Scientists review advances in ocean data assimilation

More than 170 scientists from around the world presented recent progress and challenges ahead in ocean data assimilation (DA) in a virtual event from 17 to 20 May 2021. The event, organized jointly by ECMWF and OceanPredict, was designed to meet the ever-increasing requirements of marine, weather, environmental and climate services. It was very timely since 2021 marks the beginning of the U.N. Decade of Ocean Science for Sustainable Development.

During this workshop experts from different domains addressed the multidisciplinary science underpinning climate and environmental monitoring and predictions, the exploitation of novel observations, and interactions in the ocean–atmosphere–sea-ice–biogeochemistry system at global and regional scales.

A combination of plenary talks, poster sessions, working group discussions and informal virtual breaks facilitated the exchange of information and seeding of new ideas. The working group discussions addressed questions common to all applications, including treatment of model error, the specification of short-range forecast and observation errors, the balancing of resolution and ensemble configurations, and the exploitation of machine learning. The discussions also covered infrastructure needs to share developments among different domains, and between operations and research.

Progress reported

A combination of 36 oral presentations and 29 posters reported progress on the different workshop themes, as summarized below.

- **Data Assimilation Methods**: Methods used for operational ocean data assimilation have traditionally lagged behind those used for numerical weather prediction. Work presented at the workshop indicates that the gap has narrowed significantly, notably with the development of ensemble-variational methods, and multi-scale ensemble-based methods. There has also been substantial progress on algorithms for improved ensemble generation and computational efficiency, such as developments with multiple resolution algorithms. The use of the Gaussian approximations, generally used in operational DA, appears justifiable by the fact that ocean models at present resolutions and observation operators are only weakly non-linear. The computational demands of non-Gaussian DA methodologies are unaffordable on current and planned HPC infrastructures. Fixed-lag smoothers like 4D-Var also show better Gaussian properties than Filters (Ensemble or otherwise).
- **Coupled Data Assimilation**: Progress is being made in the new field of coupled data assimilation, where both technological and scientific developments were presented at the workshop. Coupled ocean-atmosphere assimilation showed improved representation of multi-variate relations in air-sea interaction variables in reanalyses, and improved forecasts of tropical cyclones with adequate treatment of flow-dependent background errors. Efforts in coupling the physical ocean and biogeochemistry assimilation problems are also ongoing, with both variational and ensemble approaches.
- **Reanalyses**: Ongoing efforts in applying variational methods in coupled ocean-atmosphere reanalyses were presented. The latest updates on the reanalyses produced by adjoint methods with multi-decadal assimilation windows and extended control variables used for parameter estimation were reported, with ongoing efforts to include sea-ice and

biogeochemical components, and to increase the resolution via multi-resolution incremental approaches. A method to identify the main sources of uncertainty in relevant climate indices from ocean reanalyses was presented. The approach is based on an ensemble of perturbed reanalyses experiments, which allows the sensitivity of a given climate index, such as ocean heat content and meridional transports, to specific aspects of the reanalysis configuration to be quantified.

- **Applications of machine learning in data assimilation**: Data assimilation and machine learning (ML) are conceptually very close to one another as they can be viewed from the same Bayesian framework. ML techniques allow for model discovery from available observations and are particularly relevant for highly irregular or poorly understood problems where model derivation from physical principles is challenging. They can be used to characterize hard to parametrize physical processes or to replace highly non-linear observation operators that are sensitive to many variables. Parameter estimation in variational methods is essentially an ML process, and parameter tuning can be aided by ML techniques. Computationally efficient neural-network based emulators can be used to accelerate the data assimilation process. We cannot abandon developing physically based operators and completely replace them with ML; this would risk losing the existing knowledge and expertise. ML techniques have a potential to help us understand the model and observation biases and better account for them in the data assimilation process.

- **Treatment of model error**: Early results from using ML to account for model errors already show encouraging improvement in atmospheric analysis. Training data could be an issue for the data sparse ocean system though. There have been advances on the representation of random model error via stochastic physics, applied to the NEMO ocean model: a new stochastic perturbation package is now available for generating perturbations to the model tendencies, physical parameters and resolution related energy dissipation. A new approach to sample structural uncertainty, the so-called super-model ensemble, was also presented. In this approach, different models interact with each other through an EnOI scheme that assimilates pseudo-observations created by weighted ensemble mean. Errors in the representation of the diurnal cycle in the ocean have been diagnosed using data from observational campaigns, where the strong air-sea interaction is visible (SST, precipitation, stratification). In another presentation, a parameter estimation scheme applied to a high-resolution global tide model was used to estimate corrections to the bathymetry using the 1973 time-series from the FES2014 tidal model. Coastal tide gauge comparisons show an improvement in RMSE. Parameter estimation benefits from long observation time series

- **Assimilation of novel observations**: Observing system experiments to design future altimeter configurations were reported. Results show high sensitivity to the choice of Mean Dynamic Topography. There are also efforts towards observation operators for altimeter observations that include the tidal signal. Moving towards higher spatial and temporal resolutions is challenging, and requires controlling the disparity of scales among different observing platforms and the model. When observations have higher resolution than the model can resolve, filtering high frequency signals in the observations is required if an adequate representativeness error model is unavailable. Research is ongoing to combine the disparity in the sampling properties among observing platforms, such as the high resolution/low frequency of SWOT with the low resolution/high frequency of moorings. Major issues with correlated observation errors when assimilating wide-swath altimetry data such as future SWOT observations were found. The problem is somewhat alleviated by half-swath superobbing (5km). In this configuration SWOT OSSEs show improvements in eddy positions, sub-surface T/S and surface currents, but better use of the data could be made with proper representation of the observation error correlations in the data assimilation.

OSSEs to assess the impact of total surface current velocities (TSCV) data from future missions (would-be SKIM mission) were presented. The vertical background error correlation length scales for unbalanced U and V are shorter than those for T and S, and developments are required to specify background error covariances for the unbalanced (ageostrophic) velocities in NEMOVAR. Beyond OSSEs/OSEs, efforts using the variational methodology in the ROMS system reported strong and far-reaching impact of deep gliders in the Australian Western Boundary Current System when using a very high resolution regional ROMS configuration.

- **Applications**; The impact of ocean observations on 4-day forecasts on the regional ROMS configuration over the Californian Coast using the Forecasts Sensitivity to Observation (FSO)s was presented. FSO relies on the 4D-Var infrastructure to project the errors in the analyses into the forecast errors. FSO can be applied to highly targeted metrics (e.g. drifter trajectories in support of help-and-rescue). Results show that about 50% of the observations are responsible for the forecast improvements, which raises the question of whether it is possible to extract more information from existing observations. Progress has been made with the ROMS-based LETKF analysis system over the Eastern Indian Ocean. With respect to the free running model, results show improved performance against a variety of variables. Most noticeable were the improvements on the representation of surface and subsurface currents. The relative benefit of ensembles versus horizontal ocean resolution for medium-range and extended-range marine forecasts over western boundaries has been assessed. A variety of metrics, including scale-dependent verification, indicates that, beyond 1/12 degree horizontal resolution, ensemble forecasting clearly brings more benefits than increasing horizontal resolution. A novel verification technique for forecast of eddies was presented, based on tracking of individual systems, and following a similar approach to that used in weather applications for cyclone tracking. Several systems were assessed on the verification of eddies as a function of eddy-radius and eddy-amplitude. Eddies larger than 150km radius were well forecasted by the different systems, while all systems showed clear deficiencies in forecasting smaller eddies. Another presentation stated that the need for ocean reanalyses is now larger than ever, with increasing requirement of long consistent climate records for monitoring and predicting climate variability and change. The need to advance the treatment of systematic model error and balance relationships at the Equator and ocean-atmosphere boundary layer were highlighted. The potential benefits of coupled data assimilation for better estimation of the Earth energy budget were discussed, as well as the need of sufficient high resolution to resolve the air-sea interaction over frontal areas.

- **Recent assimilation infrastructure developments**: Developments on JEDI (Joint Effort on Data Assimilation Integration) and PDAF (Parallel Data Assimilation Framework) were presented. JEDI is an extension of OOPS (Object Oriented Prediction System) to include other generic components, such as interfaces across Earth-System components, workflows (EWOK) and data archive diagnostics (R2D2). As in OOPS, JEDI is a model-agnostic abstract organizational layer that follows the principle of separation of concerns. JEDI now includes SABER (System-Agnostic Background Error Representation), UFO (Unified Forward Operator), and SOCA (Sea-Ice Ocean and Coupled Assimilation), the latter being a JEDI encapsulation of MOM6 and CICE. A specific development within SABER is BUMP (Background error on an Unstructured Mesh Package) which provides tools based on ensembles of forecasts to estimate the parameters needed for hybrid modelling of the background error covariances as well as tools to model the covariances in the data assimilation system. JEDI is currently being applied to enable coupled ocean-atmosphere data assimilation to the UFS (Unified Forecast System), in preparation for future operational configurations. PDAF is a program library for ensemble modelling and data assimilation,

providing support for ensemble forecasts, DA diagnostics, and fully-implemented filter and smoother algorithms. It makes good use of supercomputers and follows the principle of separation of concerns: model, DA methods, observations. It is easy to couple to models and to code case-specific routines and it is efficient for research and operational use. It is mainly designed to support ensemble data assimilation, but there are ongoing efforts to include variational methods.

Working Group Discussions and Recommendations

Working Groups (WGs) were charged to consider the priority areas for advancing ocean data assimilation during the next decade and the primary challenges to progress. The *U.N. Decade of Ocean Science for Sustainable Development,* for which ocean analysis, reanalysis and forecasting activities are an integral part, acted as backdrop of the discussions, which were organized under the workshop four *primary* themes: *Balancing Model Resolution versus Ensembles*, *Infrastructure*, *Best Practices*, and *Data Assimilation Methods*. Within each *primary* theme, the WGs were encouraged to discuss cross-cutting aspects, namely: *opportunities for machine learning*; *education, training and outreach*; *forecasting and analysis applications*; *the use of novel observations*; *deficiencies and gaps in knowledge*; and *emerging service needs.*

Below is a summary of overarching recommendations from the WGs. The individual WG discussions and recommendations can be found on the workshop webpage.

Overarching recommendations

Coupled data assimilation: The WG discussions acknowledged the complex science of coupled data assimilation. They recommended the entrainment of cross-disciplinary expertise including those with knowledge of ocean-atmosphere boundary layer physics, and corresponding expertise for physical and biogeochemistry coupling. They also encouraged exploring machine learning solutions, and the use of targeted observations of the interface for process understanding and advanced modelling. They recommend more research to consolidate the motivation for strongly coupled ocean-atmosphere data assimilation in which changes are made to the ocean based on information from atmospheric observations and vice versa.

Resolution: Models should have sufficient resolution to resolve the relevant physical processes, such as the variability and position of western boundary currents, even if the current observing network is not sufficient to constrain the small scales. Model configurations that include targeted resolution increase in dynamically active regions or the use of two-way nested model systems could be beneficial. Novel/emerging observation platforms that sample finer scales may become more useful to the DA system. Observations that sample coarser scales (e.g. altimetry, passive microwave data) might need different treatment in a high-resolution DA system than before.

Methodology: Further development of methods for representing balanced multi-scale flow-dependent background error covariances is recommended. The time dimension within the assimilation window should be considered, both in variational and ensemble methods. When using tangent linear and adjoint models, the validity of the linear approximation should be quantified and documented. Methods to balance the need of ensembles and resolution are needed. The exploration of machine learning solutions for different aspects of data assimilation is encouraged. Data assimilation methods targeting coupled data assimilation, treatment of model error and parameter estimation are becoming increasingly relevant.

Ensembles: In principle, the quality of background error statistics diagnosed from ensembles will improve with an increasing number of ensemble members. In practice, computational resources constrain the ensemble size, and it is important to invest in enhanced ensemble generation strategies to improve the reliability of the ensemble and reduce sampling issues.

Observing System (Simulation) Evaluation Experiments (OSEs/OSSEs): These are important for understanding how well the existing observing networks constrain the models and what additional observations are needed. Comparison studies of coordinated OSSEs and OSEs will be helpful to learn from different systems. The involvement of the observational community in the design and analysis of observation impact experiments is encouraged, as well as sharing of the OSSEs/OSES outputs with the wider community for the analysis of results. The latter will provide a way to familiarize the ocean and climate experts with the data assimilation problem. For the coordinated experiments, it will be practical to have focused efforts to galvanize participation. The SynObs activity proposed by OceanPredict as a contribution to the UN Decade of Ocean Science is a good opportunity.

Evaluation: As well as standardized metrics of fit to the assimilated observations, comparing with independent data is strongly encouraged. Assessing the balance of the system (impact from one observed variable to other state variables or budget analysis) is recommended. Error growth rates can be used as indication of the goodness of the solution. Analysis of the temporal statistics of the departures and assimilation increments is encouraged, so as to gain insight into the properties of model error. Development of reliability diagnostics for ensembles should be pursued. For reanalyses, temporal consistency of the estimation should be considered. To this end, sufficiently long records of good quality observations are required.

Treatment of Observations: With the data from upcoming satellite missions such as SWOT expected to have significant spatially correlated errors, progress should be made on efficient implementation of methods to model observation error correlations in DA systems. There is a need for developing and sharing methods for automatic quality control and observation bias correction. Special attention should be paid to the quality control flags and grey lists of the Argo data, which should be regularly updated. Historical observation repositories should be updated regularly and as promptly as possible.

Infrastructure developments: Given the complexity of data assimilation algorithms, developments on modular software infrastructure that facilitates the exchange of developments and their application to different models are welcome and encouraged. These software infrastructure developments should be open source. Management of the complexity of the data assimilation infrastructure is crucial to facilitate its uptake downstream from operational centres to academia. The need for data-sharing infrastructure to facilitate collaboration was also identified. This is especially relevant for sharing output from expensive experiments that only a few operational agencies are able to conduct. Moving towards cloud computing infrastructure would facilitate sharing between operational centres and academia via cloud based containers.

Links with the modelling community: Science and infrastructure developments will benefit from stronger links between the ocean modelling and assimilation communities. The needs of data assimilation should be considered when developing ocean models (e.g. tangent linear and adjoint models, refactoring for adaptation to the emerging infrastructure, observation operators, stochastic parametrizations for ensemble generation). In turn, the capability for comparing models with

observations, the information from analysis increments, and the possibility of using data assimilation for parameter tuning are important assets for model development.

Training and recruitment: Investment in training of the next generation of scientists in data assimilation is identified as critical. This should include specific training on the use of modern software development/collaboration techniques. Beyond training, there is a need to strengthen the data assimilation activities in research institutes and universities. To this end, sustained research funding for data assimilation in non-operational centres is required to create a sufficient pool of expertise to exploit new computer architectures and observing systems.