Climate Dynamics
https://doi.org/10.1007/s00382-021-06101-3

Seasonal forecast skill of upper-ocean heat content in coupled high-resolution systems

Ronan McAdam[1] · Simona Masina[1] · Magdalena Balmaseda[2] · Silvio Gualdi[3] · Retish Senan[2] · Michael Mayer[2,4]

Received: 5 May 2021 / Accepted: 7 December 2021

Abstract

Ocean heat content (OHC) anomalies typically persist for several months, making this variable a vital component of seasonal predictability in both the ocean and the atmosphere. However, the ability of seasonal forecasting systems to predict OHC remains largely untested. Here, we present a global assessment of OHC predictability in two state-of-the-art and fully-coupled seasonal forecasting systems. Overall, we find that dynamical systems make skilful seasonal predictions of OHC in the upper 300 m across a range of forecast start times, seasons and dynamical environments. Predictions of OHC are typically as skilful as predictions of sea surface temperature (SST), providing further proof that accurate representation of subsurface heat contributes to accurate surface predictions. We also compare dynamical systems to a simple anomaly persistence model to identify where dynamical systems provide added value over cheaper forecasts; this largely occurs in the equatorial regions and the tropics, and to a greater extent in the latter part of the forecast period. Regions where system performance is inadequate include the sub-polar regions and areas dominated by sharp fronts, which should be the focus of future improvements of climate forecasting systems.

Keywords Seasonal forecasts · Ocean heat content · Essential ocean variables, forecast skill · Coupled forecasting systems · Ocean reanalysis

1 Introduction

State-of-the-art seasonal forecast systems include a coupled ocean–atmosphere (Stockdale et al. 1998, Baehr et al. 2015; Batté et al. 2019; Johnson et al. 2019; MacLachlan et al. 2015; Saha et al. 2014; Sanna et al. 2017, Takaya et al. 2018) because the main source of seasonal predictability in many climate variables, on a global scale, is the quasi-periodic ocean–atmosphere interaction known as the El Niño Southern Oscillation (ENSO). ENSO alters the atmospheric circulation across the entire tropical Pacific and, as a result, causes knock-on effects (teleconnections) which change seasonal climates across the world. The thermocline acts as a "memory bank" by providing long-term heat storage for the region (Neelin et al. 1998). The cycle of ENSO events, and therefore the teleconnections, are strongly influenced by the subsurface ocean heat content (OHC) in the tropical Pacific (Doblas-Reyes et al. 2013).

Because of this crucial role in global predictability, the initialization of the subsurface thermal structure is key for successful seasonal predictions. Initialising systems with accurate data about slowly varying regions in the subsurface improves sea surface temperature (SST) predictions (Alves et al.2004; Balmaseda & Anderson 2009; Alessandri et al. 2010; McPhaden et al. 2020). Traditionally, the focus was on the initialization of the subsurface thermal structure in the tropics (Balmaseda 2017), but more recently it has emerged that the extratropics can also have an impact on seasonal forecasts. In particular, Tietsche et al. (2020) describes the influence of the Atlantic Meridional Overturning Circulation (AMOC) on seasonal variations in North Atlantic SST.

✉ Ronan McAdam
ronan.mcadam@cmcc.it

[1] Ocean Modeling and Data Assimilation Division, Centro Euro-Mediterraneo Sui Cambiamenti Climatici, Bologna, Italy

[2] European Centre for Medium-Range Weather Forecasts, Reading, UK

[3] Climate Simulations and Predictions Division, Centro Euro-Mediterraneo Sui Cambiamenti Climatici, Bologna, Italy

[4] Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

Published online: 31 January 2022

OHC can also be used as a predictor of weather phenomena at seasonal time scales using statistical techniques. For example, in the subtropical North Atlantic, there is a correlation, with a lag of several months, between temperatures in the upper 150 m and hurricane activity (Scoccimarro et al. 2018). The correlation between subsurface temperature and extreme weather activity is also used in sub-seasonal operational systems (Mainelli et al. 2008). Hobday et al. (2011) quantified the anomaly-detection skill of the upper 100 m OHC in the Tasman Sea, in the global ocean–atmosphere forecast system and up to four months lead time, due to the importance of this depth for the distribution of tuna and the spatial management of fisheries.

There is no extensive validation of ocean heat content in seasonal forecasting systems, despite its important role in seasonal predictability and the potential applications. Existing validation in the literature is often regional and does not include a range of measures (e.g. Hobday et al. 2011; Tietsche et al. 2020). Seasonal forecasting groups validate their systems often, for example at the launch of each new version or when new validation data become available. This essential work is typically performed for other variables such as SST, precipitation, sea-ice cover and 2 m air temperature (e.g. Baehr et al. 2015; Johnson et al. 2019). As in the validation work, the uses of seasonal forecasts have typically been focused on atmospheric, land-based or surface variables (e.g. for agriculture and energy generation) (Bruno Soares & Dessai 2015). Ocean heat content, or even temperature at a particular depth below the surface, has not yet received the same level of validation or appreciation from potential uses.

Although there is some literature on the use of predictions of heat content (see references above), perhaps the potential applications are not yet widely appreciated. An exciting and urgent task for seasonal forecasting is the prediction of extreme heating events, which either occur at depth or are driven by subsurface heat anomalies. The average duration of such events is increasing globally and is crossing into the timescales of seasonal forecasts (Oliver et al. 2018; Darmaraki et al. 2019). Fortunately, events driven by subsurface warming are expected to be more predictable than those primarily driven by relatively abrupt atmospheric disturbances (Behrens et al. 2019; Holbrook et al. 2020). The early prediction of subsurface heating could be of great economic and practical benefit to several industries such as aquaculture and fishing, and could aid marine conservation efforts against mass-mortality events (e.g. Caputi et al. 2016).

Meanwhile, there is good reason to expect predictability to increase with depth; OHC anomalies are typically more persistent than SST and less responsive to daily/weekly atmospheric disturbances, thus making predictions on seasonal timescales easier. However, throughout the ocean there are many regions where SST skill is inadequate and is subject to seasonal and inter-annual changes (e.g. Weisheimer et al. 2020). Skilful surface prediction is therefore not necessarily a sign of skilful subsurface prediction, and vice versa.

A more practical reason for a lack of validation efforts could be a lack of OHC validation datasets, yet this is not the case. There exists a multitude of 3D ocean analysis and reanalysis which are already used for estimating ocean variability and for inter-annual to decadal subsurface forecast initialisation (e.g. Good et al. 2013; Balmaseda et al. 2015; Masina et al. 2017). Indeed, many studies report high levels of predictive skill for the OHC (or subsurface ocean temperatures) at these longer timescales (e.g. Robson 1990; Msadek et al. 2014; Yeager et al. 2018; Bilbao et al. 2021). Whatever the reason, the capability of current systems to forecast upper ocean heat content should not be taken for granted.

This study compares the OHC, from 0 to 300 m, against an ensemble of ocean reanalysis products. This effort is the first to cover the global ocean (except the polar seas) and all seasons. The upper 300 m is chosen because it encompasses many diverse phenomena across the ocean which are either relevant for predictability or applications. In the tropics, the upper ocean heat content is an important element of the ENSO energy cycle and recharge-discharge mechanism (e.g. Mayer et al. 2018 and references within). In the North Atlantic, Häkkinen et al. (2013) found that decadal variability of SSH is partially driven by heat transport in the upper 700 m; more generally, increasing heat content can lead to thermosteric sea level rise. Aerosol cooling from volcanic eruptions is known to cause an abrupt cooling and a later rebound in the upper 700 m temperature (Carton & Santorelli 2008), the latter of which could be predicted once aerosol data is assimilated. Tropical cyclones can induce large heat transport anomalies in the upper 100 m (Scoccimarro et al. 2011), and marine heatwaves are known to occur below the mixed layer (Elzahaby & Schaeffer 2019); early prediction of OHC anomalies may aid mitigation of extreme events. Marine wildlife is also affected by habitat displacement and shrinking occurring below the surface (Franco et al. 2020). Lastly, the ocean reanalysis products used in this study tend to agree on the upper 300 m heat content trends more than they do for deeper layers (Balmaseda et al. 2015).

There are a multitude of motivations, therefore, to include heat content in future forecasting work. This work highlights key differences between surface and upper ocean predictive skill, and identifies where, and in which seasons, skill is high. We employ two seasonal forecasting systems, both of which are fully-coupled, high-resolution, operational and multi-component. The ocean components are eddy-permitting; an uncommon trait for seasonal forecast systems yet one which is necessary to capture accurate processes such as air-sea exchanges and heat transport (Hewitt et al. 2017;

Roberts et al. 2020). We also use, as a validation dataset, a new ensemble of reanalysis products. The study of these products is very relevant for the next generation of forecast systems. To our knowledge, this is the first attempt to estimate the predictive skills of OHC at seasonal time scales and for the global ocean. We aim to provide a benchmark for future validation efforts, to explore dynamical reasons for measured forecast capability, and to highlight where forecast systems need improvement.

2 Forecast systems: CMCC-SPS3 & ECMWF-SEAS5

The two forecast systems used here are the Seasonal Prediction System Version 3 from the Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC-SPS3), and the fifth generation Seasonal Forecasting System from the European Centre for Medium-Range Weather Forecasts (ECMWF-SEAS5). Both systems were created at the turn of the century as research-only seasonal forecasting systems of the atmosphere and have undergone updates every 5–7 years since. CMCC-SPS3 became fully operational in 2016, and ECMWF-SEAS5 a year later. Since 2018 both systems have been contributing to the Copernicus Climate Change Service (C3S), which makes seasonal forecasts of precipitation, 2 m-temperature, and more, freely available online.

The model components of each system are detailed in Table 1. Both systems base their ocean model component on the eddy-permitting version 3.4 of NEMO, which has a horizontal resolution of 25 km at the equator. The ocean model grid is tripolar, introducing grid cell anisotropy north of 20 N towards the artificial poles over Canada and Siberia. The vertical resolution in ECMWF-SEAS5 is higher than in CMCC-SPS3; in the upper 300 m, there are 35 and 29 vertical levels in ECMWF-SEAS5 and CMCC-SPS3 respectively.

Both systems use versions of their respective in-house ocean reanalysis to create initial conditions. In CMCC-SPS3, the initial conditions are based on C-GLORS (Storto & Masina 2016), while in ECMWF-SEAS5 they are based on ORAS5 (Zuo et al. 2019). Both reanalyses use identical horizontal resolutions (0.25°) and the same sea ice model (LIM2), while the number of vertical levels is 75 and 50 for ECMWF-SEAS5 and CMCC-SPS3 respectively. Both use atmospheric forcing from ERA-Interim until 2016, and ECMWF's NWP analysis thereafter. Both systems used a variant of the CORE bulk formulation, although ORAS5 also includes wave forcing. Both systems assimilate temperature and salinity profiles, and altimeter derived sea-level anomalies, but the assimilation methods and observational datasets also differ. C-GLORS uses the 3D-variational data assimilation scheme OceanVar (Dobricic and Pinardi 2008; Storto & Masina 2016), while ORAS5 uses NEMOVAR. Thus, within the ocean initial conditions alone there are several factors which may contribute to differences in forecast output between the two systems.

The atmospheric model components have in common only the initial conditions (Table 1). The configuration of IFS in ECMWF-SEAS5 provides higher vertical and horizontal resolution than CAM in SPS3. CMCC-SPS3 uses the CPL7 coupler from the Community Earth System Model (CESM, Craig et al. 2012), while ECMWF-SEAS5 uses a single-executable (Mogensen et al. 2012). The coupling

Table 1 Component, resolution and initialisation details for CMCC-SPS3 and ECMWF-SEAS5 coupled forecast systems

	ECMWF-SEAS5	CMCC-SPS3
Ensemble	51	40
Forecast period	7 months (first 6 used here)	6 months
Coupler	Single-Executable	CPL7
Atmosphere		
Model	IFS	CAM
Horizontal Resolution	36 km	1°
Vertical Resolution (top)	91 levels (0.01 hPa)	45 levels (0.3 hPa)
Initialisation	ERA-Interim	ERA-Interim
Ocean		
Model	NEMO v.3.4	NEMO v3.4
Horizontal Resolution	0.25° tripolar grid	0.25° tripolar grid
Vertical Resolution	75 levels	50 levels
Initialisation	ORAS5	C-GLORS
Sea Ice	LIM2	CICE4
Waves	0.5°	N/A
Land	Embedded within IFS	CLM 4.5 1°
Rivers	Climatology	River Transport Model (RTM)

occurs every 90 min in CMCC-SPS3, every 60 min for ECMWF-SEAS5, with both capturing diurnal cycles. In both, ocean and sea-ice models are tightly coupled (i.e. they share a horizontal grid). Meanwhile, the atmosphere and wave models provide fluxes of heat, momentum, freshwater and turbulent kinetic energy to the ocean and sea ice components, while the ocean and sea-ice models provide SST, surface currents and sea-ice concentration in return.

Both ECMWF-SEAS5 and CMCC-SPS3 ensembles sample the uncertainty in the initial conditions of the land, ocean and atmosphere (Table 1). The size of the ensemble (50 for ECMWF-SEAS5 and 40 for CMCC-SPS3) ensures a high signal-to-noise ratio in the ensemble mean. In CMCC-SPS3, the ensemble is made by combining various perturbations in each initial condition set: 10 perturbations of the atmospheric component, 4 of the ocean component and 3 of the land-surface component. Then, 40 scenarios are picked from the possible 120. ECMWF-SEAS5 instead applies stochastic physics perturbations to represent uncertainty arising from missing sub-scale processes. ECMWF-SEAS5 produces 7-month forecasts, which is one more than CMCC-SPS3; we therefore use 6-month forecasts in the analysis for consistency. Further details of the ensemble generation are given in Johnson et al. (2019) and Sanna et al. (2017) for ECMWF-SEAS5 and CMCC-SPS3 respectively. The re-forecast period studied here is 1993–2016.

3 Validation datasets and methods

3.1 EN4

EN4 is an objective analysis of ocean temperature and salinity derived from many profiling instruments and is provided on a 1° grid with 42 vertical levels (Good et al. 2013). Here, we do not use EN4 as a validation dataset, against which the seasonal forecasts are compared. Instead, we compare reanalyses (Sect. 3.3) to EN4 to highlight the geographical areas in which historical ocean records disagree (Sect. 3.3). This is not to say that ocean reanalyses are a better alternative to EN4 in the validation of seasonal forecasts. Although there is a lack of a widespread network of in-situ subsurface temperature observations, such data gaps also affect reanalyses because they assimilate data.

3.2 ESA CCI sea surface temperature

The European Space Agency Climate Change Initiative (ESA CCI) is a collection of climate data records for 26 Essential Climate Variables (ECVs). The records include Essential Ocean Variables, including Version 2.1 of the SST product which is employed here (Good 2020). It is described as a “gap-filled, daily blend”. In practice, “blend” means the product is derived from several radiometers orbiting Earth since 1981 (Merchant et al.2019).

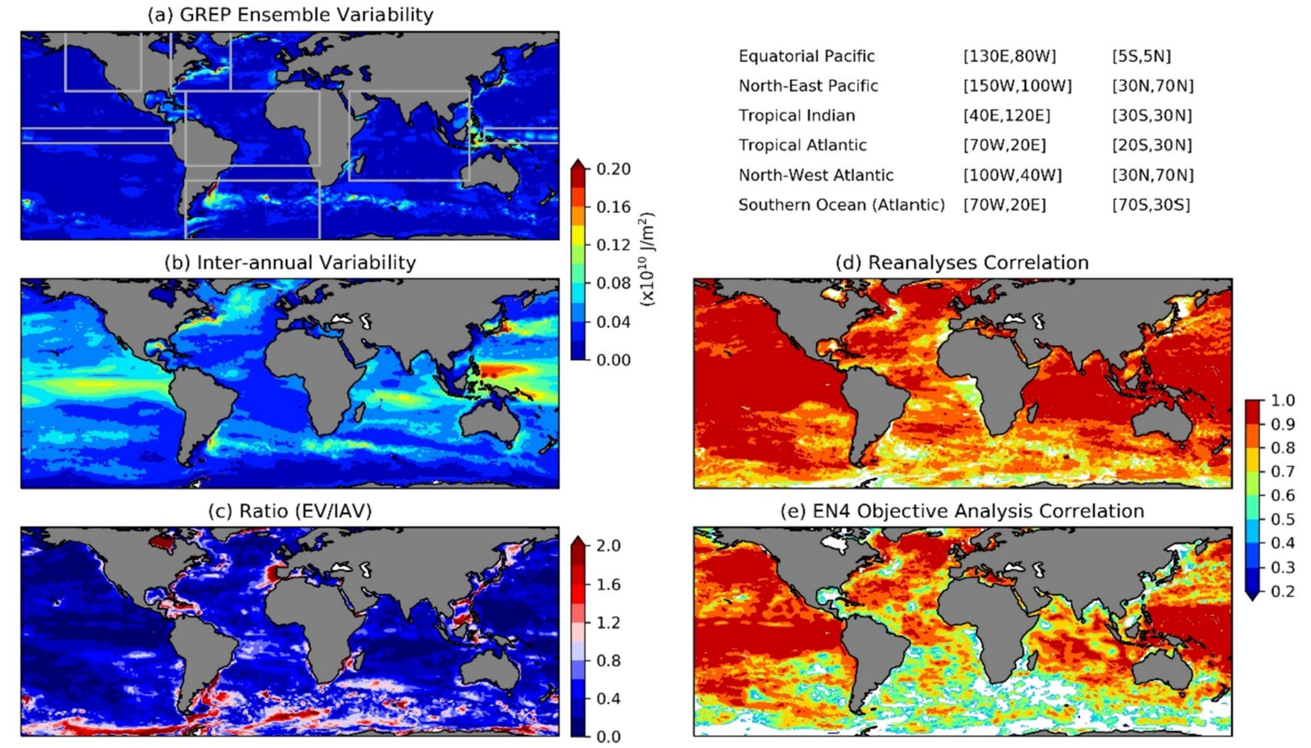
Given the diverse range of input into this product, there is some overlap with the data assimilated into the forecast systems’ initial conditions. Thus, it is not a truly independent validation dataset. Moreover, the quality of the data is not consistent throughout its availability period, although this could be said for many long-term climate datasets. Nonetheless, the ESA CCI SST provides a high-resolution, long-term product which allows for validation against 24 years of re-forecast data. Thus, interannual and, to a lesser extent, decadal variability will be included in our validations.

3.3 Global ocean reanalysis ensemble product (GREP)

GREP is an ensemble of four global 3D ocean reanalysis products: C-GLORS v7 (CMCC: Storto & Masina 2016), FOAM (Met Office UK: Blockley et al. 2014), GLORYS2V4 (Mercator: Garric et al. 2017) and ORAS5 (ECMWF: Zuo et al.2019). All products are built on version 3 of NEMO and are provided from 1993 to 2019 on the native ORCA025 tri-polar curvilinear grid. There are 75 depth levels, 34 of which are shallower than 300 m. All use the same fluxes (CORE) and atmospheric forcing (ERA-I) (with the subtle exception being ORAS5, as mentioned in Sect. 2). All products assimilate similar data streams, typically ARGO, XBT temperature profiles and AVISO Sea Level Anomaly. However, the products all have diverse assimilation schemes, observation quality control, model parameters, spin-up and surface constraints (Storto et al. 2019).

Ocean reanalyses are the unique choice for the task of global heat content validation because the ocean variables have coverage in space and time that is not matched by observations (Riser et al. 2016). Besides, ocean reanalyses integrate the observational information with that of atmospheric reanalyses via a physical ocean model (Balmaseda et al. 2013). An ensemble of ocean reanalyses, such as GREP, is more powerful than a single standalone reanalysis; the ensemble nature of the product accounts for a range of uncertainties represented by the diverse inputs and methods used in each member. Storto et al. (2019) found the ensemble mean was a significant improvement on previous single-member versions of reanalyses, across a range of marine variables.

In most parts of the ocean, the reanalysis ensemble agrees on the climatological mean; the ensemble standard deviation is typically lower than 0.02×10^{10} J/m^2 (Fig. 1a) while, in contrast, the interannual variability is typically higher than this value (Fig. 1b, c). The tropical Pacific and Indian Oceans, for example, are marked by relatively high interannual variability, but low ensemble spread (in agreement with Palmer et al. (2017)). Western boundary currents (WBCs)



Equatorial Pacific	[130E,80W]	[5S,5N]
North-East Pacific	[150W,100W]	[30N,70N]
Tropical Indian	[40E,120E]	[30S,30N]
Tropical Atlantic	[70W,20E]	[20S,30N]
North-West Atlantic	[100W,40W]	[30N,70N]
Southern Ocean (Atlantic)	[70W,20E]	[70S,30S]

(a) GREP Ensemble Variability
(b) Inter-annual Variability
(c) Ratio (EV/IAV)
(d) Reanalyses Correlation
(e) EN4 Objective Analysis Correlation
0.20
0.16
0.12
0.08
0.04
0.00
(x10^10 J/m^2)
2.0
1.6
1.2
0.8
0.4
0.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Fig. 1 Ocean Heat Content 0–300 m statistics for the 1993–2016 period. **a** Standard deviation of annual climatologies across GREP members. **b** Inter-annual variability of GREP ensemble mean (EM). **c** Ratio of inter-annual variability to ensemble member variability. **d** Average correlation of GREP members with GREP EM. **e** Correlation of GREP EM with EN4. White regions in the correlation plots indicate where correlations are statistically insignificant. All measures cover the 1993–2016 period. Grey boxes in **a** mark boundaries of regions of interest used in later Figures, and the coordinates of the bounding lines are found in the top right of the figure

and other frontal regions in the Antarctic Circumpolar Current also display greater GREP uncertainty than natural variability (Fig. 1c). The correlation between individual GREP products and the ensemble mean is lowest in these same regions (Fig. 1d), as is the agreement between GREP ensemble mean and the EN4 analysis product (Fig. 1e). Given the relatively large disagreement in the reanalyses in these regions, the forecast validation may be less reliable. While the tropical variability, being wind driven, is generally temporally coherent among ocean reanalyses, the frontal region variability, on the other hand, is dominated by chaotic nature of the ocean and is consistently high throughout the year (not shown). In addition to the chaotic variability, inherent ocean model errors imply that fronts and WBCs remain difficult to recreate in current reanalyses.

From here on in, the term "GREP" will refer to the reanalyses' ensemble mean. As our forecast systems are initialised with a version of either ORAS5 or C-GLORS, the OHC0-300 m validation dataset is not truly independent. However, as for the ESA CCI SST, the spatial and temporal coverage is unparalleled and necessary for a comparison of long-term data.

Besides acting as the validation datasets, ESA CCI and GREP are also used to construct persistence re-forecasts for SST and OHC respectively. A persistence re-forecast assumes a chosen anomaly persists over the forecast time, thus acting as a very cheap forecast. The initial anomaly, here, is the anomaly of the monthly mean from the month prior to the start-date. For example, the persistence model for the May start date propagates the April anomaly forward 6 months. Given the relative ease with which this forecast can be made, a computationally- and economically- expensive forecast system cannot justify its existence if it is outperformed by a persistence forecast. Thus, persistence forecasts are used to identify where dynamical forecast systems must be improved.

3.4 Forecast skill measures

Four skill measures are used to quantify re-forecast performance:

- Bias: $\frac{1}{Y}\left(\sum_y^Y F_y - \sum_y^Y V_y\right)$
- Normalised Root-Mean-Square-Error (N-RMSE): $\frac{1}{\sigma_v}\sqrt{\frac{1}{Y}\sum_y^Y \left(f_v - v_y\right)^2}$
- Anomaly Correlation Coefficient (between f and v).

- Amplitude Ratio (r), which is constructed from individual ensemble member statistics:
  - $C_n = \frac{1}{Y}\sum_y^Y F_{y,n}$ (ensemble member climatological mean)
  - $\sigma_n = \sqrt{\frac{1}{Y}\sum_y^Y \left(F_{y,n} - C_n\right)^2}$ (ensemble member inter-annual variability)
  - $\sigma_f = \frac{1}{N}\sum_n^N \sigma_n$ (mean of ensemble members' interannual variability)
  - $r = \frac{\sigma_f}{\sigma_V}$

F represents the SST or OHC values from the forecast output; specifically F_y represents the forecast ensemble mean values, while $F_{y,n}$ (see Amplitude Ratio) represents an ensemble member n. N is the total number of re-forecast ensemble members. V is the corresponding variable output from the validation dataset (in the case of GREP, it is the ensemble mean). The variable anomalies are represented by (lower-case) f and v; each set of anomalies is calculated against its own corresponding climatology. Y represents the total number of years (24). Each skill measure is applied to a particular season (i.e. F_y and V_y correspond to seasonal averages of May–June-July values across the 24 years). σ_V and σ_v are the standard deviations of the validation dataset value and anomalies respectively. The climatological means used here are taken from the 1993–2016 period, for the relevant seasonal (three-monthly) average.

Bias is simply the difference in climatological mean values. RMSE, which measures the average error of re-forecast anomalies relative to the validation dataset, is normalised by the standard deviation (inter-annual variability) of the validation dataset. Correlation quantifies the agreement of year-to-year fluctuations. Amplitude ratio compares the inter-annual variabilities of model and verification datasets. For example, an amplitude ratio of one means the variability in each dataset is equal, though the anomalies may not be concomitant; the correlation coefficient would be required to decide if the anomalies were in phase. Therefore, using the four skill measures together provides a fuller picture of system capability. In the following section, we begin by highlighting regions with different skill measure levels and explore what this means for forecast skill across the global ocean.

In Sects. 4.2 and 4.3, we compare correlation scores of different forecast outputs (dynamical systems versus persistence, and OHC versus SST) and include specific statistical significance testing for this task. Comparing correlation scores of OHC forecasts to SST forecasts (Sect. 4.2) involves 4 datasets: one forecast and one validation dataset for each variable. The correlations are therefore independent (note this is not the same as the datasets being independent), and the significance of the difference can be tested as such (Eq. 5 in Siegert et al.(2017)). On the other hand, comparing OHC skill in dynamical systems and persistence forecasts (Sect. 4.3) involves only 3 datasets, because both correlations are calculated relative to the same validation dataset. The correlations, in this case, would not be independent. It is therefore necessary to use an adapted significance test which incorporates the correlation between the two forecasts (Eq. 7 in Siegert et al. (2017)). In all statistical tests in the main body of this paper, statistical significance will be defined at a confidence level of 5%.

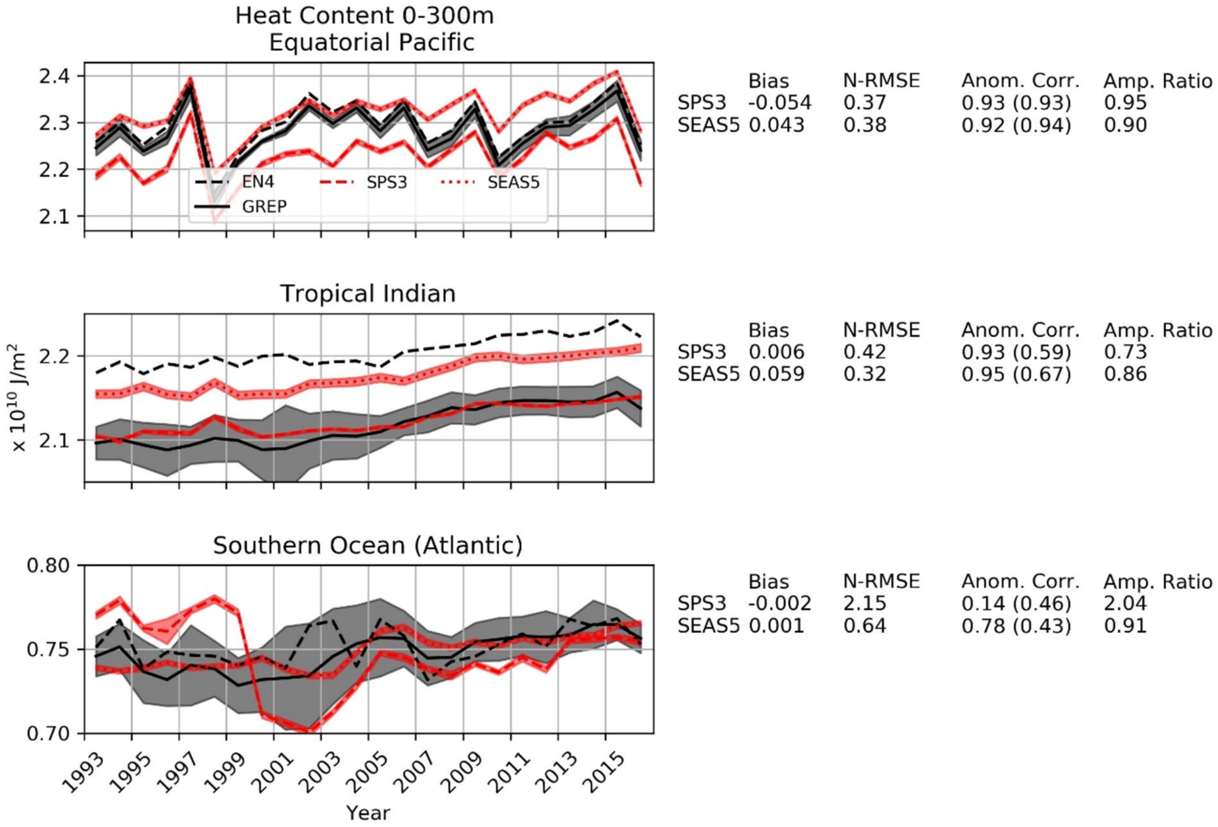
Seasonal re-forecasts from 1993 to 2016 are used in this study (November forecasts cover Feb-Mar-Apr of 2017). The considered forecast period is 6 months from initialisation and the output used here is monthly-averaged. Four start dates are used to focus the validation: February, May, August and November. "Lead" refers to the time in advance that the prediction is made; for example, for a May start, lead season 0 refers to a prediction of May–June-July, while lead season 1 refers to a prediction of August–September-October.

4 Forecast skill

4.1 Global assessment of forecasting skill

There is a diverse range of OHC forecast skill levels across the global ocean. Figure 2 shows three such examples: the Equatorial Pacific, where inter-annual variability is high and so is the skill; the Tropical Indian Ocean, which is marked by warming trends, negligible variability, and good skill; the Southern Ocean, where dynamical forecast skills are very low. Despite biases in the forecast of the Equatorial Pacific (warm in ECMWF-SEAS5, cool in CMCC-SPS3), large ENSO-driven anomalies are predicted with high skill in both systems (low errors, high anomaly correlation and amplitude ratio close to 1). The Tropical Indian Ocean, meanwhile, does not experience the sharp inter-annual variability that the Equatorial Pacific does, and displays good predictability in both systems. The correlation of the detrended time series for this region is 0.3–0.4 lower in both systems, highlighting the importance of capturing the warming trend. However, as demonstrated by the wide confidence intervals in the ensemble mean, even the individual GREP products disagree on the trend value (from 0.26 to 0.36 J/m^2 per decade).

The Southern Ocean area-average is, according to the skill measures (low correlation, forecast variability double the reanalysis variability), poorly predicted by CMCC-SPS3 yet well predicted by ECMWF-SEAS5. Although there is large uncertainty in the GREP ensemble mean (large confidence intervals, Fig. 2), CMCC-SPS3 is outside this range from 1993 to 2000, and then at the lower bound for the remainder of the period. More notably, there is an abrupt change in OHC bias in 2000; this behaviour in CMCC-SPS3



	Bias	N-RMSE	Anom. Corr.	Amp. Ratio
SPS3	-0.054	0.37	0.93 (0.93)	0.95
SEAS5	0.043	0.38	0.92 (0.94)	0.90

	Bias	N-RMSE	Anom. Corr.	Amp. Ratio
SPS3	0.006	0.42	0.93 (0.59)	0.73
SEAS5	0.059	0.32	0.95 (0.67)	0.86

	Bias	N-RMSE	Anom. Corr.	Amp. Ratio
SPS3	-0.002	2.15	0.14 (0.46)	2.04
SEAS5	0.001	0.64	0.78 (0.43)	0.91

Fig. 2 Examples of OHC 0–300 m re-forecast skill measures applied to regional averages. Time series' of August-September-October seasonally-averaged values, for the GREP reanalysis and the ensemble mean of re-forecasts initialised in May. Shading shows 95% confidence intervals in the ensemble means. The corresponding forecast skill measures are shown to the right. The units of the bias are 10^{10} J/m^2. The bracketed correlation values refer to detrended time series. The boundaries for the three regions—Equatorial Pacific, Tropical Indian Ocean, and the Atlantic Southern Ocean—are shown in Fig. 1

has previously been noted and is still being explored, but reassuringly is restricted to limited regions (parts of the Southern Ocean, and the Labrador Sea). It is related to the ingestion of observational data and how initialisation techniques respond to the introduction of Argo data in the early 2000s. ECMWF-SEAS5, on the other hand, uses an extrapolation of Argo data into the past, which renders the initial conditions less sensitive to Argo-induced changes. Thus, data sparsity is not the sole issue here; the use of initialisation methods to deal with this sparsity also has knock-on effects on forecasts.

Looking now at the global scale, as well as across the forecast period, it is important to first note that the two forecast systems produce very similar pictures of skill across the ocean (Figs. 3 & 4). Recall that the systems use the same ocean model component and horizontal resolution, albeit with different vertical resolutions and initialisations, as well as different atmospheric component resolutions. Biases are typically larger and of higher magnitude in lead season 1, although this degradation is more pronounced in CMCC-SPS3. What is a "large" OHC bias? First, we consider bias relative to the background variability; bias magnitude is lower than the GREP interannual variability in 79% and 65% of the ocean surface area between 70S and 70 N, for the first and last half of the May forecast period respectively in CMCC-SPS3. We can also compare to the validation dataset uncertainty; bias magnitude is lower than the GREP ensemble variability (uncertainty) in only 40% and 29% of the ocean. In terms of magnitude, the largest biases appear in the WBCs, the Maritime Continent and the Antarctic Circumpolar Current (reaching 3–4 × 10^9 J/m^2).

Encouragingly, there are many regions where OHC re-forecasts perform well. Specifically, these are regions marked by relatively high (close to 1) and significant correlations, accurate variability (amplitude ratio near 1) and low errors. Many regions retain this level of performance into the latter half of the forecast period: the Equatorial Pacific, the subpolar Atlantic, the North-East Pacific and even the Mediterranean Sea. Besides the bias, the skill measures largely "agree" with one another in the sense that low N-RMSE typically corresponds with high anomaly correlation and amplitude ratio near one. Given that each skill measures

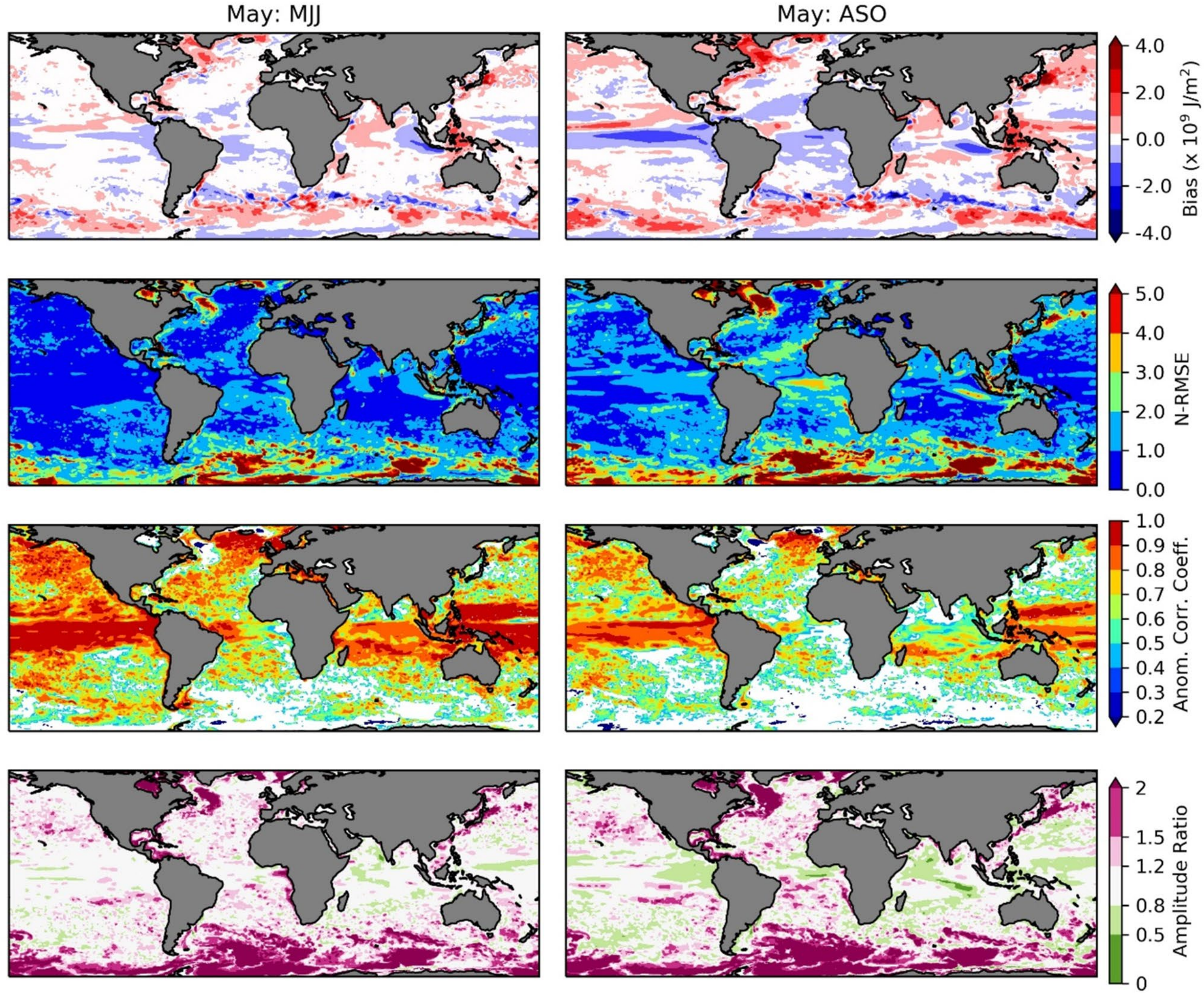


Fig. 3 Skill measure maps for CMCC-SPS3 re-forecasts of OHC 0–300 m initialised in May: Bias, Normalised-RMSE, Anomaly Correlation Coefficient and Amplitude Ratio. Left column: First lead season (May–June-July). Right column: Second lead season (August-September-October). Bias values in the range [-0.25×10^9, 0.25×10^9] are shown in white. White regions in the correlation maps indicate where correlations are not statistically significant

different forecast properties, this is even more encouraging. In particular, the Pacific tropics and extratropics are consistently among the best-scoring regions; anomaly correlation remains above 0.8 into the second lead season, even in CMCC-SPS3 where there is cool bias (below -2×10^9 J). In contrast, the Atlantic basin displays comparatively lower anomaly correlation (<0.4) over the tropics, but high skill over large areas of the North Atlantic, even in the second season.

Focusing on the poorly performing regions (the Southern Ocean and the WBCs), we see they are afflicted by overestimation of variability (amplitude ratios above 2), very low correlation and warm biases (above 3×10^9 J/m^2). Correlation values lower than a certain value are deemed statistically insignificant, depending on the test used (here, the threshold is approximately 0.4). Thus, the sample size is not large enough to provide confidence in interpreting these regions. Nonetheless, there are important lessons to learn from the differences across the forecast period and between systems. The magnitude and the geographical spread of poor skill measures in such regions are greatly reduced in ECMWF-SEAS5 compared to CMCC-SPS3 (Figs. 3, 4). In ECMWF-SEAS5, the worst-performing regions specifically match the patterns made by fronts (diagonal bands in the amplitude ratio maps throughout the Southern Ocean and along the paths of the WBCs), highlighting these features as problematic. CMCC-SPS3 displays the same level of poor skill over the wider Southern Ocean. Despite this difference, it is intriguing that both systems excessively overestimate variability in the same regions. Whether this is a

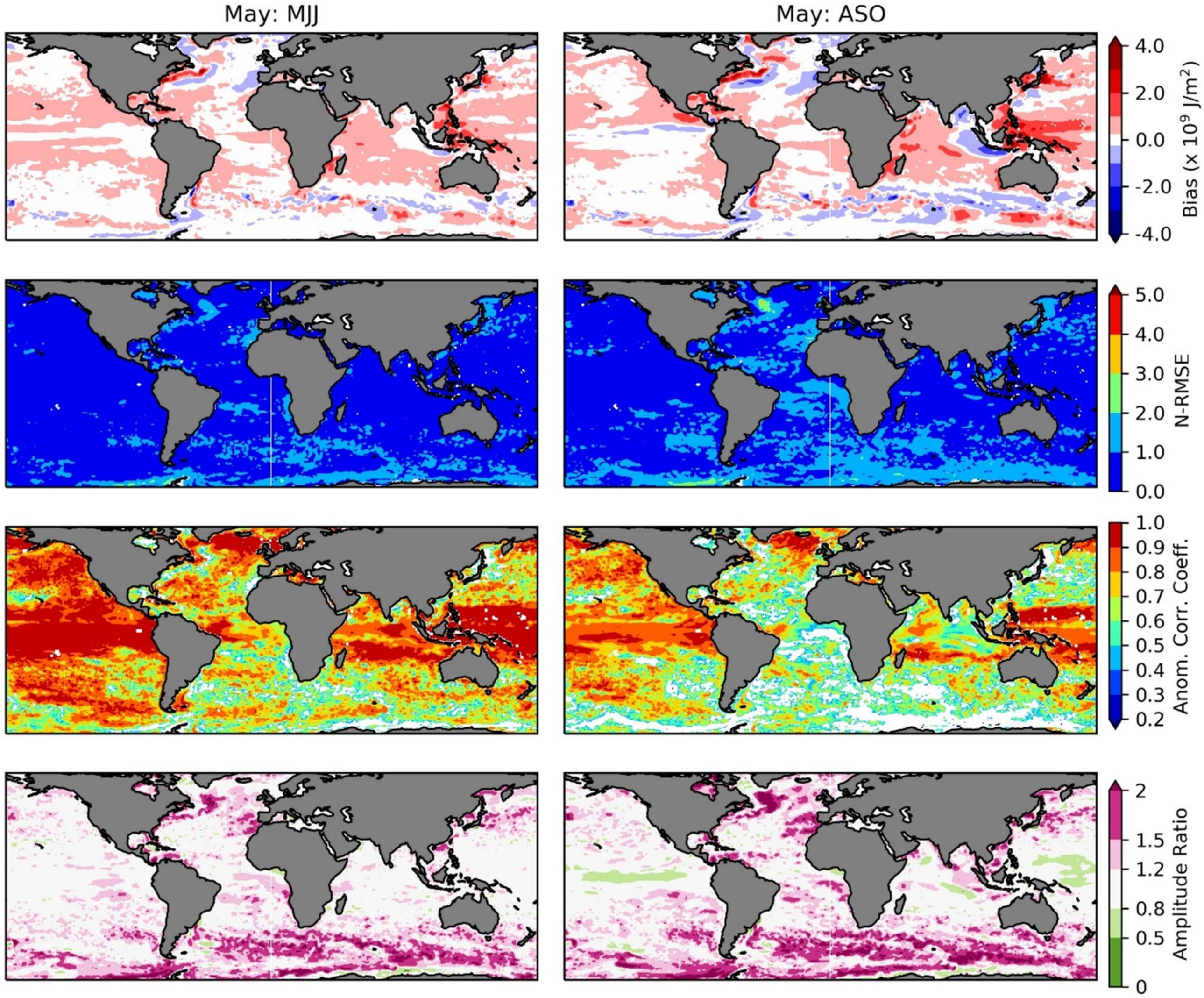


Fig. 4 Skill measure maps for ECMWF-SEAS5 re-forecasts of OHC 0–300 m initialised in May. See Fig. 3 caption for details

common problem in either the ocean initial conditions or in the dynamical modelling of either system remains an open question.

Poor skill scores in the Southern Ocean and the WBCs have also been found in the SST re-forecasts, suggesting that the drivers of poor skill affect the entire upper 300 m in a similar way (Supplementary Figs. 1, 2; Johnson et al. 2019; Sanna et al. 2017). The Southern Ocean in particular is less well covered by sub-surface in-situ observations than the rest of the global ocean (Riser et al. 2016), rendering the initial conditions less accurate. However, SST skill measures are equally poor despite SST being much better sampled in the reanalysis products used for initialisation. It may be that inaccurate sub-surface initial conditions lead to degradation of both SST and OHC. Another possibility is that the horizontal resolution in the ocean is insufficient to maintain the sharp isopycnal fronts associated with the WBCs or edge of the Antarctic Circumpolar Current, although other errors related with air-sea interaction processes cannot be discarded. The current generation of coupled-models still struggles to accurately resolve the myriad of processes and atmosphere–ocean-ice couplings which occur both at the surface and throughout the water column (Meijers 2014).

Meanwhile, in the North Atlantic, the transition from subtropical to subpolar gyres (North-Eastern Atlantic) is marked by good skill measures while, in contrast, the Labrador Sea suffers from the same inadequate skill measures as the Gulf Stream (in both systems). The Labrador Sea appears to fall inside the field of influence of whatever is driving Gulf Stream overheating (Figs. 3, 4), while the North-Eastern Atlantic does not. Previous validation work has linked SST forecast errors, in the North-Western part of the basin, with inaccurate representation of meridional transport (Tietsche et al. 2020). It is important to note also that there is a contrast in GREP ensemble variability between the Gulf Stream (high variability) and the Labrador Sea (low) (Fig. 1). Neither C-GLORS nor ORAS5 (the reanalyses used to initialise the dynamical systems) correlate well with the GREP EM OHC (not shown but both resemble Fig. 1d); this is a sign that neither set of initial conditions is more similar to the validation dataset than the other. The low skill, then, seems to be an issue with dynamical modelling. Nonetheless, accurate predictions in the WBCs and the Southern Ocean, despite their importance in global heat transport, remain an elusive target, even below the surface.

There is a greater difference between the two forecasting systems in the normalised RMSE than there is in the other skill measures. The errors in both forecasts systems are normalised with respect to the same validation dataset (GREP). The patterns in N-RMSE are mirrored by amplitude ratios (both high and low), which is expected as both provide a measure of variability. Outside of the Atlantic, ECMWF-SEAS5 errors are encouragingly within the range of natural variability (N-RMSE < 1). Meanwhile, normalised errors in CMCC-SPS3 are typically 1–2 times the natural variability. In the Southern Ocean and WBCs, CMCC-SPS3 errors can worryingly reach 5 times the natural variability, induced by the problem with initialization described at the beginning of this section.

4.2 Comparison of OHC and SST skills

The promising skill in OHC prediction indicates that it is generally more predictable than SST (Figs. 3, 4, Supplementary Figs. 1, 2). Subsurface anomalies are thought to persist longer than surface anomalies, making subsurface seasonal predictions inherently easier. To test this, we compare the skills of OHC and SST persistence forecasts (see Sect. 3.2 for details on constructing the persistence model). The skill of a persistence model is used as a proxy for the persistence (duration) of anomalies in time (i.e. persistence skill is high if the anomalies in the validation dataset do not decay). In the interest of brevity, anomaly correlation coefficient is used as the comparison skill measure. As expected, we find that OHC persistence correlation is higher than SST persistence correlation in 70% of the ocean surface area between 70S and 70 N (in both lead seasons) (Fig. 5, top; Supplementary Fig. 3). However, this difference is statistically significant in only 20% of the ocean (in both lead seasons); if a p-value threshold of 0.1 is used instead of 0.05, the area covered only increases to 30%. It is possible, then, that OHC anomalies simply do not persist for significantly longer on the seasonal timescales than SST anomalies do. It is also possible that the common time period of each product (24 years) is too small a dataset to detect significant differences between surface and subsurface anomaly persistence (Fig. 5, top). Datasets stretching further back (or forward) in time would be required to confirm this point.

Nonetheless, there are regions where OHC is (significantly) more persistent than SST (on seasonal timescales), and they are the Equatorial Pacific, north-east Atlantic and parts of the North Pacific (Fig. 5, top). In such regions, it is expected that the OHC re-forecast skill in our forecast systems would be higher than the SST re-forecast skill. This is shown to be true, as the improvements in OHC skill over SST skill in the dynamical systems match the geographical extent and magnitude of the differences in persistence (Fig. 5). In fact, at longer lead times the improvement of OHC forecasts over SST increases, again as a result of the longer persistence.

There are also many regions where OHC skill is less than or insignificantly different to SST skill in the dynamical models, particularly in the first season (Fig. 5). The dynamically active ENSO region over the Central Equatorial Pacific stands out as the area where the dynamical model comparisons differ the most from the persistence forecast comparison. The widespread similarity of scores is a sign that SST dynamical forecasts either exploit the thermal and dynamical memory of the subsurface, or benefit from improved initial conditions relative to the OHC (as a result of denser observations at the surface e.g. satellites). It is likely that both play a role, and the details of that role depend on the density of observations as well as the particular benefits created by these observations in particular regions. However, the differences (between OHC and SST in dynamical forecasts compared to persistence) indicate that, in places such as the equatorial band, SST forecasts from dynamical models implicitly include the predictive skill of OHC.

4.3 Comparison of dynamical systems and persistence

We now evaluate the added-value of dynamical seasonal forecasts of OHC by comparing their skill with the skill of an OHC persistence model. As before, the two systems used here agree on the key points (Fig. 6, Supplementary Fig. 4). In lead season 0 (of May start times), considering only where differences are significant, persistence outperforms the dynamical systems in 24% and 12% of the ocean area, for CMCC-SPS3 and ECMWF-SEAS5 respectively; this decreases to 10% and 4% in lead season 1. Dynamical systems outperform persistence in 21% (CMCC-SPS3) and 35% (ECMWF-SEAS5) of the ocean in lead season 1; this increases to 27% and 40% respectively in lead season 1. In other words, in the majority (over 50% in each case) of the ocean, there is no significant difference between persistence forecasts and the dynamical systems.

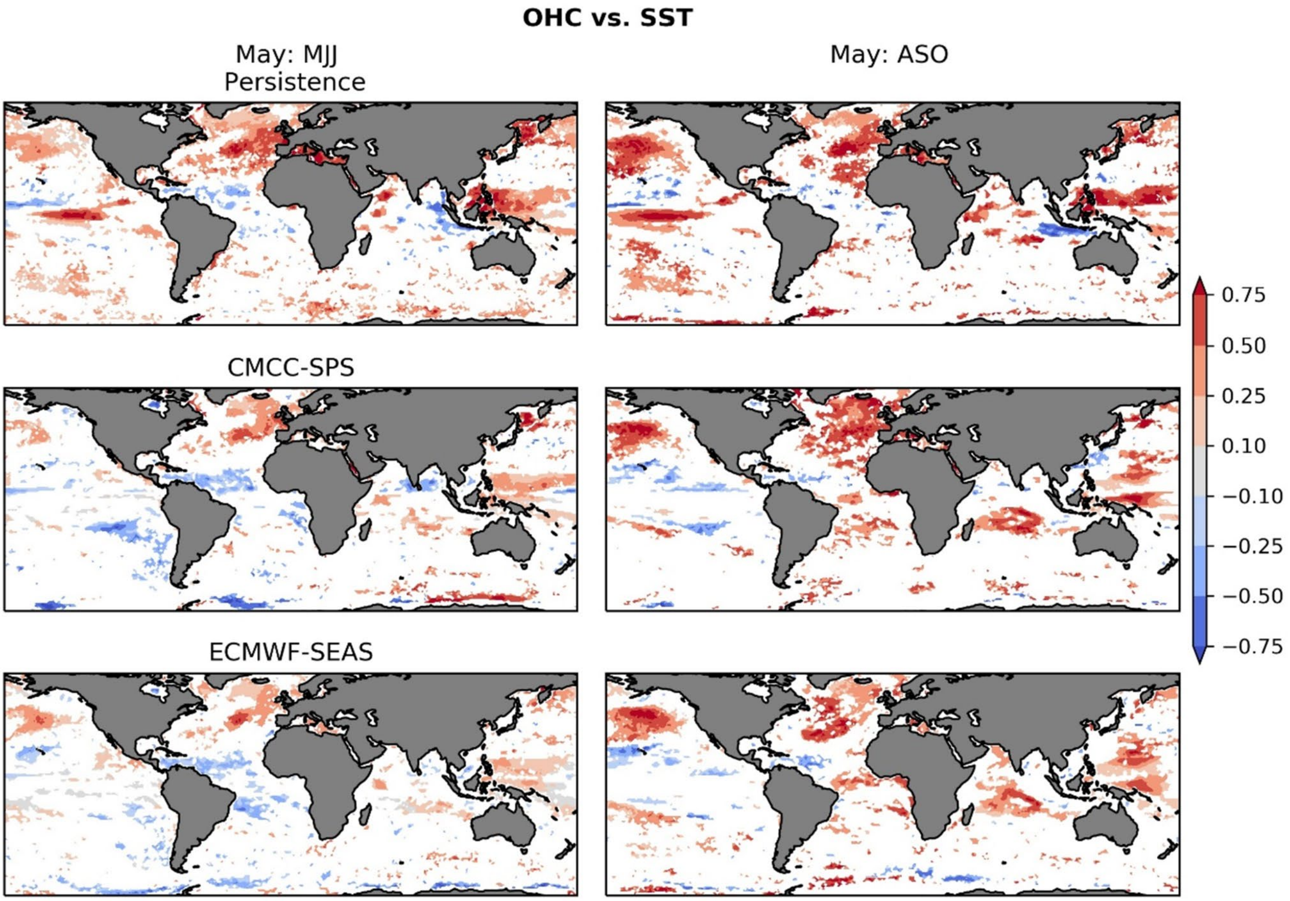


Fig. 5 Difference in anomaly correlation coefficients for OHC 0–300 m and SST in three different models: Persistence (top), CMCC-SPS3 (middle) and ECMWF-SEAS5 (bottom). Positive values show where OHC skill is greater than SST skill in the corresponding model. Re-forecasts and the persistence model are initialised in May, and the seasonal averages of Lead 0 and Lead 1 seasons are shown. White regions indicate where differences in correlations are not statistically significant

The relatively small pockets of the ocean where there is persistence skill in OHC that the dynamics systems fail to capture include the Labrador Sea and parts of the Southern Ocean. These are also regions where there is large uncertainty in the temporal variations of the ocean initial conditions and the validation dataset, as indicated in Fig. 1. This indicates that there is room for improvement in seasonal forecasts of OHC in these regions, either through improvements to the forecasting systems or in the quality and availability of the observations which are assimilated into initial conditions.

It is still reassuring, however, that the dynamical re-forecasts match or outperform persistence across the vast majority of the ocean, especially given that persistence skill is high in many regions (Fig. 7). The dynamical forecasts provide added-value over the less computationally expensive method (see Fig. 7 and following discussion for region-specific quantification), in the Equatorial regions (lead season 0) and into the tropics (lead season 1). Moreover, the areas in which the systems outperform persistence grow as the forecast period continues. In the southern tropical Pacific, for example, the systems begin with skill less than or equivalent to persistence but then show improved performance by the latter half of the period. Likewise, the regions where dynamical systems are outperformed by persistence shrink in the latter period of the forecast period. The added value of the dynamical forecasts is therefore greater for longer lead times.

To synthesise the findings on OHC skill across the seasons, Fig. 7 shows the anomaly correlation coefficients for the two systems (SPS3 and SEAS5), for four start dates, in several regions. The re-forecasts in the Equatorial Pacific have high levels of skill throughout the year and across the forecast period. They consistently beat persistence forecasts; in particular, the improvement is statistically significant and of a greater value in the latter half of the forecast period. Dynamical forecasting therefore adds most value on long lead times (i.e. a season or more in advance). The North-Eastern Pacific is a region where persistence skill is high throughout the forecast period and seasons, yet the

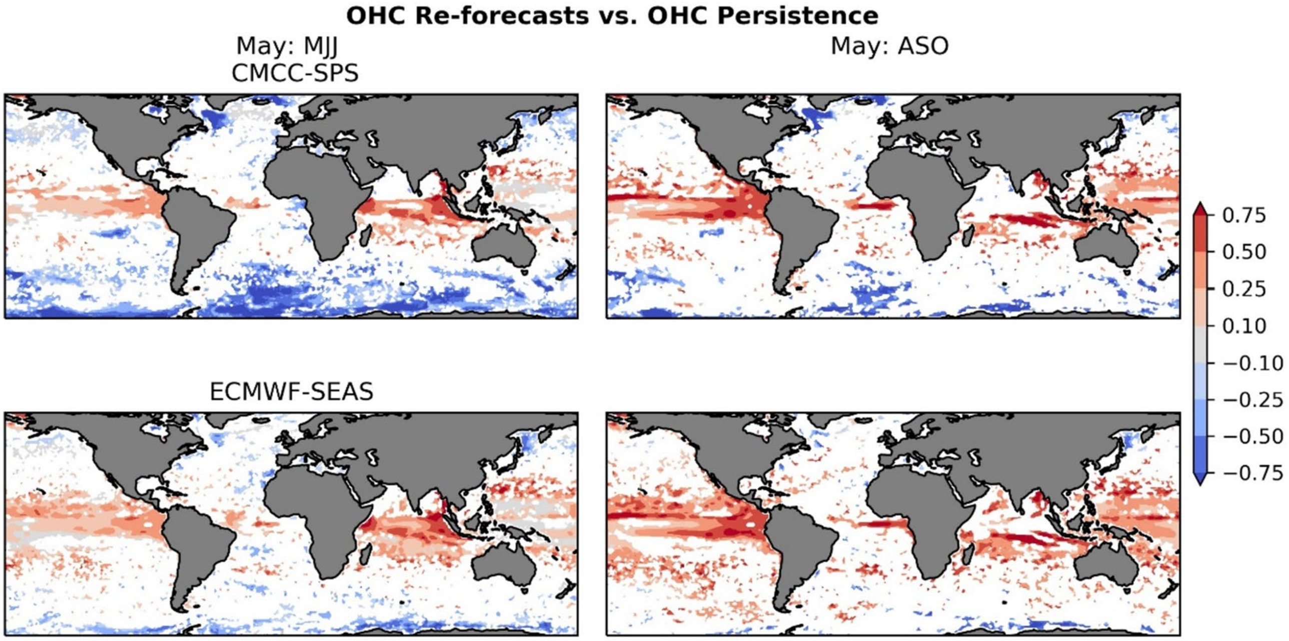


Fig. 6 Difference in anomaly correlation coefficients for OHC 0–300 m re-forecast and persistence models. Positive values show where OHC skill is greater in the forecast system than in the persistence model. Re-forecasts and the persistence model are initialised in May, and the seasonal averages of Lead 0 and Lead 1 seasons are shown. Top: CMCC-SPS3; Bottom: ECMWF-SEAS5. White regions indicate where differences in correlations are not statistically significant

dynamical systems are as skilful. In the Tropical Indian Ocean, both dynamical systems display year-round improvement over persistence, although not always large enough to be a significant improvement. Likewise, in the Tropical Atlantic, dynamical systems beat persistence in the latter half of the forecast period, although by an insignificant margin.

The forecast system skills differ in the North-Western Atlantic and the Southern Ocean, where, as mentioned, the reanalysis products show greater uncertainty (Sect. 3.2). ECMWF-SEAS5 performs better than CMCC-SPS3; the former often beats persistence in the second lead season of re-forecasts, while the latter is typically worse. Despite this, and due to the relatively poor skill in both the dynamical systems and the persistence, the differences are not significant. It should be noted that, given the large geographic area used to define the Southern Ocean (Atlantic), there is more than one issue with the systems in this region. Namely, there appear to be discrepancies in both the position of the many fronts present, as well as the temperature biases further south (shown by warm biases throughout the year, Figs. 3, 4). These performance features may be due to inaccurate representation of dynamics, in specific regions, in the current generation of models, as well as poor initial conditions.

5 Summary and discussion

This paper presents an assessment of the predictive skill of ocean heat content in the upper 300 m in two state-of-the-art seasonal forecasting systems. Here, for the first time, re-forecasts of the global ocean from a range of start-dates have been validated against high-resolution satellite data and a multi-model reanalysis ensemble. The global scope of this study provides a quantification of errors and skill measures for OHC that complement the more traditional assessment of SST skill. While it is premature to state that these findings apply to other forecast systems, the quantification of predictability provided can serve as a benchmark for future endeavours in seasonal forecasting.

Skill measures for OHC remain higher further into the forecast period than for SST, confirming its potential for long-range forecasting. However, in much of the ocean, differences between dynamical forecast skill of OHC and SST are not significant. This is due to either the inadequate sample size for the differences found, or that both variables are predicted with similar accuracy anyway. We interpret this as an indicator that dynamical seasonal forecasting systems are indeed taking advantage of the thermo-dynamical memory in the ocean initial conditions. The extent of this advantage depends on the differences in surface and sub-surface observations which serve as the input for the systems.

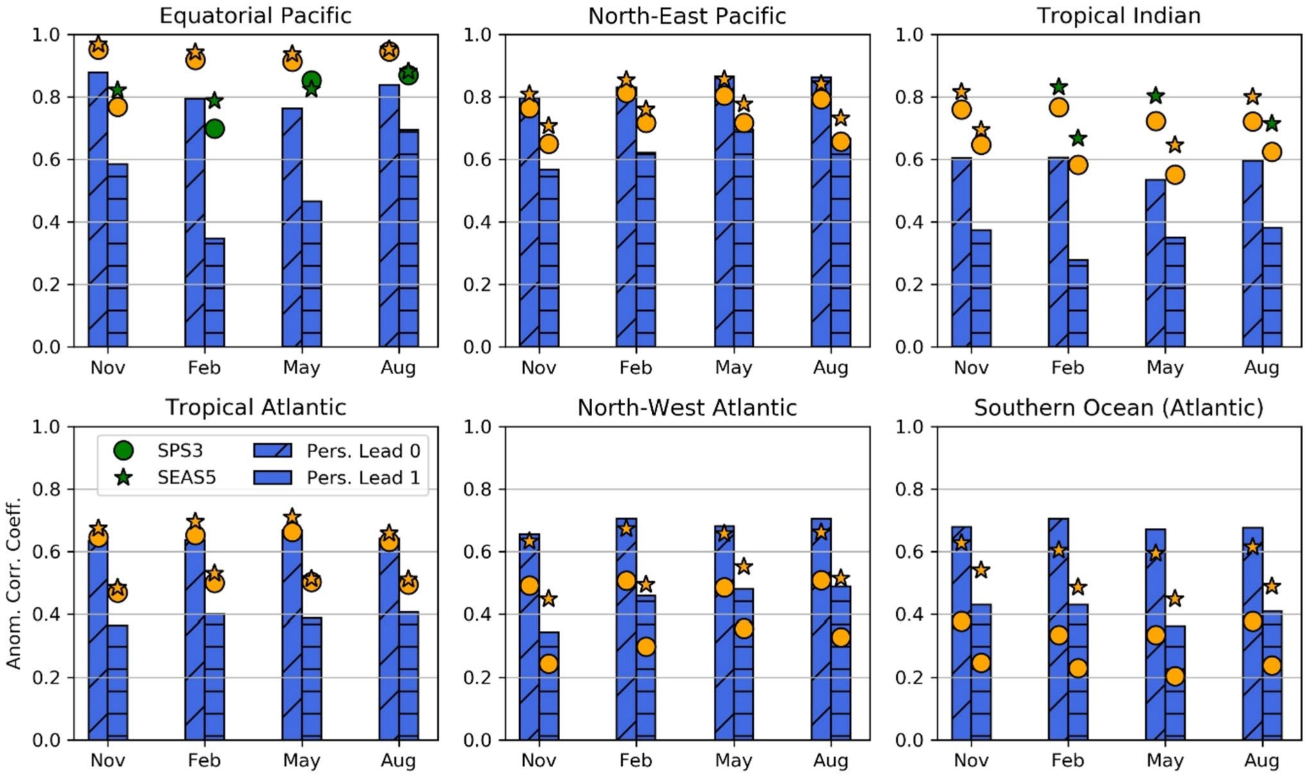


Fig. 7 Anomaly Correlation Coefficient scores for OHC re-forecasts across all start times. OHC persistence model skill is shown by bar charts. Re-forecast skill is represented by symbols. Green symbols indicate statistically-significant improvement (at the 5% confidence level) in the correlation of a dynamical system over the persistence model (there are no cases where the dynamical systems are significantly worse). The boundaries for the regions are listed in Fig. 1

Meanwhile, there are many regions which display high OHC skill measures across the forecast period, and which play key roles in local and global climate: sub-polar Arctic, semi-enclosed seas, centres of subtropical gyres, equatorial Pacific (as well as large part of the north and south basins) and the Maritime Continent. Generally, the regions of high OHC skill are independent of forecast start time. Confidence in OHC predictability has been gained by using multiple skill measures and employing a range of statistical significance measures designed specifically for correlation comparison. The four measures used—bias, normalised root-mean-square-error, anomaly correlation coefficient and amplitude ratio—all detail a different aspect of system behaviour. The use of several measures allows us to see where predictability in some regions is not overly affected by a poor value in one score alone (e.g. there are large SST biases in the north-east Pacific, but anomaly correlation remains high), and therefore provides a truer sense of which regions are predictable.

Dynamical systems are shown to beat persistence models across a range of areas, dynamical environments, seasons and start-times. We find that dynamical forecasts can be improved in the Western Boundary Currents and the Southern Ocean. Specifically, the re-forecast skill measures indicate warm biases and overestimation of the inter-annual variability compared to the validation datasets in these regions. However, in these regions there are also large uncertainties in the GREP product, between other ocean analysis and reanalysis products (Balmaseda et al. 2015), and in coupled models in general (Meijers 2014). It is not yet known whether an imprecision in seasonal forecasting in these regions comes from the dynamical modelling or a lack of constraint on the initial conditions (both surface and sub-surface).

There would also be great benefit in using a larger ensemble of seasonal forecasting systems, as differences in skill can be attributed to resolution, initialisation strategies, assimilation techniques, coupling mechanisms and model components (e.g. Latif et al. 1998; Balmaseda and Anderson 2009). Such a study could potentially explain the poor performance in certain regions (e.g. due to inaccurate initial conditions or insufficient ocean-model resolution). For example, the new version of CMCC-SPS3 (version 3.5) includes a higher resolution atmospheric component (0.5°) and early validation tests on CMCC-SPS3.5 indicate an improvement of SST/OHC skill (not shown) akin to the improvement of ECMWF-SEAS5 over CMCC-SPS3. These models still display some warm biases and exaggerated variability in the WBCs and Southern Ocean, so the reasons behind poor forecast capabilities require more attention.

Nonetheless, this validation work shows there is great potential for OHC forecasts to provide accurate predictions of marine conditions with several months lead time. An interesting question is whether subsurface forecast skill is the result of either enhanced forecast system quality below the surface or of inherent predictability of more persistent subsurface anomalies. By comparing the persistence models of both variables, we show that there is, in places, greater inherent predictability for OHC 0–300 m than for SST. Nonetheless, it will be crucial to study the change, if any, in initial condition uncertainty with depth. Improvements to SST forecasting may arise from a better understanding of OHC initialisation.

Our study has shown that there is potential to make accurate predictions of sub-surface warming up to two seasons in advance, and we believe that this gives sufficient impetus to the application of marine seasonal forecasting. For example, seasonal lead times would provide an early prediction of ocean conditions which render extreme heat events more likely, and therefore provide fisheries, aquaculture farms and marine protects areas ample time to prepare for mass mortality events (e.g. Caputi et al. 2016).

Because of the potential role such forecasts could play in socio-economic decision making, transparency about forecast error is an ethical requirement (Hobday et al. 2016). Thus, based on a validation of anomalies alone, we cannot state that the forecast systems used here are suitable for prediction of, for example, marine heat waves at depth; this would require validation of specific indices (such as the number/intensity of extreme events) in specific regions. The next step in marine seasonal forecasting work is therefore the validation of indices which have socio-economic relevance; steps are already being taken in this direction (Payne et al. 2019). Any validation of this kind will require context on how key variables, such as OHC, behave in seasonal forecast systems. Here, we have provided this first step on a global scale.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00382-021-06101-3.

Acknowledgments This work was supported by the project ‘Improving and Integrating European Ocean Observing and Forecasting Systems for Sustainable use of the Oceans’ — ‘EuroSea’ within the Work Package 4, funded by the European Union under H2020 grant agreement number 862626. We also thank C3S_330 tender and the MEDSCOPE project (MEDSCOPE is part of ERA4CS, an ERANET initiated by JPI Climate, with co-funding by the European Union (grant agreement number 690462)).

Author Contribution All authors are part of Work Package 4.6 in the EuroSea project, whose aim is to provide reliable and sustained marine forecasting services. This study contributes to the first step, a quality assessment of ocean variables from the C3S seasonal forecasts. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by RM, MB and RS. All authors contributed to interpretation of results. The first draft of the manuscript was written by RM. All authors read and approved the final manuscript.

Funding This study was partially funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 862626, as part of the EuroSea project.

Data availability The re-forecasts from ECMWF-SEAS5 and CMCC-SPS3 are publicly available from ECMWF’s Copernicus Climate Change Service (C3S), through its climate data store (https://doi.org/10.21957/p3c285, Raoult et al., 2017). Instructions on how to access these data are available from the C3S user support. L4 SST analyses from the European Space Agency Climate Change Initiative and the Copernicus Climate Service is publicly available at the Copernicus Marine Service (Product Identifier: SST_GLO_SST_L4_REP_OBSERVATIONS_010_024). The GREP ensemble mean product, along with its individual reanalysis members, is also publicly available at the Copernicus Marine Service (Product Identifier: GLOBAL_REANALYSIS_PHY_001_031).

Code availability Not applicable.

Declarations

Conflict of interest The authors have no competing interests.

Ethical approval Not applicable.

Consent to participate All authors participated willingly.

Consent for publication All authors give their consent for publication of the material submitted.

References

Alessandri A, Borrelli A, Masina S, Cherchi A, Gualdi S, Navarra A, Pietro PD, Carril AF (2010) The INGV-CMCC seasonal prediction system: improved ocean initial conditions. Mon Weather Rev 01(138):2930–2952

Alves O, Balmaseda MA, Anderson D, Stockdale T (2004) Sensitivity of dynamical seasonal forecasts to ocean initial conditions. Q J R Meteorol Soc 130:647–667

Baehr J, Fröhlich K, Botzet M, Domeisen DIV, Kornblueh L, Notz D, Piontek R, Pohlmann H, Tietsche S, Müller WA (2015) The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. Clim Dyn 44:2723–2735

Balmaseda MA (2017) Data assimilation for initialization of seasonal forecasts. J Marine Res 75(3):331–359

Balmaseda MA, Trenberth KE, Källén E (2013) Distinctive climate signals in reanalysis of global ocean heat content. Geophys Res Lett 40:1754–1759

Balmaseda M, Hernandez F, Storto A, Palmer M, Alves O, Shi L, Smith G, Toyoda T, Valdivieso M, Barnier B, Behringer D, Boyer T, Chang Y-S, Chepurin G, Ferry N, Forget G, Fujii Y, Good S, Guinehut S, Haines K, Ishikawa Y, Keeley S, Köhl A, Lee T, Martin M, Masina S, Masuda S, Meyssignac B, Mogensen K, Parent L, Peterson K, Tang Y, Yin Y, Vernieres G, Wang X, Waters J, Wedd R, Wang O, Xue Y, Chevallier M, Lemieux J-F, Dupont F, Kuragano T, Kamachi M, Awaji T, Caltabiano A, Wilmer-Becker K, Gaillard F (2015) The Ocean Reanalyses Intercomparison Project (ORA-IP). J Operation Oceanograph 8:s80–s97

Balmaseda, M. & Anderson, D. Impact of initialization strategies and observations on seasonal forecast skill. Geophysical research letters, Wiley Online Library, 2009, 36

Batté, L.; Dorel, L.; Ardilouze, C. & Guérémy, J.-F. Documentation of the METEO-FRANCE seasonal forecasting system 7. 2019

Behrens E, Fernandez D, Sutton P (2019) Meridional oceanic heat transport influences marine Heatwaves in the Tasman sea on Interannual to decadal timescales. Front Mar Sci 6:228

Bilbao R, Wild S, Ortega P, Acosta-Navarro, et al (2021) Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth. Earth Syst Dynam 12:173–196

Blockley, E.; Martin, M.; McLaren, A.; Ryan, A.; Waters, J.; Lea, D.; Mirouze, I.; Peterson, K.; Sellar, A. & Storkey, D. Recent development of the Met Office operational ocean forecasting system: an overview and assessment of the new Global FOAM forecasts. Geoscientific Model Development, 2014

Bruno Soares M, Dessai S (2015) Exploring the use of seasonal climate forecasts in Europe through expert elicitation. Clim Risk Manag 10:8–16

Caputi N, Kangas M, Denham A, Feng M, Pearce A, Hetzel Y, Chandrapavan A (2016) Management adaptation of invertebrate fisheries to an extreme marine heat wave event at a global warming hot spot. Ecol Evol 6:3583–3593

Carton JA, Santorelli A (2008) Global decadal upper-ocean heat content as viewed in nine analyses. J Clim 21:6015–6035

Craig AP, Vertenstein M, Jacob R (2012) A new flexible coupler for earth system modeling developed for CCSM4 and CESM1. Int J High Perform Comput Appl 26:31–42

Darmaraki S, Somot S, Sevault F, Nabat P (2019) Past Variability of Mediterranean Sea Marine Heatwaves. Geophys Res Lett 46:9813–9823

Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LRL (2013) Seasonal climate predictability and forecasting: status and prospects. Wires Clim Change 4:245–268

Dobricic S, Pinardi N (2008) An oceanographic three-dimensional variational data assimilation scheme. Ocean Model 22:89–105

Elzahaby Y, Schaeffer A (2019) Observational insight into the subsurface anomalies of marine heatwaves. Front Mar Sci 6:745

Franco BC, Combes V, González Carman V (2020) Subsurface ocean warming hotspots and potential impacts on marine species: the southwest South Atlantic Ocean case study. Front Mar Sci 7:824

Garric, G.; Parent, L.; Greiner, E.; Drévillon, M.; Hamon, M.; Lellouche, J.-M.; Régnier, C.; Desportes, C.; Le Galloudec, O.; Bricaud, C. & others. Performance and quality assessment of the global ocean eddy-permitting physical reanalysis GLORYS2V4. EGUGA, 2017, 18776

Good SA, Martin MJ, Rayner NA (2013) EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. J Geophys Res Oceans 118:6704–6716

Good, S.A. (2020): ESA Sea Surface Temperature Climate Change Initiative (SST_cci): GHRSST Multi-Product ensemble (GMPE), v2.0. Centre for Environmental Data Analysis, 05 August 2020

Häkkinen S, Rhines PB, Worthen DL (2013) Northern North Atlantic sea surface height and ocean heat content variability. J Geophys Res 118:3670–3678

Hewitt HT, Bell MJ, Chassignet EP, Czaja A, Ferreira D, Griffies SM, Hyder P, McClean JL, New AL, Roberts MJ (2017) Will high-resolution global ocean models benefit coupled predictions on short-range to climate timescales? Ocean Model 120:120–136

Hobday AJ, Hartog JR, Spillman CM, Alves O (2011) Seasonal forecasting of tuna habitat for dynamic spatial management. Can J Fish Aquat Sci 68:898–911

Hobday AJ, Claire M, Hartog JR (2016) Seasonal forecasting for decision support in marine fisheries and aquaculture. Fish Oceanogr 25:45–56

Holbrook NJ, Sen Gupta A, Oliver ECJ, Hobday AJ, Benthuysen JA, Scannell HA, Smale DA, Wernberg T (2020) Keeping pace with marine heatwaves. Nat Rev Earth Environ. https://doi.org/10.1038/s43017-020-0068-4

Johnson S, Stockdale T, Ferranti L, Balmaseda M, Molteni F, Magnusson L, Tietsche S, Decremer D, Weisheimer A, Balsamo G, Keeley S, Mogensen K, Zuo H, Monge-Sanz B (2019) ECMWF-SEAS5: the new ECMWF seasonal forecast system. Geosci Model Dev 12:1087–1117

Latif M, Anderson D, Barnett T, Cane M, Kleeman R, Leetmaa A, O'Brien J, Rosati A, Schneider E (1998) A review of the predictability and prediction of ENSO. J Geophys Res 103:14375–14393

MacLachlan C, Arribas A, Peterson KA, Maidens A, Fereday D, Scaife AA, Gordon M, Vellinga M, Williams A, Comer RE, Camp J, Xavier P, Madec G (2015) Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. Q J R Meteorol Soc 141:1072–1084

Mainelli M, DeMaria M, Shay LK, Goni G (2008) Application of oceanic heat content estimation to operational forecasting of recent Atlantic category 5 hurricanes. Weather Forecast 23:3–16

Masina S, Storto A, Ferry N, Valdivieso M, Haines K, Balmaseda M, Zuo H, Drevillon M, Parent L (2017) An ensemble of eddy-permitting global ocean reanalyses from the MyOcean project. Clim Dyn 49:813–841

Mayer M, Alonso Balmaseda M, Haimberger L (2018) Unprecedented 2015/2016 Indo-Pacific heat transfer speeds up tropical Pacific heat recharge. Geophys Res Lett 45:3274–3284

McPhaden, M. J.; Lee, T.; Fournier, S. & Balmaseda, M. A. ENSO Observations. El Niño Southern Oscillation in a Changing Climate, Wiley Online Library, 2020, 39–63

Meijers A (2014) The Southern Ocean in the coupled model intercomparison project phase 5. Phil Trans R Soc A 372:20130296

Merchant CJ, Embury O, Bulgin CE, Block T, Corlett GK, Fiedler E, Good SA, Mittaz J, Rayner NA, Berry D, Eastwood S, Taylor M, Tsushima Y, Waterfall A, Wilson R, Donlon C (2019) Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci Data 6:223

Mogensen, K.; Keeley, S. & Towers, P. Coupling of the NEMO and IFS models in a single executable. ECMWF Reading, United Kingdom, 2012

Msadek R, Delworth TL, Rosati A, Anderson W, Vecchi G, Chang Y-S et al (2014) Predicting a decadal shift in North Atlantic climate variability using the GFDL forecast system. J Clim 27:6472–6496

Neelin JD, Battisti DS, Hirst AC, Jin F-F, Wakata Y, Yamagata T, Zebiak SE (1998) ENSO theory. J Geophys Res 103:14261–14290

Oliver ECJ, Donat MG, Burrows MT, Moore PJ, Smale DA, Alexander LV, Benthuysen JA, Feng M, Sen Gupta A, Hobday AJ, Holbrook NJ, Perkins-Kirkpatrick SE, Scannell HA, Straub SC, Wernberg T (2018) Longer and more frequent marine heatwaves over the past century. Nat Commun 9:1324

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

Payne MR, Hobday AJ, MacKenzie BR, Tommasi D (2019) Editorial: seasonal-to-decadal prediction of marine ecosystems: opportunities, approaches, and applications. Front Mar Sci 6:100

Raoult B, Bergeron C, Alós AL, Thépaut J-N, Dee D (2017) Climate service develops user-friendly data store. ECMWF Newsletter 151:22–27

Riser SC, Freeland HJ, Roemmich D, Wijffels S, Troisi A, Belbéoch M, Gilbert D, Xu J, Pouliquen S, Thresher A (2016) Fifteen years of ocean observations with the global Argo array. Nat Clim Change 6:145–153

Roberts CD, Vitart F, Balmaseda MA, Molteni F (2020) The time-scale-dependent response of the wintertime North Atlantic to increased ocean model resolution in a coupled forecast model. J Clim 33:3663–3689

Robson, J.; Sutton, R. & Smith, D. Initialized decadal predictions of the rapid warming of the North Atlantic Ocean in the mid 1990s. Geophysical Research Letters, Wiley Online Library, 2012, 39

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Behringer D, Hou Y-T, Chuang H-Y, Iredell M (2014) The NCEP climate forecast system version 2. J Clim 27:2185–2208

Sanna, A.; A. Borrelli, P. A.; S. Materia, A. S. & S. Tibaldi, S. G. CMCC-SPS: The CMCC Seasonal Prediction System 3. Centro Euro-Mediterraneo sui Cambiamenti Climatici . 2017, CMCC Tech. Rep. RP0285, 61pp

Scoccimarro E, Gualdi S, Bellucci A, Sanna A, Giuseppe Fogli P, Manzini E, Vichi M, Oddo P, Navarra A (2011) Effects of tropical cyclones on ocean heat transport in a high-resolution coupled general circulation model. J Clim 24:4368–4384

Scoccimarro E, Bellucci A, Storto A, Gualdi S, Masina S, Navarra A (2018) Remote subsurface ocean temperature as a predictor of Atlantic hurricane activity. Proc Natl Acad Sci 115:11460–11464

Siegert S, Bellprat O, Ménégoz M, Stephenson DB, Doblas-Reyes FJ (2017) Detecting improvements in forecast correlation skill: statistical testing and power analysis. Mon Weather Rev 145(2):437–450

Stockdale TN, Anderson DL, Alves JOS, Balmaseda MA (1998) Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. Nature 392:370–373

Storto A, Masina S (2016) C-GLORSv5: an improved multipurpose global ocean eddy-permitting physical reanalysis. Earth Syst Sci Data 8:679–696

Storto A, Masina S, Simoncelli S, Iovino D, Cipollone A, Drevillon M, Drillet Y, von Schuckman K, Parent L, Garric G, Greiner E, Desportes C, Zuo H, Balmaseda MA, Peterson KA (2019) 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. Clim Dyn 53:287–312

Takaya, Y.; Hirahara, S.; Yasuda, T.; Matsueda, S.; Toyoda, T.; Fujii, Y.; Sugimoto, H.; Matsukawa, C.; Ishikawa, I.; Mori, H. 2018 Japan Meteorological Agency/Meteorological Research Institute-Coupled Prediction System version 2 (JMA/MRI-CPS2): Atmosphere--land--ocean--sea ice coupled prediction system for operational seasonal forecasting. Climate dynamics, Springer, 50, 751–765

Tietsche S, Balmaseda M, Zuo H, Roberts C, Mayer M, Ferranti L (2020) The importance of North Atlantic Ocean transports for seasonal forecasts. Clim Dyn 55:1995–2011

Weisheimer A, Befort DJ, MacLeod D, Palmer T, O'Reilly C, Strømmen K (2020) Seasonal forecasts of the twentieth century. Bull Am Meteor Soc 101:E1413–E1426

Yeager SG, Danabasoglu G, Rosenbloom NA, Strand W, Bates SC, Meehl GA, Karspeck AR, Lindsay K, Long MC, Teng H, Lovenduski NS (2018) Predicting near-term changes in the earth system: a large ensemble of initialized decadal prediction simulations using the community earth system model. Bull Am Meteor Soc 99(9):1867–1886

Zuo H, Balmaseda MA, Tietsche S, Mogensen K, Mayer M (2019) The ECMWF operational ensemble reanalysis—analysis system for ocean and sea ice: a description of the system and assessment. Ocean Sci. https://doi.org/10.5194/os-2018-154