be clear what the observations are representing before using them for evaluation.

It is important to recognise that the observations used in model evaluation are not themselves perfect representations of the true state of the ocean. Measurement techniques will introduce some error into the observations. The observations are also usually made at a specific location whereas the model represents an area-average value. This means that the model cannot represent all processes affecting the observations. These errors of representativity should also therefore be taken into account when assessing the results of any model-observation comparison.

As well as the random errors described in the previous paragraph, observations often report erroneous values. This can happen for a number of reasons such as mis-reporting of location, corruption of the observation during transmission, or instrument error. One or two bad observations can significantly impact the results of any validation/verification, so it is important that a thorough check on the quality of the data is performed prior to the evaluation. This quality control can be performed in a number of ways, but usually consists of a comparison between the observations and some reference field either from a model forecast, or from observed climatology (see for example Ingleby and Huddleston, 2007).

Evaluating ocean analyses and forecasts

The usual process for developing a new ocean forecasting system, or significant upgrades to an existing system, involves a number of stages. Scientific developments will be tested individually to ensure that they are producing the expected

change in the system. Once a number of developments are available, they will be put together into a new version of the system and this must then be thoroughly evaluated during the validation phase. This validation is usually done by means of the evaluation of a set of hindcasts of the system, where the system is run over a multi-annual period in the past. This tests that the overall changes to the system produce the expected improvements. Once the validation has been carried out, the system can be implemented operationally. At this stage it is important to continuously assess and monitor the accuracy of the system using a verification system. The results of both the validation and verification are useful for providing information to users of the system about the expected accuracy. User-specific evaluations can also be carried out to assess the suitability of the system for a given application.

A number of examples of evaluating ocean forecasting systems are given below taken from various sources (e.g. Ferry *et al.*, 2007; Oke *et al.*, 2008; Metzger *et al.*, 2008 and 2009; Storkey *et al.*, 2009), providing illustrations of some commonly used methods. The advantages and shortcomings of each method are outlined.

Evaluating the large-scale mean and variability

It is important to check that the average properties of the ocean forecasting systems are providing a good representation of the ocean climate. This is usually done by comparing multi-annual averages to climatologies generated from observational data-sets.

One example of this is a comparison between a mean dynamic topography (MDT, such as that of Rio, 2005, or Maximenko and Niiler, 2005), with the model's average sea surface height field. This provides a useful guide as to the ability of the model to represent the large scale ocean circulation (see for example Metzger *et al.* 2008).

Temperature and salinity can also be assessed using a suitable climatological data-set. In Fig. 2, the annual mean temperature anomalies from the World Ocean Atlas 2005 (Locarnini *et al.* 2006) are shown as cross-sections from two hindcasts of the FOAM system, one without assimilation and one with. This shows that the data assimilation is able to reduce the drifts of the model away from climatology. One has to be careful when performing these comparisons that any inter-annual signal is not contaminating the results. For example in Fig. 2(f), there is a clear La Nina signal, where the model is representing the true deviation from climatology.

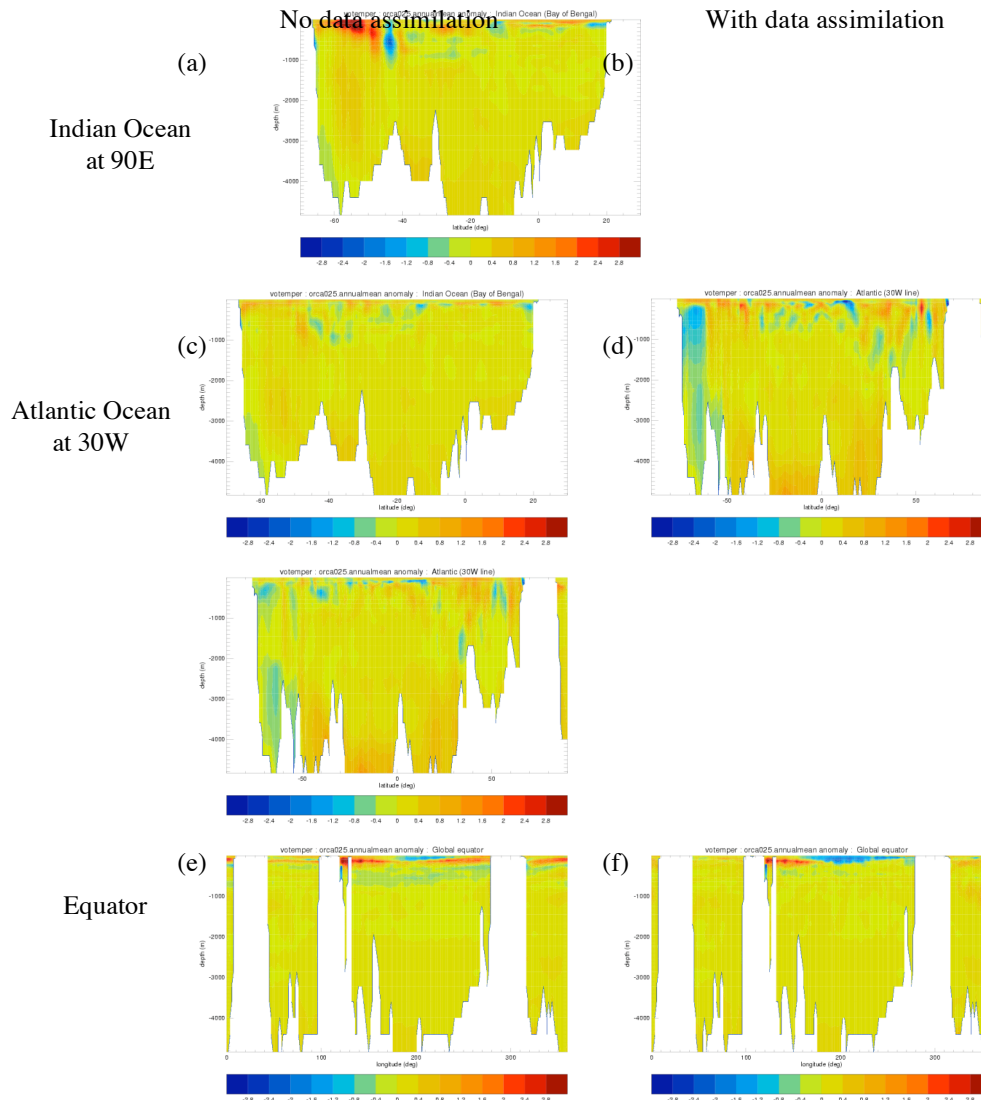


Fig. 2. Annual mean temperature anomalies from WOA05 climatology for 2008 from FOAM both without (a,c,e) and with (b,d,f) data assimilation. (a,b) show a cross-section in the Indian Ocean along 90E. (c,d) show a cross-section in the Atlantic Ocean along 30W, (e,f) show a cross-section along the equator.

The variability in the model and observations can also be assessed. For instance, sea surface height (SSH) can be used as a measure of the amount of

mesoscale activity. This can be estimated from observations provided by satellite altimeters, and also from ocean models. Figure 3 shows an example of this from the GLORYS reanalysis produced by Mercator using the $\frac{1}{4}$ degree resolution NEMO model with data assimilation. The standard deviations of the data are shown next to the standard deviation of the model fields from a 6 year period. The model analyses are reproducing the observed variability very well, including the western boundary currents which are difficult areas to accurately represent the mesoscale variability with $\frac{1}{4}$ degree resolution. The only regions where the model variability is significantly different to the observed are in the Brazil-Malvinas Confluence region and in parts of the South Pacific.

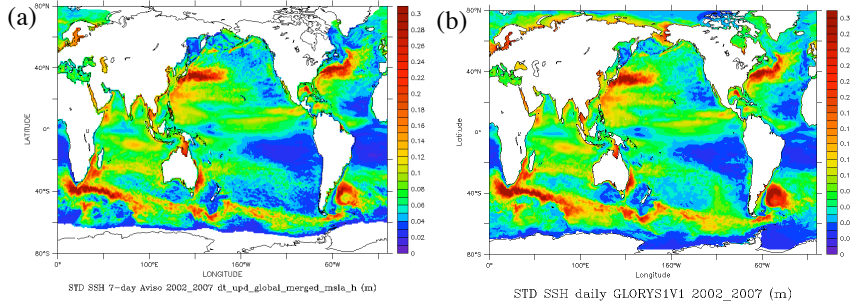


Fig. 3. Standard deviation of SSH for the period 2002 – 2007 from (a) Aviso data and (b) from the GLORYS reanalysis product.

Both the average and variability comparisons described above are useful as a first-order check on the ability of the model to represent the large-scale ocean features, and can give confidence that the model is behaving as expected. However, they do not give information about the accuracy or skill of the model and so are of limited use to most users. More detailed investigations are required for this, and are described below.

Data assimilation statistics

In the data assimilation process, the observation operator h is used to interpolate the model forecast field \mathbf{x}^f to the location in time and space of the observations, \mathbf{y} . This enables calculation of the innovations, $\mathbf{d} = [\mathbf{y} - h(\mathbf{x}^f)]$. Once the data assimilation has been performed it is also possible to calculate the equivalent using the analysis field to produce the residuals, $\mathbf{r} = [\mathbf{y} - h(\mathbf{x}^a)]$. The reduction in the errors between the analysis and the forecast can be used as an *a posteriori*

check that the data assimilation process is working as expected, and is fitting the observations to within their error (see for example Cummings, 2005).

The increments generated through the data assimilation process also provide an important source of information. The time-average of these increments can indicate areas of significant model bias. However, it is not always obvious how to diagnose the source of these biases.

For validation and verification of the model forecast, it is the innovation statistics that are of most interest, as they provide a pseudo-independent check on its accuracy. The observations being used for this comparison have not previously been assimilated so from that point of view are independent. However, previous observations of the same type will have been assimilated on previous data assimilation cycles so they cannot be viewed as completely independent.

An example of the innovation statistics from the GLORYS reanalysis system produced using the Mercator system is shown in Fig. 4. These include the mean and RMS of the innovations for SSH and for temperature. The mean errors show that the system is able to represent the global average observed SSH and temperature well. The RMS of the innovations provides a measure of the overall accuracy of the system both as a function of time, and of depth (for temperature). These time-series plots also illustrate the stability of the system, with the SSH being relatively stable, whereas the temperature errors have a clear seasonal cycle.

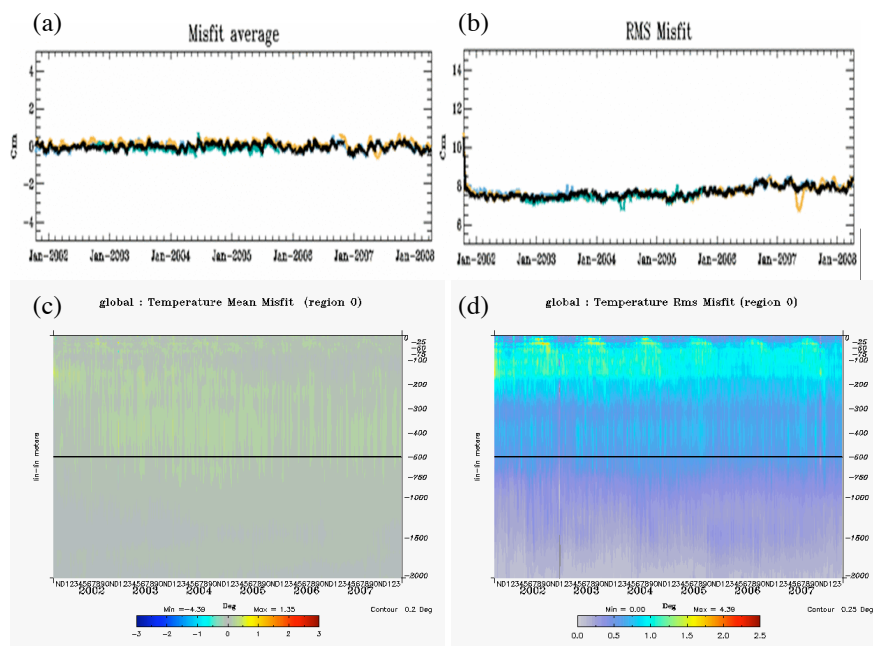


Fig. 4. (a,c) Mean and (b,d) RMS of the innovations for (a,b) SSH and (c,d) temperature for the GLORYS reanalysis system.

An example of the use of Taylor diagrams for plotting innovation statistics is shown in Fig. 5 with results from a hindcast run of the FOAM system (Storkey *et al.*, 2009). This shows the statistics for a number of different regions for both SST and SSH. The SST statistics are only shown for a comparison with the AATSR data, although other satellite SST data were assimilated. The variability in both these variables is well-reproduced by the model in all regions, but the correlations and RMS differences are clearly regionally dependent with the Mediterranean region having the largest RMS errors and lowest correlation coefficient.

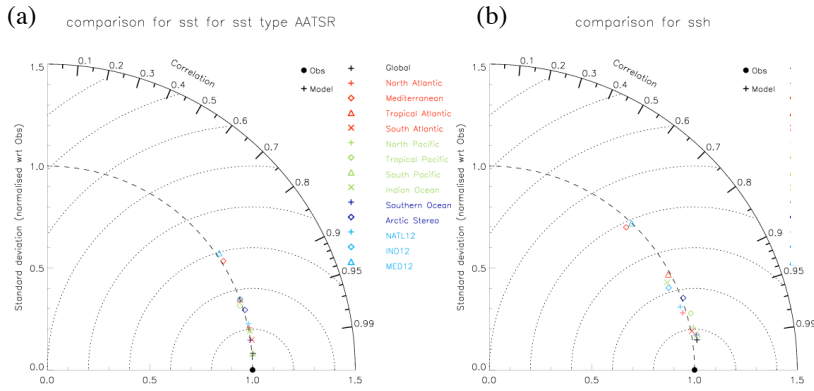


Fig. 5. Taylor diagrams from a 2-year hindcast of the FOAM system for (a) SST comparison with AATSR data and (b) SSH comparison with along-track altimeter data. The different colours and symbols represent the statistics for different geographical regions.

Evaluation of analyses and forecasts using independent data

In most operational assimilation systems, the aim is to provide the best possible estimate of the ocean state, and so all available data are assimilated. However, some data-sets are not available in real-time and so can be used in delayed mode to validate the results. An example of this is the RAPID array which measures sub-surface ocean properties in the North Atlantic in order to produce estimates of the Atlantic Meridional Overturning Circulation (AMOC).

Qualitative inter-comparisons can be made between ocean model output of SSH and satellite ocean colour data (see for example Storkey *et al.* 2009). These can help to show the performance of the systems in reproducing the position of mesoscale eddies and fronts, but it is difficult to produce robust quantitative statistics using this sort of technique.

A method which is often used to validate ocean models in a hindcast setting is to withhold certain data from the data assimilation, and use this independent data for validating the results. This is a useful technique as it provides an independent check that the data assimilation system is working as expected. It is not possible to use this to assess the overall accuracy of the system as the unassimilated data would be assimilated in the operational system, but it can give a bound on the expected accuracy.

An example of this technique is shown in Fig. 6 which shows results from the BlueLink Reanalysis (BRAN) system (Oke *et al.* 2008). Here some unassimilated Argo profiles are used to assess the RMSD in the assimilation run and the run without data assimilation. In all regions at almost all depths, the assimilation is improving the model's representation of sub-surface temperature when compared to the non-assimilating model.

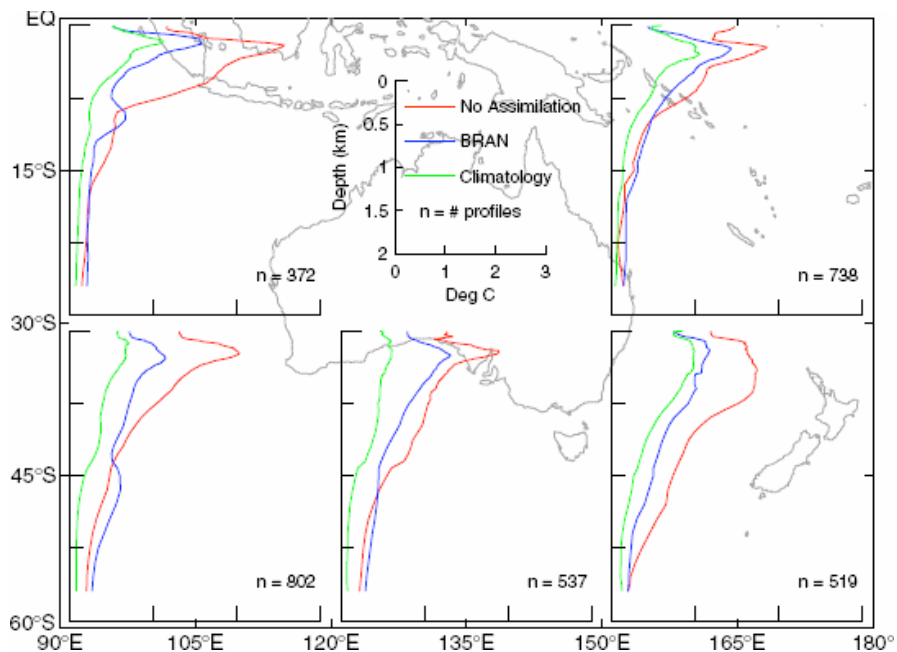


Fig. 6. Depth profiles of the RMSD between observed T profiles from Argo and the spin-up run (red; with no data assimilation), BRAN1.5 (blue) and climatology (green), using withheld profiles for the period January 2003–December 2005 (the overlapping period for BRAN1.5 and the spin-up run). Plots are shown for each 30°×30° region around Australia. The central panel over Australia shows the legend and the axis dimensions for all plots.

Some data-sets provide information about variables which are not assimilated in most ocean forecasting systems at present. For instance, most of the current operational forecasting systems don't assimilate velocity data. Direct measurements

of velocity are sparse, but there are some data in the tropical moorings and other time-series stations. There are also measurements of velocity from surface drifting buoys and these provide near-global coverage. These can be used as an independent check on the surface ocean currents, an important variable for a number of users.

Surface drifters consist of a surface buoy which is attached to a subsurface drogue. This drogue is usually centred at 15m depth. The buoy measures temperature (and sometimes other ocean properties) and the position of the drifter is usually inferred from satellite transmission information. The SST data and position of the drifter are disseminated via the global telecommunications system (GTS).

Three months of data from 1st January – 31st March 2006 were quality controlled by checking the SST against climatology using a Bayesian technique, and by checking that the average daily velocity of the floats did not exceed 2ms^{-1} . The daily mean velocity values from drifter data were calculated by estimating the distance in the latitudinal and longitudinal directions between the first and last float positions during each day, and dividing the distance by the difference in their reporting time. The modelled velocity corresponding to the observed velocity was calculated by interpolating the model's daily mean velocity fields to all of the observed drifter locations using a bilinear interpolation, and averaging the values for each day.

There are a number of issues with estimated velocities from surface drifters, for example aliasing of inertial oscillations, inaccuracy of position data, unknown drogue depths, un-drogued data and different reporting frequencies. The technique described in the previous paragraph also introduces errors as the curvature in the path of the drifter is not taken into account. Other techniques for comparing the model and observed velocities exist. For example one could input the starting position for each drifter on a particular day, run the model forward to estimate its position at the end of the day, and compare that with the final observed position of the drifter. Statistics on these position errors could then be calculated and assessed.

Various experiments were performed with the FOAM system (as it was in 2006, see Martin *et al.* 2007 for details) in order to assess the impact of different aspects of the system on the surface currents. Figure 7 shows the Taylor diagrams for a sample of these experiments for the u and v components of the velocities in the North Atlantic. The first experiment (in light blue) was a re-run of the operational FOAM system which shows that the variability in the model was close to the observed variability but that the correlation was very low with a fairly high RMSD. When not assimilating altimeter data (dark blue), the model's variability is much less, but the correlation coefficient is even worse. This implies that the altimeter assimilation is adding in variability to the model which is not naturally included in the model. One way of getting round this problem is to increase the viscosity in the model so that any spurious variability is damped. The results from a run of FOAM with an increased viscosity are shown in green. For comparison, the results from HYCOM and Mercator are also shown in yellow and orange respec-

tively. This shows improvements in the correlation and reduced RMSD compared to the other FOAM runs, giving similar results to HYCOM and Mercator.

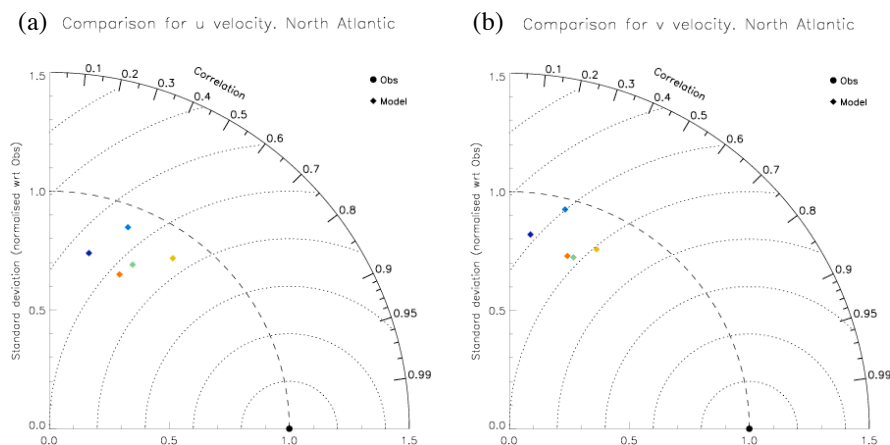


Fig. 7. Taylor diagrams for the (a) u and (b) v components of surface currents for various model runs during the period 1st Jan – 31st March 2006 compared to velocity from surface drifters. Dark blue – FOAM with no altimeter assimilation; light blue – FOAM with altimeter assimilation; green – FOAM with altimeter assimilation and increased viscosity; yellow – HYCOM; orange – Mercator.

Forecast versus analysis

In order to assess the forecasts from ocean models, one can assume that the analysis produced by the data assimilation is providing a “best estimate”. The subsequent forecast can be compared against the analysis (at the correct time), and the differences between these fields can be used, over a large number of realisations, to assess the skill in the model forecast. Various statistics can be calculated based on these differences; the most commonly used are RMSD, mean and anomaly correlations, as described previously. It should be noted that these do not give the overall magnitude of the errors, as the analysis errors are not included, but they do provide information about the evolution of errors in time. The analysis errors should be computed separately (as described previously) and used in conjunction with these errors to provide information about the overall error in the forecasts.

An example of the growth in the SSH forecast errors from the HYCOM/NCODA system (Metzger *et al.*, 2009) is shown in Fig. 8 for various regions. Here, the median ACC and RMSD statistics are plotted as a function of

forecast length out to 14 days. Globally, the model forecasts clearly have higher ACC and lower RMSD than the persistence forecasts throughout the 14-days. The picture is slightly different when looking at particular regions however. For instance, in the Kuroshio region, the forecast model is not providing much more skill than persistence due to the fact that the flow is dominated by mesoscale flow instabilities (rather than being dependent on the atmospheric forcing), although both forecast and persistence are more accurate than climatology throughout the period. In the Yellow Sea region where the ocean responds rapidly to the atmospheric forcing, a persistence forecast quickly becomes no better than climatology, whereas the forecast retains some skill out to at least 5 days.

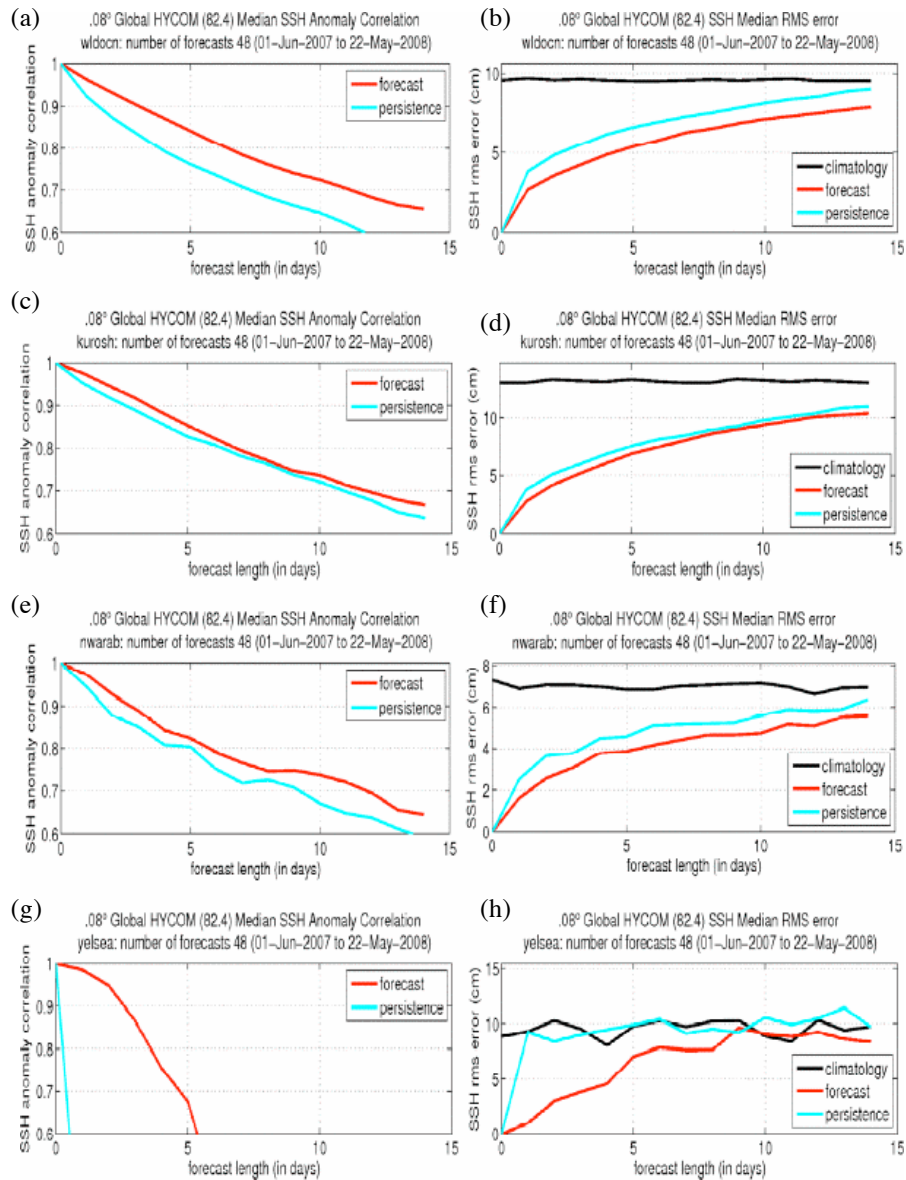


Fig. 8. Median SSH anomaly correlation (left column) and median SSH RMSD (right column) against the verifying analysis as a function of forecast length for the global ocean (entire domain – top row), the Kuroshio (120-179°E, 21-55°N – second row), the northwest Arabian Sea (51-65°E, 15-26°N – third row) and the Yellow Sea (118-127°E, 30-42°N – bottom row). The red curves are HYCOM/NCODA forecasts, the cyan curves are for persistence of the nowcast and the black curves of RMSE are for the hindcast annual mean.

Another example of comparing forecasts with analyses is shown in Fig. 9 which shows August 2009 monthly average 5-day temperature forecast-analysis differences from the FOAM system at 25m and 50m depths. A number of features are apparent in these figures, but we focus here on the main broad-scale signal: at 25m depth there is a clear negative bias in the northern mid-to-high latitudes, with a corresponding warm anomaly at 50m depth. This dipole pattern indicates that heat is being mixed too vigorously in the model. This suggests that either the wind forcing is too strong, or the mixing scheme in the model is not representing the real-world mixing correctly. It is possible to independently validate the wind forcing, for example using scatterometer data. In this case it is thought that the main problem lies with the model's mixing scheme, so the focus of model development here will be to improve this aspect of the model.

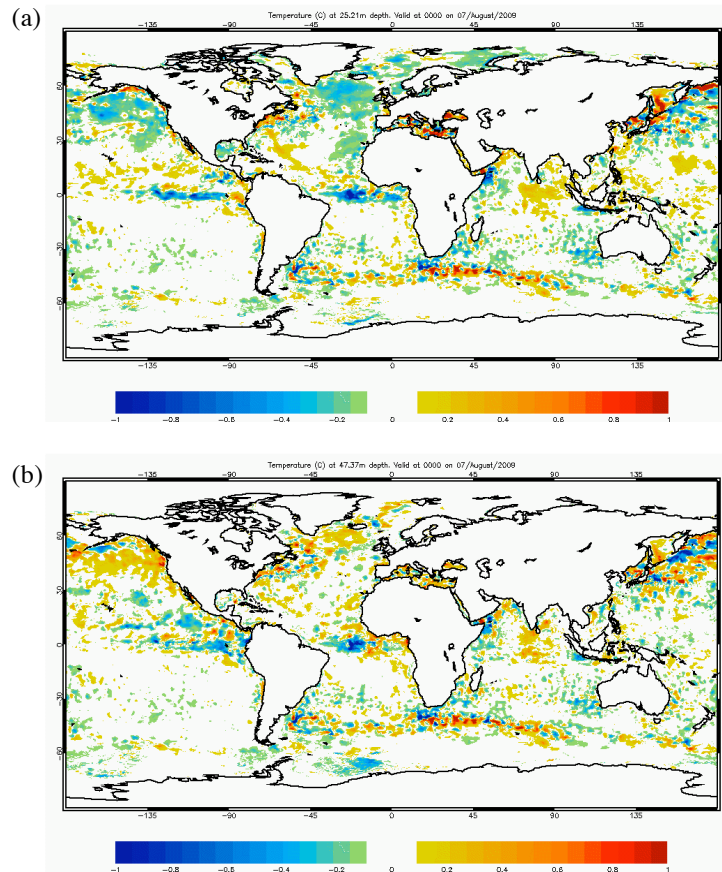


Fig. 9. Monthly average 5-day forecast temperature differences for FOAM for August 2009, compared with analyses at (a) 25m and (b) 50m depth.

Case studies for particular applications

As described previously, ocean forecasting systems serve a large number of users. Among the most significant of these are the Navies, who are interested in a number of different outputs including information about sound speed in the ocean in order to model the acoustics (Metzger, *et al.*, 2008, 2009). In order to produce accurate sound speed estimates, the temperature and salinity fields must be accurately determined, with the mixed-layer depth (MLD) and sonic-layer depth (SLD, Millero and Li, 1994) of particular interest (amongst other parameters).

Metzger *et al.* (2008, 2009) investigate the accuracy of the MLD and SLD forecasts in the HYCOM/NCODA system used by the US Navy. An example of this validation is shown in Fig. 10 which shows the mean and RMS errors in SLD as a function of forecast time for three regions. This shows that the model forecast and persistence are both producing more accurate estimates of SLD than is available from climatological estimates throughout the 14-day forecast. The skill of the model is generally similar to that of persistence, although this result is regionally dependent. The RMS errors general show a large amount of variability which is most likely due to vertical interpolation errors, and could also be due to observation sampling issues.

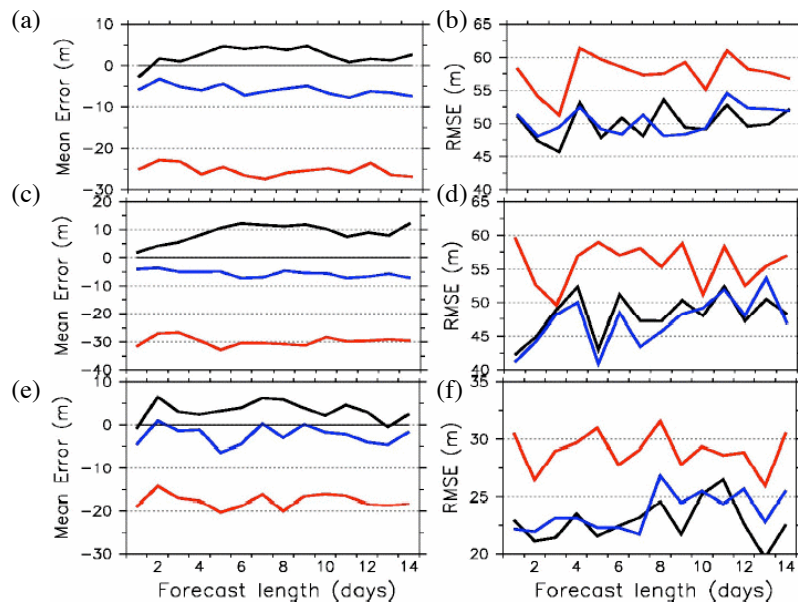


Fig. 10. Error analysis of sonic layer depth (m) as a function of forecast length based on 48 14-day forecasts by HYCOM/NCODA for regions MER4d (top), the western Pacific (middle) and the Arabian Sea (bottom). The left column shows

mean error and the right column shows RMSD. The black curves are for HYCOM/NCODA forecasts, the blue curves are for persistence of the nowcast ocean state and the red curves are for the GDEM3 climatology. Note the y-axis differs between most plots.

Summary and conclusions

An overview of methods which can be used to evaluate the accuracy and skill of ocean forecasts has been presented. Various statistical methods which can be used to perform evaluations have been defined, together with some useful diagrams for summarising related statistical information. A discussion on the importance of knowledge about the accuracy and quality of the observations used in the evaluations has also been given.

Some examples of the application of the various statistical measures to GODAE ocean forecasting systems have been given. These were used to highlight the need to evaluate the ability of the model to reproduce the large-scale ocean circulation, the accuracy of the analyses, and the accuracy of the subsequent forecasts. The use of independent data in assessing analyses and forecasts has also been presented, as has an example of validation directed at a particular user need.

Various techniques which could be used to evaluate ocean forecasting systems have not been described in detail for different reasons. For example, it is possible to estimate a formal error estimate of the analysis using the Hessian of the cost function in variational data assimilation schemes. However, this is an expensive quantity to calculate and the output of the calculation is dependent on the input error covariance information which is usually not well known. For these reasons, it is not usually calculated explicitly.

Similarly, for systems which run an ensemble of forecasts, the spread in the forecasts can be used to provide an estimate of the confidence which should be placed in the forecasts. The uncertainty in the initial conditions and the processes and parameterisations which are modelled can be sampled and the spread of the forecasts can then give statistical information on how much confidence should be placed in certain regions. However, the way in which the uncertainties in the system are sampled has a significant impact on the resulting forecast error estimates, and few operational ocean forecasting systems run an ensemble prediction system at present.

Inter-comparison with other ocean forecasting systems can also provide useful information about the skill of a particular ocean forecasting system and insight into weaknesses that can easily be corrected. For more information on this subject, the reader is directed to the separate paper on inter-comparison methods in this summer school.

The evaluation of ocean forecast products is an important aspect of all the GODAE systems, and is continually being improved. It is hoped that common

verification statistics will be produced routinely by all the systems over the coming years which will drive improvements to the systems themselves, and will also provide further insight into the most appropriate methods for their evaluation.

References

- Atger, F., 1999. The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941-1953.
- Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.
- Cummings, J.A., 2005. Operational multivariate ocean data assimilation. *Q. J. R. Meteorol. Soc.*, **131**, 3583-3604.
- Ferry, N., Rémy, E., Brasseur, P., Maes, C., 2007. The Mercator global ocean operational analysis system: Assessment and validation of an 11-year reanalysis. *J. Mar. Syst.*, **65**, 540-560.
- Ingleby, N.B., Huddleston, M.R., 2007. Quality control of ocean temperature and salinity profiles – historical and real-time data. *J. Marine Syst.*, **65**, 158-175.
- Joliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R., 2009. Summary diagrams for coupled hydrodynamic-ecosystems model skill assessment. *J. Mar. Syst.*, **76**, 64-82.
- Locarnini, R.A., Mishonov, A.V., Antonov, J.I., Boyer, T.P., Garcia, H.E., 2006. World Ocean Atlas 2005, Ed. Levitus S. NOAA Atlas NESDIS 61, U.S. Government Printing Office, Washington, D.C., 182 pp.
- Martin, M.J., Hines, A., Bell, M.J., 2007. Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. *Q. J. R. Meteorol. Soc.*, **133**, 981-995.
- Metzger, E.J., Hurlburt, H.E., Wallcraft, A.J., Shriver, J.F., Smedstad, L.F., Smedstad, O.M., Thoppil, P., Franklin, D.S., 2008. Validation Test Report for the Global Ocean Prediction System V3.0 – 1/12° HYCOM/NCODA: Phase I. Memorandum Report No. NRL/MR/7320--08-9148, Naval Research Laboratory, Oceanography Division, Stennis Space Center, MS 39529-5004.
- Metzger, E.J., Hurlburt, H.E., Wallcraft, A.J., Shriver, J.F., Townsend, T.L., Smedstad, O.M., Thoppil, P., Franklin, D.S., 2008. Validation Test Report for the Global Ocean Forecast System V3.0 – 1/12° HYCOM/NCODA: Phase II. Memorandum Report No. NRL/MR/7320--09-9236, Naval Research Laboratory, Oceanography Division, Stennis Space Center, MS 39529-5004.
- Millero, F.J., Li, X., 1994. Comments on "On equations for the speed of sound in seawater". *J. Acoust. Soc. Am.*, **95**, 2757-2759.
- Murphy, A.H., 1995. The coefficients of correlation and determination as measures of performance in forecast verification. *Wea. Forecasting*, **10**, 681-688.
- Oke, P.R., Brassington, G.B., Griffin, D.A., Schiller, A., 2008. The Bluelink ocean data assimilation system (BODAS). *Ocean Modelling*, **21**, 46-70.
- Storkey, D., Barciela, R.M., Blockley, E.W., Furner, R., Guiavarc'h, C., Hines, A., Lea, D., Martin, M.J., Siddorn, J.R., 2009. Forecasting the ocean state using NEMO: The new FOAM system. Submitted to *J. Operational Oceanogr.*
- Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geo. Res.*, **106**, 7183-7192.