Demonstration of Comparison Methods for Ocean Model Validation

Karina Hjelmervik[1,*] and Karl Thomas Hjelmervik[2]

[1]*Faculty of Techonology and Maritime Sciences, Vestfold University College, 3103 Tønsberg, Norway*

[2]*Norwegian Defence Research Establishment, 2027 Kjeller, Norway*

Abstract: In recent years an increasing amount of oceanographic data has become available. This includes observations as well as data from hydrodynamic ocean models. Validation is required for establishing the necessary confidence in new sources of data. Generally ocean models and other data sources such as satellite imagery are validated by comparing the output to conventional observations or the output of established ocean models.

Methods of comparison used in literature range from refined statistical methods to comparisons of snapshots. This work collects descriptions of some of the most widely used comparison methods. The capabilities and limitations of each method are demonstrated using examples from modelled and observed oceanographic data. The work has a particular focus on how to determine discrepancies on vertical gradients in the oceanographic parameters since acoustic propagation is sensitive to errors in the sound speed gradient.

Keywords: Ocean modeling, validation, comparison methods, depth variability.

1. INTRODUCTION

The last decades more observations have become available through satellite images and open climatology databases. In addition hydrodynamic ocean models have been refined, for instance by adding assimilation and/or ensemble modelling, resulting in increased performance and improved resolving of both large scale circulation and mesoscale phenomena. These new sources of data are often easily accessible and useful rapid environmental assessment products. However, new modelling and observation techniques sources require validation before being put to use operationally. Commonly used validation procedures include comparisons to observations or outputs from established ocean models. Either way nonambiguous comparison methods are required.

The majority of papers on ocean models contain comparisons of data, but few focus on how to compare. This paper discusses the usefulness of different comparison methods with particular focus on depth variability. Many papers are concerned with the validity of the ocean model on the sea surface, e. g. sea surface temperatures, salinities, and currents. Drifting buoys measuring temperature and conductivity [1, 2], and synthetic aperture radar (SAR) data [1, 3-6] are easily accessible and therefore ideal for such validations. However, when ocean model outputs are used for e.g. acoustic modelling, correct representation of depth gradients is important. Sound speed profiles are normally derived from temperature and salinity model data and used for applications such as sonar performance modelling. Accurate sound speed information is essential for acoustic modelling in the mid– and high–frequency ranges [7-10]. The acoustic propagation is directly controlled by the sound speed gradient [11]. Errors in the gradient may easily cause a shift in the location of acoustic shadow and convergence zones [7] and thus cause significant errors in modelled sonar performance.

Most papers on ocean validation present the data that are compared, and the main oceanography in the area of interest. When comparing oceanographic data sets, information is required on the area, such as the characteristics of present water masses, dominating currents, and special features. Such information is essential for understanding the data sets. Model output depends on certain numerical parameters and choices, e.g. the selected turbulence model, boundary conditions, and spatial and temporal resolution. Likewise observations are influenced by various premises such as measuring techniques, measurement errors, and spatial and/or temporal resolution and filtering. These conditions must be understood in order to explain the differences observed in a comparison. If a model does not take small scaled phenomena into account, then discrepancies on that scale in comparison to observed data should be expected.

Ocean models are used for a wide range of applications, e.g. forecasting, climate monitoring, and sonar performance. The choice of comparison methods should depend on the considered application, and the chosen methods should be complementary in such a way that they identify different types of discrepancies. The comparison methods are here divided into two groups; direct and statistical.

Direct comparison methods compare two data sets on a one–to–one basis e.g. comparisons of time series [1, 2, 12-15] and cross sections [2-5, 13, 14]. Direct comparison methods are useful for verifying the ocean models ability to pre-

*Address correspondence to this author at the Faculty of Technology and Maritime Sciences, Vestfold University College, 3103 Tønsberg, Norway; Tel: +47 33 03 77 38; Fax: +47 33 03 11 00;
E-mail: karina.hjelmervik@hive.no

1874-835X/12

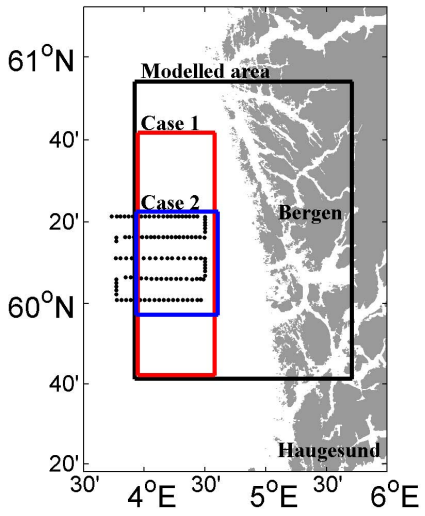

Fig. (1). The modelled data set is taken from locations inside the red box in test case 1 and inside the blue box in test case 2. Profiles are measured at the locations of the dots. Comparison methods are needed in order to handle the growing amount of oceanographic data. Vertical profiles of temperature and salinity at two positions for test case 1 (red) and test case 2 (blue).

predict the spatiotemporal distribution of water masses and oceanographic features. These aspects are important in e.g. forecasting.

Statistical comparison methods extract statistical parameters from the data sets or their difference, e.g. moments [1, 4, 6, 13, 16, 17], root–mean–square (RMS) error [1, 6, 12, 13, 15, 16, 18], and correlation coefficients [1, 12, 15, 17, 18]. Statistical comparison methods are useful for verifying the statistical distribution of water masses and oceanographic features. This is important to e. g. determine the expected stress on offshore installations, and when ocean model output is used as a priori information in acoustic inversion of the sound speed profile.

Presenting comparison results unambiguously, requires care. When comparing large amounts of data, one easily gets lost in colours, numbers, or words. Examples of comparisons using both direct and statistical methods are presented using ocean model data and CTD measurements. The capabilities and limitations of each method are highlighted and demonstrated. The ocean model data are obtained from the Norwegian Meteorological Office [19]. The observations include both temperature and salinity data and were obtained by the Norwegian Defence Research Establishment.

The intent of this review is to give an overview of widely used comparison methods with examples that highlight the main features of each method. However, this review is not exhaus- tive. More methods and variants of the methods included exist in literature. Increased focus on

2. EXAMPLE DATA SETS

In order to illustrate the aspects of different comparison methods, two test cases are defined. Observed and modelled data from the south–west coast of Norway are used (Fig. **1**). In test case 1 modelled data from two time steps 24 hours apart, 06am 24 and 25 January 2010, are compared. In test case 2 observed data are compared to modelled data from approximately the same time and place.

The modelled data are computed by the Norwegian Meteorological Institute's version of the Princeton Ocean Model (MI–POM) [19] which is a baroclinic, three-dimensional hydrodynamic ocean model. It is a primitive equation model containing conservation equations for mass and momentum. Thermodynamics are descibed by similar conservation equations for salinity and temperature. The model is set up in an area along the south–west coast of Norway (Fig. **1**) as a nested submodel with horizontal resolution of 200 meters which is sufficient to resolve mesoscale phenomena. In the vertical 14 terrain following σ–coordinates are used. Outputs are given every third hour for sea surface, salinity, temperature, and currents at selected depths. MI–POM is used for both research and forecasting purposes.

The observed data set consists of 113 CTD–profiles obtained using a Moving Vessel Profiler during a sea trial from 23 to 24 of January 2010. The observations were obtained by HU Sverdrup II, - a research vessel belonging to the Norwegian Defence Research Establishment. Each profile consists of around 500 measurements of conductivity, temperature, and depth. In order to compare to model data, the observed data are averaged over a depth range of ±2 meters in order to get representative values at selected depths. The model area does not cover the location of the westernmost observations. These are therefore left out.

Before studying the data sets one should always look at the main oceanography in the area.

In the test area chosen (Fig. **1**), two distinct water masses are present. The Norwegian Coastal Current supplied with continental fresh water runs northwards along the coastline from the Baltic Sea. A fraction of the North East Atlantic Current runs into the Norwegian Trench north of the test area, approximately following the 200 meter depth contour all the way to Skagerak [14, 20, 21]. Atlantic water is saline and relatively warm while the water from the Baltic Sea is colder and less saline due to fresh water input from fjords and rivers.

At the front separating these two water masses a complicated system of eddies and jets appear. Modelled current is shown in Fig. (**2**) for the two time steps in test case 1. During the model period the eddies move slowly northwards and change shape slightly, i.e. the eddy with the northern boundary at 60° 20' N at 06am 24 January moves its northern boundary to 60° 30' N during 24 hours. To place such eddies correctly in ocean models is difficult. MI–POM seems to reproduce the right amount and strength of eddies [14], but fails to place eddies correctly in space and time.

3. METHODS OF COMPARISON

Comparison methods are here divided into two groups: Direct and statistical methods. Direct methods are one-to-one comparions of data sets. Statistical methods compare statistical properties extracted from the data sets.

In the following sections different types of comparison methods are demonstrated using examples from the two test cases described in Sec. 2.

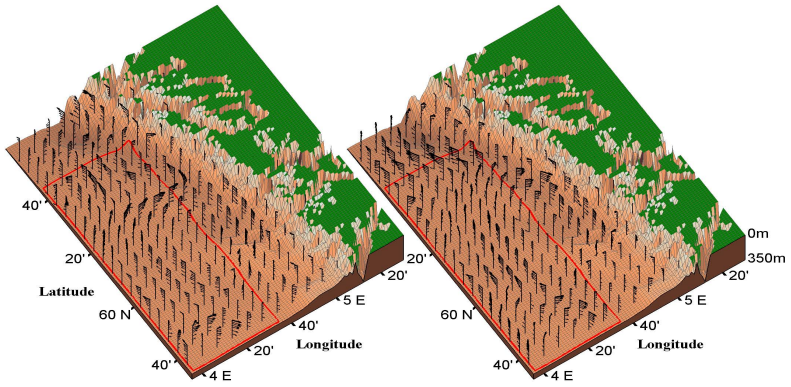

Fig. (2). Modelled current at 06 am 24 (left) and 25 (right) January 2010. The red square indicates the boundary for data from test case 1.

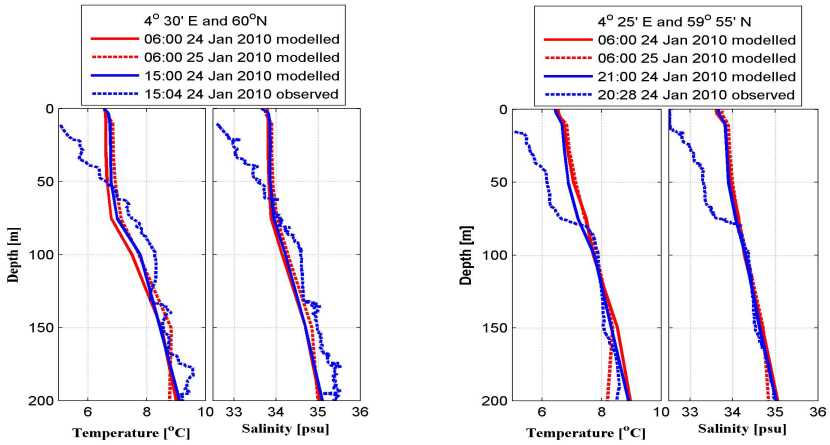

Fig. (3). Example of one dimensional plots at selected positions. Vertical profiles of temperature and salinity at two positions for test case 1 (red) and test case 2 (blue).

3.1. Direct Comparison Methods

Direct comparison methods compare state variables from two data sets at specific times and locations. Direct comparisons of either time series or two dimensional plots such as vertical or horizontal cross sections, are frequently used. This can be done either by presenting snapshots of each data set or subtractions of one set from the other [5, 14].

These intuitive methods allow for detailed studies of the ocean models' ability to locate oceanographic phenomena such as eddies and fronts.

3.1.1. One Dimensional Plots

Observations at one or several selected locations are easy to perform using buoys or other moored platforms.

Direct comparison of time series at a certain location is the most used comparison method in literature [1, 2, 12-15]. Time series may be used to verify that specific features are predicted at the observation location at the right time, and to study local effects of time variability.

Other one dimensional plots such as vertical profiles at certain locations at a given time are also compared in literature [6, 14, 22]. Vertical profiles may be used to verify the skill of the ocean model to predict depth variability which is important for acoustic purposes. However, vertical profiles at a few locations are not sufficient to validate an ocean model. Comparisons on a larger scale are required.

Example of vertical profiles at two positions are shown in Fig. (**3**). In test case 1 both modelled temperature and salinity increase slightly in the whole water column during 24 hours at both positions indicating more Atlantic water in the area. In test case 2 the observed profiles have larger depth gradients. At 4◦ 25' E and 59◦ 55' N a sudden increase in both temperature and salinity is observed identifying a distinct pycnocline.

3.1.2. Two Dimensional Plots

Two dimensional plots are commonly used in literature. Examples include horizontal cross sections (i.e. map of surface or specific depth level) [4, 5, 13, 14, 22], vertical cross sections [2, 3, 14, 22], and vertical profiles as a function of time [1, 16].

Cross section comparisons are easily understood and are efficient to locate problem areas. A downside with such visual comparison methods is that if not carefully chosen, the colours and dynamics used in the plotting may bias the interpretation, e.g. since the human eye separates some colours better than others [23].

Cross section comparisons may be performed using side–by–side comparison of snapshots and difference plots, or plots containing both observations and modelled data. A comparison by subtracting observed from modelled values, may result in relatively large discrepancies [14], since difference plots are very sensitive to phase differences. Relatively

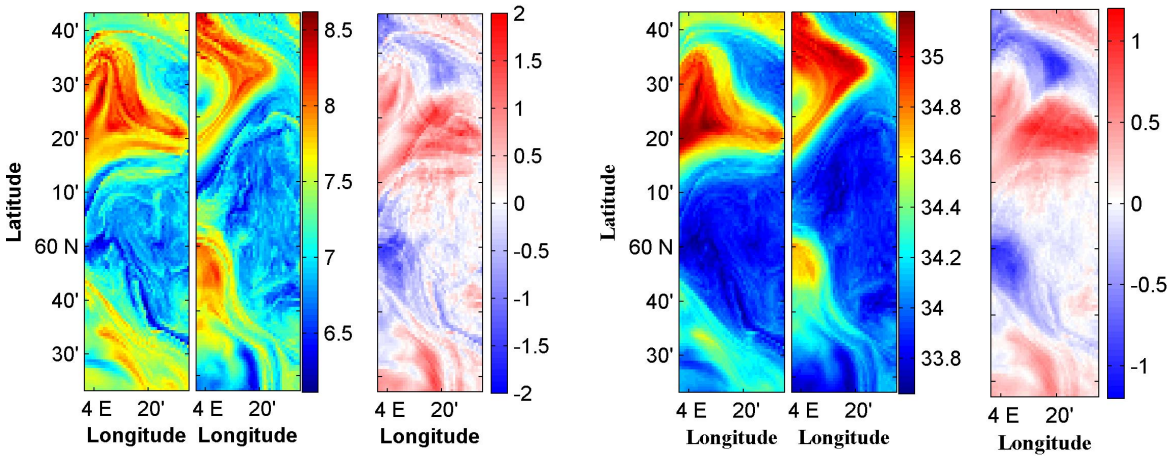

Fig. (4). Example of horizontal cross section plots for test case 1.
Horizontal cross sections at 50 m depth for temperature (left set) and salinity (right set) at 06am 24 (first) and 25 (second) January 2010, and the difference between these two fields (third).

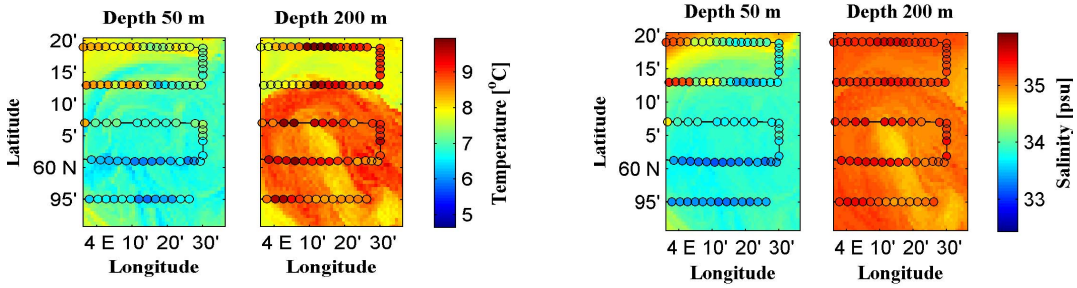

Fig. (5). Examples of horizontal cross section plots for test case 2.
Horizontal cross sections for temperature (left set) and salinity (right set) at 50 m (first) and 200 m (second) depth. The dots represent the observed data taken from 10pm 23 January to 9pm 24 January. The field represents the modelled data at 06am 24 January 2010.

small phase differences in the compared data sets may have a relatively large impact on difference plots. This method therefore efficiently determines if phenomena, e.g. eddies, are collocated in the compared data sets, but has low tolerance for errors. This method is useful for validating forecast models.

In test case 1 differences up to ±1.5° C in temperature and ±1 psu in salinity are achieved during 24 hours of model time (Fig. **4**). The pattern of the temperature and salinity field mirrors the eddy structures. Some eddies contain cold and relatively fresh water, while others contain Atlantic water with warmer and more saline water. Since the eddies move in time, discrepancies may occur due to phase errors, i.e. eddies and meanders located at the right place, but not at the right time.

In test case 2 both the observed data and modelled data show that there exists a front between warm saline water and cold less saline water (Fig. **5**). Unfortunately, the amount of observed data is limited and thereby it is difficult to determine if the data are shifted or if the data sets represent the same amount of eddies. At 60° 20' N the pycnocline lies at about 50 meters depth both in observed and in modelled data, but is more distinct in the observations (Fig. **6**). In this vertical cross section, differences up to ±2° C in temperature and ±0.5–1.5 psu in salinity are observed. The largest differences are found at the surface and at the bottom indicating discrepancies in the gradients. As with vertical profiles, vertical cross sections are useful for evaluating the ocean models' ability to predict oceanography useful for acoustic modelling purposes.

3.1.3. Scatter Plot

In scatter plots state variables in two data sets from selected locations are plotted versus each other at the same locations [6, 14]. Scatter plots are useful to investigate if a spatiotemporal correlation exists, and give a visual impression of how well two data sets agree. An identity line along y = x is often drawn as a reference. If the two data sets agree, the scatters concentrate in the vicinity of the identity line. In order to do such one–to–one comparisons, resampling the data in time and space is often required. Since location is not explicitly read from the plots, scatter plots do not uncover which parts of the domain are causing eventual deviations.

In test case 1 scatter plots show that the spatial correlation between the data at the two time steps is poor. (Fig. **7**). Phase errors result in a swarm of scatter around the linear regression lines at each depth. There seems to be a closer linear relation at more shallow depths, especially for salinity. The lack of spatiotemporal correlation is probably due to phase differences, but the reason is not revealed by scatter plots.

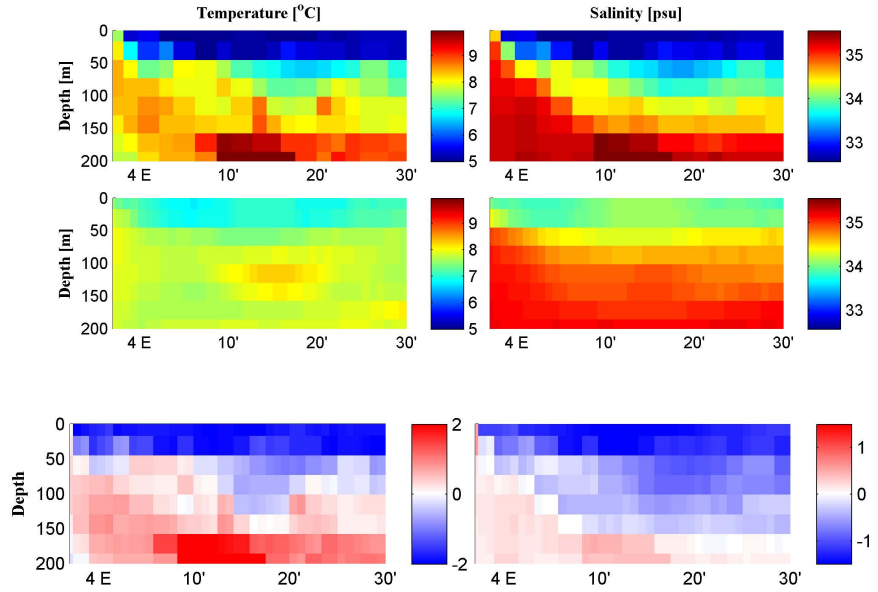

Fig. (6). Example of vertical cross section plots for test case 2.
Vertical cross sections from a line at 60° 20' N for observed (upper), modelled (middle), and the difference between observed and modelled (lower) temperature (left) and salinity (right). The observations are taken at 11pm 23 to 02am 24 January 2010. The simulations are from 00am 24 January 2010.

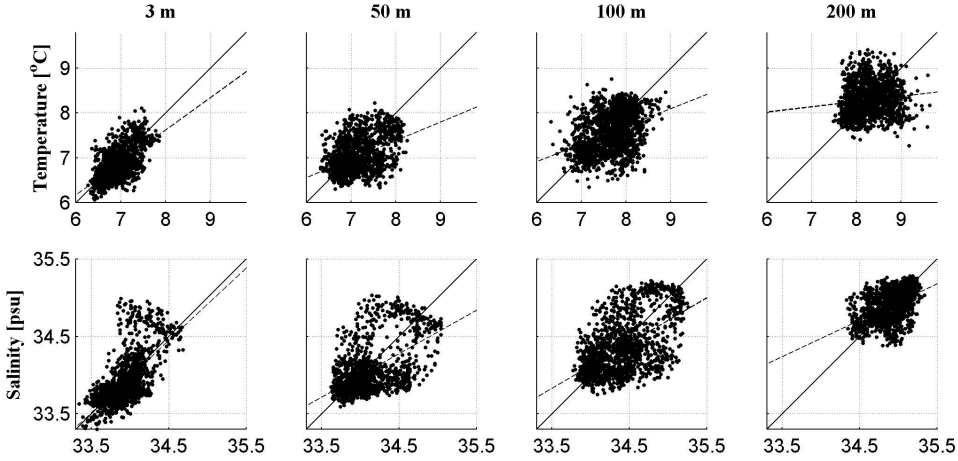

Fig. (7). Example of scatter plots for test case 1.
Scatter plots of temperature (upper) and salinity (lower) for four selected depths. Modelled data from 06am 24 January is on the first axis and from 06am 25 January on the second axis. Each plot includes two reference lines; the dashed line is a regression line based on the data, while the solid line represents the identity line.

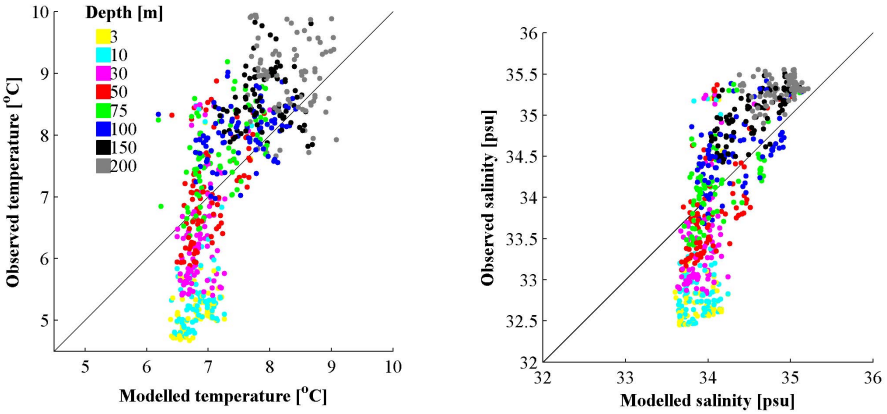

Fig. (8). Example of scatter plots for test case 2.
Scatter plots of temperature (left) and salinity (right) based on observed and modelled data at corresponding locations and times.

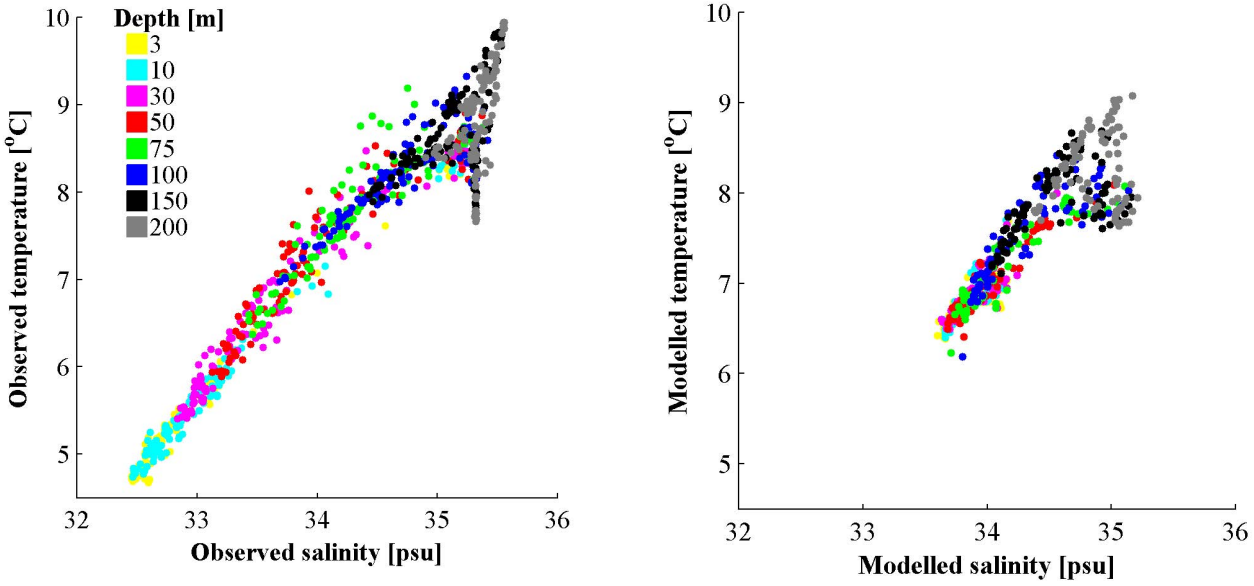

Fig. (9). Example of scatter plots for test case 2.
Scatterplots identifying the relation between temperature and salinity in the observations (left) and the model (right) at the same locations and time.

In test case 2 scatter plots show that the modelled temperature and salinity are too high in the upper layers and too low in the deeper layers (Fig. **8**). Thereby the model underestimates the depth gradients. The plots also clearly show that the variability in the observations are significantly higher than in the modelled data implying too strong mixing in the modelled data.

Scatter plots may also be used to identify the relation between two state variables, e.g. salinity and temperature. Scatter plots are able to reveal not only linear, but also nonlinear relationships. In test case 2 the observations indicate a linear relationship between temperature and salinity except for water masses with high salinity especially at larger depths (Fig. **9**). The relation in the model is similar, but with smaller interval.

3.2. Statistical Methods

Statistical methods compare statistical attributes instead of oceanographic data directly. Statistical methods may be divided into two groups; the methods that take the spatiotemporal distribution into account, and the methods that disregard the spatiotemporal distribution. Examples of the first group of methods include the correlation coefficient and root–mean–square (RMS) error. The second group include comparisons of e.g. moments and probability density functions.

Statistical methods allow for detailed studies of discrepancies in the overall statistics of the compared data sets. Some of the methods give numerical values on how well the data sets compare, e. g. moments, RMS error, and correlation coefficients, while others give visual representations, such as probability density functions and QQ–plots.

3.2.1. Moments

The first two moments are extensively used in comparisons of two data sets [1, 4, 6, 13, 16, 17]. The mean, μ_k, and standard deviation, σ_k, at the kth depth step of N data profiles may be estimated as follows:

$$\mu_k = \frac{1}{N}\sum_{n=1}^{N} p_{nk} \tag{1}$$

where p_{nk} is the kth depth step of the nth profile of a state variable. The profiles are measured or modelled in either time or space. In the case where only surface values are considered, k equals one.

The mean and standard deviation are robust and easily understood. The mean gives a quick impression of the arithmetic mean, and if plotted as a function of depth, it gives an indication of the gradient useful for acoustic purposes. Note that since the mean is an arithmetic average it is not the same as the median for skewed distributions. Standard deviation is the square root of the variance and a measurement of variability. Low standard deviation indicates that the data points are close to the mean, whereas high standard deviation indicates that the data are spread out over a larger range of values. Note that local variations causing e.g. two–peaked distributions can lead to misleading values for mean and standard deviation. Therefore the moments should always be considered in light of the corresponding distributions.

Fig. (**10**) shows an example of how statistical moments derived from different data sets may be compared. In test case 1 the means for the two data sets differ with only up to ±0.1°C in temperature and ±0.05 psu in salinity for all depths. The standard deviations are around 0.4 °C for temperature and between 0.1 and 0.4 psu for salinity in both data sets. In test case 2 the moments uncover that gradients and variances in the modelled data are underestimated. Notice that the mean profiles intersect at approximately 50 meter depth in the temperature data. This supports the decent match in the direct comparison of horizontal cross section

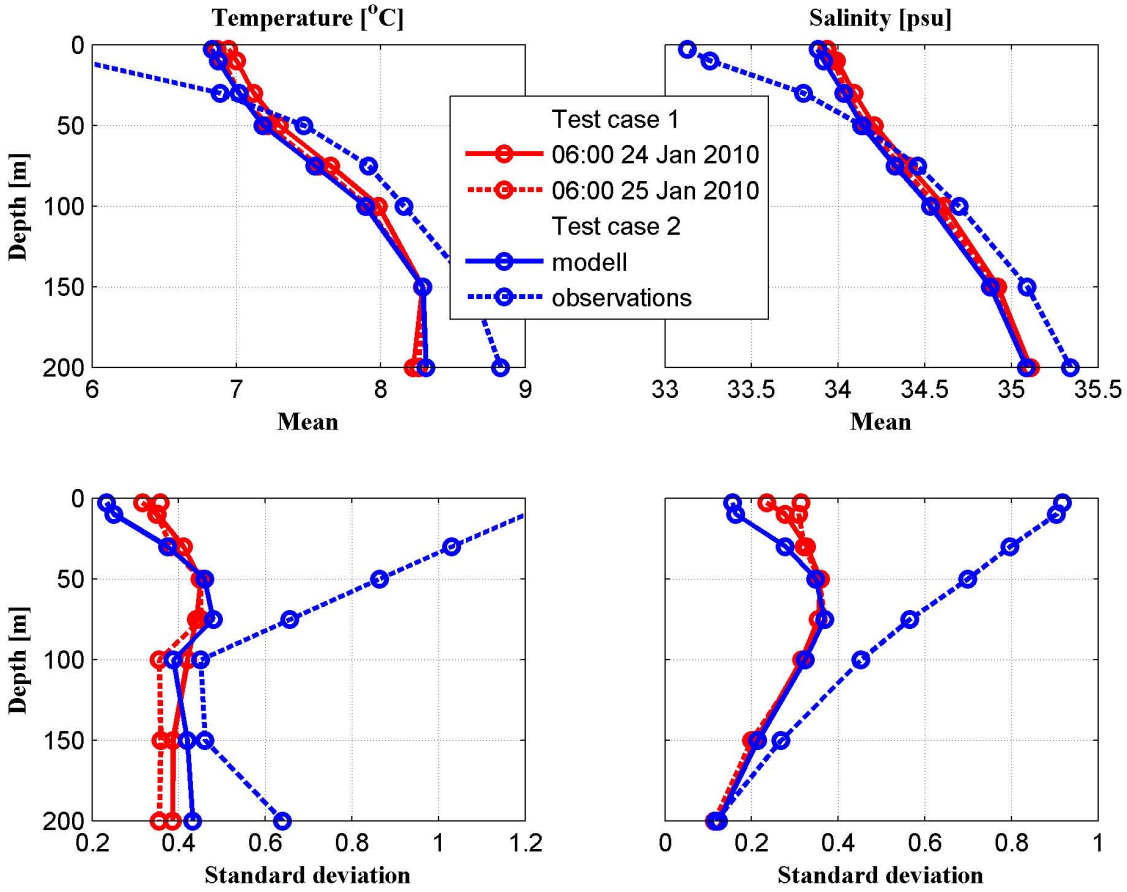

Fig. (10). Example of depth–dependent moments.
The mean (upper) and standard deviation (lower) of temperature (left) and salinity (right) in test case 1 (red) and test case 2 (blue).

data at 50 meter depth made in Fig. (**5**), and that comparisons at other depths result in more obvious differences. Horizontal cross sections at several depths are therefore necessary.

Higher order statistical moments as skewness and kurtosis which are measurements of the asymmetry of the probability distribution and extreme deviations respectively, are more seldom used in ocean modelling.

3.2.2. Depth Dependent Probability Density Functions

Comparisons of probability density functions (PDF) mirror the models' ability to reproduce the correct statistics, but is not a measure of the models' forecasting skill [14] since e.g. phase errors are not revealed in the distributions.

PDFs may be estimated by normalised histograms. In comparisons the conventional use is to generate and compare one–dimensional PDFs using observed and modelled data at selected depths or averaged over depth [14, 17]. For acoustic purposes two–dimensional PDFs that combine depth information with a state variable, are preferable. The scaling and bin widths used in the estimation of the PDFs may easily bias the interpretation, and must therefore be chosen with care.

In test case 1 the PDFs for temperature and salinity are similar for the two data sets (Fig. **11**) even though the mesoscale phenomena are shifted in time. This indicates that the two modelled data sets contain approximately the same types and quantities of water masses.

In test case 2 there is clearly more variation in the upper layer in the observed data set, than in the modelled data set (Fig. **12**). This is in accordance with the standard deviations in Fig. (**10**). The model probably mixes the two distinct water masses in the area. It is well known that acoustic modelling is very sensitive to errors in the sound speed gradient which depends on the temperature and salinity profiles. Acoustic modelling is therefore sensitive to the observed errors in this data set.

3.2.3. QQ–Plot

QQ–plots compare the distributions of two different data sets by plotting the quantiles of each data set against each other. When the probability density distributions of the data sets are known, then the quantiles may be derived by inverting the cumulative distribution function. However, for modelled and observed oceanography the exact distribution is rarely known. If the two data sets are of equal size, the data may simply be ordered and plotted against each other. In order to compare two data sets of different sizes, the data must be resampled. In test case 2 the data sets for each depth are first sorted in order of magnitude to compute the

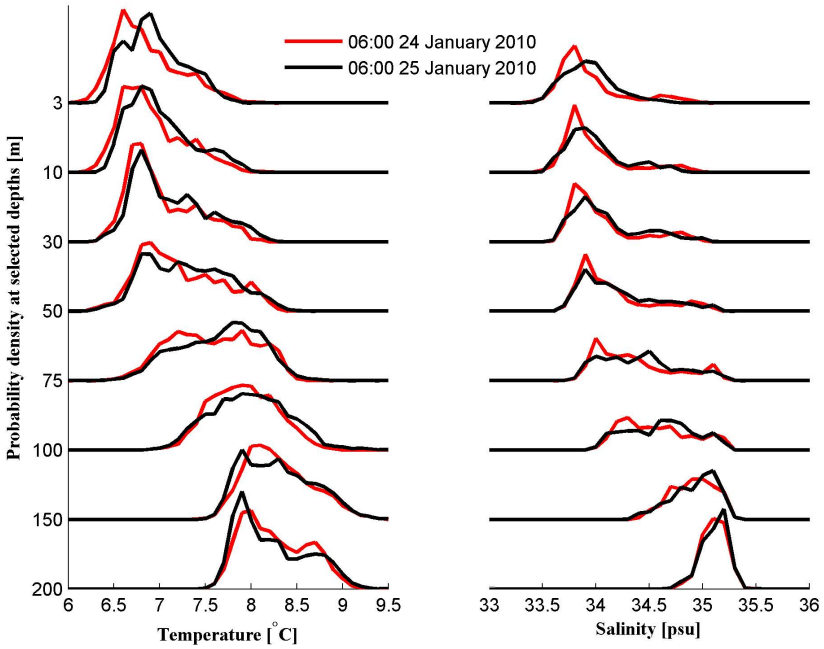

Fig. (11). Example of normalised probability density functions for test case 1.
Probability density functions at selected depths for salinity and temperature based on modelled data from two time steps, 06am 24 (red) and 06am 25 (black) January 2010. The bin widths are 0.1° C for temperature and 0.1 psu for salinity.

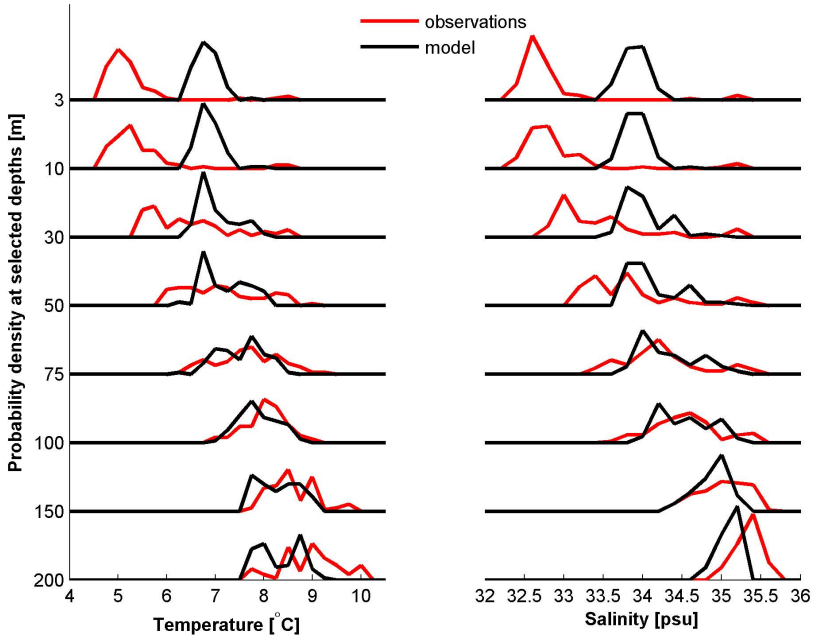

Fig. (12). Example of normalised probability density functions for test case 2.
Probability density functions at selected depths for salinity and temperature based on obser- vations (red) and modelled data (black) from January 2010. The bin widths are 0.25° C for temperature and 0.2 psu for salinity.

probability density functions. Corresponding values for different probabilities were then picked as illustrated in Fig. (**13**).

If the two distributions have a linear relationship, the points will approximately lie on a line. If the line is the identity line, y = x, the two distributions are identical. S-shaped lines indicate that one of the distributions is more skewed than the other, or has a heavier tail. Note that for comparisons of heavy tailed distributions oscillations may be observed on the tails of the QQ–plot [24]. This is expected and does not necessarily mean that the two distributions are different in nature.

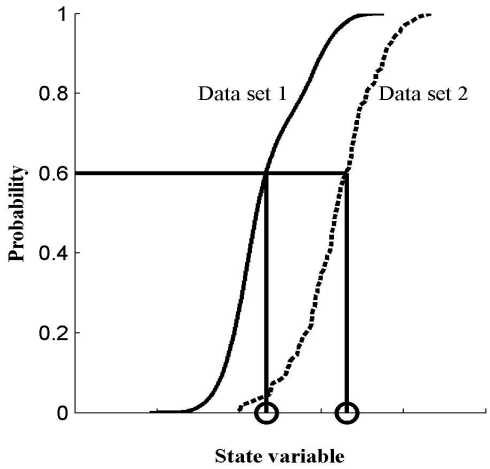

Fig. (13). Illustration of how corresponding values for two data sets with different probabilities are selected in order to create QQ–plots.

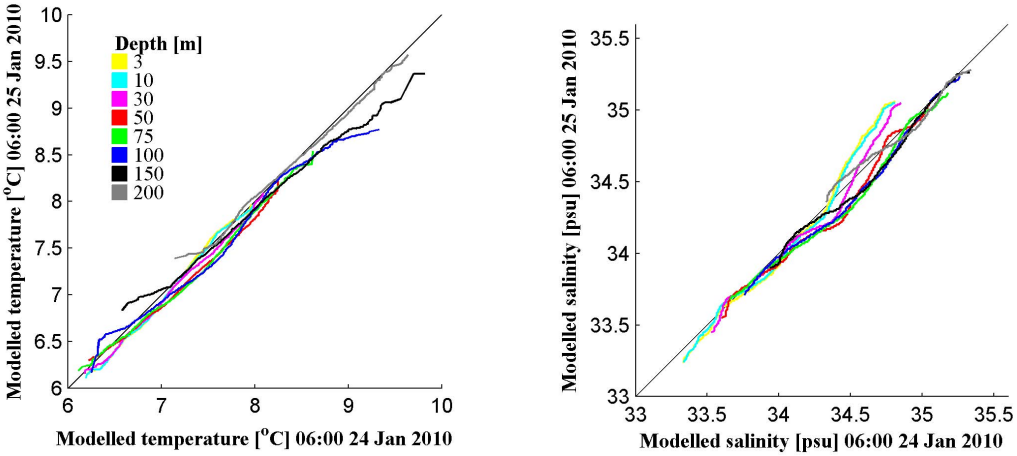

Fig. (14). Example of QQ–plots for test case 1.
QQ–plots of temperature (left) and salinity (right) based on modelled data from two time steps, 06am 24 and 25 January 2010.

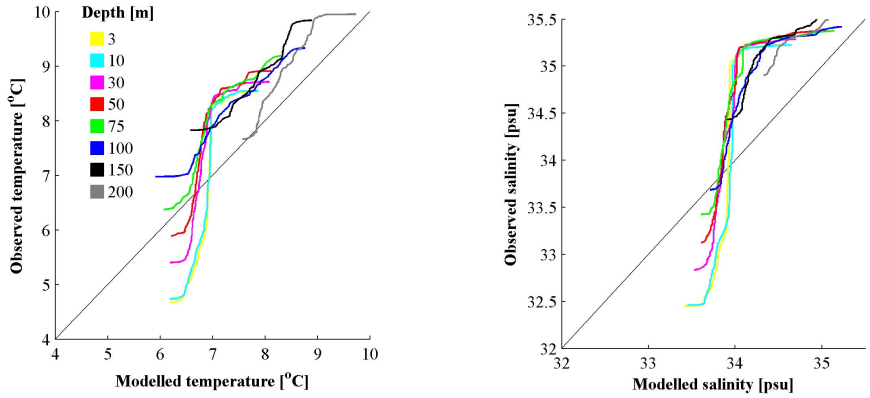

Fig. (15). Example of QQ–plots for test case 2.
QQ–plots of temperature (left) and salinity (right) based on modelled data and observations from January 2010.

In test case 1 a clear linear relationship between the distributions of the two data sets are revealed (Fig. **14**). The QQ-plot for temperature at large depths is flatter than the identity line indicating that the temperature on 25 January is more dispersed. For salinity there is less intermediate and shallow water with high salinity 25 January compared to 24 January. This is in accordance with the horizontal cross section plots (Fig. **4**) where the saline water masses in the northern part of the area move out of the model area while less saline water is introduced in the southern part of the model area.

In test case 2 the modelled temperature and salinity have too few low values and too few high values compared to the observations (Fig. **15**) indicating that the observed distribution is more skewed and/or has more extreme values than the

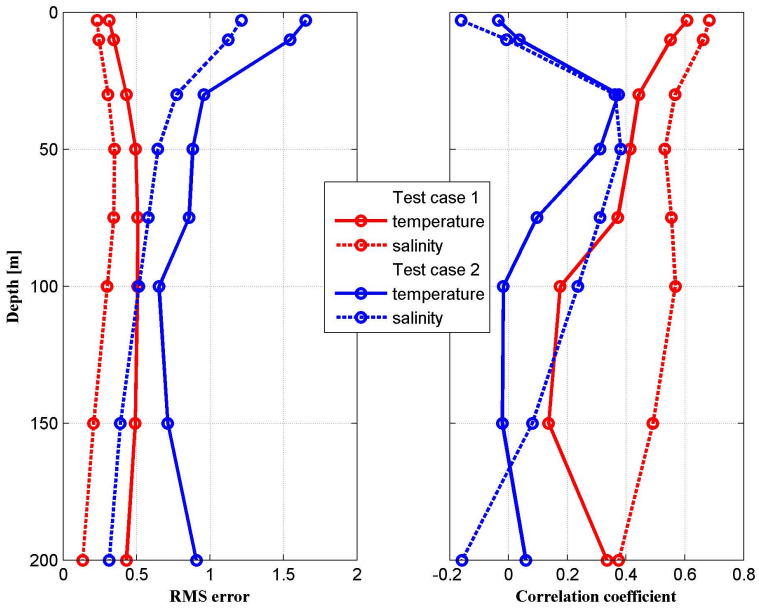

Fig. (16). Example of depth dependent RMS and correlation coefficient.
The RMS difference (left) and correlation coefficient (right) in test case 1 (red) and test case 2 (blue).

modelled distribution. This is confirmed by the probability density functions in Fig. (**12**).

Note that QQ-plots should not be confused with scatter plots. Scatter plots compare data at the same location, while QQ–plots compare distributions. In test case 1 a swarm of scatter is shown in the scatter plots (Fig. **7**) indicating low spatiotemporal correlation, while the QQ-plots reveal (Fig. **14**) a clear linear relationship between the distributions indicating that the distributions of the two data sets agree to a great extent. The QQ-plots do not contain any spatiotemporal information.

3.2.4. RMS Error

Computing the root–mean–square (RMS) error is a widely used method to determine the discrepancies between two data sets [1, 6, 12, 13, 15, 16, 18]. The RMS error is frequently used as an objective method to support hypotheses based on direct comparisons of time series or cross sections.

The RMS error of the kth depth step is estimated as follows:

$$\sigma_k = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(p_{nk} - \mu_k)^2} \tag{3}$$

where $P_{nk}^{(1)}$ and $P_{nk}^{(2)}$ represent two data sets. Typically, n represents a certain location or time.

This is similar to the computations in difference plots (Sect. 3.1.2), but here the results are averaged over all samples and outliers are exaggerated due to the squaring. Higher ordered differences can also be used, resulting in even more pronounced effects from outliers. In literature squared differences appear to be the most popular choice.

Fig. (**16**) shows an example of an RMS error plot as a function of depth for the example data sets. In test case 1 RMS errors of up to 0.5◦ C in temperature and 0.4 psu in salinity appear due to phase differences. In test case 2 larger RMS errors occur. Notice that the errors are larger at shallow depths. Concerning temperature, the smallest error is found at 100 meter depth even though the mean has the smallest difference at 30 meter depth (Figs. **3.2.1**). RMS error is influenced more strongly by outliers than the difference of the means. The standard deviation is far lower at 100 meter depth than for 30 meter depth in the observed values. This results in more outliers at 30 meter depth than at 100 meter depth which has a considerable impact on the RMS error. Studying a single statistical attribute is therefore not recommended as this may give the wrong impression.

3.2.5. Correlation Coefficient

The correlation coefficient is an often used method to compare linear trends of two data sets [1, 12, 15, 17, 18]. As with the RMS error, the correlation coefficient is used as an objective method to support hypotheses based on direct comparisons of time series.

The correlation coefficient for the kth depth step is estimated as follows:

$$R_k = \frac{1}{N-1}\sum_{n=1}^{N}\frac{(p_{nk}^{(1)} - \mu_k^{(1)})(p_{nk}^{(2)} - \mu_k^{(2)})}{\sigma_k^{(1)}\sigma_k^{(2)}} \tag{4}$$

As with the RMS error, n represents a certain location or time. The numerator contains the covariance [25] between two data sets, while the denominator contains the product of the standard deviations of each series. Two completely uncorrelated series have a covariance of zero. Two completely correlated series will have a covariance equal to the product of their standard deviations, resulting in a correlation coefficient of one. The correlation coefficient captures the amount

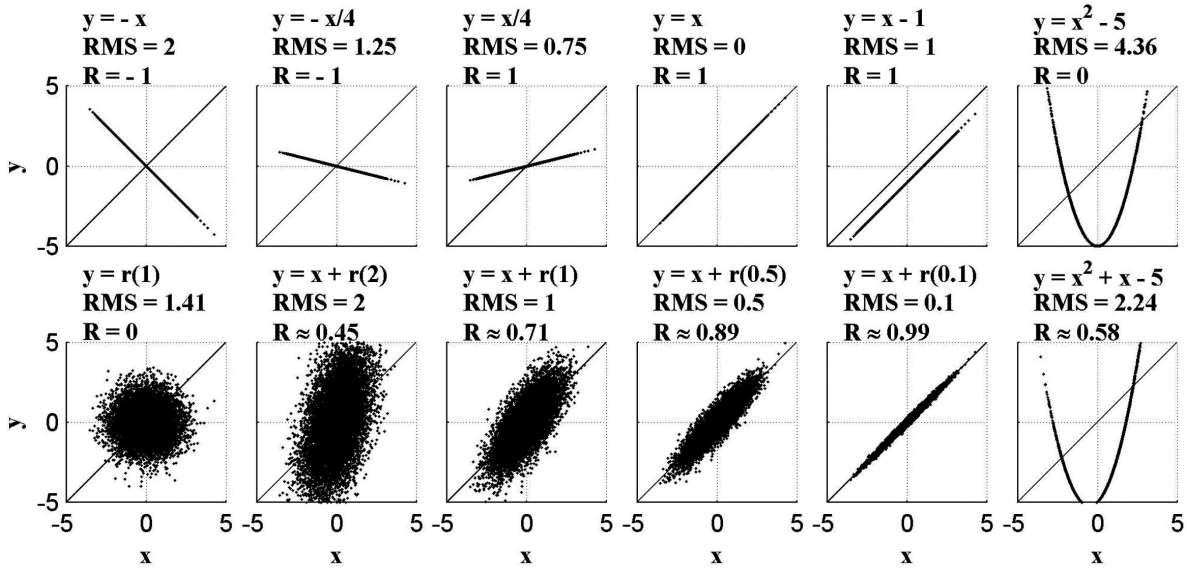

Fig. (17). Illustration showing the RMS error and correlation coefficient, R, for different scatter diagrams. The extent the scatter deviates from the identity line indicates the size of the RMS error, while the deviation from the regression line indicates the correlation coefficient. x and r(σ) are randomly distributed gaussian functions with zero mean and standard deviation one and σ respectively.

of noise in the scatter diagram and if the direction of a linear relationship, if any, is positive or negative (Fig. **17**). However, the correlation coefficient does not capture the slope of the linear relationship nor any nonlinear relationship unless the distribution is skewed.

Fig. (**16**) shows an example of the correlation coefficient as a function of depth for the example data sets. In test case 1, there seems to be a linear relationship in the scatter diagram (Fig. **7**) with less noise for shallow depths. This is captured by the correlation coefficient which is larger at shallow depths. In test case 2, there seems to be a weak nonlinear relationship in the scatter diagram (Fig. **8**). The correlation coefficient depends on the linear relationship and the amount of noise, and is therefore very low. For temperature the correlation coefficients are around zero for 100 meter depth and more. This reflects the seemingly random scatter for these depths. The negative correlation coefficient that appears for salinity at some depths, is due to the negative change in direction of the slope.

3.2.6. Taylor Diagram

A Taylor diagram compactly presents normalised standard deviations, correlations, and bias-removed RMS–differences between two data sets [15, 18]. The Taylor diagram is useful for simultaneous comparisons of different data types in the same plot, and is particularly useful for efficient comparison of performance of different ocean models [18].

The distance from the origin yields the normalised standard deviations given by $\tilde{\sigma}_K = \frac{\sigma_k}{\sigma_k^{(1)}}$ where the standard deviations are normalised with respect to data from data set one. Data with the same standard deviation as data set one, are then located on the unit circle.

The azimuth angle represents the correlation coefficient estimated using Eq. 4. Along the horizontal axis, the correlation coefficient is 1 to the right of the origin and -1 to the left of the origin. Along the vertical axis, the correlation coefficient is 0. High correlation requires a good one–to–one comparison.

The Euclidian distance from the reference point, represents the bias–removed RMS difference, which is subtly different from the RMS error described earlier. The bias–removed difference for the kth depth step is given by:

$$\widetilde{RMS}_k = \frac{1}{N}\sqrt{\sum_{n=1}^{N}\left(\left(p_{nk}^{(1)} - \mu_k^{(1)}\right) - \left(p_{nk}^{(2)} - \mu_k^{(2)}\right)\right)^2} \tag{5}$$

The bias–removed RMS difference depends on both correlation and standard deviation [18].

Fig. (**18**) shows a Taylor diagram for test case 1. Data series for each combination of depth and data type are considered here. Each comparison results in a single dot in the Taylor diagram. Almost all the dots lie in the vicinity of the unit circle, indicating that the standard deviations from the two data sets are similar. The exceptions are at shallow depths for salinity and deep depths for temperature. This is in accordance with the standard deviation in Fig. (**10**). The correlation coefficient is smallest for temperature at larger depths and highest for salinity at shallow depths in accordance with the correlation coefficient in Fig. (**16**). The poor correlation for temperature at larger depths results in a high bias–removed RMS difference due to phase differences.

Fig. (**19**) shows a Taylor diagram for test case 2. The model clearly underestimates the standard deviation, particularly at shallow depths. This is in accordance with what was observed in Sect. 3.2.1. Due to misplacement of mesoscale phenomena, the correlation is poor. The correlation for salinity is better than for temperature. Observe that the salinity at 75 meter depth has approximately the same bias–removed RMS difference as at 10 meter depth. At 75 meter the bias–removed RMS difference is due to poor correlation, while at 10 meter depth the standard deviation is too low.

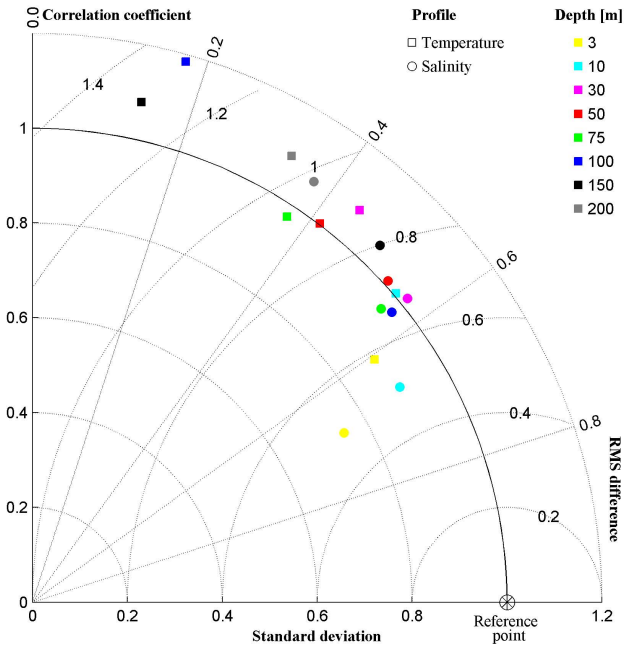

Fig. (18). Example of Taylor diagram for test case 1.

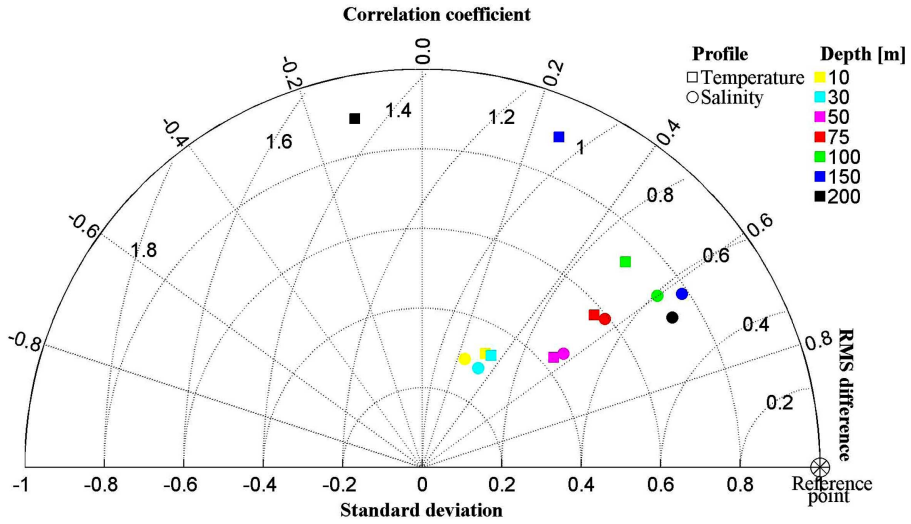

Fig. (19). Example of Taylor diagram for test case 2.

The Taylor diagram showing normalised standard deviations (distance from origin), correlations (azimuth angle), and bias removed RMS differences (euclidian distance from reference point) between modelled data from two time steps, 06am 24 and 25 January 2010. The reference point applies for data from the last time step.

The Taylor diagram showing normalised standard deviations (distance from origin), correlations (azimuth angle), and bias removed RMS differences (Euclidian distance from reference point) between modelled data and observations from January 2010. The reference point applies for all observed values.

3.3. Discussion

Various comparison methods have been applied on two different test cases taken from the south west coast of Norway. Since the methods complement each other, the results from the different methods should be discussed in combination.

In test case 1 two time steps, 24 hours apart, from the same ocean model are compared. The statistics, such as moments and distributions, should approximately be the same for the two data sets. Phase differences are expected due to propagation of fronts and mesoscale eddies.

Only small differences in statistical distributions are uncovered by the depth dependent moments (Fig. **10**), the probability density functions for each depth step (Fig. **11**), and the QQ–plots (Fig. **14**). These methods highlight different aspects. The moments are robust and easily understood, but do not fully describe the statistical distribution. Probability density functions give a visual comparison of the distributions and information on all the moments. The scaling and

bin widths used in the estimation may easily bias the interpretation, and must therefore be chosen with care. QQ–plots give a visual representation of how well the distributions compare, but include no information on the individual distributions or moments. For applications that require accurate depth variability, these comparisons should be carried out for each depth. The statistical representation of the depth gradient may be assessed this way.

Comparisons of horizontal cross section plots (Fig. **4**) clearly show the presence of phase differences, but leave the reader to interpret the extent of the differences. RMS error and correlation coefficient values compliment the cross section plots by providing objective estimates of the differences. While the RMS error gives the averaged squared difference between the two data sets, the correlation coefficient compares the trends (Sect. 3.2.4 and 3.2.5). Scatter plots efficiently show the lack of spatiotemporal correlation, but as for RMS error and correlation coefficient plots, scatter plots do not give any clear indications of the cause. Neither of these methods give any indication of how well the depth–gradients compare. Vertical cross sections or profiles (Fig. **3**) may be used to aquire information on how well the depth gradients compare at specific locations.

Taylor diagrams efficiently pile the information from bias–removed RMS–differences, correlation, and standard deviation in the same plot. The main advantage of this method is the efficient representation that allows for quick comparison of different data sources, e.g. different ocean models.

Test case 2 includes an observed and a modelled data set. The modelled data are extracted at approximately the same times and locations where the observations were made. The model is not expected to successfully determine the exact locations of eddies and fronts, thus phase differences are expected. Also larger differences in the statistical distributions are expected compared to test case 1.

Due to the low spatial resolution of the observations, it is difficult to get a clear picture of the extent of the phase differences in the horizontal cross section plot (Fig. **5**). The correlation coefficient and RMS error indicate poor comparisons. A quick study of the moments (Fig. **10**) shows that the model grossly underestimates the mean depth gradients and also the standard deviation, particularly at shallow depths. Errors in the mean will typically influence the RMS error, while high standard deviations indicate the presence of noise that lowers the correlation coefficient (Sect. 3.2.5). This indicates that the poor correlation coefficient and RMS error estimates are due to erroneous statistical representation of the oceanographic parameters in the model. The depth dependent PDFs (Fig. **12**) indicate that a probable cause is that the model mixes the different water types too strongly. Too strong mixing is also a plausible explanation for the observed discrepancy in the gradient.

For acoustic applications, correct representation of depth gradients is essential. Many of the methods discussed, revealed that the example model underestimates the vertical gradients. It is possible, but not straight forward to observe this error in the vertical cross section comparison (Fig. **6**), but the statistical comparison methods, particularly the mean moment (Fig. **10**) and depth dependent PDFs (Fig. **12**), capture this discrepancy more clearly and produce well-arranged plots that are easy to draw conclusions from. Also, the statistical comparison methods present results from a larger amount of data than the vertical cross section plot without cluttering the presentation.

4. SUMMARY

Ocean models are used for a wide range of applications. To gain confidence in an ocean model, validation is required. The choice of comparison methods should depend on the considered application, and the chosen methods should be complementary in such a way that they identify different types of aspects of the present discrepancies.

A selection of comparison methods used in literature have been illustrated and discussed. The advantages and limitations of each method are highlighted and demonstrated using data from two different test cases. In the first test case data from two different time steps, 24 hours apart, of the same hydrodynamic ocean model are compared. The compared data sets have similar statistical distributions, but due to propagation of water masses and oceanographic features, the geographical distribution differs (phase differences). In the second test case modelled data are compared to observed data. The compared data sets have both discrepancies in statistical distributions as well as phase differences. The abilities of the different comparison methods to detect and describe these differences are discussed.

Direct comparison methods are intuitive. Popular methods include visual comparisons of time series and cross sections. Such plots may be used to validate the predicted geographical distribution of water masses, fronts, and eddies. These aspects are important in e.g. forecasting.

Time series at certain locations are easy to observe. Comparison of time series is therefore perhaps the most used comparison method in literature. Vertical profiles may be used to identify local placement of e.g. the pycnocline. These one dimensional comparisons are useful to study local phenomena, but comparisons on a larger scale are required to validate an ocean model.

Cross section comparisons give a quick and intuitive picture of the presence and extent of phase differences, but do not give any indication of present statistical errors. Comparisons of vertical and horizontal cross sections and vertical profiles as a function of time can be plotted giving a visual representation of the overall results. Cross section comparisons may be performed using side–by–side comparison of snapshots, difference plots, or plots containing both observations and modelled data.

In scatter plots state variables in two data sets are plotted versus each other at the same locations. Scatter plots therefore effectively give a visual impression of how well the two data sets compare, but do not give any information of what parts of the model domain are causing the problems.

Statistical comparison methods effectively compare large data sets by reducing the dimensionality of the comparison problem, e.g. by averaging over time or a spatial dimension. Popular methods in literature include the RMS error, the correlation coefficient, and comparisons of moments. These

values are often used to objectively back up direct comparison methods.

The first two moments give the mean and standard deviation respectively. The probability density function gives a visual representation of the distribution of a state variable. QQ–plots give a visual impression of how well two probability distributions compare. For acoustic applications, and in particular inversion schemes where ocean model output is used as a priori information, comparisons of mean profiles and depth dependent probability density functions are preferred. These methods validate the predicted statistical distributions and depth behaviour of water masses, but are unsuitable for validation of predicted geographical distribution of water masses.

The RMS error is a measure of the discrepancies between two data sets, while the correlation coefficient is a numerical value of how well trends in the compared data sets follow each other. Taylor diagrams compactly present normalised standard deviations, correlations, and bias–removed RMS–differences between two data sets, and are useful for simultaneous comparisons of different data types in the same plot. The efficiency of the representation also allows for quick comparisons of more than two data sets.

This review is not exhaustive. More methods and variants of the methods included exist in literature. The intent of this work is to give the reader an overview and description of widely used comparison methods with examples that highlight the main features of each method. Further focus on comparison methods is important for validating the increasing amount of available oceanographic data, such as refined ocean models and satellite data.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGEMENTS

We would like to thank the Norwegian Meteorological Office for their easily accessible ocean model data on their website met.no. Also thanks to the crew on FFIs research vessel H. U. Sverdrup for collecting the oceanographic observations used in this work.

REFERENCES

[1] V. H. Kourafalou, G. Peng, H. Kang, P. J. Hogan, O. M. Smedstad, and R. H. Weisberg, "Evaluation of global ocean data assimilation experiment products on South Florida nested simulations with the Hybrid Coordinate Ocean Model", *J. Ocean Dyn.*, vol. 59, pp. 47-66, 2009.

[2] H. Xue, and Y. Du, "Implementation of a wetting-and-drying model in simulating the Kennebec-Androscoggin plume and the circulation in Casco Bay", *J. Ocean Dyn.*, vol. 60, pp. 341-357, 2010.

[3] M. G. Magaldi, T. M. Özgökmen, A. Griffa, and M. Rixen, "On the response of a turbulent coastal buoyant current to wind events: the case of the Western Adriatic Current", *J. Ocean Dyn.*, vol. 60, pp. 93-122, 2010.

[4] U. Grawe, J.O. Wolff, and J. Ribbe, "Mixing, hypersalinity and gradients in Hervey Bay, Australia", *J. Ocean Dyn.*, vol. 59, pp. 643-658, 2009.

[5] Y. Fujii, and M. Kamachi, "A reconstruction of observed profiles in the sea east of Japan using vertical coupled temperature-salinity EOF Modes", *J. Oceanography*, vol. 59, pp. 173-186, 2003.

[6] J. Horstmann, W. Koch, and S. Lehner, "Ocean wind fields retrieved from the advanced synthetic aperture radar aboard ENVISAT", *J. Ocean Dyn.*, vol. 54, pp. 570-576, 2004.

[7] S. E. Dosso, "Environmental uncertainty in ocean acoustic source localization", *Inverse Prob.*, vol. 19(2), pp. 419-431, 2003.

[8] S. Finette, "A stochastic representation of environmental uncertainty and its coupling to acoustic wave propagation in ocean waveguides", *J. Acoust. Soc. Am.*, vol. 120, pp. 2567-2579, 2006.

[9] K. LePage, "Modeling propagation and reverberation sensitivity to oceanographic and seabed variability", *IEEE J. Oceanic Eng.*, vol. 31, pp. 402-412, 2006.

[10] K. LePage, and B. E. McDonald, "Environmental effects of waveguide uncertainty on coherent aspects of propagation, scattering, and reverberation", *IEEE J. Oceanic Eng.*, vol. 31, pp. 413-420, 2006.

[11] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational ocean acoustics*. chap. 3, 2nd printing. Springer Verlag, New York 2000.

[12] K. Aoki, and A. Isobe, "Application of finite volume coastal ocean model to hindcasting the wind-induced sea-level variation in Fukuoka Bay", *J. Oceanography*, vol. 63, pp. 333-339, 2007.

[13] D. J. Twigt, E. D. De Goede, F. Zijl, D. Schwanenberg, and A. Y. W. Chiu, "Coupled 1D-3D hydrodynamic modelling, with application to the Pearl River Delta", *J. Ocean Dyn.*, vol. 59, pp. 1077-1093, 2009.

[14] J. Albretsen, and L.-P. Røed, "Decadal long simulations of mesoscale structures in the northern North Sea/Skagerrak using two ocean models", *J. Ocean Dyn.*, vol. 60(4), pp. 933-955, 2010.

[15] Jr G. R. Halliwell, A. Barth, R. H. Weisberg, P. Hogan, O. M. Smedstad, and J. Cummings, "Impact of GODAE products on nested HYCOM simulations of the West Florida Shelf", *J. Ocean Dyn.*, vol. 59, pp. 139-155, 2009.

[16] Alvera-Azcárate, A. Barth, and R. H. Weisberg, "A nested model of the Cariaco Basin (Venezuela): description of the basin's interior hydrography and interactions with the open ocean", *J. Ocean Dyn.*, vol. 59, pp. 97-120, 2009.

[17] J. LaCasce, and H. Engedahl, "Statistics of low frequency currents over the western Norwegian shelf and slope II: model", *J. Ocean Dyn.*, vol. 55, pp. 222-237, 2005.

[18] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram", *J. Geophys. Res.*, vol. 106, pp. 7183-7192, 2001.

[19] H. Engedahl, "Implementation of the Princeton ocean model (pom/ecom3d) at the Norwegian Meterological Institute (DNMI)", Norwegian Meteorological Institute, Oslo, Norway, Res. Rep. 5, 1995.

[20] N. G. Winther and J. A. Johannessen, "North Sea circulation: Atlantic inflow and its destination", *J. Geophys. Res.*, vol. 111, p. C12018, 2006.

[21] S. Orre, Y. Gao, H. Drange, and J. E. \O. Nilsen, "A reassessment of the dispersion properties of 99Tc in the North Sea and the Norwegian Sea", *J. Mar. Syst.*, vol. 68, pp.24-38, 2007.

[22] M. Ilicak, T. M. Özgökmen, H. Peters, H. Z. Baumert, and M. Iskandarani, "Performance of two-equation turbulence closures in three-dimensional simulations of the Red Sea overflow", *Ocean Model.*, vol. 24, pp. 122-139, 2008.

[23] B. F. Rogowitz, and L. A. Treinish, "Why should engineers and scientists be worried about color?", IBM Research Center, 1996.

[24] J. P. Nolan, "Maximum likelihood estimation and diagnostics for stable distributions", in *Lévy Processes: Theory and Applications*, O. E. Barndorff-Nielsen, T. Mikosch, and S. I. Resnick, Eds. Birkhäuser, Boston, 2001, pp. 379-400.

[25] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall, USA, 1992.

Received: January 23, 2012 Revised: February 13, 2012 Accepted: March 23, 2012