A description of Model Intercomparison Processes and Techniques for Ocean Forecasting

NEW: corrections after second review (30/01/2025):

- Changes for affiliations for Fabrice Hernandez and Marcos Garcia Sotillo
- Remove sentence: Copyright statement: the copyright statement will be included by Copernicus, if applicable.
- Acronym of authors in "Authors contributions"

RC1: 'Comment on sp-2024-39', Anonymous Referee #1

This paper I believe is intended as a brief review of past efforts and methods to compare ocean state and ocean forecast products that have been developed within the international community. While this is a useful objective I find the current version of the paper has many problems with it especially if read on its own. I understand that it would form 1 chapter of a larger report but I do suggest that someone should be reviewing the report as a whole
The paper does reference other "Chapters" and sometimes non-existent "sections" Lines29,61,66,84,100. These need to be properly checked

Initially this article is a « chapter » or « section » as part of the report published in State of the Planet: « Ocean prediction: present status and state of the art ». Inside this report, this article is strongly associated with the one entitled: «A description of Validation Processes and Techniques for Ocean Forecasting » .
The present article is going to be modified in order to change « chapter » with the exact citation. Like:

Sotillo, M. G., Drevillon, M., and Hernandez, F.: A description of Validation Processes and Techniques for Ocean Forecasting, State Planet Discuss. [preprint], https://doi.org/10.5194/sp-2024-33, in review, 2024.

—> changed at new line 35 with: Garcia-Sotillo et al. (2024, this report).
—> changed old line 61 « *(mentioned in Section 4.2.3) »* by new reference starting at new line 121: 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). 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.
—> changed at new line 144 « *also referenced in the chapter 2.15 above*« by « referenced in this report (Garcia-Sotillo et al., 2024)«
—> changed at new line 163 «(see Section 4.3.2) above « by « see Garcia-Sotillo et al. (2024) in this report for «

—> changed at new line 195 «*as reminded in Section 4.3.2*« by « as reminded in this report (Garcia-Sotillo et al., 2024). «

a supprimé: Section 4.3.2.

The text also has lots of acronyms and other notation that will mean nothing to a wider audience eg ET/OOFS, Class 4, L4 products, "go-no-go"?. I appreciate these may appear in other chapters but they should at least be defined here or cross referenced when first mentioned.

Agreed, even if defined in previous « chapters » of the report, we propose to explicitly define acronyms or « jargon », and acronyms in new text are all explicit:
—> changed at new line 121: « WMO » by « World Meteorological Organization (WMO) «
—> changed at new line 127: « GODAE » by « Global Ocean Data Assimilation Experiment (GODAE) «
—> changed at new line 145: « OOFS » by « operational ocean forecasting system (OOFS)«
—> changed at new line 152: « *OceanPredict* » by « Ocean Predict (https://oceanpredict.org/))«
—> changed at new line 155: « *ETOOFS* » by « the Expert Team on Operational Ocean Forecasting Systems (ETOOFS)«
—> changed at new line 207: « *CMEMS* » by «Coperrnicus Marine Environment Monitoring Service (CMEMS) «
—> changed at new line 299: « *when referring to L4 observation products* » by «when referring to re-processed/re-gridded observation products (also called Level 4 or L4 type of observed products). «

For « class1 » etc …. the definition is given in chapter Garcia-Sotillo et al., 2024, and it is not reproduced here. We propose the following modification at new line 142:
« 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). «

The expression « go/no-go » is self explicit and used by operational teams to decide or not to carry on some action, in this case, put in operation the new system. We propose to keep it at it is.

« NetCDF » is also commonly used in oceanography. We propose to modify new line 168 to : « NetCDF file format «

The AMIP concept is rightly introduced and a valuable concept, but seems odd then not to mention OMIP? And then finally leading to CMIP. The evolution of objectives to define the actual ocean states should then refer to the success of ERA and other atmospheric reanalyses. Emphasising the different objectives of GODAE and CLIVAR in comparing reanalyses for the ocean could then be

explained and would then follow naturally. State estimation and forecasting as different applications.

We would like to thank the reviewer for offering this insight, which we had not initially considered, in our desire to deal directly and uniquely with intercomparisons for operational oceanography. From new line 74 of our first section a full development is proposed to consider CMIP and OMIP efforts, mentioning the CORE reference framework and the recent CMIP6 experiments:

"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 supprimé: 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

Section 3.1 should start by properly justifying the value of observation space verification. The issue of independent v non-independent data comparisons should be discussed. Different kinds of metrics and their usefulness would be useful to summarise more clearly here.

Although Class4 metrics concept is reminded in Garcia-Sotillo et al. (2024) with many past references, we propose to add at the beginning of section 3.1 (new line 220) an

introduction of the Class4 metrics, and also introduce at new line 247 the « independent/non-independent » observation impact on the Class4 evaluation:

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.

Section 3.2 good topic to discuss the value of ensemble comparisons but it does not really discuss the value of ensemble product. Are biases in individual products reduced in this way? Uotila et al 2019 polar comparison is an example of this. L154 top right panel seems wrong? Include reference for Fig 1 legend.

We agree that ensemble forecast needed to be better introduced. We propose in section 3.2 to describe first ensemble forecast initiatives then going to multi-model inter comparison in terms of ensembles through the dedicated exemple of Figure 2 (previously figure 1):

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.

a supprimé: behavior

a supprimé: Obviously,

a supprimé: associated approach is to intercompare

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).

There is no reference for Fig 2 because we have produced this original figure from products extracted at the Copernicus Marine and Climate Service DataStore. We propose to upgrade the figure (now from 1980 to 2024) with a more explicit representation of the ENSEMBLE average from all reanalyses, then discuss the merit of this ensemble in section 3.2. The figure caption of Figure 2 is changed with more explicit description, because we propose to change the upper/middle left panels of Figure 2: we introduce on the top panel the differences of each product against the ENSEMBLE mean, and we move the former top panel (Box-averaged SST anomalies relative to a common climatology, called also the « SST index ») at the middle. Instead of using the OSTIA SST as a reference in the statistics of the Taylor Diagram, we use ARMOR3D, that offers longer time consistency (product till November 2024). With upgraded Figure 2, we propose to fully change the text of this section 3.2, with a reference to Uotila et al (2019) with the assessment of the ENSEMBLE estimate:

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

a supprimé: common

a supprimé:

a supprimé: (

a supprimé: ,

a supprimé: (

a supprimé:)

a supprimé: .

a supprimé: left middle panel, with

a supprimé: deviation associated with the ATL3 box averaging, indicates the shorter space-scale variability provided by

a supprimé: 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,

a supprimé: one of the observed products -here OSTIA SST- as a reference,

a supprimé: 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…

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.

a supprimé: can be better identified, compared to the ensemble pattern. The decadal warming trend is visible, with an increase in the very last years.

Sections 3.3, It would be nice if this issue of regional studies was brought more up to date? The references are all rather old? These and the following 2 sections are very brief

We agree with the review and propose for this section an extended discussion with recent references to inter comparison at regional scales, considering global versus regional system intercomparison. We also remind that inter comparison at regional scales depends on the reduced number of regional systems that overlap into a given area, and we discuss the importance of assessing the error propagation from « parents » model that feed boundary conditions for « child » regional models:

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).

a supprimé: During

a supprimé: decade

a supprimé: was

a supprimé: behavior

a supprimé: 2016).

a mis en forme : Police :Italique, Anglais (E.U.)

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.

a supprimé: were recently

a supprimé: dataserver

a supprimé: Operational centers

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

a supprimé: with

a supprimé: In the case of

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?

a supprimé: In the case of the CMEMS, several

a supprimé: scale

a supprimé: kinds

a supprimé: kinds

3.4 This section is so brief and in no way lives up to the promise of the long heading. What are the key points?

This section aims at describing the past, present and future effort on intercomparing ocean reanalyses produced by operational centres. Then suggest areas of improvements for the intercomparison activities. We propose to highlight more specifically Key points and to provide hints for future inter-comparison efforts.

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

a supprimé: 3.5 From reanalysis intercomparison to ocean state monitoring¶

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 Again very brief and the title seems to promise a future look but this is entirely absent.

We propose to change the title of the section by: « A perspective of ocean reanalyses intercomparison: the ocean state monitoring «
Then, keep the description and reference of Ocean State Monitoring. This is short, but we do not see much more to say on it.

I would say this paper needs more attention before it should, be published. It needs more careful reading through and some thought given to making it more accessible for the wider audience. For State of the Planet is "expert-based assessments of academic findings curated for a wider audience to support decision making"

We fully agree with the idea to provide an article for a large audience. Our propositions above are made in this sense: less acronyms, more explicit text, and dedicated references for more in-depth interested readers.

RC2: Comment on sp-2024-39', Anonymous Referee #2, 15 Nov 2024

This article presents an historical and exhaustive review of the different intercomparison exercises done in the Ocean Forecasting community, including ocean reanalysis intercomparison exercises. Methods and outcomes are discussed. The paper is well-written, but I would suggest few improvements to make it easier to understand to readers outside the operational forecasting community. There are many acronyms that should be defined the first time they appear in the text (GODAE, CLIVAR, OOFS, CMEMS, ORA-IP, ETOOFS) and some "internal vocabulary" that can be explained in common words as class1, class2 and class3.

We thank the reviewer for the general comments. We propose some changes all along the text in order to facilitate the reading for none specialists. We have try to be more explicit in the text, and define all acronyms, and give more explanations considering the technical vocabulary like « Class 4 » (please see answer to reviewer 1 above that made similar comment)

The addition of a concluding section highlighting the challenges and opportunities that are coming with, for example, the ensemble approaches for analysis and forecasts, the higher resolution system with an increased data volume to handle and intercomparison methods based on machine learning would make the paper more impactful even if mentioned previously in the different sections.

In order to not totally modified the frame of this article, we kept all existing sections. Section 3.4 gives a series of conclusions and recommendations. Section 3.5 is a specific perspective for Ocean State Monitoring, that inherit from intercomparison and ensemble approaches. Reason why the section 3.5 title is modified: "A perspective of ocean reanalyses intercomparison: the ocean state monitoring".

Nowadays, most intercomparison experiments carried out by the operational oceanography community are based on "multi-system" comparison, and not multi-ensemble comparison (as it can appear now in the weather or climate forecast community). Reason why this is specifically introduced in the section 3.2. Then the end of section 3.4 (new lines 613-628) has been upgraded with specific considerations on recommendations for the intercomparison framework and ensemble forecast.

The problem of the double penalty when comparing products at different resolution is not mentioned. This can be done in section 2.2, when discussing the representativity or 3.1 with the class4. It is also related to the regridding approach used in some intercomparison projects.

We thank the reviewer to raise the double penalty effect on verification procedures. We propose to mention this aspect in both sections 2.2 and 3.1 when discussing representativity issues:

- 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.

a supprimé: 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….

- 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.

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?

I found the section 3.2 confusing. First ensemble approach is related to forecast ensemble, from the same system, but then ensemble is related to an ensemble of forecasts coming from different OOFs. Intercomparison exercises will offer more opportunities if involving ensemble forecasts, with possible comparison of the different spread characteristics.

As said above, we propose to correct this section: with at the beginning (new line 302) an introduction to ensemble forecast from individual operational systems, with references on what is done in the atmospheric community. Then moving to multi-system inter comparisons through the specific example proposed in Figure 2, that has been upgraded in order to better highlight the relative merit of the ensemble-mean.

In section 3 describing intercomparison exercises in different context, I would suggest adding the UN Decade SynObs project, intercomparing different Observing Systems Experiments (OSE) to assess the impact of diverse ocean observing systems on different OOFs.

We thank the reviewer for this valuable comment. We propose to introduce the topic of "impact and importance of the observing system" and the SYNOBS project in section 3.1 (modification starting at new line 270):

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

a supprimé: observation

a supprimé: global ocean observing system

a supprimé: centers

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.

Line by line comments

l.62: I would suggest adding "ocean operational forecasting system" to differentiate from other ocean operational products based only on observations and from line 49 dealing with the 1st intercomparison also but for ocean reanalysis.

We propose to re-write all the introduction given in section 1, providing a description of the evolution of CMIP to OMIP projects under CLIVAR

l.80: the spread of the ensemble is also used as an uncertainty estimation. In the atmospheric community it may be more often seen as a reference (verification against analysis) which is less common in the oceanographic community.

We propose to re-write partly this section. Note that we do not detail here the Class4 framework, reminded in Garcia-Sotillo et al. (2024), where forecast skill are compared to persistence of the "analysis" or the "hindcast" most of the time considered as the "best estimates" that are the closest states to the "ocean truth".

l.106 to 109: You may refine the potential use of those emerging methods. This could also be addressed in a conclusion section.
l.114: to compare discrete observations?

We propose now to emphasize IA derived techniques more carefully with ad-hoc references in our chapter 3

l.131: scales and processes even for observations, especially the remote ones that are the result of complex treatments.

We propose to address more in depth issues on scales and representativity of observed products, when used as reference for inter comparison .

l.159: the definition/examples of class1 metric should be given, for example: daily 2D and 3D model fields on a common grid.
We propose to exemplify Class1 metrics

Figure 1: legends are very small.

We propose to reprocess and upgrade Figure 1

l.176: can you describe in few words the tools developed?

We propose to complement the text using Lorente et al., (2019) reference.