Uncertainty forecast from 3-D super-ensemble multi-model combination: validation and calibration

**Baptiste Mourre · Jacopo Chiggiato · Fabian Lenartz ·
Michel Rixen**

Abstract Measurements collected during the Recognized Environmental Picture 2010 experiment (REP10) in the Ligurian Sea are used to evaluate 3-D super-ensemble (3DSE) 72-hour temperature predictions and their associated uncertainty. The 3DSE reduces the total Root-Mean-Square Difference by 12 and 32% respectively with reference to the ensemble mean and the most accurate of the models when comparing to regularly distributed surface temperature data. When validating against irregularly distributed in situ observations, the 3DSE, ensemble mean and most accurate model lead to similar scores. The 3DSE temperature uncertainty estimate is obtained from the product of a posteriori model weight error covariances by an operator containing model forecast values. This uncertainty prediction is evaluated using a criterion based on the 2.5th and 97.5th percentiles of the error distribution. The 3DSE error is found to be on average underestimated during the forecast period, reflecting (i) the influence of ocean dynamics and (ii) inaccuracies in the a priori weight error correlations. A calibration of the theoretical 3DSE uncertainty is proposed for the REP10 scenario, based on a time-evolving amplification coefficient applied to the a posteriori weight error covariance matrix. This calibration allows the end-user to be confident that, on average, the true ocean state lies in the $-2/+2$ 3DSE uncertainty range in 95% of the cases.

Keywords Coastal operational oceanography · Ocean temperature · Uncertainty forecast · Super-ensemble · Decision support · Ligurian Sea · REP10

B. Mourre (✉) · J. Chiggiato
NURC, Viale San Bartolomeo, 400, 19126 La Spezia, Italy
e-mail: mourre@nurc.nato.int

F. Lenartz
GHER, Université de Liège, Allée du 6 août,
4000 Liège, Belgium

M. Rixen
World Meteorological Organization, 7bis avenue de la paix,
1211 Geneva, Switzerland

1 Introduction

Operational ocean forecasts are useful for a wide range of applications at sea, including naval operations, commercial shipping, maritime security, fishing or environmental management. Based on the predictions provided by operational centers, decisions might be taken to conduct operations, route ships or to preserve the environment in particularly sensitive areas. The accuracy of the uncertainty associated with a given ocean forecast is a critical input towards robust decision making. An underestimation of the uncertainty may lead to an underestimation of the risk associated with the decision. Inversely, an overestimation may prevent potential action.

Different operational ocean models may be found to provide divergent predictions. As an illustration, Fig. 1 shows three 48-hour operational forecasts[1] of

---

[1]From the Navy Coastal Ocean Model run at the Naval Research Laboratory Stennis Space Center, from MARS3D run at PRE-VIMER and from the Regional Ocean Modelling System run at the Nato Undersea Research Centre.
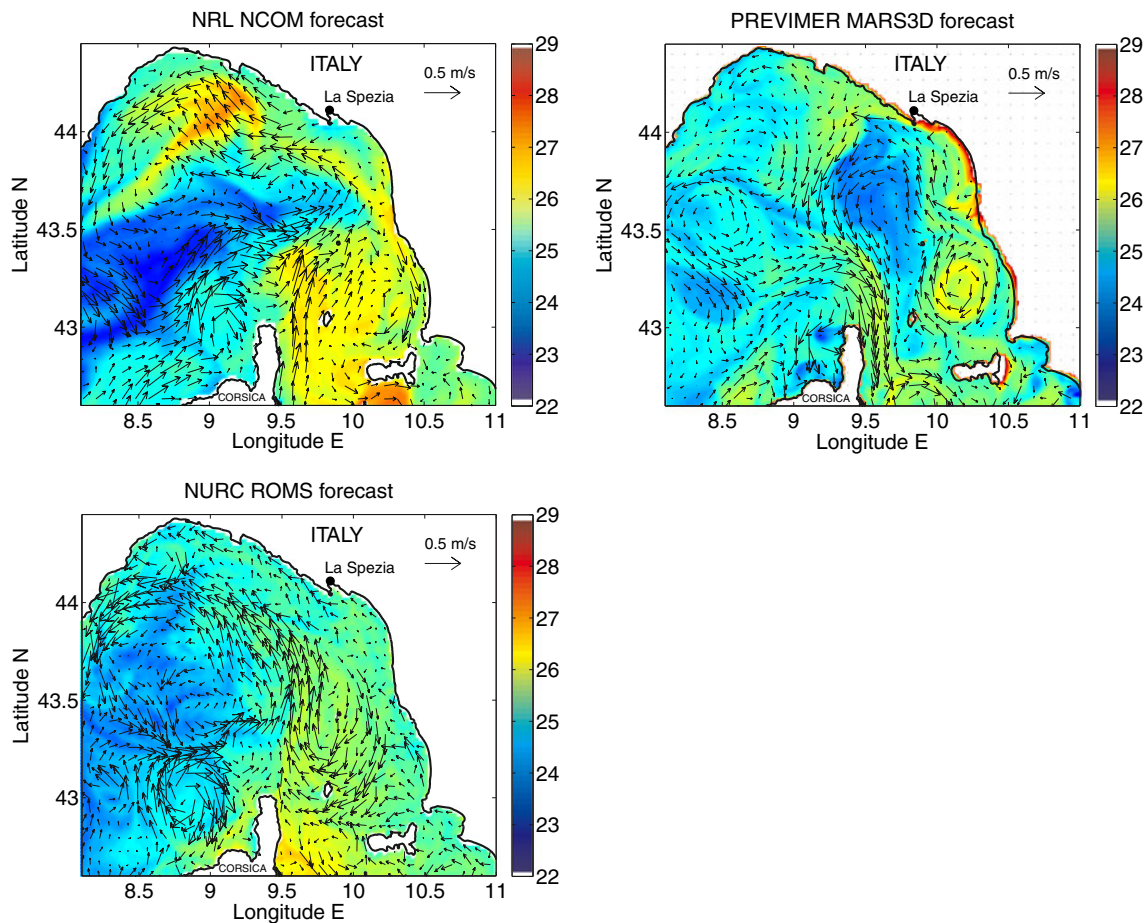
Fig. 1 NRL-NCOM, PREVIMER-MARS3D and NURC-ROMS ocean model forecasts released on 25 August 2010 00:00 and valid for 27 August 2010 00:00. The *color* and *arrows* respectively represent the surface temperature and currents

surface ocean temperature and currents valid for 27 August 2010 at 00:00 in the Ligurian Sea (Western Mediterranean). The three models, which are further described in Section 3.2, have a similar horizontal resolution around 1.5 km. The oceanographic circulation in the Ligurian Sea is characterized by the confluence of two northward currents into the so-called Northern Current, which then flows along the Italian and French coast as part of the Liguro-Provençal basin-wide cyclonic circulation (Pinardi et al. 2006). The two northward currents flow on both sides of Corsica island. While the Eastern Corsican Current is considered to weaken during summer in absence of wind forcing, the Western Corsican Current is more constant throughout the year. Moreover, an intense mesoscale variability affects this general circulation pattern. Differences in the model domain, physics, discretization, boundary conditions and atmospheric forcing generate different representations of these main circulation patterns and their associated mesoscale structures, so that surface temperature and current predictions for a given date

differ from one model to the other. From a practical perspective, the reconciliation of these forecasts is necessary to support the decision maker, who needs to make the best use of these multiple predictions.

Assuming that different models are able to capture different aspects of the true field, multi-model combination methods have been applied for forecast purposes during the last decade in meteorology and oceanography. The interested reader may refer to Krishnamurti et al. (1999), Doblas-Reyes et al. (2005), Logutov and Robinson (2005), Rixen et al. (2009) or Vandenbulcke et al. (2009). An overall finding from these authors is that forecasts from optimal multi-model combinations are, on the average, more accurate than the individual forecasts and their ensemble mean. More recently, the 3DSE formulation (Lenartz et al. 2010) was developed, aiming to cope with both the limited coverage of observations in the 3-dimensional coastal ocean and the potential spatial variability of individual model skills. The technique is based on the optimization of the spatially variable weighted linear

combination of the models during a specified learning period. The optimal weights are then used to combine the corresponding ocean model forecasts. The 3DSE was shown to reduce the surface and subsurface Root-Mean-Square Difference against temperature observations collected in 2008 in the Ligurian Sea by respectively 57% and 35% with reference to the model ensemble mean during 48-hour forecast periods (Lenartz et al. 2010).

One particularity of the 3DSE method in comparison with conventional ocean data assimilation approaches is that it minimizes error variances in the space of model weights instead of the space of the physical variables. As a consequence, the analysis step does not directly provide error covariances in the physical space, but produces error covariances associated with the model weights. An additional step is required to transfer these weight error covariances into forecast error covariances associated with the physical ocean variables. The accuracy of the resulting forecast uncertainty then needs to be properly evaluated before it can be confidently used in an operational context. This constitutes the main motivation of this paper.

The Recognized Environmental Picture 2010 experiment (hereafter REP10) was conducted in the Ligurian Sea in August-September 2010. Mainly focused on the use of remote sensing data and underwater autonomous glider vehicles for the rapid characterization of the oceanic environment, it allowed an extensive collection of satellite and in situ observations over a 2-week period in a 150×150 km^2 area offshore La Spezia, Italy. These observations are used here as a data set of opportunity to evaluate the 3DSE uncertainty prediction. The paper focuses on ocean temperature, which is the variable with the densest spatio-temporal observation sampling.

In this context, the objectives of this article are to (i) evaluate 3DSE forecast skills during REP10 experiment, (ii) evaluate the accuracy of the 3DSE uncertainty forecast, and (iii) propose a calibration of this uncertainty. The paper is organized as follows: the 3DSE method and REP10 sea trial are introduced in Sections 2 and 3 respectively, the results concerning 3DSE forecast skills, 3DSE uncertainty forecast validation and calibration are presented in Section 4, and conclusions are drawn in Section 5.

2 Method

2.1 The 3-D super-ensemble

The 3DSE prediction is a weighted linear combination of several model forecasts, with model weights spatially varying in the geographical domain. The technique uses a common 3DSE grid onto which all the models are first interpolated. The optimal weights are the result of a least-square minimization of the difference between observations and the weighted linear combination of the models. This minimization, which is carried out over a recent period called learning period, is computed over the whole modelling domain based on a priori model weight error covariances. A schematic view of the 3DSE method is represented in Fig. 2.

The 3DSE model can be expressed as follows:

$$T_{3DSE}^{p}(t) = \sum_{i=1}^{N} w_i^p \; T_i^p(t) \tag{1}$$

where T is the ocean variable under consideration (ocean temperature in this paper). Subscript i refers to the ith model forecast, superscript p to the pth point of the 3DSE spatial grid. N is the number of models. The least-square minimization is performed through a classical bayesian estimation, which may be expressed in the Optimal Interpolation formalism by defining the state vector $\mathbf{x}$ as the vector of model weights over the modelling domain:

$$\mathbf{x} = \begin{pmatrix} w_1^1 \\ \vdots \\ w_1^M \\ w_2^1 \\ \vdots \\ w_N^M \end{pmatrix} \tag{2}$$

M is the number of ocean points in the 3DSE grid. The length of the state vector $\mathbf{x}$ is MN.

The gain matrix $\mathbf{K}$, the analysis state vector $\mathbf{x}^a$ and associated error covariance matrix $\mathbf{P}^a$ are computed according to the following equations:

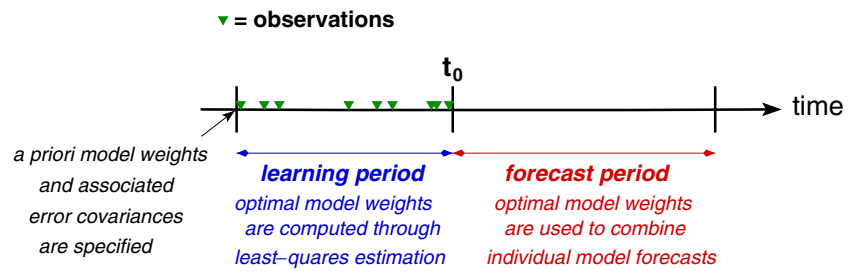$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T \left(\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R} \right)^{-1} \tag{3}$$

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K} \left(\mathbf{y}^o - \mathbf{H} \mathbf{x}^f \right) \tag{4}$$

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{K} \mathbf{H} \mathbf{P}^f \tag{5}$$

$\mathbf{P}^f$ (of size $MN \mathrm{x} MN$) is the a priori weight error covariance matrix and $\mathbf{R}$ (of size $Q \mathrm{x} Q$, where Q is the number of observations) the observation error covariance matrix. Superscript T denotes the matrix transpose. $\mathbf{y}^o$ is the vector of observations. Superscripts f and a refer to a priori and a posteriori analysis values respectively.

$\mathbf{R}$ is chosen diagonal, and contains the contributions of instrumental and representativity error variances at observation points. $\mathbf{H}$ is the observation operator

The vertical resolution is 5 m in the surface layer and decreases with depth. Regarding the temporal aspect, the duration of the learning and forecast periods are respectively fixed to 48 and 72 h.

2.2 3DSE uncertainty calculation

The a posteriori error covariances associated with the optimal model weights are obtained from Eq. 5. Forecast error covariances associated with the physical ocean variable can be inferred from these a posteriori weight error covariances after multiplication by model forecast values. At time t_0, the 3DSE error covariance matrix $\mathbf{P}^a_{3DSE}$ (describing forecast errors of the physical variable) is given by:

$$\mathbf{P}^a_{3DSE}(t_0) = \hat{\mathbf{H}}(t_0)\mathbf{P}^a\hat{\mathbf{H}}^T(t_0) \tag{6}$$

where $\hat{\mathbf{H}}$ is the operator projecting the space of model weights (space of dimension MN) onto the space of the 3DSE forecast variable (space of dimension M).

$$\hat{\mathbf{H}}(t) = \begin{pmatrix} T_1^1(t) & 0 & \cdots & T_N^1(t) & 0 \\ & \ddots & & & \ddots \\ 0 & T_1^M(t) & \cdots & 0 & T_N^M(t) \end{pmatrix} \tag{7}$$

$\hat{\mathbf{H}}(t)$ contains the values of model temperature forecasts at 3DSE gridpoints at time t. The number of rows in the matrix $\hat{\mathbf{H}}$ is the number of 3DSE spatial gridpoints (M), while the number of columns is the number of 3DSE gridpoints times the number of models (MN). Notice that spatial and cross-model weight error correlations resulting from the analysis and present in the matrix $\mathbf{P}^a$ are essential elements in this computation.

As a first approximation, and lacking a priori knowledge of the temporal evolution of the weight errors, weight error covariances are considered as stationary during the forecast period. The evolution of 3DSE error variances is then only due to the temporal variability of model temperature predictions, which is represented by the evolution of the operator $\hat{\mathbf{H}}(t)$. The evaluation of 3DSE error variances under this first approximation is presented in Section 4.2.1.

(of size $Q \times MN$), which transforms vectors from the space of model weights into the observation space. Unlike more conventional data assimilation methods, $\mathbf{H}$ does not contain the spatial interpolation coefficients from the 3DSE model grid to the position of the observations, but the product of these spatial interpolation coefficients by the values of ocean model temperatures at 3DSE gridpoints, such that $\mathbf{Hx}$ provides the 3DSE model values at observation points. Notice that the a posteriori weight error covariance matrix $\mathbf{P}^a$ does not depend on the actual observation-model mismatches, but on the position of the observations, their error covariances and the a priori weight error covariances.

The state vector $\mathbf{x}^f$ is initialized with an homogeneous value for the model weights equal to $\frac{1}{N}$, so that the 3DSE a priori solution is the ensemble mean. The initialization of $\mathbf{P}^f$ requires assumptions about the a priori weight error variances (σ_0^2) and correlations. In this study, initial weight error variances are specified as spatially homogeneous ($\sigma_0^2 = 0.01$). Spatial weight error correlations are chosen to be distance-dependent, with decorrelation scales of 50 km in the horizontal and 20 m in the vertical. These values were fixed as the result of sensitivity tests carried out with the REP10 data set. Initial cross-model weight error correlations are set to zero.

The 3DSE weight estimation can be made recursively to take into account observations prior to the learning period defined for a given simulation date. In this case, $\mathbf{x}^f$ and $\mathbf{P}^f$ are derived from a posteriori values of the previous 3DSE run. Lacking a better estimation of the evolution of the model weights and corresponding error covariances, $\mathbf{x}^f$ and $\mathbf{P}^f$ are considered as stationary between two runs.

The observations used for the learning and forecast validation are presented in Section 3.1. The three ocean models used as inputs of the 3DSE are described in Section 3.2. The 3DSE domain covers an area of approximately 150×150 km^2 in the Ligurian Sea, extending from 8.5°E to 10.6°E in longitude, from 43.1°N to 44.4°N in latitude and from 0 to 200 m in depth. The horizontal resolution of the 3DSE grid is 3 km.

In reality, weight error variances are expected to increase during the forecast period under the influence of ocean dynamics, which may change the optimal values of model weights. A calibration of the uncertainty aiming to take this temporal evolution into account is presented in Section 4.2.2.

3 Data

3.1 REP10 experiment

The REP10 experiment took place in the Ligurian Sea in August-September 2010. A wide variety of sensors and platforms were deployed at sea, including autonomous underwater glider vehicles, CTD (conductivity-temperature-depth) stations, surface temperature from Lagrangian drifters, shipborne surface CTDs and one towed vehicle (ScanFish) from the NATO Research Vessel Alliance. Four gliders sampled the study area from 0 to 200 m. In addition, one Spray glider provided ocean profiles from the surface to 1,000 m. Figure 3 shows the position of all available REP10 temperature observations at 80-meter depth from 20 August to 2 September 2010 in the area covered by the 3DSE model.

At the surface, the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA, Stark et al. 2007) was used to describe the daily observed ocean temperature conditions. OSTIA uses satellite data provided by the GHRSST project (Group for High-Resolution Sea Surface Temperature, http://www.ghrsst.org), together with in situ observations, to produce a high-resolution ($\frac{1}{20}^{\circ}$) daily analysis of Sea Surface Temperature (SST)
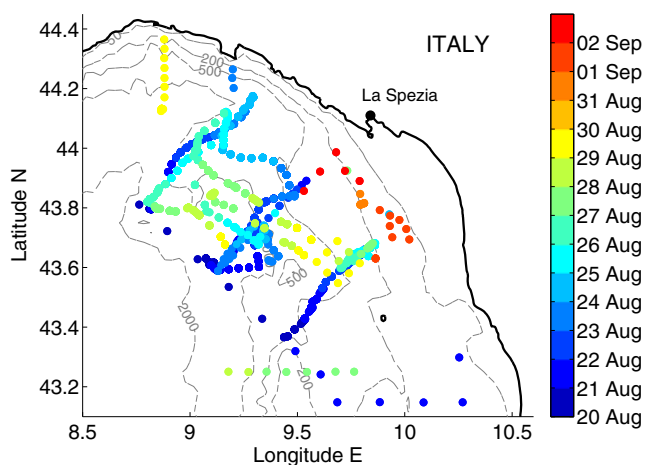
Fig. 3 Locations of REP10 in situ temperature observations at 80-meter depth. The color indicates the day of the measurement. Dashed lines represent bathymetric contours

for the global ocean. OSTIA SST represents the so-called foundation temperature, which is the surface temperature at the first time of the day before the heating impact of the solar radiation. Daily OSTIA SST were assimilated at 07:00 in the 3DSE model (notice that the term *assimilated* is used in the 3DSE context to qualify the observations used during the learning period to optimize the model weight estimates). One of the main advantages of using OSTIA products in 3DSE was that it provided a regularly distributed smoothed description of the SST over the whole modelling domain, thus providing a different data type compared to the irregularly distributed in situ subsurface profiles.

In the case of high-sampling rate observation platforms, a preprocessing was performed to reduce both the number of data assimilated during the learning period and their representativity error. In more concrete terms, (1) each glider, CTD and ScanFish profile was interpolated on the 3DSE vertical grid; (2) surface temperature measurements from Lagrangian drifters were taken at hourly rates; and (3) ship surface CTD measurements were averaged over 10-minute periods. In the 3DSE simulations, the total observation error representing both the instrumental noise and the representativity error was fixed to 1°C and 0.5°C for OSTIA products and in situ observations respectively.

The length of the data collection period allowed to repeat the 3DSE experiment in different sampling and oceanic scenarios by moving the 3DSE 5-day window (2-day learning and 3-day forecast) from August 21st to 31st. Eleven 3DSE simulations were computed starting daily at 00:00. Both the recursive (using previous a posteriori weights and errors as initial values for the next run) and the non-recursive (using default initial weights and errors for all the simulations) forms of the method were implemented during this period. Table 1 provides the mean, minimun and maximum number of observations used for 3DSE learning and validation over the set of eleven simulations carried out during the REP10. On average, approximately 6500 measurements were used to train the method. In addition, more than 2000 temperature observations were available on average for each 24-hour forecast validation period. Notice that the observations used for the validation of a given simulation were then used for the learning phase of the next simulation.

3.2 Ocean models

Three high-resolution ocean models collected during REP10 experiment were used in this study. The

Table 1 Mean, minimum and maximum number of observations used for 3DSE learning and validation over the set of 11 simulations carried out during the REP10

Number of observations	Learning	Validation (by forecast range)		
		0–24 h	24–48 h	48–72 h
Mean	6495	3151	2910	2048
Minimum	2842 (21-Aug)	890 (30-Aug)	890 (29-Aug)	906 (29-Aug)
Maximum	8757 (24-Aug)	4466 (22-Aug)	4466 (21-Aug)	4171 (23-Aug)

first operational ocean forecasting system was the NCOM model (Navy Coastal Ocean Model, Martin (2000)) coupled to the NCODA (Navy Coupled Ocean Data Assimilation, Cummings (2005)) data assimilation module. NCOM was run at the US Naval Research Laboratory NRL-SSC. A triple-nested version, similar to the one set up to support the Maritime Rapid Environmental Assessment 2007 and Ligurian Sea Cal/Val 2008 sea trials (Coelho et al. 2009), was implemented for REP10. The 3DSE used the outputs from the medium-grid model (NCOM Nest1 version). This model configuration had a horizontal resolution of 1.8 km and a vertical resolution varying from 2 m at the surface to 10 m at 200-m depth. The atmospheric forcing was provided by the Coupled Ocean Atmosphere Mesoscale Prediction System (Hodur 1997) Europe-3. The model was initialized with the NCODA analysis field after assimilation of realtime SST, sea surface height, CTD and glider data. Output data were available every hour with a 72-hour forecast range.

The second system was the French PREVIMER (http://www.previmer.org), which provided ocean forecasts over the north-western Mediterranean Sea with a 1.2-km resolution and 30 vertical sigma-coordinate levels. The core ocean model was MARS3D (Lazure and Dumas 2008). The surface forcing was provided by MM5 run at ACRI-ST. The north-western Mediterranean domain was nested in the Mediterranean forecasting system (MFS, Oddo et al. 2009). No data assimilation was performed in this operational forecast system. At the time of this research, output data were available every 3 h with forecast range exceeding 72 h.

The third system was the NURC operational ocean forecasting system based on the ocean model ROMS (Haidvogel et al. 2008) with a setup dedicated to the REP10 sea trial framework. The horizontal resolution was 1.8 km, with 32 vertical s-coordinate levels. The ROMS Ligurian Sea domain was nested in the MFS model (Oddo et al. 2009). The atmospheric model COSMO-ME (Bonavita and Torrisi 2005) of the Italian Air Force National Meteorological Center (Centro Nazionale per la Meteorologia e Climatologia Aeronautica - CNMCA) provided the surface forcing. No

data assimilation was performed in this operational forecast system. Output data were available every 3 h with a 72-hour forecast range.

4 Results

4.1 3DSE forecast skills

Figure 4 illustrates the temperature forecast skills of the individual models, their ensemble mean and the 3DSE. The x- and y-axis respectively represent the unbiased Root-Mean-Square Difference (RMSD) and the absolute bias (or mean difference) between models and observations during the 72-hour forecast period. The color represents the linear correlation coefficient between models and observations. The dashed lines represent the total RMSD, which encompasses both the
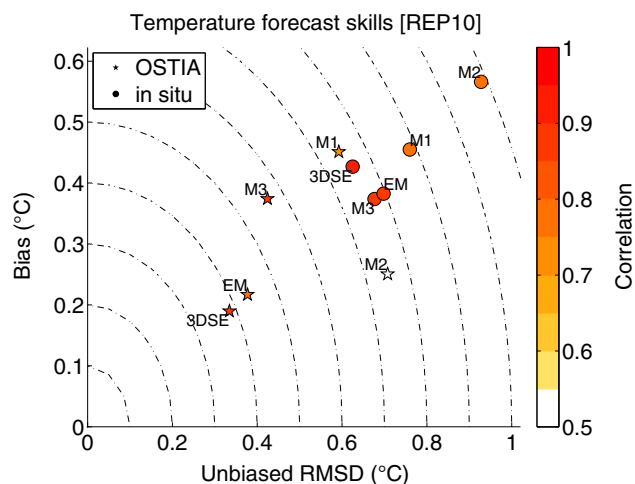
Fig. 4 Averaged temperature forecast skills during REP10. M1, M2 and M3 represent the three ocean models used in 3DSE. EM is their ensemble mean. The *colorbar* displays the linear correlation coefficient between models and observations. Deviations from OSTIA SST values are represented by a *star marker*, while deviations from in situ observations are displayed by a *circle marker*

unbiased RMSD and the bias. The model deviations from (1) OSTIA values and (2) in situ observations are treated separately since they provide two different validation scenarios. OSTIA products are smooth and regularly distributed SST data, while in situ measurements (from gliders, CTDs, ship surface CTD, towed ScanFish, Lagrangian drifters) are irregularly distributed over the spatial and temporal domains during both the learning and forecast periods. Statistics plotted in this diagram are averaged over the set of eleven simulations from 21 to 31 August 2010 using the recursive form of the method.

When compared to OSTIA SST products, the 3DSE forecast exhibits a bias and an unbiased RMSD of 0.19°C and 0.33°C respectively. The total RMSD of 0.38°C is reduced by respectively 12% and 32% with respect to the ensemble mean (EM) and the most accurate of the models. This good performance shows that the 3DSE trained using, among other data, regularly distributed measurements is still in good agreement with this source of observation during the forecast period. The main contributions to this 0.38°C mismatch come from (i) observation errors, (ii) the concurrent assimilation of measurements from other observing platforms in the surface mixed layer and (iii) the evolution of the ocean state due to ocean dynamics during the forecast period. The correlation between 3DSE forecast and OSTIA is 0.88, showing that the large scale surface temperature patterns described in OSTIA analysis are properly represented by the method. This good overall agreement with these regularly distributed data constitutes a first and necessary validation of the 3DSE, before it can be evaluated in the case of more irregular observations, where the particular spatio-temporal distribution may also impact the results.

As expected, the 3DSE forecast skills are degraded when comparing to in situ observations. The total RMSD obtained from the 3DSE (0.76°C) is similar to that obtained with EM (0.80°C) and with the most ac-

curate of the models (0.77°C), even if the contributions to this total RMSD slightly differ. The 3DSE forecast has a larger bias than EM and the most accurate of the models (0.43°C vs 0.38 and 0.37°C), but a reduced unbiased RMSD (0.62°C vs 0.70 and 0.68°C). The linear correlation coefficient with respect to these data is 0.91, which mainly indicates that the sign of the vertical gradients is properly represented in 3DSE predictions. These RMSD scores are in reasonable agreement with the results obtained by Lenartz et al. (2010). During LSCV08 experiment in the Ligurian Sea, authors reported a 3DSE total RMSD of 0.30°C at the surface and 0.68°C with respect to subsurface observations. The relative 3DSE skill improvement compared to the individual models and their ensemble mean is weaker here than in this previous study due to the consideration during REP10 of models with better average skills than those used during LSCV08.

4.2 3DSE uncertainty

Figure 5 (right panel) shows the 3DSE temperature uncertainty prediction at 80-meter depth produced on 29 August 2010 and valid for 1 September 00:00. The position of observations assimilated during the learning period is displayed by the yellow squares. The corresponding temperature forecast is plotted on the left panel. The magnitude of the uncertainty estimate (from 0.1 to 1.7°C) is consistent with the magnitude of the total RMSD illustrated in Fig. 4 (average value of 0.76°C). The spatial shape of the 3DSE uncertainty estimate is strongly impacted by the position of assimilated observations. The closer to the observations, the more the weights used to combine the models are applicable and hence the error associated with the 3DSE forecast is reduced. The 50-km decorrelation scale used to define a priori horizontal weight error correlations leads to larger uncertainties at the edges of the 3DSE domain, where there is no measurement in a 50-km radius.

Fig. 5 *Left*: 3DSE temperature forecast at 80-meter depth produced on 29 August 2010 and valid for 1 September 00:00 (t0+72 h). *Right*: associated uncertainty (*yellow squares* display the position of assimilated observations at this depth)
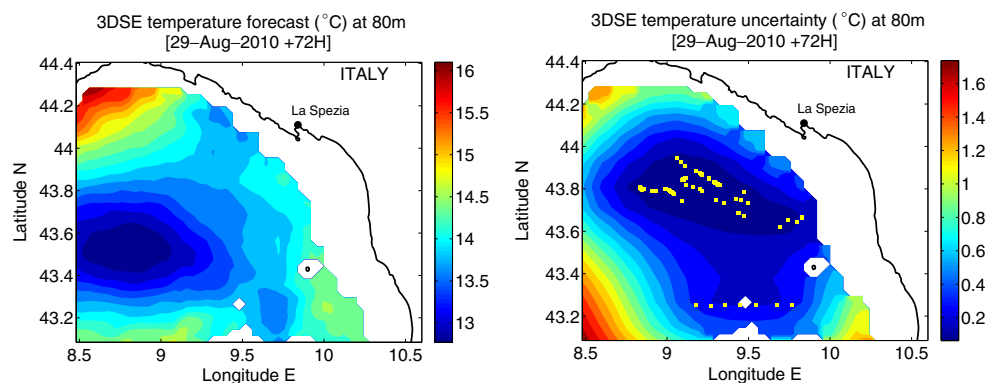
The large number of observations taken in the central part of the domain around 43.8°N leads to a reduced uncertainty of about 0.05°C.

4.2.1 Validation

The indirect calculation of the 3DSE uncertainty, the impact of a priori assumptions about weight error co-variances and the unknown influence of ocean dynamics during the forecast period make the evaluation of the 3DSE uncertainty estimate necessary. The non-recursive simulations of the 3DSE are considered in this Section since they provide a larger number of independent test cases for this evaluation. The criterion used for the validation of the uncertainty is based on the 2.5th and 97.5th percentiles of the distribution of the predicted random variables. Assuming the symmetry of these distributions and making reference to the Gaussian distribution, we expect approximately 95% of observations (within the range of their observation error) to fall between -2 and $+2$ error standard deviations (σ_{3DSE}) from the 3DSE value. When this percentage is found to be lower (respectively larger) than 95%, 3DSE error variances are considered to be underestimated (respectively overestimated).

Let us first evaluate the 3DSE uncertainty computed during the learning period at the position of assimilated observations. Figure 6 displays the percentage of observations falling outside the $-2/+2$ σ_{3DSE} range during the learning period for the eleven 3DSE simulations, as well as the average value over the whole set of simulations. The variability of this percentage with the date of the simulation is related to the heterogeneous distribution of the observations throughout the REP10
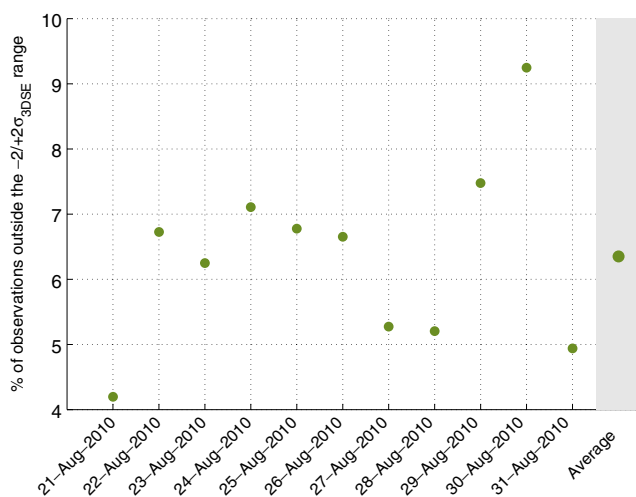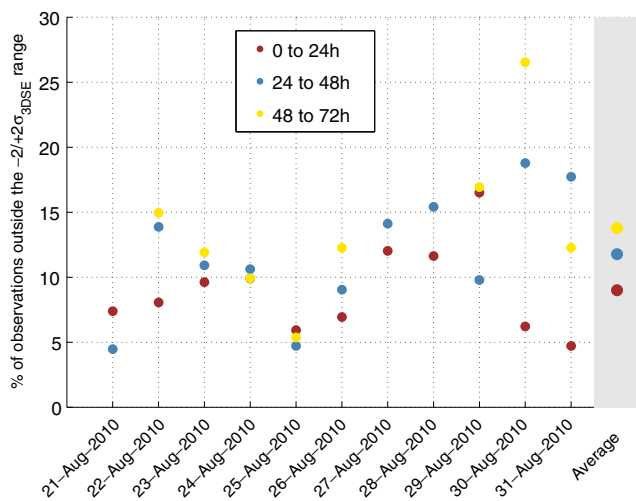

Fig. 6 Percentage of observations falling outside the $-2/+2$ σ_{3DSE} range during the learning period

cruise. Values range from 4.2 to 9.2%, with an average 6.3% of observations falling outside the $-2/+2$ σ_{3DSE} range. Given the variety of observation platforms used in this experiment and the inevitable departure of the statistics from the normal distribution, we consider this value sufficiently close to 5% to validate the 3DSE uncertainty estimate at observation points during the learning period. This first validation aspect is an important prerequisite since it indicates that the method properly propagates the uncertainty from the space of model weights onto the space of the 3DSE physical variable.

The uncertainty prediction during the forecast period implies additional mechanisms for which the quantitative impacts cannot be readily evaluated: (i) the validation is carried out at arbitrary locations, sometimes far away from the location of assimilated observations, so that it may be affected by the inaccuracies in the a priori spatial weight error correlations, and (ii) ocean variablity is expected to increase the model weight uncertainty by changing the state of the system. Only a full and repetitive coverage of observations during the forecast period would allow to properly characterize the accuracy of both the spatial variability of the uncertainty, which is linked to the spatial correlations of the weight errors, and its temporal variability, which is related to ocean dynamics. In a real scenario such as the REP10 experiment, the limited and heterogeneous distribution of the measurements over the spatio-temporal domain prevents from properly distinguishing between the spatial and temporal variability of the uncertainty. The focus is made here on the temporal variability and the observations are binned by 24-hour windows over the whole modelling domain.

Figure 7 shows the percentage of observations falling outside the $-2/+2$ σ_{3DSE} uncertainty range during the forecast period for the eleven simulations, as well as the average value. Again, the heterogeneous distribution of the measurements in space and time generates a strong variability of the results from the different simulations. Values range from 4.5 to 26.5%. Even if this percentage is not systematically lower during the first 24 h of the forecast (7 occurences over 11 simulations), the average value indicates an increase of this percentage with time. Respectively 9.0, 11.8 and 13.8% of the observations fall outside the $-2/+2$ σ_{3DSE} uncertainty range for the time windows 0–24 h, 24–48 h and 48–72 h. This increase is due to the dynamical evolution of the system which makes the model weights and their uncertainty computed during the learning period gradually less suited to the description of the ocean as time goes on during the forecast period. As expected, this proportion is larger than 5%, meaning that the

Fig. 7 Percentage of observations falling outside the $-2/+2$ σ_{3DSE} range during the forecast period (results are provided by 24-hour forecast ranges). The statistics for the 48–72 h time segment on August 21th, 27th and 28th are not available due to the failure of one of the input models

3DSE theoretical uncertainty is underestimated during the forecast period and requires calibration.

A detailed analysis focusing on the magnitude of the forecasted errors shows that the smallest 3DSE uncertainties are more reliably predicted than the largest ones. While 10.1% of all the observations located in a region with a 3DSE uncertainty lower than 0.2°C (i.e. in the vicinity of the assimilated observations) fall outside the $-2/+2$ σ_{3DSE} uncertainty range, this percentage rises to 35.2% when considering observation locations with a 3DSE uncertainty larger than 1.0°C. This difference points out the inaccuracies in the initial spatial weight error correlations, which propagate the uncertainty from the locations of the assimilated measurements into the rest of the domain.

4.2.2 Calibration

The underestimation of the 3DSE uncertainty during the forecast period can be corrected by amplifying the weight error covariance matrix $\mathbf{P}^a$:

$$\mathbf{P}^{a[amplified]}(t) = \mathbf{A}(t)\mathbf{P}^a\mathbf{A}(t) \tag{8}$$

where $\mathbf{A}$ is a diagonal amplification matrix allowing to introduce spatially variable amplification coefficients.

This formulation allows to increase error variances without modifying error correlations, preserving thus the rank of the covariance matrix. Given the rise of the 3DSE error with time during the forecast period (see Fig. 7), the amplification matrix is expected to be time-dependent, with an increasing tendency. Ideally,

a full and repetitive coverage of observations should allow to properly characterize the diagonal terms of the amplification matrix, as well as their temporal evolution. In a real scenario such as the REP10 experiment, the limited and heterogeneous distribution of the measurements over the spatio-temporal domain prevents from properly distinguishing between the spatial and temporal variability. In order to still get an insight into the temporal variability of the error, we consider an average homogeneous amplification coefficient α over the whole domain, and investigate its evolution over three 24-hour time windows during the forecast period. Under these conditions, Eq. 8 becomes:

$$\mathbf{P}^{a[amplified]}(t) = \alpha^2(t)\mathbf{P}^a \tag{9}$$

For a given 24-hour window, the optimal coefficient α is the coefficient leading to the percentage of observations falling outside the $-2/+2$ σ_{3DSE} range the closest as possible as 5%. This computation is achieved by finding the amplification of the 3DSE spread which minimizes the difference between this percentage of observations and 5%. Figure 8 shows the values of the optimal amplification coefficient α for the different simulations, as well as the average values. The coefficient α exhibits the same kind of variability as the one displayed in Fig. 7. Values are spread between 1.0 and 5.4 when considering the various simulation dates and forecast ranges. This large variability confirms that the result from a single simulation strongly depends on the particular observational situation, so that only the average values over a large number of simulations can provide more robust indicators. The average values of α for the three 24-hour time windows are ranked consistently compared to the average values shown in
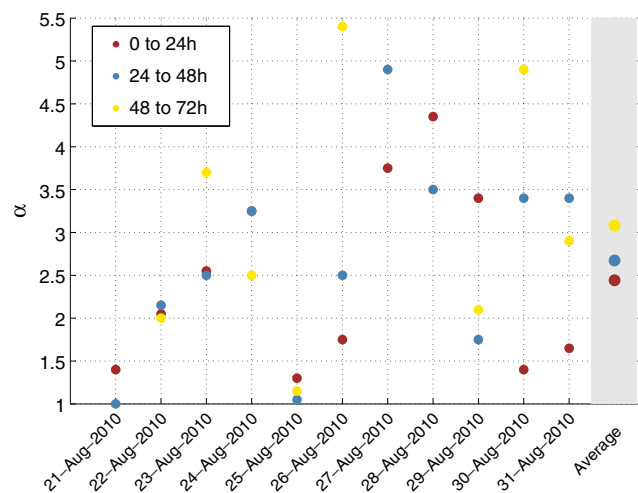
Fig. 8 Alpha coefficient leading to a calibrated 3DSE uncertainty during REP10

Fig. 7. The average values of α of 2.4, 2.7 and 3.1 for the time windows 0–24 h, 24–48 h and 48–72 h respectively, indicate the temporal increase of the amplification coefficient. The linear model best fitting these average values in a least-square sense is:[2]

$$\alpha(t) = 2.25 + \frac{0.32}{86400}t \tag{10}$$

where t is expressed in seconds since the beginning of the forecast.

Notice that this linear model provides an amplification coefficient α of 2.25 for $t = 0$, while a value close to one could have been expected given the reliability of the 3DSE uncertainty at the locations of the assimilated observations during the learning period. This underestimation of the 3DSE uncertainty at the beginning of the forecast period indicates that the spatial propagation of the weight error reduction from the locations of the assimilated measurements to the rest of the modelling domain still needs to be improved through a better characterization of the initial covariances. This empirical model for α also highlights the departure from the ideal case with homogeneous observation sampling and normal probability distributions. A proportion of 9% of observations outside the $-2/+2 \ \sigma_{3DSE}$ uncertainty range would indeed lead to an amplification factor of[3] 1.18 in the case of exact normality of statistics with an homogeneous sampling for their validation. Here, the diversity and limited coverage of observations lead to significant asymmetries in the validation sampling.

Finally, the potential refinement of the calibration by areas was also investigated, either based on the ocean dynamics or on the distance to assimilated observations. Unfortunately, no robust result could be obtained due to the high observational requirements for such a classification.

5 Conclusion

The 3DSE technique was evaluated with regards to ocean temperature prediction using data from the REP10 experiment. This experiment provided an extensive data collection over a 2-week period in a 150 × 150 km^2 area in the Ligurian Sea. The 3DSE provided

a total RMSD of 0.38°C against repetitive and regularly distributed surface temperature data. This RMSD was reduced by 12 and 32% respectively with reference to the ensemble mean and the most accurate of the models. When validating against in situ observations spread over the whole domain and taken at different depths from different sensors, the average total RMSD increased up to 0.76°C, and was similar to the one provided by the ensemble mean and the most accurate of the models. These RMSD scores are in agreeement with the 3DSE forecast skills presented by Lenartz et al. (2010) based on a previous cruise in the Ligurian Sea. Overall, the 3DSE method was shown to be an efficient statistical technique to combine forecasts from individual models and thereby improve forecast skills, provided that sufficient data has been collected in order to specify the appropriate weights during the learning period.

The associated 3DSE temperature uncertainty estimate was inferred from the product of the weight error covariances issued from the analysis by the operator $\hat{\mathbf{H}}$ containing model forecast values. The position of assimilated observations clearly defined the spatial patterns of the 3DSE uncertainty estimate due to the hypothesis of distance-dependent a priori weight error correlations. The confidence in the model weights, and as a consequence in the 3DSE temperature prediction, was larger in close proximity to the assimilated observations.

The evaluation of this 3DSE error forecast was made necessary by (1) the indirect calculation of this uncertainty, (2) the impact of a priori assumptions made about weight error covariances, and (3) the unknown influence of ocean dynamics during the forecast period. To address the first point, the uncertainty estimate was first validated during the learning period against assimilated observations. Our validation approach used a criterion based on the 2.5th and 97.5th percentiles of the distribution of the predicted random variables. By reference to the normal distribution, approximately 95% of the observations (within their own uncertainty) should fall within the $-2/+2 \ \sigma_{3DSE}$ uncertainty range (σ_{3DSE} being the 3DSE error standard deviation). The results indicate that the theoretical uncertainty was valid at the locations of the assimilated observations during the learning period, but underestimated on average compared to the true 3DSE error during the forecast period. This underestimation was due to both the system evolution due to ocean dynamics and the consideration of validation measurements outside the area spanned by the assimilated observations. Over eleven simulations run at different dates, the average percentage of observations falling outside the $-2/+2$

---

[2] The determination coefficient r^2 for this regression is 0.97.

[3] Value obtained from the tables of the error function erf: $1.18 = \frac{2}{\sqrt{2} \, \mathrm{erf} \, inv\left(1 - \frac{9}{100}\right)}$.

σ_{3DSE} uncertainty range was found to be 9.0, 11.8 and 13.8% respectively for the 0–24 h, 24–48 h and 48–72 h forecast time windows. This increase, which corresponds to the temporal growth of the 3DSE forecast error, illustrates the gradual maladaptation of the optimal model weights computed during the learning period to the description of the future ocean state.

This evaluation of the 3DSE uncertainty revealed inaccuracies of the a priori weight error covariances specified in the 3DSE. The distance-based spatial correlations did not include the potential anisotropic characteristics of the ocean dynamics, and therefore limited the accuracy of the 3DSE uncertainty estimate outside the area spanned by the assimilated observations. The inclusion of dynamical information coming from the individual models in the a priori weight error covariances should allow to improve the uncertainty estimate by better representing the influence of the regional oceanic features. However, this step is not straighforward since temperature error covariances, which can be deduced from model runs, can not be simply transferred in terms of model weight error covariances, which are needed by the 3DSE. The proper specification of initial weight error covariances still constitutes an open challenge to improve the 3DSE.

A simple calibration of the 3DSE uncertainty estimate was proposed to allow to meet, on average, the above-mentioned criterion with the present formulation of the initial covariances. This calibration consisted in an amplification of the a posteriori weight error covariance matrix. This amplification should theoretically be variable in both (i) space to highlight dynamically active areas and (ii) time to account for the growth of the uncertainty due to the evolution of the system. Unfortunately, only a full and repetitive coverage of observations could allow to properly describe the spatiotemporal variability of the 3DSE forecast error. In the REP10 observation scenario, and despite the extensive collection of measurements, the heterogeneity of data distribution and sensors precludes the separation of the spatial and temporal variability of the error. In this context, our approach focused on the temporal variability. Considering the average values over the eleven simulations, the simple linear model $\alpha(t) = 2.25 + \frac{0.32}{86400}t$ was proposed to amplify the weight error covariance matrix, and hence the associated 3DSE uncertainty, to obtain an error estimate which was consistent, on average, with the REP10 validation measurements. The application of this amplification coefficient allows confidence that, on average, the true ocean state lay within the $-2/+2\,\sigma_{3DSE}$ uncertainty range in 95% of the cases. This confidence is necessary before being able to use 3DSE forecasts as a potential support in any decision system. The proposed uncertainty calibration is inevitably linked to the REP10 observation scenario. Further studies are needed to evaluate its validity in other regions, for other time periods and using other data sets.

References

Bonavita M, Torrisi L (2005) Impact of a variational objective analysis scheme on a regional area numerical model: the Italian air force weather service experience. Meteorol Atmos Phys 88:39–52

Coelho E, Peggion G, Rowley C, Jacobs G, Allard R, Rodriguez E (2009) A note on NCOM temperature forecast error calibration using the ensemble transform. J Mar Syst 78S1: S272–S281. doi:10.1016/j.jmarsys.2009.01.028

Cummings J (2005) Operational multivariate ocean data assimilation. Q J R Meteorol Soc 131:3583–3604

Doblas-Reyes F, Hagedorn R, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. Tellus 57A:234–252

Haidvogel DB, Arango H, Budgell WP, Cornuelle BD, Curchitser E, Di Lorenzo E, Fennel K, Geyer WR, Hermann AJ, Lanerolle L, Levin J, McWilliams JC, Miller AJ, Moore AM, Powell TM, Shchepetkin AF, Sherwood CR, Signell RP, Warner JC, Wilkin J (2008) Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the regional ocean modeling system. J Comput Phys 227:3595–3624

Hodur RM (1997) The naval research laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). Mon Weather Rev 125:1414–1430

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, Gadgil S and Surendran S (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285(5433):1548–1550. doi:10.1126/science.285.5433.1548

Lazure P, Dumas F (2008) An external-internal mode coupling for a 3D hydrodynamical model for applications at regional scale (MARS). Adv Water Resour 31:233–250

Lenartz F, Mourre B, Barth A, Beckers JM, Vandenbulcke L, Rixen M (2010) Enhanced ocean temperature forecast skills through 3-D super-ensemble multi-model fusion. Geophys Res Lett 37(19):L19606. doi:10.1029/2010GL044591

Logutov OG, Robinson AR (2005) Multimodel fusion, error parameter estimation. Q J R Meteorol Soc 131:3397–3408

Martin PJ (2000) Description of the navy coastal ocean model version 1.0, NRL/FR/7322-00-9962. Naval Research Laboratory, 42 pp

Oddo P, Adani M, Pinardi N, Fratianni C, Tonani M, Pettenuzzo D (2009) A nested Atlantic-Mediterranean Sea general circulation model for operational forecasting. Ocean Sci 5:461–473. doi:10.5194/os-5-461-2009

Pinardi N, Arneri E, Crise A, Ravaioli M, Zavatarelli M (2006) The physical, sedimentary and ecological structure and variability of shelf areas in the Mediterranean Sea. In: Robinson AR, Brink K (eds) The sea, vol 14. Harvard University Press, Cambridge, pp 1245–1330

Rixen M, Book JW, Carta A, Grandi V, Gualdesi L, Stoner R, Ranelli P, Cavanna A, Zanasca P, Baldasserini G, Trangeled A, Lewis C, Trees C, Grasso R, Giannechini S, Fabiani A, Merani D, Berni A, Leonard M, Martin P, Rowley C, Hulbert M, Quaid A, Goode W, Preller R, Pinardi N, Oddo P, Guarnieri A, Chiggiato J, Carniel S, Russo A, Tudor M, Lenartz F, Vandenbulcke L (2009) Improved ocean prediction skill and reduced uncertainty in the coastal region from multimodel superensembles. J Mar Syst 78:S282–S289. doi:10.1016/j.jmarsys.2009.01.014

Stark JD, Donlon CJ, Martin MJ and McCulloch ME (2007) OSTIA : An operational, high resolution, real time, global sea surface temperature analysis system. Oceans '07 IEEE Aberdeen, conference proceedings. Marine challenges: coastline to deep sea. IEEE, Aberdeen, Scotland

Vandenbulcke L, Beckers JM, Lenartz F, Barth A, Poulain PM, Aidonidis M, Meyrat J, Ardhuin F, Tonani M, Fratianni C, Torrisi L, Pallela D, Chiggiato J, Tudor M, Book JW, Martin P, Peggion G, Rixen M (2009) Super-ensemble techniques: application to surface drift prediction. Prog Oceanogr 82(3):149–167. doi:10.1016/j.pocean.2009.06.002