https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

STATE OF THE PLANET Discussions

A description of Model Intercomparison Processes and Techniques for Ocean Forecasting

Fabrice Hernandez[1], Marcos Garcia Sotillo[2], Angelique Melet[3]

[1]LEGOS, Institut de Recherche pour le Développement, Toulouse, France
[2]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 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 for addressing monitoring of ocean state in operations.

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 modeling 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 Chapter 2.15. 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 AMIP simulations). This process involved analyzing a series of targeted

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

key variables extracted from the model state to provide an overview of the model's skill in representing various atmospheric aspects.

The ocean modeling research community adopted a similar approach 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 scale simulated by these models.

An obvious aspect of these intercomparison exercises was the community effort, trying to commonly define a modelling strategy, conduct the simulations individually, and be able to store them, in order to enable exchanges among participants. Then design 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 CLIVAR: the Global Synthesis and Observation Panel (GSOP) project aimed to intercompare different ocean reanalysis computed over several decades, and provide "ocean synthesis" on ocean state estimation for 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 factors of discrepancies among them. The idea being that multi-system ensemble approaches should be useful to obtain better estimates of the ocean. 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 data 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 behavior 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 WMO framework (mentioned in Section 4.2.3). A first dedicated intercomparison of ocean operational systems, operated on routine, was achieved by the 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 centers; (c) evaluate the overall scientific quality of the different GODAE operational systems. The preliminary task was to define the

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

validation concepts and methodologies (Hernandez et al., 2015a), also referenced in the chapter 2.15 above, 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 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 1) the implementation of a 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 OceanPredict. And 2) proposing the validation and intercomparison methodology, improved and tested regularly since, until being adopted as "best practices" and recommended by 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 Section 4.3.2) above 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 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 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).

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

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 centers. 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 Section 4.3.2.
- Interoperability and common standards is key: the large amount of products offered by the different centers can not be spread in every single center.
- Representativity is a central aspect of intercomparison: scales and processes represented in each product (observations and models) need to be correctly documented to reduce mis-interpretation when intercompared. Moreover, the CMEMS clearly shows that for a given Essential Ocean Variables (EOV), a large amount of products (from observations or models) are now provided, and should be properly "used" in an intercomparison exercise.
- 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 consensus clustering, machine learning, and other tools developed in the frame of ensemble estimation and forecast.
- User oriented metrics and process oriented metrics are more and more implemented in operational centers. They are also new insight for establishing the performance of intercompared OOFS into the user oriented framework.

3 Ongoing model intercomparison activities

3.1 Class4 metrics: model intercomparison in the observation space for verification forecast

Class 4 metrics aims to compare observations, and the equivalent model forecast at the same time and place, for different lead-time (Hernandez et al., 2015a). These metrics, for different kinds of ocean variables, were first designed to measure 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., 2015). Within GODAE OceanView, the Class 4 project has been 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 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 shared multi-system intercompared Class 4 information to enhance its own daily basis verification procedures. The Class 4 intercomparison is still routinely performed (Figure 1).

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

Another issue of Class 4 comparison to observation was the routine evaluation of the overall quality of the global ocean observing system. 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 centers that could carry out a second loop of data corrections, for the benefit of all data users (Hernandez et al., 2015b).

Indirectly, comparison to observations raises the key issue of representativity, both from the observation and the modelling side. What are the scales sampled by a given observing system? What are the effective scales and ocean processes represented by a given OOFS? The classical example is comparison of satellite altimetry and/or tide gauge observation 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 L4 observation 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 behavior of atmospheric simulations. Obviously, the associated approach is to intercompare forecast members in order to 1) identify common patterns 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.

Figure 1 illustrates the assessment of a common 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. The left middle panel, with the standard deviation associated with the ATL3 box averaging, indicates the shorter space-scale variability provided by each product in the box. This also gives an indication on the confidence level on the box-mean estimation, considering classical Gaussian distribution statistics. In addition, using one of the observed products -here OSTIA SST- as a reference, the Taylor diagram quantifies the relative value of each individual product in terms of differences and correlations. Seasonal climatology removed; anomalies of each product can be used to infer the "Atlantic Niño" index (top right panel). Spanning on a shorter temperature range, these time series show more discrepancies, and outliers can be better identified, compared to the ensemble pattern. The decadal warming trend is visible, with an increase in the very last years.

Note that ensemble clustering, a recent methodology, also called consensus clustering aims at producing a synthesis among an identified cluster from a given dataset. The construction of the clusters from the initial dataset (here the different members of

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

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). 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.

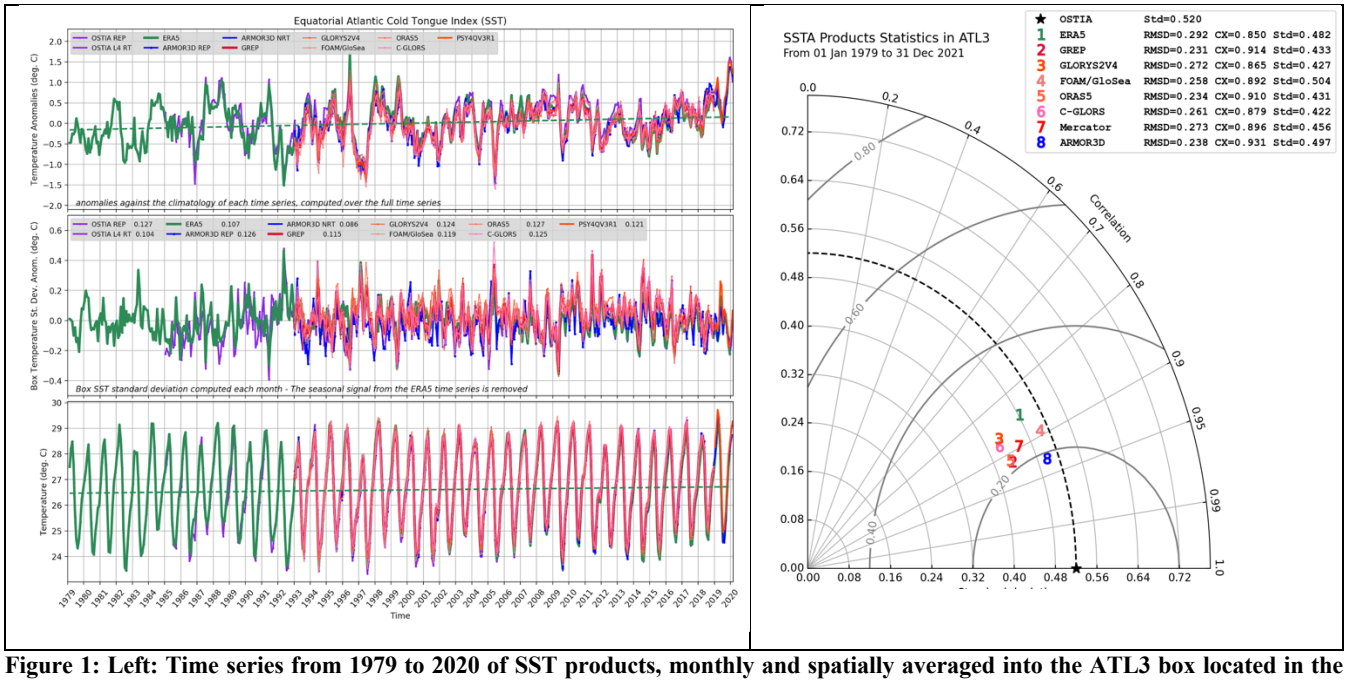

Figure 1: Left: Time series from 1979 to 2020 of SST products, monthly and spatially averaged into the ATL3 box located in the eastern equatorial band [20°W-0°E / 2.5°S-2.5°N] of the Tropical Atlantic (bottom the total averaged values, middle the standard deviations associated with the monthly mean values, top the anomalies relative to seasonal climatologies computed for each product). Right: the associated Taylor diagram using as reference the OSTIA SST anomaly product.

3.3 Regional forecast intercomparison and nesting strategy evaluation

During the last decade, the validation methodology proposed by the GODAE global ocean community was adopted by 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 behavior of the ocean under tropical cyclone conditions using several OOFS and ad hoc metrics (Zhu et al., 2016).

On a regional basis, systematic validation tools were recently developed (e.g., Lorente et al., 2016). These tools, operated by a given operational centre, are efficient essentially if an inter-operable dataserver 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

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

coastal system evaluation rely on specific local observing systems, like HF-radar, offering an "ocean truth" at higher resolution (Kourafalou et al., 2015), that cannot be represented by global OOFS. Operational centers, or programs like the CMEMS with both regional and global systems started to intercompare their different systems. In the case of 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). In the case of the CMEMS, several overlapping regional systems can be compared to the global solution (Lorente et al., 2019). In this case, using similar metrics, typically Class 4, for evaluating all these systems brings a series of questions. Which are the scale represented by the child system that is lacking in the parent system, or in the observations? What is the impact of the different kinds of forcings and different kinds 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?

3.4 Evaluating retrospective views of the ocean dynamics: dedicated ocean reanalysis 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 behavior through comprehensive intercomparison projects require large resources and preparation. Most are conducted outside of routine operations by forecasting centers. 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.

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

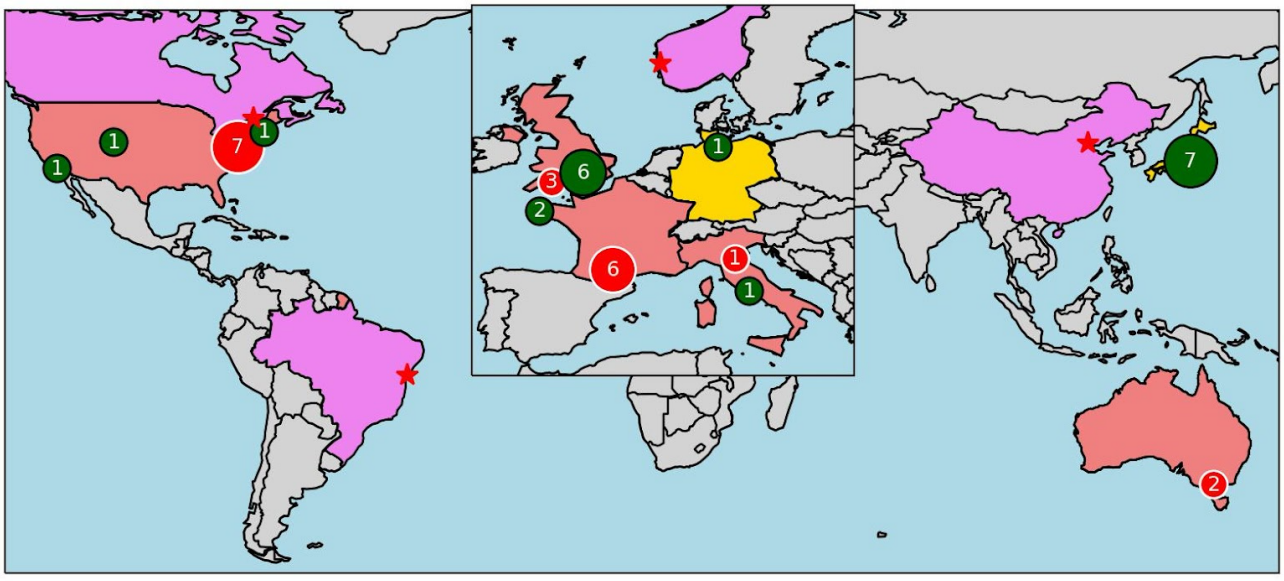

Figure 2: Operational centers and countries involved in a common intercomparison international framework during the last 20 years. Circles indicated with their size and numbers the products/locations participating in the ORA-IP (Balmaseda et al., 2015). Green circles for ORA-IP only, and red circles for centers that are contributing in addition to the Class 4 routine intercomparison (Hernandez et al., 2015). Red stars indicate centers solely participating in the Class 4 intercomparison. Countries in violet, yellow, orange, contribute respectively to Class 4, ORA-IP, or both exercises.

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 2). 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.

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

3.5 From reanalysis intercomparison to 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 centers 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 an annual basis the Ocean State Report (von Schuckmann et al., 2018).

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 https://doi.org/10.48670/ETOOFS, 2022.

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.

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.

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.

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 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.

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., Bertino, L., Brassington, G., Chassignet, E., Cummings, J., Davidson, F., Drévillon, M., Garric, G., Kamachi, M., Lellouche, J.-M., Mahdon, R., Martin, M.J., Ratsimandresey, 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., 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., 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.

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.

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.

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.

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.

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

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.

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.

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.

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.

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.

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.

https://doi.org/10.5194/sp-2024-39
Preprint. Discussion started: 7 October 2024

STATE OF THE PLANET Discussions

315 **Competing interests**

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

Data and/or code availability

This can also be included at a later stage, so no problem to define it for the first submission.

Authors contribution

320 This can also be included at a later stage, so no problem to define it for the first submission.

Acknowledgements

This can also be included at a later stage, so no problem to define it for the first submission.