A description of Model Intercomparison Processes and Techniques for Ocean Forecasting

Fabrice Hernandez[1], Marcos Garcia Sotillo[2], 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 to: Fabrice Hernandez (fabrice.hernandez@ird.fr)

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 launched only in the recent period dedicated actions aimed at identifying robust patterns in eddy-permitting simulations: it required definition of modelling configurations, execution of dedicated experiments that deal also with the storing of the outputs and the implementation of evaluation frameworks. Starting from this baseline, numerous initiatives like CLIVAR for climate research and GODAE for operational systems have raised and are actively promoting best practices through specific intercomparison tasks, aimed at demonstrating the efficient use of the Global Ocean Observing System and the operational capabilities, sharing expertise and increase the scientific quality of the numerical systems. Examples, like the ORA-IP, or the Class 4 near real time GODAE intercomparison are introduced and commented, discussing also on the ways forward on 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 centre.

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 Garcia-Sotillo et al. (2024, 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.

The same atmospheric community, involved in the climate studies, settled in 1996 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 CO2 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 through future CMIP phases were re-processed historical simulations defined in the AMIP will allow to track changes and benefits of more elaborated components of the Earth System Models (ESM).

The ocean modelling research community adopted a similar approach than the AMIP when the first global or basin-scale eddy-permitting ocean simulations were achieved in the 1990's. 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 parametrisation were identified. The DYNAMO project, dedicated to offer intercomparison among three classes of ocean models of the North Atlantic Ocean in a similar numerical experiment framework (Meincke et al., 2001) allowed to identify patterns of the North Atlantic Ocean circulation that were robust, and others that were sensitive to model parameterization. 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 as an updated knowledge of the North Atlantic Circulation. In other terms, identifying the "ensemble pattern" from the simulations 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 at providing common references for consistent assessment from a multi-model perspective (Griffies et al., 2009). The CORE-I intends at evaluating 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 eighteen time-dependant ocean numerical simulations have yet been performed, 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/biogeochemistry simulations forced by common atmospheric data sets (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 ESM. 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, an OMIP version 2 is ongoing using the more recent JRA-55 reanalysis forcings (Kobayashi et al., 2015). Metrics ocean –diagnostics– endorsed by the OMIP are those recommended for the assessment of the 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, then intercompare the simulations in order to evaluate model's performance with regard to observed realistic references. Bringing better characterisation of model errors and weaknesses considering specific ocean processus, from the physical to the biogeochemical aspects, over decadal, interannual and seasonal time scales. Implicitly, these efforts have involved strategies for distributing, storing, sharing simulations and metrics, under constraints of computing server limitations in capacity and communication bandwidth. In practice, this brought to the common technical definition of standards shared by all participants, and a fit-for-purpose evaluation framework, to be applied in similar ways on every simulation. And finally, carry out a common synthesis effort 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 provide "ocean synthesis" on ocean state estimation through a chosen series of Ocean Essential Variables (EOV) 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 being 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; (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 eighteen 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 maintain 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. 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; (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 (Garcia-Sotillo et al., 2024), and that 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), then to carry out the evaluation through intercomparison and validation against "truth references" (Hernandez et al., 2011).

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

2.1 From academia to operation: adoption of best practices

The legacy of the first ten 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 being adopted as "best practices" 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 OOFS and models brought the following aspects to address:

- Characterize the performance of individuals OOFS of the same kind relatively to a given "truth", identify outliers, give clues for further OOFS improvements
- Allow "ensemble estimation" that provides qualitatively more robust and reliable estimates: the "ensemble mean" approach. In some cases, if the "ocean truth" is missing, the ensemble mean can be considered as a reference, and be used to validate individually the systems.
- Provide an *ad hoc* methodology for operational qualification, see Garcia-Sotillo et al. (2024) in this report 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, 2015b), and pre-figured the FAIR best practices (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-Paris AirFrance plane in 2009, and the subsequent intensive search for the wreckage in the Tropical Atlantic. Evaluating the accuracy of current fields from OOFS 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). 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 OOFS and their products is now a common practice in operational centres. However, various aspects need to be reminded and addressed:

- Common validation/verification methodology needs to be adopted by all participants, preferably adopting recommendations, as reminded in this report (Garcia-Sotillo et al., 2024).
- Interoperability, shareability of products and common standards is key: the large number of products offered by the different centres cannot be spread in every single centre. The FAIR best practices are essentials.
- 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:
    - Re-gridding by downscaling or upscaling ocean products toward a common grid might generates 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) datastore 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: there is a promising way in using novel approaches based in 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 more and more implemented in operational centres. They are also new insight for establishing the performance of intercompared OOFS into the user oriented framework.

3 Ongoing ocean models and forecasting systems intercomparison activities

3.1 Class4 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 programs such as Argo, 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 Garcia-Sotillo et al. (2024) in this report. In summary, Class 4 metrics aims to compare observations with the equivalent model forecast at the same time and place, for different lead-time (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 OOFS can similarly be compared to the same data, in the same manner. Hence Class 4 metrics since the beginning were used when comparing several OOFS 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/time of the observation used as reference.

Within GODAE OceanView, the Class 4 intercomparison project is operating since 2013. A first set of intercomparison of 6 global OOFS (Ryan et al., 2015) was an opportunity to present new metrics (radar plot, Taylor Diagrams, best systems mapping, bar charts, rank histograms…). The same Class 4 information was also used with more specific metrics around Australia (Divakaran et al., 2015), with the objective for 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 inhouse routine validation in Australia was taken 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 (Figure 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 forecast and many other drift dependant applications.

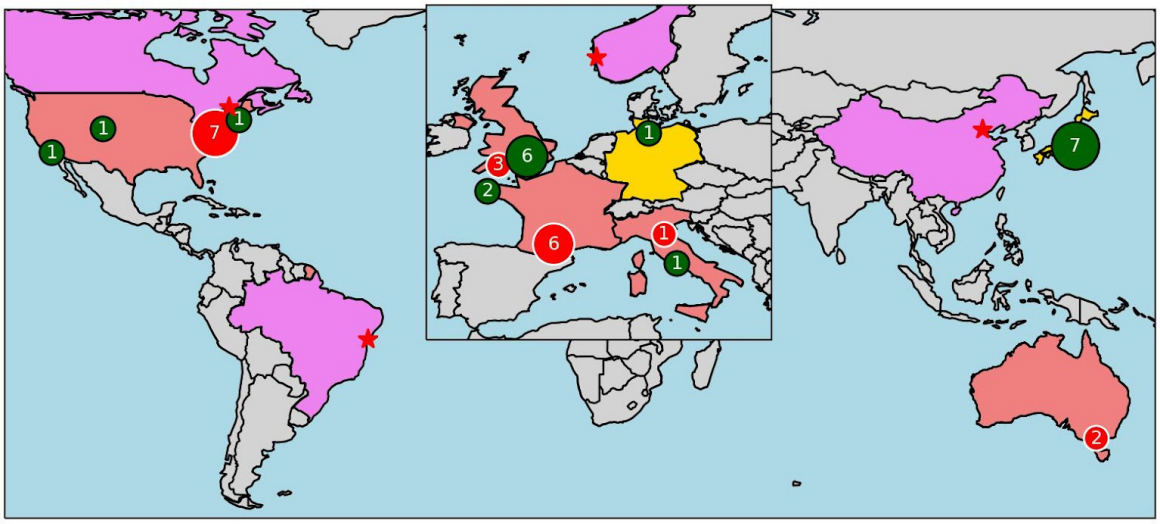

Figure 1: Operational centres and countries involved in a common intercomparison international framework during the last 20 years. Circles indicate their size and numbers the products/locations participating in the ORA-IP (Balmaseda et al., 2015). Green circles for ORA-IP only, and red circles for centres 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 violet, yellow, orange, contribute respectively to Class 4, ORA-IP, or both exercises.

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 OOFS 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 Nation Ocean Decade Program

(Fujii et al., 2019 ; Fujii et al., 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, to take into account double penalty effects when measuring the skill of a given product for given scales or ocean regimes. With the necessity 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 later does not represent the tidal dynamics, obviously, observations need to be pre-processed to filter-out tidal signals. This is the reason why 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 fit-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, a particular attention needs to be addressed on the representativity and the uncertainty of observations. It is mandatory to take them into account while comparing several OOFS 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 in parallel a series of deterministic forecasts where some initial or forcing conditions were stochastically modified between members. With the purpose of performing the intercomparison of the forecast members in order 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 centre in charge of marine meteorology forecast, i.e., wind and wave forecast. 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, that concluding, among other results, that ensemble wave skill score at day 10 outperformed deterministic one at day 7 (Campos et al., 2018). Other example, the recent intercomparison of seasonal ensemble forecasts from two centres contributing to the Copernicus Climate Change Service (C3S) quantifying 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 forecast routinely (e.g., Lisæter et al., 2003 ; Keppenne

et al., 2008 ; Seo et al., 2009). However, there is not yet achieved large community effort dedicated to intercomparisons of ensemble forecasts produced by different centres.

We propose to illustrate here ensemble approach benefits with a multi-system intercomparison as proposed by the CLIVAR/GSOP initiative (mentioned above) and the ORA-IP project (detailed in section 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). All products –observation-derived-only and reanalysis estimates (see Balmaseda et al. 2015, for product's 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 degrees 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 averaging of various reanalyses (Masina et al., 2015). The Fig 2b, shows the time series of the so called "SST 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 time scales. 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 "SST index" with the ensemble average is given by Fig 2a, quantified by their respective root-mean-square differences. Before 1993, the ensemble-ensemble average is computed only with the ERA5 reanalysis and the OSTIA observation-derived-only product, covering this period. Consequently, Fig 2a exhibits the 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 impact of their respective representativity of SST in the ATL3 box, and possibly mapping/observation errors to be further investigated. 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. Then, 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", that 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 in similar ways OOFS variables on specific model grids and layers (Hernandez et al., 2015b). In the Class 1 approach, OOFS outputs are regridded and resampled in a common grid and timeframe (e.g., daily 2D model fields) for being 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 OOFS 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.

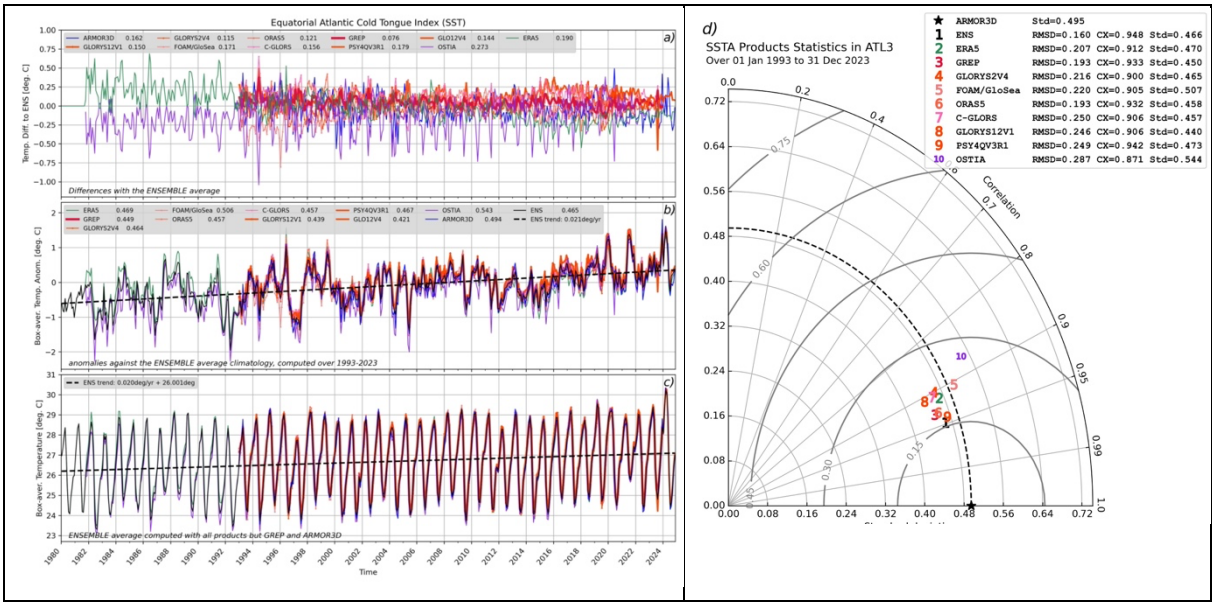

Figure 2: Left: Time series from 1980 to 2024 of SST products, monthly and spatially averaged into the ATL3 box located in the eastern equatorial band [20°W-0°E / 2.5°S-2.5°N] of the Tropical Atlantic. a) Differences relative to the ensemble-average (root-mean-square differences indicated in label).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

3.3 Regional forecast intercomparison and nesting strategy evaluation

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

On a regional basis, specific systematic multi-product validation tools are gradually developed (e.g., Lorente et al., 2016 ; Lorente et al., 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 rely on specific local observing systems, like HF-radar, offering an “ocean truth” representing the ocean dynamics at higher resolution (Kourafalou et al., 2015), that cannot be represented by global OOFS.

However, it is worth noting that comprehensive multi-product operational intercomparison is not common at regional scales. Unlike global OOFS, there are rarely many fine scales regional OOFS 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 OOFS would provide little information compared with the global intercomparison initiatives already underway.

But there is an increasing number of operational centres, or programs like the CMEMS, that operate over the same area both regional and global systems, and that started to intercompare their different systems. For instance, comparing two OOFS of the same kind, Mercator and MFS, in the Mediterranean Western Basin, and evaluating their respective strengths and weaknesses over specific subdomains (Juza et al., 2015). Or measuring the benefit of improving the resolution of a regional OOFS 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 proposing 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 be also be given for the Canadian Arctic and North Atlantic regional OOFS (Dupont et al., 2015), the USA East Coast OOFS and reanalyses (Wilkin et al., 2022), or the Australian global and regional OOFS evaluations that focus on specific case studies and applications: disaster/search and rescue, defence/acoustic, or 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 errors propagate from the global to the nested system and degrade the expected seamless transition from open ocean to coastal dynamics? How specific ocean processes of interest are represented in the different systems? How reliable they can appear for end-user needs in the 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 by the academia to study ocean processes over long periods, with the support of the increasing amount of ocean observations over time, and the improvement of assimilation techniques. Evaluation of such reanalysis representing decades of ocean behaviour through comprehensive intercomparison projects require large resources and preparation. Most are conducted outside of routine operations by forecasting centres. They represent a milestone in progress in the field, both from the point of view of the evaluation of the system/reanalysis itself, and of the new validation methodologies 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 from 20 to 50 years, and focusing on eight EOV - Ocean Heat Content, Steric Height, Sea Level, Surface Heat Fluxes, Mixed Layer Depth, Salinity, Depth of the 20°C Isotherm, and Sea Ice (Figure 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 and that cannot be sampled with a single system. Uncertainty sensitive to the temporal variations of the observing system, 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 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 EOV. Results highlighted impacts of model resolution, components of the observing system assimilated, complexity of the ocean models, of the data assimilation scheme, and quality of external forcing (Palmer et al., 2017; Shi et al., 2017; Storto et al., 2017; Toyoda et al. 2017a, 2017b; Valdivieso et al., 2017; 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 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 help in identifying the impact of their absence in the past, and define where they are most decisive in the quality of present and future reanalyses. And consequently, provide suggestions for improvements of the GOOS.

Figure 2 shows that multi-product intercomparisons allow to infer key indicator of the ocean environment changes together with estimates of their uncertainties. Beyond reanalyses assessment based on EOV, next stage of ocean reanalysis intercomparison should first target key ocean processes that affect the climate system, identify their past occurrences, better unravel their mechanisms and interactions, in order to estimate their present and future impacts. Machine learning approaches are expected to explore more systematically ocean variability in a multi-system framework 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 decisive: more realistic clusters of ocean reanalyses with better characterisation 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 Ocean Prediction 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 technical challenges in term 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: the ocean state monitoring

An important outcome of the ORA-IP has been the development of the Real Time Multiple Ocean Reanalysis Intercomparison, carried on routine every month by NOAA/NCEP, which 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 a growing importance in operational oceanography: it offers through key EOV 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 (OMI), whose reliability and uncertainty are estimated through intercomparison of multiple products. Using the OMI, the CMEMS started in 2018 to produce on annual basis the Ocean State Report (von Schuckmann et al., 2018) now on its 8th edition (https://marine.copernicus.eu/access-data/ocean-state-report , last access 29 January 2025).

References

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

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.

Balmaseda, M. A., Hernandez, F., Storto, A., Palmer, M. D., Alves, O., Shi, L., … Gaillard, F.: The Ocean Reanalyses Intercomparison Project (ORA-IP). Journal of Operational Oceanography, 8(sup1), s80-s97. https://doi.org/10.1080/1755876X.2015.1022329 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, Frontiers in Marine Science, 11, doi: 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, Frontiers in Marine Science, 11, doi: 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(3). 14-21. http://dx.doi.org/10.5670/oceanog.2009.62 http://dx.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, W. Krauss, Editor. Gebrüder Borntraeger: Berlin - Stuttgart, 91-128. , 1996.

Chevallier, M., Smith, G.C., Dupont, F. et al.: Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project. Climate Dynamics, 49, 1107-1136. https://doi.org/10.1007/s00382-016-2985-y https://doi.org/10.1007/s00382-016-2985-y, 2017.

Cailleau, S., Bessières, L., Chiendje, L., Dubost, F., Reffray, G., Lellouche, J.-M., van Gennip, S. J., 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, doi: 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, doi: 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, Frontiers in Marine Science, 7, doi: 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, Meteorological Applications, 15, 3-18, doi: 10.1002/met.52, 2008.

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, 2019.

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

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, doi: 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. http://dx.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 the GODAE Coastal and Shelf Seas Working Group Community: Application in Coastal Modelling and Forecasting. Oceanography, 22(3). p. 198-205. https://doi.org/10.5670/oceanog.2009.79 https://doi.org/10.5670/oceanog.2009.79, 2015.

Divakaran, P., Brassington, G. B., Ryan, A. G., Regnier, C., Spindler, T., Mehra, A., … Davidson, F.: GODAE OceanView Inter-comparison for the Australian Region. Journal of Operational Oceanography, 8(sup1), s112-s126. https://doi.org/10.1080/1755876X.2015.1022333 https://doi.org/10.1080/1755876X.2015.1022333, 2015.

Drévillon, M., Greiner, E., Paradis, D. et al.: 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. https://doi.org/10.1007/s10236-012-0580-2, 2012.

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, doi: 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, Journal of Operational Oceanography, 14, 1-12, doi: 10.1080/1755876X.2019.1602102, 2021.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics, 53, 343-367, doi: 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, doi: 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, Frontiers in Marine Science, 6, doi: 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, Frontiers in Marine Science, 11, doi: 10.3389/fmars.2024.1476131, 2024.

Garcia-Sotillo, M., Drevillon, M., and Hernandez, F.: A description of Validation Processes and Techniques for Ocean Forecasting, State Planet Discuss., 2024, 1-11, doi: 10.5194/sp-2024-33, 2024.

Gates, W.L.: AN AMS CONTINUING SERIES: GLOBAL CHANGE--AMIP: The Atmospheric Model Intercomparison Project. Bulletin of the American Meteorological Society, 73(12), 1962-1970. https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2, 1992.

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, doi: 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, 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, doi: 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., 2010.

Hartigan, J.A., and Wong, M.A.: Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108. https://doi.org/10.2307/2346830, 1979.

Hernandez, F.: Performance of Ocean Forecasting Systems - Intercomparison Projects. In: Schiller, A., Brassington, G. (eds) Operational Oceanography in the 21st Century. 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., … Zhang, A.: Recent progress in performance evaluations and near real-time assessment of operational ocean products. Journal of Operational Oceanography, 8(sup2), 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(3), 128-143. https://doi.org/10.5670/oceanog.2009.71, 2015b.

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

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. Mar. Sys., 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, Journal of Operational Oceanography, 9, s155-s166, doi: 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, Monthly Weather Review, 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, Journal of the Meteorological Society of Japan. Ser. II, 93, 5-48, doi: 10.2151/jmsj.2015-001, 2015.

Kourafalou, V. H., De Mey, P., Le Hénaff, M., Charria, G., Edwards, C. A., He, R., … Zhu, X.: Coastal Ocean Forecasting: system integration and evaluation. Journal of Operational Oceanography, 8(sup1), 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, Climate Dynamics, 33, 341-364, doi: 10.1007/s00382-008-0441-3, 2009.

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

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.

Lorente, P., Piedracoba, S., Sotillo, M. G., Aznar, R., Amo-Baladrón, A., Pascual, Á., … Álvarez-Fanjul, E.: Ocean model skill assessment in the NW Mediterranean using multi-sensor data. Journal of Operational Oceanography, 9(2), 75-92. https://doi.org/10.1080/1755876X.2016.1215224, 2016.

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, Climate Dynamics, 1-29, doi: 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, Transactions American Geophysical Union, 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). Progress in Oceanography, 48(2-3), 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 Forecasting. Weather and Forecasting, 8(2). 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, doi: 10.5194/gmd-10-2169-2017, 2017.

Palmer, M.D., Roberts, C.D., Balmaseda, M. et al.: Ocean heat content variability and change in an ensemble of ocean reanalyses. Climate Dynamics, 49, 909-930. https://doi.org/10.1007/s00382-015-2801-0, 2017.

Ryan, A. G., Regnier, C., Divakaran, P., Spindler, T., Mehra, A., Smith, G. C., … Liu, Y.: GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison. Journal of Operational Oceanography, 8(sup1), s98-s111. https://doi.org/10.1080/1755876X.2015.1022330, 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, Quarterly Journal of the Royal Meteorological Society, 137, 879-893. https://doi.org/10.1002/qj.816, 2011.

Salman, A.: Editorial: Application of machine learning in oceanography and marine sciences, Frontiers in Marine Science, 10, doi: 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, Journal of Operational Oceanography, 13, 1-18, doi: 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, 2009.

Shi, L., Alves, O., Wedd, R. et al.: An assessment of upper ocean salinity content from the Ocean Reanalyses Inter-comparison Project (ORA-IP). Climate Dynamics, 49, 1009-1029. https://doi.org/10.1007/s00382-015-2868-7, 2017.

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, Environmental Research Letters, 16, 073008, doi: 10.1088/1748-9326/ac0eb0, 2021.

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. Proceedings of OceanObs09: Sustained Ocean Observations and Information for Society (Vol. 2), Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306, 2009.

Storto, A., Masina, S., Balmaseda, M. et al.: Steric sea level variability (1993–2010) in an ensemble of ocean reanalyses and objective analyses. Climate Dynamics, 49, 709-729. https://doi.org/10.1007/s00382-015-2554-9, 2017.

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, Frontiers in Marine Science, 6, 418, doi: 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, Frontiers in Marine Science, 6, doi: 10.3389/fmars.2019.00471, 2019.

Toyoda, T., Fujii, Y., Kuragano, T. et al.: Intercomparison and validation of the mixed layer depth fields of global ocean syntheses. Climate Dynamics 49, 753-773. https://doi.org/10.1007/s00382-015-2637-7, 2017a.

Toyoda, T., Fujii, Y., Kuragano, T. et al.: Interannual-decadal variability of wintertime mixed layer depths in the North Pacific detected by an ensemble of ocean syntheses. Climate Dynamics, 49, 891-907. https://doi.org/10.1007/s00382-015-2762-3, 2017b.

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, Climate Dynamics, doi: 10.1007/s00382-018-4242-z, 2018.

Valdivieso, M., Haines, K., Balmaseda, M. et al.: An assessment of air–sea heat fluxes from ocean and coupled reanalyses. Climate Dynamics, 49, 983-1008. https://doi.org/10.1007/s00382-015-2843-3, 2017.

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. Geophysical Research Letters, 47(8), e2020GL087108. https://doi.org/10.1029/2020GL087108, 2020.

von Schuckmann, K., Le Traon, P. Y., Smith, N., Pascual, A., Brasseur, P., Fennel, K., … Zuo, H.: Copernicus Marine Service Ocean State Report. Journal of Operational Oceanography, 11(sup1), 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, doi: https://doi.org/10.1016/j.pocean.2022.102919, 2022.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.: 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. et al.: A real-time ocean reanalyses intercomparison project in the context of tropical pacific observing system and ENSO monitoring. Climate Dynamics, 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.

Competing interests: The authors declare that they have no conflict of interest.

Data and/or code availability

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

ERA5 DOI: 10.24381/cds.f17050d7

OSTIA DOI : 10.48670/moi-00165

GLORY12V1 DOI: 10.48670/moi-00021

ARMOR3D DOI : 10.48670/moi-00052

GLO12V4 and PSY4QV3R1: 10.48670/moi-00016

GREP and FOAM/GloSea and C-GLORS and ORAS5 and GLORYS2V4 DOI: 10.48670/moi-00024

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

Authors contribution

FH wrote the article, and produced Figure 1 and Figure 2. MGS and AM participated to the initial redaction and the overview of the article.

Acknowledgements

Fabrice Hernandez contributed as IRD researcher to this publication initiative of the Decade Collaborative Centre for Ocean Prediction as part as an in-kind IRD contribution to Mercator Ocean International in 2024 and 2025. Fabrice Hernandez thanks the Ocean Predict Intercomparison and Validation Task Team for the continuous efforts on providing Class 4 files for routine intercomparaison that can be discussed in this article. Angélique Melet is supported by SEACLIM and FOCCUS, which has received funding from the European Union's Horizon Europe programs under grants agreement no 101180125 and 101133911 respectively. 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 (funded in the framework of the E.U. CMEMS Contract 21002L6-COP-MFC IBI-5600). The authors thank the reviewers for their helpful comments that suggested an extended and updated view on intercomparison activities.