CHAPTER 29

Measuring Performances, Skill and Accuracy in Operational Oceanography: New Challenges and Approaches

Fabrice Hernandez[1], Greg Smith[2], Katrijn Baetens[3], Gianpiero Cossarini[4], Isabel Garcia-Hermosa[1], Marie Drévillon[1], Jan Maksymczuk[5], Angélique Melet[1], Charly Régnier[1], and Karina von Schuckmann[1]

[1]MetOcean Mercator Océan, Ramonville St Agne, France; [2]Environment Canada, Montréal, Canada; [3]Royal Belgian Institute of Natural Sciences, Brussels, Belgium; [4]National Institute of Oceanography and Experimental Geophysics, Sgonico, Italy; [5]Met Office, Exeter, UK

Operational oceanography is now established in many countries, focusing on global, regional, or coastal areas, and targeting different aspects of the « blue », « white » or « green » ocean processes in order to provide reliable information to users. There are nowadays a large variety of interests and users, with different disciplines and levels of expertise. Validation and verification of operational products and systems are evolving in order to anticipate user's needs, and better quantify the level of confidence on all these variety of ocean products. Operational oceanography evaluation development is in front of key issues: Ocean models are reaching the submesoscale description, which is currently not adequately observed; many products are available now for a given ocean variable, and often discrepancies are larger than similarities; real time forecasting systems are also challenged by reanalyses or reprocessed time series; operational systems are getting more complex, with coupled modelling, where errors from the different compartment need to be carefully addressed in order to measure their performance and provide further improvements. In parallel, the global ocean observing system is continuously completed with additional satellites in the constellation, with innovative sensors on new satellite missions, with efforts to better integrate the global, regional and coastal in-situ observing capabilities, and the design of new instrument, like the BGC-Argo that should bring an enhanced description of the ocean biogeochemical variability. This book chapter provides an overview of the existing, mature, validation and verification science in operational oceanography; discusses the ongoing efforts and new strategies; presents some of the structured groups and outcomes; and lists a series of challenges on the field.

Introduction

Operational oceanography has reached a mature stage. Now established in many countries, operational centres began by providing ocean products to a small, select group of experts. However, over the past several years, operational oceanography has expanded and now provides services and ocean monitoring to a wide community of users. Operational Ocean Forecasting and Monitoring Systems (OOFMS) can focus on global, regional, or coastal areas, and target different aspects of the « blue » physical ocean processes, « green » biological/biogeochemical low trophic level processes, or « white » sea ice and cryosphere

Hernandez, F., et al., 2018: Measuring performances, skill and accuracy in operational oceanography: New challenges and approaches. In "*New Frontiers in Operational Oceanography*", E. Chassignet, A. Pascual, J. Tintoré, and J. Verron, Eds., GODAE OceanView, 759-796, doi:10.17125/gov2018.ch29.

processes over ocean, in order to inform a large variety of interests and marine users within different disciplines and with varying levels of expertise (see, for example, Fig. 1 of Schiller et al., 2016).

Operational ocean global and regional initiatives expanded a great deal over the past two decades, working to overcome major issues under the auspices of community of experts such as those involved in the Global Ocean Data Assimilation Experiment (GODAE), which was followed by GODAE OceanView (Bell et al., 2015; Tonani et al., 2015). Operational oceanography is based on three pillars: 1) ocean observing systems; 2) modelling tools; and 3) data assimilation or other estimation and control techniques. These are structured to provide descriptions and predictions in the marine environment and offer dedicated services to marine stakeholders. The 2017 GODAE OceanView summer school addresses many of the advances and challenges that arise in these three areas, in particular the main limitations and weaknesses that directly impact contemporary performances of OOFMS. The reader is invited to look at all of these GODAE OceanView (GOV) summer school contributions where observing system limitations, model errors, and data assimilation performances are discussed as a way to begin to understand assessment and evaluation approaches designed and implemented by the operational oceanography community.

Although initially, operational oceanography developments were heavily science-driven (Schiller et al., 2015), they evolved and are now more user-driven (Schiller et al., 2016). This has broadened the prediction and monitoring capabilities in a seamless way closer to the coast (Kourafalou et al., 2015) in order to fulfil the UN sustainable development goals for the marine ecosystems and taking into account the huge needs of the permanently growing "blue economy."

From very local applied operational systems to global monitoring and forecasting ocean centres, the main goal of operational oceanography is to *provide timely and accurate information, including prediction and projections, about the marine environment*. Consequently, validation and verification of ocean numerical simulations and estimations are core activities in operational oceanography in order to anticipate user needs and better quantify the level of confidence on a variety of ocean products (Hernandez et al., 2015).

This chapter provides an overview of the validation and verification framework in operational oceanography. It also presents the standards and methods adopted by the GOV community since the beginning of GODAE. It then introduces different evaluation approaches and key issues based on the evaluation framework raised by the European Union (EU) Copernicus Marine Environment Monitoring Service (CMEMS) program. And finally, it presents some recent validation approaches and new metrics.

General Validation and Verification Background

Introductory considerations on evaluating performance of operational systems and quality of products

The ocean operational community is challenged in the way assessment is performed. Evaluation tools are now widely implemented in operational centres. But consider the overview proposed by

Hernandez et al. (2015), as a companion or introductory work to this chapter. Hernandez et al. (2015) reviewed the principles and main concepts driving evaluation approaches in operational oceanography, specifically the methodology raised by GODAE and GOV, with standardized methods such as Class 1, 2, 3, and 4 metrics, which are largely implemented now (e.g., Ryan et al., 2015). Hernandez et al. (2015) also detailed a series of recent metrics designed to assess specific variables (e.g., sea ice, chlorophyll, water masses) or focus on particular OOFMS assessment aspects (e.g., long-term forecast, assimilation performance, regional systems efficiency, upstream quality control, and/or ensemble assessment).

Most operational centres have implemented an assessment framework dedicated to:

- Evaluating and monitoring the performance of operational systems, considering the:
  - impact of the observing system,
  - model errors, and
  - data assimilation efficiency.
- Evaluating the accuracy of products such as the:
  - products derived from observation (real time or reprocessed),
  - routine hindcast and forecast (and their predictive skill), and
  - reanalyses.
- Measuring the strengths and weaknesses of the system operated in order to make further improvements
- Assessing each product's reliability in light of user needs

It is important to keep in mind the way that these errors and existing observations that represent the "ocean truth" are part of the "three pillars" of the operational oceanography structure mentioned above. Fig. 29.1 describes the main errors associated with each product and notes some of the difficulties in using observations for evaluation of an ocean product quality. It becomes evident that, for the complete OOFMS, many factors and errors limit a product's quality. In particular, observation distribution in space and time as well as sparseness strongly impact our capacity to assess how efficient operational systems are in representing ocean processes. An effectual validation and verification framework should take into account these factors and characterize the contribution of each error type.

Finally, considering user needs is a new aspect in the operational oceanography evaluation framework that has required consideration of several "internal" and "external" metrics. These include:

***Internal metrics*: verification that the systems satisfy initial requirements**. In other words, verify that a system reacts and behaves as expected with regard to its own representativity. For example, let's look at evaluation of the eddy-permitting system representation of the Gulf Stream path and the associated gradient. It would not be appropriate to evaluate the capability of this system in producing correctly tidal fronts, since tidal dynamics are not part of the "ocean model engine" used for this system. But it would be appropriate to measure the Gulf Stream frontal position against near real time satellite sea surface temperature images which indicates its surface thermal signature.

Other appropriate examples include position and seasonal variability of the subtropical gyres, which should be verified in a 1,000-year coupled climate simulation, while M2 harmonic phase and amplitude should be assessed in a barotropic tidal model.

***External metrics*: reliability of the product based on user requirements**. In many cases, external metrics look to measure the departures of the products against real ocean processes, whatever the representativity of the ocean model. So, for instance, if a user is looking at the extension of harmful algae blooms, proper assessment of the hourly rate of nutrients, oxygen, turbidity, mixed layer dynamics, fresh water run-off on the shelf is needed. Obviously, a basin scale eddy-permitting biogeochemical system would not be suitable for this purpose; a high-resolution regional model with high-frequency forcings and the relevant local ecosystem dynamics would offer a user better predictions in a case such as this.

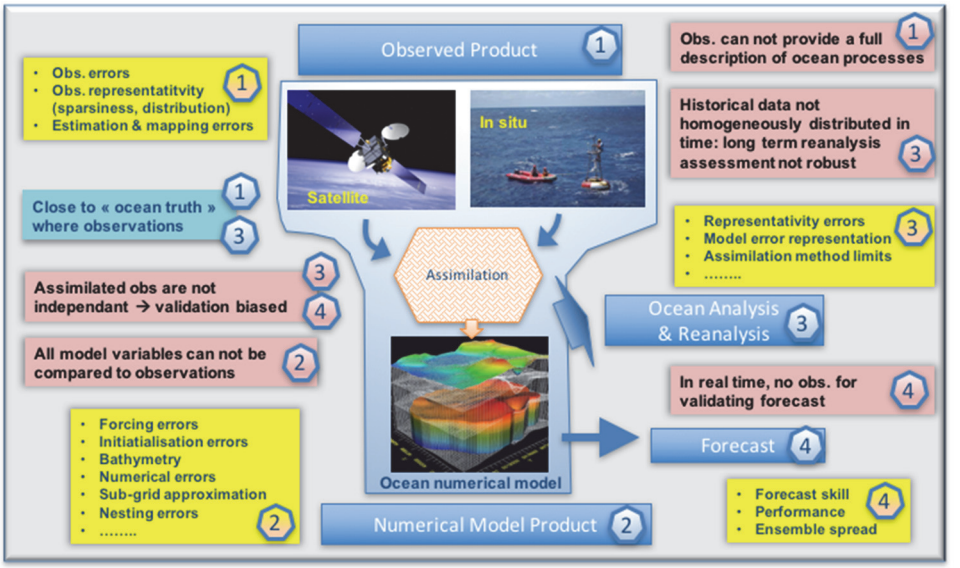


Figure 29.1. Different operational oceanography products (blue), their associated errors (yellow), positive aspects (light blue), and some drawbacks in using observations for the evaluation (pink).

The operational ocean assessment framework is essentially built upon comparison to observations. It is thus worth to note that a given set of observations might be used both for internal or external assessment. For instance, let's consider a metrics aiming at measuring sea level differences at high-frequency (e.g., a few minutes) between a tide gauge along the coast and model solutions. The low-pass filtering of the sea level time series would allow an "internal" assessment of any eddy-permitting system. This is what is carried out when GOV global systems are compared to the international tide gauge network (e.g., see Fig. 7 of Zuo et al., 2015). Now, let's consider the same tide gauge time series non-filtered on an hourly basis; comparing it to the same GOV global system would demonstrate the inability of this GOV system to represent coastal tidal harmonics. This would be an "external" evaluation of this global product for a user interested in high-frequency, on-shelf circulation. Meanwhile, if the tested system was a high-frequency, high-resolution regional shelf model that included tidal dynamics, then the comparison between hourly tide gauge data and

the model output would again be considered an "internal" evaluation, where one tries just to assess what such a model should represent (e.g., tidal assessment in the CMEMS IBI configuration by Maraldi et al., 2013). Finally, considering the complete non-filtered tide gauge time series, e.g., at five-minute frequency, and comparing the extreme sea level events with this regional high-resolution regional system would show the capability of this system in predicting and phasing extreme storm surges. This would be a valuable "external" assessment for any decision-maker in charge of coastal management and warnings.

Operational oceanography community effort on evaluation and verification

Through GOV, some in the international community work to address common OOFMS assessment challenges. The GOV Intercomparison and Validation Task Team, together with other task teams, is focusing on some of these issues and experimenting with several Class 1 and Class 4 metric approaches (Divakaran et al., 2015; Hernandez et al., 2015; Ryan et al., 2015). The GOV Coastal and Shelf Seas Task Team has proposed new approaches to the evaluation of regional operational systems (De Mey et al., 2017; Kourafalou et al., 2015), and highlighted coastal modelling, coastal observations, and nesting assessment issues. As part of the GOV 2017 summer school lectures, Mourre et al. (2018) for the Balearic Sea OOFMS and Roughan et al. (2018) for the New South Wales Australian coast integrated observing system, offered a comprehensive overview of regional assessment objectives, methods, and challenges of the operational ocean framework.

Complementary other assessment challenges have been tackled by GOV community efforts. Recently, the GOV Marine Ecosystem Analysis and Prediction Task Team began to address validation issues of the biogeochemical component of global or regional operational systems (Gehlen et al., 2015). In the GOV framework, task teams such as the Observing System Evaluation Task Team and the Coupled Prediction Task Team have also shared and rely upon evaluation approaches proposed by the other task teams (Brassington et al., 2015; Oke et al., 2015a; 2015b).

Moreover, some international initiatives such as the CLIVAR Global Synthesis and Observations Panel (GSOP) Ocean Reanalyses Intercomparison Project (http://www.clivar.org/panels-and-working-groups/gsop/gsop.php), along with EU COST Action Evaluation of Ocean Syntheses (EOS) project (http://www.eos-cost.eu/the-action/about-eos), allow for exploration of evaluation approaches for ocean reanalyses (Balmaseda et al., 2015) at both global and regional scales. The GOV 2017 summer school lecture by Haines (2018) proposed an updated overview of reanalyses development and assessment.

In parallel, the EU CMEMS is pioneering many aspects of global and operational oceanography development and issues, supported by the expertise and organization of most "marine" nations in Europe (Drévillon et al., 2018, this book, as part of GOV summer school contribution). The CMEMS developed a dedicated, cross-cutting activity on validation and verification activities, with a long-term strategy and plans to address the most recent issues on the validation, verification, and performance of OOFMS (Hernandez and Melet, 2016; Le Traon et al., 2017).

Evaluation in operational oceanography: some identified challenges

Above we provide some examples of the operational oceanography community's ongoing efforts to organize and develop adequate evaluation tools. Their main achievements are reviewed in Hernandez et al. (2015), and some challenging issues are listed in Schiller et al. (2015). However, the community is facing new emerging questions:

First, operational oceanography is continuously evolving toward more complex systems:

- Global and regional open ocean models are reaching the submesoscale description, typically less than 10 km (see, for example, discussion in the lecture by Jacobs et al., 2018), and the ocean dynamics at these fine scales are not adequately observed;
- Through nesting strategies, increasingly, local, coastal modelling tools are exchanging poorly controlled information with larger-scale systems at their boundaries;
- The diversity of products available for a given ocean interest/variable increases among operational ocean catalogues (e.g., gridded observed products from in situ or satellite measurements, or products merging both in situ and remote sensing data; model forecasts, model reanalyses from different kind of models; global or regional products etc.). Sometimes these discrepancies are larger than their similarities; assessment is needed to evaluate their quality and to inform users about their usefulness for a given application.
- Real-time forecasting systems are also challenged by reanalyses or reprocessed time series. All of these products need to be evaluated and their respective quality communicated to users. Then, real-time system performance must be revisited with regard to better quality reanalyses, with the goal being to motivate improvements.
- Operational systems are getting more complex with coupled modelling (e.g., Brassington et al., 2015; Harris, 2018; Tonani et al., 2015), where errors from the different compartments need to be carefully addressed in order to measure performance and provide further improvements, but also in order design ways to limit impacts of these errors from one compartment to another.

Second, operational oceanography, originally a product of academia, must now communicate more efficiently with a wide range of user communities and practices:

- Traditionally, for a given ocean domain, evaluation in research mode tried to assess the performance of the system as a whole. Now, in the same domain, different types of users and different types of applications (economic, security, health, biodiversity, regulation etc.) can have entirely different and dedicated assessment procedures carried out in parallel.
- Representativity in a given OOFMS needs to be taken into account and quantified in terms of "representativity errors" (also called "representation" or "representativeness" errors) when communicating a product's reliability: through a comprehensive assessment based on external metrics, one can inform users expecting the full

discrepancies and errors from what really happens at sea, and not providing confidence levels through few metrics related to the specific capabilities of the system.

- Informing users about product reliability also requires adopting new approaches: communicating error and accuracy levels must take into account the user's expertise, awareness, and their effective use of the products.

This chapter, associated with Hernandez et al. (2015), addresses validation, verification, and assessment concepts by structuring most of the evaluation methods and tools implemented in operational oceanography centres. For additional practical examples, the reader is invited to read other chapters in this book that describe specific OOFMS and their identified errors from different components (e.g., chapters by Bouillon et al., 2018; Ford et al., 2018; Harris, 2018; Jacobs et al., 2018; Le Sommer et al., 2018) as well as their validation framework (e.g., chapters by Lellouche et al., 2018; Mourre et al., 2018; Roughan et al., 2018; Wilkin et al., 2018).

The GODAE OceanView Common Evaluation Framework

Any ocean operational centre can design, implement, and perform evaluation of its forecasting tools on its own if worldwide ocean observations are easily accessible. However, when GODAE began, it was clear that the operational oceanography community could follow in the footsteps of the weather forecast and climate communities in the way that they work together under the patronage of the World Meteorological Organisation (WMO). For validation and verification activities, this allows for the 1) sharing of best practices and innovations; 2) inter-comparing performances of operational systems and evaluating the GODAE system against other systems; 3) using other estimates to improve the operational service; 4) responding consistently, particularly when requesting information, tools, and support from other parties (e.g., observations from space agencies, national marine institutions, other expert groups etc.); and 5) understanding common requests from users and applications. The ability to act in unison has allowed GODAE to establish and engage in legitimate dialogs with other communities. Over the past several years, the GOV Intercomparison and Validation Task Team has begun to exchange with expert groups, such as the Working Group on Numerical Experimentation from the WMO and World Climate Research Program, and the WMO Joint Working Group on Forecast Verification Research (https://www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html), whose expertise in weather forecast verification offers numerous valuable approaches (e.g., Casati et al., 2008; Ebert et al., 2013; Gilleland et al., 2009).

With regard to understanding common user requests and applications, one of the unfortunate illustrations are airplane accidents at sea. As discussed in Hernandez et al. (2015), the AF447 Air France Rio-Paris airplane crash in June 2009 led to a series of published improvements in ensemble estimation and assessment from OOFMS in order to better specify rescue and search activities at sea (e.g., Drévillon et al., 2013). This permitted several organizations to initiate their search activities after the crash of the Malaysia Airline MH370 plane in 2014 and to refine their techniques using model simulations and observations while debris and evidences were appearing (Griffin and

Oke, 2017; Griffin et al., 2016). Despite the fact that the crash location was not yet known, this work resulted in new insights into ways to deal with the reliability of ocean products and give confidence to the relevant users. Specific events such as the Malaysia Airline crash are valuable as case studies for evaluating the performance of incoming OOFMS. Similarly, the AF447 accident was used by Mercator Océan to evaluate the surface dispersion provided by the new high-resolution global system, and showed large improvements and skill in positioning search areas (see the chapter in this book by Lellouche et al., 2018).

Clearly, the performance of common model and forecast evaluations requires the adoption of common objectives and principles. Thus, from the start GODAE validation experts adopted approaches in line with those of the weather forecast community. First, assessment is sought and expected in terms of "consistency," then "accuracy," and then "performance" evaluation (see Murphy, 1993, and Hernandez et al. [2015] for details on these evaluation principles). Finally, the "fit-for-purpose" principle associated with "external" metrics discussed previously is applied with the goal being to provide a more user-oriented evaluation.

Additionally, a common framework and inter-comparison implies the use of a common vocabulary and standardized tools. To accomplish this, Class 1, 2, 3, and 4 metrics categories were defined in order to implement a technical framework for model-to-model and model-to-observation comparisons (see details in Hernandez et al., 2009). Initially, this framework was defined for open ocean, physical eddy-permitting modelling assessments. Subsequently, this framework has been used for regional, higher-frequency, higher-scale forecasting systems as well as for reanalyses intercomparison. And most recently, it has been extended to sea ice or biogeochemical variables.

As part of the GOV Intercomparison and Validation Task Team activities, intercomparison tasks based on Class 1, 2, and 4 metrics have been conducted, resulting in a number of global and regional published results (Hernandez et al., 2015; Divakaran et al., 2015; Oke et al., 2012; Ryan et al., 2015). A particular effort on Class 4 metrics is ongoing, based on model-to-observation comparisons for routine monitoring of hindcasts and forecasts. This approach compares water column temperature and salinity assessment to profilers, sea level data to along-track satellite altimetry data, and sea surface temperature (SST) to drifter measurements. The initiative has been expanded to include sea ice concentrations using satellite data, and efforts are ongoing to extend it to comparisons of surface velocity to drifter trajectories. The goal is to develop global and regional OOFMS with eddy-permitting to eddy-resolving capability. However, the scales (observability) provided by these data do not allow assessment of shorter scales (see further discussion below).

Centres that have adopted this international intercomparison framework on a routine basis are now using it for their internal system evaluation (e.g., Blockley et al., 2014; Lellouche et al., 2013). As described in Hernandez and Melet (2016), this framework has also been adopted for the CMEMS product quality assessment framework (see below).

Additionally, the GOV Intercomparison and Validation Task Team intercomparisons initiatives are providing experiences and feedback with regards to community best practices, in particular how to structure the activity among international partners and what to ask, in a context of where an operational centre's expertise, tools, and goals are evolving rapidly. The Class 4 intercomparison

projects called attention to the global OOFMS involved and strengthened their comparisons with regional OOFMS at their national levels (e.g., Australia and the US east coast; chapters by Roughan et al., 2018; Wilkin et al., 2018). That said, the GOV intercomparison framework has also generated a great deal of interest among regional operational centres worldwide that are working to enhance certain comparison partnerships between global OOFMS and their regional system (e.g., the China Sea; Zhu et al., 2016).

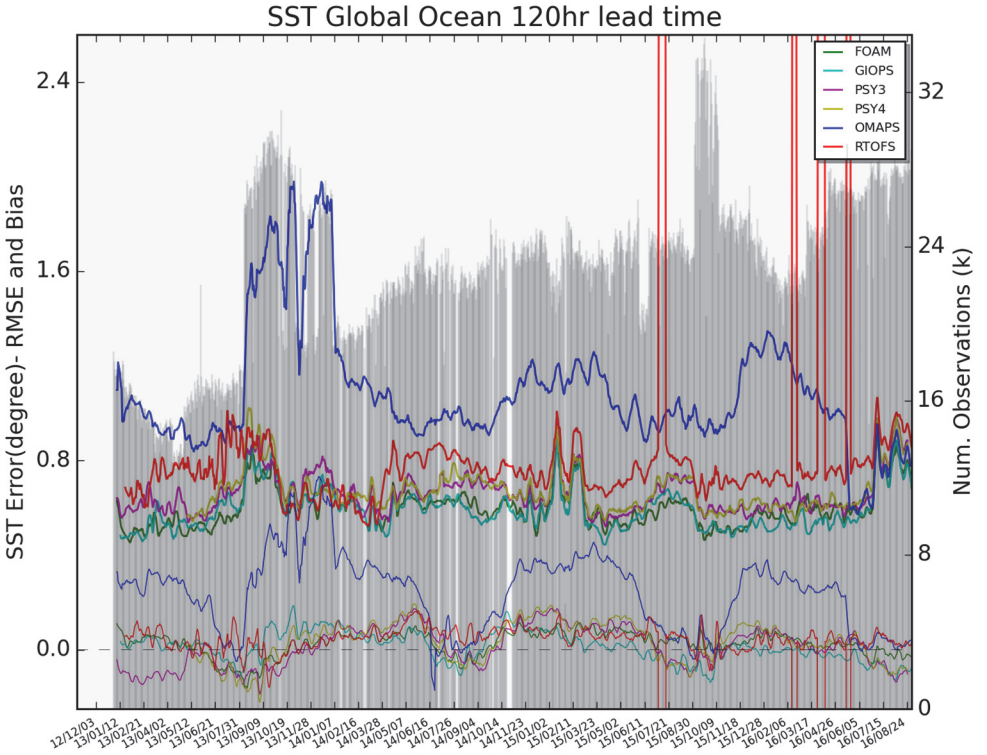


Figure 29.2. Time series of SST Class 4 global statistics from the 2013-2016 period. Six operational systems' five-day forecast real-time evaluation comparing SST to drifters. FOAM (¼°, UK-Met) , GIOPS (¼° coupled, ECCC, Canada), PSY3 & PSY4 (1/4° & 1/12° Mercator Océan, France), OMAPS (1/4°-1/10°, Bureau of Met. Australia), RTOFS (1/12°, NOAA/NCEP USA). Statistics of differences between model forecasts and drifter measured temperature: global root mean square differences (thick lines) and global biases (or mean differences, thin lines). Shaded gray bars corresponds to the number of drifter data. This synthetic assessment in the frame of GOV Intercomparison and Validation Task Team was computed by the Department of Fisheries and Ocean and the Environment and Climate Change Canada.

The GOV Intercomparison and Validation Task Team Class 4 intercomparison involves monitoring the overall performance of the participating global system. In recent years, an annual review was presented to GOV scientists, addressing the strengths and weaknesses of these systems and revealing problems with the evaluation approach. Fig. 29.2 shows that most of the systems involved (with the exception of the Australian system, OMAPS) offer similar skill scores meaning resolution was a strong penalty compared to SST from drifters. Notably, the OMAPS system was upgraded in 2016 and Class 4 monitoring shows that it now offers the same overall quality as the others. Also noteworthy are peaks associated with the RTOFS time series: here, the computation of Class 4 differences was performed without quality control of the drifter SST data and outliers were

not removed, thus greatly increases the measure of misfit. Based on this, the evaluation procedure at NOAA/NCEP has been upgraded as well. This example demonstrates another key objective of the GOV intercomparison tasks: strengthening and maturing the operational oceanography community toward organized, standardized, and shared practices. Later on, this evaluation framework could also be adopted by international organizations such as the WMO and the Intergovernmental Oceanographic Commission's Joint Technical Commission for Oceanography and Marine Meteorology (e.g., Schiller et al., 2016).

Overview of Existing Evaluation Approaches in Operational Oceanography with the CMEMS Product Quality Policy

General evaluation concepts in CMEMS operational centres

The CMEMS aims to provide regional and global products of the "blue" (physical variables), "green" (biological/biogeochemical low trophic level variables), and "white" (sea ice variables) ocean (Drévillon et al., 2018, this issue, as part of the GOV summer school contribution). The product scales cover from global mesoscale (20-50 km) to regional submesoscale (2-10 km) in the European seas, at frequencies of hourly to monthly. The CMEMS delivers real-time: hindcasts, short-term forecasts (three to 15 days). The service also delivers three-dimensional ocean temperature, salinity, currents, chlorophyll, nutrient content, dissolved oxygen, as well as other parameters at the surface such as sea level, waves, and sea ice variables. And, in delayed mode, the CMEMS provides ocean reanalyses and reprocessed observed products for the same parameters.

The CMEMS put a tremendous amount of effort into the assessment of the OOFMS' performance and skill as well as the evaluation of the product's accuracy (Hernandez and Melet, 2016; Le Traon et al., 2017). The CMEMS evaluation approach and the tasks performed can be split in several distinct categories:

Calibration of ocean models and estimation tools. This task is carried out when models or estimation tools are revisited, and when their algorithms need to be adjusted. Most often, comparison to the "ocean truth" based on observations or reference data is performed. Also very often, these situations (in time and space) are chosen in the most favorable way. For instance, a model will be compared to a comprehensive dataset from various oceanographic campaigns, where a large amount data has been recorded. This permits testing of new algorithms under different sea conditions and accommodates a large range of ocean process behaviors of interest.

Pre-operational qualification of the OOFMS. This task is performed in the CMEMS when the existing OOFMS is going to be replaced by a new and improved one. In such cases, the new system is tested over a given period in pre-operational conditions and compared to the existing system in order to measure improvements and potential benefits of the upgrade. Observations are most often used to provide an "ocean truth," where the existing and new OOFMS are compared. Of course, "non-regression" is an important criteria and it is expected that a new OOFMS will beat the

performance of an existing one. Typically, this qualification is completed over one year or more of simulations in order to test the OOFMS for various seasons and to take into account the ocean variability. This type of evaluation is "internal."

Routine validation of OOFMS. This task is carried out in real-time or near real-time. The goal is to monitor performance of the system on a daily basis in order to alert operational teams to major mismatches of the system against the "ocean truth." Every observation available in real-time conditions is usually taken into account, and most of the time the same observation is used by the assimilation procedure. When observations are not available, reference information such as climatologies can be used. Additionally, dedicated metrics can be applied to provide a specific user assessment in order to characterize the reliability of ocean products against "ocean truth" or the fit-for-purpose when considering specific applications (see earlier discussion on external metrics).

Off-line validation of reprocessed products and ocean reanalyses. For all real-time products, the CMEMS associates provision of reprocessed, observation-based products or reanalyses. The objective of this task is to offer an accurate description of past ocean conditions by providing reprocessed information rather than an accumulation of real time hindcasts. Dedicated validation tasks are performed with two main goals in mind: 1) characterizing the accuracy level of these products in order to communicate their reliability, and 2) measuring their quality as compared to real-time products. For comparison to the "ocean truth" off-line validation usually benefits from using a reprocessed, more accurate and more comprehensive dataset. Many types of observations not available in real-time, and specifically quality-controlled and calibrated, offer a more complete description of the three-dimensional ocean variability for many physical and biogeochemical variables such as datasets of vessel-mounted acoustic Doppler current profilers, tide gauge data not transmitted in real-time, CTDs from sea experiments, fluorescence measurements, etc. Moreover, as discussed later in this chapter, several products from different origins can provide past period estimations for the same ocean variables. In such cases, through intercomparison and by taking into account their relative strengths and weaknesses, one can infer their relative accuracy or, at least, propose an overall accuracy level.

Quality control of upstream information used by the OOFMS. Both for real-time forecasts and for production of reanalyses, errors caused by external information used to run the systems are considered, tracked, and sometimes reduced with adapted corrections. Several types of information are currently subject to dedicated quality control: forcing fields, observations used in the assimilation, and boundary conditions from other systems.

Internal control of coupled component of the OOFMS. Coupled systems are used in the most recent versions of the CMEMS, initially for the biogeochemical forecast and coupled with the physical ocean model, but also coupling atmospheric, waves and ocean models; or coupling optical and biogeochemical models. The newest approach consists of monitoring the exchanged variables at the interface or some key parameters (e.g., the mixed-layer depth variations, whose errors may exaggerate the vertical fluxes of nutrients and impact the quality of the primary production by the biogeochemical model).

For all categories, assessment is always sought and expected in terms of "consistency," but the majority of efforts are dedicated to evaluating "accuracy" and "performance" (see Murphy, 1993 and Hernandez et al. 2015 for details on these evaluation criteria). The "fit-for-purpose" criterion is associated with "external" metrics evaluation and aims, through routine and off-line validation tasks, to provide a more user-oriented assessment of the CMEMS products.

Whenever possible, the CMEMS metrics are based on comparisons to observations and upstream information is sought internally during production through Thematic Assembly Centre deliveries. This guarantees that 1) all observations used are quality-controlled and sometimes reprocessed (for off-line validation); 2) producers of observation datasets are known internally by all CMEMS; and 3) in cases where spurious observations are detected when used for validation, a blacklisting procedure can inform Thematic Assembly Centre experts and trigger a corrective task.

These observations may be used through assimilation by the CMEMS OOFMS. However, the CMEMS product quality strategy does not rely on assimilation statistics (i.e., statistics on misfits and increments) for evaluating accuracy and performance. Even if observations are used through assimilation, dedicated Class 2 and Class 4 metrics comparing model values and observations, are carried out (see Hernandez et al., 2015 for an introduction to these metrics). In such cases, the evaluation is not fully independent. However, these approaches limit the filtering effects of the observation operator used by the assimilation scheme that take into account the model's representativity and compute misfits.

Providing quality information to CMEMS users

The CMEMS product quality strategy is primarily dedicated to informing users about the reliability of the products available for download. A series of difficulties have been identified in this communication effort:

- Most calibration, qualification, off-line, or routine validation tasks are based on "internal" metrics that characterize the accuracy and performance of the OOFMS with respect to its representativity. Following the example above: a sea level assessment will filter out tidal signals in the observation dataset if the operating system does not contain tidal dynamics. Consequently, it is important that users be informed of that limitation.
- As reviewed in Hernandez et al. (2015), sometimes there is a failure of a model to accurately represent a specific ocean process. However, sometimes this is a failure of the evaluation approach, in that it is not able to assess the effective skill of the OOFMS because it is not using the appropriate metrics. This issue may arise when the "geography" of an assessed area contains various scales and processes (e.g., open ocean and coastal zones), with a metrics design to characterize some of these scales and processes. For instance, the use of global SST-mapped products is not adequate to identify coastal fronts when assessing both the open Celtic Sea area and the Bay of Biscay in some of the regional CMEMS production centres.
- One aspect of the above issue is the "traditional" use of metrics based on Gaussian statistics of model-to-observation comparisons. That is, mean differences (also called biases) and

root mean square differences (also called root-mean-square errors) usually tend to hide the real nature of the discrepancies. For instance, a model field may contain well-shaped fronts and eddies that are not properly phased in time, which may create very large errors. Users would need to know that the forecast provides appropriate scales and features, but at the wrong time. In a case such as this, a metrics that provides some uncertainty on the phase lag of these predicted features would be valuable. Conversely, low bias or root-mean-square difference statistics over large areas may not reflect some high and localized errors, unidentified due to their relative weight in the overall metric's computation.

- Alternatively, "tradition" can also be a drawback to user expectations. For example, users expecting an evaluation based on comparison to observations consider it a paradigm to evidence departure from the "ocean truth." But given the scarcity of ocean observations, evaluation sometimes must rely on other approaches such as ensemble assessment. In such cases, there is no "ocean truth" but rather the comparison of several estimates of a given ocean process with the assumption that some of the estimate's errors are not correlated. Therefore, a probability level of accuracy is provided, but users need to be properly informed on how to manage it.

Over time, the CMEMS has adopted various methods for communicating product quality to users. First, it provides a quality information document (QuID for every product in the CMEMS' product catalogue (http://marine.copernicus.eu/services-portfolio/access-to-products/). The QuID offers a comprehensive description of the OOFMS itself, the way the system's performance and the product's accuracy are assessed, and it discusses results of this validation. In order to offer a quick understanding to non-expert users, QuIDs include an executive summary that provides tables of estimated accuracy numbers. The goal of these estimations is to communicate some overall error level for a given product. The QuIDs are produced each time the OOFMS is upgraded or when the quality or reliability of the associated products is changed (typically every one to two years).

For the most recent accuracy numbers, users can visit the CMEMS website (http://marine.copernicus.eu/services-portfolio/scientific-quality/). There, using Class 4 metrics, every production centre provides model-to-observation statistics for their region of interest on a quarterly basis. Users can learn about recent changes, time series, quality improvements over time, and successive upgrades of the operational systems (going back to 2013, when this service was established). However, this monitoring is standardized for all types of products across all areas, so it is complemented by specific product quality information available from the different CMEMS production centres.

Soon, links to regional websites will be added in order that will connect site visitors to related regional overviews and expertise. These websites are designed to emphasize local aspects of the CMEMS product accuracy or highlight specific aspects of the quality, reliability, or added value of the products in light of a particular downstream user's activities. For instance, the Baltic Operational Oceanography System website (www.boos.org) provides a comprehensive overview of operational oceanography tools around the Baltic Sea (e.g., the community and tools, real-time observation information, forecasts, etc.). In particular, it provides a daily multi-model assessment that informs

the Baltic stakeholders about the reliability of Baltic ocean products with a real-time warning system. A similar initiative has been started by the Baltic Monitoring and Forecasting Centre (MFC) for the North Sea (http://noos.eurogoos.eu/model-results/), with plans to build a dedicated regional quality monitoring website. Forecasting centre regional verification websites have also been developed for the Arctic (http://cmems.met.no/ARC-MFC/V2Validation/index.html) and the Mediterranean Sea (http://medforecast.bo.ingv.it/mfs-copernicus-evaluation/ [illustrated in Fig. 29.3] and http://medeaf.inogs.it/nrt-validation). At the same time, the CMEMS Thematic Assembly Centres are developing ways to provide users with online information about their observed products, e.g., for sea level (https://duacs.cls.fr) or for ocean colour (http://octac.acri.fr). Plans call for these websites to be integrated into the CMEMS framework.

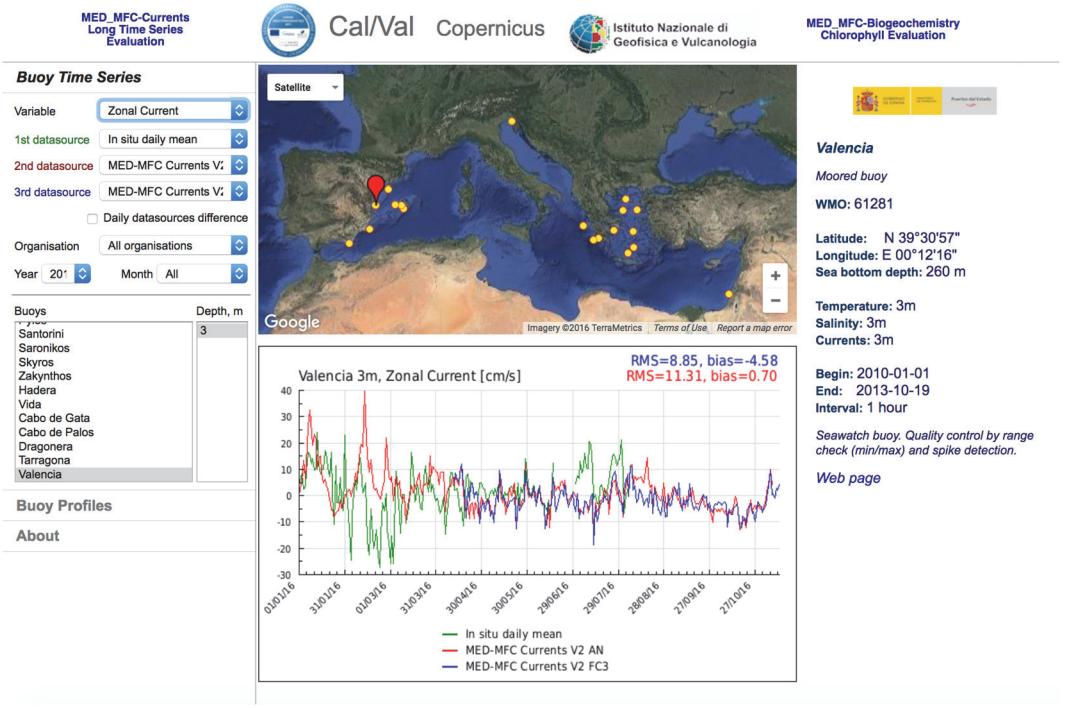


Figure 29.3. Screen copy of the monitoring website developed for the Mediterranean Sea MFC assessment. Against fixed moored platform observations, the time series of measured zonal currents are compared to analysis and three-day forecast. Courtesy of G. Coppini, CMCC and N. Pinardi, INGV, Italy.

On an annual basis, the CMEMS has started to provide regular description of the ocean climate changes and variability dedicated to the scientific community, the decision makers or the general public. As shown with the first initiative: the annual Ocean State Report 2015 (von Schuckmann et al., 2016), the accuracy levels of the CMEMS product used are detailed. For this ocean climate assessment annual reporting it was decided to: 1) use only verified and quality-controlled CMEMS products for inferring and discussing the ocean state; 2) provide a level of confidence, whenever possible (i.e., error bars for every ocean indicator illustrated in the report); and 3) move toward reliance on an ensemble assessment in order to be more confident of error levels.

Continuing to improve the way that product quality information is shared with users is critical. While users are certainly interested in product quality and reliability, on a dedicated user forum set

up by the CMEMS they appear to prioritize timeliness and continuous delivery as their most immediate concerns (D. Obaton, CMEMS Service Manager, personal communication, 2017). This makes clear how important it is that users understand and be aware of any potential areas for product anomalies.

Estimated accuracy numbers, while informative, should be considered by users as the most basic level of product quality information They speak to the overall quality of a product based on first and second order Gaussian statistics. However, these statistics do not allow for characterizing specific anomalous behaviours of the OOFMS in specific areas or events. For example, estimated accuracy numbers for satellite altimetry reflect aggregate information provided by satellites (Table 29.1) deduced from cross-over difference statistics but they should be complemented with a specific error analysis for along-track noise or large wavelength errors (e.g., Le Traon, 2013) as illustrated in Fig. 29.4. For model products, estimated accuracy numbers are typically obtained through comparisons to observations or climatology over one-year periods or longer (as illustrated for the North West Shelf Regional MFC in Table 29.2). However, nothing in the estimated accuracy numbers would alert a user to potential product error(s). In Table 29.2, values are given with two-digit precision, which is not necessary for an overall estimation of errors; however, this level of accuracy does allow experts to identify further improvements when these numbers are compared in time over successive versions of the operational systems.

More process-oriented metrics are now being tested to increase the value of estimated accuracy numbers. The CMEMS Arctic MFC is pioneering a new approach for sea ice extent, sea ice concentration, and sea ice type assessment contingency table metrics that characterizes the number of good forecasts or occurrences compared to observations, and discussed against persistence score, as it is also done worldwide by other operational centres (e.g., in Canada, see Smith et al., 2016).

Altimeter	NRT errors (cm rms)	DT errors (cm rms)
OSTM/Jason-2	< 4	< 3
AltiKa	< 4	<3
Cryosat2	< 6	<5
HY-2A	< 6	< 5

Table 29.1. Estimated accuracy numbers provided for satellite altimetry in the CMEMS Sea Level QuID, for Near Real Time (left) and Delayed Time (right) products.

In weather forecasting, verification tools are used to characterize the occurrence of specific events such as rain with a confidence level. The WMO Joint Working Group on Forecast Verification Research puts forth diagnostic metrics, but many of these meteorological assessment methods are based on ensemble forecasts (Ebert et al., 2013). Currently, the CMEMS products considered to be “core products” of the marine environment may not reach the scales necessary to described many of the synoptic events of interest, in particular extreme events that may develop on short timescales near the coast.

Variable	Location	Supporting observations	RMS error	Mean error
M2 tidal harmonic (amplitude)	Whole region	Tide gauge data	12 cm	0.59 cm
M2 tidal harmonic (phase)	Whole region	Tide gauge data	12.3°	1.9°
SST	Full domain	In situ observations	0.51°C	-0.025°C
	Continental shelf	In situ observations	0.52°C	-0.032°C
T profiles	Full domain	In situ observations	0.65°C	-0.076°C
S profiles	Full domain	In situ observations	0.15	-0.0
Bottom temperature	Continental shelf	Climatology	1.5°C	0.95°C
Mixed layer depth	Full domain	In situ observations	121.77 m	21.7m

Table 29.2. Estimated accuracy numbers available for the CMEMS North West Shelve Regional MFC, for different variables, from QuID.

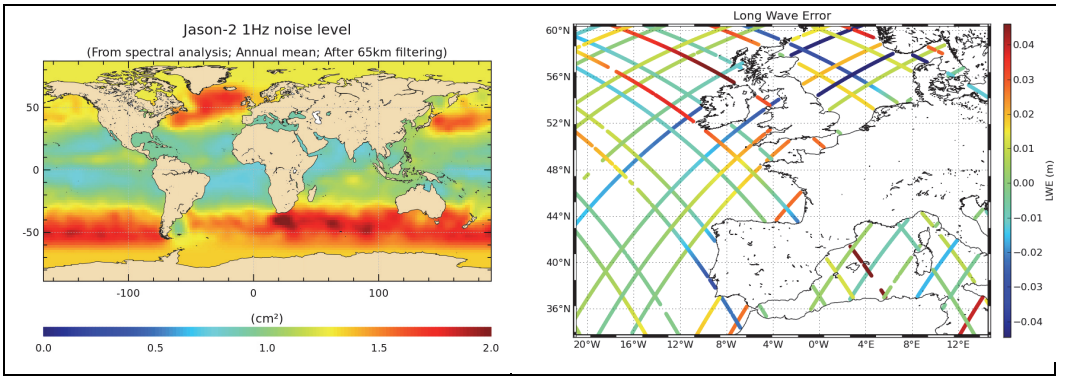


Figure 29.4. At left: Jason-2 along-track sea level anomaly noise as provided by the CMEMS Sea Level Thematic Assembly Centre (I. Pujol, CLS, as part of CMEMS, 2017). At right: Example of along-track sea level anomaly long wavelength errors for Jason-1 (C. Dufau, CLS, as part of CMEMS, 2017).

Finally, when it comes to informing users about product reliability, the CMEMS does this mainly through static documents such as the QuIDs. But ideally, these should be complemented with near real-time information published online. That said, dedicated user-oriented applications and downstream services that use the CMEMS products and provide services on websites, tablets, and smartphones are now available (e.g., www.sea-condition.com; personal communication, Coppini, 2017). Users must be informed of the potential errors and reliability of these services based on the CMEMS product quality information.

Identified path of improvements in the CMEMS product quality assessment framework

Within the CMEMS framework, areas for improvement have been identified and described by Hernandez and Melet (2016).

Reference information used for evaluation.

Due to some national, institutional or private sector data policy, ocean observations are not freely available to all operational centres or the oceanographic community. Some production centres may access and use them for validation while others do not. Sharing data through CMEMS Thematic Assembly Centres would ensure free access to all (particularly to historical datasets), if data owner give access to them. If not, and in cases difficulties arise, the growing influence of the CMEMS at the European level may help to push owners to share them.

Ocean observations used in the CMEMS validation framework are quality controlled. Furthermore, the implementation of black and grey listing internal mechanisms through assimilation procedures from MFCs toward Thematic Assembly Centres is also a way to evidence observation anomalies. Both procedures allow Thematic Assembly Centres to better manage outliers or instrumental problems.

Issues related to fixed measurement data also need to be addressed. Historical data from moorings, tide gauges, other platforms and instruments usually located near the coast is often difficult to access. Many countries and groups do not submit their historical or real-time observations to a common database or a Global Data Assembly Centre. When subjected to the standardization of a Global Data Assembly Centre, fixed instrument measurements can be quality-controlled, which allows for multi-model assessments. Such is the case with the BOOS framework in the Baltic Sea, which makes it possible to compare forecasts from the Danish, Swedish, Finish, and Norwegian regional operational systems. A global system could be handled similarly; this would allow us to measure the benefit and added value of these regional 1-3 km resolution systems in contrast to the CMEMS global 1/12° eddy-resolving system, forced by global European Centre for Medium-Range Weather Forecasts atmospheric fluxes and not considering the tides. Fig. 29.3 illustrates this type of OOFMS monitoring based on fixed platforms. Zonal current time series are compared to a three-day forecast and hindcast from the Mediterranean Sea MFC. The metrics used are biases and root-mean-square differences. Other CMEMS velocity products could also be compared and their skill characterized in a Taylor diagram (Taylor, 2001), with other metrics evidencing specific weaknesses of the OOFMS (e.g., rotary spectrum, threshold and contingency table metrics, etc.). Because many of these platforms are located near the coast where many applications can be anticipated, they could support the design of user-oriented external metrics.

Strengthen the validation and verification activities on most-used variables.

Surface variables (such SST, mixed layer depth, surface current, sea level, chlorophyll) are among the most used because they have numerous applications. Therefore, it makes sense to focus the quality assessment of these surface variables on high-frequency variability and assessment of the

diurnal cycle. Until recently, the CMEMS typically provided product delivery on a daily basis whereas today, it provides hourly products for surface parameters.

The CMEMS is a multidata framework. This means that for a parameter of interest (e.g., SST), several products are available in a given area: from the regional and global MFCs (for model products), to Thematic Assembly Centres (for observed products), as well as, of course, from real-time or delayed-mode reanalysis/reprocessing. Hence, comparison between these different estimates of a given parameter will allow for a better characterization of the accuracy and weaknesses of the products and ultimately evaluation of the reliability of a given product for a given application. The CMEMS multi-data framework also encourages the sharing of best practices. For instance, the SST satellite community may share their metrics with MFCs to the benefit of modellers, who sometimes lack the expertise that SST producers have when considering SST assessment.

Better monitoring of the information used in the OOFMS.

Data for assimilation should be systematically quality-controlled against predefined criteria or using a departure from the model forecast. If rejected, grey or black listing has to be considered, taking into account model forecast errors and representativity. External forcing functions (atmospheric, bathymetry, river run-off.…) also need to be monitored and possibly corrected, in cases of anomalies.

To improve nesting OOFMS, boundary conditions should be monitored. Errors from the large-scale model need to be identified in order to infer their impact in the nested system. Moreover, inconsistencies between large-scale and regional-scale dynamics will first appear at the boundary, and also need to be characterized.

Furthermore, coupling strategies are becoming widespread in operational centres and seamless atmosphere-wave-ocean modelling approaches are being developed. Likewise, all CMEMS MFCs now offer biogeochemical modelling, coupling the physical components on a daily basis. But biogeochemical models need dedicated assessments based on biogeochemical observations, which presents a challenge because of the sparseness of the measurements and the fact most variables in the biogeochemical models are not observed. Also, the performance of a biogeochemical model can be modelled with consideration for the accuracy of the physical forcing model. Some physical variables, such as vertical fluxes, have a large influence on the biogeochemical model. Therefore, a number of key physical parameters have been identified for dedicated monitoring including: bottom temperature, stratification and length of the stratification period, mixed layer depth variability, vertical velocities and diffusivity, euphotic layer depth, solar radiation and its penetration into the sea water, and wind stress. Dedicated monitoring of these parameters would allow for better characterizations of the behaviour of a biogeochemical model in direct response to the physical forcing, which has the potential to be erroneous.

It is also important to remember that the CMEMS is a distributed network of centres and experts who sometimes collaborate on validation/verification development in other frameworks. Therefore, when innovations are transferred to the CMEMS processing chains, it directly benefits the overall product quality of the CMEMS organisation. For example, the CMEMS global systems are

benefitting from developments carried on in the GOV intercomparison framework. A similar benefit is seen when the CMEMS products in the Arctic are able to be part of the evaluations performed in the frame of the Year of Polar Prediction project (YOPP) and in the CMEMS ocean reanalyses being part of the international intercomparison framework lead by the CLIVAR/GSOP community.

Novel Evaluation Approaches in Operational Oceanography

Benefiting from existing observations

As discussed above, observations need to be considered first with regard to what part of the "ocean truth" they represent, taking into account the type of sampling they offer (their observation representativity). Then, their reliability (i.e., accuracy and precision) at different scales of time and space should be considered.

Table 29.3 summarizes the use of existing observation datasets in operational oceanography for validation/verification purposes. The first and second columns list the instruments and measured parameters, the third column details the characteristics of each type of measurement, and the fourth column indicates the type of assessment that is performed by each instrument or measured parameter.

When validating OOFMS, it is important to remember that most raw measurements require dedicated expertise and permanent quality control monitoring in order to be used for validation purposes. Most of these measurements depend on a lot of additional information in order to be derived into the final observations. This is particularly true for satellite measurements, which need to account for the satellite platform behaviour (e.g., orbit, rolling), noise of the instrument, modification of the measured information through the atmosphere and at the surface, and other ancillary information.

Even though real-time data quality is not fully satisfactory given there is no time to obtain the complementary information needed to correct raw data, provision of measurements in real-time is a key element in operational oceanography. For most users, observations collected in real-time provide some evidence of the reliability of estimates and forecasts. Of course, corrected datasets are more precise, thus allowing us to calibrate operational systems, evaluate ocean reanalyses in greater detail, and evaluate the performance of the operational system in delayed-mode. But all measurements cannot be collected in real-time, which means that real-time assessment suffers from less accurate reference data and consists of only some of the observations that describe the ocean processes.

Note also that scales and representativity of observations need always be considered before OOFMS validation. As mentioned earlier in this chapter, editing, filtering, or averaging might be necessary to compare model values and measurements at the same time and space scales. However, for external metrics, the full measurements are preferred. Observations might also be rejected. For instance, coastal data that describe very local processes that cannot be filtered out before comparisons with larger-scale model.

Observation type	Parameters	Measurement characteristics	Use for OOFMS evaluation
CTD + additional sensor on rosette (ADCP…)	• T/S • Additional parameters (U/V)	• Vertical profile • Non frequent • Not systematic real time transmission • High resolution, high quality (high precision instrument, then quality control on profiles and possible corrected values)	• Real time (RT) –sometimes- validation of water masses and stratification • Delayed mode (DM) precise assessment of water masses and stratification • RT/DM validation of additional parameter • Unless for a dense section, where synoptic scales can be evaluated, used for large scale assessment
Water samples from experiments	• T/S • Chemical properties of sea water • Biogeochemical properties of sea water • Biological analysis	• Non frequent, sparse • Time and space sampling depending on experiments • Mostly processes off-line in labs and available with substantial delay • Top quality measures	• DM Validation or dedicated calibration of physical or biogeochemical parameters • Unless for a dense section, where synoptic scales can be evaluated, used for large scale assessment
XBT / XCTD	• T • T/S	• Vertical profile • Repeat sections / sea experiments • Mostly real time transmission • Low quality temperature and salinity profiles	• RT/DM validation of temperature/MLD/thermocline • RT/DM validation of water masses and stratification • For frequent and dense section, used for synoptic scales assessment
Argo profiler	• T/S	• Vertical profile • 5-10 day cycling • real time transmission	• RT/DM validation of water masses and stratification deep to 2000m • Global • Depending on density, can be used for synoptic or large scales assessment
BGC Argo profiler	• T/S • $N0_3$, O_2, Chl-a (fluorescence), downward irradiances and derived optical properties	• Vertical profile • 5-10 day cycling real time transmission • Bio sensors quality still under improvements	• RT/DM validation of low trophic levels / Dissolved oxygen • At present, used for specific assessment

Observation type	Parameters	Measurement characteristics	Use for OOFMS evaluation
Gliders	• T/S • Additional parameters	• High frequency vertical profiles • Real time transmission	• RT/DM validation of water masses and stratification deep to 1000m • RT/DM validation using additional sensors • At specific locations of interest • Used for synoptic assessment
On board TSG or FerryBox	• T/S • Fluorescence • Turbidity • pH • Oxygen • Phyto/zoo	• Along the route of ship (merchant or oceanographic vessels) • Need careful calibration and processing using water samples, in particular for S and biogeochemical measurements	• DM (sometimes RT) validation of surface T/S properties • DM validation of biogeochemical properties • Used both for synoptic and large scale assessment
Miscellanous opportunistic T sensors (Recopesca, net sensor)	• T/S • Turbidity	• Fishermen nets : follow their route (mostly tested over continental shelves) • Low quality T/S sensors attached to the net • Real time transmission	• Need dedicated QC • DM validation of T/S, dense measurements • Tested on coastal waters
Sensors on sea mammals	• T/S • Additional parameters	• Depending on sea mammal, vertical profiles at various depth and frequencies • Instruments may be biased • Transmission not always real time	• Need dedicated QC • RT/DM validation of water masses and additional sensors • Useful to cover high latitudes and near sea-ice areas. • Used like Argo profilers
Ice tethered - profiler	• Ice temperature • Near-surface water temperature	• Over the sea-ice: may move with the ice, or in water if melting • Very few profilers • Profile down to 800m • Real time transmission	• DM/RT validation of water mass below the ice
Drifters	• Trajectories • Temperature • Air pressure • Salinity (specific) • Wind (specific) • Rain (specific) • Wave (specific)	• Real time transmission • Global array • Different type of buoys, different depth of drogues • Post-processing of trajectories to infer U/V (drogue loss estimation)	• RT/DM validation of U/V at given depth, Eulerian and Lagrangian assessment • RT/DM validation of T • RT/DM validation of using specific sensors • Unless dense coverage, used for large scale assessment
VM-ADCP	• U/V	• Ship route measurement below the hull • Not real time • Need post-processing	• DM validation of U/V for surface layers

Observation type	Parameters	Measurement characteristics	Use for OOFMS evaluation
Fixed/moored ADCP	• U/V	• Fixed location measurement • Often not real time • High frequency measures stored on memory, often transmission of time-averages • Need post-processing	• DM validation of U/V for specific layer • Local measurement
Tide gauges	• Sea level	• Fixed location measurement • On the coast (most of the time) • Real-time • High frequency measures stored on memory, often transmission of time-averages • Need post-processing for exact positioning (e.g. land vertical displacement)	• RT/DM sea level validation • Local measurement, but used for large scale assessment of sea level
Bottom pressure gauges	• Bottom pressure	• Fixed location measurement on sea floor • Often not real time (unless cable transmission) • High frequency measures stored on memory • Need post-processing	• Bottom pressure or sea level assessment • DM validation
Wave buoys	• SWH • Dominant part of the wave spectrum (peak, period and direction)	• Fixed location measurement • Real-time • High frequency measures stored on memory, often transmission of time-averages	• RT/DM validation of wave parameters
High frequency (HF) radar	• Waves • U/V	• Fixed location measurement • Real-time • High frequency measures • Coastal measurements from few km to few hundred km • Need post-processing • Near-real time	• U/V DM validation on very specific locations along the coast

Observation type	Parameters	Measurement characteristics	Use for OOFMS evaluation
Satellite nadir radar altimeter & laser altimeter (ICEsat)	• Sea level • SWH • Wind intensity • Ice thickness	• Along-track measurements (e.g., Earth revolution in 100 minutes) • Global coverage • Different repeat or non-repeat satellites and track distances • Different high latitude coverage • Real-time transmission • Specific post processing • Sea surface height scale resolved: 50 km at best	• RT/DM validation of sea level, SWH and wind • DM validation of ice thickness
Satellite IR radiometer	• Surface brightness temperature	• Along swath (variable length) or geostationary measurements • Different orbits • Global coverage • Real time transmission • Specific post processing • Cloudiness problem • Transform in skin, bulk or foundation SST values	• RT/DM SST validation • Used for meso- and large-scale assessment
Satellite microwave radiometer & spectrometer	• Surface brightness temperature • Surface salinity • Surface roughness • Sea Ice Thickness (thin)	• Along swath (variable length) measurements • Less precise than IR radiometers for SST • Different orbits • Global coverage • Real time transmission • Specific post processing • Transform in skin, bulk or foundation SST values • Large scale SSS • Thin Ice thickness	• RT/DM SST validation • DM SSS validation (large scale) • DM validation for ice thickness
Satellite scatterometer	• Surface roughness (wind)	• Along swath (variable length) measurements • Different orbits • Global coverage • Real time transmission • Specific post processing	• RT/DM validation of wind
Satellite Synthetic Aperture Radar (SAR)/InSAR	• Sea Ice concentration • Wave spectrum • U/V	• Along swath (variable length) measurements • Different orbits • Global coverage • Real time transmission • Specific post processing	• RT validation of sea ice • DM validation of sea-ice • DM validation for U/V

Observation type	Parameters	Measurement characteristics	Use for OOFMS evaluation
Satellite Imager	• Reflectances (ocean colour) for different visible and near-IR spectral bands	• Along swath (variable length) measurements • Different orbits • Global coverage • Real time transmission • Specific post processing for Chl content • Strong dependence on type of waters and algorithms • Depending on euphotic depth and turbidity, different colours at different layers	• RT/DM validation of reflectances, and chlorophyl content • Used for meso- and large-scale assessment

Table 29.3. List of observations, the parameters measured, and uses for OOFMS validation and verification.

Since the 1970s, satellite instruments and constellations have been constantly improving the observability of increasingly more parameters from space. That said, there has been the strong limitation of measuring only the ocean surface. Although some measurements, such as sea surface height from altimetry or earth potential/geoid from gravimetry, offer integrated estimates of the full water column.

The in situ observing system offers the full three-dimensional description of ocean parameters, although data is more sparse. Nevertheless, the Argo program provided a noticeable improvement of the observability of water masses. The Biogeochemical (BGC)-Argo initiative is expected to make equally significant contributions to the in situ system and to bring new, potentially revolutionary insight by allowing for a synoptic assessment of some key parameters of the biogeochemical processes. The other "new player" will be the coastal high-frequency (HF) radar network. Effort to allow a continuous monitoring of coastal waters is difficult but ongoing, and in the future we might expect HF radar networks in some areas (the US/Canada east and west coasts, European coasts, Australia, China Sea) that will provide surface velocity and sea-state measurement to monitor the reliability of coastal operational systems.

Along-track satellite altimetry: finer scales to be observed

Since the launch of Geosat in 1985 (the earlier SeaSat mission was more a pioneering effort), the sea level and surface geostrophic flow assessment is traditionally performed using along-track satellite radar altimeters with the so-called "conventional" nadir low resolution mode. The recent availability of Sentinel-3A (S-3A) data using an along-track Synthetic Aperture Radar (SAR) mode that reduces noise level of along track sea surface height retrievals (see details in the chapter by Morrow et al., 2018) changes the sea level spatial scales that can be observed and compared to model products. Hence, instrumental noise of the 1 Hz "conventional" data since Geosat needed to be filtered out and was giving access to wavelength larger than 70-80 km. Due to changes of frequency and a finer footprint, the AltiKA Saral altimeter provides access to 40-50 km closer to

the coast, and the S3 SAR mode allows us to reach the 30-50 km along-track resolution (see also the Mesoscale Capability Determination estimation by Dufau et al., 2016).

Satellite altimetry repetitivity and time sampling have always been the main limitations for describing the ocean two-dimensional turbulence. However, the constellation of satellites has increased and the combination of several satellite passages allows for better two-dimensional descriptions through mapping techniques. However, two-dimensional reconstructed maps of sea level cannot be considered as reference observations for exact validation. Mapping techniques offer gridded fields, such as ocean models, but also present weaknesses and erroneous extrapolation features, even if new optimal interpolation dynamical techniques are now proposed (Rogé et al., 2017; Ubelmann et al., 2015). This is the reason why the Class 4 approach (see GOV intercomparison above) is based on along-track comparison that provides the evaluation only on one direction, but that allows assessing in a rigorous way specific aspects of the mesoscale dynamics (fronts, size of meanders and eddies, strength of currents etc…).

Recently, in order to assess dynamical turbulence in model fields, comparisons between along-track satellite altimetry spectrum of sea level and similar quantities from the model have been conducted. For a given area, sea surface height wavenumber spectra can be computed and spectral slopes can be determined and mapped globally (Xu and Fu, 2012) or by season (Dufau et al., 2016). The power law of the spectrum can be compared to the turbulence theory, as well as compared between model and satellite data (see the chapter by Morrow et al., 2018). Fig. 29.5 illustrates an evaluation performed between Jason-1 along-track data and best estimates of the CMEMS 1/12° global forecasting system. In this example, when compared to Fig. 1 of Dufau et al. (2016), the model sea surface height presents sharper slopes in the equatorial band.

The incoming Surface Water and Ocean Topography (SWOT) mission's new technology will offer a new paradigm for sea level and ocean turbulence assessments. The interferometric SAR (InSAR) will measure 70 km on both sides of the satellite course, with a 10 km gap at the nadir that will be partially compensated by measurements from a classical altimeter. The swath resolution will be 1 km, but in practice, SWOT is expected to provide a 15 km-scale two-dimensional resolution. This will allow us to better infer vertical movements associated with quasi-geostrophic turbulence at mesoscale (see chapter by Morrow et al., 2018). It will also allow us to monitor errors on coastal models and forecasts, and small mesoscale behaviors of eddy-resolving operational systems.

Sea ice parameters assessment

Sea ice concentration, thickness, drift, and type are operational products of great interest for navigation and off-shore industry. To a lesser extent, these parameters are also needed to close the budget of the Earth climate system, the water budget, and its monitoring on short-to-medium timescales in terms of seasonal forecasts. Ice modelling also needs to progress (see chapter by Bouillon et al., 2018). Sea ice concentrations in the Arctic and Antarctic basins are typically assessed using satellite microwave measurements from instruments such as the Special Sensor Microwave Imager (SSM/I) or, more recently, the Advanced Microwave Scanning Radiometer 2 (AMSR2). The latter is used to perform Class 4 metrics in the GOV Intercomparison and Validation

Task Team's intercomparison framework. However, in real time, marine navigation users prefer to rely on ice charts deduced by ice centres from direct observations or advanced very-high-resolution radiometer (AVHRR), Moderate Resolution Imaging Spectroradiometer (MODIS), Radarsat, or Sentinel-1 SAR imagery. Users look at ice model based products for prediction of ice edge, extent, and thickness. Let's look at why ice operational centres pay particular attention to the validation of these last parameters.

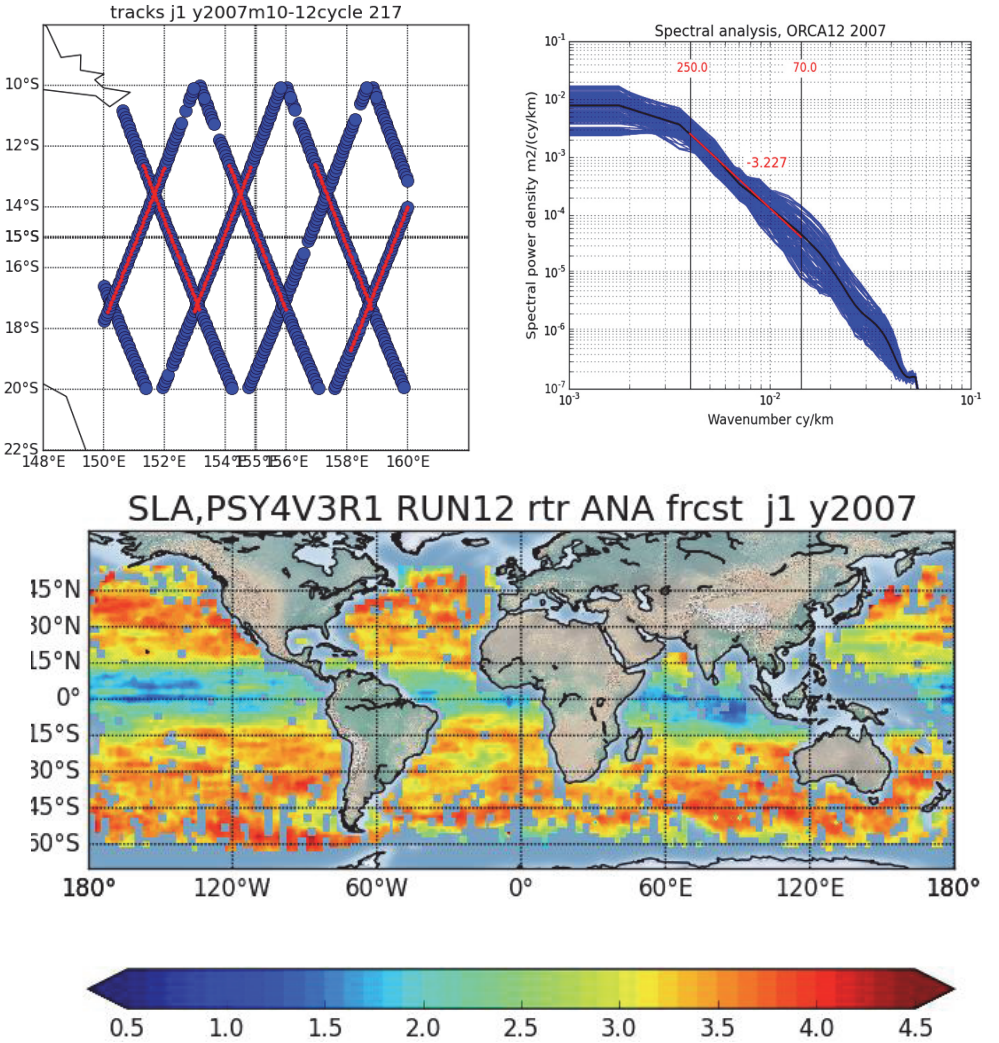


Figure 29.5. Top: Wave number spectra computation over 10°x10° boxes (only tracks longer than 560 km). Top left: example of satellite tracks from Jason-1. Top right: Corresponding spectrum on the along track interpolated SSH fields from the CMEMS 1/12° global model in 2007, with the estimated slope (in red). Bottom: global map of the wave number spectra slope from the CMEMS 1/12° global system.

Hernandez et al. (2015) presented contingency metrics designed to measure skill prediction for correct water or correct ice areas, looking specifically at scores on marginal ice zone. Sea ice edge position is subject to particular assessment due to high user interest in these zones. For instance, the Arctic forecasting system of the CMEMS provides near real-time validation bulletins (Bertino and

Melsom, pers. comm., 2017) in which sea ice edge forecasts are compared to persistence. Moreover, sea ice concentration is assessed using different categories of ice fraction cover and then compared to satellite data (see http://cmems.met.no/ARC-MFC/V2Validation/index.html). Fig. 29.6 compares a Class 4 metrics assessment against AMSR2 data for the global CMEMS operational system. Scores, in terms of mean biases and root-mean-square differences, are plotted for different operational products (these are best estimates that gives the maximum performance of the system). Then, forecast and persistence scores are compared for one to 10 days lead time. In both the Arctic and Antarctic basins, forecast beat persistence in winter, which is not the case in summer (this demonstrates that the system still need to be improved). Next, nominal and former CMEMS global systems are compared, looking at the proportion of correct hits versus the total (PCT, sum of the correct ice and correct water, or CP and CN in the table of Fig. 29.7 below, divided by the total number of samples, i.e., [a+d]/[a+b+c+d]). Improvements are observed in summer. This is due to an enhancement of the full PSY4V3R1 system as well as the implementation of the assimilation of sea ice concentration (see chapter by Lellouche et al., 2018).

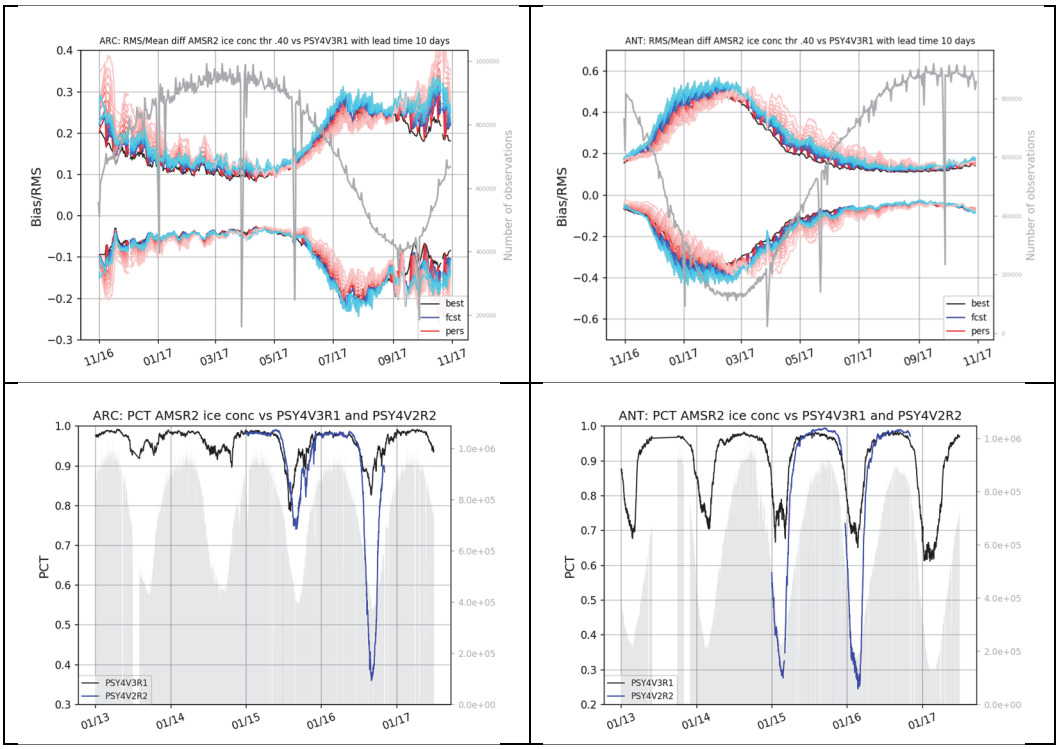


Figure 29.6. Sea ice concentration assessments based on contingency table for the Arctic (left column) and the Antarctic (right column) basins. Top row: Time series of Class 4 metrics for the global operational 1/12° CMEMS system PSY4V3R1 against AMSR2 data. Grey curve: number of non-zero observations. Lower curves: biases. Upper curves: root mean square differences. The black curve for the best estimate. The blue/cyan curves the forecast from 1- to-10 day lead time. The associated persistence score from 1- to 10-day are plotted in red/magenta. Bottom row: time series comparison of PCT (see text) for the nominal CMEMS global system PSY4V3R1 (black) and the former PSY4V2R2 operational system (blue), with a sea ice concentration threshold of 0.4.

New approaches propose to combine ice concentration and ice type or ice thickness information in order to evaluate for which category of ice the operational system is skillful. This is the way Smith et al. (2016) performed their recent evaluation of the Canadian Global Ice Prediction System using a contingency table on sea ice concentration per sea ice thickness categories, and characterizing performances for the different seasons. With the availability of altimeter Cryosat-2 data together with the SMOS satellite retrievals, which provide reliable estimates for ice thicknesses smaller than 50 cm (Tian-Kunze et al., 2014), other new ice thickness assessments are emerging. Dedicated assessment focuses on first-year ice and multi-year ice categories, mostly due to larger ice melting in the Arctic during recent summers, in order to measure the capability of models to correctly estimate multi-year ice location and drifts (Melsom et al., 2017).

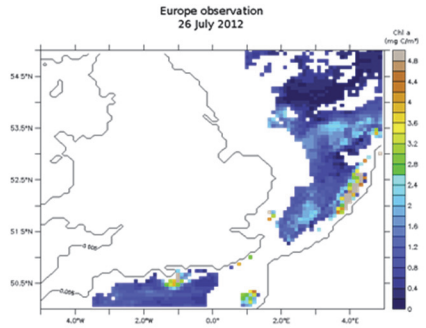
Assessment of biogeochemical parameters and models

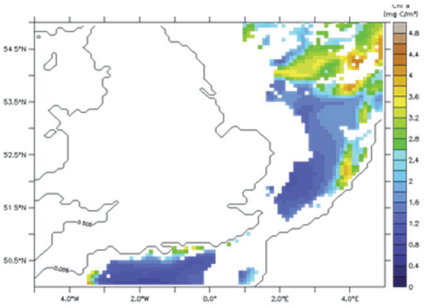
As discussed in Hernandez et al. (2015), biogeochemical operational products that provide information about marine ecosystems and marine life state and dynamics are increasingly in demand by a wide range of users and policy makers, particularly in the context of global change (e.g., Gehlen et al., 2015). Hernandez et al. (2015) also points out that until now available observations did not allow for a detailed assessment of biogeochemical and ecosystem models. Moreover, these models, while complex, with many variables, and specifically tailored for the different domains, still require significant improvement (see chapter by Ford et al., 2018). In practice, biogeochemical modelling key parameters, such as nutrients (e.g. nitrogen, phosphorus, silicon and iron), oxygen, and carbon biomass of plankton compartments, are very sparsely measured across the global ocean. Meanwhile, in situ historical datasets, which are not always distributed, cannot provide comprehensive and reliable climatology (e.g., the main physical parameters).

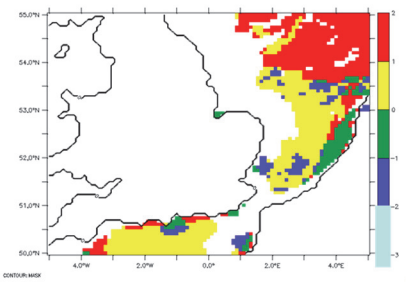
An alternative has been to rely on satellite measurements of ocean colour in order to assess the concentration of chlorophyll that can be linked to the carbon biomass of phytoplankton. Ocean colour reflectance measurements can be combined with sea water optical property parametrisations to provide estimates of chlorophyll concentration. In real-time, ancillary data (e.g., meteorological data) are not available and ocean colour estimates are less accurate (for more discussion, see chapter by Volpe et al., 2018). Furthermore, up until now, in situ data were not available for real-time validation of the satellite reflectance measurements, as can be done in delayed-mode. The CMEMS Ocean Colour Production Centres have developed a method for real-time comparison to a chlorophyll climatology, as presented in Hernandez et al. (2015) and Volpe et al. (2018). This method makes it possible, through consistency assessments, to evaluate if ocean colour estimates suffer from peculiar peaks and outliers. Moreover, a multi-sensor comparison (e.g., between VIIRS and MODIS or between VIIRS and OLCI reflectances) performed between the satellite swath make it possible to evidence biases and apply corrections (Garnesson and Mangin, as part of the CMEMS validation, pers. comm. 2017).

Comparison of model chlorophyll content against satellite ocean colour retrieval is widely used. However, due to the very specific aspect of distribution in time and space of phytoplankton blooms, the classical statistics can be complemented by image oriented assessment. Size and intensity of the

observed bloom can be compared with time lag or space shift through windowing techniques, and then contingency metrics can be applied in order to measure if and when the model could totally or partially reproduce the bloom. Imagery techniques are being used for process-oriented metrics in weather forecast evaluation (e.g., Gilleland et al., 2009). These new approaches have been tested in the CMEMS framework (see http://marine.copernicus.eu/services-portfolio/scientific-quality/#novelmetrics|novelmetrics|biogeochemistry). Fig. 29.7 illustrates the daily performance through a contingency table, also called the ROC (Relative Operational Characteristics) index, for a given threshold value of the chlorophyll field of 2 mg/m3 in the North Sea. Performance time series can then be drawn using the Hanssen-Kuipers Discriminant, also known as the Peirce Skill Score (e.g., Gordon, 1982), where the False Alarm Rate (FAR, the incorrect negative divided by the sum of the incorrect negative and the correct negative, IN/[IN+CN]) is subtracted from the Hit Rate (HR, the correct positive divided by the sum of the correct positive and the incorrect positive, CP/[CP+IP]). Both the HR and the FAR range from 0 to 1. If higher than 0.5, the rates are meaningful; thus, when the Hanssen-Kuipers Discriminant is higher than 0.5, it implies that a meaningful hit rate (HR>0.5) is not neutralized by a high and meaningful false alarm rate. Since the value of the threshold is arbitrary, the Hanssen-Kuipers Discriminant needs to be calculated for a range of thresholds between the minimum and maximum chlorophyll values.







Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

a: Correct Positive
b: Incorrect Positive
c: Incorrect Negative
d: Correct Negative
Hanssen-Kuipers Discriminant = (ad-bc) / (a+b)(c+d)

Figure 29.7. Top left: Chlorophyll concentration (mg/m3) in the North Sea from the 1 km MyOcean satellite ocean colour product merging several sensors the 26/07/2012. Top left: chlorophyll content for the same day from the 7 km FOAM ERSEM (UK-Met Office) operational system. Bottom left: Contingency metrics map, with a chlorophyll threshold of 2 mg/m3, comparing the model and observations. Green (CP): hit, observed and forecast. Yellow (CN): correct non-event, observed and forecast. Blue (IP): false alarm, forecast but not observed. Red (IN): miss, observed but not forecast.

The development of Argo floats equipped with biogeochemical sensors, the so-called BGC-Argo floats, provides access to vertical profiles of downward irradiances, photosynthetically available radiation, turbidity, coloured dissolved organic matter, chlorophyll, dissolved oxygen, and nitrate (NO_3). For more details, see Xing et al. (2011). These data allow us to analyse a model's performance in reproducing key biogeochemical parameters (e.g., oxygen, nitrate, chlorophyll) and the vertical profile shapes that are the result of the interaction between physical and biogeochemical processes. Further, BGC-Argo data can be used to intercalibrate ocean colour variables such as chlorophyll, diffuse attenuation coefficient for a given downwelling irradiance wavelength, and turbidity, as proposed by the CMEMS Ocean Colour experts (http://octac.acri.fr/ and http://seasiderendezvous.fr/matchup.php). Fig. 29.8 shows an observed (i.e., a BGC-Argo float in the Balearic Sea) and forecasted (i.e. CMEMS MED-MFC model) vertical time series of chlorophyll concentration. Interestingly, we can see that the chlorophyll maximum occurs below 50 m most of the time, except during winter when mixed layer depth deepens. Obviously, such a subsurface maximum could not be observed through remote sensing, which means that ocean colour data are not (at least for some periods of the year) able to provide adequate information for assessing key biogeochemical processes. BGC-Argo profiles, once validated, would allow us to design more efficient metrics: the surface chlorophyll and timing of the surface blooms, as well as the total chlorophyll content in the top layers (i.e., photic layer or 0-200m), the deep chlorophyll maximum in the summer, and the depth of the vertically mixed bloom in the winter. Then, since vertical biogeochemical profile shapes are tightly linked to physical vertical processes, these novel metrics might help to identify possible mismatches of physical processes. Fig. 29.9 shows the nitrate comparison for the same model system with another BGC-Argo float in the western Mediterranean Sea. One can see, the shallowing of the nitracline is captured by the BGC-Argo float during vertical mixing winter events, but it stays in the 50-100 m depth layers the rest of the time. Here, metrics are used to compare nitrate concentration at the surface, nitracline depth, and the vertical integrated content, which is a measure of the potential fertilization occurring during winter mixing. Similar comparisons using the global CMEMS system BIOMER and BGC-Argo floats in the Labrador Sea (not shown here) allow measurements of dissolved oxygen, characterizing the winter ventilation, the oxygen uptake, and then the loss due to respiration processes along the water column.

The CMEMS strategy to assess the performance of biogeochemical systems focuses also on monitoring physical parameters that have a strong impact on the coupled biogeochemical model, and with "physical forcing" errors that may strongly alter the biogeochemical results. In particular, the vertical mixing near the surface and the mixed-layer depth changes used to evidence when erroneous chlorophyll concentration forecasts, might be caused by unrealistic surface dynamics behaviour.

Another emerging aspect is the uncertainty assessment of important essential climate variables such as pH and pCO_2. The CMEMS product accuracy can be indirectly estimated by evaluating prognostic model carbonate system variables (e.g., dissolved inorganic carbon and alkalinity) with historical datasets and climatology, or directly by using BGC-Argo floats data planned to measure also the pH.

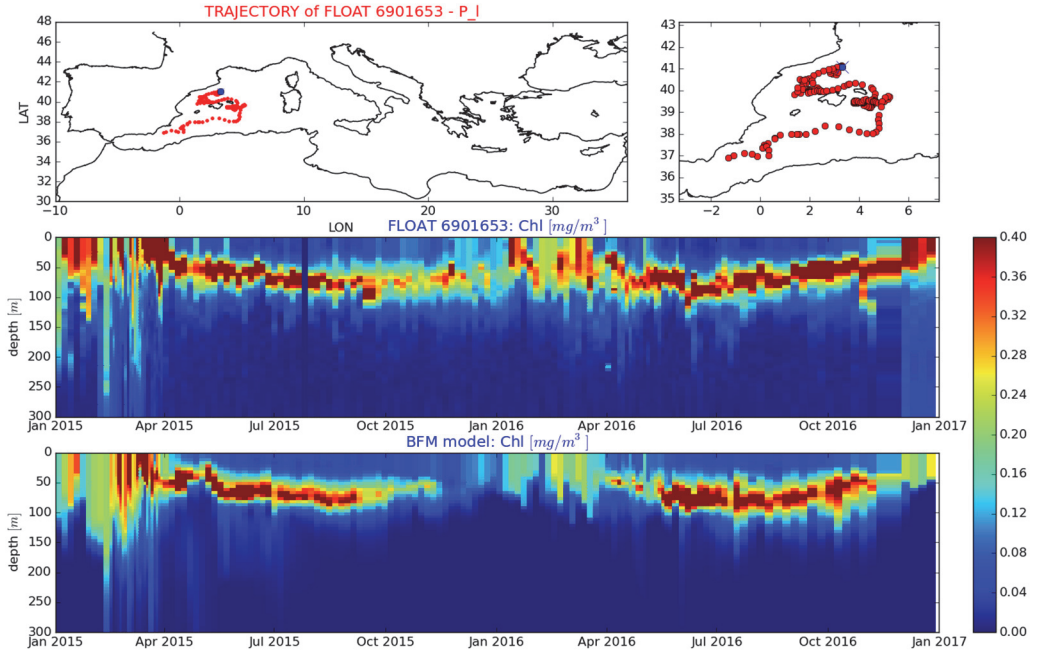


Figure 29.8. Time series of vertical chlorophyll profiles: comparison between the CMEMS Mediterranean biogeochemical system BFM (bottom) and a BGC-Argo float (middle panel), in mg/m3, in the Balearic Sea (float location at the top figures).

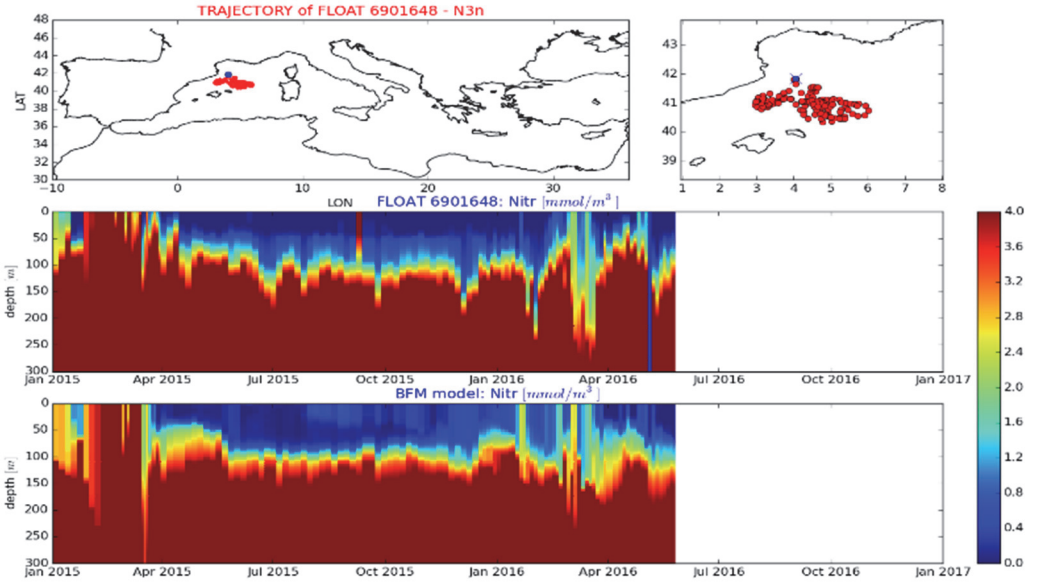


Figure 29.9. Time series of vertical profiles of nitrate: comparison between the CMEMS Mediterranean biogeochemical system BFM (bottom) and a BGC-Argo float (middle panel), in mmol/m3, in the Western Mediterranean Sea (float locations in the top figure).

Multi-model and ensemble assessment

For a given ocean parameter, OOFMS may provide a variety of estimates: observed values, gridded products based on observed values, gridded products from models corrected by assimilation (hindcast, nowcast), and forecasts. The OOFMS can also be designed to perform ensemble forecasts and to provide an estimate of forecast uncertainties through the spread of predicted values. In addition to real-time products, there are also offline products such as ocean reanalyses.

Consequently, the following ensemble approaches provide different ways to evaluate uncertainties and performance of systems:

- The operational systems based on ensemble assimilation allows us, through an ensemble of analyses, to explore the analyses error patterns in space and time (Hernandez et al., 2015).
- The ensemble forecast system, initiated from several analyses or perturbations of analyses, also allows us to explore statistically the forecast errors patterns and their variations over lead time (short- to medium-range) and over different seasons, etc. (Hernandez et al., 2015).
- The collection of estimates (hindcast, nowcast, or forecast) from several models and estimating techniques can be compared in order to characterise their uncertainties.

With the first two approaches, a preliminary work is typically carried out in order to provide users with the best options and the most relevant products from the ensemble, associated with some ways to understand uncertainty level deduced from the spread.

The third approach is becoming more relevant. In practice, in a given area, users are now in front of many ocean products for a requested parameter. Which implies that evaluation strategies need to 1/ characterise the accuracy of each product and estimate how close these products represent ocean “truth” and 2/ how relevant is a given product for a given application. Evaluating as a given ensemble several products allows to carry out the same metrics for all of them, compute an average of this ensemble, and characterise the departure of each product from this average. If product’s errors are non-correlated, the spread and the distance to this average offers an objective estimate of the relative quality of each product. If errors are correlated, the mean biases should be estimated separately, ideally with independent reference data. All these products may also offer different spatial and temporal spectral content. So it is important in the evaluation mentioned above to perform metrics that separate or filter out scales that are non-shared by the all products under evaluation. Hernandez et al. (2015) mentioned the consensus forecast estimation methods, where an ensemble average computed in an adapted way (e.g., clustering) offers increased accuracy compared to every individual member.

Operational centres are preparing to share forecasts and collaborate on ensemble assessments at global and regional scales. A good example of this already underway are the Baltic and North West Shelf CMEMS multi-system projects (http://www.boos.org, http://noos.eurogoos.eu/model-results/), which allow every contributing system to make its own evaluations of real-time departures from others and from the average.

Several ongoing real-time ocean monitoring initiatives are also using ensemble evaluations. For instance, on a monthly basis, the NOAA/NCEP ocean monitoring group gathers updates of

temperature from different operational global systems and uses spreading and signal-to-noise ratios to perform ocean analysis (Xue et al., 2017). This monitoring is a direct outcome of the Ocean Reanalyses Intercomparison Project and its focus is on thermal content (Balmaseda et al., 2015; Xue et al., 2012). Following this same strategy, the CMEMS is now providing several eddy-permitting ocean reanalyses as well as its suite of global ocean reanalysis multi-model ensemble products (GREP-V1; see http://marine.copernicus.eu/documents/QUID/CMEMS-GLO-QUID-001-026.pdf) average, similar to the SST-gridded products which are averaged into the Global Ocean Sea Surface Temperature Multi Product Ensemble (GMPE, Martin et al., 2012). Fig. 29.10 illustrates the evaluation of each reanalyses product as compared to observed values, averaged monthly. This example shows global salt content and the score of the GREP-V1 ensemble average as compared to the climatology. Clearly the GREP-V1 offers better data than the individual estimates that fall below the ensemble spread. It is noteworthy that in 2002, with addition of Argo data, discrepancies between each reanalyses salt content estimates were reduced. But the significant increase of salinity data (gray shading on right figure) does not appear to improve noticeably after 2008. A similar assessment can be performed for other model variables and can be carried out for derived quantities whenever reference quantities are available.

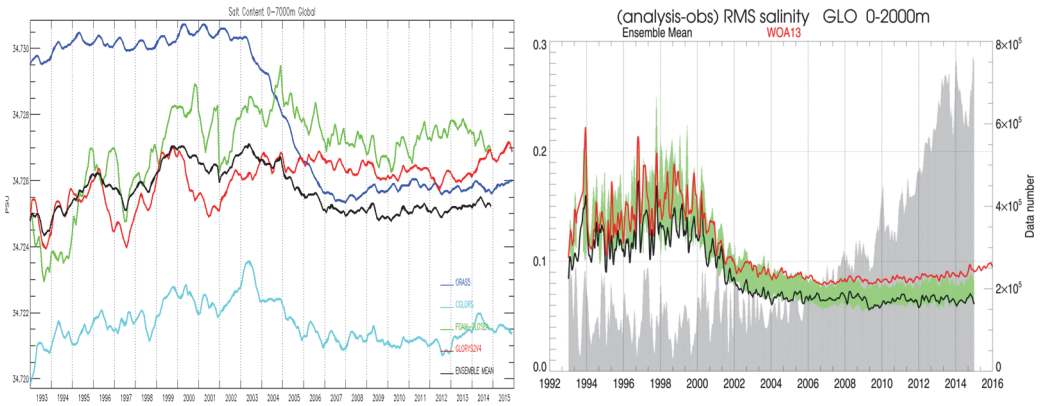


Figure 29.10. Left: Time evolution of the salt content [0-bottom] global average (psu) for the ensemble mean (black), GLORYS2V4 (red), ORAS5 (blue), CGLORS (cyan), FOAM (green) reanalyses. Right: time evolution of the root-mean-square differences with respect to observations from the Coriolis database, computed in the observations' space, between monthly ocean reanalyses estimates and daily observations for salinity in psu: for the ensemble mean GREP-V1 (black) and using the WOA 13 monthly climatology for the 2005-2012 decade (red). The green shading indicates the spread of the RMS of the four members of GREP-V1.

Concluding remarks

In support of wider user interest, particularly in the marine industry and among policy makers, operational oceanography is now able to deliver products on longer timescales and at finer space- and timescales, covering the "blue", the "green" and the "white" ocean. Validation and verification are core functions of operational centres, measuring the system performance, accuracy of ocean products, and the reliability of these products for targeted applications.

A review of validation and verification approaches designed and implemented by operational oceanography was proposed in Hernandez et al. (2015). Here, we update this review and discuss the primary principles and strategies followed in developing these evaluation tools.

Ocean observations are key elements of validation and verification approaches, although the sparseness of in situ data limits the robustness of this method. Another challenge to using observations is timeliness, which limits the capability of real-time validation in operational centres. Furthermore, operational products are now provided at higher resolution (i.e., full eddy-resolving and submesoscale representations at hourly frequencies). Observation representativity is thus an aspect that needs to be taken into account with regards to validation and verification practices. The good news is that the global observing system continue to become more effective: coastal networks of high-frequency radar integrated into coastal observing strategies are emerging; the use of gliders on regular routes is becoming more common; BGC-Argo floats are now equipped with more robust and reliable biogeochemical sensors; and opportunity measurements using sea mammals or other animals are taking place. At the same time, remote sensing capabilities are expanding: constellations are larger, and new sensors are coming online (for the future SWOT or SKIM satellite missions) that should generate new insights into surface circulation. Of course, these new datasets will not impact our ability to evaluate ocean reanalyses for the past. For that, we will need to rely on new techniques, such as multi-model assessments, to help us better understand errors and accuracy of historical ocean estimations.

Operational systems are also becoming more complex, coupling various models along the full range of causalities and mechanisms: atmosphere, wave and ocean dynamics, physical and biogeochemical processes, optical, biological, sea floor sedimentology, chemistry, etc. As such, the performance and robustness of these more complex systems must be able to pay attention to and monitor the efficiency of interfaces. Verification techniques are now systematically taking into account the information shared by the different compartments of an overall full system.

Evaluation of the accuracy of operational ocean products is also evolving, focusing more on user requests and areas of specific interest. Here, we introduced the idea of “external” assessment and metrics in contrast to the more classical “internal” test-bed academic evaluation of an ocean model’s performance. This new “external” or “user-oriented” metrics strategy need to be complemented by new and expanded ways of communicating the reliability of products to users given new technological options (e.g., smart phones) and the fact that the user community is more diverse, less academic, and more oriented toward societal decision-making.

The validation and verification activities in operational oceanography are maturing, first in terms of more structuration. Initiatives such as the CMEMS or GOV groups, which are associated with standardization mechanisms and the endorsement of best practices by the Joint Technical Commission for Oceanography and Marine Meteorology, demonstrate that operational centres are organizing and developing networks in order to leverage data and expertise, and to associate their efforts through multi-model assessment. These groups allow for the sharing of experiences with user interactions and expectations. They also function as a bridge between the oceanographic and other scientific communities, such as the atmospheric and weather prediction verification groups,

often resulting in the adoption if innovative approaches and metrics that are more user- or process-oriented.

Finally, this chapter provides an overview of the validation and verification of operational system performance evaluation. As discussed, it is important to keep in mind that validation techniques and metrics must be continually revisited in order to become more robust in characterizing products accuracy and reliability.

Acknowledgements

The authors want to thank the GODAE OceanView Intercomparison and Validation Task Team members as well as the CMEMS Product Quality Working Group experts for the valuable exchanges and discussions on metrics development and validation/verification approaches over the years. Work on this review is partly funded by Mercator Océan, as part of the CMEMS activities.

References

Balmaseda, M. A., and Coauthors, 2015: The Ocean Reanalyses Intercomparison Project (ORA-IP). Journal of Operational Oceanography, 8:sup1, s80-s97. doi:10.1080/1755876X.2015.1022329

Bell, M. J., A. Schiller, P.-Y. Le Traon, N. R. Smith, E. Dombrowsky, and K. Wilmer-Becker, 2015: An introduction to GODAE OceanView. Journal of Operational Oceanography, 8, s2-s11, doi: 10.1080/1755876X.2015.1022041.

Blockley, E. W., and Coauthors, 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, 2613-2638, doi:10.5194/gmd-7-2613-2014.

Bouillon, S., P. Rampal, and E. Olason, 2018: Sea ice modelling and forecasting. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Brassington, G. B., and Coauthors, 2015: Progress and challenges in short- to medium-range coupled prediction. Journal of Operational Oceanography, 8, s239-s258. doi:10.1080/1755876X.2015.1049875.

Casati, B., and Coauthors, 2008: Forecast verification: current status and future directions. Meteorological Applications, 15, 3-18. doi:10.1002/met.52.

Coppini, G., 2017: New operational tools at sea: www.sea-conditions.com. Personal Communication.

De Mey, P., E. Stanev, and V. H. Kourafalou, 2017: Science in support of coastal ocean forecasting—part 1. Ocean Dynamics, 67, 665-668, doi:10.1007/s10236-017-1048-1

Divakaran, P., and Coauthors, 2015: GODAE OceanView Class 4 inter-comparison for the Australian Region. Journal of Operational Oceanography, 8:sup1, s112-s126, doi:10.1080/1755876X.2015.1022333

Drévillon, M., and Coauthors, 2013: A strategy for producing refined currents in the Equatorial Atlantic in the context of the search of the AF447 wreckage. Ocean Dynamics, 63, 63-82, doi:10.1007/s10236-012-0580-2

Drévillon, M., and Coauthors, 2018: Learning about Copernicus Marine Environment Monitoring Service “CMEMS”: a practical introduction to the use of the European operational oceanography service. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Dufau, C., M. Orsztynowicz, G. Dibarboure, R. Morrow, and P.-Y. Le Traon, 2016: Mesoscale resolution capability of altimetry: Present and future. Journal of Geophysical Research: Oceans, 121, 4910-4927, doi: 10.1002/2015JC010904.

Ebert, E., and Coauthors, 2013: Progress and challenges in forecast verification. Meteorological Applications, 20, 130-139, doi:10.1002/met.1392.

Ford, D., S. Kay, R. McEwan, I. Totterdell, and M. Gehlen, 2018: Marine biogeochemical modelling and data assimilation for operational forecasting, reanalysis and climate research. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Gehlen, M., and Coauthors, 2015: Building the capacity for forecasting marine biogeochemistry and ecosystems: recent advances and future developments. Journal of Operational Oceanography, 8, s168-s187, doi:10.1080/1755876X.2015.1022350.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. Weather Forecast., 24, 1416-1430, doi:10.1175/2009waf2222269.1.

Gordon, N. D., 1982: Evaluating the Skill of Categorical Forecasts. Monthly Weather Review, 110, 657-661, doi:10.1175/1520-0493(1982)110<0657:etsocf>2.0.co;2.

Griffin, D. A., and P. R. Oke, 2017: The search for MH370 and ocean surface drift – Part III. (Report number EP174155), ed. by CSIRO Oceans and Atmosphere, Australia. [Available at https://www.atsb.gov.au/media/5773371/mh370_csiro-ocean-drift-iiil.pdf.]

Griffin, D. A., P. R. Oke, and E. Jones, M., 2016: The search for MH370 and ocean surface drift – Part II. (Report number EP167888), ed. by CSIRO Oceans and Atmosphere, Australia. [Available at https://www.atsb.gov.au/media/5772119/mh370_ocean_driftv29.pdf.]

Haines, K., 2018: Ocean Reanalysis. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Harris, C. M., 2018: Coupled atmosphere-ocean modelling. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Hernandez, F., and A. Melet, 2016: Product Quality Strategic Plan. in CMEMS, (CMEMS-PQ-StrategicPlan), ed. by Mercator Océan, Toulouse.

Hernandez, F., and Coauthors, 2009: Validation and intercomparison studies within GODAE. Oceanography Magazine, 22, 128-143, doi:10.5670/oceanog.2009.71.

Hernandez, F., and Coauthors, 2015: Recent progress in performance evaluations and near real-time assessment of operational ocean products. Journal of Operational Oceanography, 8, s221-s238, doi: 10.1080/1755876X.2015.1050282.

Jacobs, G. A., and Coauthors, 2018: Operational Ocean Data Assimilation. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Kourafalou, V. H., and Coauthors, 2015: Coastal Ocean Forecasting: system integration and evaluation. Journal of Operational Oceanography, 8, s127-s146, doi:10.1080/1755876X.2015.1022336.

Le Sommer, J., E. Chassignet, and A. Wallcraft, 2018: Ocean circulation modelling for operational oceanography: current status and future challenges. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Le Traon, P.-Y., 2013: From satellite altimetry to Argo and operational oceanography: three revolutions in oceanography. Ocean Sci., 9, 901-915, doi:10.5194/os-9-901-2013.

Le Traon, P.-Y., and Coauthors, 2017: The Copernicus Marine Environmental Monitoring Service: Main Scientific Achievements and Future Prospects. Mercator Ocean Journal, 56, 100.

Lellouche, Jean-Michel, and E. Greiner, 2018: The Mercator Ocean Global High Resolution Monitoring and Forecasting System. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Lellouche, J.-M., and Coauthors, 2013: Evaluation of global monitoring and forecasting systems at Mercator Océan. Ocean Sci., 9, 57-81, doi:10.5194/os-9-57-2013.

Maraldi, C., and Coauthors, 2013: NEMO on the shelf: assessment of the Iberia-Biscay-Ireland configuration. Ocean Sci., 9, 745-771, doi:10.5194/os-9-745-2013

Martin, M., and Coauthors, 2012: Group for High Resolution Sea Surface temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE). Deep Sea Research Part II: Topical Studies in Oceanography, 77-80, 21-30, doi:10.1016/j.dsr2.2012.04.013.

Melsom, A., S. Eastwood, J. Xie, S. Aaboe, and L. Bertino, 2017: Challenges in validating model results for first year ice. EGU 2017 General Assembly, EGU.

Morrow, R. A., D. Blumstein, and G. Dibarboure, 2018: Fine-scale altimetry and the future SWOT mission. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Mourre, B., and Coauthors, 2018: Assessment of high-resolution regional ocean prediction systems using multi-platform observations: illustrations in the Western Mediterranean Sea. GODAE Oceanview International School in “New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Murphy, A. H., 1993: What is a Good Forecast - An essay on the nature of goodness in weather forecasting. Weather Forecast, 8, 281-293, doi:10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2.

Oke, P. R., G. B. Brassington, J. A. Cummings, M. J. Martin, and F. Hernandez, 2012: GODAE inter-comparisons in the Tasman and Coral Seas. Journal of Operational Oceanography, 5, 11-24.

Oke, P. R., and Coauthors, 2015a: Assessing the impact of observations on ocean forecasts and reanalyses: Part 1, Global studies. Journal of Operational Oceanography, 8, s49-s62, doi: 10.1080/1755876X.2015.1022067.

Oke, P. R., and Coauthors, 2015b: Assessing the impact of observations on ocean forecasts and reanalyses: Part 2, Regional applications. Journal of Operational Oceanography, 8, s63-s79, doi: 10.1080/1755876X.2015.1022080.

Rogé, M., R. Morrow, C. Ubelmann, and G. Dibarboure, 2017: Using a dynamical advection to reconstruct a part of the SSH evolution in the context of SWOT, application to the Mediterranean Sea. Ocean Dynamics, 67, 1047-1066, doi:10.1007/s10236-017-1073-0.

Roughan, M., C. Kerry, and P. McComb, 2018: Shelf and Coastal Ocean Observing and Modelling Systems – A New Frontier for Operational Oceanography. GODAE Oceanview International School in "New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Ryan, A. G., and Coauthors, 2015: GODAE OceanView Class 4 forecast verification framework: Global ocean inter-comparison. Journal of Operational Oceanography, 8:sup1, s98-s111, doi: 10.1080/1755876X.2015.1022330

Schiller, A., F. Davidson, P. M. DiGiacomo, and K. Wilmer-Becker, 2016: Better Informed Marine Operations and Management: Multidisciplinary Efforts in Ocean Forecasting Research for Socioeconomic Benefit. Bul. Amer. Met. Soc., 97, 1553-1559, doi:10.1175/bams-d-15-00102.1.

Schiller, A., and Coauthors, 2015: Synthesis of new scientific challenges for GODAE OceanView. Journal of Operational Oceanography, 8, s259-s271, doi:10.1080/1755876X.2015.1049901.

Smith, G. C., and Coauthors, 2016: Sea ice forecast verification in the Canadian Global Ice Ocean Prediction System. Quarterly Journal of the Royal Meteorological Society, 142, 659-671. doi: 10.1002/qj.2555.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106, 7183-7192, doi:10.1029/2000JD900719.

Tian-Kunze, X., L. Kaleschke, N. Maaß, M. Mäkynen, N. Serra, M. Drusch, and T. Krumpen, 2014: SMOS-derived thin sea ice thickness: algorithm baseline, product specifications and initial verification. The Cryosphere, 8, 997-1018, doi:10.5194/tc-8-997-2014.

Tonani, M., and Coauthors, 2015: Status and future of global and regional ocean prediction systems. Journal of Operational Oceanography, 8, s201-s220, doi:10.1080/1755876X.2015.1049892

Ubelmann, C., P. Klein, and L.-L. Fu, 2015: Dynamic Interpolation of Sea Surface Height and Potential Applications for Future High-Resolution Altimetry Mapping. Journal of Atmospheric & Oceanic Technology, 32, 177, doi:10.1175/JTECH-D-14-00152.1.

Volpe, G., B. Buongiorno Nardelli, S. Colella, and R. Santoleri, 2018: An Operational Interpolated Ocean Colour Product in the Mediterranean Sea. GODAE Oceanview International School in "New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

von Schuckmann, K., and Coauthors, 2016: The Copernicus Marine Environment Monitoring Service Ocean State Report. Journal of Operational Oceanography, 9, s235-s320, doi: 10.1080/1755876X.2016.1273446.

Wilkin, J. L., J. Levin, A. Lopez, H. Arango, E. J. Hunter, and J. Zavala-Garay, 2018: A coastal ocean forecast system for U.S. Mid-Atlantic Bight and Gulf of Maine. GODAE Oceanview International School in "New Frontiers in Operational Oceanography", E. P. Chassignet, A. Pascual, J. Tintore, and J. Verron, Eds.

Xing, X., A. Morel, H. Claustre, D. Antoine, F. D'Ortenzio, A. Poteau, and A. Mignot, 2011: Combined processing and mutual interpretation of radiometry and fluorimetry from autonomous profiling Bio-Argo floats: Chlorophyll a retrieval. Journal of Geophysical Research: Oceans, 116, doi: 10.1029/2010JC006899.

Xu, Y., and L.-L. Fu, 2012: The Effects of Altimeter Instrument Noise on the Estimation of the Wavenumber Spectrum of Sea Surface Height. J. Phys. Oceanogr., 42, 2229, doi:10.1175/JPO-D-12-0106.1.

Xue, Y., and Coauthors, 2012: A Comparative Analysis of Upper-Ocean Heat Content Variability from an Ensemble of Operational Ocean Reanalyses. J. Climate, 25, 6905-6929. doi:10.1175/jcli-d-11-00542.1.

Xue, Y., and Coauthors, 2017: A real-time ocean reanalyses intercomparison project in the context of tropical pacific observing system and ENSO monitoring. Climate Dynamics, 49, 3647-3672, doi:10.1007/s00382-017-3535-y.

Zhu, X., and Coauthors, 2016: Comparison and validation of global and regional ocean forecasting systems for the South China Sea. Nat. Hazards Earth Syst. Sci., 16, 1639-1655. doi: 10.5194/nhess-16-1639-2016.

Zuo, H., M. Balmaseda, and K. Mogensen, 2015: The new eddy-permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals. Climate Dynamics, 1-21, doi:10.1007/s00382-015-2675-1.