iGOR - emulating Global Ocean Reanalyses for ocean digital twins

Revised version, as submitted on Nov. 29 2024, with changes in blue color

1. Description of the task:

The purpose of the proposed challenge is to provide a framework for developing and intercomparing neural emulators capable of reproducing the temporal evolution of state of the global ocean as described by ocean reanalyses, over different time horizons and resolutions. The prediction task consists in the forecast of the time evolution of the global 3D ocean state vector (U, V, T, S) and relevant 2D variables (SSH, SIC) on a fixed grid from perfectly known initial conditions, under different time-varying surface heat and freshwater forcings. Training and evaluation data will be based on the latest GLORYS12 reanalysis. Two versions of the challenges will be considered, a first version will be set-up on the native GLORYS12 model grid resolution (1/12°, 50 levels) another version will be set-up at coarser resolution for prototyping (1/2°, 30 levels).

2. Datasets:

The main data source that will be used for this challenge will be the global ocean reanalysis GLORYS12 produced by Mercator Ocean International for the Copernicus marine service (Lellouche et al. 2021, see also the data product description on CMEMS portal[1]). This dataset will be used for both training and evaluation. This choice is in line with the strategy adopted by several groups (see for instance Wang et al. 2024). We propose to use the latest version of the GLORYS12 reanalysis that has been produced in 2024, which covers the period 1993-2024. In addition, because the typical downstream application use-case for the challenge is the design of AI-native ocean prediction systems, we propose to train the emulators with the initial conditions and surface forcings (heat and freshwater) used in the context of operational predictions. The data challenge training data will therefore include nowcasts from CMEMS global prediction system and atmospheric fields from ECMWF analyses. Given the sheer size of these datasets, we will a priori only share a reduced resolution version of the fields (1/2° grid, daily averaged) as part of the data-challenge. How to optimally share native 1/12° grid resolution fields in a full resolution version of the challenge will be discussed with the PPR Ocean & Climate technical team. The first 26 (1993~2019) years of the reanalysis will be used for training, the last 5 years will be kept for a posteriori evaluation.

3. Evaluation metrics:

We propose to consider different evaluation metrics that can be divided into three categories:

- ***Short-term accuracy metrics in 3D state space:*** this first series of metrics will be used for training and for a posteriori evaluation during prototyping phases. It will assess the ability of the emulator to reproduce the dynamics of the training dataset over short time horizons (typically <30 days). In practice, this will consist in Root Mean Square Errors (RMSE) and Anomaly Cross Correlations (ACC) computed for different variables, and split over different geographical areas (ocean basins, latitude

[1] GLORYS12 reanalysis data product description : https://doi.org/10.48670/moi-00021

bands), and/or regions with specificities in terms of dynamics or relevance for applications (western boundary currents, coastal areas, polar oceans). Short term metrics relevant to specific targets (as for instance the prediction of marine heat waves) will also be discussed with the broader PPR Ocean & Climate data challenge team.

- ***Short-term accuracy metrics in observation space :*** this second series of metrics will be used for a posteriori evaluation only. It will assess the quality of the predictions provided by the emulators with state-of-the-art metrics used in a short term operational prediction center. This series of metrics will be inspired by GODAE OceanView Intercomparison and Validation Task team (GOV IV-TT) Class-4 metrics (Ryan et al. 2015). In practice, it will be based on emulator predictions at different lead times interpolated in observation space (nadir altimeter, ARGO floats, …). We think that this set of metrics should be based on a metric package to be shared across several data challenges.

- ***Physical consistency metrics :*** this third series of metrics is inspired from the CLIVAR Ocean Model Intercomparison protocols (OMIP) (Griffies et al. 2016). It is aiming at assessing the potential of emulators to be used for longer simulations (from seasonal forecasts to decadal predictions). In practice, this will consist of a time-series of aggregated physical diagnostics (heat and freshwater content, meridional heat and freshwater transport, …) computed from multidecadal simulations based on the OMIP forcing and initialisation protocol.

All the metrics will be computed from predictions provided on a fixed grid (referred to as the *emulator state space*), with the two possible target resolutions (1/12° and 1/2°). They will consist of time-series as a function of prediction lead time.

4. Baselines:

Several global emulators of 3D ocean models or ocean reanalyses have been developed over the past two years : as for instance Xihe (Wang et al. 2024), AI-GOMS (Xiong et al. 2023) and ORCA (Guo et al. 2024). Here we propose to use as the main baseline solution the emulator of GLORYS12 reanalysis developed by A. El Aouni at Mercator Ocean International[2]. We also propose to include predictions of Xihe in the data-challenge leaderboard (provided an agreement is found with the developers of Xihe). As described in section 7, we also think that the iGOR data challenge will be used for designing ML-based bias corrections for hybrid ocean models (Bora et al. 2023; Storto et al. 2024). We therefore propose to develop a first baseline of ML-based bias correction to be used in hybrid ocean models.

5. Related scientific references:

We here gather the list of references to the scientific publications mentioned in the proposal :

- Bora et al. (2023) : https://arxiv.org/abs/2302.03173

[2] Referred to as GloNet, see : https://agu.confex.com/agu/agu24/meetingapp.cgi/Paper/1524960

- Chattopadhyay et al. (2024) : https://doi.org/10.1038/s41598-024-72145-0
- de Burgh-Day and Leeuwenburg (2023): https://doi.org/10.5194/gmd-16-6433-2023
- Farchi et al. (2023) : https://doi.org/10.1029/2022MS003474
- Griffies et al. (2016) : https://doi.org/10.5194/gmd-9-3231-2016
- Guo et al. (2024) : https://doi.org/10.48550/arXiv.2405.15412
- Heimbach et al. (2024) : https://doi.org/10.5194/sp-2024-18
- Lellouche et al. (2021) : https://doi.org/10.3389/feart.2021.698876
- Rasp et al. (2020) : https://arxiv.org/abs/2002.00469
- Ryan et al. (2015) : https://doi.org/10.1080/1755876X.2015.1022330
- Subel and Zanna (2024) : https://arxiv.org/abs/2402.04342
- Storto et al. (2024) : https://doi.org/10.5194/gmd-2024-185
- Wang et ail. (2024) : https://arxiv.org/abs/2402.02995
- Xiong et al (2023) : https://arxiv.org/abs/2308.03152

6. Relevance of the proposal in terms of AI-native solutions :

Machine Learning (ML) based emulators trained from reanalysis data are currently challenging atmospheric models for both short term prediction and climate applications (de Burgh-Day and Leeuwenburg, 2023). Whether ocean models will be challenged in the same way is still to be demonstrated and the purpose of the proposed data challenge is to establish a reference benchmark in this field as WeatherBench for atmospheric models (Rasp et al. 2020). The main ambition of iGOR is to foster the adoption of the data challenge by ML practitioners, this is why we have chosen to formulate the prediction problem on fixed grids for both input and output. Our team is fully aware that end-to-end pipelines where input fields and evaluation metrics are formulated in observations space may eventually lead to better prediction. Still we think that problem formulated in iGOR data challenge is a necessary first step for our community.

The overall ambition of the iGOR initiative is to provide a framework for assessing whether emulators trained from ocean reanalyses have the potential to replace current generation physics-based models for short term forecast and seasonal prediction. The AI-native models that will be developed in response to this data challenge will be readily available for producing ocean forecasts at reduced computational cost, with similar skill as current generation physics-based models. They will also be available for producing ensemble simulations with perturbed initial conditions, therefore allowing to better sample the uncertainty related initial conditions in forecasting systems.

The proposed data challenge is formulated at global scale but we stress that the proposed benchmark and metrics will also be useful for groups involved in designing high resolution emulators at regional scale, which is also a very active field for research (see for instance Chattopadhyay et al. 2024; Subel and Zanna 2024).

7. Relevance of the proposal in terms of topical demonstrations:

As described by Heimbach et al. (2024) in their review, machine learning is currently accelerating research in ocean science. Besides the direct expected benefit for short term forecasts described above, the iGOR initiative will also provide a framework for developing hybrid ocean models based on a combination of physics-based components and ML-components. Our data challenge can indeed also be leveraged for learning corrections to existing physics-based models. This idea of leveraging reanalysis data and ML for learning state dependent bias correction to geoscientific models has been proposed by several authors over the past years (see for instance Farchi et al. 2023 and Bora et al. 2023). We therefore expect that the training data and evaluation metrics provided as part of iGOR will accelerate this line of research too. We also stress that the choice of developing physical consistency metrics should also foster research on the conditions for ML-based emulators and model corrections to be leveraged in longer simulation in the context of climate simulation (Griffies et al. 2016; Guo et al. 2024; Subel and Zanna 2024)

8. Scientific and technical staff involved:

The project team involves researchers and engineers from two research institutes (IGE and IMT Atlantic) and two operational institutions (MOI and Datlas). The members of the team, with their respective expertise and contribution to the project, are listed below in alphabetical order.

Name	Institute	Position	Expertise	Contribution
Anass El Aouni	MOI	Research scientist	Machine learning, ocean forecasting	ML baseline solution
Aurélie Albert	IGE	Research engineer	Ocean modeling, data science	Data management and OMIP metrics
Sammy Metref	Daltas	Research engineer	Data challenges, ocean forecast	Metrics, code and pipelines
Julien Le Sommer	IGE	DR CNRS	Ocean modeling, data assimilation, machine learning	Coordination, evaluation metrics
Andrea Storto	CNR	Research Scientist	Data assimilation, machine learning, ocean reanalyses	evaluation metrics, ML baseline for hybrid models
Said Ouala	IMT-Atl.	Ass. Prof	Machine learning, dynamical systems	ML baseline for hybrid ocean model

9. Response to the points raised during the initial review process:

Our response to each of the points raised during the initial review process are provided below, with reference to the changes in the previous section when relevant.

A. *Could evaluation metrics address PPR scientific priorities (eg, polar oceans, extremes,...)?*

Thank you very much for this suggestion. Indeed, developing metrics specifically adapted to PPR Ocean & Climate scientific priorities would increase the potential impact of our proposition. We have therefore modified the description of the metrics accordingly in the text above. We think that the definition of these evaluation metrics should ideally be discussed and coordinated at the level of the broader PPR data Ocean & Climate challenge team.

B. *Is the proposed evaluation for short-term accuracy relevant for global short-term ocean forecasts (DC 11) and/or ocean reconstruction (DCs 10, 12)?*

As described in the text above, we think the the evaluation metrics in observation space should be based on tools and software packages shared across several data challenges. Provided the metadata conventions and software APIs are carefully defined, such a package could be used for formulating probabilistic metrics (as for instance CRPS) as part or DC11 or reconstruction metrics as part of DC10 and DC12. Our team is ready to contribute to the definition and implementation of such a core metrics package, provided the effort is coordinated at the level of the broader PPR data Ocean & Climate challenge team.