

## Query Details

1. Author: Please check and confirm that the authors and their respective affiliations have been correctly identified and amend if necessary.

# The added value of the multi-system spread information for ocean heat content and steric sea level investigations in the CMEMS GREP ensemble reanalysis product

Andrea Storto, <sup>1,7</sup>✉

Email [andrea.storto@cmre.nato.int](mailto:andrea.storto@cmre.nato.int)

Simona Masina, <sup>1,2</sup>

Simona Simoncelli, <sup>2</sup>

Doroteaciro Iovino, <sup>1</sup>

Andrea Cipollone, <sup>1</sup>

Marie Drevillon, <sup>3</sup>

Yann Drillet, <sup>3</sup>

Karina von Schuckman, <sup>3</sup>

Laurent Parent, <sup>3</sup>

Gilles Garric, <sup>3</sup>

Eric Greiner, <sup>4</sup>

Charles Desportes, <sup>3</sup>

Hao Zuo, <sup>5</sup>

Magdalena A. Balmaseda,<sup>5</sup>

K. Andrew Peterson,<sup>6</sup>

<sup>1</sup> Euro-Mediterranean Center on Climate Change (CMCC)  
Foundation, Bologna, Italy

<sup>2</sup> National Institute for Geophysics and Volcanology (INGV), Bologna, Italy

<sup>3</sup> Mercator Ocean, Ramonville Saint-Agne, France

<sup>4</sup> Collecte Localisation Satellites (CLS), Ramonville Saint-Agne, France

<sup>5</sup> European Center for Medium Range Weather Forecasts  
(ECMWF), Reading, UK

<sup>6</sup> Met Office, Exeter, UK

<sup>7</sup> Andrea Storto, Centre for Maritime Research and Experimentation  
(CMRE), viale San Bartolomeo 400, 19126 La Spezia, Italy

Received: 30 April 2018 / Accepted: 10 December 2018

---

## Abstract

Since 2016, the Copernicus Marine Environment Monitoring Service (CMEMS) has produced and disseminated an ensemble of four global ocean reanalyses produced at eddy-permitting resolution for the period from 1993 to present, called GREP (Global ocean Reanalysis Ensemble Product). This dataset offers the possibility to investigate the potential benefits of a multi-system approach for ocean reanalyses, since the four reanalyses span by construction the same spatial and temporal scales. In particular, our investigations focus on the added value of the information on the ensemble spread, implicitly contained in the GREP ensemble, for temperature, salinity, and steric sea level studies. It is shown that in spite of the small ensemble size, the spread is capable of estimating the flow-dependent uncertainty in the ensemble mean, although proper re-scaling is needed to achieve reliability. The GREP members also exhibit larger consistency (smaller spread) than their predecessors, suggesting advancement with time of the reanalysis vintage. The uncertainty information is crucial for monitoring the climate of the ocean, even at regional level, as GREP shows consistency with CMEMS high-

resolution regional products and complement the regional estimates with uncertainty estimates. Further applications of the spread include the monitoring of the impact of changes in ocean observing networks; the use of multi-model ensemble anomalies in hybrid ensemble-variational retrospective analysis systems, which outperform static covariances and represent a promising application of GREP. Overall, the spread information of the GREP product is found to significantly contribute to the crucial requirement of uncertainty estimates for climatic datasets.

---

## Keywords

Ocean synthesis, reanalysis accuracy  
Uncertainty  
Hybrid data assimilation  
Observation impact

## Electronic supplementary material

The online version of this article (<https://doi.org/10.1007/s00382-018-4585-5>) contains supplementary material, which is available to authorized users.

---

## 1. Introduction

During the last decade, ocean reanalyses became a fundamental tool for climate investigation (e.g. Masina and Storto 2017), testified by the growing number of studies that make use of them. Ocean reanalyses have been largely adopted for evaluating key climate oceanic diagnostics that are not directly observed, such as the assessment of the deep ocean warming (Balmaseda et al. 2013), the reconstruction of the overturning circulation (Jackson et al. 2016) and the study of the Arctic energy variability (Mayer et al. 2016). This picture is likely justified by the increasing maturity that reanalyses acquired with time, as the ocean models, data assimilation systems, and atmospheric forcing and observational datasets (Stammer et al. 2016) improve. One of the most significant advances is the improved spatial sampling of the ocean observational network, mostly through Argo float data (Riser et al. 2016). Additionally, more sophisticated strategies of accounting for model biases and drifts (e.g. Lea et al. 2008), advanced quality control of observations (e.g. Storto 2016), and improved initialization methods (e.g. Brix et al. 2015) have been developed during recent years. Reanalyses thus became a tool for Earth's System energy investigations (e.g. Trenberth et al. 2016) and also a reference dataset for the verification of climate model simulations such as CMIP (Coupled Model Intercomparison

project, e.g. Taylor et al. 2012), see for instance the CREATE initiative (Potter et al. 2018; <https://esgf.nccs.nasa.gov/projects/create-ip/>). This increased confidence has recently allowed reanalyses to span centennial time periods, demonstrated by the emergence of datasets covering the entire twentieth century in several institutions (e.g. Yang et al. 2017; de Boissésou et al. 2017).

#### AQ1

The ocean reanalysis community has recently devoted significant efforts to quantify strengths and weaknesses of the reanalyses, through performing systematic inter-comparison of state-of-the-art products, and using ensemble-based diagnostics. Indeed, from a multi-system ensemble perspective, the consistency among reanalyses approximates the fidelity of the reanalyses themselves, their diversity suggesting instead structural deficiencies. The ORA-IP project (Balmaseda et al. 2015) has followed this approach, suggesting that the inter-annual variability and trends of many well-constrained ocean variables (upper ocean temperature, Palmer et al. 2017; steric sea level; Storto et al. 2017; sea-ice concentration; Chevallier et al. 2017) are well captured by reanalyses, other parameters (e.g. salinity in the 1990s, Shi et al. 2017; deep ocean variability; Storto et al. 2017; air-sea fluxes; Valdivieso et al. 2017; and sea-ice thickness; Chevallier et al. 2017) that are not sufficiently constrained show large spread, while some others not directly constrained (e.g. mixed layer depth, Toyoda et al. 2017; sea-ice velocities; Chevallier et al. 2017) exhibit realistic inter-annual variability. Seasonal variability is in general well captured by reanalyses for all inter-comparisons. For certain parameters, the ensemble mean emerges as the most robust estimate and may outperform the best individual member (for instance for steric sea level, see Storto et al. 2017). Indeed, the mean of the ensemble can be expected to outperform individual members when the systematic errors of the members are fairly independent and the ensemble mean operator thus cancels them out. Thus, the multi-model ensemble approach proves a promising method because of both the robustness of its mean and the implicit quantification of uncertainty by means of the spread. However, when there exists mutual dependence between model errors, as in most real-world multi-model ensemble, it is non-obvious whether the ensemble mean shall outperform the best individual member. Further to inter-comparison activities, complementary observation impact studies and ocean model simulations are necessary tools for disentangling the relative importance of error sources (e.g. Karspeck et al. 2017), i.e. whether reanalyses deficiencies stem from lack of observations, approximations in data assimilation schemes, model parameterizations and inaccuracies in the atmospheric forcing.

The Copernicus Marine Environment Monitoring Service (CMEMS) of the European Union provides real-time and retrospective analysis of the variability of the ocean and marine ecosystems for the global ocean and the European regional seas (see Le Traon et al., 2017, for an overview of CMEMS products and research activities). In particular, global ocean reanalyses are identified as a fundamental product to monitor the state of the ocean climate and its temporal evolution, along with the possibility of using ocean reanalyses for a large variety of downstream and downscaling applications, ranging from initialization of seasonal and decadal prediction systems, provision of boundary forcing for regional applications, to forcing fields for biogeochemical models. To this end, CMEMS supported since 2016 the production of a four-member ensemble reanalysis called GREP (Global ocean Reanalysis Ensemble Product). The GREP product consists of four global state-of-the-art reanalyses at eddy-permitting resolution (approximately  $1/4$  degree of horizontal resolution and 75 depth levels), which are all based on the NEMO ocean model (Madec et al., 2012), forced by the European Centre for Medium Range Weather Forecasts (ECMWF) ERA-Interim (Dee et al. 2011) atmospheric forcing albeit with different bulk formulas, and employ different observational datasets and data assimilation schemes. Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), UK Met Office, Mercator Ocean and ECMWF produced the four reanalyses, and Mercator Ocean has compiled the GREP ensemble product. GREP represents a prototypical multi-system reanalysis at eddy-permitting resolution, which will be updated in near real-time to primarily serve climate monitoring, initialization of long-range prediction system, and downstream applications.

The ensemble members of GREP share the ocean modeling core and the atmospheric forcing dataset. The spread comes obviously from observational and assimilation differences. However, there are also many important differences in the reanalysis initial states, air-sea flux formulations, sea-ice models, parameterizations, which all concur to the dispersion of the reanalysis realizations. Classifying the most important sources of uncertainty is not easily achievable, as it would require additional set of ensemble experiments. As extensively discussed by Masina et al. (2017) for the previous MyOcean ensemble of reanalyses, the use of the same ocean model might in principle lead to under-estimated ensemble dispersion, in spite of the diversity of the individual configurations, providing a further motivation to assess the reliability of the spread, which is one of the main goals of this article. Indeed, Masson and Knutti (2011) discuss how the use of the same model induces strong similarities among the members, thus questioning the ensemble reliability itself. Nevertheless, sea-ice models, NEMO versions, physical and numerical parameterizations, spin-up

strategies, and revisions of bulk formula differ among the four reanalyses, providing an important diversity of implementations. Moreover, GREP weighs each reanalysis equally, which may further hide possible model hierarchies within the ensemble (Chandler 2013). More sophisticated “super-ensemble” techniques, where members are weighed unevenly depending on their prior accuracy such as the ones proposed by Krishnamurti et al. (2000), are not considered here.

Here, further to assess the GREP mean and spread reliability, we provide several examples where information about spread is beneficial to ocean and climate investigations. The structure of the paper is as follows: Sect. 2 presents the GREP product; Sect. 3 presents the results concerning the assessment of the ensemble spread reliability, and other applications of the spread for climate monitoring, observing network assessment, hybrid data assimilation and assessment of reanalysis advancements. Section 4 discusses the main results and concludes. Finally, temperature and salinity fields are assessed in Appendix 1 and 2, in terms of statistical scores against in-situ profiles and spread comparison with the previous vintage of reanalyses, respectively.

## 2. Data and methods

### 2.1. The GREP ocean reanalysis systems

The four global ocean reanalyses included in GREP run at eddy-permitting resolution (approximately 1/4 degree of horizontal resolution, all using the same tri-polar ORCA grid, Madec and Imbard 1996), covering the altimetry period 1993–2016. It is worth noting that some diagnostics presented in this study were conducted before the completion of the year 2016, and thus cover the period 1993–2015. GREP monthly means are currently released on a coarse resolution  $1 \times 1$  regular grid, along with the ensemble mean and spread through the CMEMS catalogue (product reference GLOBAL\_REANALYSIS\_PHY\_001\_026). Only three-dimensional temperature and salinity were released for the time being. Recently, also ocean currents are available from the CMEMS catalogue. The release of the data at higher spatial and temporal resolution and including additional parameters such as sea-ice and two-dimensional variables is foreseen for 2019. Throughout the article, we will make use of only monthly mean data.

Table 1 summarizes the characteristics of the GREP members. The four reanalyses were run using the NEMO ocean model, albeit in different versions. They all use the tripolar ORCA025 grid with Arakawa C-grid staggering and 75 depth levels with partial steps (Barnier et al., 2006). The resolution of the model

allows eddies to be represented approximately between 50°S and 50°N (Penduff et al. 2010). The horizontal resolution increases with latitude, with about 12 km in the Arctic region. Three reanalyses are coupled with the LIM2 thermodynamic-dynamic sea-ice model. Among these, one reanalysis implements visco-plastic rheology while two implement elasto-visco-plastic rheology (Bouillon et al. 2009). Foam is coupled to the CICE (v4.1) sea-ice model (Hunke et al. 2013). All models are forced by the ECMWF ERA-Interim reanalysis (Dee et al. 2011). The air-sea exchange processes are based on the formulation of the CORE bulk formulas (Large and Yeager 2004), although oras5 has revised the bulk formulas, for inclusion of the wave effects (Zuo et al. 2017b). However, slight changes in the use of forcing among the members occur (different temporal frequency, ad-hoc corrections, interpolation procedures, different bulk formulation implementation). Many physical and numerical schemes and parameterizations are similar in the four reanalyses, nevertheless a number of significant changes, including the ocean model version and a few parameterizations, occurs, thus introducing differences in the four ocean model configurations.

**Table 1**

Main characteristics of the four ocean reanalysis systems included in GREPv1

<b>GREP member</b>	<b>Institute and version</b>	<b>Ocean and sea-ice model version</b>	<b>Surface constraint</b>	<b>Assimilation Scheme</b>	<b>Assimilated observations</b>
Cglo	CMCC C-GLORS v7	NEMO3.6 + LIM2	SST, SSS, SIC, and Arctic SIT Nudging	OceanVar (3DVAR) + Large-scale bias correction [7 days]	EN4 In-situ profiles AVISO SLAs

<b>GREP member</b>	<b>Institute and version</b>	<b>Ocean and sea-ice model version</b>	<b>Surface constraint</b>	<b>Assimilation Scheme</b>	<b>Assimilated observations</b>
Foam	Met Office GLOSEA5v13	NEMO3.4 + CICE4.1	SSS Nudging	NEMOVAR (3DVAR) [1 day]	EN4 In-situ profiles, ICOADS In-situ SST, NOAA Pathfinder L2P AVHRR (PFV52) and ESA L2P/L3U ATSR, AVISO SLA, OSISAFv1.2 sea ice concentration
Glor	Mercator Ocean GLORYS2V4	NEMO3.1 + LIM2	Large scale correction for precip. and radiative flux Ice shelf and iceberg melting	SEEK + 3DVAR Bias correction [7 days]	CMEMS/CORA In-situ profiles NOAA SST, AVISO SLA (CMEMS), Mean Dynamical Topography, CERSAT sea-ice concentration
Oras	ECMWF ORAS5	NEMO3.4 + LIM2	SST, and SSS Nudging and surface waves	NEMOVAR (3DVAR) + Bias correction [5 days]	EN4 In-situ HadISSTv2 SST AVISO SLAs, OSTIA sea-ice concentration



There exist large differences in the data assimilation methods applied in terms of data assimilation scheme (three-dimensional variational data assimilation, 3DVAR, versus Singular Evolutive Extended Kalman Filter, SEEK), code (OceanVar versus NEMOVAR versus Systeme d'Assimilation Mercator, SAM), input observational datasets, surface data assimilation and nudging, frequency of



analysis and assimilation time-windows, bias correction schemes and error definitions for use in data assimilation, which introduce a large number of uncertainties as ensemble spread. Table 1 reports the details of each reanalysis member, highlighting all the aforementioned differences, together with reference publications for each GREP member. Additionally, initial conditions valid on 01-01-1993 differ among the GREP members, implying that also this kind of uncertainty is considered in the GREP ensemble. Note that Table 1 cannot easily encompass the variety of settings in the reanalyses. For details on the bulk formulation, bias correction methods and other details, refer to the documentation of the individual reanalyses provided in the table.

A preliminary assessment of GREP is performed through the use of CLASS4 metrics (statistics of observation minus model fields, in observation space, against EN4 in-situ profiles) and reported in detail in Appendix 1. The assessment provides an outlook about how close to the in-situ dataset the four reanalyses are, also compared to several objective analyses and climatology. Such assessment indicates that the reanalysis ensemble is generally closer to the profiles than objective analyses during well-observed period (after the Argo float deployment, i.e. since around 2005). In particular, for temperature, GREP-EM performs significantly better than objective analyses below 700 m. The ensemble mean always beats the individual members in approaching the in-situ profile dataset.

A second assessment is contained in Sect. 3.5, where GREP is compared to the ORA-IP subsample (OIS, vintage 2013, see Balamseda et al. 2015) of the reanalyses from the same four producing centers as the GREP members, i.e. the four predecessor reanalyses of GREP. This comparison allows us to assess the advances achieved during approximately five years of reanalysis production at global scale.

## 2.2. Definitions

In what follows, the ensemble mean (EM, hereafter GREP-EM for the GREP ensemble) is the arithmetic average performed over the GREP members. The ensemble spread ES (hereafter GREP-ES for the GREP ensemble) is defined as the standard deviation over the members. For ocean heat content and steric sea level, it is calculated on the anomalies with respect to the long-term mean field, while for temperature and salinity on the full fields. For horizontal and vertical averages, the resulting spread is the horizontal average of the spread of the vertically integrated values (e.g. the spread of the steric sea level in the Southern extra-Tropics 0–700 m as in Figure 11 is the horizontal average over the 60°S to 20°S region of the spread of the 0–700 m steric sea level). When used for

background-error covariance calculation (Sect. 3.4), the grid-point spread is calculated over anomalies collected also from neighboring grid-points and temporal records, in order to artificially increase the ensemble size.

The ratio between ensemble mean and ensemble standard deviation is used within several diagnostics of this article to quantify the significance of the trend. Trends are computed as linear regressions from yearly mean values. The ensemble mean (e.g. of trends) represents the signal, while the ensemble standard deviation is the noise that quantifies the fluctuations around the signal, implied by the sources of errors in reanalyses. In particular, if the distribution of trends is assumed to follow a Normal distribution, the significance is achieved at about 95% confidence level if the mean exceeds twice the spread, and at 99% confidence level if it exceeds thrice the spread.

As in Storto et al. (2017), we define the steric sea level as the local or global effect of the seawater mass expansion or contraction (Griffies and Greatbatch 2012), calculated as the vertical integral of the density anomaly over the water column (the time- and space- varying density divided by the reference density, here equal to  $1026 \text{ kg m}^{-3}$ ). Thermosteric (halosteric) contributions are calculated with the same approach but imposing the salinity (temperature) equal to the yearly three-dimensional climatological field. For calculating ocean heat content, we use values of  $3991.87 \text{ J K}^{-1}$  for the seawater heat capacity, and  $1026 \text{ kg m}^{-3}$  for the seawater density. Finally, except for the observation space-based diagnostics (Sect. 3.1 and Appendix 1), all the diagnostics are calculated from the  $1^\circ \times 1^\circ$  degree horizontal resolution dataset as released through CMEMS.

## 3. Applications

### 3.1. Quantifying the reanalysis uncertainty

In this section, we investigate the feasibility of using the ensemble spread for uncertainty quantification. The spread is compared with the RMSE of the ensemble mean, which in turn quantifies the GREP-EM accuracy. Before analyzing the temporal variability of the GREP-ES, we assess the 1993–2016 total ensemble spread, reported in Figure S1 in terms of steric sea level and its components for the top 700 m vertical layer, and calculated from both monthly (left panels) and yearly (right panels) means. Spread from monthly means peaks at more than 10 cm around mesoscale active regions, the eastern Tropical Pacific and Pacific warm pool, Equatorial Atlantic and Indian Ocean (especially in the middle of the Indian gyre). Similar patterns are found for the total steric and the thermosteric sea level spread. The halosteric sea level spread is generally much smaller, peaking at around 8 cm only in mesoscale active areas and in the Arctic

Ocean. The spread from yearly means is obviously smaller than that from monthly means. It presents however slightly different characteristics: on the inter-annual scale, the spread in the western Tropical Pacific ocean and in the Arctic Ocean becomes prominent. In general, further to eddy active regions (Gulf Stream, Kuroshio, extension, Antarctic Circumpolar Current, and other western boundary regions), where the high variability induces high ensemble spread, the reanalyses tend mostly to disagree in the Equatorial band, in the Indian ocean, and in the Arctic region. The time-mean ensemble anomalies from the individual reanalyses (not shown) indicate that cglo and foam tend to be warmer (higher steric sea level) and glor and oras cooler. These features persist during both decades (1993–2004 and 2005–2016), although during 2005–2016 the anomalies reduce compared to the first decade.

Both RMSE and ensemble spread are computed in observation space (i.e. at observation location, and later grouped into monthly RMSE and spread), following an approach similar to that of Yamaguchi et al. (2016). The RMSE is evaluated against EN4 in-situ profiles in observation space. Although these profiles are not independent since they are assimilated by three out of four reanalyses, thus representing a sub-optimal validation dataset opposed to independent observations (Yamaguchi et al. 2016), we assume that the different temporal scale of the datasets (monthly mean data from reanalyses, instantaneous values from observations) helps limit the correlation among the two datasets. Furthermore, there is no availability of independent validation data for both temperature and salinity, at global scale and during the entire study period, unless reanalyses were conceived with the withholding of certain observations, which was not the case for the GREP reanalyses. However, this limitation should be taken into account.

The comparison between RMSE and ensemble spread is commonly adopted (Fortin et al. 2014) to evaluate whether the ensemble is or is not under-dispersive (over-dispersive), namely when spread under-estimates (over-estimates) the uncertainty. More sophisticated diagnostics are available in general for ensemble prediction systems (e.g. Candille and Talagrand 2005), which are not exploited here due to the limited ensemble size. The purpose of these diagnostics is to assess whether the spread itself can be used for uncertainty quantification, even for a small ensemble size as GREP.

Figure 1 shows the monthly time-series of the GREP-EM RMSE of temperature and the GREP-ES, over the global ocean, along with yearly values (thick lines) and normalized values, i.e. subtracting the long-term mean value from each of the two timeseries and dividing by its temporal standard deviation, in a way

similar to the normalization of anomalies (Hart and Grumm 2001). The purpose of the normalization is to visualize the coherence of the timeseries, i.e. excluding any possible under- or over- dispersion issue when the timeseries are compared. We focus here on temperature only for sake of simplicity; salinity diagnostics are summarized in the following figures. Since only four members are considered, the GREP ensemble is clearly under-dispersive, providing a time-mean spread of 0.47 °C against 0.88 °C for the GREP-EM RMSE. The ratio of the two timeseries standard deviation, equal to 1.8, provides a way to quantify the degree of under-dispersion of the GREP ensemble over time. The correlation between the two timeseries is equal to 0.86 (0.94 for yearly values), which is significant above the 99% confidence level. Once the normalization is applied, it appears that the two timeseries closely resemble each other, with an initial increase of both spread and RMSE, likely due to the different initial conditions used by the four reanalyses in association with the poor observing network and the different observation bias correction methodologies. Since the beginning of 2000s, the two timeseries exhibit a consistent decrease due to the Argo float deployment. Sporadic increases occur in 2011 and mid-2013, the former probably due to the strong La Nina event (see also Sect. 3.4).

### **Fig. 1**

Monthly (thin lines) and yearly (thick lines) timeseries of GREP-EM RMSE (black lines) and ensemble spread (GREP-ES, red lines) in observation space for temperature. Top panel: raw values; Bottom panel: normalized as explained in the text. Blue bars represent the monthly number of in-situ single-level observations (in millions)

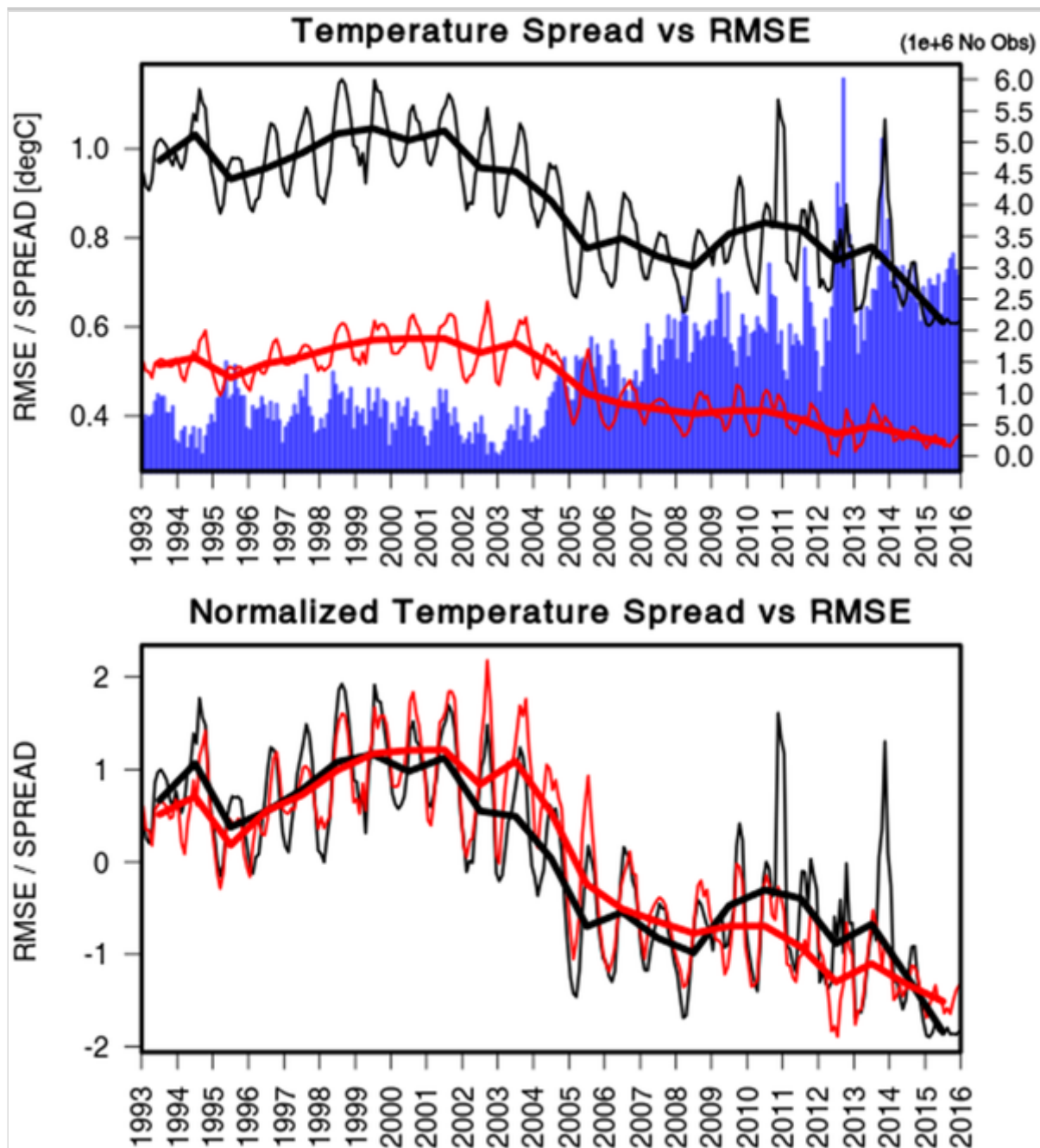


Figure 2 shows the correlation matrix between ensemble spread and RMSE for different horizontal and vertical regions, thus summarizing how well the spread behavior with time reproduces flow dependent errors of reanalyses. Correlations are calculated for temperature and salinity separately, both for monthly and yearly timeseries. With the exception of yearly salinity in the 800–1500 layer, correlations are all significant at 99% confidence level, meaning that the spread captures the temporal evolution of skill. Generally, temperature correlations are higher than salinity ones, and yearly scores are higher than the monthly ones, due to some differences in the seasonality between GREP-EM RMSE and GREP-ES (not shown). Furthermore, global and northern hemisphere values of correlations are generally higher at the monthly time-scale, while the tropics exhibit the largest correlations at yearly time scale. Generally, the southern

hemisphere exhibits the smallest correlations, suggesting that the observing network and eventual model deficiencies may worsen the temporal consistency between spread and RMSE, due to the outlier behavior of some members.

**Fig. 2**

Matrix plot for temperature and salinity correlations of GREP-ES with GREP-EM RMSE, for monthly and yearly values over the period 1993–2015, in categories of horizontal and vertical layers (N. ET: North extra-Tropics, from 60°N to 20°N; TROP: Tropics, from 20°N to 20°S; S. ET: South extra-Tropics, from 20°S to 60°S)

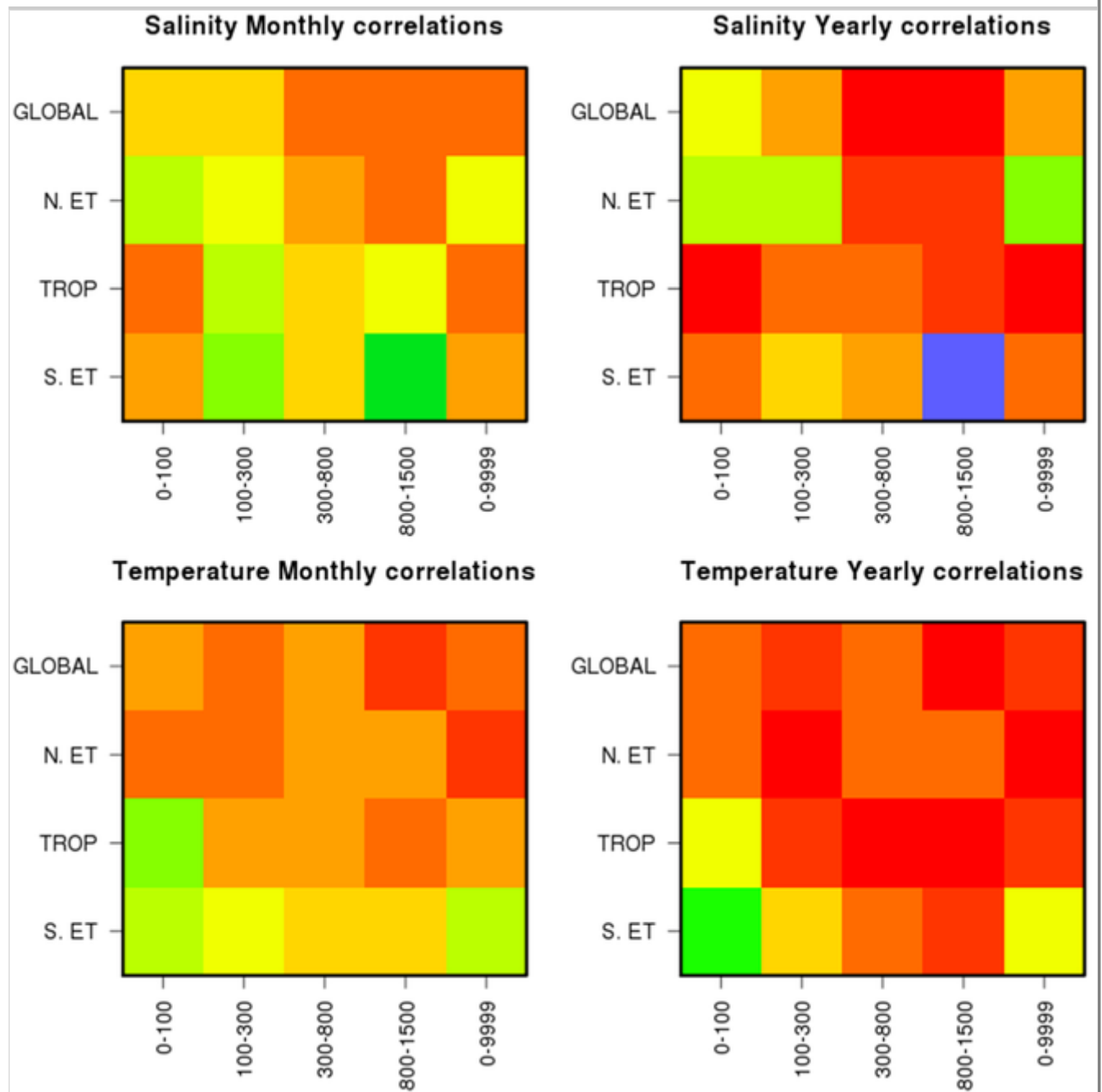


Figure S2 reports the ratio between the ensemble mean RMSE standard deviation divided by the ensemble spread standard deviation. All temperature diagnostics show values greater than 1, i.e. the ensemble system is under-dispersive, especially at the sea surface and in the Southern Ocean. On the contrary, salinity ratios show values close or smaller than 1, namely the system is over-dispersive, likely associated with the facts that (1) the lack of salinity observations and the sensitivity of reanalyses to model parameterizations and boundary conditions (atmospheric forcing, river and ice runoff) may cause the spread to exceed the ensemble mean error and (2) reanalysis systems may over-fit temperature observations, especially near the surface where two reanalyses implement a nudging scheme to SST analyses, leading to such discrepancy of ratios between temperature and salinity. This analysis suggests that, by taking into account the degree of dispersion of the ensemble with respect to the ensemble mean error, the spread in all layers and regions is able to quantify the temporal evolution of the GREP uncertainty, although re-scaling must be considered, as usually done in ensemble data assimilation (e.g. Rainwater and Hunt 2013). Particular care must be taken to poorly observed regions, where the lack of observational constraints induces over-dispersion in the GREP ensemble.

### 3.2. Applications to ocean monitoring indexes

One major activity undertaken by CMEMS is the yearly production of the Ocean State Report (OSR, von Schuckman et al. 2017, 2018), which has the objective of providing a report on the state of the global ocean and European regional seas for the ocean community and decision-makers. It will be released on yearly basis through publication in scientific journal. The latest release (von Schuckman et al. 2018) is organized in sections, each describing essential ocean variables (see [http://www.goosocean.org/index.php?option=com\\_content&view=article&id=14&Itemid=114](http://www.goosocean.org/index.php?option=com_content&view=article&id=14&Itemid=114)), climate monitoring indexes, regional in-depth analyses and special events that occurred during the last year, all reproduced by CMEMS products.

This latest OSR has made extensive use of the GREP reanalyses for many observed and unobserved key climate ocean parameters, with the underlying idea that the use of GREP provides an uncertainty envelope that helps quantify the confidence of the inter-annual variability, such as for instance decadal trends in ocean heat content, transports at relevant section, Arctic and Antarctic sea-ice volume, etc. For instance, freshwater transports, ocean heat content and steric sea level trends and 2016 anomalies, along with many additional global and regional diagnostics were assessed through the use of GREP ensemble mean and spread. A novel approach in the OSR analyses was the recourse to super-

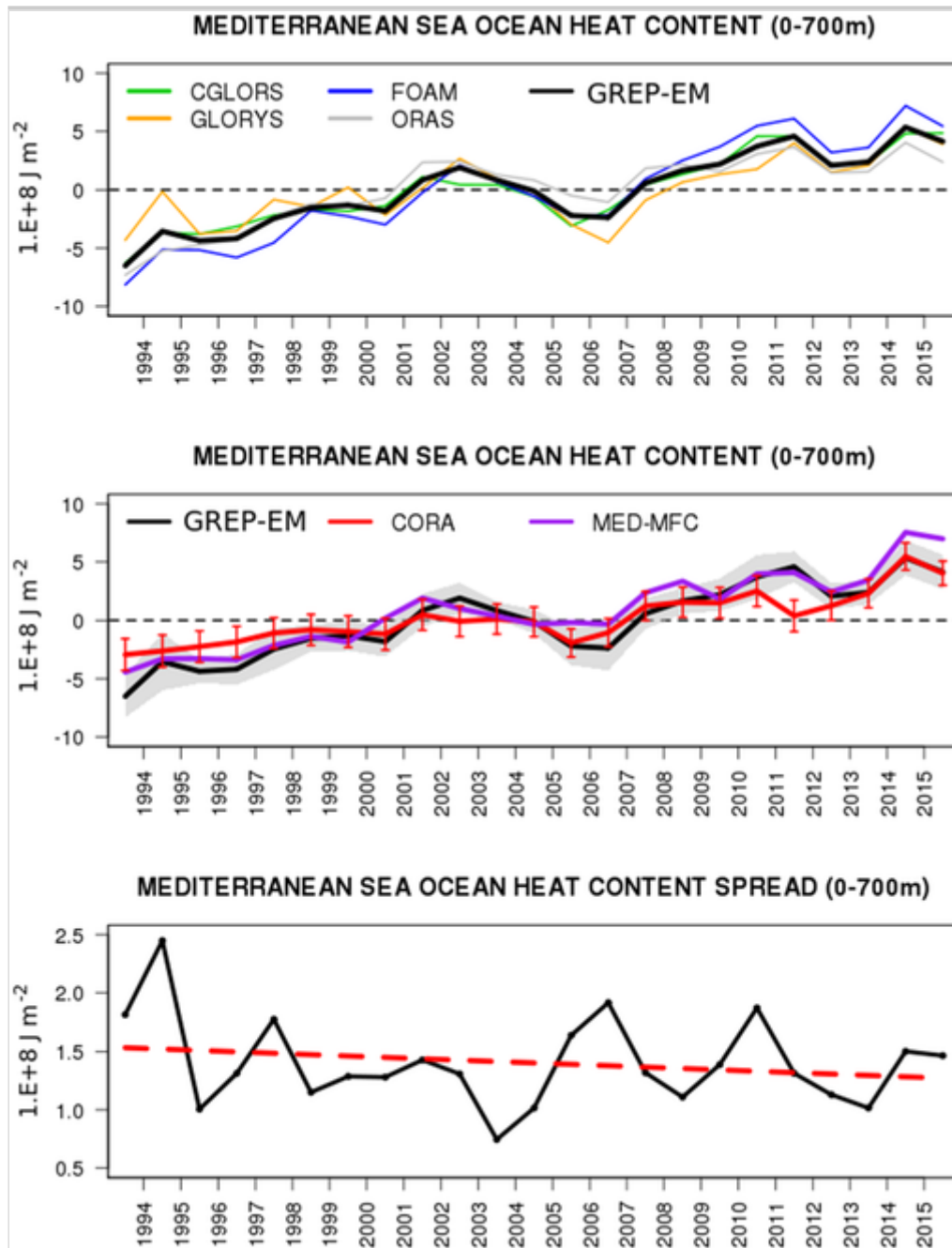
ensemble, an approach that combines GREP members with regional reanalyses and observation-only products for regional assessment. In this context, while high-resolution regional products may in principle better explain the small-scale variability compared to coarser global products, these latter provide uncertainty estimates. Thus, there exists an intrinsic complementarity between high-resolution regional and coarse resolution global reanalyses.

This exercise offers the possibility to investigate the consistency between global and regional products for selected areas. For instance, we focus here on the representation of ocean heat content variability simulated by global and regional reanalyses in the Mediterranean Sea. We detail here the assessment in the Mediterranean Sea, because of its data abundance and consolidated knowledge, other European seas being assessed in the OSR. Figure 3 shows the time series of the yearly heat content anomalies in the Mediterranean upper layer (top 700 m) for the period 1993–2015. The top panel presents the four global reanalyses along with the GREP ensemble mean, indicating the large consistency among the products. In particular, the four global reanalyses and GREP-EM present negative ocean heat content anomalies before 2001 and in 2005–2006, while after 2008 the anomalies are positive, reaching the maximum value in 2014.

### **Fig. 3**

Ocean heat content anomaly timeseries (0–700 m) in the Mediterranean Sea from the four GREP members and the ensemble mean (top panel); from GREP-EM, the CMEMS Mediterranean Sea Monitoring and Forecasting Center (MED-MFC) regional reanalysis and the CORA observation-only product (middle panel), with +/- one ensemble standard deviation represented by the grey shading and the CORA error estimated by the red bars; the ensemble spread from GREP (bottom panel) and the linearly fit line (dashed red line). Values are yearly means





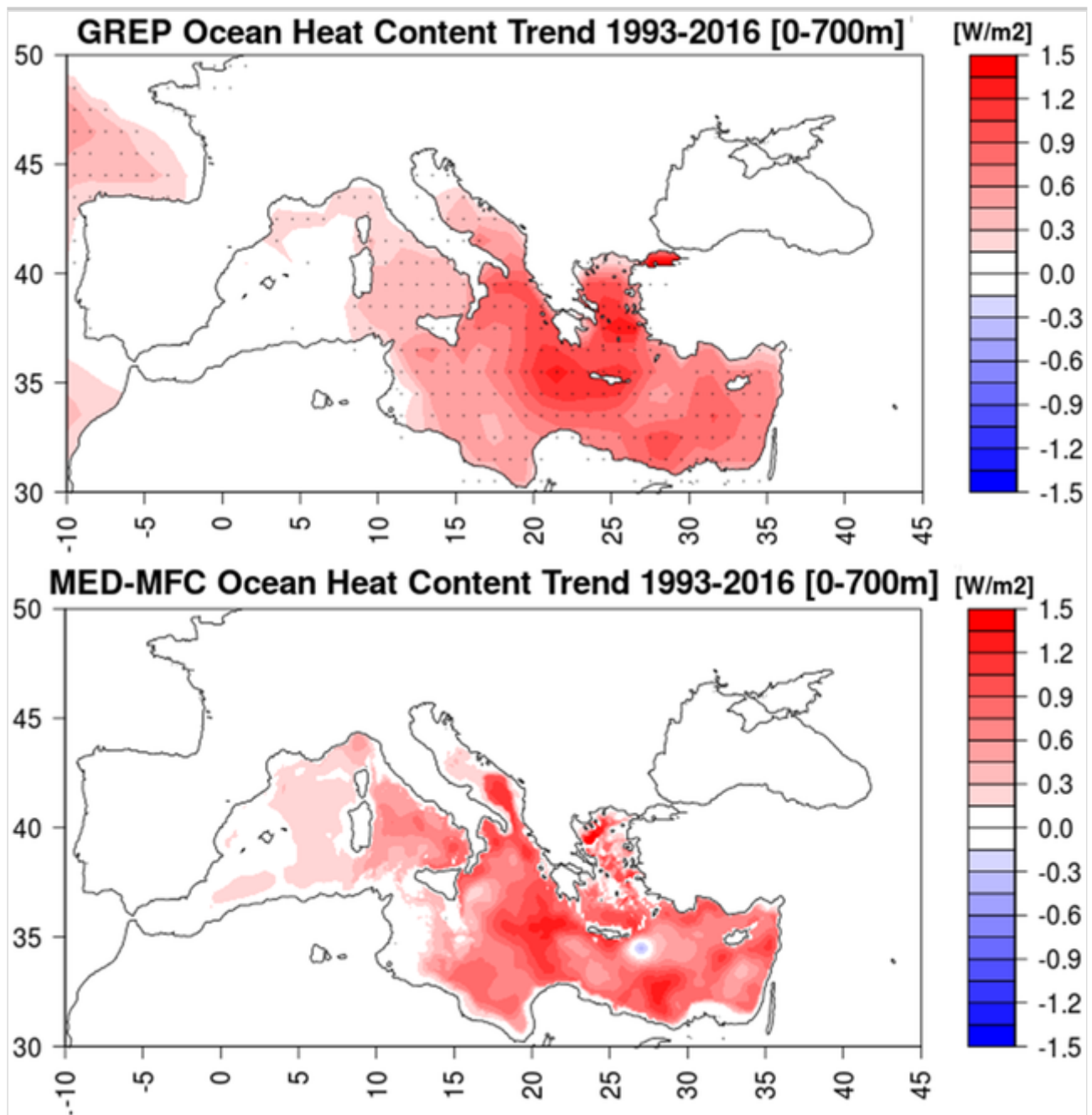
In the middle panel, time series of GREP-EM heat content is compared to the regional reanalysis product (Mediterranean Sea Monitoring and Forecasting Center, MED-MFC, Simoncelli et al. 2014, 2016) and the Coriolis Ocean database ReAnalysis (CORA, Cabanes et al. 2013). The regional MED-MFC reanalysis presents negative anomalies from 1993 to 1999, consistently with GREP-EM, but its positive anomaly starts one year before, peaking in 2001 and remaining slightly positive in 2002 and 2003. Between 2004 and 2006, the MED-MFC anomaly is neutral and increases again in 2006, coherently with the GREP-EM. The estimate from CORA (Sect. 2.1 in von Schuckman et al. 2018), only based on observations, presents negative anomalies before 2001, smaller than GREP-EM and MED-MFC, then remains close to zero until 2004, and has slightly negative values in 2005–2006. In 2007, CORA estimates become

positive too, peaking in 2014 as the other products. In 2011, CORA anomaly is close to zero, differently from GREP-EM and MED-MFC. The regional MED-MFC product resides among the GREP ensemble spread and the CORA uncertainty except for the last two years, when it is the largest. von Schuckman et al. (2018) relate this to the large temperature anomalies that characterize the surface and intermediate layers of the Eastern Mediterranean, in particular in the Southern Adriatic, the Ionian and the Aegean seas. Finally, the bottom panel shows the GREP ensemble spread as a function of time, with the dashed red line representing its linear trend. A decrease of the spread with time is found (linear trend equal to  $-0.04 \pm 0.04 \text{ W m}^{-2}$ ), likely due to the improved observational sampling, although not statistically significant.

Figure 4 reports the map of the heat content trend in upper 700 m, computed for the period 1993–2016, for the GREP-EM, with significant trend from spread analysis over-plotted, together with the MED-MFC product. While the regional product exhibits finer scale structures than the global ensemble, the spatial patterns of warming areas are in close agreement, confirming a large consistency between the two products. The products consistently identify the eastern Mediterranean Sea as the region with enhanced warming, consistently with observation-based independent estimates (e.g. Marbà et al. 2015), and due to the warming and drying of the regional climate, which favors the formation of warmer and saltier Levantine intermediate waters from surface waters in either the Levantine or Cretan Sea (Schroeder et al. 2017). Warming peaks occur west of Cyprus and north of the Egyptian shoreline, while damped warming is found in correspondence of the Ierapetra gyre. Moreover, the use of the ensemble information provides the trend significance, which represents a crucial information complementary to the high-resolution MED-MFC deterministic trend. The trend confirms the warming of the Eastern Mediterranean basin with the largest tendencies in the Southern Adriatic, the Ionian and the Aegean seas.

#### **Fig. 4**

Maps of ocean heat content (0–700 m) trend in the Mediterranean Sea from GREP (top panel) and the CMEMS Mediterranean Sea Monitoring and Forecasting Center (MED-MFC) regional reanalysis (bottom panel). In the top panel, dots correspond to points that show significant trends, defined as the ratios between ensemble mean and spread of trends exceeding 2 (95% confidence level)



### 3.3. Quantifying the observing networks impact

Ensemble spread from multi-system ocean reanalyses may capture changes in the observing networks, provided that additional (or missing) observation types lead to a decrease (or increase) of ensemble spread. This idea is implicitly contained in any ensemble data assimilation system (e.g. Burgers et al. 1998) and has been recently exploited by Xue et al. (2017) that stressed the importance of the long-term stability of observing systems for anomaly monitoring by showing the increase of multi-model ensemble spread associated to the TAO (Tropical Atmosphere Ocean observing array) crisis in 2012. Ensemble systems might also be used within ensemble data-denial experiments to quantify the impact of different observing networks. Some existing strategies are (1) the

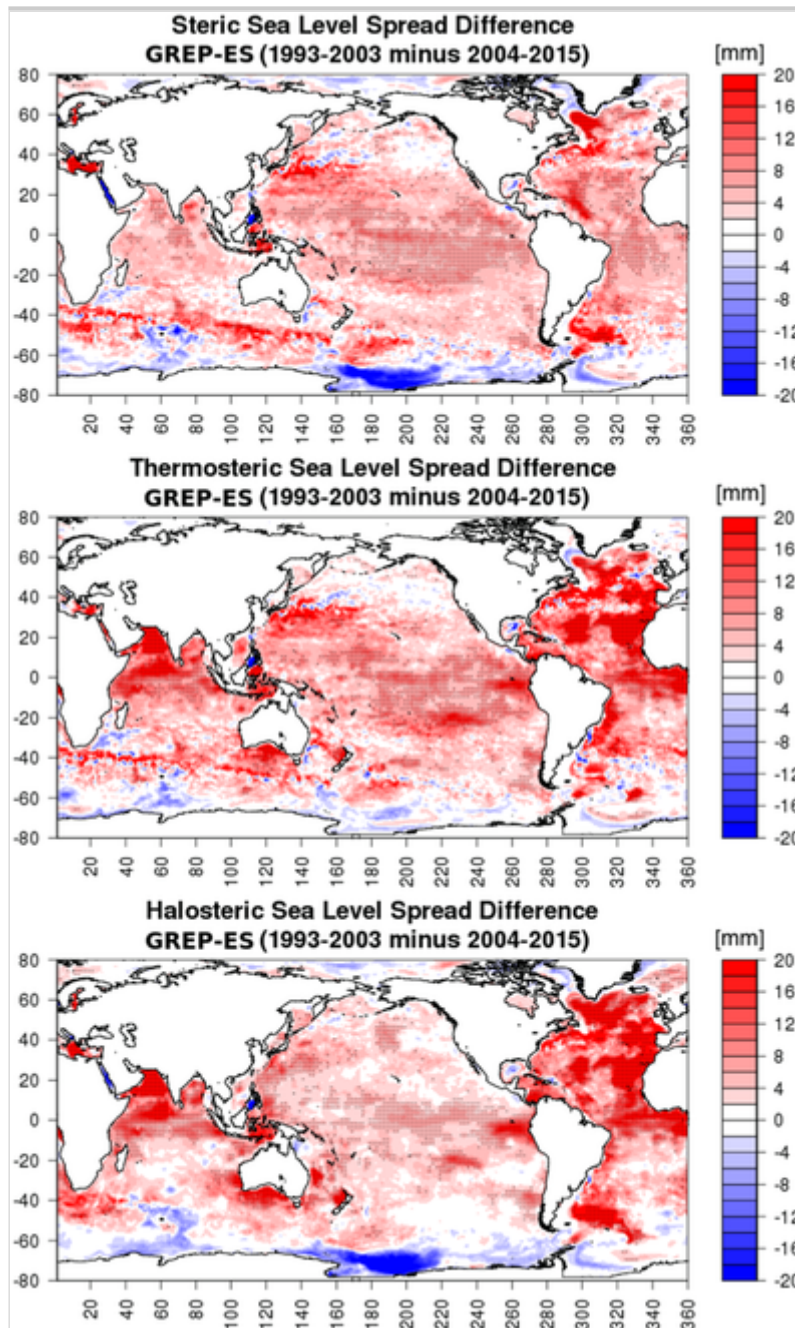
assessment of spread increase corresponding to individual observing network withholdings (Storto et al. 2013); or (2) the use of recently proposed metrics for ensemble-based observation impact—such as Ensemble Forecast Sensitivity to Observations (EFSO, Ota et al. 2013) that quantify the sensitivity of forecasts to individual observing network within ensemble data assimilation. Nonetheless, here we limit our study to more qualitative assessment, due to the practical limitation that the four reanalyses assimilate different observational datasets.

While the consistency of ocean reanalyses depends not only on the observing systems, but also on the accuracy of ocean models, atmospheric forcing and simplification in data assimilation schemes, which are extremely hard to disentangle, abrupt changes with time of the ensemble spread are to a first approximation due to changes in the ocean observing network (Hu and Kumar 2015). Changes in the atmospheric observing networks may also influence the reanalysis consistency (Xue et al. 2011; Storto and Masina 2017), therefore care should be taken in this kind of assessments.

In this section, we examine two notable cases where the effect of changes in observation sampling on ensemble spread is evaluated. First, time-averaged ensemble spread is considered for steric, thermo- and halosteric contribution to sea level, in the top 700 m of the global ocean. Figure 5 shows the difference of the spread computed in the Argo-poor (1993–2003) and Argo-rich (2004–2015) periods. It is worth mentioning that we subjectively choose the cut-off at the beginning 2004, which roughly corresponds to the year where 25% of floats were deployed compared to the last year of the reanalysis. Positive values are associated to spread decrease, while dots are in correspondence of grid-boxes that exhibit statistically significant differences, based on the ratio with respect to the ensemble spread as explained in Sect. 2.2.

### **Fig. 5**

Differences of steric, thermosteric and halosteric ensemble spread in the Global Ocean between the periods 1993–2003 and 2004–2015. Dots correspond to points that show significant differences, defined as those exceeding 2 times the average spread (95% confidence level), calculated over the entire period. Spread difference is between poorly (1993–2003) and well (2004–2015) observed periods: red (blue) indicates spread increase (decrease) associated with poorly developed Argo observing networks



The impact of the availability of observations on the total steric sea level is evident all over the World Ocean, with large differences in the Tropical areas of all basins, in the Antarctic Circumpolar Current (ACC), in the western boundary current, and within semi-enclosed basins such as the Mediterranean, the Baltic and the Labrador seas. The spread increases with time only in a few areas, mainly located at high-latitudes (from 60° poleward), but these are regions that the Argo floats do not reach, and where the inconsistencies of freshwater anomalies lead to increased halosteric sea level spread (not shown).

Interestingly, the spread reduction for thermo and halo steric sea level in many areas is larger than the spread reduction for total steric sea level (middle and bottom panels). For instance, the Atlantic (especially the North) and the Indian

oceans, and a few others areas such as the Great Australian Bight and the Eastern Pacific Ocean, show larger spread decrease than the total steric, suggesting that Argo floats, besides contributing to the total steric sea level accuracy increase, also contribute to reducing the uncertainty of the thermo and halo steric partition in these regions. This in turn means that before Argo float deployment the salinity is largely unconstrained, leading to wrong thermo-halo steric partitions. Argo floats thus represent a crucial observing network for capturing temperature and salinity compensation and correlations.

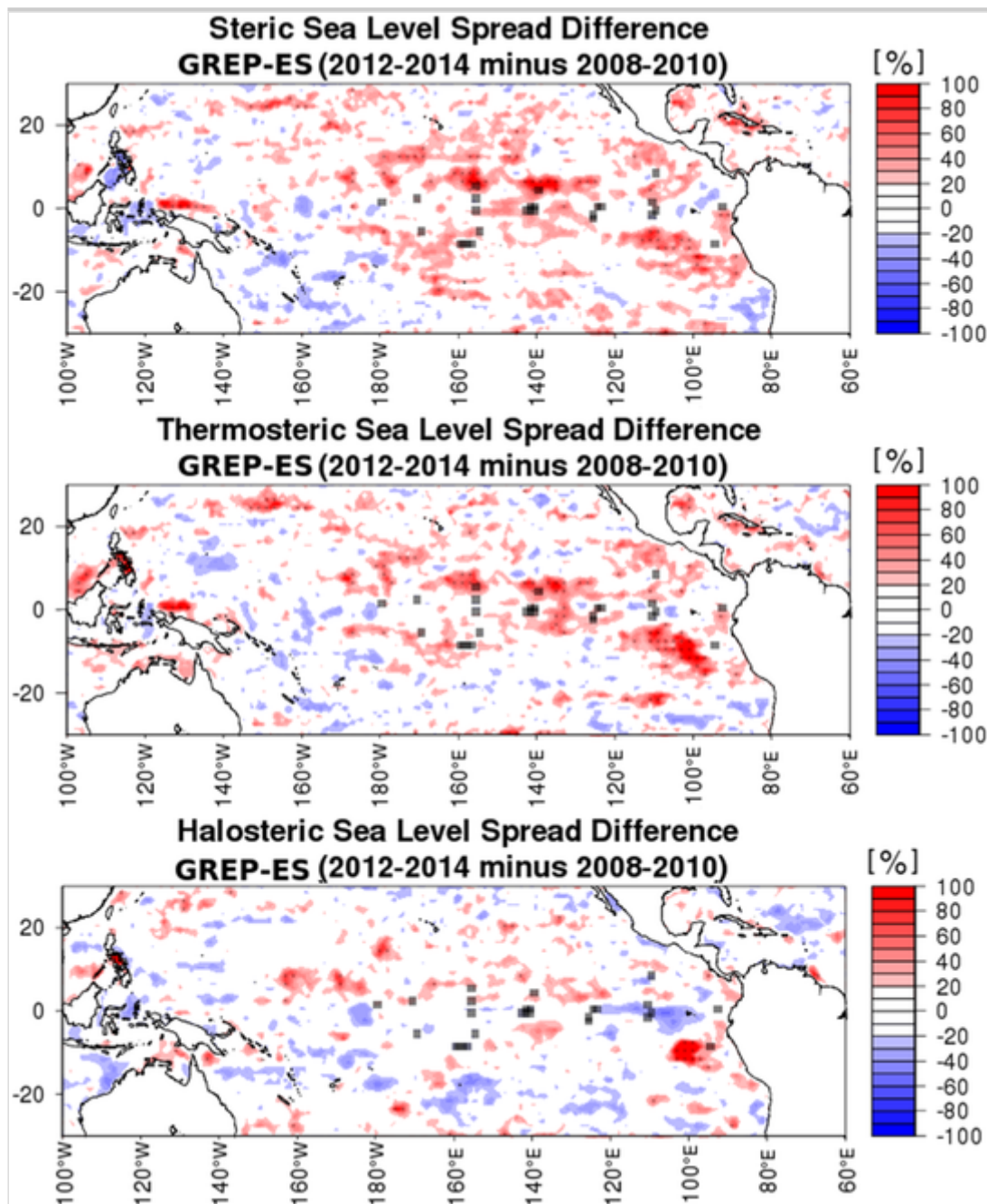
The second example where the multi-model ensemble spread reproduces the observing system changes concerns the impact of the reduction of TAO/TRITON mooring data after 2012, to some extent in agreement with the study by Xue et al. (2017). TAO/TRITON moorings have crucial impact on predictions at both short, seasonal and decadal time scales (e.g. Fujii et al. 2015), and here we monitor their impact on the GREP reanalysis ensemble spread. Similar diagnostics on steric sea level spread differences shown in Fig. 5 is presented in Fig. 2 for the Tropical Pacific region considering the 2008–2010 and 2012–2014 as period pre- and post- TAO crisis occurrence, respectively. The figure presents this time results in terms of percentage difference, and also shows the location of the missing TAO moorings. Positive values are associated with spread increase corresponding to the decline of TAO observations during the 2012–2014 period, while dots are in correspondence of grid-boxes that exhibit statistically significant differences. Between 2012 and 2014 several moorings located between  $150^{\circ}$  W and  $95^{\circ}$  W failed, causing a relatively sudden lack of data in the eastern part of the Tropical Pacific Ocean. Consequently, these areas are characterized by steric sea level spread increase, particularly significant in the region  $180^{\circ}$ W– $130^{\circ}$ W,  $2^{\circ}$ – $16^{\circ}$ N, and  $110^{\circ}$ W– $100^{\circ}$ W,  $5^{\circ}$ – $15^{\circ}$ S. Using the GREP ensemble, the augmented uncertainty is mostly attributable to temperature spread increase in the central Pacific area and, mostly, in the eastern Pacific area, except for a local significant effect of salinity spread increase around  $100^{\circ}$ W,  $10^{\circ}$ S. Note that the figures present quite noisy features that may be ascribed to the acknowledged spurious variability introduced by TAO data assimilation, in both oceanic and atmospheric reanalyses (Josey et al. 2014). We argue that for this study period, the TAO moorings mostly improved the heat content variability representation, and only to a lesser extent the salinity variability, based on the results of Fig. 6, and noting that there is however a small amount of salinity data available from TAO moorings. However, different climate regimes occurring in these two short periods (2008–2010 versus 2012–2014), for instance related to ENSO variability, may affect the spread, so that results cannot be generalized. Nonetheless, spots of main differences are rather localized and in proximity of the failed TAO moorings with displacement due to local advection

effects, thus confirming the capability of the spread in capturing the change of the mooring observing network.

**Fig. 6**

Percentage differences of steric, thermosteric and halosteric ensemble spread in the Tropical Pacific Ocean between the periods 2012–2014 and 2008–2010. Small gray squares are located in correspondence of TAO moorings that have much less or no data disseminated during 2012–2014 compared to 2008–2010. The percentage difference is defined as spread difference normalized by the spread of the first period and multiplied by 100. Dots correspond to points that show significant differences, defined as those exceeding 2 times the average spread (95% confidence level). Spread difference is between poorly (2012–2014) and well (2008–2010) observed periods: red (blue) indicates spread increase (decrease) associated with failed TAO moorings





Such analyses exemplify the potential of multi-model ensemble spread evaluation for monitoring the changes of the ocean observing network, for potential use also in near real-time monitoring applications. In particular, the example that investigates the TAO crisis reveals the different importance of observed parameters (temperature versus salinity) depending on the specific region (central versus eastern Tropical Pacific), and indicates that spread assessment can be therefore used to identify the relative importance of the observed parameters, once the spread has been evaluated as reliable.

### 3.4. Multi-model flow-dependent error statistics for use in hybrid data assimilation



A relatively novel direction in geophysical data assimilation is the recourse to schemes where the forecast errors (so called, background errors, through which a forecast is weighed onto the analysis relative to the observations and their errors,) are formulated merging (hybridization) both variational and ensemble scheme formulations. Hybrid ensemble-variational data assimilation schemes have been introduced in Numerical Weather Prediction systems (Hamill and Snyder 2000) and have become popular during the last decade since they succeed in introducing time-varying (i.e. flow-dependent) error statistics, which are usually neglected in variational data assimilation, in a relatively straightforward way. In practice, hybrid data assimilation combines together stationary (time-invariant) background-error statistics, usually adopted in variational schemes, with ensemble-derived error statistics that bear information about the “errors of the day”. Their power resides in merging advantages of both variational formulations, generally considered very robust, with ensemble data assimilation, generally considered effective in capturing the variability of the errors. Recent attempts of using hybrid schemes in ocean analysis systems include the global hybrid ensemble-variational scheme introduced by Penny et al. (2015), and the regional hybrid assimilation system proposed by Oddo et al. (2016).

In spite of the GREP ensemble size, the availability of multi-model ensemble statistics provides a flow-dependent error dataset that can be used to estimate time-varying error covariances. Indeed, using a multi-model ensemble has important advantages for reanalysis applications, as several reanalyses were produced worldwide (Balmaseda et al. 2015) and may span a larger variety of uncertainty than single-model ensembles. Such an approach can be easily adopted for future reanalyses. It does not require to perform an ensemble system but only to use reanalyses already available, making this application very attractive, although with obvious technical limitations for near real-time applications.

Recently, Storto et al. (2018) proposed a possible approach to include hybrid covariances into the CMCC OceanVar data assimilation system through the strategy of augmenting the control vector, originally proposed by Lorenc (2003). The formulation is here recalled very briefly. In terms of the control vector transformation, the ocean state increments (**dx**) are PLEASE MAKE THE dx as in the following equation (should be bold and italic for consistency) decomposed in:

$$\mathbf{dx} = \beta_c \mathbf{dx}_c + \beta_e \mathbf{dx}_e$$

where  $\mathbf{dx}_c$  and  $\mathbf{dx}_e$  are the two components of the control vector, associated with the climatological (static) and ensemble-derived (flow-dependent) covariances, respectively. The  $\beta_c$  and  $\beta_e$  parameters determine the relative weight given to static and flow-dependent covariances. Following Wang et al. (2007), such a formulation is equivalent to considering a hybrid background-error covariance matrix  $\mathbf{B}$  of the form:

$$\mathbf{B} = \alpha \mathbf{B}_c + (1 - \alpha) \mathbf{B}_e$$

with  $\beta_c^2 = \alpha$  and  $\beta_e^2 = (1 - \alpha)$ , assuming independence of the two control vectors. The static component is usually retained to provide full-rank covariance matrix, primarily for robustness purposes, provided the limited ensemble size and the non-optimality of the uncertainty spanned by the GREP ensemble.

OceanVar decomposes the background-error covariances in multi-variate vertical EOFs of temperature and salinity, for the vertical covariance component, and horizontal correlations achieved through the application of recursive filters. In these regards, vertical EOFs and spatially-varying horizontal correlation length-scales were estimated from the GREP monthly mean ensemble anomalies for each month. In order to limit spurious covariances due to the small ensemble size, flow-dependent covariances are calculated for each month using the monthly mean ensemble anomalies of the GREP members of the current month, together with the previous and the following month. Furthermore, following Raynaud et al. (2008), each grid-point uses ensemble anomalies that fall in a  $9 \times 9$  grid-point box centered in the grid-point itself (spatial moving average) in order to calculate vertical covariances and thus recover from the small ensemble size. This might slightly smooth out the covariance fields but it is a necessary step to artificially increase the ensemble size; alternative approach based on localization may also be considered. Correlation length-scales are calculated through fitting the empirical correlation curve, as a function of the distance, to a Gaussian curve. Thus, the multi-model ensemble covariances feed both the vertical EOFs and the horizontal correlation length-scales embedded in the  $\mathbf{B}_e$  covariance matrix, and complement the stationary  $\mathbf{B}_c$ . Both components use the dynamic height balance operator to map temperature and salinity increments onto increments of sea level, as in Storto et al. (2011).

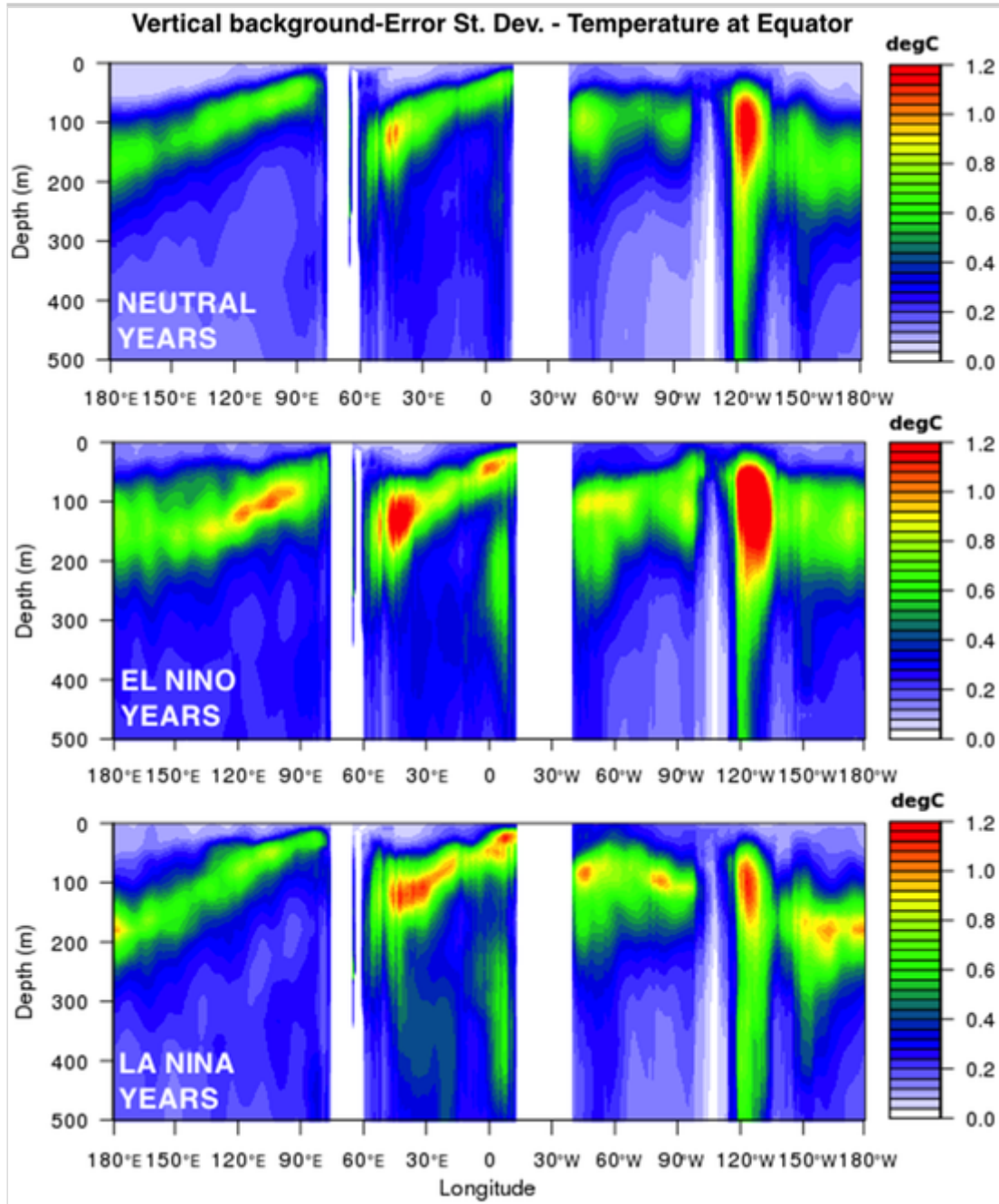
The rationale behind the use of this approach is the ability of the time-varying ensemble spread in capturing the uncertainty. This has been discussed previously in the manuscript in terms of the correlation between the RMSE ensemble mean and the ensemble spread (Sect. 3.1) and in terms of the impact of the observing network sampling on the ensemble spread (Sect. 3.3). Furthermore, the multi-

model ensemble may bear information useful to characterize the reanalysis errors depending on the climate regime. For each month, three-dimensional background-error covariances are calculated as covariance of the ensemble anomalies from the three-month period centered over the chosen month, and considering also the eight nearest grid-points, for a total of 108 anomalies (4 members, 3 months, 9 grid-points). This procedure is adopted in order to increase artificially the ensemble size.

As notable proof-of-concept, we compare the ensemble-derived background-error standard deviations (from  $B_e$ ) in the Tropics (meridionally averaged along 10°S–10°N), estimated from the multi-model ensemble during ENSO events, i.e. separating neutral, El Niño and La Niña years (Fig. 7). El Niño occurrences are characterized by an increase and broadening of vertical covariances with respect to the neutral years (especially in the eastern Tropical Pacific, see also Figure S3), due to the larger thermocline variability during El Niño compared to neutral and La Niña years. The El Niño and La Niña years bear larger errors in the Tropical Atlantic Ocean as well. The comparison of composite error covariances thus motivates the use of multi-model ensemble approach for the flow-dependent component.

### Fig. 7

Composite ensemble-derived ( $B_e$ ) background-error standard deviation of temperature in the Tropics as a function of longitude and depth. From top to bottom, for neutral, El Niño and La Niña months, respectively. The years are defined according to the Nino3.4 index: an El Niño or La Niña month is identified if the 3-month running-average of the Nino3.4 (5°N to 5°S, 170°W to 120°W) mean SST exceeds + 0.4 °C for El Niño or – 0.4 °C for La Niña. SST monthly data are extracted from HadISST (Rayner et al. 2003). There are 174, 39 and 62 occurrences for neutral, El Niño or La Niña months, respectively



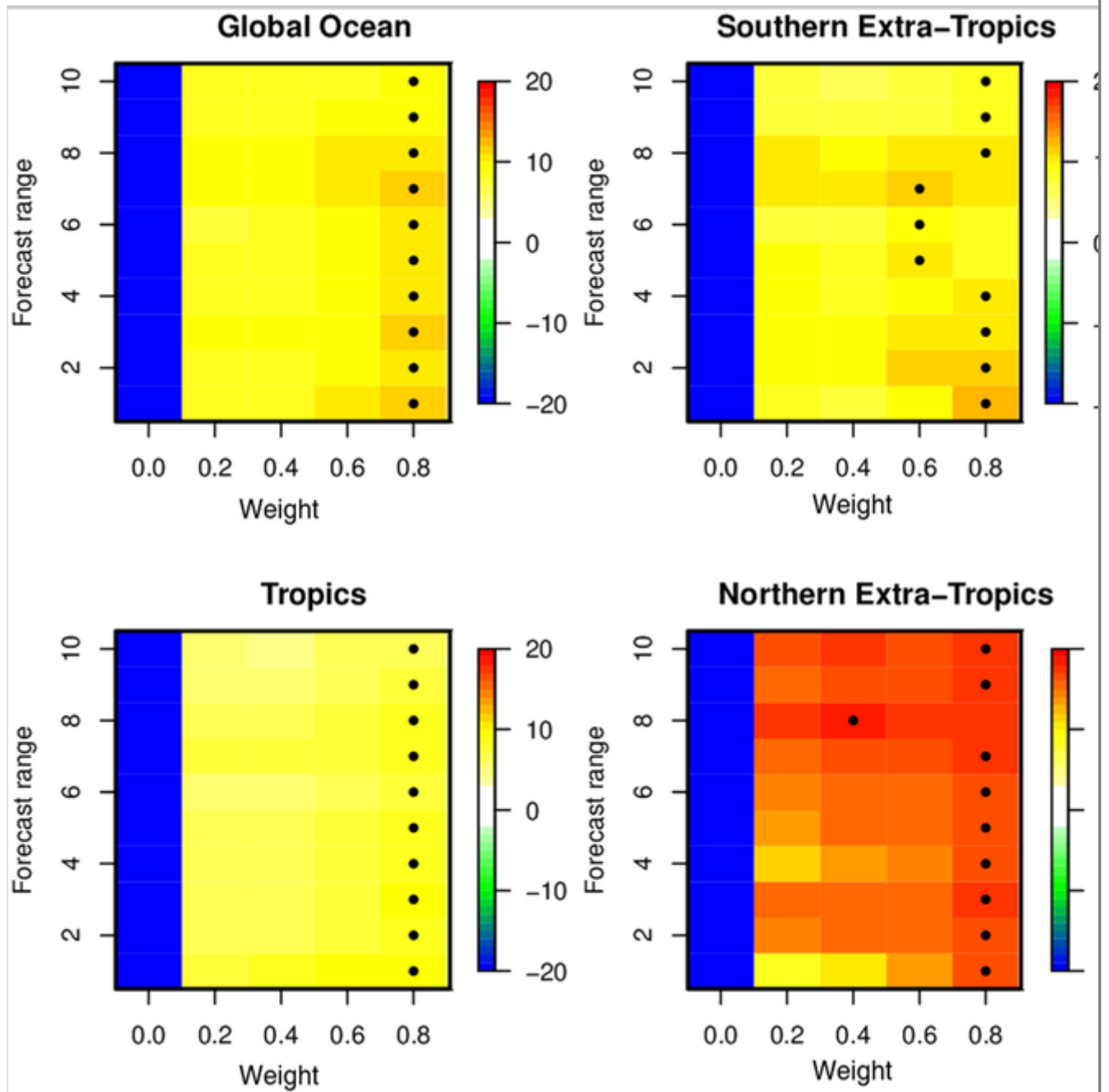
The adoption of the hybrid data assimilation with multi-model covariances has been tested in a coarse-resolution analysis system, performed with NEMO v3.6 in the global ORCA2L31 configuration for the period 1994–2014, similarly to the experimental configuration used by Storto et al. (2018). The experimental setup assimilates all in-situ hydrography profiles from the UK Met Office EN4 dataset (Good et al. 2013). Although there exist methods to optimally estimate the hybrid weight  $\alpha$ , here we performed several experiments varying  $\alpha$  from 0 (fully ensemble covariances) to 1 (fully static covariances) with step 0.2, in order to identify the weight that leads to the best verification skill scores.

Results are reported in Fig. 8 in terms of RMSE decrease (percentage) with respect to the pure 3DVAR RMSE (i.e.  $\alpha = 1$ ), as a function of the forecast

range (from 1 to 10 days) and the hybrid weight, separately for the global ocean, and the Southern and Northern extra-Tropics and the Tropics. The statistics include both temperature and salinity observations (normalized by their nominal errors, in a way similar to Storto et al. 2018) in order to provide an overall accuracy diagnostics. The larger is the RMSE reduction, the more beneficial is the adoption of the hybrid formulation. In particular, the RMSE error metric includes simultaneously all available in-situ temperature and salinity observations, scaled by their respective nominal observational error, to provide an overall measure of accuracy. For the RMSE computation, note that differences between model equivalent and observations are computed before their ingestion in the analyses, i.e. they represent an independent validation dataset if one neglects the temporal correlations of the observation errors.

### **Fig. 8**

RMSE percentage reduction in hybrid ensemble-variational data assimilation with respect to the pure 3DVAR (hybrid weight equal to 1), as a function of hybrid weight (0 fully ensemble, 1 fully static) and forecast lead time in days. The black dots corresponds to the maximum RMSE reduction (North extra-Tropics from 60°N to 20°N; Tropics from 20°N to 20°S; South extra-Tropics from 20°S to 60°S). Note that the percentage reduction for the hybrid weight equal to 1 is zero by construction and is not shown



The largest reduction of error is found for  $\alpha$  equal to 0.8 in all basins, which provided reduction up to 10% globally and 20% in the Northern extra-Tropics. The recourse to hybrid covariances is beneficial to the reanalysis system for all intermediate weights that are generally clustered together. Differences of RMSE between them is not statistically significant in most cases. The use of pure ensemble covariances, due to the small ensemble size, leads to statistically non-robust covariances and detrimental impact. In particular, the Northern extra-Tropics exhibit large improvement for all intermediate ranges, due to the concurrent facts that static covariances may under-estimate the errors in the CMCC reanalysis system (Storto and Masina, 2016), the observational sampling is larger than in other bands and amplifies the impact of the background-error covariances, and GREP-ES is large here (Figure S1), indicating pronounced

variability that can be better captured with the introduction of flow-dependent covariances than stationary ones. Note that the highest reduction corresponding to the hybrid weight around 0.8 is in agreement with the results found by Storto et al. (2018) for the case of covariances coming from a single-system ensemble.

### 3.5. Measuring the evolution of reanalyses accuracy

A final application of GREP concerns the measure of the accuracy evolution with versioning (i.e. vintage), in terms of ensemble spread. As the ensemble spread measures the consistency between ocean reanalyses, one may expect that the consistency increases with a new vintage of ocean reanalyses, because of potentially improved use of ocean observations (better pre-processing procedures, data assimilation schemes and error characterization), improved model parameterizations, increased horizontal and/or vertical resolution and improved external forcing (e.g. atmospheric forcing, river runoff, etc.). The comparison of the spread between successive vintages of reanalyses provides a straightforward tool to qualify the advances of the ocean reanalysis community, which is of fundamental importance to show and acknowledge the gained increase of accuracy and confidence in reanalyses and foster further developments.

Since 2012, the Ocean Re-analyses Inter-comparison Projects (ORA-IP), undertaken by the GODAE Ocean View (GOV) jointly with CLIVAR Global Synthesis and Observations Panel (GSOP), has performed systematic comparisons of key ocean parameters from about 15 to 20 ocean-only and coupled reanalyses and observation-only products. The inter-comparison was generally performed on a  $1 \times 1$  degree regular grid. Preliminary outcomes of the inter-comparison are presented in the paper of Balmaseda et al. (2015), followed by a suite of studies in the Ocean Reanalysis Inter-comparison Special Issue (Balmaseda 2017), hosted by Climate Dynamics.

Steric sea level and its thermo- and halo- steric components were evaluated in ORA-IP from 16 ocean reanalyses and 4 objective analyses over the period 1993–2010 and compared to partly independent satellite estimates (altimetry minus gravimetry) over the period 2003–2010 by Storto et al. (2017). The four products of GREP had older predecessors included in the ORA-IP inter-comparison, thus offering a unique way to assess ocean reanalysis advances in the latest five years. Table 2 reports the predecessors of cglo, foam, glor and oras, along with the main improvements included in the present version since then. We refer to this older vintage of the reanalyses as OIS, which stands for ORA-IP Subsample, to distinguish them with the current GREP ensemble or the entire ORA-IP ensemble that is formed by many more members. It should be



noted that the OIP reanalyses are predecessors of the MyOcean reanalysis ensemble presented by Masina et al. (2017) and are not all at eddy-permitting resolution as in MyOcean. Indeed, two of the reanalyses (cglo and oras) have finer horizontal and vertical resolution, while all re-tuned the data assimilation system and the model parametrizations. Because of that, a gain in accuracy is expected from GREP, compared to OIS, reflecting the aforementioned advancements made. Note that because of the different resolution of some members, the comparison between OIS and GREP should be taken cautiously, as the processes resolved by the two ensemble may be significantly different and difficult to compare. Here, we aim to identify differences between them at basin scale and in a statistical sense, in order to assess the advances achieved with the latest vintage of reanalyses.

**Table 2**

For each GREP member, the table identifies the version included in the previous comparison (OIS, ORA-IP Sub-sampled), along with the main upgrades included in GREP

<b>GREP member</b>	<b>Version in OIS</b>	<b>Main upgrades in GREP w.r.t. OIS</b>
Cglo	C-GLORS05 v3	Increased horizontal and vertical resolution; Tuning of background-error covariances; Ocean model upgrade; Variational quality control of observations
Foam	Met Office	Ocean model upgrade (Megann et al. 2014); Sea ice model upgrade (Rae et al. 2015); Dual horizontal correlation scale introduce to sub-surface tracer assimilation (Mirouze et al., 2015). Updated ocean model to UK's Joint Ocean Modelling Programme (JOMP) Global Ocean configuration version 5.0 (GO5) (Megann et al. 2014)
Glor	Mercator Ocean	Revised surface mass budget and steric increments, update of: CORA in-situ profiles, Mean Dynamical Topography, initial conditions, runoff fluxes and large scale correction of atmospheric forcing
Oras	ORAS4	Increased horizontal and vertical resolution; wave effects; updated bias correction scheme; revised BGE covariance specification; up-to-date observation datasets and improved QC process

To summarize the spread differences, Fig. 9 shows the yearly mean timeseries of the spatially averaged ensemble spread for the total steric, thermosteric and halosteric components over different basins in the top 300 and top 700 m layers, during the period 1993–2010. The reanalysis upgrade yields a significant reduction of spread for all parameters in all basins, except for the thermosteric component in the southern extra-tropics, where differences are not significant.



However, in this region, the steric ensemble spread of GREP is smaller than OIS, due to the large decrease of spread for the halosteric component. Large reduction of spread is also visible in the tropical region timeseries. Around 1998, OIS presents a sudden spread increase that is completely absent in GREP, likely due to the strong El Nino occurrence in 1997–1998 not well captured by some of the OIS members. In the northern extra-tropics, the spread reduction is significant for the total steric and the two components, and both vintages increase the consistency with time.

**Fig. 9**

Timeseries of steric, thermosteric and halosteric sea level basin-averaged ensemble spread from OIS (solid lines) and GREP (dashed lines) in the top 300 (black lines) and top 700 (red lines) meters

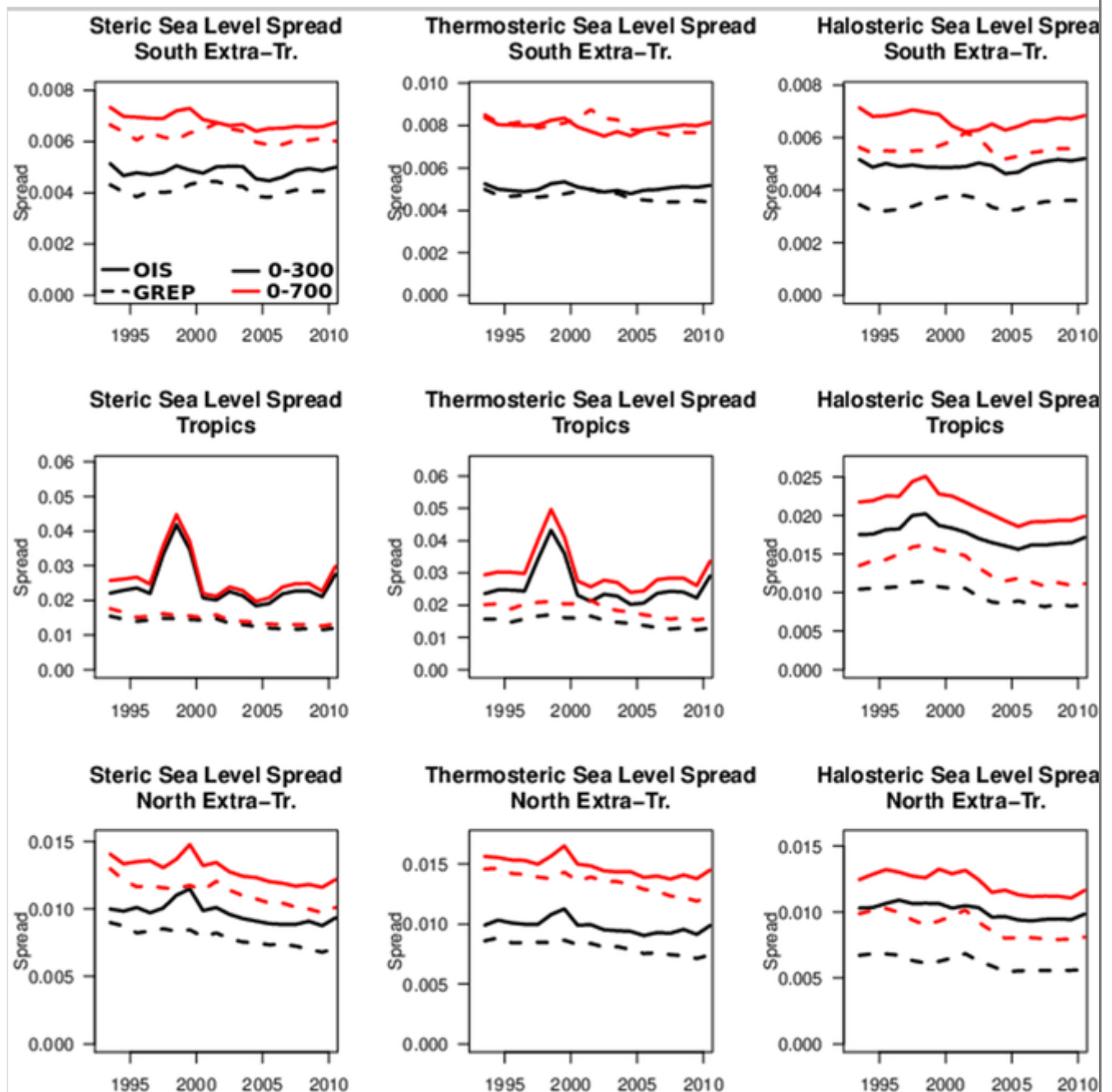


Figure S4 provides a map of the difference of the total steric spread between OIS and GREP, for two periods, 1993–2001 and 2002–2010. Results confirm the large reduction of uncertainty in the Tropical Pacific and Indian oceans and, to a lesser extent, in the tropical Atlantic and South Pacific oceans. Some spread increases are visible along the ACC and in the mid-latitude Northern hemisphere. Locally, the differences in spread may also be related to displacements of western boundary currents or fronts between the two ensembles, likely a consequence of the increased resolution in GREP, plus other differences. The improvements occur for both decades, although the first decade 1993–2001 exhibits larger spread difference than the second one, with larger areas in the Tropical Pacific and Atlantic Oceans with significant spread

differences, meaning that the improvements in the new vintage are greatest for the poorly observed decade.

As we are interested in assessing the consistency of the inter-annual signal between the two vintages, Table 3 shows the mean and standard deviation of the global steric sea level linear trend for the period 1993–2010. The reduction in the thermosteric trend uncertainty appears substantial. The uncertainty in the halosteric trend is also reduced, but to a lesser extent. The uncertainty of the total steric sea level spread is reduced accordingly, from  $\pm 0.60$  to  $\pm 0.44$  mm/year. The trend itself differs among the two vintages: considering a reference value of about  $1.1 \text{ mm/yr} \pm 0.4 \text{ mm/year}$  (from Hanna et al. 2013), it turns out that OIS underestimates the steric sea level due to a much too negative halosteric contribution, while GREP overestimates it because of the large thermosteric trend.

**Table 3**

2003–2010 linear trends of the OIS and GREP ensemble means (OIS-EM and GREP-EM) and their uncertainty (trend standard deviation) and correlation against ALT-GRV (see text for details) over the period 2003–2010, for the total, interannual and seasonal signal

Diagnostic	GREP Version	Steric Sea Level 0-700m		
	Component	Total	Thermo	Halo
Linear Trend 1993–2010 (mm/year)	OIS-EM	0.86+/-0.60	1.18+/-0.56	- 0.33+/-0.42
	GREP-EM	1.64+/-0.44	1.64+/-0.26	0.00+/-0.36
Correlation with ALT-GRV (2003–2010)	Signal	Total	Interannual	Seasonal
	OIS-EM	0.68	- 0.06	0.85
	GREP-EM	0.83	0.85	0.88
The seasonal signal is computed as total minus the linear trend, while the interannual signal as total minus mean seasonal cycle				

Finally, we validate the global steric sea level signal against partly independent estimates taken from altimetry minus gravimetric data, as in Storto et al. (2017). In particular, global sea level is provided by Nerem et al. (2010), while global eustatic (mass) component comes from GRACE data following Johnson and Chambers (2013). We calculated the correlation of the total, the seasonal and the inter-annual signal separately. Results (in Table 3) indicate the high skill of the

new vintage in capturing the inter-annual signal of the altimetry minus gravimetry dataset (0.85), as opposed to the non-significant correlation of OIS (−0.06). The resulting correlation of the full signal is significantly larger for GREP (0.83 versus 0.68), based on Steiger’s Z test with 99% confidence level (Steiger 1980), confirming that the new vintage represents a clear improvement with respect to the previous one.

## 4. Summary and discussion

In this study, we have investigated the feasibility of using a small ensemble of global ocean reanalyses for a range of applications such as uncertainty quantification at both global and regional scales. The four reanalyses considered here share the same ocean model and atmospheric forcing dataset, but employ different data assimilation systems and observational datasets, different air-sea flux formulation, initialization strategy, sea-ice model, model suite version and to some extent ocean model configuration parameters. All these factors contribute to the ensemble dispersion, which thus accounts for several sources of uncertainty, although some of them may be under-sampled.

The accuracy of the GREP ensemble mean was assessed through the use of the so-called CLASS4 metrics. This analysis showed that the quality of GREP is comparable to that of objective analyses, noting however that the validating observations are not independent and results must be taken cautiously. Based on this assessment approach, reanalyses tend to beat objective analyses for well-observed period (i.e. after the Argo float deployment). In particular, for temperature, GREP significantly outperforms the objective analyses below 700 m. Due to the cancellation of systematic errors, the ensemble mean always beats the individual members, as investigated in this study through the comparison of the RMSE of the ensemble mean versus the ensemble mean of the individual RMSEs from the members.

Our analysis also showed that the global ocean reanalyses gain accuracy through the new production: the comparison with a previous vintage for the steric sea level and its components indicates that a greater consistency has been achieved. GREP has a smaller steric sea level spread than the ORA-IP subsample of members in almost all vertical and horizontal regions investigated, meaning that the reanalyses acquired consistency in the new vintage, compared to the old one. Furthermore, globally averaged steric sea level exhibits a larger self-consistency (smaller spread) than OIS, and an increased consistency with a partly independent verifying dataset, deduced from altimetry minus gravimetry data. The latter comparison shows in particular that a large improvement was obtained

in the representation of the inter-annual signal. Although it is difficult to identify the main factor leading to consistency increase, the decrease in spread is likely the result of the convergence on ocean model configuration and forcing fluxes, as well as the introduction of bias correction schemes. This kind of exercise can be relevant to evaluate the reanalysis community advances and eventually to increase the confidence in reanalyses. Such an assessment is also crucial to provide large visibility to the reanalyses user community and motivate further developments in methods and practices for multi-decadal reanalyses.

The assessment of the ensemble system has been performed here with a caveat: independent observation datasets for skill score metrics and observation space-based evaluation of ensemble reliability are practically not available in the global ocean, given the fact that all available observational information is already ingested in the reanalysis systems. Nevertheless, the potential of the reanalysis ensemble emerged for a number of applications. In particular, the multi-system ensemble proves beneficial for:

- Complementing the metrics from high-resolution regional products with confidence levels;
- Identify the reanalysis error behavior with time through timeseries of ensemble spread;
- Monitoring the observing network through analysis of the ensemble spread tendency;
- Using the multi-model ensemble spread to form flow-dependent error covariances in hybrid ensemble-variational data assimilation systems;

Despite the small ensemble size, the spread is able to capture the temporal variability of the errors, quantified here by the RMSE of the ensemble mean. Proper re-scaling of the spread is needed because the use of only four members induces under-dispersive temperature spread. Salinity is however often over-dispersive, linked with the large uncertainty in freshwater content and poorly sampled salinity fields before the full Argo float deployment. We showed that correlation with the ensemble mean RMSE is high and always significant. As the GREP product is released together with the ensemble standard deviation, we have thus demonstrated its capability in representing the global ocean uncertainties during the reanalyzed period. Future studies should focus on understanding whether a larger ensemble size than GREP (e.g. such as that previously used within ORA-IP), where possibly additional ocean modeling

cores are used, leads to even improved RMSE-ES relationships and attenuated temperature under-dispersion. Moreover, routinely running and assessing assimilation-free control simulations, at present not available for all members, should better enlighten us about the relative impact of the different modeling configurations and associated spread with respect to the total ensemble spread of GREP.

The impact of the GREP horizontal resolution on the low-frequency variability and ensemble spread shall also be investigated in detail. In this case for example, it would be interesting to assess which would be the impact of using the GREP at its native resolution, instead of degrading it as it is currently done.

Furthermore, releasing information about data assimilation analysis increments, bias correction increments, sea surface nudging fluxes may help the reanalysis users and the climate community to improve the understanding of regional, budget and process-oriented assessments or comparisons that use ocean reanalyses. Including analysis increments as explicit term e.g. in heat budget analyses has already proven largely informative (e.g. Yang et al. 2016). Because compensation effects driven by data assimilation may occur differently between reanalyses (Storto et al. 2016c), it is desirable that reanalysis data catalogues are complemented with assimilation output fields, which will also shed further light on the relative performance of individual ensemble members.

Interestingly, global reanalyses show a large consistency also regionally, at least in the comparison performed over the Mediterranean Sea. This was also obtained for the other European seas included in the 2017 Ocean State Report (not shown but available in von Schuckman et al. 2018). Such a result fosters the adoption of super-ensemble products for ocean monitoring indexes, where regional high-resolution products are combined with observation-only and global products. This strategy may lead in particular to a better uncertainty quantification that spans different sources of uncertainty, provided that the addition of global products augments the regional ensemble size.

The novel idea to use a multi-system ensemble to derive flow-dependent background error-covariances for use in hybrid data assimilation proved very successful, suggesting that this approach may be further exploited in the future. It is worth noting that even at a relatively coarse temporal resolution as the monthly one used in our experiments, the ensemble anomalies are able to provide flow-dependent information related to observation sampling change with time and climate regimes that modulate the structure of the error covariances. Even without information about high-frequency error covariances, the adoption of GREP for flow-dependent covariances has a highly positive impact. Clearly,

such a scheme may be implemented easily for reanalyses, where the members of the ensemble are generally run retrospectively at once, likewise the reanalysis that exploits the multi-model ensemble-derived covariances. For real-time applications, this approach requires more technical efforts in order to estimate error covariances on-line from real-time products and adapt to any operational change in model, data assimilated, etc. Improvements up to 20% with respect to the use of static covariances suggest that the use of this approach might be also considered in designing future ensemble reanalysis products, i.e. next releases of GREP. Alternatively, simply using the multi-system ensemble to account for regime-dependent covariances (e.g. using ENSO index to embed covariance structures typical of positive, neutral or negative ENSO events) might be a doable way for ingesting to some extent flow-dependent error structures without requiring real-time computations.

The use of ensemble reanalyses for climate monitoring enables the use of physically consistent ocean state solutions with estimates of uncertainty through ensemble dispersion. This study suggests that routinely use of global ocean multi-system reanalysis ensemble as the GREP product proves promising for a number of applications, fostering the continuation of its design and use, and possibly the increase of the ensemble size.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Acknowledgements

Data from the reanalyses presented in this work are available from the Copernicus Marine Environment Monitoring Service (CMEMS, <http://marine.copernicus.eu/>). Part of this work was supported by the EOS COST Action (“Evaluation of Ocean Synthesis”, <http://eos-cost.eu/>) through its Short Term Scientific Missions program. The full C-GLORS dataset is available at <http://c-glors.cmcc.it>. This work has received funding from the Copernicus Marine Environment Monitoring Service (CMEMS). The EN4 subsurface ocean temperature and salinity data were quality-controlled and distributed by the U.K. Met Office. The authors declare no conflicts of interest. We are grateful to four anonymous reviewers for their help in improving the quality of the manuscript.

## 5. Appendix 1: Assessment of GREP

The GREP product has been extensively validated and the main outcomes are included in the CMEMS QUality Information Document (QUID), available at

<http://marine.copernicus.eu/documents/QUID/CMEMS-GLO-QUID-001-026.pdf>. Here, we focus only on skill score statistics of the monthly mean fields of temperature and salinity, which are the dataset used throughout this work. We base our assessment on the so-called GODAE OceanView CLASS4 metrics (Ryan et al. 2015), i.e. observation based statistics that estimate the reanalysis accuracy in observation space, inherited by the MERSEA project (Crosnier and Le Provost 2007) and then adopted by GODAE near real-time inter-comparison exercises (Hernandez et al. 2009). In accordance with disseminated data availability, we use monthly mean data for extracting the model equivalents from the GREP reanalyses. For sake of comparison, we introduce three observation-only products that do not make use of any dynamical model—also referred to as objective analyses, OA, with their ensemble mean (OA-EM)—and the World Ocean Atlas 2013 (WOA13) monthly climatology (Locarnini et al. 2013; Zweng et al. 2013) for the whole period 1955–2012. In particular, regarding the OAs, we calculated CLASS4 metrics for the Met Office EN4 objective analyses (Good et al. 2013), the Ifremer CORA objective analyses (Cabanes et al. 2013) and the CLS ARMOR objective analyses (Guinehut et al. 2012), within the GREP period. While there exists a large number of validation metrics (see e.g. Hernandez et al. 2015, for a detailed discussion), the goal of this exercise is to provide a basic assessment of the performances of the GREP-EM temperature and salinity through commonly adopted observation-based skill score metrics.

The Met Office EN4 (v4.2.0) observational data are used for this evaluation. EN4 includes profiles of MBTs and XBTs, Argo floats, CTDs, moorings and sea-mammal data. Only observations flagged as “good” are used. Note that these data are not independent strictly speaking, as they are assimilated by three out of four reanalyses and one objective analysis. Observational dataset may indeed differ notably because of different data sub-sampling, quality control procedure, and correction procedures (e.g. XBT fall rate corrections), especially before the Argo floats deployment. Therefore, the assessment presented hereafter has obvious limitations and serves only the purpose to verify how close to a reference dataset the reanalysis ensemble mean is, rather than quantifying the accuracy of either the reanalyses or objective analyses.

First, we validate the use of monthly means for the observation misfit statistics, which might seem inappropriate due to the higher than monthly temporal resolution of the model fields actually used as background in the data assimilation systems. To simplify, we focus on one product only, cglo, to verify the impact of sub-monthly variability on the skill scores during an observation-rich period. Figure S5 compares the cglo RMSE timeseries for two periods (1993–1998 and 2010–2015), of either daily or monthly mean fields of

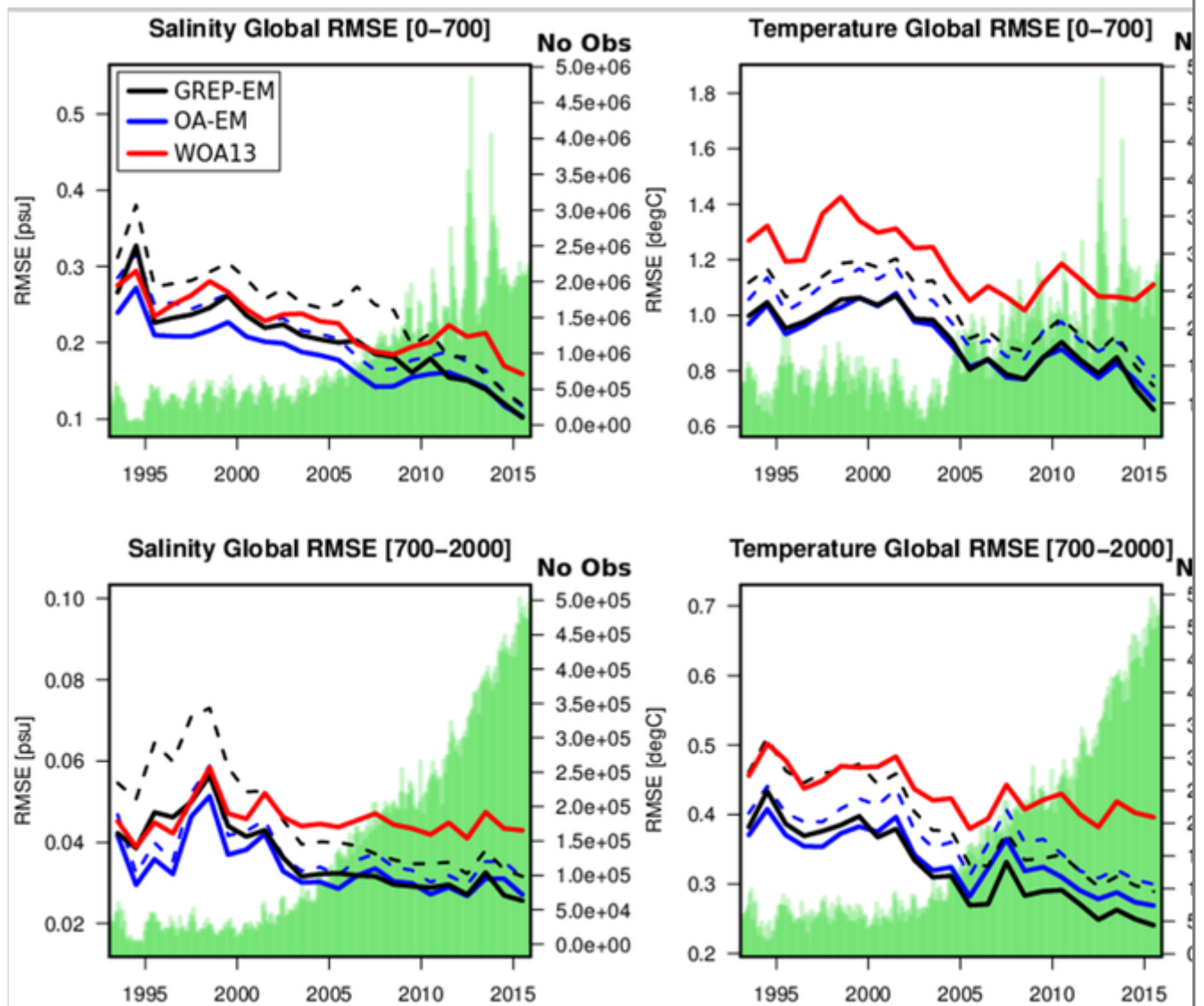


temperature and salinity. For both variables, differences are small, generally because the spatial representativeness error dominates the RMSE budget, rather than the temporal representativeness error. During the early period 1993–1998, differences are always less than 5% and 10% for salinity and temperature, respectively. During the recent period, the relative difference is of the order of 2% for salinity with occasional peaks in 2013–2014 up to 8%. Differences in temperature are slightly larger, with average value equals to 11% and peaking up to 22% in 2014. The behavior with time of the curves is almost identical, which indicates, together with the small differences, that the use of monthly means does not compromise the statistics. Note also that relative differences are greater during the latest years—in spite of smaller absolute values of RMSE, linked to the much denser observational sampling in both time and space.

In Fig. 10, we show yearly values (1993–2015) of the global RMSE statistics for temperature and salinity in the top 700 m and in the layer 700–2000 m, for the GREP-EM, OA-EM, and the WOA13 monthly climatology, along with monthly number of observations (green bars). Dashed lines represent for GREP and OA separately the ensemble average of the RMSE timeseries from individual members. The RMSE of the ensemble mean always outperforms the ensemble mean of the RMSEs, especially for the first decade and the deep layer. That confirms the effectiveness of the ensemble approach within data-sparse regions or periods.

### **Fig. 10**

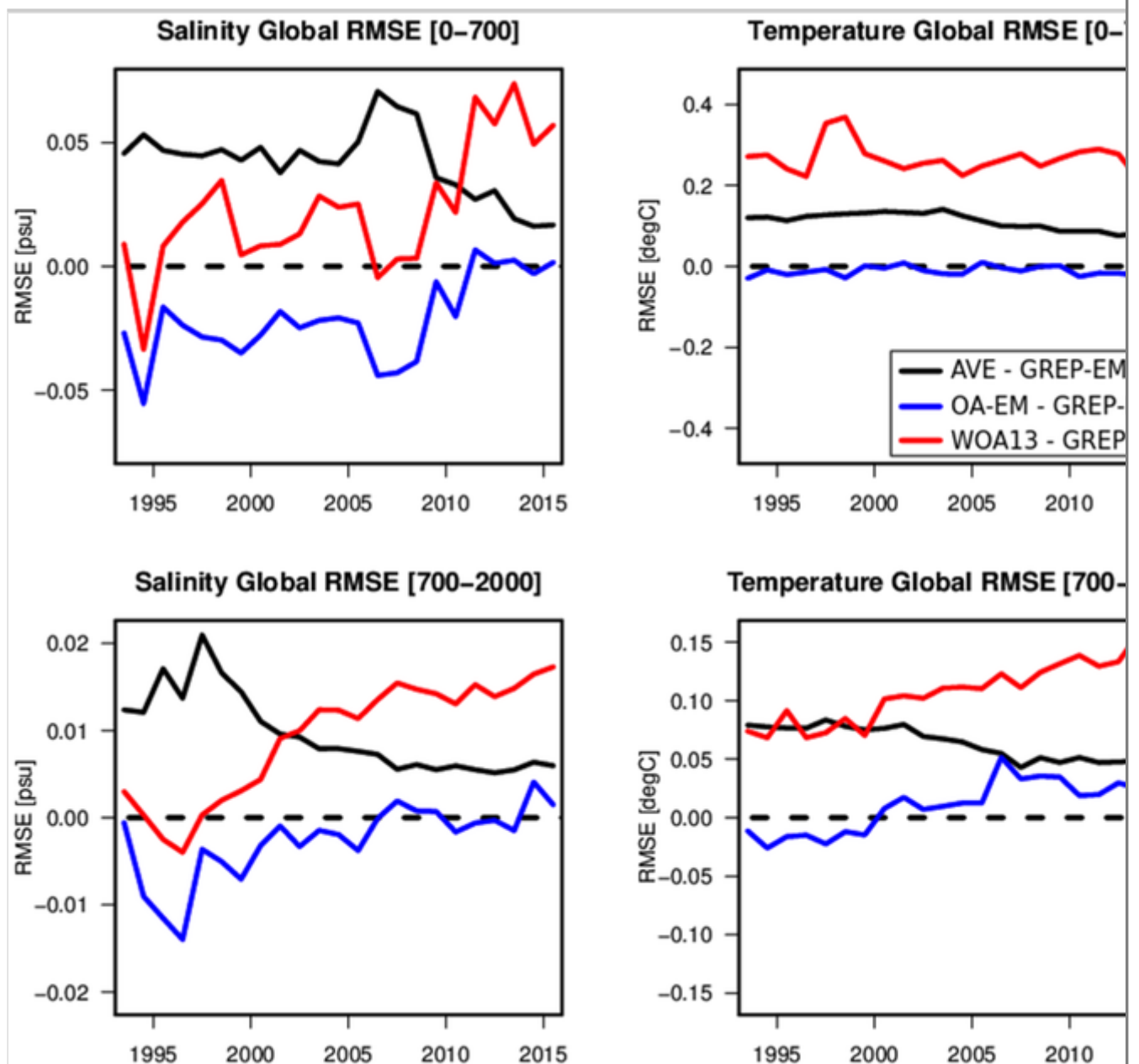
RMSE of in-situ observations against monthly mean fields from the GREP-EM (black), the objective analyses ensemble mean (OAM), which is the average of CORA, EN4 and ARMOR objective analyses, and the WOA13 monthly climatology, for temperature and salinity in the layer 0–700 m and 700–2000 m. The dashed lines report for both GREP and OAM the ensemble average of the RMSE timeseries from individual members (opposite to the RMSE of the ensemble mean). Values are yearly. The green bars (right side axis) show the total number of monthly observations. Reanalysis, objective analysis and climatology data are monthly means, while the validating in-situ profiles are extracted from the Met Office EN4 dataset



To clearly identify the skill scores behavior, Fig. 11 shows the differences between the average RMSE of the individual ORAs (AVE) and the RMSE of the ensemble mean (black), between RMSE of OA-EM and RMSE of GREP-EM (blue) and between RMSE of WOA13 and RMSE of GREP-EM (red). Positive values indicate that GREP-EM outperforms the other timeseries.

**Fig. 11**

Differences between the average of the RMSE timeseries from individual reanalyses and the RMSE of GREP (AVE—GREP-EM), between the RMSEs of OA-EM and GREP-EM (OA-EM—GREP-EM) and between the RMSEs of WOA13 and GREP-EM (WOA13—GREP-EM). Values are yearly. Reanalysis, objective analysis and climatology data are monthly means, while the validating in-situ profiles are extracted from the Met Office EN4 dataset



In the upper ocean, salinity skill scores are characterized by OA-EM errors smaller than GREP-EM up to around 2010. This suggests that the lack of in-situ observations is crucial for salinity RMSE. For temperature, WOA13 RMSE is significantly larger than GREP-EM and OA-EM, which present very similar behavior within the top 700 m layer. In the deeper ocean from 700 to 2000 m, OA-EM accuracy is higher than GREP-EM for the first simulated years, until 2000; then GREP-EM outperforms the ensemble of objective analyses. Figure S6 and S7 show similar analysis as Fig. 10 but for the three latitudinal bands Southern Extra-Tropics (60°S–20°S), Tropics (20°S–20°N) and Northern Extra-Tropics (20°N–60°N). The figures suggest that for temperature, especially in the Tropics, the reanalyses significantly outperform the OA-EM ensemble. Salinity skill scores of GREP-EM in the Southern Extra-Tropics, particularly in the upper

ocean, are worse than OA-EM and the WOA13 climatology during the first decade of the reanalysis.

Overall, within data sparse periods, OA-EM generally shows the smallest errors, likely due to the use of climatology background within the objective analyses, while GREP-EM behaves at least as good as the objective analyses after 2000, when the deployment of Argo floats started. For the deep ocean, the added value of Argo floats after 2006, able to constrain the heat content evolution, results in better scores for GREP-EM than OA-EM. Note however that different time periods imply different spatial sampling of the observations—notably before Argo deployment the RMSE is representative mostly of the skill in the Northern Hemisphere—and should be considered accordingly.

## 6. Electronic supplementary material

Below is the link to the electronic supplementary material.

Supplementary material 1 (DOCX 1409 KB)

## References

Balmaseda MA (2017) Editorial for ocean reanalysis intercomparison special issue. *Clim Dyn* 49:707. <https://doi.org/10.1007/s00382-017-3813-8>

Balmaseda MA, Trenberth KE, Källén E (2013) Distinctive climate signals in reanalysis of global ocean heat content. *Geophys Res Lett* 40:1754–1759. <https://doi.org/10.1002/grl.50382>

Balmaseda MA, Hernandez F, Storto A, Palmer MD, Alves O, Shi L, and Coauthors (2015) The ocean reanalyses intercomparison project (ORA-IP). *J Oper Oceanogr* 8(sup1):s80–s97. <https://doi.org/10.1080/1755876X.2015.1022329>

Blockley EW, Martin MJ, McLaren AJ, Ryan AG, Waters J, Lea DJ, Mirouze I, Peterson KA, Sellar A, Storkey D (2014) Recent development of the Met Office operational ocean forecasting system: an overview and assessment of the new Global FOAM forecasts. *Geosci. Model Dev* 7:2613–2638. <https://doi.org/10.5194/gmd-7-2613-2014>

Bouillon S, Morales Maqueda M, Legat V, Fichefet T (2009) An elastic–viscous–plastic sea ice model formulated on Arakawa B and C grids. *Ocean Model* 27:174–184

Brix H, Menemenlis D, Hill C, Dutkiewicz S, Jahn O, Wang D, Bowman K, Zhang H (2015) Using Green’s Functions to initialize and adjust a global, eddying ocean biogeochemistry general circulation model, *Ocean Model*. <https://doi.org/10.1016/j.ocemod.2015.07.008>

Burgers G, van Leeuwen PJ, Evensen G (1998) Analysis Scheme in the Ensemble Kalman Filter. *Mon Weather Rev* 126:1719–1724

Cabanes C, Grouazel A, von Schuckmann K, Hamon M, Turpin V, Coatanoan C, Paris F, Guinehut S, Boone C, Ferry N, de Boyer Montégut C, Carval T, Reverdin G, Pouliquen S, Traon L (2013) The CORA dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements. *Ocean Sci* 9:1–18. <https://doi.org/10.5194/os-9-1-2013>

Candille G, Talagrand O (2005) Evaluation of probabilistic prediction systems for a scalar variable. *QJR Meteorol Soc* 131:2131–2150. <https://doi.org/10.1256/qj.04.71>

Chandler RE (2013) Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Phil Trans R Soc A* 371:20120388. <https://doi.org/10.1098/rsta.2012.0388>

Chevallier M, Smith GC, Dupont F, Lemieux J-F, Forget G, Fujii Y, Hernandez F, Msadek R, Peterson KA, Storto A, Toyoda T, Valdivieso M, Vernieres G, Zuo H, Balmaseda M, Chang Y-S, Ferry N, Garric G, Haines K, Keeley S, Kovach RM, Kuragano T, Masina S, Tang Y, Tsujino H, Wang X (2017) Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project. *Clim Dyn* 49:1107–1136. <https://doi.org/10.1007/s00382-016-2985-y>

Crosnier L, Le Provost C (2007) Inter-comparing five forecast operational systems in the North Atlantic and Mediterranean basins: The MERSEA-strand1 methodology. *J Mar Syst* 65:354–375. <https://doi.org/10.1016/j.jmarsys.2005.01.003>

de Boissésou E, Balmaseda MA, Mayer M (2017) Ocean heat content variability in an ensemble of twentieth century ocean reanalyses. *Clim Dyn*.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thépaut J-N, Vitart F (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q J R Meteorol Soc* 137:553–597

Fortin V, Abaza M, Anctil F, Turcotte R (2014) Why should ensemble spread match the RMSE of the ensemble mean? *J Hydrometeorol* 15:1708–1713

Fujii Y, Cummings J, Xue Y, Schiller A, Lee T, Balmaseda MA, Rémy E, Masuda S, Brassington G, Alves O, Cornuelle B, Martin M, Oke P, Smith G, Yang X (2015) Evaluation of the tropical pacific observing system from the ocean data assimilation perspective. *QJR Meteorol Soc* 141:2481–2496.  
<https://doi.org/10.1002/qj.2579>

Garric G, Parent L, Greiner E, Drévillon M, Hamon M, Lellouche JM, Régnier C, Desportes C, Le Galloudec O, Bricaud C, Drillet Y, Hernandez F, Dubois C, Le Traon P-Y (2018) Performance and quality assessment of the global ocean eddy-permitting physical reanalysis GLORYS2V4. Operational Oceanography serving Sustainable Marine Development. Proceedings of the Eight EuroGOOS International Conference. 3–5 October 2017, Bergen, Norway. E. Buch, V. Fernandez, G. Nolan and D. Eparkhina (Eds.) EuroGOOS. Brussels, Belgium. 2018. ISBN:978-2-9601883-3-2. 516

Good SA, Martin MJ, Rayner NA (2013) EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J Geoph Res* 118:6704–6716.  
<https://doi.org/10.1002/2013JC009067>

Griffies S, Greatbatch R (2012) Physical processes that impact the evolution of global mean sea level in ocean climate models. *Ocean Model* 51:37–72

Guinehut S, Dhomp A-L, Larnicol G (2012) High resolution 3-D temperature and salinity fields derived from in situ and satellite observations. *Ocean Sci* 8:845–857. <https://doi.org/10.5194/os-8-845-2012>

Hamill TM, C. Snyder (2000) A hybrid ensemble kalman filter—3D variational analysis scheme. *Mon Wea Rev* 128:2905–2919, [https://doi.org/10.1175/1520-0493\(2000\)128%3C2905:AHEKFV%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128%3C2905:AHEKFV%3E2.0.CO;2)

Hanna E et al (2013) Ice-sheet mass balance and climate change. *Nature* 498:51–59

Hart RE, Grumm RH (2001) Using normalized climatological anomalies to rank synoptic-scale events objectively. *Mon Weather Rev* 129(9):2426–2442

Hernandez F, Bertino L, Brassington G, Chassignet E, Cummings J, Davidson F, Drévillon M, Garric G, Kamachi M, Lellouche J-M, Mahdon R, Martin MJ, Ratsimandresy A, Regnier C (2009) Validation and intercomparison studies within GODAE. *Oceanography* 22(3):128–143. <https://doi.org/10.5670/oceanog.2009.71>

Hernandez F, Blockley E, Brassington GB, Davidson F, Divakaran P, Drévillon M et al. (2015) Recent progress in performance evaluations and near real-time assessment of operational ocean products, *J Oper Oceanogr* 8(2):s221–s238, <https://doi.org/10.1080/1755876X.2015.1050282>

Hu Z-Z, Kumar A (2015) Influence of availability of TAO data on NCEP ocean data assimilation systems along the equatorial Pacific. *J Geophys Res Oceans* 120:5534–5544. <https://doi.org/10.1002/2015JC010913>

Hunke EC, Lipscomb WH, Turner AK, Jeffery N, Elliott SM (2013) CICE: the Los Alamos Sea Ice Model, Documentation and Software, Version 5.0. Los Alamos National Laboratory Tech. Rep. LA-CC-06-012. <http://oceans11.lanl.gov/trac/CICE>

Jackson L, Peterson KA, Roberts C, Wood R (2016) Recent slowing of Atlantic overturning circulation as a recovery from earlier strengthening. *Nat Geosci* 9:518–522. <https://doi.org/10.1038/ngeo2715>

Johnson G, Chambers D (2013) Ocean bottom pressure seasonal cycles and decadal trends from GRACE Release-05: ocean circulation implications. *J Geophys Res* 118:4228–4240

Josey SA, Yu L, Gulev S, Jin X, Tilinina N, Barnier B, Brodeau L (2014) Unexpected impacts of the Tropical Pacific array on reanalysis surface

meteorology and heat fluxes. *Geophys Res Lett* 41:6213–6220.  
<https://doi.org/10.1002/2014GL061302>

Karspeck AR, Stammer D, Köhl A, Danabasoglu G, Balmaseda M, Smith DM, Fujii Y, Zhang S, Giese B, Tsujino H, Rosati A (2017) Comparison of the Atlantic meridional overturning circulation between 1960 and 2007 in six ocean reanalysis products. *Clim Dyn* 49:957–982.  
<https://doi.org/10.1007/s00382-015-278>

Krishnamurti TN, Kishtawal CM, Shin DW, Williford CE (2000) Multi-model superensemble forecasts for weather and seasonal climate. *J Clim* 13:4196–4216. [https://doi.org/10.1175/1520-0442\(2000\)013%3C4196:MEFFWA%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013%3C4196:MEFFWA%3E2.0.CO;2)

Laloyaux P, Balmaseda M, Dee D, Mogensen K, Janssen P (2015) A coupled data assimilation system for climate reanalyses. *Q J R Meteorol Soc* 142:65–78

Large WG, Yeager SG (2004) Diurnal to decadal global forcing for ocean and sea-ice models: the data sets and flux climatologies, NCAR Technical report NCAR/TN-460, Boulder, Colorado, USA

Le Traon PY, Nadal F, Ducet N (1998) An improved mapping method of multisatellite altimeter data. *J Atmos Oceanic Technol* 15:522–534,  
[https://doi.org/10.1175/1520-0426\(1998\)015%3C0522:AIMMOM%3E2.0.CO;2](https://doi.org/10.1175/1520-0426(1998)015%3C0522:AIMMOM%3E2.0.CO;2)

Le Traon PY et al. (2017) The copernicus marine environmental monitoring service: main scientific achievements and future prospects. *Mercator Ocean Journal* (Special Issue CMEMS), pp 2–101. <http://www.mercator-ocean.fr/en/science-publications/mercator-ocean-journal/mercator-ocean-journal-56-special-issue-cmems>

Lea DJ, Drecourt J-P, Haines K, Martin MJ (2008) Ocean altimeter assimilation with observational- and model-bias correction. *QJR Meteorol Soc* 134:1761–1774. <https://doi.org/10.1002/qj.320>

Locarnini RA, Mishonov AV, Antonov JJ, Boyer TP, Garcia HE, Baranova OK, Zweng MM, Paver CR, Reagan JR, Johnson DR, Hamilton M, Seidov D (2013) *World Ocean Atlas 2013, Volume 1: Temperature*. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 73, 40 pp



Loeb NG et al (2009) Towards optimal closure of the earth's top-of-atmosphere radiation budget. *J Clim* 22:748–766

Lorenc AC (1986) Analysis methods for numerical weather prediction. *QJR Meteorol Soc* 112:1177–1194. <https://doi.org/10.1002/qj.49711247414>

Lorenc AC (2003) The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *QJR Meteorol Soc* 129:3183–3203. <https://doi.org/10.1256/qj.02.132>

Madec G, Imbard M (1996) A global ocean mesh to overcome the North Pole singularity. *Clim Dyn* 12:381–388. <https://doi.org/10.1007/BF00211684>

Madec G, the NEMO team (2012) “NEMO ocean engine”. Note du Pole de modélisation de l’Institut Pierre-Simon Laplace, France, No 27 ISSN No 1288–1619

Marbà N, Jordà G, Agustí S, Girard C, Duarte CM (2015) Footprints of climate change on Mediterranean Sea biota. *Front Mar Sci* 2:56. <https://doi.org/10.3389/fmars.2015.00056>

Masina S, Storto A (2017) Reconstructing the recent past ocean variability: status and perspective. *J Mar Res* 75:727–764. <https://doi.org/10.1357/002224017823523973>

Masina S, Storto A, Ferry N, Valdivieso M, Haines K, Balmaseda M, Zuo H et al (2017) An ensemble of eddy-permitting global ocean reanalyses from the MyOcean project. *Clim Dyn* 49:813–841. <https://doi.org/10.1007/s00382-015-2728-5>

Masson D, Knutti R (2011) Climate model genealogy. *Geophys Res Lett* 38:L08703. <https://doi.org/10.1029/2011GL046864>

Mayer M, Haimberger L, Pietschnig M, Storto A (2016), Facets of Arctic energy accumulation based on observations and reanalyses 2000–2015, *Geophys Res Lett*. <https://doi.org/10.1002/2016GL070557>

Meehl GA, Boer GJ, Covey C, Latif M, Stouffer RJ (1997) Intercomparison makes for a better climate model. *Eos Trans AGU* 78(41):445–451. <https://doi.org/10.1029/97EO00276>

Megann A, Storkey D, Aksenov Y, Alderson S, Calvert D, Graham T, Hyder P, Siddorn J, Sinha B (2014) GO5.0: the joint NERC–Met Office NEMO global ocean model for use in coupled and forced applications. *Geosci Model Dev* 7:1069–1092. <https://doi.org/10.5194/gmd-7-1069-2014>

Mirouze I, Blockley EW, Lea DJ, Martin MJ, Bell MJ (2016) A multiple length scale correlation operator for ocean data assimilation. *Tellus A: Dyn Meteorol Oceanogr* 68:1. <https://doi.org/10.3402/tellusa.v68.29744>

Nerem R, Chambers D, Choe C, Mitchum G (2010) estimating mean sea level change from the TOPEX and jason altimeter missions. *Mar Geodesy* 33:435–446

Oddo P, Storto A, Dobricic S, Russo A, Lewis C, Onken R, Coelho E (2016) A hybrid variational-ensemble data assimilation scheme with systematic error correction for limited-area ocean models. *Ocean Sci* 12:1137–1153. <https://doi.org/10.5194/os-12-1137-2016>

Ota Y, Derber JC, Kalnay E, Miyoshi T (2013) Ensemble-based observation impact estimates using the NCEP GFS. *Tellus A* 65, <https://doi.org/10.3402/tellusa.v65i0.20038>

Palmer MD, Roberts CD, Balmaseda M, Chang Y-S, Chepurin G, Ferry N, Fujii Y, Good SA, Guinehut S, Haines K, Hernandez F, Köhl A, Lee T, Martin MJ, Masina S, Masuda S, Peterson KA, Storto A, Toyoda T, Valdivieso M, Vernieres G, Wang O, Xue Y (2017) Ocean heat content variability and change in an ensemble of ocean reanalyses. *Clim Dyn* 49:909–930. <https://doi.org/10.1007/s00382-015-2801-0>

Penduff T, Juza M, Brodeau L, Smith GC, Barnier B, Molines J-M, Treguier A-M, Madec G (2010) Impact of global ocean model resolution on sea-level variability with emphasis on interannual time scales. *Ocean Sci* 6:269–284. <https://doi.org/10.5194/os-6-269-2010>

Penny SG, Behringer DW, Carton JA, Kalnay E (2015) A Hybrid Global Ocean Data Assimilation System at NCEP. *Mon Wea Rev* 143:4660–4677. <https://doi.org/10.1175/MWR-D-14-00376.1>

Potter GL, Carriere L, Hertz J, Bosilovich M, Duffy D, Lee T, Williams DN (2018) Enabling reanalysis research using the collaborative reanalysis

technical environment (CREATE). Bull Amer Meteor Soc. In press,  
<https://doi.org/10.1175/BAMS-D-17-0174.1>

Rae JGL, Hewitt HT, Keen AB, Ridley JK, West AE, Harris CM, Hunke EC, Walters DN (2015) Development of the global sea ice 6.0 CICE configuration for the Met Office Global Coupled model. *Geosci Model Dev* 8:2221–2230.  
<https://doi.org/10.5194/gmd-8-2221-2015>

Rainwater S, Hunt BR (2013) Ensemble data assimilation with an adjusted forecast spread. *Tellus A: Dyn Meteorol Oceanogr* 65(1):19929.  
<https://doi.org/10.3402/tellusa.v65i0.19929>

Raynaud L, Berre L, Desroziers G (2008) Spatial averaging of ensemble-based background-error variances. *Q J R Meteorol Soc* 134:1003–1014

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J Geophys Res* 108:4407. <https://doi.org/10.1029/2002JD002670>, D14

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. *J Clim* 20:5473–5496. <https://doi.org/10.1175/2007JCLI1824.1>

Riser SC et al. (2016) Fifteen years of ocean observations with the global Argo array. *Nat Clim Change* 6:145–150.  
<https://doi.org/10.1038/nclimate2872>

Ryan AG, Regnier C, Divakaran P, Spindler T, Mehra A, Smith GC, Davidson F, Hernandez F, Maksymczuk J, Liu Y (2015) GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison. *J Oper Oceanogr* 8(sup1):s98–s111. <https://doi.org/10.1080/1755876X.2015.1022330>

Schroeder K, Chiggiato J, Josey SA, Borghini M, Aracri S, Sparnocchia S (2017) Rapid response to climate change in a marginal sea. *Sci Rep* 7:4065.  
<https://doi.org/10.1038/s41598-017-04455-5>

Shi L, Alves O, Wedd R, Balmaseda MA, Chang Y, Chepurin G, Ferry N, Fujii Y, Gaillard F, Good SA, Guinehut S, Haines K, Hernandez F, Lee T, Palmer M, Peterson KA, Masuda S, Storto A, Toyoda T, Valdivieso M, Vernieres G, Wang X, Yin Y (2017) An assessment of upper ocean salinity

content from the Ocean Reanalyses Inter-comparison Project (ORA-IP). *Clim Dyn* 49:1009–1029. <https://doi.org/10.1007/s00382-015-2868-7>

Simoncelli S, Fratianni C, Pinardi N, Grandi A, Drudi M, Oddo P, Dobricic S (2014) Mediterranean Sea physical reanalysis (MEDREA 1987–2015). Copernicus Monitoring Environment Marine Service (CMEMS). [https://doi.org/10.25423/medsea\\_reanalysis\\_phys\\_006\\_004](https://doi.org/10.25423/medsea_reanalysis_phys_006_004)

Simoncelli S, Masina S, Axell L, Liu Y, Salon S, Cossarini G, Bertino L, Xie J, Samuelsen A, Levier B et al (2016) MyOcean Regional Reanalyses: Overview of Reanalyses Systems and Main Results. *Mercator Ocean Journal* n.54: Special Issue on Main Outcomes of the MyOcean2 and MyOcean Follow-on projects. <https://www.mercator-ocean.fr/wp-content/uploads/2016/03/JournalMO-54.pdf>

Stammer D, Balmaseda M, Heimbach P, Köhl A, Weaver A (2016) Ocean data assimilation in support of climate applications: status and perspectives. *Ann Rev Mar Sci* 8:491–518. <https://doi.org/10.1146/annurev-marine-122414-034113>

Steiger J (1980) Tests for comparing elements of a correlation matrix. *Psychological Bull* 87:245–251

Storto A (2016) Variational quality control of hydrographic profile data with non-Gaussian errors for global ocean variational data assimilation systems. *Ocean Model* 104:2016, 226–241. <https://doi.org/10.1016/j.ocemod.2016.06.011>

Storto A, Masina S (2016a) C-GLORSv5: an improved multipurpose global ocean eddy-permitting physical reanalysis. *Earth Syst Sci Data* 8:679–696. <https://doi.org/10.5194/essd-8-679-2016>

Storto A, Masina S (2017) Objectively estimating the temporal evolution of accuracy and skill in a global ocean reanalysis. *Met Apps* 24:101–113. <https://doi.org/10.1002/met.1609>

Storto A, Dobricic S, Masina S, Di Pietro P (2011) Assimilating along-track altimetric observations through local hydrostatic adjustments in a global ocean reanalysis system. *Mon Weather Rev* 139:738–754. <https://doi.org/10.1175/2010MWR3350.1>

- Storto A, Masina S, Dobricic S (2013) Ensemble spread-based assessment of observation impact: application to a global ocean analysis system. *QJR Meteorol Soc* 139:1842–1862. <https://doi.org/10.1002/qj.2071>
- Storto A, Masina S, Dobricic S (2014) Estimation and impact of non-uniform horizontal correlation length-scales for global ocean physical analyses. *J Atmos Ocean Tech* 31:2330–2349. <https://doi.org/10.1175/JTECH-D-14-00042.1>
- Storto A, Masina S, Navarra A (2016b) Evaluation of the CMCC eddy-permitting global ocean physical reanalysis system (C-GLORS, 1982–2012) and its assimilation components. *Q J Roy Meteorol Soc* 142:738–758. <https://doi.org/10.1002/qj.2673>
- Storto A, Yang C, Masina S (2016c) Sensitivity of global ocean heat content from reanalyses to the atmospheric reanalysis forcing: A comparative study. *Geophys Res Lett* 43:5261–5270. <https://doi.org/10.1002/2016GL068605>
- Storto A, Masina S, Balmaseda M, Guinehut S, Xue Y, Szekely T (2017) Steric sea level variability (1993–2010) in an ensemble of ocean reanalyses and objective analyses. *Clim Dyn* 49(3):709–729. <https://doi.org/10.1007/s00382-015-2554-9>
- Storto A, Oddo P, Cipollone A, Mirouze I, Lemieux-Dudon B (2018) Extending an oceanographic variational scheme to allow for affordable hybrid and four-dimensional data assimilation. *Ocean Modeling Ocean Modelling* 128:67–86,. <https://doi.org/10.1016/j.ocemod.2018.06.005>
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc* 93:485–498
- Toyoda T, Fujii Y, Kuragano T, Kamachi M, Ishikawa Y, Masuda S, Sato K, Awaji T, Hernandez F, Ferry N, Guinehut S, Martin M, Peterson KA, Good SA, Valdivieso M, Haines K, Storto A, Masina S, Köhl A, Zuo H, Balmaseda M, Yin Y, Li Shi O, Alves G, Smith Y-S, Chang G, Vernieres X, Wang G, Forget P, Heimbach O, Wang I, Fukumori T, Lee (2017) Intercomparison and validation of the mixed layer depth fields of global ocean syntheses. *Clim Dyn* 49:753–773. <https://doi.org/10.1007/s00382-015-2637-7>
- Trenberth KE, Fasullo JT, von Schuckmann K, Cheng L (2016) Insights into earth's energy imbalance from multiple sources. *J Clim* 29:7495–7505.

<https://doi.org/10.1175/JCLI-D-16-0339.1>

Valdivieso M,K, Haines M, Balmaseda Y-S, Chang M, Drevillon N, Ferry Y, Fujii A, Köhl,A. Storto,T, Toyoda Xang,J, Waters Y, Xue Y, Yin B, Barnier F, Hernandez A, Kumar T, Lee S, Masina K (2017) An assessment of air–sea heat fluxes from ocean and coupled reanalyses. *Clim Dyn* 49:983–1008. <https://doi.org/10.1007/s00382-015-2843-3>

Vancoppenolle M, Fichefet T, Goosse H, Bouillon S, Madec G, Morales Maqueda MA (2009) Simulating the mass balance and salinity of Arctic and Antarctic sea ice. 1. Model description and validation. *Ocean Model* 27(1–2):33–53

von Schuckmann K, Le Traon P-Y, Alvarez-Fanjul E, Axell L, Balmaseda M, Breivik L-A et al. (2017) The copernicus marine environment monitoring service ocean state report, *J Oper Oceanogr*, 9, Issue sup2: The Copernicus Marine Environment Monitoring Service Ocean, s235-s320

von Schuckmann K et al. (2018) Copernicus marine service ocean state report. *J Oper Oceanogr* 11:s1–s142. <https://doi.org/10.1080/1755876X.2018.1489208>

Wang X, Snyder C, Hamill TM (2007) On the Theoretical Equivalence of Differently Proposed Ensemble–3DVAR Hybrid Analysis Schemes. *Mon Wea Rev* 135:222–227. <https://doi.org/10.1175/MWR3282.1>

Wunsch C (2016) Global ocean integrals and means, with trend implications. *Ann Rev Mar Sci* 8:1–33. <https://doi.org/10.1146/annurev-marine-122414-034040>

Xue Y, Huang B, Hu Z-Z, Kumar A, Wen C, Behringer D, Nadiga S (2011) An assessment of oceanic variability in the NCEP climate forecast system reanalysis. *Climate dynamics* 37:2511–2539

Xue Y, Wen C, Kumar A et al (2017) A real-time ocean reanalyses intercomparison project in the context of tropical pacific observing system and ENSO monitoring. *Clim Dyn* 49:3647. <https://doi.org/10.1007/s00382-017-3535-y>

Yamaguchi M, Lang STK, Leutbecher M, Rodwell MJ, Radnoti G, Bormann N (2016) Observation-based evaluation of ensemble reliability. *QJR Meteorol*

Soc 142:506–514. <https://doi.org/10.1002/qj.2675>

Yang C, Masina S, Bellucci A, Storto A (2016) The rapid warming of the North Atlantic Ocean in the Mid-1990s in an eddy-permitting ocean reanalysis (1982–2013). *J Clim* 29:5417–5430. <https://doi.org/10.1175/JCLI-D-15-0438.1>

Yang C, Masina S, Storto A (2017) Historical ocean reanalyses (1900–2010) using different data assimilation strategies. *QJR Meteorol Soc* 143:479–493. <https://doi.org/10.1002/qj.2936>

Zuo H, Balmaseda MA, de Boisseson E, Hirahara S, Chrut M, de Rosnay P (2017a) A generic ensemble generation scheme for data assimilation and ocean analysis. ECMWF Tech Memo 795, 46 pp, European Centre for Medium-Range Weather Forecasts, Reading, UK.  
<https://www.ecmwf.int/en/elibrary/technical-memoranda>

Zuo H, Balmaseda MA, Mogensen K (2017b) The new eddy-permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals. *Clim Dyn* 49:791. <https://doi.org/10.1007/s00382-015-2675-1>

Zuo H, Balmaseda MA, Tietsche S, Mayer M, Robert CD (2018), Evaluation of the ECMWF ocean reanalysis system ORAS5, *Q.J.R.M.S.*, in preparation

Zweng MM, Reagan JR, Antonov JJ, Locarnini RA, Mishonov AV, Boyer TP, Garcia HE, Baranova OK, Johnson DR, D.Seidov MM (2013) World Ocean Atlas 2013, Volume 2: Salinity. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 74, 39 pp