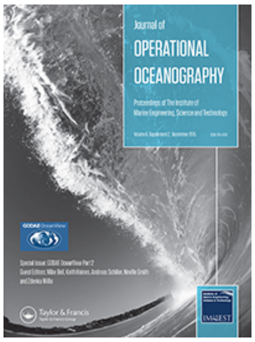


Taylor & Francis
Taylor & Francis Group

Journal of Operational Oceanography

ISSN: 1755-876X (Print) 1755-8778 (Online) Journal homepage: http://www.tandfonline.com/loi/tjoo20

Recent progress in performance evaluations and near real-time assessment of operational ocean products

Fabrice Hernandez, Edward Blockley, Gary B. Brassington, Fraser Davidson, Prasanth Divakaran, Marie Drévillon, Shiro Ishizaki, Marcos Garcia-Sotillo, Patrick J. Hogan, Priidik Lagemaa, Bruno Levier, Matthew Martin, Avichal Mehra, Christopher Mooers, Nicolas Ferry, Andrew Ryan, Charly Regnier, Alistair Sellar, Gregory C. Smith, Sarantis Sofianos, Todd Spindler, Gianluca Volpe, John Wilkin, Edward D. Zaron & Aijun Zhang

Journal of Operational Oceanography, 2015
Vol. 8, No. S2, s221–s238, http://dx.doi.org/10.1080/1755876X.2015.1050282

Taylor & Francis
Taylor & Francis Group

Recent progress in performance evaluations and near real-time assessment of operational ocean products

Fabrice Hernandez[a,e*], Edward Blockley[b], Gary B. Brassington[c], Fraser Davidson[d], Prasanth Divakaran[c], Marie Drévillon[e], Shiro Ishizaki[f], Marcos Garcia-Sotillo[g], Patrick J. Hogan[h], Priidik Lagemaa[i], Bruno Levier[e], Matthew Martin[b], Avichal Mehra[j], Christopher Mooers[j], Nicolas Ferry[e†], Andrew Ryan[b], Charly Regnier[e], Alistair Sellar[b], Gregory C. Smith[l], Sarantis Sofianos[m], Todd Spindler[j], Gianluca Volpe[n], John Wilkin[o], Edward D. Zaron[k] and Aijun Zhang[p]

[a]Institut de Recherche pour le Développement (IRD), LEGOS, Toulouse, France; [b]Met Office, Ocean Forecasting Research & Development, Exeter, UK; [c]Centre for Australian Weather and Climate Research, Australian Bureau of Meteorology, Melbourne, Australia; [d]Fisheries and Oceans, St Johns, Canada; [e]Mercator Océan, Ramonville St Agne, France; [f]Japan Meteorological Agency (JMA), Ohtemachi, Tokyo, Japan; [g]Puertos del Estado, Madrid, Spain; [h]Navy Research Laboratory/Stennis Space Center, Mississippi, USA; [i]Marine Systems Institute at Tallinn University of Technology, Tallinn, Estonia; [j]Environmental Modeling Center, NOAA/NWS/NCEP, College Park, Maryland, USA; [k]Portland State University, Portland, Department of Civil and Environmental Engineering, Oregon, USA; [l]Environment Canada, Montréal, Canada; [m]Ocean Physics and Modeling Group, University of Athens, Athens, Greece; [n]Istituto di Scienze dell'Atmosfera e del Clima, Rome, Italy; [o]Institute of Marine and Coastal Sciences, Rutgers, State University of New Jersey, USA; [p]Center for Operational Oceanographic Products and Service, NOAA, Silver Spring, Maryland, USA

Operational ocean forecast systems provide routine marine products to an ever-widening community of users and stakeholders. The majority of users need information about the quality and reliability of the products to exploit them fully. Hence, forecast centres have been developing improved methods for evaluating and communicating the quality of their products. Global Ocean Data Assimilation Experiment (GODAE) OceanView, along with the Copernicus European Marine Core Service and other national and international programmes, has facilitated the development of coordinated validation activities among these centres. New metrics, assessing a wider range of ocean parameters, have been defined and implemented in real-time. An overview of recent progress and emerging international standards is presented here.

Introduction

Operational ocean forecast systems (OOFSs) now provide a wide range of analyses and forecasts of the marine environment that can be exploited by many users. The value of the products to any particular user depends not only on the quality and skill of the products but also on the user's knowledge (and understanding) of the quality, skill and reliability of the products for his or her particular application. Since the initial implementation of OOFSs during the late 1990s, continuous efforts have been made to evaluate hindcast and forecast accuracy and skill (Hernandez 2011; Martin 2011). Accuracy and skill here are defined respectively as the OOFS products degree of closeness to the 'ocean truth' (Hernandez 2011) and the OOFS's usefulness for a given application (Jolliff et al. 2009). An overview of skill assessment using observations and other reference datasets representing this truth is given by Stow et al. (2009).

The calibration, validation, verification and quality control of OOFS products are core activities in ocean operational centres (Lellouche et al. 2013; Oke et al. 2013; Blockley et al. 2014) (OOCs). Usually calibration refers to a task in which model parameters are optimized. Here, the calibration phase refers to the last comprehensive scientific assessment of the new OOFS version before operation. The calibration phase is also often used to demonstrate that the new system performance is better than the existing system. Validation refers to the OOFS performance assessment while in operation. Verification is defined here as the quantification of OOFS skill based on independent data, i.e. not used to generate the products.

Methods for assessing OOFS reliability (Crosnier & Le Provost 2007) were defined in the early days of the Global Ocean Data Assimilation Experiment (GODAE) experiment (Bell et al. 2009), based on (1) consistency, (2) quality (or accuracy) and (3) added value as proposed and defined by weather forecast skill verification approaches (Murphy 1993; Murphy & Winkler 1987). The first two types of assessments are undertaken routinely by OOCs as 'internal metrics'. The third is considered as user-oriented, and requires use of 'external metrics' measuring the fitness for purpose (provision of dependable, reliable and repeatable information), or the value of ocean

*Corresponding author. Email: fhernandez@mercator-ocean.fr forecast services. This is also addressed by some OOCs in parallel with verification tasks performed by users.

Experts in OOCs across the world who are assessing the skill of OOFSs face similar issues with the observational data sets available for validating their products. In general, *in situ* and satellite measurements are collected by dedicated data assembly centres (DACs) that pre-process the data and make it available for OOFSs. Hence, for similar components of operational systems, methods and tools for assessing the representation of ocean processes can be shared. Owing to the nature and quality of the observations, validation experts also face comparable issues, such as the validation of mesoscale chlorophyll or primary products using ocean colour satellite data, or the use of Lagrangian approaches and drifters to verify the realism of eddies in regional models for oil-spill forecast skill. As a result, there is a great potential for collaboration within the scientific community in this area.

Naturally, working groups were set up to tackle, as a community, these validation issues. This started earlier within the ocean observation community, which raised expert groups to develop guidelines and standards for providing state-of-the-art ocean observation products. Some examples of these are the Ocean Surface Topography Science Team for sea surface height (SSH) and satellite altimetry (www.aviso.altimetry.fr), the Group for High Resolution Sea Surface Temperature (GHRSST) for sea surface temperature (SST) from various satellite and *in situ* sensors (www.ghrsst.org/), the Argo team for *in situ* vertical profiles of primarily temperature (T) and salinity (S) (www.argo.ucsd.edu), the Global Ocean Surface Underway Data group for sea surface salinity (SSS) (www.gosud.org) and the International Ocean Color Coordinating Group (www.ioccg.org).

This paper aims to highlight recent progress in near-real-time monitoring of OOFS performance, and to describe different validation strategies and their limitations. For the sake of completeness, we also present DACs validation procedure for advanced observed-based products. Some recent examples are presented and discussed in the first section. The next section illustrates progress by OOCs in integrating validation in their systems. More specifically, since GODAE (Bell et al. 2009), the operational community has maintained a partnership to share and standardize validation methodologies. This community has gained mutual benefit from inter-comparing their ocean products and inferring the relative strength and weaknesses of the operational systems (Oke et al. 2012). These issues are addressed in the framework of the ongoing GODAE OceanView program (Schiller et al. 2015) (GOV) (www.godae-oceanview.org) by the Intercomparison and Validation Task Team (IV-TT). Three initiatives have started and that are ongoing: an Ocean Reanalysis Intercomparison (Balmaseda et al. 2015); the organization of a multi-model ensemble forecast approach for ocean surface parameters; and the organization of the near-real-time operational product 'Class 4 metrics' inter-comparison against observations, described later in this paper and detailed in two companion papers (Ryan et al. 2015; Divakaran et al. 2015).

Recent improvements in near-real time and operational assessment

New metrics

Presently, real-time OOFS skill assessment focuses on various aspects of the dynamics of physical and biogeochemical processes of the ocean, at different time-scales, over different areas, and with different purposes and uses. Evaluation metrics have evolved in order to synthesize different aspects of system performance together. For example, Taylor (2001) and target (Jolliff et al. 2009) diagrams consider root-mean-square error (RMSE) or root-mean-square differences (RMSD) together with anomaly correlations versus observations. Similarly, cost functions and model efficiency values (Hyder et al. 2012) can provide a synthesis of model performance indicators.

Furthermore, new metrics have been designed to characterize other properties. For example, in the case of search and rescue, ensemble predictions and clouds of dispersion (Melsom et al. 2012) have been used to evaluate the contribution of uncertainty in ocean currents to drift projections. Dispersion is also assessed using multi-model approaches (such as Fukushima Cesium 137 concentration estimates; Masumoto et al. 2012). New metrics have also been defined for sea-ice, such as contingency tables and distribution statistics used over ensemble coupled model seasonal forecast experiments (Benestad et al. 2011). Skill assessment of ocean biogeochemical models has also been addressed recently. In particular, Lynch et al. (2009) point out the failure of a model to accurately represent the 'ocean truth', but also the failure to correctly assess the effective skill of the model using appropriate metrics.

Assimilation performance assessment

In parallel, the monitoring of the performance of analysis systems has been continuously improved. Statistics derived from innovations (observation minus background) and residuals (observation minus analysis) are used to assess the consistency of the assimilation framework (including the model background and observation error covariances). In the case of ensemble analysis systems, these statistics can be used to verify the adequacy of forecast spread (Balmaseda et al. 2013; Desroziers et al. 2005; Desroziers & Ivanov 2001). Verifying and reducing ocean model biases is also an important issue, as many assimilation methods are based on the assumption that models have no bias, which can reduce the efficiency of analysis methods and even lead to unphysical increments if biases are present and not handled correctly.

Rigorous skill assessment in the assimilation framework is a difficult task: most available observations are used to adjust models and reduce analysis errors. Thus, independent assessment is only possible by: (1) withholding part of the dataset for statistical quantification of errors (a trade-off between a sufficient population size to estimate a statistic while not significantly impacting the quality of the system performance being measured); or by (2) using other sources of data that have not been assimilated (Gregg et al. 2009). The latter is generally employed with data not available in near-real time, which is useful for reanalysis (or hindcast) evaluation, but not for operational routine verification.

Longer-term forecast assessment

Most of the OOCs provide short-term forecasts (from a few days, to 1–2 weeks), but some have begun providing longer monthly forecasts, like the Japanese Meteorological Agency MOVE/MRI.COM-WNP OOFS. It covers a large part of the Northwest Pacific (15°N–65°N, 117°E–160°W), with a specific zoom (1/10° resolution) over 15°N–50°N, 117°E–160°E (Usuii et al. 2006), and uses a multivariate Three-Dimensional Variational (3DVAR) data-assimilation scheme (Fujii & Kamachi 2003). Persistence and 1- to 30-day forecasts are compared against analyses to provide RMSE statistics. A forecast skill metric is used, whereby the ratio of forecast RMSE over the persistence RMSE is calculated for a given forecast lead-time. Using this skill score, the forecast provides useful skill compared with persistence if the ratio is below 1. Results from the MOVE OOFS are shown in Figure 1 for the velocity field at 50 m depth over the period February 2006 to January 2008. Even for 30-day forecasts, the system performs better than persistence in most areas around Japan, and fails only in the vicinity of the Kuroshio Extension area [Figure 1(a–c)]. Moreover, Kuroshio dynamics and predictability are assessed using a specific metric based on the Kuroshio main axis. The Kuroshio axis error is defined as the distance between a forecast and the 'true' axis position over the 133–139°E, 30–35°N region, where the axis position is determined using the position of the 15°C isotherm at 200 m depth from the analysis. Figure 1(d) indicates that the dynamical forecasting system is consistently better than persistence at all lead times. This type of metric is also useful to convey forecast skill to users, as the error is expressed in tangible terms (here a distance in kilometres) rather than an abstract unit-less skill score or RMSE value.

Specific approaches for regional operational systems

Recent improvements have also been made in terms of evaluating specific regional and mesoscale dynamics. For example, the Gulf of Mexico Pilot Prediction Project (GOMEX-PPP, http://abcmgr.tamu.edu/gomexppp/) is investigating the OOFSs performance for predicting the evolution of the Loop Current in the Gulf of Mexico. The

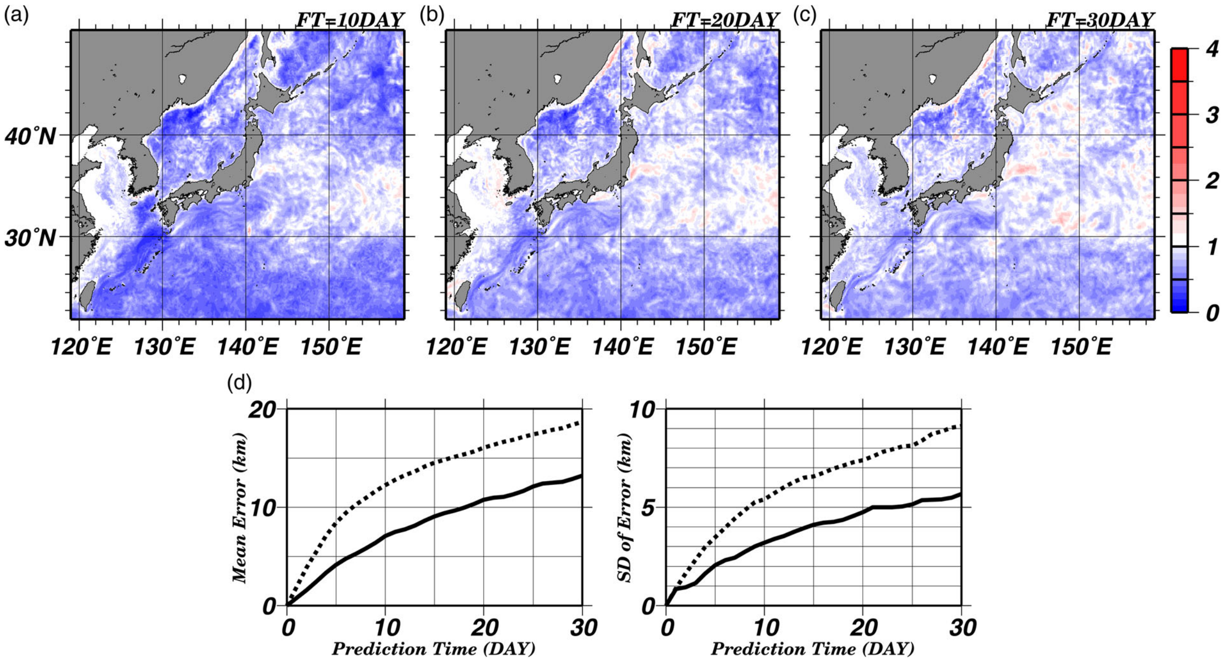


Figure 1. MOVE/MRI maps of forecast skill, comparing forecast and persistence RMSE statistics against analyses for velocity field at 50-m depth, for 10-, 20- and 30-day forecast lead-time, respectively [(a), (b) and (c)]. Forecast beats persistence for values below 1. (d) Performance assessment of the Kuroshio axis position (in kilometres; see text for definition) for the forecast (solid line) and persistence (dashed line), for 0- to 30-day lead-time.

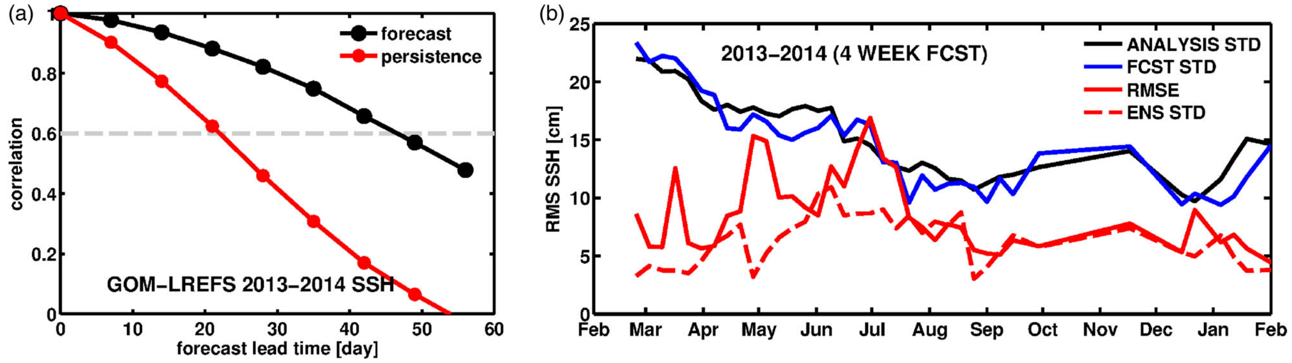


Figure 2. Real-time SSH assessment of the GOM-LERFS Gulf of Mexico OOFS against satellite altimetry data. Comparison with analysis is restricted to water deeper than 200 m in the subdomain, 82W to 89W and 22N to 28N. (a) Correlation of SSH anomalies for persistence (red) and forecast (black). (b) RMSE of the ensemble mean forecast (red) and ensemble spread standard deviation (red dashed) for SSH. Also shown is the spatial variability in term of standard deviation (STD) of the analysis (black) and the 4-week ensemble mean forecast (blue).

Long Range Ensemble Forecasting System (GOM-LERFS) developed at the Naval Research Laboratory (Stennis, USA), has been providing 2-month forecasts since January 2013, with the intention of supporting end users impacted by strong currents associated with the Loop Current and its eddies, and to provide boundary conditions for coastal ocean models. This 3-km resolution OOFS performs weekly 32-member 60-day forecasts. It is initialized using an analysis provided by the Navy Coupled Ocean Data Assimilation (NCODA) scheme that uses a 5-day assimilation window to ingest satellite altimetry, SST and *in situ* data obtained from the global telecommunication system. A verification of SSH is performed in which forecasts are compared in real-time with along-track SSH data following the Class 4 metrics approach (i.e. in observation space; discussed below). Then, in model space, the ensemble is used to assess the probabilistic forecast skill. In Figure 2, statistics of weekly comparisons against analyses for the period January–September 2013 show that forecasts remain skilful for approximately twice as long as persistence. The SSH anomaly variance agrees closely in the forecast and verifying analysis, but the ensemble standard deviation does not appear to predict the forecast error, suggesting that the ensemble spread does not fully capture the forecast error patterns. Adequately sampling the uncertainty in initial conditions, model physics and forcing is an important aspect of ocean ensemble prediction that requires further study.

Another example of forecast skill assessment is the dynamical feature-based validation approach used for the Experimental System for Predicting Shelf/Slope Optics (ESPreSSO), operated in real-time by Rutgers University over the New Jersey coast Mid-Atlantic Bight (Wilkin & Hunter 2013). This OOFS is based on the 7-km horizontal resolution Regional Ocean Modeling System (ROMS) using boundary conditions from the HYCOM-NCODA global OOFS. The system is initialized using daily analyses from a Four-Dimensional Variational (4DVAR) analysis system with a 3-day analysis window (Moore et al. 2011), which assimilates a large set of data [including glider T/S profiles and CODAR HF-Radar measurements (http://www.myroms.org/espresso/)]. The 4DVAR approach allows a better quantification of model errors by assessing the impact of the assimilated data, thereby permitting the correction of large-scale biases. Slope currents and water masses used for real-time applications are evaluated using all available data. A specific off-line verification is performed using independent surface drifters, moored data and SSH. Moreover, a dedicated multi-model real-time assessment has been performed, comparing estimates from ESPreSSO together with three other regional OOFS, and three global OOFS (including HYCOM-NCODA) in order to evaluate the OOFS' prediction skill for subtidal currents and shelf water mass changes. This assessment is comprehensively discussed in Wilkin and Hunter (2013) including a presentation of performance improvements through downscaling strategies. In their figures 4 and 5, improvements to Taylor and Target diagrams are proposed, to represent individual vs average performance, and seasonal model biases respectively.

Validation of biogeochemical products

In the field of ecosystem modelling and marine-resources management, *in situ* data for adequate validation of operational products are sparse. Hence, satellite ocean colour (OC) products remain the main source of information for estimates of phytoplankton pigment concentration distribution [i.e. chlorophyll, CHL; Figure 3(a)]. The OC Thematic Assembly Centre (TAC), within the European MyOcean project (www.myocean.eu), has developed specific processing chains to operationally distribute state-of-the-art, quality-checked daily OC observations over both global and regional domains. The need for

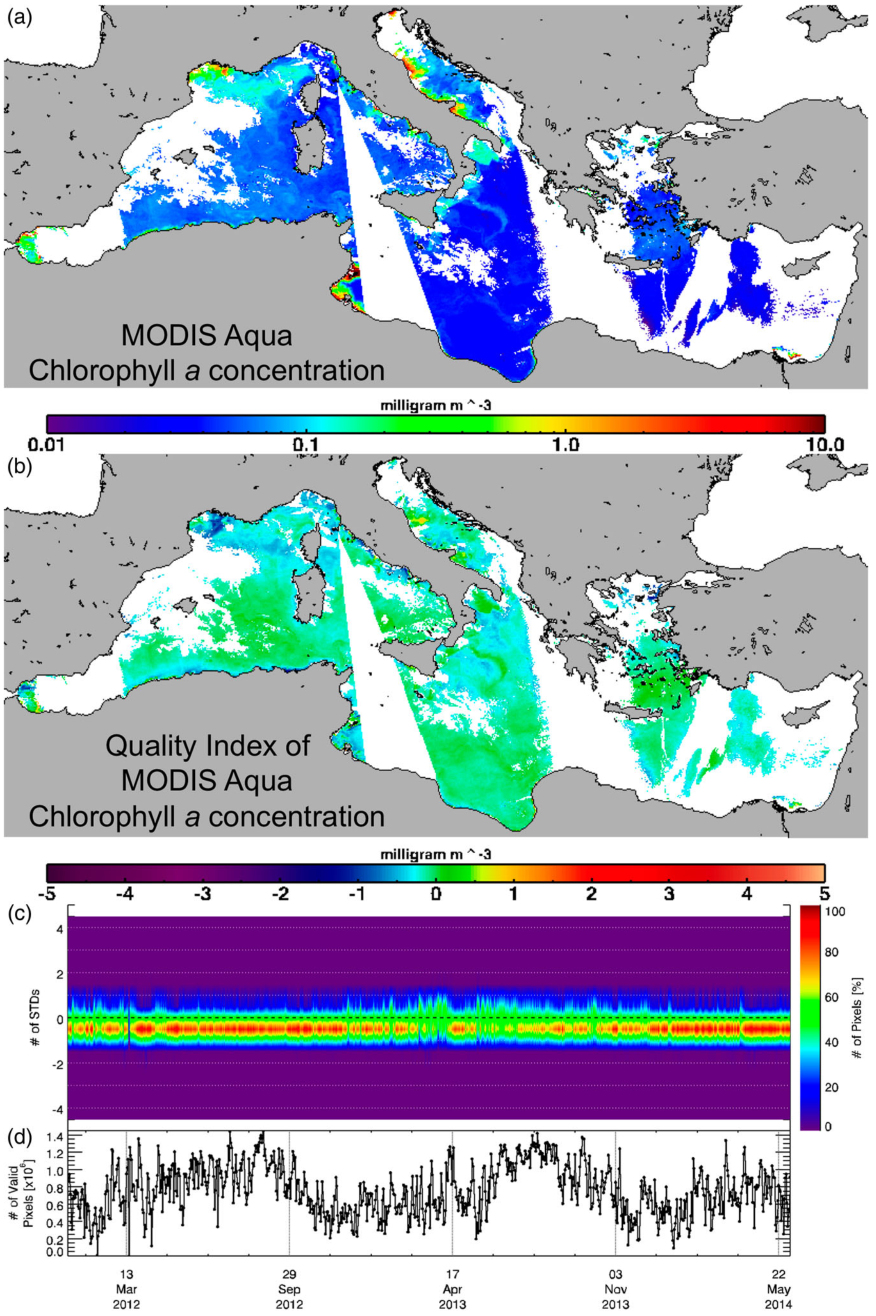


Figure 3. MODIS Aqua daily Level-3 CHL product, processed via the MedOC3 algorithm, for 30 May 2014 (a), with the corresponding Quality Index: normalized departure from the daily climatology (b). The full online validation statistics time series is given by the shaded plot in panel c, with time on the x axis and the histogram bins of the Quality Index on y axis; colours show the percent occurrence with respect to the total number of valid pixels (classified as reliable by the operational processing and not flagged as cloud or other contamination factor), as described by the time series on panel d.

regional processing comes from the demonstrated inadequacy, at regional scales, of global algorithms to generate reliable products of sufficient accuracy (Volpe et al. 2012). For example, the oligotrophic waters of the Mediterranean Sea were shown to be significantly less blue and more green than the global ocean (Volpe et al. 2007). The OC TAC provides value-added products (generally not distributed by space agencies) such as: (1) daily merged fields from different sensors; (2) Level 4 products without data voids owing to clouds generated using both Optimal Interpolation and Empirical Orthogonal Function approaches; and (3) products that account for the two broad classes of bio-optical regimes (open ocean and coastal waters). Level 4 and Level 3 (L4 and L3) mentioned here refer to product levels as defined by the Committee on Earth Observation Satellites (www.ceos.org). Typically, L4 products are regular maps of a given parameter, obtained by merging and processing similar measurements from different sources, and using specific estimation methods (optimal interpolation, krigging, etc.). For these observation-based products, both offline and online quality assessments are performed by the OC TAC. The former refers to the comparison of space–time co-located *in situ* and satellite derived products for quantities such as spectral remote sensing reflectance, total suspended matter, coloured dissolved organic matter and chlorophyll concentration. In real-time, such data are not sufficiently robust, and validation is limited to a consistency assessment (Hernandez 2011), where the daily climatology for each region is used as a reference to make pixel-based comparison (online validation). A Quality Index, based on the normalized departures from climatology, is computed from the SeaWiFS sensor [no longer operational; Figure 3(b)]. The overall statistics and distribution are analysed, as illustrated for the Mediterranean Sea CHL [Figure 3(c) and (d)]. In parallel, the monitoring of input data (number, quality …) has increased the reliability of the products.

Many initiatives have led to progress on skill assessment of biogeochemical modelling (Stow et al. 2009). For example, as part of the Ocean Carbon Model Intercomparison Project, univariate metrics were proposed to quantify both physical and biogeochemical parameters of the coupled simulations (Doney et al. 2009). Multivariate metrics (i.e. quantifying the reliability of both the parameters and their relation to observed processes) (Allen & Somerfield 2009), or map-based validation (Rose et al. 2009), is also emerging in this field. However, for real-time assessment of ecosystem-biogeochemical forecasts, most of the OOFSs can only rely on references given by OC satellite products. Moreover, the dynamics of biogeochemical systems is strongly characterized by the patchiness of its properties generated by oceanic mesoscale, which causes heterogeneity in concentration fields (Levy & Martin 2013). Consequently, most forecast verifications mimic OC product assessment, by analysing in a similar way, at the pixel level, the model equivalent to CHL and optical satellite data (Lazzari et al. 2012), as described in the previous paragraph.

Validation of sea-ice products

Growing interest in polar regions has driven the need for improved sea-ice verification metrics to demonstrate the capacity and quality of sea-ice forecast skill to potential users. This effort has been hindered by the reliability and availability of observational datasets together with a lack of knowledge of how to adequately account for nonlinearities in the verification metrics. Contingency table-based metrics, introduced in the early twentieth century (Pearson 1904), have been re-popularized, as well as the root-mean-square distance of ice edge. However, these metrics may not be relevant for regional or process-dependent verification. In particular, errors in ice edge location assessment are sensitive to the definition of 'ice edge', as multiple ice edges may be present and the total error will be sensitive to the length of the ice edge. Hence, the metrics defined for the Arctic might not be suitable in the Baltic Sea, and definitions should consider sub-regional scaling (e.g. size of a gulf) (Lagemaa 2013). However, even if the ice metric is not properly defined, it still gives valuable user information for the dense marine traffic regions like the Baltic Sea.

An example of a contingency table-based metric from the Canadian Meteorological Centre Global Ice-Ocean Prediction Systems (GIOPSv1.0) is shown in Figure 4. GIOPSv1.0 uses a 3DVAR ice concentration analysis for correcting the Los Alamos sea-ice model by assimilating satellite data together with daily ice charts from the Canadian Ice Service. The reference dataset is given by the Interactive Multisensor Snow and Ice Mapping System (IMS) analyses from the National Oceanic and Atmospheric Administration (NOAA) National Ice Centre that provide binary fields of ice/open water on a 4 km grid. Sea-ice analyses suffer from an incomplete coverage of observations, with data-reliability issues and the mis-representation of leads. A particular issue is the high sensitivity of passive microwave retrievals to surface melt, often resulting in erroneous values of open water in summer. Contingency table statistics produced using IMS analyses (applying a threshold of 0.4 to determine binary ice/water values from the GIOPSv1.0 ice concentration forecasts) are used in order to evaluate the proportion of correct ice, or correct water. These contingency scores are computed separately for forecast and persistence fields. Then, differences of scores for forecasts and persistence are computed. These metrics are mapped for 2011 in Figure 4 showing skilful 7-day forecasts along most of the ice edge. In other words, 7-day forecasts beat persistence considering the prediction of correct proportion of sea ice and correct water.

Reliability assessment of input information

Another recent aspect in OOFS validation strategy is the systematic feedback of errors and anomalies to providers of input data. For instance, validation of atmospheric forcing fields is now carried out for some wave-prediction systems (Feng et al. 2006). Moreover, inputs of ocean assimilation systems, such as *in situ* data collected by TACs or DACs, can suffer in real-time from incomplete levels of quality control. While automatic procedures are applied for the rapid distribution of the observations in real-time, more detailed visual analysis is often left for delayed-time datasets. Other analyses usually depend on the level of expertise of the provider. In near-real time,

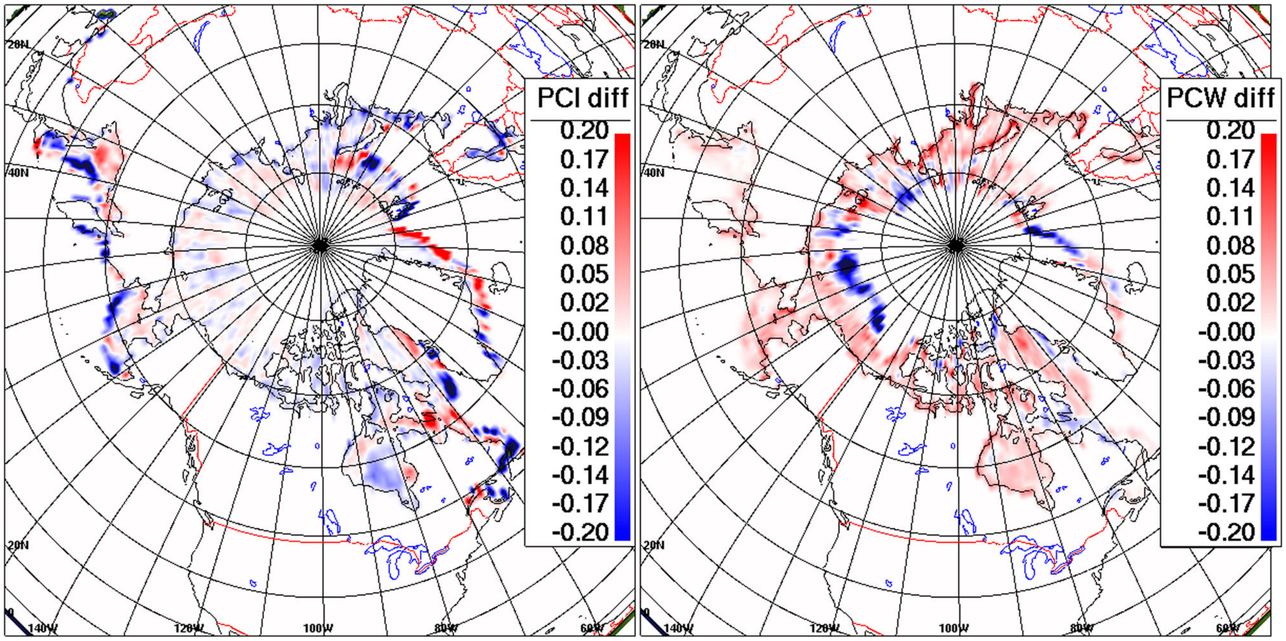


Figure 4. Contingency analyses of GIOPSv1.0 sea-ice forecast over 2011 for 7-day forecasts, using IMS data as reference. Forecast minus persistence (skilful/erroneous, red positive/blue negative resp.) for proportion of correct ice (left) and correct water (right)

erroneous *in situ* profiles can drastically impact the quality of ocean analyses. As a result, systematic quality control has been implemented in many OOFSs to prevent this. At Mercator Océan, two techniques are applied for *in situ* T/S profiles. First, innovations (guess/forecast minus observation) are tested against a threshold envelope. This envelope is defined using statistics of innovations from an ocean reanalysis and is used to detect anomalous observations (e.g. blue dots between 500 and 1000 m depth in Figure 5). Second, dynamic heights are computed from the T/S increments, and then probability density functions are constructed for consistent dynamical areas, in order to detect points outside from the normal distribution. In some cases, feedback to producers is organized through blacklisting (Cabanes et al. 2013).

Development of integrated operational verification systems

Several OOCs have recently taken steps to structure calibration, validation, and verification activities, in real-time or delayed mode, as an integrated component of the OOFSs. In the USA, the national backbone of real-time data, tidal predictions, data management and operational modelling supporting NOAA's missions (http://tidesandcurrents.noaa.gov) under the National Operational Coastal Modeling Program now performs quality control and forecast skill verification in a centralized way for all OOFS through the Continuous Operational Real-Time Monitoring System.

Similarly, the MyOcean IBI (Irish Biscay Iberian shelves) OOFS team (Puertos del Estado, Spain, and Mercator Océan, France) has developed a comprehensive tool called NARVAL (Numeric Assessment for Regional VALidation) to check its operational performance, in terms of consistency, accuracy and reliability. NARVAL uses available observations, such as: satellite-derived Sea Level Anomalies (SLA), SST and SSS (from both L3 and L4 products), *in situ* T/S profiles, HF-radar surface currents and tide gauge sea level. This tool builds on the MyOcean project structure such that the input data are quality checked by the TACs. NARVAL is modular and extendible for any new data sources as a reference (measurements, climatologies or model estimates). All validation information produced is archived for further evaluation. Additionally, the 'On-line Mode Validation' provides an automated quality and consistency assessment, and is routinely performed for each forecast bulletin (from the previous day's hindcast up to 5-day forecasts). It generates Class 1–4 metrics (Hernandez et al. 2009) that provide daily statistics and an evolution of the skill score for each parameter over the past two weeks [Figure 6(a)]. Furthermore, a 'Delayed Mode Validation' provides an overall review of the IBI product quality over longer time periods (i.e. monthly, seasonal and annual). Real-time statistics are accumulated to provide a synthesis assessment over longer periods, while dedicated metrics using off-line datasets, can focus on particular ocean phenomena or parameters. Metrics are performed over the whole domain (26°N–56°N, 19°W–5°E), but also over specific sub-regions of interest – both for users and for verification teams, for example: Strait of

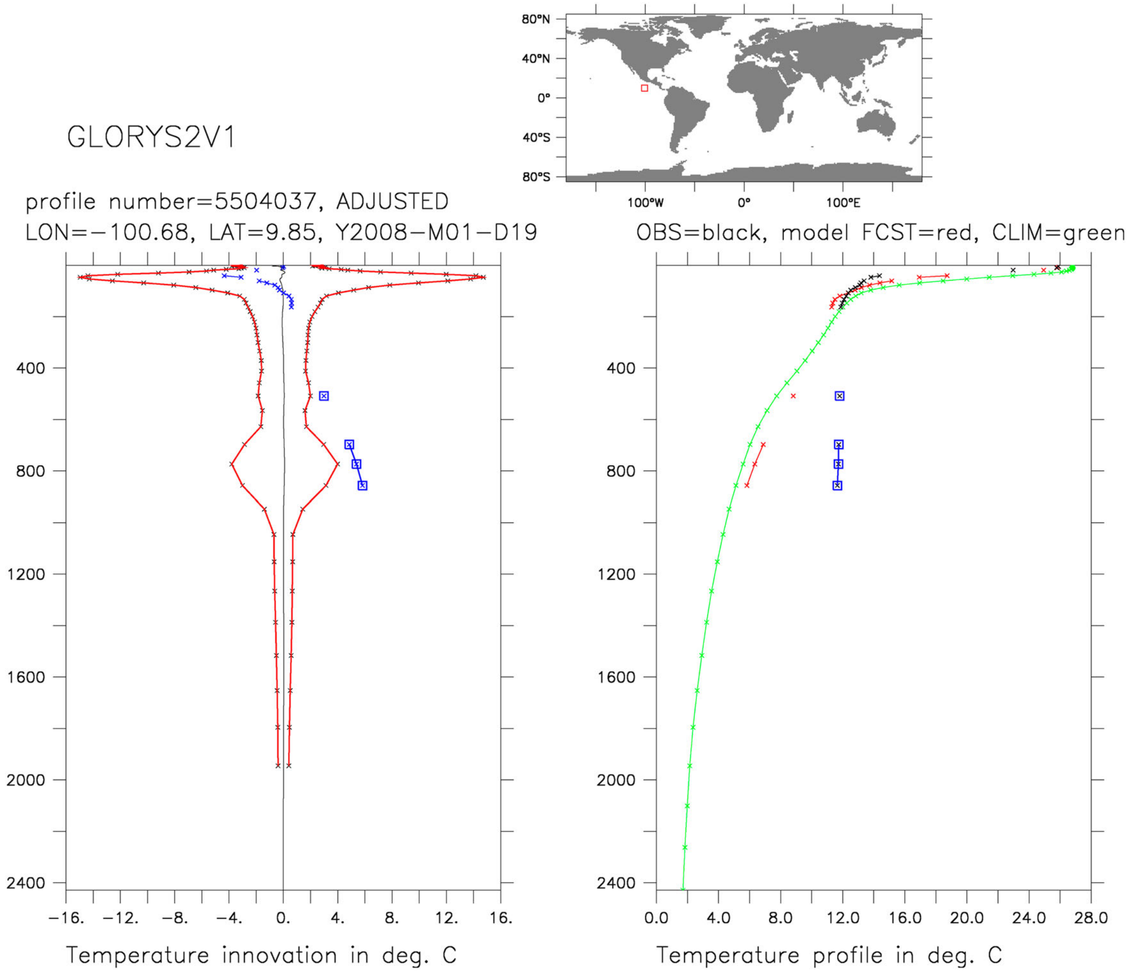
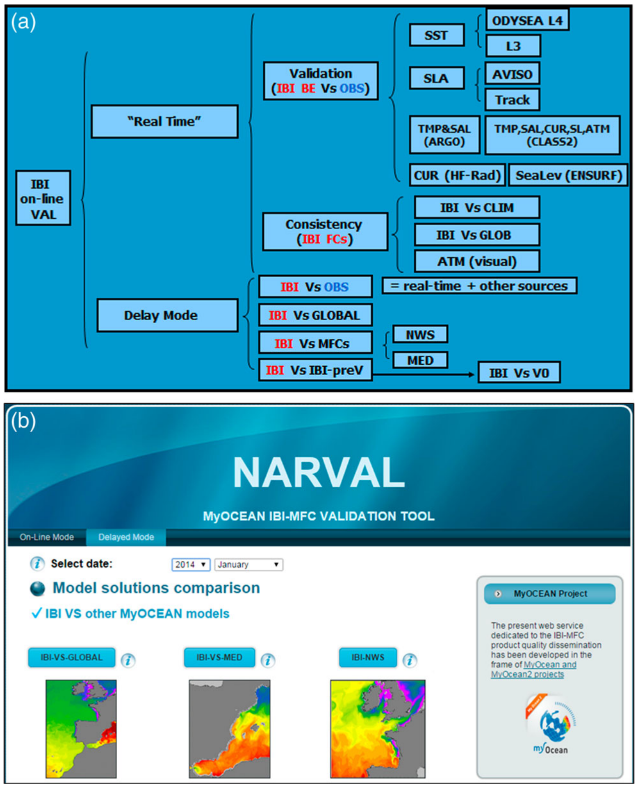


Figure 5. Detection of *in situ* profile anomalies before assimilation in the GLORYS2V1 Mercator Océan reanalysis. Left: innovation (blue dots) and threshold envelop (red). Right: temperature-profile observations (blue dots), model forecast for the corresponding observations (red) and climatology (green). In this case, the cluster of blue dots in the depth range 500–1000 m has not passed the test. Top: location in the equatorial Pacific of the profile.

Gibraltar, English Chanel, Western Mediterranean Sea, Gulf of Biscay, Western and Northern Iberian shelves, the Canary Islands area and the Irish Sea.

Using NARVAL, the performance can be monitored for specific areas, dynamics and OOFS, as illustrated for SST using a Taylor diagram (Figure 7). This type of figure is obviously complex, but it is used by validation teams to monitor, at a glance, several systems' SST scores over various areas. NARVAL has been designed to allow automatic inter-comparison between IBI and adjacent regional OOFS within the MyOcean framework for the Mediterranean Sea and the North-West-Shelf [Figure 6(b)]. Comparison with adjacent OOFSs aims primarily to maintain consistency in products and user delivery. Comparison with the global OOFS (within which the IBI OOFS is nested) quantifies added value of the regional shelf system (Figure 7), that representing tides and high-frequency upper ocean dynamics. There are also comparisons against coastal systems over key areas, such as the SAMPA (*Sistema Autónomo de Medición, Predicción y Alerta*) system around the Gibraltar Strait (Lorente et al. 2014).

In the MyOcean framework, a similar methodology is implemented for the Baltic Sea by Danish, Estonian, Finnish, German and Swedish OOCs, with a comprehensive validation toolbox designed to cover all available data with various metrics. It provides detailed outputs for expert users and model developers. However, for less experienced users and decision makers, the system provides a more general reliability output. Routines are adapted for mapped, on-track and time-series reference data covering the sea level, ice thickness and concentration, T, S, transports, CHL, oxygen, nitrate and phosphate metrics (Lagemaa et al. 2013). Moreover, the five contributing Baltic OOCs have organized a multi-model verification and comparison process, together with a multi-model ensemble estimate assessment. For some parameters, results are regularly posted to the Baltic Operational



The present web service dedicated to the IBI-MFC product quality dissemination has been developed in the frame of MyOcean and MyOcean2 projects

Figure 6. (a) Summary of comparisons computed by NARVAL (on daily, monthly, quarterly and yearly basis) to make an on-line and delayed mode validation of the Irish – Bay of Biscay – Iberian shelves (IBI) products. Abbreviations: IBI BE, IBI best estimates; IBI FCs, IBI forecast products at different forecast horizons; CLIM, climatologies; ATM, atmospheric fields used as IBI forcing. (b) NARVAL Delayed-Modes web pages: SST comparison of IBI fields with OOFSs from adjacent areas and the global forecasting system.

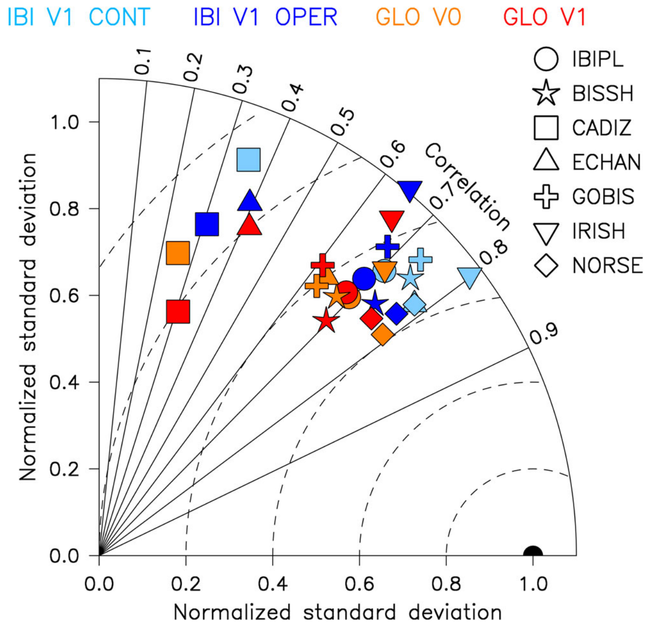


Figure 7. SST Class 4 metrics against MyOcean super collated SST_EUR_SST_L3S_NRT_OBSERVATIONS_010_009_a. With IBI free-run (cyan), IBI operational (blue), global old (yellow) and new version (red), for daily hindcast fields, in different domains defined in the text. Y-axis scale also corresponds to normalized standard deviation. 0.2 isocontours for RMSD (inner dashed circles) are associated with the normalized standard deviation.

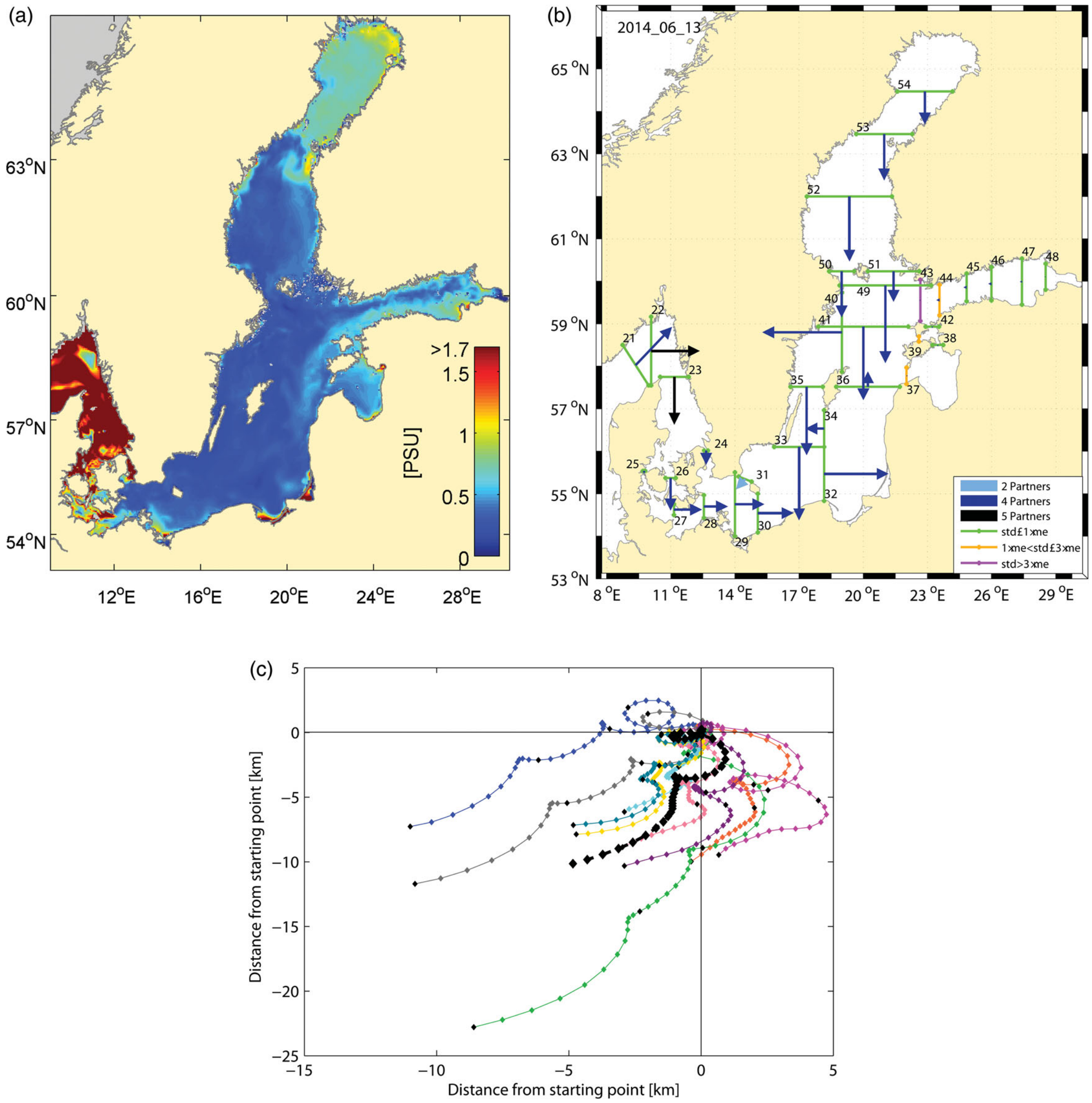


Figure 8. Example of the multi-model standard deviation for SSS (a). Multi-model mean total water transports in Sverdrup from 4 or 5 OOFS (blue/black arrows, numbers correspond to sections) for 13 June 2014. (b) Green transect indicating that the multi-model standard deviation transport is lower than the mean value, and yellow comprise between 1 and 3 times the mean value. (c) Progressive vector diagram of model sea surface currents from nine Baltic Sea forecast products. All trajectories starting from the same location on the Gulf of Finland. Similar figures are available and discussed at www.boos.org. Inga Golbeck, Xin Li and Frank Janssen (German Maritime and Hydrographic Agency), pers. comm.

Oceanographic System server (www.boos.org). Beyond the extended information on forecast scores, the intercomparison of different OOFS is adding value to the near real-time validation routines, in addition to the usual evaluation against observations and climatology. The multi-model standard deviations from different forecast products (Figure 8) provides valuable information about their uncertainties, which is difficult to assess using regular model-reference approaches owing to sparse coverage of observations. These figures are available daily at www.boos.org. Interestingly, nine OOFS are assembled to show the reliability of surface current forecast, using vector diagrams [Figure 8(c)]. This strategy has also been adopted in MyOcean by the North West European Shelf Operational Oceanographic System, covering the North Sea and English Channel regions, presenting a multi-model assessment from Belgium, Denmark, Germany, Norway, Sweden and UK OOCs (www.noos.cc).

Reducing uncertainties by ensemble approach: ocean surface parameters multi-model estimation

Major incidents, such as the AF447 Air France Rio-Paris airplane crash in June 2009, the DeepWater Horizon oil

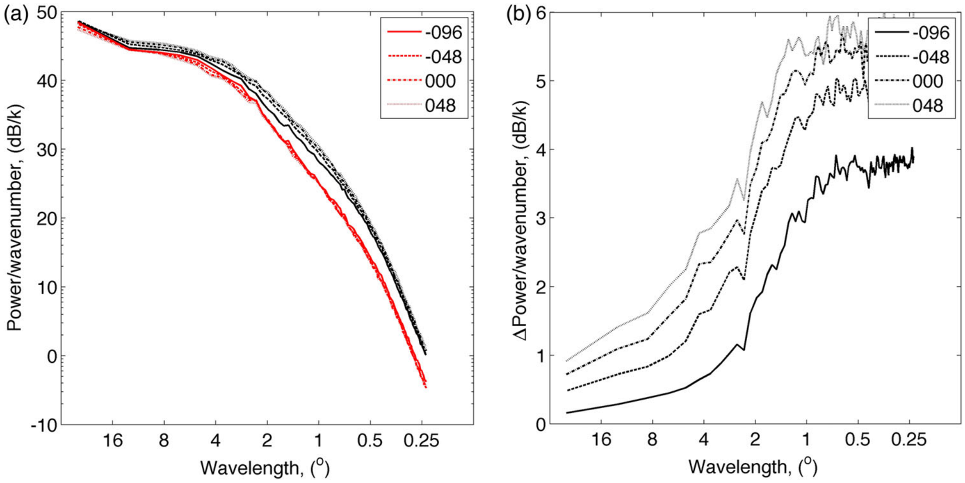


Figure 9. (a) Power spectrum for SST anomaly in the Tasman Sea for zonal sections (38–32S) and temporally averaged from 1 March to 31 August 2012 from the Australian OceanMAPS OOFS. The black (red) lines represent the 0-lag latest forecast and the ensemble mean, respectively. The periodograms are shown for the forecast hours −096 (4-day before) solid), −048 (2-day before, dashed), 00000 (dash-dot) and 048 (2-day forecast, dotted). (b) Difference in power between the 0-lag latest forecast and the weighted ensemble mean for the forecast −096 (solid), −048 (dashed), 000 (dash-dot) and 048 (dotted).

platform accident in April 2010, the Fukushima nuclear power plant catastrophe in March 2011 or the search for the missing MH370 Malaysia Airlines flight in March 2014 in the Indian Ocean, have highlighted operational oceanography's capacity to provide relevant information for decision makers (Masumoto et al. 2012; Kawamura et al. 2011). For all these events, national authorities have made requests to their respective OOCs in order to provide some assistance in near-real time or offline, to carry out dedicated studies to complement risk assessment.

Ocean studies performed in support of the effort to find the Air France plane wreckage relied on several new aspects: (1) an international effort to collect forecasts from different OOCs and to provide different ocean datasets to assist rescue activities in real-time; (2) the use of multi-model datasets and ensemble approaches to reduce errors of ocean surface dynamics in hindcasts and forecasts, with the implementation of dedicated high resolution model simulations in the area, nested into global OOFSs; (3) a retrospective statistical analysis of the accuracy of ocean currents and, in particular, the reliability of mixing and transport properties; (4) the formation of an international task team, with contributions from many ocean experts from both the *in situ* and modelling communities (Scott et al. 2012; Drévillon et al. 2013). Performance gains were also made during the search for MH370 through the use of ensemble mean products that improved the representation of buoy trajectories.

At the Australian Bureau Of Meteorology, deterministic forecast errors of the OceanMAPS OOFS are assessed and reduced by implementing time-lagged ensemble forecast, also called a multicycle ensemble (Brassington 2013). Over four successive days, forecasts are performed each day, starting from background fields independent from each other. Weighted ensemble averages are then computed, and forecast errors are assessed using spectral methods that quantify the impact of ensemble averaging as a function of wavenumber. For instance, for SST, Figure 9 demonstrates the increase in power for random information (Brassington 2013). By comparing the power spectrum at different forecast periods, the growth in random error relative to wavelength is also captured.

For marine pollution in the Northern Aegean Sea, studies based on a 48-h oil-spill dispersion forecast have been performed recently. The system is based on atmospheric, wave and ocean circulation models coupled with the operational systems using the Aegean-Levantine Eddy Resolving Model (nested in the MyOcean OOFS) and SKIRON of the University of Athens and oil-spill dispersion models (http://diavlos.oc.phys.uoa.gr). A Lagrangian-based verification east of the Limnos Island (Northeastern Aegean Sea) was conducted during October 2012 where 25 drifting buoys and special oil-spill drifting instruments were compared with drift predictions. The area was characterized by a very strong front, and in many cases a small error in the prediction of the frontal line resulted in very large errors in the oil-spill prediction (Figure 10, left). This experiment shows that forecasts beat persistence over the first 20 h (Figure 10, right). Moreover, in these areas of varying dynamical features (fronts, eddies), forecast errors grow significantly, emphasizing the need for more advanced prediction systems such as ensemble forecasts. Ensemble approach are considered now at regional scale, as in the Ligurian Sea, where a multi-model strategy is tested against an

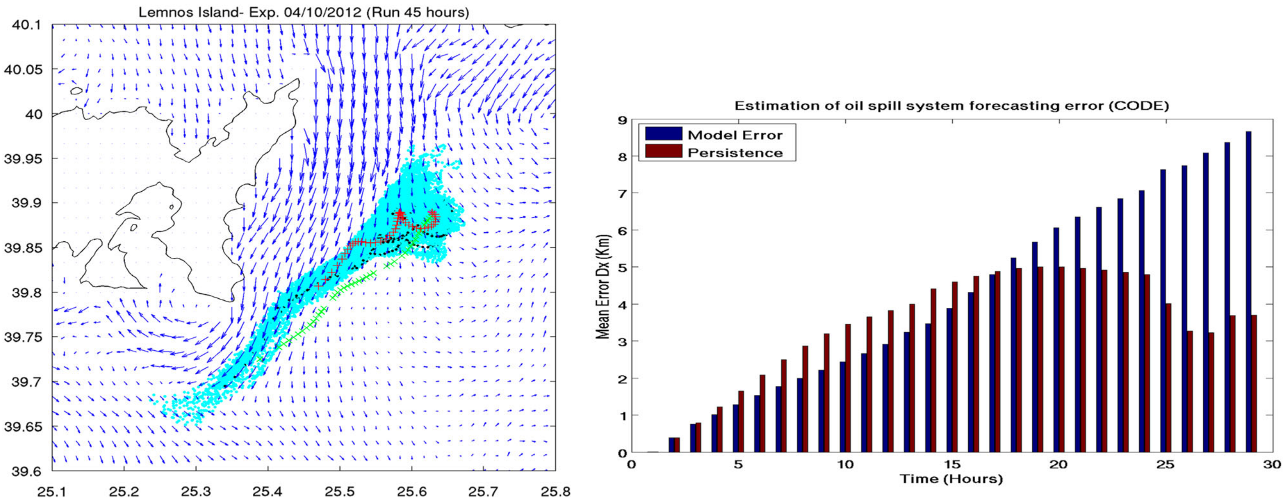


Figure 10. Left: experimental area, east of the Limnos Island; surface velocities given by the model (blue arrows). Simulated oil spills (cyan and black), centre of mass of the oil spills (red x) and drifter tracks (green x) are plotted. Right: time evolution of the oil-spill forecasting error (in kilometres, blue bars), derived as the distance between the drifter location and the centre of mass of the predicted oil-spill, compared with the persistence of the centre of mass (also as the difference with drifters in km, red bars).

ensemble predicting system, showing the respective merit of each approach (Mourre & Chiggiato 2014). Additionally, ensemble approaches proposed by the operational SST community at global scales have been shown to provide promising results, where the ensemble is usually more reliable than individual estimates (Martin et al. 2012; Dash et al. 2012; Xie et al. 2008).

Several OOFSs involved in GOV activities contributed to the rescue actions carried out for the dramatic events mentioned above. The IV-TT proposed to strengthen this multi-model approach, by organizing the real-time provision of operational hindcasts and forecasts among several GOV OOFSs. Recent experiences have shown that (1) surface ocean parameters were the most needed products, and (2) higher resolution improved the estimation (e.g. for drift, dispersion, mixing, sinking, etc.). As a result, since 2013, four global OOCs (US Navy with HYCOM-NCODA, UK Met Office with the Forecast Ocean Assimilation Model (FOAM), NOAA/NCEP with the Real-Time Ocean Forecast System (RTOFS) and the Climate Forecast System (CFS), and Mercator Océan with PSY3) have been providing a daily rolling archive of model native grid fields of best estimates and forecasts of T, S and currents at the surface. From these nowcasts/forecasts, a first multiple/ensemble assessment has been made, focusing on SST, together with two observation-only datasets chosen as reference: the NCEP Real-Time Global (RTG) (Thiébaux et al. 2003) and the GHRSST NAVO K10 level-4 (Martin et al. 2012) dataset. Three ensemble computations are defined using daily OOFS outputs: (1) simple arithmetic mean average; (2) weighted average, based on root-mean-square (RMS) daily differences of each member with respect to the reference SST field; and (3) clustered average, based on a k-mean algorithm (Hartigan & Wong 1979). A first hindcast comparison has now been performed for July 2013 compared with NCEP RTG (Figure 11). In this evaluation, Mercator_PSY3, FOAM and CFS perform better than the two other OOFSs. Note also that FOAM biases are slightly different using the NAVO K10 product (not shown). This highlights the sensitivity to uncertainty in the observational dataset used as 'truth'. Above all, Figure 11 shows that the use of an ensemble results in an improvement over each of the members, with the k-means clustered average performing the best of the three ensemble methods. Similar ensemble scores are obtained against the NAVO K10 SST (not shown). RMS scores seem dependent on the number of clusters: preliminary tests (not shown) from one to 10 clusters indicate significant improvements. This assessment is ongoing, with further analysis of the ensemble mean computation for forecasts and for other ocean parameters. One of the key aspects of this community effort is the real-time provision of these OOFS outputs.

Forecast skill: intercomparison of ocean parameters against observations: Class 4 metrics assessment

The Class 4 metrics approach, developed during the EU MERSEA Strand1 project (Crosnier & Le Provost 2007) and improved during the EU MERSEA-Integrated Project, was adopted at the international level by the GODAE community (Hernandez et al. 2009). This approach is based on comparison with reference measurements, from space or *in situ*, to assess the OOFS forecasting skill. Reference data, providing ocean 'truth', are used in a similar way, to infer the accuracy of both the best estimate/analyses and the forecasts at different lead times. Additionally, to evaluate the added value provided by the model and the OOFSs'

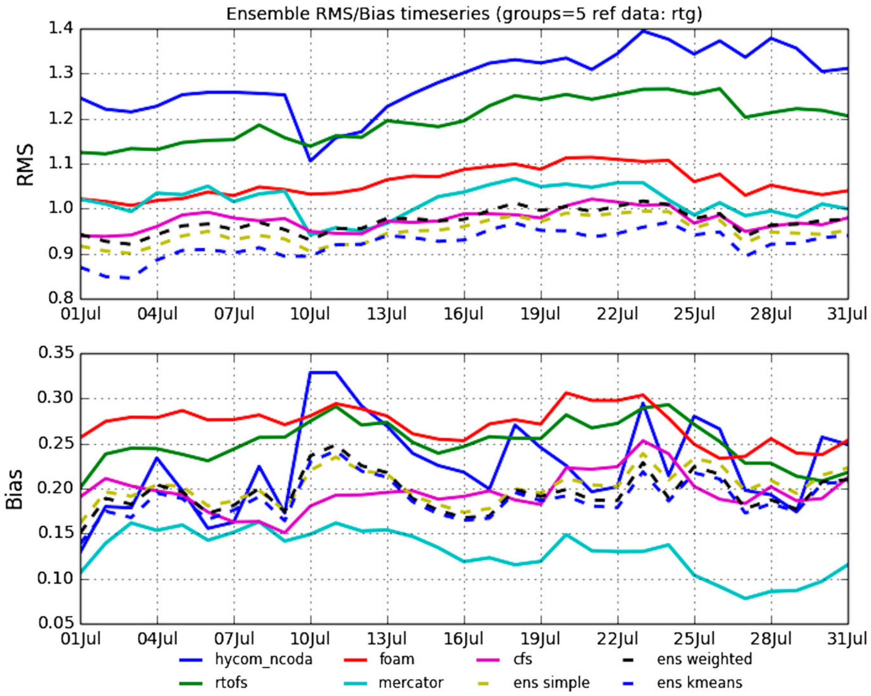


Figure 11. Global SST hindcast comparison statistics with RTG SST, computed daily, in July 2013. RMS (top) and biases (bottom) time series for HYCOM-NCODA (blue), RTOFS (green), FOAM (red), Mercator_PSY3 (cyan), CFS (purple) and ensemble averages: simple (dashed yellow), weighted (dashed black), clustered (dashed blue). Units in Kelvin.

short-term prediction efficiency, tests of the skill with respect to climatological fields and persistence are made.

Class 4 metrics performed in near-real-time are limited by the availability and quality of observations with several important consequences. First, owing to the scarcity of ocean measurements, the real-time assessment relies on observations that are also used by the assimilation system. These observations can be considered approximately independent (neglecting the autocorrelation of observation error in time) for forecast assessment though – in particular when considering short time-scale ocean transient features. A second consequence is the larger error budget for real-time observations that are not fully cross-calibrated, verified and corrected as in the delayed mode (Cabanes et al. 2013; Le Borgne et al. 2012). A third consequence relates to the overall quality of reference product in near-real time. If some biases exist between these products and information of the same kind that is assimilated, then validation scores can be impacted or wrong.

Assessment using Class 4 metrics can be distinguished from assimilation diagnostics in several ways. By assimilation diagnostics, we refer to statistics in observation space derived from the 'background minus observation' (i.e. the so-called misfits); from the 'increments' (i.e. the correction applied to the background to obtain the analysis); or from the 'analysis residual' (i.e. analysis minus observations). For most assimilation systems, there is a pre-processing of Global DAC (GDAC) data through editing, filtering or thinning, in order to limit the number of assimilation observations. This is done (1) because of a limit in the maximum number of observations that can be used by the assimilation scheme (computational requirements), (2) because some observations are considered a priori redundant (i.e. thinning provides a means to avoid having to consider correlated observation error) or (3) because the observations include features and dynamical processes and scales not represented by the forecasting system. 'Super-obing' can also be used in this assimilation pre-processing. In any case, the associated assimilation statistics and metrics often result in a net reduction in the number of observations considered. This is obviously not the purpose of the Class 4 metrics: ideally, all 'good' data from the GDAC can be compared with the OOFS fields and, by the way, measure the accuracy, forecasting skill and scales not represented by the OOFS. Thus, this approach is not OOFS dependent (i.e. the way observations are assimilated, the ocean model gridding, etc.). The same observation can be used in the evaluation of several OOFSs. That is, provided the reference data are independent, exact inter-comparison is possible, considering a specific ocean process or parameter.

The 'Class 4' strategy is now in place in several OOFSs, for global (Lellouche et al. 2013; Blockley et al. 2012) or regional assessment (Maraldi et al. 2013). In the framework of the GOV IV-TT, the Class 4 metrics project aims to stimulate the inter-comparison of OOFSs by verifying different aspects of ocean processes captured by the available observations in real time. A near-real-time intercomparison activity has been ongoing since January 2013. Five OOCs are involved, and six OOFSs are compared, looking at SST, T/S at depth and SSH – with sea ice concentration in preparation. A companion paper (Ryan et al. 2015) presents the global inter-comparison performed over basin scale areas. This exercise also allows each of the partners to assess more carefully the forecast capability of every OOFS in the region of interest and measure the efficiency of each system. The Australian group has performed regionally this multi-system assessment presented in a second companion paper (Divakaran et al. 2015).

Based on the statistics of the comparison with the same observations, this Class 4 assessment allows the following questions to be addressed:

- What is the relative reliability of each system for a given parameter in near-real-time?
- What is the performance of each system in forecast mode (5 days ahead)?
- What is the added value of the system compared with climatology or persistence?
- What benefits could be obtained through an ensemble approach, compared with each individual system?

As part of the Class 4 intercomparison, interesting new metrics and ways of representing the information graphically have been proposed to better synthesize the information. For instance, radar charts provide the score of each system at different forecast lead times, for all parameters evaluated (Figure 12). Note that the terminology 'hindcast' is used here for analysis, nowcast, hindcast or 'best estimate'. Owing to the details of their real-time operational assimilation scheme, every centre is providing what it considers as its 'best field' in near-real-time with minimum delay. Scores are defined by RMSE, based on differences between observation and model values for each parameter normalized by the largest RMSE. Reference observations are fully described in the companion paper (Ryan et al. 2015). Using this approach, one can characterize the relative score of each OOFS for each parameter. Missing parameters are not problematic: SLA is not evaluated for RTOFS (Ryan et al. 2015), and the radar chart can still be used for plotting scores from the other four systems. Moreover, this approach allows us to assess specific features, such as resolution, by comparing the global eddy permitting (PSY3, ¼°) and eddy-resolving (PSY4, 1/12°) Mercator Océan OOFS. The radar charts indicate that PSY3 skill scores are always better than PSY4 scores, even for SLA. It is worth noting that PSY3 and PSY4 are run in parallel every day at Mercator Océan (Lellouche et al. 2013), using the same forcing fields and assimilating the same set of observations. In this case, one may ask whether the observations used to derive the Class 4 metrics are capable of assessing the eddy-resolving capability of the PSY4 system, and if this Class 4 metric is able to infer the mesoscale predictive capabilities of these global high-resolution systems. In this case, SLA assessment could be performed using along-track satellite observations filtered differently, in order to capture more mesoscale features. Similarly, model SST could be compared with the highest resolution and most reliable SST products provided in near-real-time by DACs.

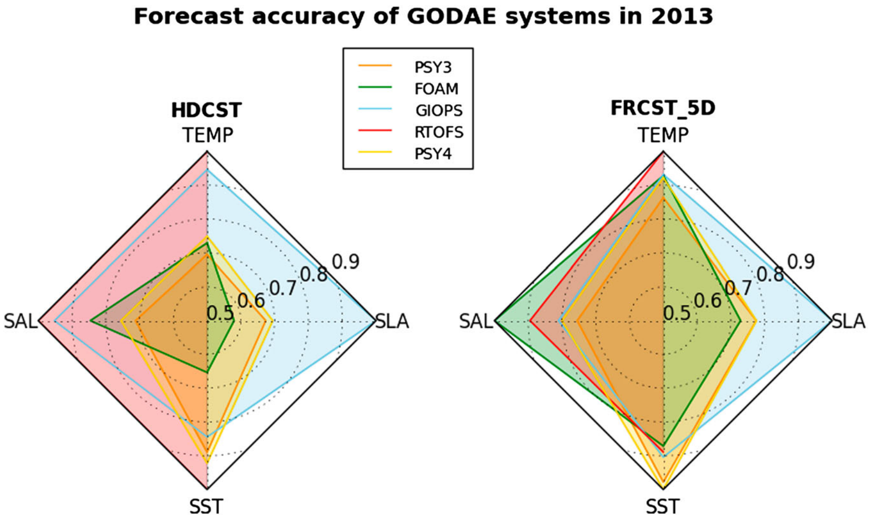


Figure 12. Class 4 global assessments over 2013. Radar charts for hindcast (HDCST), and 5-day forecasts (FRCST_5D). Four parameter evaluations are displayed: 5–100 m depth temperature (TEMP), and salinity (SAL), then SST, and SLA. Each score (between 0 and 1) is normalized by the largest RMSE value among the five evaluated OOFSs.

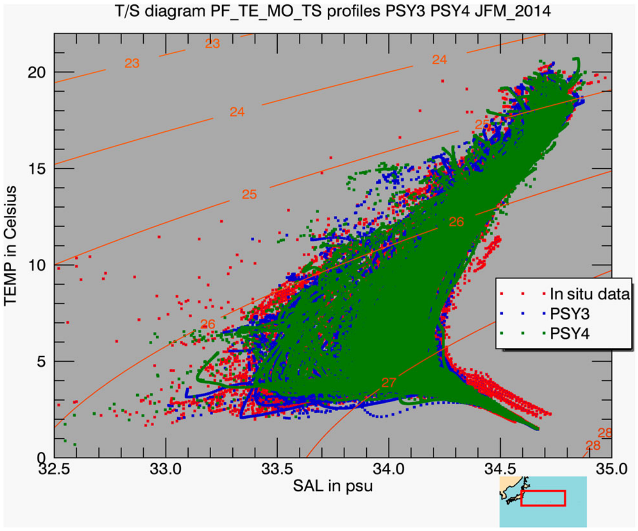


Figure 13. T–S diagram using Class 4 metrics applied on PSY3 (blue) and PSY4 (green) Mercator Océan OOFSs. Hindcast fields for the January–March 2014 period over the Kuroshio Extension area. The observed values of temperature and salinity are plotted in red.

Class 4 metrics over a given period and space domain are also presented through lead-time summary plots (see all figures of Ryan et al. 2015, and figure 7 from Divakaran et al. 2015), or Taylor diagrams. Note that for the Australian regional seas assessment (Divakaran et al. 2015), this diagram (figure 8 from Divakaran et al. 2015) also contains shaded values of the skill score, as defined by Taylor (2001). This score merges the accuracy (RMSE) and the pattern (correlation) evaluation of the parameter variability. Note also that this assessment of vertical parameters (here T, S) is presented separately for biases (consistency assessment), RMSE (quality, or accuracy assessment) and anomaly correlation (pattern of the variability). This allows OOCs to measure at which depth, and for which water masses, the OOFS is reliable. At this stage, Class 4 metrics are univariate, but alternatively, these metrics can be used in more ‘ocean oriented’ figures such as T–S diagrams (Figure 13). For this T–S diagram only hindcasts (i.e. not forecasts, persistence or climatology) are plotted together with the observed values, for the sake of clarity. Figure 13 shows that both PSY3 and PSY4 systems present inaccurate dense waters at depth, while the rest of the water column is qualitatively well represented for this 3-month period.

Inter-comparison of several OOFSs using Class 4 metrics also allows OOCs to address the added value that using an ensemble approach might bring. Ryan et al. (2015) show that the ensemble mean outperforms individual OOFS scores in most cases. Interestingly, in their Figure 6, they propose a synthetic global view of the most reliable OOFS for the four parameters tested. Other new approaches mentioned above (IBI OOFS) involve inter-comparing the same diagnostic for several OOFSs in order to show the added value of regional versus global, free model simulation versus assimilation or old versus new system estimates (Figure 7).

Summary

Significant progress has been made in ocean-model skill assessment during the last 5–10 years. Under the constraints of real-time operation, many forecasting centres have implemented more mature validation and performance-assessment procedures. The most advanced examples are operationally integrated, modular and able to use any available reference dataset. Based on a large number of metrics, they permit a diverse validation strategy: (1) comparing old and new systems to measure potential improvements and degradations; (2) comparing coarse resolution ‘father’ and nested high-resolution ‘son’ systems to quantify the added value of downscaling; (3) comparing adjacent or overlapping systems to verify the consistency of adjacent forecasts; (4) multi-model comparison to better characterize model error growth using different systems running in parallel; and (5) ensemble approaches to assess the benefit of ensemble versus individual system estimates.

Real-time assessments suffer from limitations imposed owing to observation availability and quality, as many high-quality reference datasets can only be used off-line – meaning that the routine monitoring skill evaluation is less efficient. To avoid spurious effects from erroneous real-time data (for assimilation or validation), quality checking and control of input information (observations, forcing fields) is performed by most OOFS. Moreover, the systematic feedback of quality control information and observation ‘blacklists’ to providers is starting to be integrated into OOFSs.

More complex metrics that are better suited to assessing physical, ecosystem and biogeochemical forecast processes are being progressively adopted in operational centres. Multivariate metrics now complement univariate techniques in order to enhance the parameters-oriented assessments to full ocean process evaluations. In parallel, metrics including Taylor diagrams, target diagrams and radar charts are used to provide a more enhanced quantification of model skill. Additional user-oriented metrics are also being developed, complementing the basic assessment of OOFSs with more detailed information about skill for specific applications.

Operational ocean forecasting systems are evolving toward higher horizontal resolution and eddy-resolving capability, and offer finer mesoscale representation. For instance, AVISO SSH or Reynolds SST L4 mapped products offer 50–100-km resolution. Hence, these products are no longer suitable for evaluating 5-km-resolution global eddy-permitting OOFS. For regional and coastal OOFSs providing sub-mesoscale description, this issue is even more crucial. Their evaluation using the existing observing system presents new issues: are the metrics currently used reliable, and do they provide pertinent information?

The L4 observation-based products provided by operational DAC and their evaluation also have to be considered carefully. First, these products can be used directly by the scientific community or other users instead of model-based products. Second, many OOFS validation procedures rely on these products and can be deficient if they are erroneous.

Finally, multi-model inter-comparison and ensemble approaches offer several potential benefits. For example, forecast spread can be used for forecast error evaluation and is particularly efficient if individual model errors are not correlated (e.g. for models using different forcing). In many studies, ensemble estimates are seen to benefit from qualities of each individual OOFS and to reduce errors. With the initiatives carried out by the GOV IV-TT, operational oceanography is following a strategic path similar to that of the weather-forecast community 30 years ago, the goal being to routinely exchange information among OOFS in a multi-model framework, and enhance both system predictability and skill assessments, for the eventual benefit of OOFS users.

Acknowledgements

During the final preparation of this article, our co-author Nicolas Ferry passed away. He was an active member of GOV and MyOcean, and was a significant innovator in the validation activity at Mercator Océan, mentioned in the present work. We will greatly miss the scientific expert, the gentle colleague and the good friend.

We thank the reviewers for their helpful comments and assistance in improving the clarity of this synthesis article. This work was supported by the European Commission funded projects MyOcean (FP7-SPACE-2007-1) and MyOcean2 (FP7-SPACE-2011-1).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Allen JI, Somerfield PJ. 2009. A multivariate approach to model skill assessment. J Marine Syst. 76(1–2):83–94. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.009

Balmaseda MA, Hernandez F, Storto A, Palmer MD, Alves O, Shi L, Smith GC, Toyoda T, Valdivieso M, Barnier B, Behringer D, Boyer T, Chang Y-S, Chepurin GA, Ferry N, Forget G, Fujii Y, Good S, Guinehut S, Haines K, Ishikawa Y, Keeley S, Köhl A, Lee T, Martin MJ, Masina S, Masuda S, Meyssignac B, Mogensen K, Parent L, Peterson KA, Tang YM, Yin Y, Vernieres G, Wang X, Waters J, Wedd R, Wang O, Xue Y, Chevallier M, Lemieux J-F, Dupont F, Kuragano T, Kamachi M, Awaji T, Caltabiano A, Wilmer-Becker K, Gaillard F. 2014. The Ocean Reanalyses Intercomparison project (ORA-IP). J Oper Oceanogr. 8(S1):s80–s97.

Balmaseda MA, Mogensen K, Weaver AT. 2013. Evaluation of the ECMWF ocean reanalysis system ORAS4. Q J Roy Meteor Soc. 139(674):1132–1161. doi:10.1002/qj.2063

Bell MJ, Lefebvre M, Le Traon P-Y, Smith N, Wilmer-Becker K. 2009. GODAE, The global ocean data experiment. Oceanogr Magazine. 22(3):14–21. doi:http://dx.doi.org/10.5670/oceanog.2009.62

Benestad RE, Senan R, Balmaseda MA, Ferranti L, Orsolini Y, Melsom A. 2011. Sensitivity of summer 2-m temperature to sea ice conditions. Tellus A. 63(2):324–337. doi:10.1111/j.1600-0870.2010.00488.x

Blockley EW, Martin MJ, Hyder P. 2012. Validation of FOAM near-surface ocean current forecasts using Lagrangian drifting buoys. Ocean Sci. 8(4):551–565. doi:10.5194/os-8–551–2012

Blockley EW, Martin MJ, McLaren AJ, Ryan AG, Waters J, Lea DJ, Mirouze I, Peterson KA, Sellar A, Storkey D. 2014. Recent development of the Met Office operational ocean forecasting system: An overview and assessment of the new global FOAM forecasts. Geosci Model Dev. 7(6):2613–2638. doi:10.5194/gmd-7-2613-2014

Brassington GB. 2013. Multicycle ensemble forecasting of sea surface temperature. Geophys Res Lett. 40 (23):2013GL057752. doi:10.1002/2013GL057752

Cabanes C, Grouazel A, von Schuckmann K, Hamon M, Turpin V, Coatanoan C, Paris F, Guinehut S, Boone C, Ferry N, de Boyer Montégut C, Carval T, Reverdin G, Pouliquen S, and Le Traon P-Y. 2013. *The CORA dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements*. Ocean Sci. 9(1):1–18. doi:10.5194/os-9–1–2013

Crosnier L, Le Provost C. 2007. Inter-comparing five forecast operational systems in the North Atlantic and mediterranean basins: The MERSEA-strand1 methodology. J Marine Syst. 65(1–4):354–375. doi:10.1016/j.jmarsys.2005.01.003

Dash P, Ignatov A, Martin M, Donlon CJ, Brasnett B, Reynolds RW, Banzon V, Beggs H, Cayula J-F, Chao Y, Grumbine R, Maturi E, Harris A, Mittaz J, Sapper J, Chin TM, Vazquez-Cuervo J, Armstrong EM, Gentemann CL, Cummings JA, Piollié J-F, Autret E, Roberts-Jones J, Ishizaki S, Høyer JL, and Poulter D. 2012. Group for high resolution sea surface temperature (GHRSST) analysis fields inter-comparisons, Part 2: Near real time web-based level 4 SST quality monitor (L4-SQUAM). Deep Sea Res Part II. 77–80:31–43. doi:http://dx.doi.org/10.1016/j.dsr2.2012.04.002

Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. Q J Roy Meteor Soc. 131(613):3385–3396. doi:10.1256/qj.05.108

Desroziers G, Ivanov S. 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. Q J Roy Meteor Soc. 127(574):1433–1452. doi:10.1002/qj.49712757417

Divakaran P, Brassington GB, Ryan AG, Regnier C, Spindler T, Mehra A, Hernandez F, Smith GC, Liu YY, Davidson F. 2015. GODAE OceanView Class 4 inter-comparison for the Australian region. J Oper Oceanogr. 8(S1):s112–s126.

Doney SC, Lima I, Moore JK, Lindsay K, Behrenfeld MJ, Westberry TK, Mahowald N, Glover DM, Takahashi T. 2009. Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data. J Marine Syst. 76(1–2):95–112. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.015

Drévillon M, Greiner E, Paradis D, Payan C, Lellouche J-M, Reffray G, Durand E, Law-Chune S, Cailleau S. 2013. A strategy for producing refined currents in the equatorial Atlantic in the context of the search of the AF447 wreckage. Ocean Dynam. 63(1):63–82. doi:10.1007/s10236–012–0580–2

Feng H, Vandemark D, Quilfen Y, Chapron B, Beckley B. 2006. Assessment of wind-forcing impact on a global wind-wave model using the TOPEX altimeter. Ocean Eng. 33(11–12):1431–1461. doi:http://dx.doi.org/10.1016/j.oceaneng.2005.10.015

Fujii Y, Kamachi M. 2003. Three-dimensional analysis of temperature and salinity in the equatorial Pacific using a variational method with vertical coupled temperature-salinity empirical orthogonal function modes. J Geophys Res. 108(C9):3297. doi:10.1029/2002JC001745

Gregg WW, Friedrichs MAM, Robinson AR, Rose KA, Schlitzer R, Thompson KR, Doney SC. 2009. Skill assessment in ocean biological data assimilation. J Marine Syst. 76(1–2):16–33. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.006

Hartigan JA, Wong MA. 1979. Algorithm AS 136: A K-means clustering algorithm. J Roy Stat Soc C. 28(1):100–108. doi:10.2307/2346830

Hernandez F, Bertino L, Brassington GB, Chassignet EP, Cummings JA, Davidson F, Drévillon M, Garric G, Kamachi M, Lellouche J.-M., Mahdon R, Martin MJ, Ratsimandresy A, Regnier C. 2009. Validation and intercomparison studies within GODAE. Oceanogr Magazine. 22 (3):128–143. doi:http://dx.doi.org/10.5670/oceanog.2009.71

Hernandez F. 2011. Performance of ocean forecasting systems – Intercomparison projects. In: Schiller A, Brassington GB, editor. Operational oceanography in the 21st century. Springer Science+Business Media B. V.; p. 633–655. doi:10.1007/978-94-007-0332-2_23

Hyder P, Storkey D, Blockley EW, Guiavarc'h C, Siddorn J, Martin M, Lea D. 2012. Assessing equatorial surface currents in the FOAM Global and Indian Ocean models against observations from the global tropical moored buoy array. J Oper Oceanogr. 5(2):25–39.

Jolliff JK, Kindle JC, Shulman I, Penta B, Friedrichs MAM, Helber R, Arnone RA. 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. J Marine Syst. 76(1–2):64–82. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.014

Kawamura H, Kobayashi T, Furuno A, In T, Ishikawa Y, Nakayama T, Shima S, Awaji T. 2011. Preliminary numerical experiments on oceanic dispersion of 131I and 137Cs discharged into the ocean because of the Fukushima Daiichi nuclear power plant disaster. J Nucl Sci Technol. 48 (11):1349–1356. doi:10.1080/18811248.2011.9711826

Lagemaa P, Janssen F, Jandt S, Kalev K. 2013. Pers. Comm. General Validation Framework for the Baltic Sea in GODAE OceanView Symposium Poster. Baltimore, USA.

Lagemaa P. 2013. Pers. Comm. Ice model metrics and scores for the Baltic Sea in HIROMB-BOOS Scientific Workshop. oral presentation. Tallinn, Estonia.

Lazzari P, Solidoro C, Ibello V, Salon S, Teruzzi A, Béranger K, Colella S, Crise A. 2012. Seasonal and inter-annual variability of plankton chlorophyll and primary production in the Mediterranean sea: A modelling approach. Biogeosciences. 9(1):217–233. doi:10.5194/bg-9–217–2012

Le Borgne P, Marsouin A, Orain F, Roquet H. 2012. Operational sea surface temperature bias adjustment using AATSR data. Remote Sens Environ. 116(0):93–106. doi:http://dx.doi.org/10.1016/j.rse.2010.02.023

Lellouche J-M, Le Galloudec O, Drévillon M, Régnier C, Greiner E, Garric G, Ferry N, Desportes C, Testut C-E, Bricaud C, Bourdallé-Badie R, Tranchant B, Benkiran M, Drillet Y, Daudin A, De Nicola C. 2013. Evaluation of global monitoring and forecasting systems at Mercator Océan. Ocean Science. 9(1):57–81. doi:10.5194/os-9-57-2013

Levy M, Martin AP. 2013. The influence of mesoscale and submesoscale heterogeneity on ocean biogeochemical reactions. Global Biogeochemical Cycles. 27(4):1139–1150. doi:10.1002/2012gb004518

Lorente P, Soto-Navarro J, Alvarez Fanjul E, Piedracoba S. 2014. Accuracy assessment of high frequency radar current measurements in the Strait of Gibraltar. J Oper Oceanogr. 7 (2):59–73.

Lynch DR, McGillicuddy Jr DJ, and Werner FE. 2009. Skill assessment for coupled biological/physical models of marine systems. J Marine Syst. 76(1–2):1–3. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.002

Maraldi C, Chanut J, Levier B, Ayoub N, De Mey P, Reffray G, Lyard FH, Cailleau S, Drévillon M, Alvarez Fanjul E, Garcia Sotillo M, Marsaleix P, the Mercator Research Development T. 2013. NEMO on the shelf: Assessment of the Iberia-Biscay-Ireland configuration. Ocean Science. 9 (4):745–771. doi:10.5194/os-9–745–2013

Martin M, Dash P, Ignatov A, Banzon V, Beggs H, Brasnett B, Cayula J-F, Cummings J, Donlon C, Gentemann C, Grumbine R, Ishizaki S, Maturi E, Reynolds RW, Roberts-Jones J. 2012. Group for high resolution sea surface temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE). Deep Sea Res Part II. 77–80:21–30. doi: http://dx.doi.org/10.1016/j.dsr2.2012.04.013

Martin M. 2011. Ocean forecasting systems: Product evaluation and skill. In: Schiller A, Brassington GB, editors. Operational oceanography in the 21st century. Springer Science+Business Media B. V.; p. 611–632. doi:10.1007/978-94-007-0332-2_22

Masumoto Y, Miyazawa Y, Tsumune D, Tsubono T, Kobayashi T, Kawamura H, Estournel C, Marsaleix P, Lanerolle L, Mehra A, Garraffo ZD. 2012. Oceanic dispersion simulations of 137Cs released from the Fukushima Daiichi nuclear power plant. Elements. 8(3):207–212. doi:10.2113/gselements.8.3.207

Melsom A, Counillon F, LaCasce J, Bertino L. 2012. Forecasting search areas using ensemble ocean circulation modeling. Ocean Dynam. 62(8):1245–1257. doi:10.1007/s10236-012-0561-5

Moore AM, Arango HG, Broquet G, Powell BS, Weaver AT, Zavala-Garay J. 2011. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I – System overview and formulation. Prog Oceanogr. 91(1):34–49. doi:http://dx.doi.org/10.1016/j.pocean.2011.05.004

Mourre B, Chiggiato J. 2014. A comparison of the performance of the 3-D super-ensemble and an ensemble Kalman filter for short-range regional ocean prediction. Tellus: Series A. 66:1–14. doi:10.3402/tellusa.v66.21640

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. Mon Weather Rev. 115(7):1330–1338. doi:10.1175/1520–0493(1987)115<1330:agfffv>2.0.co;2

Murphy AH. 1993. What is a good forecast – an essay on the nature of goodness in weather forecasting. Weather Forecast. 8(2):281–293. doi:10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2

Oke PR, Brassington GB, Cummings JA, Martin MJ, Hernandez F. 2012. GODAE inter-comparisons in the Tasman and Coral Seas. J Oper Oceanogr. 5(2):11–24.

Oke PR, Griffin DA, Schiller A, Matear RJ, Fiedler R, Mansbridge J, Lenton A, Cahill M, Chamberlain MA, Ridgway KR. 2013. Evaluation of a near-global eddy-resolving ocean model. Geosci Model Dev. 6(3):591–615. doi:10.5194/gmd-6-591-2013

Pearson K. 1904. Mathematical Contributions to the theory of Evolution. On the Theory of Contingency and Its Relation to Association and Normal Correlation. Research Memoirs Biometric Series I. Department of Applied Mathematics, University College, University of London. London, UK. http://ia600408.us.archive.org/18/items/cu31924003064833/cu31924003064833.pdf

Rose KA, Roth BM, Smith EP. 2009. Skill assessment of spatial maps for oceanographic modeling. J Marine Syst. 76 (1–2):34–48. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.05.013

Ryan AG, Regnier C, Divakaran P, Spindler T, Mehra A, Smith GC, Liu YY, Davidson F, Hernandez F, Maksymczuk J, Lui Y. 2015. GODAE OceanView Class 4 forecast verification framework: Global ocean inter-comparison. J Oper Oceanogr. 8(S1):s98-s111.

Schiller A, Bell M, Brassington G, Brasseur P, Barciela R, De Mey P, Dombrowsky E, Gehlen M, Hernandez F, Kourafalou V, Larnicol G, Le Traon P-Y, Martin M, Oke P, Smith GC, Smith N, Tolman H, Wilmer-Becker K. 2015. Synthesis of New scientific challenges for GODAE OceanView. J Oper Oceanogr doi: 10.1080/1755876X.2015.1049901.

Scott RB, Ferry N, Drévillon M, Barron CN, Jourdain NC, Lellouche J-M, Metzger EJ, Rio M-H, Smedstad OM. 2012. Estimates of surface drifter trajectories in the equatorial Atlantic: A multi-model ensemble approach. Ocean Dynam. 62(7):1091–1109. doi:10.1007/s10236–012–0548–2

Stow CA, Jolliff J, McGillicuddy Jr DJ, Doney SC, Allen JI, Friedrichs MAM, Rose KA, Wallhead P. 2009. Skill assessment for coupled biological/physical models of marine systems. J Marine Syst. 76(1–2):4–15. doi:http://dx.doi.org/10.1016/j.jmarsys.2008.03.011

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. J Geophys Res. 106 (D7):7183–7192. doi:10.1029/2000JD900719

Thiébaux J, Rogers E, Wang W, Katz B. 2003. A new high-resolution blended real-time global sea surface temperature analysis. B Am Meteorol Soc. 84(5):645–656. doi:10.1175/BAMS-84–5–645

Usuii N, Ishizaki S, Fujii Y, Tsujino H, Yasuda T, Kamachi M. 2006. Meteorological research institute multivariate ocean variational estimation (MOVE) system: Some early results. Adv Space Res. 37(4):806–822. doi:10.1016/j.asr.2005.09.022

Volpe G, Colella S, Forneris V, Tronconi C, Santoleri R. 2012. The mediterranean ocean colour observing system – System development and product validation. Ocean Sci. 8(5):869–883. doi:10.5194/os-8–869–2012

Volpe G, Santoleri R, Vellucci V, Ribera d'Alcalà M, Marullo S, D'Ortenzio F. 2007. The colour of the Mediterranean Sea: Global versus regional bio-optical algorithms evaluation and implication for satellite chlorophyll estimates. Rem Sens Environ. 107(4):625–638. doi:http://dx.doi.org/10.1016/j.rse.2006.10.017

Wilkin JL, Hunter EJ. 2013. An assessment of the skill of real-time models of Mid-Atlantic Bight continental shelf circulation. J Geophys Res-Oceans. 118(6):2919–2933. doi:10.1002/jgrc.20223

Xie J, Zhu J, Li Y. 2008. Assessment and inter-comparison of five high-resolution sea surface temperature products in the shelf and coastal seas around China. Cont Shelf Res. 28(10–11):1286–1293. doi:http://dx.doi.org/10.1016/j.csr.2008.02.020