Probabilistic short-term forecasting of 3D ocean dynamics over the global ocean

Description of the task:

The task associated with this challenge is the probabilistic short-term (up to a a few tens of days) forecasting of upper ocean dynamics (SSH, SST, SSS, U, V, MLD) from real altimetry data (2D + T, with sparsely sampled data in space and time), as well as ARGO data, with potential extension to a full 3D prediction of ocean state (T, S, U, V).

The surface ocean is highly constrained by the atmospheric conditions and its inherent synoptic scale (predictable for a week to 10 days). However the internal ocean dynamics evolve slower than the atmosphere: a few days to a week is required for the ocean evolution to matter. Hence accurate prediction of the ocean, beyond its persistence, is a current outstanding challenge of ocean physics. This project aims at addressing it. We hypothesize that there exists an ocean forecast horizon beyond the atmospheric forecast horizon, if one acknowledges the probabilistic aspect of the synoptic atmosphere and hence aims for a probabilistic ocean forecast.

Recent advances in weather prediction (GraphCast [1], GenCast [2]…) have recently set new standards for short term global forecasts of geophysical dynamics, including for probabilistic forecasts. Similar demonstrations are in their early development stage for the ocean (eg, [3], [4]), and are deterministic. They are not able to provide an uncertainty estimate in the predictions, which severely limits practical applications. This proposal aims to set up a reference benchmark for such probabilistic forecasts of the upper level ocean dynamics.

This project will be done in two steps. The first step will focus on predicting the ocean surface (strongly constrained by atmospheric conditions). The second step will be to predict the full 3D ocean, including the ocean interior. The latter being more integral and less impacted by atmospheric conditions, it is somehow more predictable. Hence, here the challenge is to filter the relevant surface dynamics for extrapolating to the ocean interior, rather than a higher-dimension prediction problem. Because of its many operational needs, the first step is more timely, and thus will be the initial focus of the project. Generalisation to a 3D structure will be done in synergy with other Data Challenges that specifically focus on ocean interior reconstruction.

Inputs to be given to the prediction systems to be benchmarked will include past observations of key variables related to upper surface ocean dynamics (altimetry data, SST, SSS, Sea Surface Currents), but also information about the water column (Mixed Layer Depth obtained from ARGO profiles). Simulation data may also be leveraged for training. The target output is a global 3D ocean state for short lead times, i.e. in the order of a few tens of days.

Probabilistic forecasts have to be evaluated based on more than just the mean prediction error, to assess whether the spread in the estimates behaves reasonably as a function of the bias, requiring the use of specific metrics. The focus is on developing methodologies that can produce accurate, uncertainty-quantified short-term predictions of upper ocean dynamics and then 3D dynamics, leveraging sparsely sampled observational data. The challenge encourages innovative AI-based approaches and highlights the importance of uncertainty quantification in real-world forecasting.

Datasets, descriptions of training data/inputs/outputs for the solutions :

Description of the proposed training, validation and test datasets (types, resolution, volume, etc.), of the required preprocessing steps

Expected outputs should be standardized and identical for all solutions, to allow for quantitative comparison. We also propose to standardize inputs available at inference time, though this point may be open to discussion among participants, if relevant observables we have not thought about are identified. However, learning based solutions may leverage additional data modaties (observables, scales, simulated data) during training depending on the strategy, if useful to improve forecasting *in fine* from the identified input variables.

Outputs (DC V1):

- Observables: SST, SSH, SSS, U, V, MLD in 2D+T
- Scale : Global
- Spatial/Temporal resolutions: 1/4° in space, daily

Outputs (DC V2):

- Observables: T, H, S, U, V, MLD in 3D+T
- Scale : Global
- Spatial/Temporal resolutions: 1/4° in space, daily

Proposed inputs at inference (DC V1):

The test period ranges from 2023 to 2024.

- Nadir altimetry, SWOT Data, multi sensor L3 data (sparse in space and time) for sea surface variables.
- Mixed Layer Depth obtained pointwise from in-situ vertical profiles from ARGO floats.
- Gridded atmospheric forecasts up to 10 days.

Proposed inputs at inference (DC V2):

The test period ranges from 2023 to 2024.

- Nadir altimetry, SWOT Data, multi sensor L3 data (sparse in space and time) for sea surface variables.
- Mixed Layer Depth, T, S, temporally integrated U,V for different depths, all obtained pointwise from in-situ ARGO float data.
- Gridded atmospheric forecasts up to 10 days.

Tentative proposal of usable training data:

At least all the variables used during inference should be leveraged, however some additional relevant data may be considered, (e.g. Chlorophyll concentration). However, we propose to exclude from usable training data gridded products for surface oceanic variables (e.g. DUACS outputs) that could be used to supervised training, using only sparsely sampled data as inputs.

Here is a minimal list of data that can be leveraged for training:

DC V1 : Nadir altimetry + SWOT for SSH, SST, SSS Level 2 data ARGO float data (MLD) for subsurface ocean measurements.

DC V2 : Nadir altimetry + SWOT for SSH, SST, SSS Level 2 data ARGO float data (T,S,temporally integrated U,V for different depths,MLD) for subsurface ocean measurements.

Past observations of these will constitute the inputs to be fed to the competing models, but can also be leveraged as training data. However, these quantities are sparsely sampled in space and time, and the acquisition geometry is specific to each sensor. So raw datasets will require significant preprocessing steps (to be shared to all participants once set up) in order to be exploitable to train solutions.

An important point is that **for training, simulated (model) datasets can also be leveraged**, e.g. eORCA36 simulations.

Evaluation metrics:

Description of the proposed evaluation metrics, including the associated computational complexity and preprocessing steps if any.

We will mainly consider standard metrics used for probabilistic forecasting, see e.g. [5]:

- ○ Continuous Rank Probability Score (CRPS) : to assess the relevance of an ensemble of predictions wrt a single ground truth observation. It reduces to mean absolute deviation (MAE) when the forecast is deterministic. Scales in O(N log N) with the number of members N in the prediction. Does not extend immediately to multivariate predictions, but scales may be used on all marginal distributions.
- ○ Spread-Skill ratio : This metric measures how the variability of the predictions (spread) relate to the bias in the prediction (skill). A good model should have a ratio of 1, meaning that the uncertainty grows at the same rate as the bias.

Any of those metrics may be computed globally (averaging over the surface and/or the vertical axis) for different lead times (e.g. 5-10, 15, 20, 15, 30 days), or spatialized at each pixel/time frame to evidentiate spatial/temporal patterns in the errors and uncertainties. Other metrics may include probabilistic versions of Energy Spectra [5].

Those metrics will be computed and compared between solutions for every output variable.

Baselines:

description of the proposed baselines, both in terms of operational and learning-based baselines if any.

There are two scientific hurdles to pass to propose successful solutions for the DC compared to current learning based solutions: the step from deterministic to stochastic predictions, which is probably the most fundamental challenge to overcome, and also the step from 2D to 3D.

This justifies the division of the DC into two parts, a first on Sea Surface variables and MLD, that integrates some partial information about the vertical axis. This first version will allow us to focus on proposing strong stochastic baselines. The second version integrates the water column, for which new baselines will have to be proposed. As a first extension of proposed surface level methods, one can use projection methods to map 2D solutions to 3D using vertical profiles, thus creating simple baselines as a start [6] for the second step. Apart from the higher dimensional problem, 3D prediction is expected to be also less dependent on the state of the atmosphere.

Baselines for the 2D case:

- operational forecasting products
    - DUACS (2023-2024)
    - GLO Forecast from Mercator
- Deterministic ML baselines (training criterion based on MSE or similar metrics):
    - XiHe [3]
    - 4DVarNet [7]
- Stochastic learning based baselines at a global scale are not available yet, but are to be developed for the data challenge. We will notably propose stochastic versions of the 4DVarnet deterministic framework [7]. These probabilistic solutions will require moving to a probabilistic training criterion, such as e.g. the CRPS or likelihoods depending on the parametrization of the models.

Apart from projection methods, AI native baselines for the second step that directly predict the 3D ocean state are to be developed during the challenge.

Answers to the questions raised by the coordination committee (CC):

As suggested by the CC, we have merged the proposal with the team from DC3 (PROOFCAST), and also integrated the proposal of DC8 (PROCAST) with the proposing team, which had a lot of common features to our initial proposal.

From a scientific point of view, the CC suggested looking into metrics related to extremes which are indeed critical events to be captured. There are several relevant ways to define an extreme event for ocean processes. Some of them are very precisely defined, e.g. oceanic heat waves, or rogue waves; but both seem out of reach for characterization due to our choices of operational temporal/spatial prediction horizons. However, the probabilistic solutions proposed could evidentiate phenomena such as locally rare intense currents. In this case, we could introduce metrics that are able to assess how well high quantiles of the data are reconstructed, but this would require comparing two distributions (reference distribution over time or space of an observable and forecast distribution) rather than a distribution (forecast) and a punctual realization (test observation). Still, efficient ML inspired-metrics related to higher order moments [8] or MSEs on quantiles could be leveraged. Bifurcations in some well identified cases may also be captured using a probabilistic framework. In these situations, case-by-case metrics based on regional variations of transition probabilities could be used. These metrics could be integrated to the DC after the classical metrics have been implemented due to their more exploratory nature.

The CC also asked about relationships between the proposed DC and other DCs related to reconstruction tasks. We indeed acknowledge the proximity between our task and metrics and the case of reconstruction: we essentially propose to generalize deterministic metrics (e.g. CRPS generalizes MAE, likelihoods generalize MSEs), and also switch 2D observations to 3D observations of similar observables, and at the same time consider a particular case of reconstruction (forecasting from past observations, with several lead times). Indeed, if selected, the proposed DC will call for coordination with selected reconstruction DCs, both from the point of view of data pipelines and software building blocks (shared codebase and structure to compute metrics or preprocess similar data, for instance).

Related scientific references:

[1] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416-1421.

[2] Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., ... & Willson, M. (2023). GenCast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*.

[3] Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., ... & Song, J. (2024). Xihe: A data-driven model for global ocean eddy-resolving forecasting. *arXiv preprint arXiv:2402.02995*.

[4] Chattopadhyay, A., Gray, M., Wu, T., Lowe, A.B. and He, R., 2024. OceanNet: A principled neural operator-based digital twin for regional oceans. *Scientific Reports*, *14*(1), p.21181.

[5] Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., ... & Sha, F. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS004019.

[6] Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A. M., Roquet, F., ... & Tandeo, P. (2022). Four-dimensional temperature, salinity and mixed-layer depth in the Gulf Stream, reconstructed from remote-sensing and in situ observations with neural networks. *Ocean Science*, *18*(4), 1221-1244.

[7] Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., & Rousseau, F. (2021). Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002572.

[8] Arbel, M., Korba, A., Salim, A., & Gretton, A. (2019). Maximum mean discrepancy gradient flow. Advances in Neural Information Processing Systems, 32.

Relevance of the proposal in terms of AI-native solutions for the DTO: (5-10 lines)

The proposed data challenge will show limitations of existing AI-native baselines, which are mostly deterministic, as opposed to traditional sequential assimilation-based solutions that are able to produce uncertainty estimates. The challenge aims at fostering research around uncertainty quantification in AI-based forecasting systems, which is still largely an open problem, especially for high-dimensional states and partial observations in 3D + T. Models that are both performant and provide reliable uncertainties could evidentiate multimodal distributions, evaluate the relevance of short term evolution scenarii, or potentially better capture extreme events which by definition are not supposed to lie close to the mean prediction of the state.

Relevance of the proposal in terms of topical demonstrations: (5-10 lines)

The challenge and the solutions that will be proposed will naturally contribute to the challenges of the PPR such as « Défi 1 » revolving around the better prediction of extreme events, in particular tropical cyclones, but also and mainly to « Défi 6 » which involves breaking jamlocks in the way of the design and development of DTOs, among which uncertainty quantification constitutes a crucial feature. More specifically, the challenge will drive operational forecasting improvements, with a focus on short-term ocean dynamics, critical for maritime navigation, fisheries, climate research, and disaster management. The challenge aims to deliver new methods for high-resolution, probabilistic predictions of ocean state variables, which could eventually transform decision-making in global ocean monitoring and forecasting.

Scientific and technical staff involved:

Description of the key personnel for the proposal.

We anticipate a need for 6 p.m for the research engineer time to gather all sources of data, preprocess them, and build the codebase for metrics and their interface with solutions to be proposed (in coordination with other DCs).

Key personnel, will be involved in all regular project meetings (2p.m each)

Lucas Drumetz	Associate Prof. IMT Atlantique, Odyssey	ML/IA signal/image processing for remote sensing applications	DC coordination, definition of evaluation metrics, solution design
Florian Sévellec	DR CNRS, LOPS, Odyssey	ocean physics, climate dynamics, ocean and climate predictability	definition of key variables and data, evaluation metrics
Julien Le Sommer	DR CNRS, IGE	Ocean modeling, data assimilation, machine learning	coordination of the DC, definition of key variables and data, evaluation metrics
Sammy Metref	Datlas Research	Data challenges, ocean engineer forecast	Metrics, code and pipelines

Additional personnel that will take part in the design/implementation of the DC, and/or the design/development of solutions (1p.m for each permanent researcher, 2p.m for PhD students)

Hugo Georgenthum	PhD Student, IMT Atlantique, Odyssey	AI and data assimilation	development of the forecasting solutions, 4DVarNet development
Pierre Haslée	PhD Student, IMT Atlantique, Odyssey	AI and data assimilation	development of the forecasting solutions, 4DVarNet development
Bertrand Chapron	Researcher, IFREMER, Odyssey	physical oceanography and remote sensing	definition of key variables, evaluation metrics
Clément Boyer de Montégut	Researcher, IFREMER, Odyssey	physical oceanography and in situ observations	definition of key variables, evaluation metrics
Pierre Tandeo	Associate Prof, IMT Atlantique, Odyssey	stats/ML for ocean remote sensing data	definition of key variables, evaluation metrics, solution design
Said Ouala	Associate Prof, IMT Atlantique, Odyssey	AI and data assimilation	definition of key variables, evaluation metrics, solution design
Carlos Granero	Associate Prof, IMT Atlantique, Odyssey	AI, signal/image processing, ocean dynamics	definition of key variables, evaluation metrics, solution design