Geosci. Model Dev., 16, 179–197, 2023
https://doi.org/10.5194/gmd-16-179-2023

Geoscientific Model Development
Open Access
EGU

The Euro-Mediterranean Center on Climate Change (CMCC) decadal prediction system

Dario Nicolì[1], Alessio Bellucci[1,a], Paolo Ruggieri[1,b], Panos J. Athanasiadis[1], Stefano Materia[1], Daniele Peano[1], Giusy Fedele[1], Riccardo Hénin[1], and Silvio Gualdi[1]

[1]Climate Simulation and Prediction (CSP) division, Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna 40127, Italy
[a]now at: Consiglio Nazionale delle Ricerche, Istituto di Scienze dell'Atmosfera e del Clima (CNR-ISAC), Bologna 40129, Italy
[b]now at: Department of Physics and Astronomy, University of Bologna, Bologna 40126, Italy

Correspondence: Dario Nicolì (dario.nicoli@cmcc.it)

Received: 12 July 2022 – Discussion started: 21 July 2022
Revised: 15 November 2022 – Accepted: 25 November 2022 – Published: 5 January 2023

Abstract. Decadal climate predictions, obtained by constraining the initial condition of a dynamical model through a truthful estimate of the observed climate state, provide an accurate assessment of near-term climate change and are a useful tool to inform decision-makers on future climate-related risks.

Here we present results from the CMIP6 (Coupled Model Intercomparison Project Phase 6) Decadal Climate Prediction Project (DCPP) decadal hindcasts produced with the operational CMCC (Euro-Mediterranean Center on Climate Change) decadal prediction system (DPS), based on the fully coupled CMCC-CM2-SR5 dynamical model. A 20-member suite of 10-year retrospective forecasts, initialized every year from 1960 to 2020, is performed using a full-field initialization strategy.

The predictive skill for key variables is assessed and compared with the skill of an ensemble of non-initialized historical simulations so as to quantify the added value of the initialization. In particular, the CMCC DPS is able to skillfully reproduce past climate surface and subsurface temperature fluctuations over large parts of the globe. The North Atlantic Ocean is the region that benefits the most from initialization, with the largest skill enhancement occurring over the subpolar region compared to historical simulations. On the other hand, the predictive skill over the Pacific Ocean rapidly decays with forecast time, especially over the North Pacific. In terms of precipitation, the skill of the CMCC DPS is significantly higher than that of the historical simulations over a few specific regions, including the Sahel, northern Eurasia, and over western and central Europe.

The Atlantic multidecadal variability is also skillfully predicted, and this likely contributes to the skill found over remote areas through downstream influence, circulation changes, and teleconnections. Considering the relatively small ensemble size, a remarkable prediction skill is also found for the North Atlantic Oscillation, with maximum correlations obtained in the 1–9 lead year range.

Systematic errors also affect the forecast quality of the CMCC DPS, featuring a prominent cold bias over the Northern Hemisphere, which is not found in the historical runs, suggesting that, in some areas, the adopted full-field initialization strategy likely perturbs the equilibrium state of the model climate quite significantly.

The encouraging results obtained in this study indicate that climate variability over land can be predictable over a multiyear range, and they demonstrate that the CMCC DPS is a valuable addition to the current generation of DPSs. This stresses the need to further explore the potential of the near-term predictions, further improving future decadal systems and initialization methods, with the aim to provide a reliable tool to inform decision-makers on how regional climate will evolve in the next decade.

Published by Copernicus Publications on behalf of the European Geosciences Union.

1 Introduction

Climate fluctuations are the end result of a number of processes acting on a multitude of timescales. Prior to the year 2000, century-scale climate change projections, initialized with a physical state of the climate system obtained from a long simulation of the preindustrial period and subject to prescribed anthropogenic and natural forcings, have been the only available product to inform decision-makers on future climate-related risks. A major limitation of non-initialized climate projections is their lack of information about the ongoing natural variability that may affect climate changes in the near future, which is, at least in part, linked to the current state of the Earth's climate system. Decadal predictions, obtained by constraining the initial condition of a dynamical model (coupled global circulation model/Earth system model) through a realistic estimate of the observed climate state, provide a more accurate assessment of climate change in the near-term (decadal) range, where both external and internal drivers contribute to the climate evolution (Smith et al., 2007; Kushnir et al., 2019).

Starting from the 2000s, initialized decadal predictions have been assessed in multiple projects, from the first pioneering efforts up to the fifth Coupled Model Intercomparison Project (CMIP5; Smith et al., 2007; Keenlyside et al., 2008; Pohlmann et al., 2009; Meehl et al., 2009; Doblas-Reyes et al., 2011), in which coordinated experiments allowed a multi-system comparison to reduce single-model uncertainties (Taylor et al., 2012; Bellucci et al., 2015b), contributing to the Intergovernmental Panel on Climate Change Fifth Assessment Report (AR5, chap. 11; Kirtman et al., 2013).

Years of coordinated research and development have led to an established experiment protocol that has overcome some of the limitations (e.g., limited ensemble size and initialization every 5 years) of the decadal prediction simulations produced in the CMIP5 framework. This protocol is extensively described in the CMIP6 (Coupled Model Intercomparison Project Phase 6) Decadal Climate Prediction Project (DCPP), a coordinated multimodel effort within the World Climate Research Programme (WCRP), which aims to investigate climate predictions, predictability, and variability from annual to decadal timescales (Boer et al., 2016). The DCPP is intended to make skillful forecasts and predictions on these timescales using state-of-the-art climate models and statistical approaches. The core of the DCPP is component A, which includes a set of retrospective forecasts (hindcasts). This framework has laid the groundwork for a number of single-model (Bethke et al., 2021; Bilbao et al., 2021; Kataoka et al., 2020; Robson et al., 2018; Sospedra-Alfonso et al., 2021; Xin et al., 2019; Yang et al., 2021; Yeager et al., 2018) and multimodel studies (Borchert et al., 2021a, b; Delgado-Torres et al., 2022).

Climate anomalies on annual-to-multidecadal timescales are determined from both the internal and the externally forced variability (Boer et al., 2016). External contributions derive from solar irradiance variations, volcanic aerosols, and anthropogenic activities, including land use, aerosols, and greenhouse gas emissions, accounting for the global warming trend. On the other hand, the oceans, and to a lesser degree the land surface, sea ice, and stratosphere (Bellucci et al., 2015a) and their interaction with the atmosphere, are the primary source of internal variability within the climate system on decadal timescales. Low-frequency fluctuations in the North Atlantic sea surface temperature (SST), known as the Atlantic multidecadal variability (AMV), affect the global climate through local impacts and remote teleconnections (e.g., Sutton and Hodson, 2005; Zhang and Delworth, 2005; Knight et al., 2006; Sun et al., 2015; Nicolì et al., 2020; Ehsan et al., 2020; Ruprich-Robert et al., 2021). The long-term AMV evolution is well captured in most state-of-the-art forecast systems and represents one of the primary sources of predictability and, possibly, of skill at a decadal timescale that is generally attributed to the initialization of the Atlantic meridional overturning circulation (AMOC; Zhang et al., 2019). Other predictability sources arise from the Pacific Ocean. Interestingly, decadal El-Niño–Southern Oscillation (ENSO) impacts may be modulated by the Interdecadal Pacific Oscillation (IPO), which features both the ENSO SST region and extends to other extratropical areas (Henley et al., 2015). In addition, the initial state of some land surface characteristics, stratosphere, snow cover, and sea ice may also impact the predictability of the climate system (e.g., Bellucci et al., 2015b; Meehl et al., 2021). The aforementioned initialized components provide additional predictability, for the atmospheric circulation and, in particular, for the North Atlantic Oscillation (NAO) affecting boreal winter climate over Europe (Smith et al., 2019; Athanasiadis et al., 2020; Dunstone et al., 2022).

In this paper, we present the decadal prediction system (DPS) developed at the Euro-Mediterranean Center on Climate Change (CMCC) using the CMCC-CM2-SR5 state-of-the-art climate model (Cherchi et al., 2019) and contributing to the CMIP6 DCPP project. In particular, the study aims to assess the skill in predicting the observed anomalies in key meteorological variables, thereby testing the ability of the DPS to simulate the main climate variations from annual to the decadal timescale.

The article is structured as follows: Sect. 2 provides details on the model configuration, the experimental protocol, the evaluation metrics, and the data used to check the benefits of the initialization. Section 3 presents results on the predictions' skill for key quantities and their evolution in time, using deterministic and probabilistic approaches, and assesses the evolution of some relevant model biases. Section 4 focuses on the skill for selected regional climate variability indices. Finally, Sect. 5 summarizes and discusses the main findings of the study and also draws some conclusions.

2 Data and methodology

2.1 Description of the CMCC DPS model

The CMCC decadal prediction system is based on the CMCC-CM2-SR5 coupled model, shortly described below (see Cherchi et al., 2019, for additional details). The atmospheric component is the Community Atmosphere Model version 5 (CAM5) with a regular grid of 0.9–1.25° and 30 hybrid levels, including 17 levels below 200 hPa and extending up to 2 hPa. The finite-volume configuration has been chosen for the dynamical core. The ocean model is the Nucleus for European Modelling of the Ocean version 3.6 (NEMO v3.6), using a tripolar ORCA grid with a horizontal resolution of about 1° (with a varying latitudinal resolution ranging from 1/3° near the Equator up to 1° at high latitudes) and 50 levels in the vertical. The sea ice component is the Community Ice CodE in its fourth version (CICE4). The DPS configuration of CICE4 uses a single category to characterize the sea ice thickness, for consistency with the respective reanalysis used for the initialization. The Community Land Model version 4.5 (CLM4.5) is used for the simulation of the land surface at the same horizontal grid used by the atmospheric component. Finally, the River Transport Model (RTM; Branstetter, 2001) routes liquid and ice runoff from the land surface model to the active ocean to simulate a closed hydrological cycle.

A suite of retrospective forecasts (hindcasts) consisting of 20-member ensembles of 10-year-long hindcasts, initialized every year from 1960 to 2020, has been completed, following the CMIP6 DCPP-A protocol (Boer et al., 2016). As summarized in Table 1, all the members are initialized on 1 November, starting from direct, full-field estimates of the observed state of the ocean, sea ice, land surface, and atmosphere, without any coupled assimilation runs. For each start date, two initial conditions for the atmosphere are obtained from the ERA-40 (1960–1978; Uppala et al., 2005), ERA-Interim (1979–2018; Berrisford et al., 2011), and ERA5 (2019 onwards; Hersbach et al., 2020) reanalyses, taking the atmospheric states of 1 and 2 November. The ocean, sea ice, and land surface states are initialized with an ensemble of global data assimilation products (ocean and sea ice) and analyses constrained with observed fluxes (land surface). Specifically, land surface is initialized using two different analyses obtained from a land-only configuration of the CLM4.5 land model integrated offline with two different atmospheric forcing datasets, namely CRUNCEP version 7 (Viovy, 2016) and GSWP3 (Kim, 2017). These datasets provide the land model with instantaneous 2 m air temperature and humidity, 10 m winds and surface pressure every 6 h, and accumulated radiation and precipitation every 3 h. Ocean initial states are derived from CHOR (for the period 1960–2010; Yang et al., 2016) and CGLORSv7 reanalysis (for the period 2011–2020; Storto and Masina, 2016) and are performed with a three-dimensional variational data assimilation system, a surface nudging, and a bias correction scheme. It is worth noting that the reanalysis is performed with the same ocean model used in the CMCC DPS (i.e., NEMO v3.6). An ensemble of five ocean initial states is used to initialize the ocean and sea ice components, where three initial estimates originate from global ocean reanalysis characterized by different assimilation strategies of SST and in situ profiles of temperature and salinity in a 0.5° configuration of the NEMO ocean model, while the remaining two initial states are derived through linear combinations of the former three initial states. The ocean initial states provide three-dimensional fields of temperature, salinity, and horizontal currents. The sea ice model has been initialized, starting from sea ice temperature, sea ice volume, sea ice area, and snow volume. Our full-field initialization approach is similar to the ones adopted by other DPSs (e.g., Bilbao et al., 2021; Sospedra-Alfonso et al., 2021) in which the a posteriori correction is applied, since the simulations deviate to their own model attractors. Other DPSs (e.g., Bethke et al., 2021; Brune and Baehr, 2020; Kataoka et al., 2020) are initialized using the observed anomalous variability superimposed on the model climatology to avoid the initial shock. Both the techniques are deemed valid in CMIP6 DCPP protocols (Boer et al., 2016) and present some drawbacks. For example, bias correction in the former may remove part of the variability signal, while the latter has the assumption that the model variability is independent from the model mean state. Nevertheless, several studies have proven that differences in skill are small and localized (Smith et al., 2013; Bellucci et al., 2015b; Volpi et al., 2017).

The time-evolving radiative forcings (including solar radiation, greenhouse gas concentrations, and anthropogenic and volcanic aerosols) are prescribed during the historical period (1960–2014) and follow the Shared Socioeconomic Pathways (SSPs) scenario of SSP2-4.5 (O'Neil et al., 2016) from 2015 onwards, which is in agreement with the CMIP6 DCPP protocol.

2.2 Verification data

Uninitialized historical simulations covering the 1850–2014 period are used to assess the added value of realistic model initialization in decadal predictions. We use a 10-member ensemble of historical simulations initialized with different states of a multi-century preindustrial climate simulation. Each member of the historical ensemble is extended until 2030 under the SSP2-4.5 scenario, thus allowing a fair comparison with the decadal forecast ensemble initialized in the year 2020. Since only 10 historical members are available, the 20-member hindcast ensemble has been scaled down using a random sub-sampling with 100 combinations in order to allow a fair comparison to the historical ensemble in the skill assessments.

The predictive skill for both initialized reforecasts and uninitialized projections is assessed against observational products. The temporal coverage of lead year 1 is 1961–2020, since not every observational product used in this study

Table 1. List of initial conditions used for the generation of the 20-member hindcast ensemble.

	Data source	No. of initial conditions (ICs)	Procedures
Land	Land-only analyses forced by two different atmospheric datasets, namely CRUNCEPv7 (Viovy, 2016) and GSWP3 (Kim, 2017). Note that, from 2015 onwards, the atmospheric fluxes to force the land-only analysis are taken from the National Centers for Environmental Prediction (NCEP) reanalysis (instead of CRUNCEPv7) and from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 (instead of GSWP3)	Two ICs (two runs forced by two different datasets, providing instantaneous 2 m air temperature and humidity, 10 m winds, and surface pressure every 6 h, as well as accumulated radiation and precipitation every 3 h)	Direct interpolation on target grid from land restarts
Atmosphere	ERA40 (Uppala et al., 2005) for 1960–1978 start dates, ERA-Interim (Berrisford et al., 2011) for 1979–2018 start dates, and ERA5 (Hersbach et al., 2020) from 2019 onwards	Two ICs (derived from time-lagging perturbations, using 1 and 2 November)	Direct interpolation on a target grid from the atmospheric 3D state of temperature, specific humidity, and horizontal wind components
Ocean	CHOR (Yang et al., 2016) for 1960–2010 start dates and CGLORSv7 (Storto and Masina, 2016) for 2011–present start dates	Five ICs (from three realizations of the global ocean/sea ice reanalysis and two ICs from linear combinations of the former three ICs)	Direct interpolation on target grid from 3D state of temperature, salinity, and horizontal components of the ocean currents
Sea ice			Direct interpolation on target grid of sea ice temperature, sea ice volume, sea ice area, and snow volume

covers the year 2021 onwards. Lead years 1–5 and 6–10 consider, respectively, the periods 1961–2015 and 1966–2020. To verify the skill for SST, we rely on the Met Office Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset version 1.1 (Rayner et al., 2003), while, for 2 m air temperature (T2m), the CRU TS v4.05 dataset (Harris et al., 2021) is used. Precipitation is assessed by means of the GPCC Full Data Monthly Product Version 2020 (Schneider et al., 2020), while, for mean sea level pressure, the HadSLP2 dataset (Allan and Ansell, 2006) is used.

2.3 Verification metrics

Initializing decadal predictions from estimates of the observed states of the Earth system may generate spurious responses, since the climate model used to produce the simulations, after initialization, tends to drift towards its own attractor (mean climate), deviating from the observed climatology, which is a consequence of the model's systematic error (bias). This issue is particularly pronounced in the prediction systems adopting a full-field initialization strategy, as in the present case. The spurious drift can be removed a posteriori by subtracting a lead-time-dependent climatology at each grid point, assuming a constant drift throughout the time record (Goddard et al., 2013; Boer et al., 2016).

To evaluate the skill of the prediction system, both deterministic and probabilistic metrics are used. The anomaly correlation coefficient (ACC) and the mean square skill score (MSSS) are deterministic metrics, measuring the accuracy of the ensemble mean prediction in reproducing the observed variability over the 1961–2020 period targeted by the decadal reforecasts. More specifically, the ACC is a dimensionless measure evaluating the phase agreement between predicted and observed anomalies, ranging from -1 to 1 (Wilks, 2011). The MSSS, additionally, quantifies the magnitudes between the predicted and observed anomalies (Goddard et al., 2013). This metric evaluates the skill of the ensemble mean prediction with respect to a reference prediction.

Specifically, the MSSS is defined as follows:

$$\mathrm{MSSS_{HPO}} = 1 - (\mathrm{MSE_{HO}}/\mathrm{MSE_{PO}}), \quad (1)$$

where $\mathrm{MSE_{HO}}$ ($\mathrm{MSE_{PO}}$) is the mean square error evaluated for the initialized (uninitialized) ensemble mean against observations. The MSSS takes a maximum value of one (1.0), while it does not have a lower limit. Positive MSSS values mean more accurate predictions in the initialized runs, and one may speculate that the opposite is also true. However, since the MSSS is not symmetric around zero, the positive and negative MSSS absolute values do not have the same meaning in terms of variance.

Probabilistic skill scores provide a useful complement to deterministic metrics in assessing the quality of a prediction system. In this study, relative operating characteristic (hereafter ROC) score maps have been assessed for the hindcasts (Kharin and Zwiers, 2003; Wilks, 2011). Each grid point in these maps shows the area under the ROC curve, equal to the probability of a certain anomaly to exceed a specific threshold. When the ROC score approaches the perfect forecast (i.e., equal to one), then the DPS is able to discriminate the occurrence of predetermined events. On the other hand, no skill emerges when the score is close to 0.5. Here, we have considered three equiprobable categories, i.e., upper tercile, lower tercile, and between lower and upper terciles (neutral). Note that the ROC score outcome is not dependent on forecast biases (i.e., calibration; Kharin and Zwiers, 2003).

The DPS's ability to reproduce the dominant climate variability patterns is also tested, focusing in particular on the North Atlantic and North Pacific sectors. Decadal variability in the Atlantic region is well described by the AMV, estimated as the detrended anomalies of SSTs area-weighted over the North Atlantic basin, following the definition adopted in Trenberth and Shea (2006). The skill in predicting the NAO index is also tested using the definition in Li and Wang (2003; Fig. S1 in the Supplement).

We characterize the low-frequency variability in the Pacific basin through the IPO, which is, in turn, expressed in terms of the IPO tripolar index (TPI), accounting for the difference between the averaged SST anomalies over the equatorial zone and over the extratropical lobes of the IPO (Henley et al., 2015). At shorter timescales, the ENSO prediction is evaluated through the Niño 3.4 index, representing the spatially averaged SST anomaly over the respective region of the equatorial Pacific.

The statistical significance is assessed with a one-tailed Student's t test (Wilks, 2011), accounting for autocorrelation in the time series (Eq. 30 from Bretherton et al., 1999). The anomalies of the observations and the historical simulations are computed with respect to their climatologies (reference period 1981–2010). In the initialized runs, the climatology for each forecast year is computed considering the highest possible number of initialization years. This approach allows us to maximize the statistical robustness, even if the skill may depend on the targeted verification years.

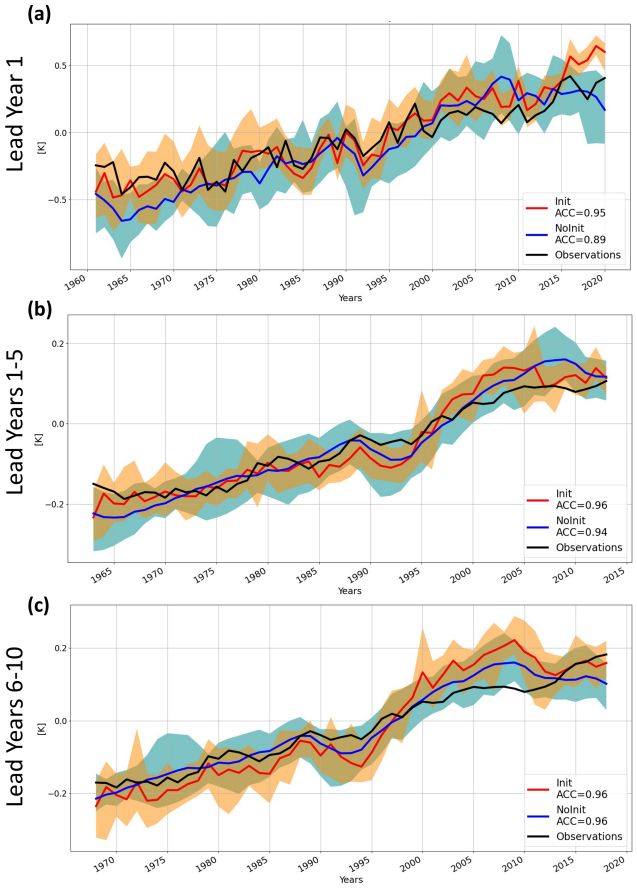

Figure 1. Global mean near-surface temperature (T2m plus SST) annual average anomaly time series (K) for the hindcast (Init; in red), historical and SSP2 scenario (NoInit, in blue), and CRU TS v4.05 and HadISST1.1 (in black), for **(a)** forecast years 1, **(b)** 1–5, and **(c)** 6–10. The orange (cyan)-colored envelope denotes the intra-ensemble spread for Init (NoInit). The time series are centered at the lead year interval, e.g., 1963 corresponds to the 1961–1965 mean in panel **(b)**.

3 Skill evaluation

3.1 Near-surface air temperature

Skill in predicting the global mean surface temperature (GMST; based on 2 m air temperature over land and SST over the ocean) is assessed against observed anomalies combining CRU TS v4.05 (Harris et al., 2020) over land and HadISST 1.1 for SSTs (Rayner et al., 2003). Figure 1 shows GMST for initialized hindcasts (in red; hereafter Init), non-initialized historical simulations (in blue; hereafter NoInit) and observations (in black).

At lead year 1, the initialized ensemble reproduces the observed GMST anomalies quite closely, showing a higher correlation (ACC = 0.95) compared to NoInit (ACC = 0.89), which is mainly explained by the strong impact of the imposed initial state at the beginning of the forecasts. Looking

at the lead year range 1–5, Init still resembles the observed variability (ACC = 0.96) within a range of 0.05 °C. Even if NoInit displays relatively high correlation (ACC = 0.94), its skill is substantially due to the global warming trend driven by external forcings, and in this case, the anomaly time series does not reproduce the observed interannual variability. The time evolution of the near-surface temperature over the 6–10 lead year range exhibits comparable correlations for both initialized and historical ensembles (ACC = 0.96), indicating that the radiative forcing has a dominant role at longer lead times.

The ensemble spread envelope of predicted GMST (denoting maximum and minimum range of the members variability; shown in orange) encompasses the observations, especially at lead years 1 and 1–5, successfully capturing the multiyear variability, including the cooling effect of major volcanic eruptions, such as the El Chichón and Pinatubo eruptions that occurred in 1982 and 1991, respectively. The initialization contributes to the reduction in the Init ensemble spread, which is about half the envelope of the NoInit for lead year 1, due to the beneficial impact of synchronizing observed and model internal climate variability.

3.2 Deterministic metrics

Predictive skill at the regional scale is assessed through ACC maps. Figure 2 shows ACC for annual surface temperatures evaluated at different lead year intervals and the corresponding differences with respect to NoInit. In order to remove the skill impact on different ensemble sizes, we compare the skill of the historical (10 available members) with the skill of random subsampling of the initialized run with 10 members out of the 20 available members. The MSSS maps provide further details on the skill improvement determined by initialization (Fig. 3), assessing the consistency between the magnitudes of the predicted and the observed anomalies (see Sect. 2.3).

At lead year 1, significant predictive skill is found over most of the globe, reaching the highest values (ACC = 0.80) over the tropical Indian Ocean, northern and equatorial Africa, northeastern part of South America, subpolar North Atlantic, and western tropical Pacific. Lack of skill, instead, characterizes the western subtropical North Atlantic, eastern Europe, central part of South America, and part of the western North Pacific and Southern Ocean. The added value of initialization (Fig. 2b) is particularly prominent over the tropical and the eastern subpolar North Atlantic and over the tropical and the extratropical North Pacific. In addition, Init exhibits higher skill (up to 0.5) over the North American continent, central Africa, and the Indian subcontinent. The corresponding MSSS pattern (Fig. 3a) clearly indicates that Init outperforms NoInit in reproducing the magnitude and the sign of the observed anomalies over approximately the same areas, showing improved ACC with respect to NoInit (Fig. 2b).

In the 1–5 lead year range, the skill is generally higher than for lead year 1, likely due to the effect of averaging over a longer interval (5 years) and to the emerging warming trend. In contrast, the skill undergoes a clear deterioration over the tropical and northern part of the Pacific Ocean when a multiyear range of prediction skill is considered (Fig. 2c). Significant skill is found over the continental areas of North America, Eurasia, Africa, and over the Maritime Continent. A large fraction of the skill seems to derive from the warming trend that increases predictability, at this lead year range, over land and over the Indian Ocean (Van Oldenborg et al., 2012). Over the North Atlantic Ocean, the emerging AMV footprint is recognizable, with high predictive skill associated with the typical horseshoe pattern emerging from the Init vs. NoInit comparison (Fig. 2d). This pattern is also noticeable in the relative MSSS map (Fig. 3c), suggesting improved predictability for the AMV tropical lobe, while the extratropical lobe may be affected by strong biases as it is characterized by high ACC values and neutral MSSS. Near-term prediction skill is improved especially over the eastern Mediterranean and the Arabian Peninsula (ACC = 0.3), reaching high correlation values (ACC = 0.90 in Fig. 2c) also reflected in the MSSS (Fig. 3c).

The pattern exhibited in the lead year range 6–10 is very similar to that shown in lead years 1–5, although some regional changes, such as those in eastern Europe and Siberian region, may be easily spotted. Areas with non-statistically significant skill cover part of the eastern Pacific Ocean (Fig. 2e). The generally higher skill attributable to initialization (Fig. 2f) is substantially consistent with the pattern obtained for the lead year range 1–5, even if it is not reproduced in the MSSS analysis (Fig. 3e), suggesting that surface temperature variations are not well captured.

To corroborate the skill analysis of surface temperature at decadal timescales, we assess the skill for the ocean heat content integrated over the top 300 m of the water column (hereafter OHC300). The ACC pattern computed for the OHC300 anomalies (Fig. 4) is similar to, and thus consistent with, the results obtained for the SST (Fig. 2). At lead year 1, significant ACC covers most parts of the oceans, except for the eastern Atlantic and Southern Ocean. The anomalies' values are also well captured north of 30° N. The OHC300 area exhibiting significant skill is reduced when higher lead time ranges are considered. At lead years 1–5 and 6–10, the ACC is significant over the tropical Pacific, excluding the equatorial band due to the poor long-term predictability of ENSO, as also found in other DPSs (e.g., Bilbao et al., 2021). Positive ACC values also cover part of the Indian Ocean and South and North Atlantic regions. The MSSS shows positive values mainly localized over the midlatitudes in the North Atlantic (Fig. S6) and is found to be quite consistent with SST MSSS (Fig. 3). The lack of skill over the subpolar gyre may be partly due to the erroneous representation of the AMOC in the DPS, altering the local ocean circulation and heat content. A complementary analysis reveals that

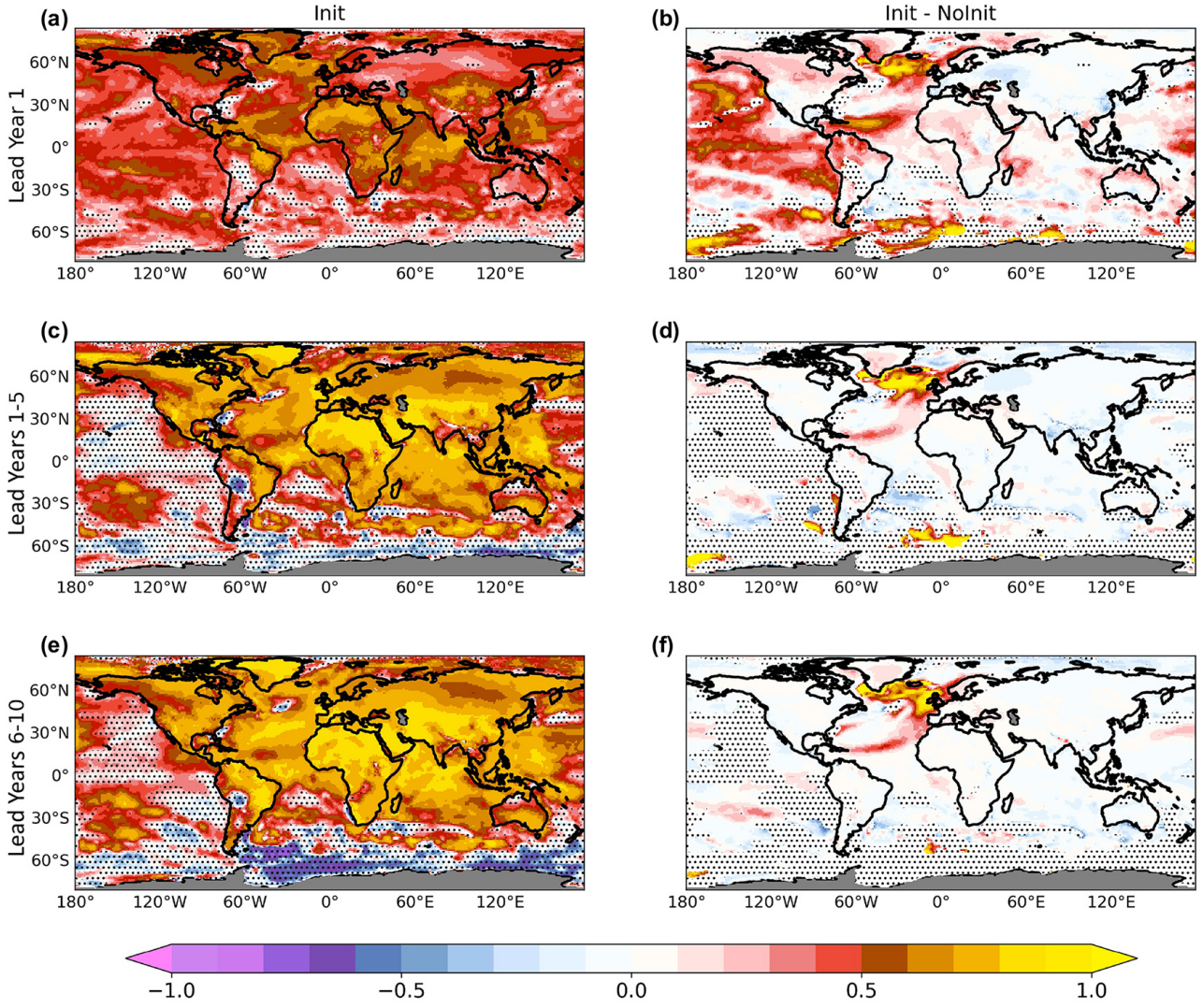

Figure 2. Near-surface temperature (T2m plus SST) anomaly correlation coefficient (ACC) of the hindcast ensemble (Init; **a, c, e**) and its difference with the NoInit ensemble (Init − NoInit; **b, d, f**) for lead years 1 **(a, b)**, 1–5 **(c, d)**, and 6–10 **(e, f)**. The stippling denotes points where 95 % statistical significance is not reached, according to a one-tailed t test. Effective degrees of freedom have been computed following Eq. (30) in Bretherton et al. (1999).

the mean AMOC cell in the DPS is quite well reproduced in terms of structure although its maximum is located too far south (below 20° N) at lead year 1 (Fig. S3), as compared to other AMOC reconstructions based on different oceanic reanalyses (e.g., Karspeck et al., 2017). At lead years 1–5 and 6–10 the maximum moves northwards, due to the model adjustment towards its own climatology, resembling the structure reported in other studies (e.g., Tsujino et al., 2020). The initialization shock may lead to the AMOC slowdown up to lead year 2 (Figs. S4 and S5), underestimating the maximum by about 2 Sv at 26.5° N in the period covered by RAPID array (Moat et al., 2022). The slightly negative trend of the observed AMOC occurring during the last few decades is reproduced just at lead year 1 in the hindcasts (Figs. S4 and S5), while the simulated low-frequency variability is consistent with the observed one also at longer lead years.

Compared to surface temperatures, skill in precipitation is generally lower and less spatially coherent (Collins, 2002; Doblas-Reyes et al., 2013). At lead year 1, significant skill is found only in limited areas (Fig. 5a), including northeastern Brazil, southwestern U.S., southern Africa, eastern Australia, the Republic of Türkiye, and the Balkan peninsula, as reflected also by the MSSS values (Fig. 3b). For the lead year ranges 1–5 (Fig. 5c) and 6–10 (Fig. 5e), significant ACC values can be attributed to the northern part of the Eurasian continent, the Sahel, and Europe, including the Iberian Peninsula, the British Isles, and central Europe. However, comparing Init with NoInit reveals that the skill is largely due to

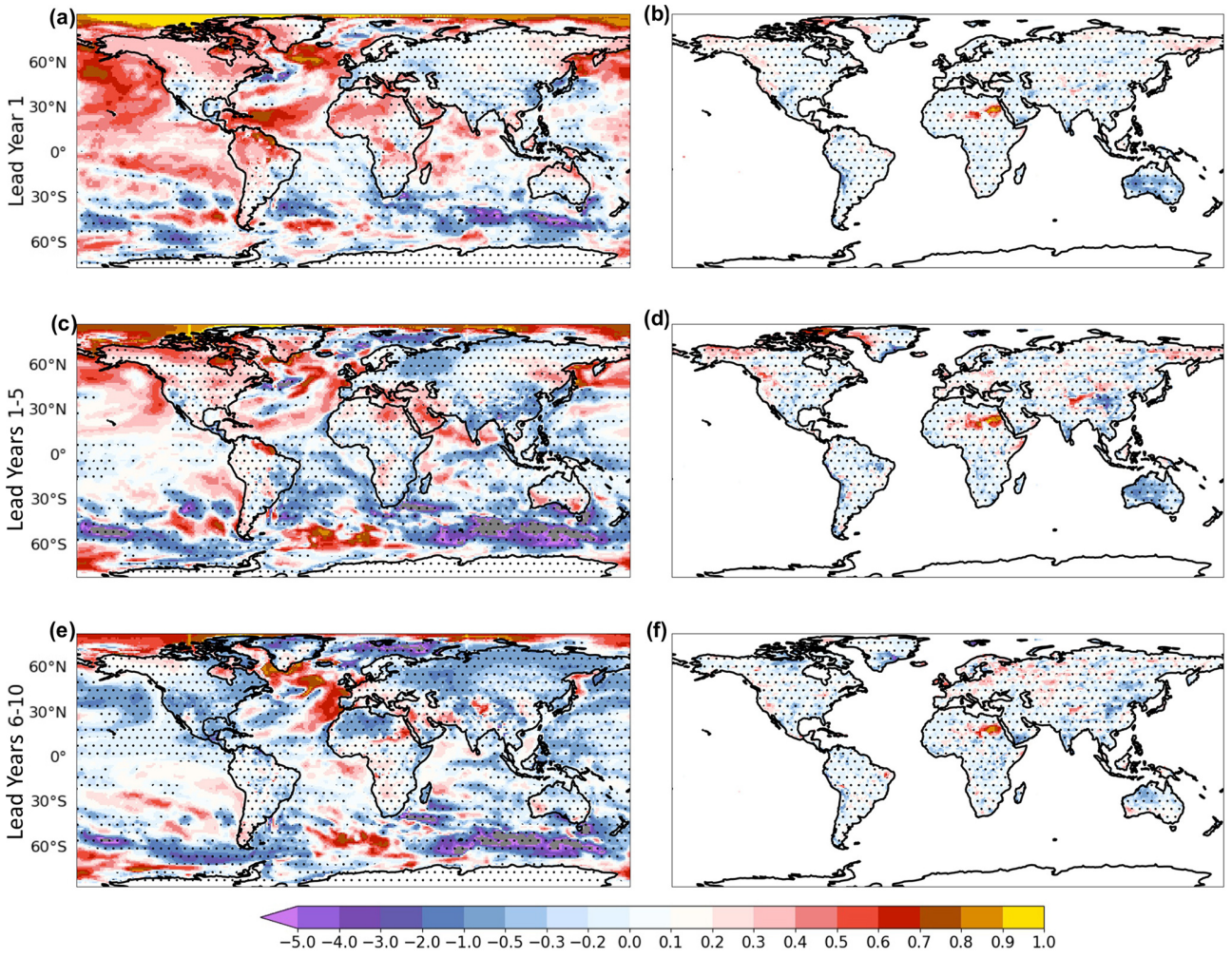

Figure 3. Near-surface temperature (T2m plus SST; **a, c, e**) and precipitation **(b, d, f)** mean squared skill score (MSSS) of the hindcasts, using NoInit runs as the reference forecast to beat. Note that the color bar is not symmetric around zero. Stippling is used to indicate points where 95 % statistical significance is not reached, according to a one-tailed t test. Effective degrees of freedom have been computed following Eq. (30) in Bretherton et al. (1999).

trends in the radiative forcing, with slight improvements associated with initialization (Gaetani and Mohino, 2013; Bellucci et al., 2015b).

3.3 Mean bias assessment

The full-field strategy is used to initialize the forecasts, providing the best estimates of the observed state to each model component. It does have an important drawback in that it generates spurious, transient signals determined by the model's tendency to drift towards its own climatological mean state after being initialized from a realistic state around the observed climatology.

Following the recommendation of the International CLIVAR Project Office (ICPO, 2011), the mean bias is defined as the lead-time-dependent ensemble mean deviation from the observed mean state defined throughout the whole time record (1960–2020). Assessing the mean bias is an important part of evaluating decadal predictions. The time-dependent SST bias for decadal hindcasts (lead years 1, 2–5, and 10) and the bias in the historical simulations are shown in Fig. 6.

The SST bias in Init is very rapidly established during year 1, followed by a slower adjustment occurring in the following years, since the Init curves of zonal mean bias for lead years 1, 2–5, and 10 remain relatively close to each other (Fig. 6e). Bias patterns featured by Init and NoInit differ substantially over the Northern Hemisphere, with the former presenting a prominent cold bias, which is not found at all longitudes in NoInit. In the Southern Hemisphere, Init and NoInit are much more similar. This lack of agreement between Init and NoInit suggests that initialization likely perturbs the equilibrium state of the model climate quite significantly. Interestingly, the same kind of departures from the observed state have been found also in several other