State Planet, 5-opsr, 17, 2025
https://doi.org/10.5194/sp-5-opsr-17-2025

A description of model intercomparison processes and techniques for ocean forecasting

Fabrice Hernandez[1], **Marcos Garcia Sotillo**[2], **and Angélique Melet**[3]

[1]Laboratoire d'Études en Géophysique et Océanographie Spatiales (LEGOS), Institut de Recherche pour le Développement (IRD) et Université de Toulouse, CNRS, CNES, Toulouse, France
[2]NOW Systems (Nologin Oceanic Weather Systems), Santiago de Compostela, Spain
[3]Mercator Ocean International, Toulouse, France

Correspondence: Fabrice Hernandez (fabrice.hernandez@ird.fr)

Received: 1 October 2024 – Discussion started: 7 October 2024
Revised: 29 January 2025 – Accepted: 30 January 2025 – Published: 2 June 2025

Abstract. The availability of numerical simulations for ocean dynamics past estimates or future forecast worldwide at multiple scales is opening new challenges in assessing their realism and predictive capacity through an intercomparison exercise. This requires a huge effort in designing and implementing a proper assessment of models' performances, as already demonstrated by the atmospheric community that was pioneering in that sense. Historically, the ocean community has only recently launched dedicated actions aimed at identifying robust patterns in eddy-permitting simulations: it required definition of modelling configurations, execution of dedicated experiments that also deal with the storing of the outputs and the implementation of evaluation frameworks. Starting from this baseline, numerous initiatives like WCRP/Climate Variability and Predictability (CLIVAR) for climate research and the Global Ocean Data Assimilation Experiment (GODAE) for operational systems have arisen and are actively promoting best practice through specific intercomparison tasks, aimed at demonstrating the efficient use of the Global Ocean Observing System and its operational capabilities, sharing expertise, and increasing the scientific quality of the numerical systems. Examples, like the Ocean Reanalysis Intercomparison Project (ORA-IP) or the Class 4 near-real-time GODAE intercomparison, are introduced and commented on, also discussing ways forward for making this kind of analysis more systematic using artificial intelligence approaches for addressing monitoring of ocean state in operations or facilitating in-house routine verification in ocean forecasting centres.

1 Historical development of model intercomparisons

Historically, in oceanography, model comparisons began with evaluations of "free" and "forced" numerical simulations of ocean circulation over the same space and time frames, assessing their differences within comparable situations. The international Atmospheric Model Intercomparison Project (AMIP), under the World Climate Research Programme (WCRP), played a pioneering role in guiding the oceanic modelling community (Gates, 1992). AMIP's primary objective was to comprehensively evaluate each model's performance and document systematic errors. From an academic standpoint, this intercomparison aimed to identify avenues for enhancing future atmospheric models and driving further developments. Consequently, this approach aligns clearly with the validation framework outlined in Sotillo et al. (2025, in this report). To provide an objective assessment of each "competing" model's performance, a common "reference truth" was selected, such as climatology or atmospheric reanalysis (deemed more realistic than the AMIP simulations). This process involved analysing a series of targeted key variables extracted from the model state to provide an overview of the model's skill in representing various atmospheric aspects.

In 1996, the same atmospheric community, involved in climate studies, settled the basis of the Coupled Model Intercomparison Project (CMIP) under the auspice of the WCRP/Climate Variability and Predictability (CLIVAR) panel to document systematic errors of global couple climate simulations in support of the Intergovernmental Panel on Climate Change (IPCC) framework (Meehl et al., 1997). Over the six phases of the CMIP, intercomparisons have refined the assessments, increasingly including the physical, biochemical and ecosystem components of the Earth system, by testing various climate scenarios of past, present and future CO_2 emissions. In the current phase, the CMIP6, the variety of models, simulations and their objectives have led the community to redefine the federated structure through a common specific framework, the Diagnostic, Evaluation and Characterization of Klima (DECK) experiments, which set out the simulations and scientific questions to be addressed. The DECK is the new acceptance criterion for a climate intercomparison project in the CMIP (Eyring et al., 2016). The evolution of the CMIP has been accompanied by the gradual adoption by the climate community of common standards, coordination, infrastructure and documentation, accessible to all. This persistent framework aims to ensure continuity in climate model performance assessment of future CMIP phases in which re-processed historical simulations defined in the AMIP would allow changes and benefits of more elaborated components of the Earth system models (ESMs).

The ocean modelling research community adopted a similar approach to the AMIP when the first global- or basin-scale eddy-permitting ocean simulations were achieved in the 1990s. The US–German Community Modelling Effort (CME), in support of the World Ocean Circulation Experiment (WOCE), started to infer model parametrization and sensitivity studies in modelling the North Atlantic basin (Böning and Bryan, 1996). Sources of errors like ocean boundaries or vertical mixing parametrization were identified. The DYNAMO project, dedicated to offering intercomparison among three classes of ocean models of the North Atlantic Ocean in a similar numerical experiment framework (Meincke et al., 2001), allowed patterns of the North Atlantic Ocean circulation to be identified that were robust and other patterns that were sensitive to model parametrization. In this case, the intercomparison approach brought another benefit than just identifying performances among the simulations: the common and matching patterns represented by the simulations were considered updated knowledge of the North Atlantic circulation. In other terms, the "ensemble pattern" from the simulations is identified as a robust representation of the "ocean truth" at the scales simulated by these models.

This first initiative led to the development of a common ocean modelling framework from the ocean community also involved in the CMIP projects, the Coordinated Ocean-ice Reference Experiments (COREs), aiming to provide common references for consistent assessment from a multi-model perspective (Griffies et al., 2009). CORE-I intends to evaluate model mean biases under a normal year forcing, using a prescribed series of metrics (e.g. Danabasoglu et al., 2014). The CORE-II framework extends the ocean model evaluation under the common interannual forcing – starting in 1948 – proposed initially by Large and Yeager (2009). It offers more direct comparison to ocean observations and to the effective ocean interannual variability. An intercomparison of 18 time-dependent ocean numerical simulations have been performed so far, with useful outcomes for global ocean model improvements. The CORE-II approach is the foundation of the Ocean Model Intercomparison Projects (OMIPs) carried out in support of the successive CMIPs, with a coordinated evaluation of the ocean, sea ice, tracer and biogeochemistry simulations forced by common atmospheric datasets (Eyring et al., 2016). The OMIP version 1 contribution to CMIP6, with ocean simulations' intercomparisons over the 1948–2009 period, is described by Griffies et al. (2016) and contains a comprehensive list of metrics and guidance to evaluate ocean–sea ice model skills as part of ESMs. A companion article by Orr et al. (2017) proposes the evaluation framework for the biogeochemical coupled model simulations in CMIP6. Under the CLIVAR Ocean Model Development Panel (OMDP) coordination, OMIP version 2 is ongoing using the more recent JRA-55 reanalysis forcings (Kobayashi et al., 2015). Metrics of the ocean (equivalent here to diagnostics) endorsed by the OMIP are those recommended for the assessment of ocean climate behaviour, impacts and scenarios in the CMIP DECK.

These first ocean intercomparison projects witness the community effort, trying to commonly define modelling strategies; conduct the simulations individually; and then intercompare the simulations in order to evaluate the model's performance with regard to observed realistic references. The projects bring better characterization of model errors and weaknesses considering specific ocean processes, from physical to biogeochemical aspects, over decadal, interannual and seasonal timescales. Implicitly, these efforts have involved strategies for distributing, storing and sharing simulations and metrics, under constraints of computer server limitations in capacity and communication bandwidth. In practice, this added to the common technical definition of standards shared by all participants and a fitness-for-purpose evaluation framework to be applied in similar ways for every simulation. And finally, a common synthesis effort is carried out in order to provide valuable conclusions.

The first intercomparison project that involved the operational oceanography has been carried out in the frame of the CLIVAR Global Synthesis and Observation Panel (GSOP). In practice, this involved intercomparing different ocean reanalyses computed over several decades and providing "ocean synthesis" on ocean state estimation through a chosen series of essential ocean variables (EOVs) considered in climate research (Stammer et al., 2009). A step was taken since it was no longer comparison of model outputs but of products issued from the more complex system producing each reanalysis (observation + model + assimilation), increasing the factors of discrepancies among them. The idea is that multi-system ensemble approaches should be useful to obtain better estimates of the ocean evolution. The GSOP objectives were (1) to assess the consistency of the synthesis through intercomparison; (2) to evaluate the accuracy of the products, possibly by comparison to observations; (3) to estimate uncertainties; (4) to identify areas where improvements were needed; (5) to evaluate the lack of assimilated observations that directly impacted the synthesis and propose future observational requirements; and (6) to work on new approaches, like coupled data assimilation. One of the outcomes was to highlight common behaviour among some products, that is, evidence "clusters" and correlated patterns that sometimes had just inappropriate biases.

In the atmospheric and weather-forecast side, usually responsible for marine meteorology predictions, routine intercomparison for wave forecast has been settled for many years under the World Meteorological Organization (WMO) framework. The European Centre for Medium-Range Weather Forecasts (ECMWF) hosts the ongoing WMO Lead Centre for Wave Forecast Verification where 18 regional and global wave forecast systems are compared (https://confluence.ecmwf.int/display/WLW, last access: 29 January 2025). Beyond wave forecasts' verification and quality monitoring, the ECMWF commits to maintaining an archive of the verification statistics to allow the generation and display of trends in performance over time.

A first dedicated intercomparison of ocean operational systems, operated on routine, was achieved by the Global Ocean Data Assimilation Experiment (GODAE) community (Bell et al., 2009), through an intercomparison of hindcasts over 2008. The main objectives were to (a) demonstrate GODAE operational systems in operations, (b) share expertise and design validation tools and metrics endorsed by all GODAE operational centres, and (c) evaluate the overall scientific quality of the different GODAE operational systems. The preliminary task was to define the validation concepts and methodologies (Hernandez et al., 2015a), with the so-called "Class 1 to 4 metrics" described in this report (Sotillo et al., 2025), and those directly inherited from the weather forecast verification methods (Murphy, 1993). A demanding task was to provide similar "Class 1", "Class 2" and "Class 3" files from each Operational Ocean Forecasting System (OOFS) and then to carry out the evaluation through intercomparison and validation against "truth references" (Hernandez, 2011).

2 Key findings for state-of-the-art model intercomparison

2.1 From academia to operation: adoption of best practice

The legacy of the first 10 years of GODAE was the implementation of an expert community for OOFS intercomparison: the Intercomparison and Validation Task Team (IVTT). This group was created during GODAE, continuing its activity in GODAE OceanView and, up to present day, in Ocean Predict (https://oceanpredict.org/, last access: 29 January 2025). A second benefit was the development of an ad hoc validation and intercomparison methodology, improved and tested regularly since, until it was adopted as best practice and recommended by the Expert Team on Operational Ocean Forecasting Systems (ETOOFS; Alvarez-Fanjul et al., 2022).

As a result of these activities, it was found that performing intercomparison of OOFSs and models brought the following aspects to address:

- Characterize the performance of individual OOFSs of the same kind relatively to a given "truth", identify outliers and give clues for further OOFS improvements.

- Allow "ensemble estimation" that provides qualitatively more robust and reliable estimates, i.e. the "ensemble mean" approach. In some cases, if the "ocean truth" is missing, the ensemble mean can be considered a reference and be used to validate individually the systems.

- Provide an ad hoc methodology for operational qualification; see Sotillo et al. (2025) for detailed explanation on OOFS qualification or "calibration". In other words, when the OOFS is upgraded, inter-comparing the old and new systems informs on the benefits of the upgrade and justifies "go/no-go" decisions.

- Adopt or refine technologies supporting large exchanges of information among the community: in this sense, the NetCDF file format and climate forecast standardization has greatly facilitated the "shareability" (Hernandez et al., 2015a, b) and prefigured the FAIR best practice (Findability, Accessibility, Interoperability and Reuse of digital assets), proposed more recently (Wilkinson et al., 2016).

An exceptionally illustrative intercomparison example emerged from the tragic crash of the Rio de Janeiro–Paris Air France plane in 2009 and the subsequent intensive search for the wreckage in the tropical Atlantic. Evaluation of the accuracy of current fields from OOFSs and observed products, a user-centric approach based on dispersion and Lagrangian metrics, was employed within an intercomparison framework. It was demonstrated that the ensemble mean yielded more reliable results compared to individual estimates (Drévillon et al., 2013). A similar approach was also adopted to identify the crash area for the March 2014 Malaysia Airlines flight MH370 in the Indian Ocean (Griffin and Oke, 2017; Durgadoo et al., 2021).

2.2 Intercomparison: key aspects to be addressed

Intercomparing routinely or during specific phases OOFSs and their products is now common practice in operational centres. However, various aspects need to be reiterated and addressed:

- Common validation/verification methodology needs to be adopted by all participants, preferably adopting recommendations, as reiterated in this report (Sotillo et al., 2025).
- Interoperability, shareability of products and common standards are key: the large number of products offered by the different centres cannot be spread in every single centre. The FAIR principles of best practice are essential.
- Representativity is a central aspect of intercomparison: scales and ocean processes represented in each product (observations and models) need to be correctly documented to reduce mis-interpretation when intercompared. In particular, the following points should be noted.
  - Re-gridding by downscaling or upscaling ocean products toward a common grid might generate errors and not conservative effects of ocean dynamics.
  - Comparing ocean re-gridded products with re-gridded observations containing different ocean scales might create double penalty scores.
  - Due to operational oceanography growing activity, it is worth remembering that an increasing number of products are available for each EOV, for each area. The Copernicus Marine Environment Monitoring Service (CMEMS) Data Store is a good illustration of this, with a large number of products derived from models or from space or in situ observations for a given EOV. This reinforces the importance of an a priori assessment of the representativity of each product before any intercomparison.
- Intercomparison is a first path toward ocean state estimation from various sources and products: it is promising to use novel approaches based on data mining, consensus clustering, machine learning, and other tools developed in the frame of ensemble estimation and forecast (e.g. Sonnewald et al., 2021).
- User-oriented metrics and process-oriented metrics are increasingly being implemented in operational centres. They are also new insight for establishing the performance of intercompared OOFSs into the user-oriented framework.

3 Ongoing ocean models and forecasting systems intercomparison activities

3.1 Class 4 metrics: model intercomparison in the observation space for verification forecast

Ocean observations provide an accurate estimation of the "ocean truth". However, the Global Ocean Observing System (GOOS) provides a sparse representation over time of three-dimensional ocean dynamics. Their quantity and quality have increased substantially with permanent mooring and programmes such as Argo and the Global Drifter Program, together with satellite measurements (e.g. Tanhua et al., 2019). The GOOS is providing these recent years a valuable representation of the large-scale dynamics and some aliased representation of the ocean fine scales where measurements are performed. This led to the evaluation of OOFS performance by direct comparison with observations and to the definition of the Class 4 metrics detailed in Sotillo et al. (2025).

In summary, Class 4 metrics aims to compare observations with the equivalent model forecast at the same time and place, for different lead times (Hernandez et al., 2015a). Thus, these metrics, for different kinds of ocean variables, characterize the performance of a given OOFS against observations in the observation space. One of the advantages of using the observations as the reference frame is that other OOFSs can similarly be compared to the same data, in the same manner. Hence Class 4 metrics have been used since the beginning when comparing several OOFSs and their performance with the same "truth" (Hernandez et al., 2015a). When the observations are not assimilated by the OOFS, one can get a fully independent error assessment that can be statistically representative of the overall quality of the OOFS. Otherwise, one can consider that the overall error level is underestimated. However, this still provides an objective measure of the actual gap between the OOFS estimate and the "ocean truth" at the exact location and time of the observation used as reference.

Within GODAE OceanView, the Class 4 intercomparison project has been operating since 2013. A first set of intercomparison of six global OOFSs (Ryan et al., 2015) was an opportunity to present new metrics (radar plot, Taylor diagrams, best system mapping, bar charts, rank histograms, etc.). The same Class 4 information was also used with more specific metrics around Australia (Divakaran et al., 2015), with the objective of the Australian Bureau of Meteorology to identify a path of improvements for its own OOFS. This was also a first demonstration of one of the benefits of such intercomparison: the in-house routine validation in Australia took advantage of the internationally shared and compared multi-system Class 4 information to enhance its own daily basis verification procedures. The Class 4 intercomparison is still routinely performed (Fig. 1) and is continuously extended. A recent intercomparison based on Class 4 for surface velocity using drifters by Aijaz et al. (2023) offers an additional evaluation of OOFS surface dynamics performance, key for applications like search and rescue, marine pollution forecasts, and many other drift-dependent applications.

Another issue of Class 4 comparison to observations was the routine evaluation of the overall quality of the GOOS. Performing comparisons with observations of several OOFSs also gives more confidence in identifying observation outliers and incorrect measurements: a feedback procedure was proposed to inform data centres that could carry out a second loop of data corrections, for the benefit of all data users (Hernandez et al., 2015b). This approach is now considered in the frame of the recent project SYNOBS endorsed by the United Nations Ocean Decade programme (Fujii et al., 2019, 2024). SYNOBS aims at evaluating the best combinations of ocean-observing platforms through observing system design carried out by different operational centres (e.g. Balmaseda et al., 2024a). The existing intercomparison framework will allow faster common assessment among the different contributors.

Mentioned above, comparison to observations raises the key issue of representativity, both from the observation and the modelling side. And subsequently, double penalty effects must be taken into account when measuring the skill of a given product for given scales or ocean regimes. It is necessary to carefully address the following questions: what are the scales sampled by a given observing system? What are the effective scales and ocean processes represented by a given OOFS? What ocean processes do they represent? The classical example is comparison of satellite altimetry and/or tide gauge observations with the sea surface height given by an OOFS: if the latter does not represent the tidal dynamics, obviously, observations need to be pre-processed to filter out tidal signals. This is the reason that the concept of “internal” metrics, aiming to measure the efficiency of the OOFS at the expected scales, was distinguished from the concept of “external” metrics, where operational products’ reliability and fitness for purpose need to be assessed in the light of the user’s requirements (Hernandez et al., 2018) and taken into account while performing intercomparisons. In addition, particular attention needs to be given to the representativity and uncertainty of observations. It is mandatory to take them into account while comparing several OOFSs with observations, in particular when referring to re-processed/re-gridded observation products (also called Level 4 or L4 type of observed products).

3.2 Ensemble forecast comparison: assessment through ensemble mean, ensemble spread and clusters

The atmospheric community developed ensemble forecasts, first to represent uncertainties of seasonal predictions considering the stochastic behaviour of atmospheric simulations. This was done using an individual forecasting system, by running a series of deterministic forecasts in parallel where some initial or forcing conditions were stochastically modified between members. The purpose of performing the intercomparison of the forecast members was to (1) identify common patterns from the probability distribution for eventually defining clusters, (2) compute probabilistic occurrences of specific events, and (3) use the ensemble spread as a proxy for forecast skill and performance assessment and try to separate outliers. The associated verification framework has been largely documented (e.g. Casati et al., 2008) and defined for the atmospheric components of the seasonal forecast activities (e.g. Coelho et al., 2019). For the ocean environment, this approach is currently used by weather prediction centres in charge of marine meteorology forecasting, i.e. wind and wave forecasts. For instance, the evaluation exercise performed by the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP), evaluating ensemble and deterministic forecasts, concluded, among other results, that the ensemble wave skill score at day 10 outperformed the deterministic one at day 7 (Campos et al., 2018). Another example is the recent intercomparison of seasonal ensemble forecasts from two centres contributing to the Copernicus Climate Change Service (C3S), which quantified their respective skill on sea surface height, ocean heat content and sea surface temperature (Balmaseda et al., 2024b).

At this stage, unlike weather prediction centres, ensemble forecasting from individual systems is not generalized in operational oceanography, although dedicated experiments are carried out in many areas (e.g. Pinardi et al., 2011; Schiller et al., 2020). And through specific data assimilation methods like the ensemble Kalman filter (Evensen, 2003), several centres are producing ensemble forecasts routinely (e.g. Lisæter et al., 2003; Keppenne et al., 2008; Seo et al., 2009). However, a large community effort dedicated to intercomparisons of ensemble forecasts produced by different centres has not yet been achieved.

Here we propose to illustrate ensemble approach benefits with a multi-system intercomparison as proposed by the CLIVAR/GSOP initiative (mentioned above) and the Ocean Reanalysis Intercomparison Project (ORA-IP) (detailed in Sect. 3.4 below and also discussed by Storto et al., (2019)). Figure 2 illustrates the assessment of a commonly used indicator for the so-called “Atlantic Niño” regimes in the tropical Atlantic, associated with the “Atlantic zonal mode” and targeting the equatorial cold tongue that develops in the Gulf of Guinea from April to July (Vallès-Casanova et al., 2020).

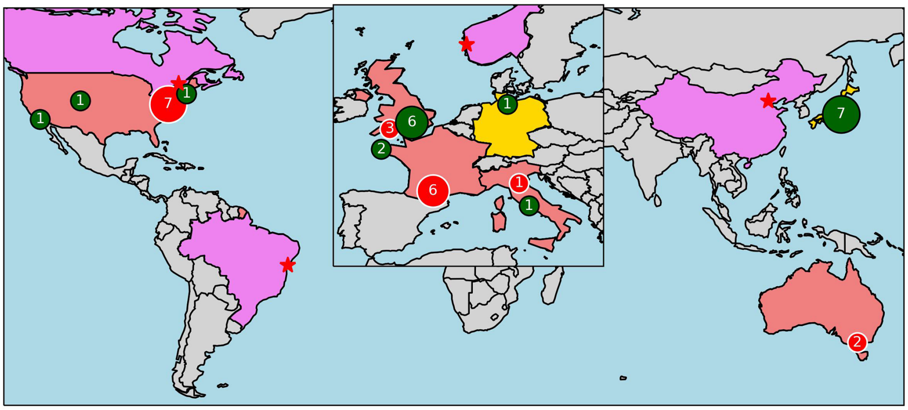

Figure 1. Operational centres and countries involved in a common intercomparison international framework during the last 20 years. Circles indicate their size and numbers the products/locations participating in the ORA-IP (Balmaseda et al., 2015). Green circles for ORA-IP only and red circles for centres that are contributing in addition to the Class 4 routine intercomparison (Hernandez et al., 2015a). Red stars indicate centres solely participating in the Class 4 intercomparison. Countries in pink, yellow and orange contribute, respectively, to Class 4, ORA-IP and both exercises.

All products –observation-derived-only and reanalysis estimates (see Balmaseda et al., 2015, for products' details) – give a consistent representation of the seasonal and interannual variability, from which an interannual trend can be deduced over the 1980–2024 period (ensemble-average trend in Fig. 2c of 0.02 °C per year). The ensemble average is computed like the multi-product-mean in Uotila et al. (2018) and without ARMOR3D, the observation-derived-only product used as "ground truth" (Guinehut et al., 2012), and without the GREP reanalysis, already an ensemble average of various reanalyses (Masina et al., 2015). Figure 2b shows the time series of the so-called SST (sea surface temperature) index: the box-averaged temperature anomalies relative to the annual climatology (computed with the ensemble average). All products exhibit the same interannual patterns, although some discrepancies are observed at intra-seasonal timescales. This is reflected by the small differences in the standard deviations computed for each time series over the denser period (1993–2023). A more precise view of the differences of each product's SST index with the ensemble average is given by Fig. 2a, quantified by their respective root-mean-square differences. Before 1993, the ensemble average is computed only with the ERA5 reanalysis and the OSTIA-observation-derived-only product, covering this period. Consequently, Fig. 2a exhibits a large discrepancy of these two products with respect to the ensemble average. The 1993–2023 period is chosen to assess the relative merit of each product, quantified using the ARMOR3D observation-derived-only product, not included in the ensemble-average computation in the Taylor diagram (Fig. 2d). First, one can see very large differences with OSTIA, the other observation-derived-only product, suggesting that the impact of their respective representativity of SST in the ATL3 box and possibly mapping/observation errors should be investigated further. The lesson here is that the "ground truth" also presents subjective drawbacks that need to be taken into account while measuring the relative merit in this multi-product ensemble assessment. The Taylor diagram reflects the very close performances of all products altogether in a cluster. The ensemble average performs better than individual reanalyses. The GREP multi-reanalyses product presents also good performances in representing the ATL3 index relatively to ARMOR3D. This confirms previous findings (e.g. Masina et al., 2015; Uotila et al., 2018; Storto et al., 2019) showing the "bias-reduction" benefits of ensemble averaging. In practice, the ensemble average provides a valuable estimate of the decadal SST trend in the ATL3 box. The ensemble-average estimate is also useful in identifying outliers.

Note that in recent methodologies, ensemble forecast comparison is performed using "ensemble clustering", also called "consensus clustering", which aims at producing a synthesis among an identified cluster from a given dataset (e.g. Hakobyan, 2010). The construction of the clusters from the initial dataset (here the different members of the ensemble forecast) can use many criteria. In the frame of GODAE OceanView, the Class 1 metrics were designed to compare OOFS variables on specific model grids and layers in similar ways (Hernandez et al., 2015b). In the Class 1 approach, OOFS outputs are re-gridded and resampled in a common grid and time frame (e.g. daily 2D model fields) and compared to a common reference (e.g. a regular L4 mapping of sea surface temperature from satellite retrievals). In this intercomparison, Class 1 files from various global OOFSs were used to compare and evaluate the quality of the ensemble mean; the weighted ensemble mean; and the k-mean clustering algorithm mean (Hartigan and Wong, 1979), which proved to be the more accurate (Hernandez et al., 2015b). Consensus clustering is now used for machine learning, and this might be one of the next stages associated with model products' intercomparison and ocean state estimation in the near future.

3.3 Regional forecast intercomparison and nesting strategy evaluation

Over recent years, the validation methodology proposed by the GODAE global ocean community has been adopted by many operational regional centres (some examples are given by Hernandez et al., 2015b), in particular because the coastal community started to relate inside GODAE OceanView with the IVTT. Specific assessments also started to be carried out, like assessing the behaviour of the ocean under tropical cyclone conditions using several OOFSs and ad hoc metrics (Zhu et al., 2016) or predicting the beaching of *Sargassum* in the Caribbean using global and regional OOFSs (Cailleau et al., 2024).

On a regional basis, specific systematic multi-product validation tools are gradually developed (e.g. Lorente et al., 2016, 2019). These tools, operated by a given operational centre, are efficient essentially if an inter-operable data server policy is implemented among the operational ocean community, in order to allow the real-time intercomparison of different sources of products. In parallel, regional and coastal system evaluation relies on specific local observing systems, like high-frequency (HF) radar, offering an "ocean truth", representing the ocean dynamics at higher resolution (Kourafalou et al., 2015), which cannot be represented by global OOFSs.

However, it is worth noting that comprehensive multi-product operational intercomparison is not common at regional scales. Unlike global OOFSs, there are rarely many fine-scale regional OOFSs that overlap in a given coastal region, even along the well-covered European marginal seas (Capet et al., 2020). And conducting a regional intercomparison gathering essentially global OOFSs would provide little information compared with the global intercomparison initiatives already underway.

But there is an increasing number of operational centres, or programmes like the CMEMS, that operate both regional and global systems over the same area and that have started to intercompare their different systems. For instance, two OOFSs of the same kind, Mercator and MFS (Mediterranean Forecasting System), are compared in the western Mediterranean basin, and their respective strengths and weaknesses are evaluated over specific subdomains (Juza et al., 2015). The benefit of improving the resolution of a regional OOFS is measured by comparing the coarse and fine grid systems using the same metrics (Crocker et al., 2020). In the CMEMS, most regional systems are nested into the global system. Hence, intercomparison between "parent" and "child" systems started to arise with the objective of measuring the benefit and added value for users of proposed regional and coastal products (De Mey et al., 2009). Several overlapping regional systems in the CMEMS can be compared to the global solution (Juza et al., 2016; Lorente et al., 2019). Examples can also be given for the Canadian Arctic and North Atlantic regional OOFS (Dupont et al., 2015), the US East Coast OOFS and reanalyses (Wilkin et al., 2022), and the Australian global and regional OOFS evaluations that focus on specific case studies and applications: disaster/search and rescue, defence/acoustic, and sea level/coastal management (Schiller et al., 2020). Some of these intercomparisons compare the regional OOFS of interest with several global products in order to measures both the local and global forecast skill, considering fine scales. In this case, using similar metrics, typically Class 4, for evaluating all these systems brings a series of questions. Which are the scales represented by the child system that is lacking in the parent system or in the observations? What is the impact of the different kind of forcings and different kind of assimilated dataset? How do errors propagate from the global to the nested system and degrade the expected seamless transition from the open ocean to coastal dynamics? How are specific ocean processes of interest represented in different systems? How reliable are they for end users' needs in different systems?

3.4 Evaluating retrospective views of the ocean dynamics: dedicated ocean reanalyses intercomparison project and ways to improve intercomparison methodologies

Past numerical simulations and ocean reanalyses were naturally the first step built in academia to study ocean processes over long periods, with the support of the increasing number of ocean observations over time and the improvement of assimilation techniques. Evaluation of such reanalysis representing decades of ocean behaviour through comprehensive intercomparison projects requires considerable resources and preparation. Most are conducted outside of routine operations by forecasting centres. They represent a milestone in progress in the field, from the point of view of both the evaluation of the system/reanalysis itself and the new validation methodologies that have been tested and implemented.

In the direct line of the GSOP project, the Ocean Reanalysis Intercomparison Project (ORA-IP) brought together more than 20 operational centres in order to intercompare more than 25 products (including observed products) spanning 20 to 50 years and focusing on eight EOVs – ocean heat content, steric height, sea level, surface heat fluxes, mixed layer depth, salinity, depth of the 20 °C isotherm and sea ice (Fig. 1). One of the objectives was to infer a new ocean state estimation of the global ocean, trying to reduce the so-called structural uncertainty, i.e. the uncertainty associated with the state estimation methodology, which cannot be sampled with a single system. Uncertainty is sensitive to the temporal variations of the observing system and to the errors of the ocean model, atmospheric fluxes and assimilation system, which are often flow-dependent and not easy to estimate. Following the Class 1 metrics approach, the ORA-IP is based on common

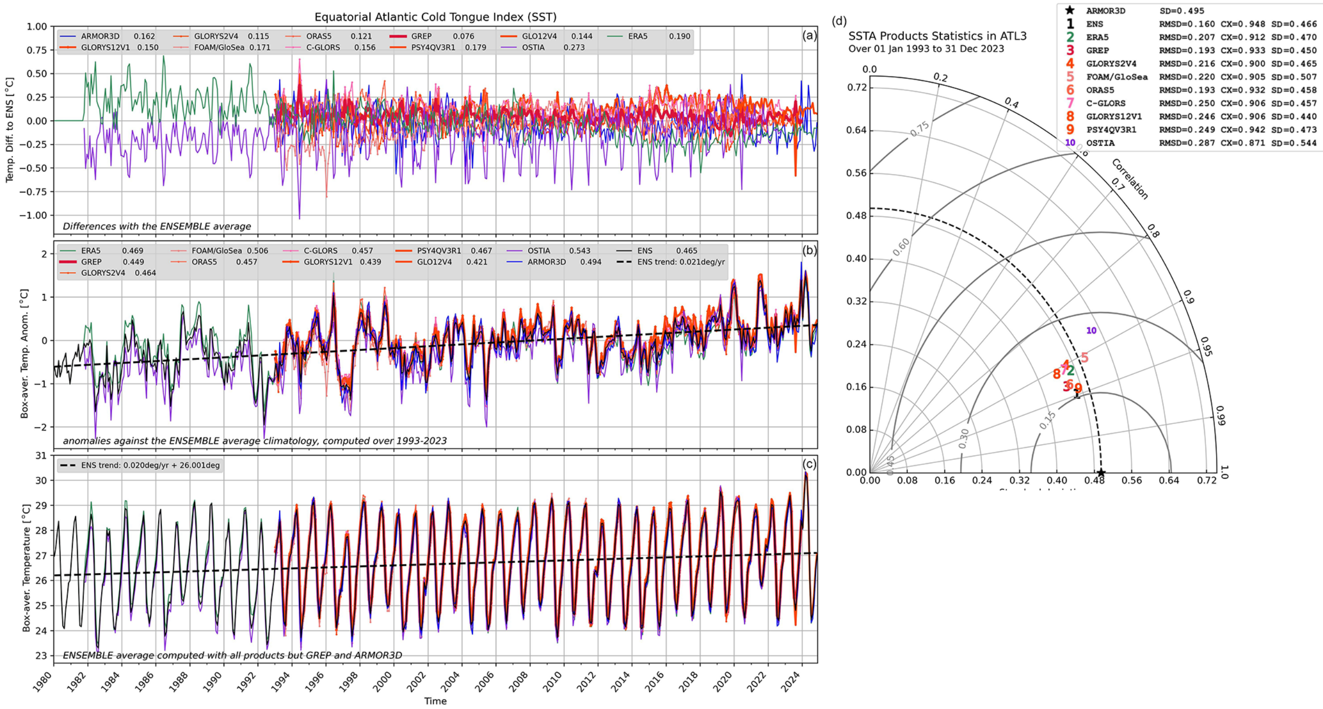

Figure 2. **(a)**–**(c)** Time series from 1980 to 2024 of SST products, monthly and spatially averaged into the ATL3 box located in the eastern equatorial band [20° W–0° E, 2.5° S–2.5° N] of the tropical Atlantic. **(a)** Differences relative to the ensemble average (root-mean-square differences (RMSDs) indicated in the legend). **(b)** The ATL3 index computed as anomalies relative to the climatology mean (standard deviations indicated in the legend). **(c)** The time series of box-averaged SST in the ATL3 box. **(d)** The associated Taylor diagram of the ATL3 index, using the ARMOR3D product as a reference. Statistics of root-mean-square differences, correlation with ARMOR3D and standard deviations for each product are given in the legend. All products were extracted from the Copernicus Marine Data Store and Climate Data Store.

grid re-interpolated products and monthly averages that were compared similarly over the 1993–2010 period under the responsibility of a leading expert for each of the eight EOVs. Results highlighted impacts of model resolution, components of the observing system assimilated, complexity of the ocean models and the data assimilation scheme, and quality of external forcing (Palmer et al., 2015; Shi et al., 2015; Storto et al., 2015; Toyoda et al., 2015a, b; Valdivieso et al., 2015; Chevallier et al., 2016).

New independent metrics were tested and used to evaluate each product and also the ensemble mean. The ensemble spread was identified as a measure of uncertainty. Following Storto et al. (2019), ocean reanalyses offer state-of-the-art representation of the past and present state of the global and regional oceans. Their accuracy depends on many factors, one of the most important being the observations available and the constraints they provide. Intercomparison helps in identifying the impact of their absence in the past and defines where they are most crucial in the quality of present and future reanalyses. And consequently, suggestions for improvements of the GOOS are provided.

Figure 2 shows that multi-product intercomparisons allow key indicator of the ocean environment changes to be inferred together with estimates of their uncertainties. Beyond reanalyses assessment based on EOVs, the next stage of ocean reanalyses intercomparison should first target key ocean processes that affect the climate system, identify their past occurrences, and better unravel their mechanisms and interactions, in order to estimate their present and future impacts. Machine learning approaches are expected to explore ocean variability in a multi-system framework more systematically and disentangle ocean key mechanisms for further identification in ocean simulations (e.g. Ahmad, 2019; Sonnewald et al., 2021; Salman, 2023). In particular, in ESM simulations, initial conditions are crucial: more realistic clusters of ocean reanalyses with better characterization of their errors and limitations (with or without the support of artificial intelligence) would ensure more reliable global and regional climate projections and associated skill assessment. Following this framework, ocean reanalyses intercomparison initiatives should also target end users' applications and societal impacts and identify requirements in terms of OOFS resolution, frequency and complexity, together with adequate observing systems, able to provide reliable and useful answers. Emerging international panels like the OceanPrediction Decade Collaborative Centre should help in providing intercomparison standards and recommendations from the user's point of view (Ciliberti et al., 2023). As already commented above, large and comprehensive multi-reanalyses intercomparisons are demanding and bring technical challenges in terms of storage, access, distribution and shareability. Cloud computing, ad hoc data mining technics and other artificial intelligence approaches will be needed to obtain valuable outcomes from the increasing number of available numerical ocean products resolving finer scales over longer periods.

3.5 A perspective of ocean reanalyses intercomparison: ocean state monitoring

An important outcome of the ORA-IP has been the development of the Real Time Multiple Ocean Reanalysis Intercomparison, carried out routinely every month by NOAA/NCEP, whose main objective is to gather operational hindcasts in order to perform ocean state monitoring (OSM) over the tropical Pacific, inferring the state of the ocean by computing the ensemble mean and identifying robust patterns using the ensemble spread (Xue et al., 2017). Note that OSM has growing importance in operational oceanography: through key EOVs it offers an assessment of the evolution of the ocean component as part of the real-time climate system evolution. Validation performed in the frame of OSM also provides a level of uncertainty for seasonal forecasts performed every month by many centres nowadays. OSM activity brought the CMEMS into routine calculation of Ocean Monitoring Indicators (OMIs), whose reliability and uncertainty are estimated through intercomparison of multiple products. Using OMIs, in 2018 the CMEMS started to produce the Ocean State Report (von Schuckmann et al., 2018) on an annual basis, now on its eighth edition (https://marine.copernicus.eu/access-data/ocean-state-report, last access: 29 January 2025).

Data availability. Ocean products used to produce Fig. 2 were downloaded in November 2024 from the Copernicus Marine Data Store and Climate Data Store (https://marine.copernicus.eu/ and https://climate.copernicus.eu/, last access: 29 January 2025).

- ERA5: https://doi.org/10.24381/cds.f17050d7 (Hersbach et al., 2023).
- OSTIA: https://doi.org/10.48670/moi-00165 (CMEMS, 2023a; Good et al., 2020).
- GLORYS12V1: https://doi.org/10.48670/moi-00021 (CMEMS, 2023b; Lellouche et al., 2021).
- ARMOR3D: https://doi.org/10.48670/moi-00052 (CMEMS, 2024a; Guinehut et al., 2012).
- GLO12V4 and PSY4QV3R1: https://doi.org/10.48670/moi-00016 (CMEMS, 2024b; Lellouche et al., 2018).
- GREP and FOAM/GloSea and C-GLORS and ORAS5 and GLORYS2V4: https://doi.org/10.48670/moi-00024 (CMEMS, 2024c; Masina et al., 2015).

Figures 1 and 2 are produced using Python 3.6 Matplotlib modules.

Author contributions. FH wrote the article and produced Figs. 1 and 2. MGS and AM participated in the initial redaction and the overview of the article.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. Fabrice Hernandez contributed as IRD researcher to this publication initiative of the OceanPrediction Decade Collaborative Centre as part as an in-kind IRD contribution to Mercator Ocean International in 2024 and 2025. Fabrice Hernandez thanks the Mercator Ocean International validation team, Marie Drévillon and Charly Regnier for their helpful comments ; the Ocean Predict Intercomparison and Validation Task Team and Greg Smith for the continuous efforts in providing Class 4 files for routine intercomparison that could be discussed in this article ; the Copernicus Marine Monitoring and Forecasting Centre Product Quality Working Group for their pioneering efforts in validation and intercomparison approaches. Marcos Garcia Sotillo contributed to this publication as leader and product quality expert of the Copernicus Marine Monitoring and Forecasting Centre for the IBI region. The authors thank the reviewers for their helpful comments that suggested an extended and updated view on intercomparison activities.

Financial support. This research has been supported by grants of the European Union's Horizon Europe programme (project SEACLIM, grant no. 101180125, and project FOCCUS, grant no. 101133911), awarded to Angélique Melet, and by an EU CMEMS grant (contract 21002L6-COP-MFC IBI-5600), awarded to Marcos Garcia Sotillo.

Review statement. This paper was edited by Enrique Álvarez Fanjul and reviewed by two anonymous referees.

References

Ahmad, H.: Machine learning applications in oceanography, Aquatic Research, 2, 161–169, https://doi.org/10.3153/AR19014, 2019.

Aijaz, S., Brassington, G. B., Divakaran, P., Régnier, C., Drévillon, M., Maksymczuk, J., and Peterson, K. A.: Verification and intercomparison of global ocean Eulerian near-surface currents, Ocean Model., 186, 102241, https://doi.org/10.1016/j.ocemod.2023.102241, 2023.

Alvarez Fanjul, E., Ciliberti, S., and Bahurel, P.: Implementing Operational Ocean Monitoring and Forecasting Systems, IOC-UNESCO, GOOS-275, https://doi.org/10.48670/ETOOFS, 2022.

Balmaseda, M. A., Hernandez, F., Storto, A., Palmer, M. D., Alves, O., Shi, L., G. C. Smith, Toyoda, T., Valdivieso da Costa, M., Barnier, B., Behringer, D. W., Boyer, T. P., Y.-S. Chang, Chepurin, G. A., Ferry, N., Forget, G., Fujii, Y., Good, S., Guinehut, S., Haines, K., Ishikawa, Y., Keeley, S., Köhl, A., Lee, T., Martin, M., Masina, S., Masuda, S., Meyssignac, B., Mogensen, K., Parent, L., Peterson, A. K., Tang, Y. M., Yin, Y., Vernieres, G., Wang, X., Waters, J., Wedd, R., Wang, O., Xue, Y., Chevallier, M., Lemieux, J.-F., Dupont, F., Kuragano, T., Kamachi, M., Awaji, T., Caltabiano, A. C., Wilmer-Becker, K., and Gaillard, F.: The Ocean Reanalyses Intercomparison Project (ORA-IP), J. Oper. Oceanogr., 8, s80–s97, https://doi.org/10.1080/1755876X.2015.1022329, 2015.

Balmaseda, M. A., Balan Sarojini, B., Mayer, M., Tietsche, S., Zuo, H., Vitart, F., and Stockdale, T. N.: Impact of the ocean *in situ* observations on the ECMWF seasonal forecasting system, Front. Mar. Sci., 11, https://doi.org/10.3389/fmars.2024.1456013, 2024a.

Balmaseda, M. A., McAdam, R., Masina, S., Mayer, M., Senan, R., de Bosisséson, E., and Gualdi, S.: Skill assessment of seasonal forecasts of ocean variables, Front. Mar. Sci., 11, https://doi.org/10.3389/fmars.2024.1380545, 2024b.

Bell, M. J., Lefebvre, M., Le Traon, P.-Y., Smith, N., and Wilmer-Becker, K.: GODAE, The Global Ocean Data Experiment, Oceanography, 22, 14–21, https://doi.org/10.5670/oceanog.2009.62, 2009.

Böning, C. W. and Bryan, F. O.: Large-Scale Transport Processes in High-Resolution Circulation Models, in: The Warmwatersphere of the North Atlantic Ocean, edited by: Krauss, W., Gebrüder Borntraeger: Berlin – Stuttgart, 91–128, ISBN 3-443-01033-4, 1996.

Cailleau, S., Bessières, L., Chiendje, L., Dubost, F., Reffray, G., Lellouche, J.-M., van Gennip, S., Régnier, C., Drevillon, M., Tressol, M., Clavier, M., Temple-Boyer, J., and Berline, L.: CAR36, a regional high-resolution ocean forecasting system for improving drift and beaching of Sargassum in the Caribbean archipelago, Geosci. Model Dev., 17, 3157–3173, https://doi.org/10.5194/gmd-17-3157-2024, 2024.

Campos, R. M., Alves, J.-H. G. M., Penny, S. G., and Krasnopolsky, V.: Assessments of Surface Winds and Waves from the NCEP Ensemble Forecast System, Weather Forecast.., 33, 1533–1546, https://doi.org/10.1175/waf-d-18-0086.1, 2018.

Capet, A., Fernández, V., She, J., Dabrowski, T., Umgiesser, G., Staneva, J., Mészáros, L., Campuzano, F., Ursella, L., Nolan, G., and El Serafy, G.: Operational Modeling Capacity in European Seas—An EuroGOOS Perspective and Recommendations for Improvement, Front. Mar. Sci., 7, https://doi.org/10.3389/fmars.2020.00129, 2020.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18, https://doi.org/10.1002/met.52, 2008.

Chevallier, M., Smith, G. C., Dupont, F., Lemieux, J.-F., Forget, G., Fujii, Y., Hernandez, F., Msadek, R., Peterson, K. A., Storto, A., Toyoda, T., Valdivieso, M., Vernieres, G., Zuo, H., Balmaseda, M., Chang, Y.-S., Ferry, N., Garric, G., Haines, K., Keeley, S., Kovach, R. M., Kuragano, T., Masina, S., Tang, Y., Tsujino, H., and Wang, X.: Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project, Clim. Dynam., 49, 1107–1136, https://doi.org/10.1007/s00382-016-2985-y, 2016.

Ciliberti, S. A., Alvarez Fanjul, E., Pearlman, J., Wilmer-Becker, K., Bahurel, P., Ardhuin, F., Arnaud, A., Bell, M., Berthou, S., Bertino, L., Capet, A., Chassignet, E., Ciavatta, S., Cirano, M., Clementi, E., Cossarini, G., Coro, G., Corney, S., Davidson, F., Drevillon, M., Drillet, Y., Dussurget, R., El Serafy, G., Fennel, K., Garcia Sotillo, M., Heimbach, P., Hernandez, F., Hogan, P., Hoteit, I., Joseph, S., Josey, S., Le Traon, P.-Y., Libralato, S., Mancini, M., Matte, P., Melet, A., Miyazawa, Y., Moore, A. M., Novellino, A., Porter, A., Regan, H., Romero, L., Schiller, A., Siddorn, J., Staneva, J., Thomas-Courcoux, C., Tonani, M., Garcia-Valdecasas, J. M., Veitch, J., von Schuckmann, K., Wan, L., Wilkin, J., and Zufic, R.: Evaluation of operational ocean forecasting systems from the perspective of the users and the experts, in: 7th edition of the Copernicus Ocean State Report (OSR7), edited by: von Schuckmann, K., Moreira, L., Le Traon, P.-Y., Grégoire, M., Marcos, M., Staneva, J., Brasseur, P., Garric, G., Lionello, P., Karstensen, J., and Neukermans, G., Copernicus Publications, State Planet, 1-osr7, 2, https://doi.org/10.5194/sp-1-osr7-2-2023, 2023.

Coelho, C. A. S., Brown, B., Wilson, L., Mittermaier, M., and Casati, B.: Chapter 16 – Forecast Verification for S2S Timescales, in: Sub-Seasonal to Seasonal Prediction, edited by: Robertson, A. W. and Vitart, F., Elsevier, 337–361, ISBN 978-0-12-811714-9, 2019.

Crocker, R., Maksymczuk, J., Mittermaier, M., Tonani, M., and Pequignet, C.: An approach to the verification of high-resolution ocean models using spatial methods, Ocean Sci., 16, 831–845, https://doi.org/10.5194/os-16-831-2020, 2020.

Danabasoglu, G., Yeager, S. G., Bailey, D., Behrens, E., Bentsen, M., Bi, D., Biastoch, A., Boning, C., Bozec, A., Canuto, V. M., Cassou, C., Chassignet, E., Coward, A. C., Danilov, S., Diansky, N., Drange, H., Farneti, R., Fernandez, E., Fogli, P. G., Forget, G., Fujii, Y., Griffies, S. M., Gusev, A., Heimbach, P., Howard, A., Jung, T., Kelley, M., Large, W. G., Leboissetier, A., Lu, J., Madec, G., Marsland, S. J., Masina, S., Navarra, A., George Nurser, A. J., Pirani, A., Salas y Melia, D., Samuels, B. L., Scheinert, M., Sidorenko, D., Treguier, A.-M., Tsujino, H., Uotila, P., Valcke, S., Voldoire, A., and Wang, Q.: North Atlantic simulations in Coordinated Ocean-ice Reference Experiments phase II (CORE-II). Part I: Mean states, Ocean Model., 73, 76–107, https://doi.org/10.1016/j.ocemod.2013.10.005, 2014.

De Mey, P., Craig, P., Davidson, F., Edwards, C. A., Ishikawa, Y., Kindle, J. C., Proctor, R., Thompson, K. R., Zhu, J., and GODAE Coastal and Shelf Seas Working Group: Application in Coastal Modelling and Forecasting, Oceanography Magazine, 22, 198–205, https://doi.org/10.5670/oceanog.2009.79, 2009.

Divakaran, P., Brassington, G. B., Ryan, A. G., Regnier, C., Spindler, T., Mehra, A., Hernandez, F., Smith, G. C., Liu, Y. Y., and Davidson, F.: GODAE OceanView Class 4 inter-comparison for the Australian Region, J. Oper. Oceanogr., 8, s112–s126, https://doi.org/10.1080/1755876X.2015.1022333, 2015.

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

Dupont, F., Higginson, S., Bourdallé-Badie, R., Lu, Y., Roy, F., Smith, G. C., Lemieux, J.-F., Garric, G., and Davidson, F.: A high-resolution ocean and sea-ice modelling system for the Arctic and North Atlantic oceans, Geosci. Model Dev., 8, 1577–1594, https://doi.org/10.5194/gmd-8-1577-2015, 2015.

Durgadoo, J. V., Biastoch, A., New, A. L., Rühs, S., Nurser, A. J. G., Drillet, Y., and Bidlot, J.-R.: Strategies for simulating the drift of marine debris, J. Oper. Oceanogr., 14, 1–12, https://doi.org/10.1080/1755876X.2019.1602102, 2021.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean OSTIA Sea Surface Temperature and Sea Ice Analysis, Marine Data Store (MDS) [data set], https://doi.org/10.48670/moi-00165, 2023a.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean Physics Reanalysis, Marine Data Store (MDS) [data set], https://doi.org/10.48670/moi-00021, 2023b.

E.U. Copernicus Marine Service Information (CMEMS): Multi Observation Global Ocean 3D Temperature Salinity Height Geostrophic Current and MLD, Marine Data Store (MDS) [data set], https://doi.org/10.48670/moi-00052, 2024a.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean Physics Analysis and Forecast, Marine Data Store (MDS) [data set], https://doi.org/10.48670/moi-00016, 2024b.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean Ensemble Physics Reanalysis, Marine Data Store (MDS) [data set], https://doi.org/10.48670/moi-00024, 2024c.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynam., 53, 343–367, https://doi.org/10.1007/s10236-003-0036-9, 2003.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Fujii, Y., Rémy, E., Zuo, H., Oke, P., Halliwell, G., Gasparin, F., Benkiran, M., Loose, N., Cummings, J., Xie, J., Xue, Y., Masuda, S., Smith, G. C., Balmaseda, M., Germineaud, C., Lea, D. J., Larnicol, G., Bertino, L., Bonaduce, A., Brasseur, P., Donlon, C., Heimbach, P., Kim, Y., Kourafalou, V., Le Traon, P.-Y., Martin, M., Paturi, S., Tranchant, B., and Usui, N.: Observing System Evaluation Based on Ocean Data Assimilation and Prediction Systems: On-Going Challenges and a Future Vision for Designing and Supporting Ocean Observational Networks, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00417, 2019.

Fujii, Y., Remy, E., Balmaseda, M. A., Kido, S., Waters, J., Peterson, K. A., Smith, G. C., Ishikawa, I., and Chikhar, K.: The international multi-system OSEs/OSSEs by the UN Ocean Decade Project SynObs and its early results, Front. Mar. Sci., 11, https://doi.org/10.3389/fmars.2024.1476131, 2024.

Gates, W. L.: An AMS Continuing Series: Global Change–AMIP: The Atmospheric Model Intercomparison Project, B. Am. Meteorol. Soc., 73, 1962–1970, https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2, 1992.

Good, S., Fiedler, E., Mao, C., Martin, M. J., Maycock, A., Reid, R., Roberts-Jones, J., Searle, T., Waters, J., While, J., and Worsfold, M.: The Current Configuration of the OSTIA System for Operational Production of Foundation Sea Surface Temperature and Ice Concentration Analyses, Remote Sens., 12, 720, https://doi.org/10.3390/rs12040720, 2020.

Griffies, S. M., Biastoch, A., Böning, C., Bryan, F., Danabasoglu, G., Chassignet, E. P., England, M. H., Gerdes, R., Haak, H., Hallberg, R. W., Hazeleger, W., Jungclaus, J., Large, W. G., Madec, G., Pirani, A., Samuels, B. L., Scheinert, M., Gupta, A. S., Severijns, C. A., Simmons, H. L., Treguier, A. M., Winton, M., Yeager, S., and Yin, J.: Coordinated Ocean-ice Reference Experiments (COREs), Ocean Model., 26, 1–46, https://doi.org/10.1016/j.ocemod.2008.08.007, 2009.

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., Chassignet, E. P., Curchitser, E., Deshayes, J., Drange, H., Fox-Kemper, B., Gleckler, P. J., Gregory, J. M., Haak, H., Hallberg, R. W., Heimbach, P., Hewitt, H. T., Holland, D. M., Ilyina, T., Jungclaus, J. H., Komuro, Y., Krasting, J. P., Large, W. G., Marsland, S. J., Masina, S., McDougall, T. J., Nurser, A. J. G., Orr, J. C., Pirani, A., Qiao, F., Stouffer, R. J., Taylor, K. E., Treguier, A. M., Tsujino, H., Uotila, P., Valdivieso, M., Wang, Q., Winton, M., and Yeager, S. G.: OMIP contribution to CMIP6: experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project, Geosci. Model Dev., 9, 3231–3296, https://doi.org/10.5194/gmd-9-3231-2016, 2016.

Griffin, D. A. and Oke, P. R.: The search for MH370 and ocean surface drift – Part III, CSIRO Oceans and Atmosphere, Australia, https://www.atsb.gov.au/sites/default/files/media/5773371/mh370_csiro-ocean-drift-iiil.pdf (last access: 24 April 2025), 2017.

Guinehut, S., Dhomps, A.-L., Larnicol, G., and Le Traon, P.-Y.: High resolution 3-D temperature and salinity fields derived from in situ and satellite observations, Ocean Sci., 8, 845–857, https://doi.org/10.5194/os-8-845-2012, 2012.

Hakobyan, M.: Ocean data mining with application to climate indices of the North Atlantic, Master, Department of Computational Science Memorial University of Newfoundland, St John's, 154 pp., https://research.library.mun.ca/10594/1/Hakobyan_Madlena.pdf (last access: 24 April 2025), 2010.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, J. R. Stat. Soc. C-Appl., 28, 100–108, https://doi.org/10.2307/2346830, 1979.

Hernandez, F.: Performance of Ocean Forecasting Systems - Intercomparison Projects, in: Operational Oceanography in the 21st Century, edited by: Schiller, A. and Brassington, G., Springer, Dordrecht, https://doi.org/10.1007/978-94-007-0332-2_23, 2011.

Hernandez, F., Blockley, E., Brassington, G. B., Davidson, F., Divakaran, P., Drévillon, M., Ishizaki, S., Sotillo, M. G., Hogan, P. J., Lagemaa, P., Levier, B., Martin, M., Mehra, A., Mooers, C., Ferry, N., Ryan, A., Regnier, C., Sellar, A., Smith, G. C., Sofianos, S., Spindler, T., Volpe, G., Wilkin, J., Zaron, E. D., and Zhang, A.: Recent progress in performance evaluations and near real-time assessment of operational ocean products, J. Oper. Oceanogr., 8, s221–s238, https://doi.org/10.1080/1755876X.2015.1050282, 2015a.

Hernandez, F., Bertino, L., Brassington, G., Chassignet, E., Cummings, J., Davidson, F., Drévillon, M., Garric, G., Kamachi, M., Lellouche, J.-M., Mahdon, R., Martin, M. J., Ratsimandresy, A., and Regnier, C.: Validation and Intercomparison Studies Within GODAE, Oceanography, 22, 128–143, https://doi.org/10.5670/oceanog.2009.71, 2015b.

Hernandez, F., Smith, G. C., Baetens, K., Cossarini, G., Garcia Hermosa, I., Drévillon, M., Maksymczuk, J., Melet, A., Régnier, C., and von Schuckmann, K.: Measuring performances, skill and accuracy in operational oceanography: New challenges and approaches, in: New Frontiers in Operational Oceanography, edited by: Chassignet, E., Pascual, A., Tintoré, J., and J. Verron, GODAE OceanView, 759–796, https://doi.org/10.17125/gov2018.ch29, 2018.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J.-N.: ERA5 monthly averaged data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.f17050d7, 2023.

Juza, M., Mourre, B., Lellouche, J.-M., Tonani, M., and Tintoré, J.: From basin to sub-basin scale assessment and intercomparison of numerical simulations in the Western Mediterranean Sea, J. Marine Syst., 149, 36–49, https://doi.org/10.1016/j.jmarsys.2015.04.010, 2015.

Juza, M., Mourre, B., Renault, L., Gómara, S., Sebastián, K., Lora, S., Beltran, J. P., Frontera, B., Garau, B., Troupin, C., Torner, M., Heslop, E., Casas, B., Escudier, R., Vizoso, G., and Tintoré, J.: SOCIB operational ocean forecasting system and multi-platform validation in the Western Mediterranean Sea, J. Oper. Oceanogr., 9, s155–s166, https://doi.org/10.1080/1755876X.2015.1117764, 2016.

Keppenne, C. L., Rienecker, M. M., Jacob, J. P., and Kovach, R.: Error Covariance Modeling in the GMAO Ocean Ensemble Kalman Filter, Mon. Weather Rev., 136, 2964–2982, https://doi.org/10.1175/2007MWR2243.1, 2008.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Jpn. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Kourafalou, V. H., De Mey, P., Le Hénaff, M., Charria, G., Edwards, C. A., He, R., Herzfeld, M., Pascual, A., Stanev, E. V., Tintoré, J., Usui, N., van der Westhuysen, A. J., Wilkin, J. L., and Zhu, X.: Coastal Ocean Forecasting: system integration and evaluation, J. Oper. Oceanogr., 8, s127–s146, https://doi.org/10.1080/1755876X.2015.1022336, 2015.

Large, W. G. and Yeager, S. G.: The global climatology of an interannually varying air–sea flux data set, Clim. Dynam., 33, 341–364, https://doi.org/10.1007/s00382-008-0441-3, 2009.

Lellouche, J.-M., Eric, G., Romain, B. B., Gilles, G., Angelique, M., Marie, D., et al.: The copernicus global 1/12° oceanic and sea ice GLORYS12 reanalysis, Frontiers of Earth Science, 9, 698876, https://doi.org/10.3389/feart.2021.698876, 2021.

Lellouche, J.-M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.-E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., and Le Traon, P.-Y.: Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system, Ocean Sci., 14, 1093–1126, https://doi.org/10.5194/os-14-1093-2018, 2018.

Lisæter, K. A., Rosanova, J., and Evensen, G.: Assimilation of ice concentration in a coupled ice–ocean model, using the Ensemble Kalman filter, Ocean Dynam., 53, 368–388, https://doi.org/10.1007/s10236-003-0049-4, 2003.

Lorente, P., Piedracoba, S., Sotillo, M. G., Aznar, R., Amo-Baladrón, A., Pascual, Á., Soto-Navarro, J., and Álvarez-Fanjul, E.: Ocean model skill assessment in the NW Mediterranean using multi-sensor data, J. Oper. Oceanogr., 9, 75–92, https://doi.org/10.1080/1755876X.2016.1215224, 2016.

Lorente, P., García-Sotillo, M., Amo-Baladrón, A., Aznar, R., Levier, B., Sánchez-Garrido, J. C., Sammartino, S., de Pascual-Collar, Á., Reffray, G., Toledano, C., and Álvarez-Fanjul, E.: Skill assessment of global, regional, and coastal circulation forecast models: evaluating the benefits of dynamical downscaling in IBI (Iberia–Biscay–Ireland) surface waters, Ocean Sci., 15, 967–996, https://doi.org/10.5194/os-15-967-2019, 2019.

Masina, S., Storto, A., Ferry, N., Valdivieso, M., Haines, K., Balmaseda, M., Zuo, H., Drevillon, M., and Parent, L.: An ensemble of eddy-permitting global ocean reanalyses from the MyOcean project, Clim. Dynam., 49, 813–841, https://doi.org/10.1007/s00382-015-2728-5, 2015.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: Intercomparison makes for a better climate model, Eos T. Am. Geophys. Un., 78, 445–451, https://doi.org/10.1029/97EO00276, 1997.

Meincke, J., Le Provost, C., and Willebrand, J.: Dynamics of the North Atlantic Circulation: Simulation and Assimilation with High-Resolution Models (DYNAMO), Prog. Oceanogr., 48, 121–122, https://doi.org/10.1016/S0079-6611(01)00002-7, 2001.

Murphy, A. H.: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecast.ing, Weather Forecast., 8, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Orr, J. C., Najjar, R. G., Aumont, O., Bopp, L., Bullister, J. L., Danabasoglu, G., Doney, S. C., Dunne, J. P., Dutay, J.-C., Graven, H., Griffies, S. M., John, J. G., Joos, F., Levin, I., Lindsay, K., Matear, R. J., McKinley, G. A., Mouchet, A., Oschlies, A., Romanou, A., Schlitzer, R., Tagliabue, A., Tanhua, T., and Yool, A.: Biogeochemical protocols and diagnostics for the CMIP6 Ocean Model Intercomparison Project (OMIP), Geosci. Model Dev., 10, 2169–2199, https://doi.org/10.5194/gmd-10-2169-2017, 2017.

Palmer, M. D., Roberts, C. D., Balmaseda, M., Chang, Y. S., Chepurin, G. A., Ferry, N., Fujii, Y., Good, S. A., Guinehut, S., Haines, K., Hernandez, F., Köhl, A., Lee, T., Martin, M. J., Masina, S., Masuda, S., Peterson, K. A., Storto, A., Toyoda, T., Valdivieso, M., Vernieres, G., Wang, O., and Xue, Y.: Ocean heat content variability and change in an ensemble of ocean reanalyses, Clim. Dynam., 49, 909–930, https://doi.org/10.1007/s00382-015-2801-0, 2015.

Pinardi, N., Bonazzi, A., Dobricic, S., Milliff, R. F., Wikle, C. K., and Berliner, L. M.: Ocean ensemble forecasting. Part II: Mediterranean Forecast System response, Q. J. Roy. Meteor. Soc., 137, 879–893, https://doi.org/10.1002/qj.816, 2011.

Ryan, A. G., Regnier, C., Divakaran, P., Spindler, T., Mehra, A., Hernandez, F., Smith, G. C., Liu, Y. Y., and Davidson, F.: GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison, J. Oper. Oceanogr., 8, s98–s111, https://doi.org/10.1080/1755876X.2015.1022330, 2015.

Salman, A.: Editorial: Application of machine learning in oceanography and marine sciences, Front. Mar. Sci., 10, https://doi.org/10.3389/fmars.2023.1207337, 2023.

Schiller, A., Brassington, G. B., Oke, P., Cahill, M., Divakaran, P., Entel, M., Freeman, J., Griffin, D., Herzfeld, M., Hoeke, R., Huang, X., Jones, E., King, E., Parker, B., Pitman, T., Rosebrock, U., Sweeney, J., Taylor, A., Thatcher, M., Woodham, R., and Zhong, A.: Bluelink ocean forecasting Australia: 15 years of operational ocean service delivery with societal, economic and environmental benefits, J. Oper. Oceanogr., 13, 1–18, https://doi.org/10.1080/1755876X.2019.1685834, 2020.

Seo, G.-H., Kim, S., Choi, B.-J., Cho, Y.-K., and Kim, Y.-H.: Implementation of the Ensemble Kalman Filter into a Northwest Pacific Ocean Circulation Model, in: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, edited by: Park, S. K. and Xu, L., Springer Berlin Heidelberg, Berlin, Heidelberg, 341–351, ISBN 978-3-540-71056-1, 2009.

Shi, L., Alves, O. J., Wedd, R., Balmaseda, M. A., Chang, Y.-S., Chepurin, G. A., Ferry, N., Fujii, Y., Gaillard, F., Good, S. A., Guinehut, S., Haines, K., Hernandez, F., Lee, T., Palmer, M. D., Peterson, K. A., Masuda, S., Storto, A., Toyoda, T., Valdivieso, M., Vernieres, G., Wang, X., and Yin, Y.: An assessment of upper ocean salinity content from the Ocean Reanalyses Intercomparison Project (ORA-IP), Clim. Dynam., 49, 1009–1029, https://doi.org/10.1007/s00382-015-2868-7, 2015.

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V.: Bridging observations, theory and numerical simulation of the ocean using machine learning, Environ. Res. Lett., 16, 073008, https://doi.org/10.1088/1748-9326/ac0eb0, 2021.

Sotillo, M. G., Drevillon, M., and Hernandez, F.: A description of Validation Processes and Techniques for Ocean Forecasting, in: Ocean prediction: present status and state of the art (OPSR), edited by: Álvarez Fanjul, E., Ciliberti, S. A., Pearlman, J., Wilmer-Becker, K., and Behera, S., Copernicus Publications, State Planet, 5-opsr, 16, https://doi.org/10.5194/sp-5-opsr-16-2025, 2025.

Stammer, D., Köhl, A., Awaji, T., Balmaseda, M., Behringer, D., Carton, J., Ferry, N., Fischer, A., Fukumori, I., Giese, B., Haines, K., Harrison, E., Heimbach, P., Kamachi, M., Keppenne, C., Lee, T., Masina, S., Menemenlis, D., Ponte, R., Remy, E., Rienecker, M., Rosati, A., Schröter, Jens , Smith, D., Weaver, A., Wunsch, C., and Xue, Y.: Ocean Information Provided Through Ensemble Ocean Syntheses, in: Proceedings of OceanObs09: Sustained Ocean Observations and Information for Society (Vol. 2), Venice, Italy, 21–25 September 2009, edited by: Hall, J., Harrison D. E., and Stammer, D., ESA Publication WPP-306, https://pure.mpg.de/rest/items/item_2144438/component/file_2144439/content (last access: 24 April 2025), 2009.

Storto, A., Masina, S., Balmaseda, M., Guinehut, S., Xue, Y., Szekely, T., Fukumori, I., Forget, G., Chang, Y.-S., Good, S. A., Köhl, A., Vernieres, G., Ferry, N., Peterson, K. A., Behringer, D., Ishii, M., Masuda, S., Fujii, Y., Toyoda, T., Yin, Y., Valdivieso, M., Barnier, B., Boyer, T., Lee, T., Gourrion, J., Wang, O., Heimback, P., Rosati, A., Kovach, R., Hernandez, F., Martin, M. J., Kamachi, M., Kuragano, T., Mogensen, K., Alves, O., Haines, K., and Wang, X.: Steric sea level variability (1993–2010) in an ensemble of ocean reanalyses and objective analyses, Clim. Dynam., 49, 709–729, https://doi.org/10.1007/s00382-015-2554-9, 2015.

Storto, A., Alvera-Azcarate, A., Balmaseda, M. A., Barth, A., Chevallier, M., Counillon, F., Domingues, C. M., Drévillon, M., Drillet, Y., Forget, G., Garric, G., Haines, K., Hernandez, F., Iovino, D., Jackson, L. C., Lellouche, J.-M., Masina, S., Mayer, M., Oke, P. R., Penny, S. G., Peterson, A. K., Yang, C., and Zuo, H.: Ocean reanalyses: Recent advances and unsolved challenges, Front. Mar. Sci., 6, 418, https://doi.org/10.3389/fmars.2019.00418, 2019.

Tanhua, T., McCurdy, A., Fischer, A., Appeltans, W., Bax, N., Currie, K., DeYoung, B., Dunn, D., Heslop, E., Glover, L. K., Gunn, J., Hill, K., Ishii, M., Legler, D., Lindstrom, E., Miloslavich, P., Moltmann, T., Nolan, G., Palacz, A., Simmons, S., Sloyan, B., Smith, L. M., Smith, N., Telszewski, M., Visbeck, M., and Wilkin, J.: What We Have Learned From the Framework for Ocean Observing: Evolution of the Global Ocean Observing System, Front. Mar. Sci., 6, https://doi.org/10.3389/fmars.2019.00471, 2019.

Toyoda, T., Fujii, Y., Kuragano, T., Kamachi, M., Ishikawa, Y., Masuda, S., Sato, K., Awaji, T., Hernandez, F., Ferry, N., Guinehut, S., Martin, M. J., Peterson, K. A., Good, S. A., Valdivieso, M., Haines, K., Storto, A., Masina, S., Köhl, A., Zuo, H., Balmaseda, M., Yin, Y., Shi, L., Alves, O., Smith, G., Chang, Y.-S., Vernieres, G., Wang, X., Forget, G., Heimbach, P., Wang, O., Fukumori, I., and Lee, T.: Intercomparison and validation of the mixed layer depth fields of global ocean syntheses, Clim. Dynam., 49, 753–773, https://doi.org/10.1007/s00382-015-2637-7, 2015a.

Toyoda, T., Fujii, Y., Kuragano, T., Kosugi, N., Sasano, D., Kamachi, M., Ishikawa, Y., Masuda, S., Sato, K., Awaji, T., Hernandez, F., Ferry, N., Guinehut, S., Martin, M., Andrew Peterson, K., Good, S., Valdivieso, M., Haines, K., Storto, A., Masina, S., Köhl, A., Yin, Y., Shi, L., Alves, O., Smith, G., Chang, Y.-S., Vernieres, G., Wang, X., Forget, G., Heimbach, P., Wang, O., Fukumori, I., Lee, T., Zuo, H., and Balmaseda, M.: Interannual-decadal variability of wintertime mixed layer depths in the North Pacific detected by an ensemble of ocean syntheses, Clim. Dynam., 49, 891–907, https://doi.org/10.1007/s00382-015-2762-3, 2015b.

Uotila, P., Goosse, H., Haines, K., Chevallier, M., Barthélemy, A., Bricaud, C., Carton, J., Fučkar, N., Garric, G., Iovino, D., Kauker, F., Korhonen, M., Lien, V. S., Marnela, M., Massonnet, F., Mignac, D., Peterson, K. A., Sadikni, R., Shi, L., Tietsche, S., Toyoda, T., Xie, J., and Zhang, Z.: An assessment of ten ocean reanalyses in the polar regions, Clim. Dynam., 52, 1613–1650, https://doi.org/10.1007/s00382-018-4242-z, 2018.

Valdivieso, M., Haines, K., Balmaseda, M., Chang, Y.-S., Drevillon, M., Ferry, N., Fujii, Y., Köhl, A., Storto, A., Toyoda, T., Wang, X., Waters, J., Xue, Y., Yin, Y., Barnier, B., Hernandez, F., Kumar, A., Lee, T., Masina, S., and Andrew Peterson, K.: An assessment of air–sea heat fluxes from ocean and coupled reanalyses, Clim. Dynam., 49, 983–1008, https://doi.org/10.1007/s00382-015-2843-3, 2015.

Vallès-Casanova, I., Lee, S.-K., Foltz, G. R., and Pelegrí, J. L.: On the Spatiotemporal Diversity of Atlantic Niño and Associated Rainfall Variability Over West Africa and South America, Geophys. Res. Lett., 47, e2020GL087108, https://doi.org/10.1029/2020GL087108, 2020.

von Schuckmann, K., Le Traon, P. Y., Smith, N., et al.: Copernicus Marine Service Ocean State Report, J. Oper. Oceanogr., 11, S1–S142, https://doi.org/10.1080/1755876X.2018.1489208, 2018.

Wilkin, J., Levin, J., Moore, A., Arango, H., López, A., and Hunter, E.: A data-assimilative model reanalysis of the U.S. Mid Atlantic Bight and Gulf of Maine: Configuration and comparison to observations and global ocean models, Progr. in Oceanogr., 209, 102919, https://doi.org/10.1016/j.pocean.2022.102919, 2022.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

Xue, Y., Wen, C., Kumar, A., Balmaseda, M., Fujii, Y., Alves, O., Martin, M. J., Yang, X., Vernieres, G., Desportes, C., Lee, T., Ascione, I., Gudgel, R., and Ishikawa, I.: A real-time ocean reanalyses intercomparison project in the context of tropical pacific observing system and ENSO monitoring, Clim. Dynam., 49, 3647–3672, https://doi.org/10.1007/s00382-017-3535-y, 2017.

Zhu, X., Wang, H., Liu, G., Régnier, C., Kuang, X., Wang, D., Ren, S., Jing, Z., and Drévillon, M.: Comparison and validation of global and regional ocean forecasting systems for the South China Sea, Nat. Hazards Earth Syst. Sci., 16, 1639–1655, https://doi.org/10.5194/nhess-16-1639-2016, 2016.