Systematic investigation of skill opportunities in decadal prediction of air temperature over Europe

Giovanni Sgubin (giovanni.sgubin@lsce.ipsl.fr)
EPOC - University of Bordeaux https://orcid.org/0000-0002-0190-0188
Didier Swingedouw
EPOC - University of Bordeaux
Leonard F. Borchert
LOCEAN - IPSL
Matthew B. Menary
LOCEAN - IPSL
Thomas Noël
The Climate Data Factory
Harilaos Loukos
The Climate Data Factory
Juliette Mignot
LOCEAN - IPSL

Research Article

Keywords: Climate Variability, Decadal Climate Predictions, De-biasing, Atlantic Multidecadal Variability, Climate Service

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-544705/v1

Abstract

Decadal Climate Predictions (DCP) have gained considerable attention for their potential utility in promoting optimised plans of adaptation to climate change and variability. Their effective applicability to a targeted problem is nevertheless conditional on a detailed evaluation of their ability to simulate the near-term climate evolution under specific conditions. Here we explore the performance of the IPSL-CM5A-LR DCP system in predicting air temperature over Europe, by proposing a systematic assessessment of the prediction skill for different time windows (periods of the calendar time, forecast years and months/seasons). In this framework, we also compare raw and de-biased hindcasts, in which the temperature outputs have been corrected using a quantile matching method. The systematic analysis allows to discern certain conditions conferring larger predictability, which we find to be intermittent in time. The predictions appear more skilful around the 1960s and after the 1980s, in coincidence with large shifts of the Atlantic Multidecadal Variability, which are well reproduced in the hindcasts. Averages on longer forecast periods also generally imply better prediction skill, while the best predicted months appear to be mainly those between late spring and early autumn. Moreover, we find an overall added value due to initialisation, while de-biased predictions significantly outperform raw predictions only for a few specific time windows. Finally, we discuss the potential implications of the proposed systematic exploration of skill opportunities in DCPs for integrated applications in climate sensitive sectors.

Introduction

One of the biggest scientific challenge for the 21st century concerns the capacity of simulating the future climate evolution through numerical models (Dutton, 2002). Climate change and variability is threatening many natural ecosystems (e.g. IPCC, 2014), and is having a progressively stronger impact on human society by affecting a wide range of sectors (Arent et al., 2014), like agriculture, fishery, energy, tourism, and transport to name but a few. Knowing a few years in advance how the climate will vary with a certain degree of accuracy is crucial for stakeholders and policymakers (Trenberth et al., 2016), as it potentially allows for well-timed adaptation efforts. Predictions up to 10 years ahead, hereafter Decadal Climate Predictions (DCP), are thus increasingly demanded (Kushnir et al., 2019) as they deal with time ranges typically relevant for infrastructures, long-term investments and other business plans (Dessai and Bruno Soares, 2015). For this reason, both public and private sectors are growingly fostering the development of operational climate services based on near-term predictions (Vaughan and Dessai, 2014; Buontempo et al., 2014; Street, 2016). Climate services aim at providing customized products for decision makers in climate-sensitive sectors, thus creating a bridge between the academic world and the end-users by translating scientific outcomes into targeted information (Goddard, 2016; Giannini et al., 2016). Yet, the development of a climate service needs to be based on trustable climate predictions (Lemos et al., 2012, Mehta et al., 2013), and so its effective operability requires a scrupulous evaluation of the actual skill of the existing prediction systems in simulating the near-term climate evolution, notably in those specific contexts that are relevant for the targeted analysis (Bruno Soares et al., 2017). As an example, let us consider the tourism sector: a climate service for ski resorts in the Alpine regions should demonstrate

skilful predictions of snow fall during the winter in mountains areas, while a climate service for seaside activities in the Mediterranean region should rather demonstrate skilful prediction of temperature and precipitation in coastal areas during summer. This implies that forecasts based on the same DCP system may show a wide range of confidence when applied to these very different scopes, as the prediction skill considerably depends on a multitude of contingent factors like the relevant climatic variable, the specific season, the particular period, and the region under investigation. The rationale behind this work is therefore to propose a prototype for the individuation of the optimal "opportunities" for integrated applications of DCPs. Here, we explore the potential expressed in this respect by the IPSL-CM5A-LR model in decadal predictions of air temperature over Europe, but the same approach might be potentially used to evaluate the potential reliability of other DCP systems in simulating any other climatic variables. In this context, the Copernicus Climate Change Services (https://climate.copernicus.eu/) will soon include operational decadal predictions implemented by different Institutes, which will possibly allow for an optimal selection of model experiments as a function of the specific study.

The way in which the future climate is simulated depends on the considered time scale. On short time scales in the order of days, i.e. O(days), meteorological forecasts are so-called initial condition problems since the prediction primarily depends on the internal state of the climate system at the beginning of the prediction. On long time scales, i.e. O(century), climate projections are essentially so-called boundary condition problems since they primarily depend on the response to an external forcing. The forcing includes both anthropogenic changes in atmospheric greenhouse gas and aerosol concentrations and natural changes such as modulations of the solar insolation and volcanic eruptions. In between these time ranges, near-term DCP are a mix of initial and boundary conditions problem (Murphy et al., 2010). Indeed, the climate evolution over a time horizon of 1–10 years is basically the combined result of (i) the external forcing, and (ii) the unforced internal variability, coming from the intrinsic variations of the climate system. Over Europe, the climate variability on decadal timescales is largely modulated by the Atlantic Multidecadal Variability (AMV) and the North Atlantic Oscillation (NAO). The AMV (Kerr, 2000; Sutton and Hudson, 2005) is a mode of climate variability affecting the Sea Surface Temperature (SST), characterised by fluctuations between anomalously warm and anomalously cool phases, with enhanced energy in the inter-decadal band (Dima and Lohmann, 2007). While the AMV appears to be linked with the variability of the Atlantic Meridional Overturning Circulation (AMOC) (Yeager and Robson, 2017, Oelsmann et al., 2020), its main drivers are still uncertain. Indeed, over the historical era, AMV-like decadal fluctuations can be potentially attributable to intrinsic variability in absence of external forcing (Knight et al., 2006; Ting et al., 2009), or to a response to external forcing, e.g. changes in aerosol and greenhouse gas concentrations of anthropogenic origin (Booth et al., 2012, Bellucci et al., 2017) and natural events like volcanic eruptions (Ottera et al., 2010, Swingedouw et al. 2015, 2017, Borchert et al. 2020). The switch between positive and negative phases of the AMV causes significant climatic impacts over Europe, leading to distinct mean temperature and precipitation patterns, most notably during summer (Sutton and Hudson, 2005). Aside from the AMV, the NAO is an atmospheric mode of variability of the flow patterns over the North Atlantic Ocean, which has important impacts on the weather and climate in Europe, notably during winter (Hurrel et al., 2013). The NAO variability has a weak red spectrum (Wunsch,

1999), and it is characterised by statistically significant decadal variations, which in general appear to be out of phase with the AMV signal (Li et al., 2013; Omrani et al., 2013). These modes of decadal variability overlap with the long-lasting warming signal due to increased greenhouse gas emissions, thus shaping the near-term climate variations by either amplifying, smoothing or even inverting the long-term trend.

DCPs are designed to account for both internal modes of variability and the effects of changing anthropogenic emissions and natural phenomena. They consist of $O(10)$ years experiments forced by external boundary conditions, and starting from a specific climate state constrained by observations, which represents the imposed initial conditions of the dynamical system. This initialisation supplies the main potential added value of DCPs with respect to climate projections (Boer et al., 2004). A variety of techniques and methodologies of initialisation have been developed by the different modelling groups, influencing the predictability of a specific phenomenon (Matei et al. 2012; Menary et al., 2015). In general, DCPs mostly rely on the large thermal inertia provided by the ocean for the climate system, which may provide a “memory” of O(10) years at the surface and of $O(10^3)$ years at depth. Initial conditions are primarily obtained by assimilating “full-field” or “anomalies” of a set of oceanic parameters from observational data (Smith et al., 2013). For example, the IPSL-CM5-LR DCP system (Swingedouw et al., 2013, Mignot et al. 2016) is based on initial conditions constructed from observed SST anomalies, while both SST and Sea Surface Salinity (SSS) anomalies are used for the IPSL-CM6A-LR DCP (Estella-Perez et al., 2020). More sophisticated methods of initialisation also imply the assimilation of atmospheric variables (see Meehl et al., 2014 for a detailed summary of the different methods of initialisation used within CMIP5 models). By exploiting the slowly evolving interactions within the climate system, the synchronisation of the model initial conditions with the actual climate state might constrain, at least for a certain time window, the stochastic evolution of the climate system in the simulation. Therefore, the key question in the evaluation of DCP is whether the initialization process is able to effectively align the phase of the modelled and observed internal variability, thus producing in principle more skilful simulations than climate projections over a time scale of one decade.

Prediction skill is typically assessed by comparing the hindcasts, i.e. retrospective predictions initialised at a given past climate state, with the corresponding observational data over the common period. Pioneering studies showed the potential skill improvement through initialisation (Collins et al., 2006; Smith et al., 2007, Keenlyside et al. 2008), thus favouring the development of a coordinated protocol for decadal prediction systems within the Coupled Model Intercomparison Project (CMIP5, Taylor, 2012). This first multi-model approach allowed a more robust assessment of the impact of the different initialization strategies. In this framework, added value with respect to climate projections have been clearly shown for the 1–10 years predictions of different climatic variables over different regions (van Oldenborgh et al., 2012; Kim et al., 2012; Doblas-Reyes et al., 2013; Bellucci et al., 2015), and notably over the North Atlantic (Branstator and Teng, 2012), where both the AMV and AMOC show large potential predictability (Garcia-Serrano et al., 2015). Also, it has been shown that predictions based on a multi-model ensemble mean are more skilful than predictions with individual models (Bellucci et al., 2015). However, multi-model analyses and inter-model comparisons imply a standard procedure for DCP evaluation, which may prevent a

detailed exploration of the full range of potential predictability shown by a single DCP system as well as gaining understanding in the physical sources of predictability. Indeed, in previous studies, the skill metrics are typically (but not exclusively) calculated on an annual basis, i.e. by considering the yearly means, over standard lead-time years - usually 1–5 years and 6–10 years - and for the whole period of the available hindcasts, which typically start in 1960 (e.g. Mignot et al. 2016). Yet, the prediction skill can be strongly dependent on (i) the specific period considered, (ii) the prediction lead-time used and (iii) the specific seasons of the year in focus. In this regard, the notion of “windows of opportunity” has been recently adopted for DCPs (Dessai and Bruno Soares, 2013; Mariotti et al., 2020), pointing out that specific periods in the past may confer greater predictability than others (Brune et al., 2018, Borchert et al, 2019, Mariotti et al., 2020). Furthermore, the predictability of a specific phenomenon has been shown to be strongly dependent on the region and on the considered time scale (Boer, 2004; Boer, 2011; van Oldenborgh, 2012). At the same time, the predictions’ quality may be strongly dependent on the considered season (e.g. Yeager et al., 2018). For example, over Europe, as the summer climate is mainly modulated by the AMV, while in winter it is largely influenced by NAO variations, the different model abilities in reproducing and predicting these two modes of variability and the associated teleconnections may yield to different predictability over Europe for the different seasons. In turn, the predictability of these modes of variability has been shown to be linked with the background climate mean state and variability (Qasmi et al., 2017), leading to an enhanced prediction skill over Europe for specific periods and forecast times. The confluence of factors implying a larger DCP skill may enhance its usability for specific contexts, with implications for the optimisation of climate services. While the existence of some of these factors has been identified in previous studies (e.g. Qasmi et al., 2017; Yeager et al., 2018; Mariotti et al., 2020), their systematic evaluation is missing so far.

The scope of the present study is to assess the potential offered by the DCP system based on the IPSL-CM5A-LR model in predicting air temperature over Europe. Previous studies using predictions with IPSL-CM5A-LR DCP system already showed skill in predicting the AMOC (Swingedouw et al., 2013) and the AMV (Mignot et al., 2016). Due to their influence on the European climate (e.g. Persechino et al 2013), this suggests the likely existence of some skill also in predicting the near-term temperature over the continent. Here, we explore the possible existence of specific time windows for which the DCP exhibits a higher predictability. These potential windows of opportunity are defined by particular time periods, whose background climate mean state and variability can favour the DCP predictability. Moreover, we analyse the benefits of the initialisation procedure and its time duration, as well as the impact of seasonality on predictability. In this way we extend the notion of windows of opportunity to forecast periods and months of the year.

The expected outcomes of the study are to provide supporting information for the development of climate services. In this framework, potential limitations are the systematic biases in mean state and variance that intrinsically affect the climate models (e.g. Haerter et al., 2011; Maraun, 2012, Bilbao et al., 2021), which can produce imprecise assessments of climate evolution and its impacts. To address such a potential limitation, different bias-adjustment techniques have been developed (e.g. Michelangeli et al., 2009). De-biasing consists in adjusting raw model data to calibrate their statistical properties with those

of the corresponding observational data. Its benefit has been already tested for long-term impact analysis of climate change over specific regions, e.g. West Africa (Famien et al., 2018). When applied to DCP, this data adjustment may also potentially have implications on the prediction skill. For this reason, we also systematically compare, for the first time, raw and de-biased hindcasts, and we evaluate whether the data adjustment, beyond correcting the mean state bias of the predictions, is also beneficial in terms of prediction skill. Our approach, detailed in Sect. 2, is based on a systematic analysis of the prediction skill of raw and debiased hindcasts when simultaneously varying the initialisation periods, the prediction lead times and the predicted months of the year. The main features of skill over Europe and its 7 sub-regions are illustrated in *Sect. 3*, as well as the pattern of the added-value due to initialisation and the skill improvement due to de-biasing. In Sect. 4 we finally discuss the potential implications of our main findings and stress the utility of this type of analysis for the optimal development of climate services based on DCPs.

Methodology

2.1 The climate model

The decadal predictions analysed in this study have been performed with the IPSL-CM5A-LR model (Dufresne et al., 2013) developed at the Institut Pierre Simon Laplace (Paris). It is a global general circulation model consisting of the coupling between atmospheric and oceanic systems. The atmospheric component is based on the LMDZ5A model (Hourdin et al., 2013) which, for the Low Resolution (LR) configuration, consists in 96 x 95 grid points corresponding to a resolution of 3.75 x 1.875 and 39 vertical levels. The ocean component is based on the NEMOv3.2 model (Madec, 2008), with a horizontal resolution varying from 0.5◦ to 2◦ and 31 depth levels varying from 10 m thickness near the surface to 500 m at depth. The model also includes the sea-ice module LIM2 as well as the biogeochemical module PISCES (Aumont and Bopp, 2006). The IPSL-CM5A-LR model has been set to produce both an ensemble of climate projections and an ensemble of decadal predictions. The latter employs the progressive imposition of initial conditions, which have been produced by means of a simple assimilation technique consisting of nudging to observed surface SST anomalies (cf. Mignot et al. 2016). It is important to specify that, while decadal predictions with the latest IPSL-CM6A-LR model version have been recently released, this latest version was not available at the time of this study, when de-biasing adjustment was carried out. Since one of the aims of this study is to assess the potential improvement due to a quantile de-biasing adjustment, our analysis here is exclusively based on the IPSL-CM5A-LR version.

2.2 Simulations and validation dataset

We analyse monthly averaged temperature from a set of 2 different model experiments, namely (i) the non-initialised historical experiments (HIS), and (ii) the initialised decadal predictions experiments (DCP). For the skill metrics calculation (see Section 2.5), we compare these temperature model outputs with observation-based temperature data (OBS), *i.e.* NOAA-20CR reanalyses data (Slivinski et al., 2019),

interpolated on the model grid. NOAA-20CR is a global gridded data set consisting of 56 different members. Here we use their ensemble mean. While reanalysis data imply the use of a model and thus the possible inclusion of errors due to intrinsic model biases, it is worth stressing that NOAA-20CR temperature data over Europe are significantly consistent with other observational data not implying the use of a model, e.g. the correlation with HadCRUT4 interpolated temperature observational dataset is 0.95 ($p<0.05$) for 1960-2014 de-trended monthly anomalies averaged over Europe (not shown).

For both simulations and validation datasets we consider monthly temperature anomalies, which have been computed by removing the corresponding monthly climatology calculated over the common period 1961-2014.

The historical experiments (HIS) are extracted from the CMIP5 database and consist of 3 members running from 1850 to 2005. The initial conditions are obtained randomly from a 1000-year control simulation based on stationary preindustrial climatological forcing. The HIS external boundary conditions consist of the prescribed radiative forcing estimated from observed aerosol and greenhouse gases concentrations in the atmosphere since 1850 as well as changes in ozone and land-use, and the effects of solar radiation and volcanic eruptions over this period. The HIS ensemble will be referred to as “non-initialised”, since it does not start from an observed climate state. In conformity with the hindcasts, we only consider the part of the experiments after 1961. Moreover, after 2005, we prolonged this experiment until 2014 by using the RCP4.5 scenario (Taylor et al., 2012).

The initialised DCP experiments are directly derived from nudged experiments, which are based on a data assimilation aimed at adjusting the SST anomalies towards observational *i.e.* ERSST (Reynolds et al., 2007), anomalies. The nudged experiment is a “constrained” experiment in which, under the same boundary conditions as in the HIS experiments, a SST anomaly term is added into the conservation equations for SST to adjust the heat flux at each model time-step t^. The restoring term is expressed as follows:

$$Q(t^{\wedge})=-\gamma[SST'_{MOD}(t^{\wedge})-SST_{ERSST}(t^{\wedge})]$$

where $\gamma=40 Wm^{-2} K^{-1}$ is the restoring coefficient corresponding to a relaxation time-scale of around 60 days for a mixed layer of 50 m, SST'_{MOD} is the modeled SST anomaly, and SST'_{ERSST} is the measured SST anomaly with respect to the climatological mean obtained from ERSST over the overlapping period. From the constrained nudged simulations after 1960, ensembles of 3 members of 10-year free-running simulations have been launched from December 31 of every year until 2013, without any constraint applied on the SST. These free-running simulations are still constrained by the external forcing as in the corresponding portions of the HIS experiments. The DCP members are separated by adding a white noise perturbation to the SST field at the time of initialization, chosen randomly at each grid point between -0.05 and 0.05 °C, thereby mimicking the unpredictable part of the climate signal. This specific protocol is described in further details in Mignot et al. (2016).

2.3 De-biased predictions

In this study we analyse both raw temperature predictions and de-biased temperature predictions obtained through a bias adjustment of the raw DCP dataset. Here we use adjusted data provided by "the Climate Data Factory" (https://theclimatedatafactory.com/), whose de-biasing procedure relies on the Cumulative Distribution Function transform (CDF-t) method (Michelangeli et al., 2009, Vrac et al., 2016, Famien et al., 2018). It is based on the quantile mapping method (Vrac et al., 2015), consisting of an adjustment of the raw simulated temperature through a transfer function, such that its cumulative distribution function (CDF) matches the observation-based one over a calibration period. Also, the CDF-t represents a variant of the quantile-quantile method as it also accounts for changes of CDF between the correction period and the calibration period (Michelangeli et al., 2009), and was already adopted and validated for various applications (Oettli et al., 2011; Lavaysse et al., 2012; Vautard et al, 2013; Vigaud et al., 2013). For the present evaluation, the reference data for the calculation of the transfer function have been obtained by interpolating the NOAA-20CR reanalysis (Slivinski et al., 2019) over the IPSL-CM5A-LR spatial grid. Moreover, in order to make the different periods used for skill assessment completely independent on the calibration period of the de-biasing process, data correction has been performed separately for two different periods, whose transfer functions were, in turn, calculated over two independent periods, e.g. 1961-1987 and 1988-2014 for lead-time of 1 year.

2.4 Data organisation

While HIS and OBS datasets are continuous time-evolving data, raw and de-biased DCP experiments are multiple 10-year simulations starting every year. Therefore, in order to make all the datasets conform to each other, HIS and OBS datasets have been first organized to mimic the DCP outputs, by decoupling them as multiple 10-year pseudo-predictions according to the start dates. In this way, the OBS temperature over a generic year Y, corresponds to the OBS pseudo-prediction starting from year Y-LT, where LT is the lead time. For example, the OBS temperature in 1981 corresponds to the OBS pseudo-prediction initialised in 1980 for LT=1 year, in 1979 for LT=2 years and so on. Finally, over the common period 1961-2014, we compute all the dataset as continuous time-evolving monthly temperature in function of their individual lead-time years LT and months M. The resulting time-series, hereinafter named principal time series, are at the base of the systematic calculation of skill scores (see Sections 2.6 and 2.7). Indeed, for each dataset, the different linear combinations of these 120 principal time series, along with the selection of the period of initialisation P, allows to define all the possible combinations of time windows, i.e. forecast periods and multi-months periods (see the table in the Appendix and further discussion in Section 2.7).

2.5 De-trending

In order to exclude the contribution of the long-term radiative forcing from the skill assessment, which explains a large part of the skill in decadal forecasts (van Oldenborgh, 2012), our analysis is exclusively based on de-trended datasets. For each dataset, we remove a linear trend calculated over the overlapping period from all the principal time series defined above. Note however that both the external radiative forcing and the climate response are in reality non-linear (see discussion in Garcia Serrano et al., 2015),

so the residual signal represents just an approximation of the un-forced component of the near-term temperature evolution.

2.6 Skill metrics

The skill evaluation is based on the comparison between the monthly temperature anomalies of the simulations dataset, i.e. HIS and DCP, and validation dataset, i.e. OBS (see Section 2.2). Here we define the prediction skill scores by means of two different verification metrics, namely the anomaly correlation coefficient (ACC), and the root mean square error (RMSE). For the comparison of datasets that have been averaged over different time windows, the latter metric is calculated after having standardised the variance of the time series. Indeed, the averaging process over different forecast times and different months corresponds to a linear combination of different so-called principal time series, so that the variance of time series averaged over longer time windows (e.g. for LT=2-9 years, for given M and P) is, on average, intrinsically lower than the variance of time series averaged over shorter time windows (e.g. for LT=2-3 years, for given M and P). This prevents the direct comparison of the RMSE calculated over different LT and M, as their difference may be just the result of this numerical artefact. Therefore, we adjust all the time series that are a combination of different principal time series such that their variance matches the mean of the variances of the principal time series composing it. In this way, the RMSE calculated for these scaled time series (which hereinafter we will refer to as RMSE*) is not affected by the intrinsic differences of variance due to the averaging process. Note that this manipulation is not necessary for ACC, as its calculation implies a standardisation of the data.

The statistical significance of the ACC metric is evaluated through a one-sided Student's t-test, for which the effective degrees of freedom have been calculated taking into account the serial autocorrelation (Bretherton et al. 1999). The test on the significance of the difference between two correlations values is based on the Fisher z-transformation. Finally, the statistical significance of the difference between the RMSE from two different datasets is evaluated through the Welch's t-test.

2.7 Systematic analysis

The de-trended principal time series for DCP, HIS and OBS datasets defined in Sections 2.4 and 2.5 are at the base of the systematic analysis of the DCP skills. Starting from them, (i) we partition the whole period of initialisation 1960-2012 to obtain 28 different 26-year moving initialisation periods P; (ii) we combine the 10 different individual forecast years to obtain 55 different combinations of consecutive (single or multiple) lead-time windows LT, i.e. the so-called forecast periods; (iii) we extract all the 78 different combinations of consecutive predicted months M. This procedure defines a three-dimensional matrix of time windows accounting for all the combined configurations of P, LT and M, hereinafter referred to as contexts (see table in the Appendix for their definition). The systematic approach proposed here is aimed at evaluating the temperature prediction skill score S of DCP (both raw and de-biased) and HIS for each of the defined contexts. In other words, we analyse the function S=f(M,LT,P), where the time windows P, LT and M are considered as independent variables. The choice of 26-year moving periods for the P variable is justified by the fact that it is the largest length of years for which the first initialisation period (1960-

1985) has no common time step with the last initialisation period (1987-2012). Results with different moving periods (35-year and 44-year lengths) have been also analysed, and are qualitatively similar to those that will be presented here (not shown).

The skill metrics S in the systematic analysis have been calculated after having spatially averaged the temperature over different specific regions, i.e. over whole Europe (23°W-65°E, 33°N-70°N), and 7 European sub-regions, namely Scandinavia (4°E-32°E, 57°N-70°N), Central Europe (4°E-32°E, 44°N-57°N) North-eastern Europe (32°E-55°E, 50°N-70°N), North Atlantic sector (23°W-4°E, 44°N-67°N), Iberian Peninsula (12°W-4°E; 35°N-44°N), Mediterranean (0°E-32°E, 35°N -44°N), South-eastern Europe (32E-55E, 35N-50N). This procedure produces a total of sixteen 3-dimensional matrices (8 regions for two skill metrics) of 120,120 skill values (78x55x28) corresponding to each possible combination of time windows. From these matrices, we extract the conditions exhibiting the best prediction skill, and then re-calculate the skill scores for each grid point of the European domain for that context. Finally, we compare these best prediction skills with the skill calculated for a reference context, which is defined by following the approach of a classical, non-optimised, assessment of DCP skill, i.e. LT=1-5 years and M=Jan-Dec, and, arbitrary, for P=1960-1985.

Results

3.1 Prediction skill for a reference context

To establish a reference point in our systematic analysis, we evaluated the skill of the IPSL-CM5A-LR model in predicting air temperature over Europe for the reference context. In Figure 1 we thus show the spatial distribution of the ACC and RMSE scores for the (i) period P identifying predictions initialised every year from 1960 to 1985, (ii) forecast time from the first to fifth year of prediction (LT=1-5 years) and (iii) predicted annual temperature means (M=Jan-Dec). Non-initialised historical simulations (Figs. 1a,1d) exhibit only limited skill concentrated over the Mediterranean sector, although the ACC is not statistically significant at the 95% level (and thus not visible in Fig. 1a), as for the rest of Europe, which is characterised by both low ACC and relatively high RMSE. Raw predictions (Figs. 1b,1e) clearly appear more skilful than the historical experiments. Added value with respect to HIS simulation is statistically significant over most of the land surface north of 45°N when ACC is considered, while RMSE is significantly lower in DCP than in HIS over three main spots: the U.K., the central part of Europe and the region between the Black and Caspian Seas. The ACC becomes significantly positive over most of the Central sectors of Europe, also including the southern part of Scandinavia as well as the peninsular part of Italy, the Balkans and the regions surrounding the Black and Caspian Seas. Finally, de-biased predictions (Figs. 1c, 1f) exhibit a further improvement of the skill with respect to the raw predictions. The ACC skill for the de-biased DCP is generally higher than for the raw DCP, although such an improvement is statistically significant just for a few grid points, e.g. in Scandinavia and in the Hellenic Peninsula. At the same time, the RMSE for the de-biased DCP is generally lower than for the raw DCP, notably over Scandinavia where this difference is statistically significant.

To extend the detection of the skill opportunities beyond this reference context, we systematically calculate the skill metrics for different time windows, i.e. for different combinations of P, LT and M. We first focus this analysis on the raw DCP dataset, while we extensively evaluate the added value due to initialisation and the effects of de-biasing in Section 3.4, where the performance of the raw DCP have been systematically compared with the performance, respectively, of the HIS simulations and of the de-biased hindcasts.

3.2 Skill at varying P, LT and M

We now analyse how the prediction skill illustrated in Fig. 1 changes when the independent variables P, LT and M are successively varied (Fig.2). For this, we consider time series of air temperature averaged over Europe (see Section 2.7 for the definition of its boundaries) and we calculate ACC and standardised RMSE* (see Section 2.6) for all possible combinations of consecutive P, LT, and M. For the fixed standard LT=1-5 years and M=Jan-Dec, ACC skill is statistically significant for most of the 26-yr initialisation periods P (Fig. 2a), but for those starting from 1972 and 1979. For P=1972-1997 ACC score is the lowest (0.26, $p>0.05$) while best ACC is found for P=1961-1986 (0.52, $p<0.05$). Concomitantly, RMSE varies between 0.75 and 0.9 (Figs. 2a, 2d). The modulation of skill on P (Figs. 2a, 2d) appears to be relatively less marked than the one on LT (Figs. 2b, 2e) and M (Figs. 2c, 2f), at least for the time windows analysed so far. This partly reflects the fact that the different initialisation periods P largely overlap. For P=1960-1985 and M=Jan-Dec (Figs. 2b, 2e), the best ACC along the LT axis (for P=1960-1985 and M=Jan-Dec) is found for LT=1-8 years where it reaches 0.74 ($p<0.05$), while the worst ACC skill is found for LT=7 years where it is -0.24 ($p>0.05$). Similarly, the best RMSE* is also found for LT=1-8 years where it measures 0.59, while the worst RMSE* skill is found for LT=10 year where it is 1.17. On the M axis, the evolutions of ACC and RMSE* (Figs. 2c, 2f) appear as a sequence of parabolic-like curves with vertexes centred over those combinations including the late spring and early autumn months. Best ACC and RMSE* scores along the M axis (for P=1960-1985 and LT=1-5 years) are found for predicted months M=May-Sep when they respectively measure 0.78 ($p<0.05$) and 0.42, thus evidencing a certain degree of conformity between the two metrics used. The worst ACC skill is found for predicted months M=Dec, for which it is -0.22 ($p>0.05$), while the worst RMSE* skill is found for predicted months M=Jan, for which it is 1.51.

A more comprehensive view on the function S=f(P,LT,M) is given in Fig. 3, where two of the independent variables are changed while the third is held constant. These two-dimensional representations qualitatively confirm most of the features shown in Fig. 2. The predictions appear in general more skilful over the extended summer season (from late spring to early autumns) and over longer forecast periods (Figs. 3a, 3c, 3d, 3f). Furthermore, for forecast periods implying an average over the same number of years, the skill appears to be unsurprisingly larger for those lead times that are closer to initialisation time, *e.g.* LT=1-5 years shows better skills than LT=6-10 years. When considering the summer months (Fig. 3b), there is significant ACC skills for the 26-yr periods starting around 1965 and 1980, punctuated by a skill degradation for P beginning between about 1970 and 1980. For example, for LT=1-5 years and M=May-Sep, the ACC is 0.77 ($p<0.05$) for the period 1961-1986, 0.17 ($p>0.05$) for the period 1972-1997 and 0.65 ($p<0.05$) for the period 1982-2007 (Fig. 3b). For the same periods P, the RMSE* is respectively 0.41, 0.74

and 0.51 (Fig. 3e). When fixing M=Jan-Dec (right panels of Fig. 3), higher ACC and RMSE* skills around 1965 and 1980 are also found for those contexts implying longer time averages and including the first prediction years (Figs. 3c, 3f). In general, the peaks and troughs of skill in the three-dimensional matrix identify clusters of points characterised by high or low predictability, as all their adjacent points exhibit similar scores. This feature gives confidence in the robustness of the results. The modulation of the predictability for various initialisation periods will be interpreted in the light of the background variability in Section 3.3.

Overall, Fig. 3 evidences that, for specific time windows, the prediction skill is significantly higher than for the reference context (cf. black circles in the Fig. 3). In turn, the prediction skill for the reference context is also significantly higher than the skill for other specific time windows (cf. crosses in the Fig. 3). Yet, only part of all the possible 120,120 time windows are shown in Fig. 3. From the three-dimensional matrix of skill scores over the full European domain, it is possible to extract numerically the conditions of best performance for the prediction of near-term air temperature. When the entire continent and all the possible combinations of P, LT and M are considered, the best skill is found for predictions over the period 1980-2005, for LT=1-9 years and considering the months of the year from June to October. This is largely consistent with the previous findings evidencing that prediction skill scores are higher for an extended summer season, and for forecast periods of several years. Specifically, at the grid point level (Fig. 4a), the ACC scores for this optimal configuration are generally higher than the scores under the standard context (Fig. 1b), with the area characterised by a significant correlation showing an expansion over the Atlantic sector of Europe and part of Iberian Peninsula. Indeed, over these regions, the ACC score is significantly higher than the ACC for the reference context. Most of the western and central part of the continent shows a significant ACC, while poor skill persists over the north-eastern regions and over the southern part of Iberian Peninsula.

It is important to note that the procedure of averaging the air temperature over a relatively large area gives an indication of the predictability over Europe, yet does not allow capturing all regional features of such predictability. Therefore, following the same systematic approach used so far for the whole European region, we additionally considered 7 different sub-regions of Europe (Fig. 4b-4h). Such partition shows that significant skill can be found everywhere in Europe under certain conditions of P, LT and M (Fig. 4b-4h). The predictability is significantly dependent not only on these independent variables, but also on the specific area considered, as the conditions for the best skill vary from one region to another. Nevertheless, a common feature is that their best performance coincides with the simulation of summer months. Indeed, the conditions for the best ACC always include at least the months from June to September for all the selected regions. Also, apart from north-eastern Europe, the best skill is associated with forecast periods averaged at least over 7 years and including the first lead-time years. This agrees with what was already shown in Fig. 2 and Fig. 3. We interpret this feature as due to (i) the larger impact of initialisation for short lead times and (ii) to a better imprints on climate of the oceanic variations and external forcing for longer forecast periods. Nevertheless, some region may show unexpected higher skills for long lead times or for predictions averaged over just a few years, which are possibly linked to a delayed atmospheric response or a response to external forcing. Such a re-emergence of skill has been, for

example, illustrated in the oceanic context by Matei et al. (2012) and Brune et al. (2018). This could possibly explain the peculiar peak of skill found for LT=3-7 years over north-eastern Europe (Fig. 4d), whose robust interpretation, however, would need a dedicated study.

3.3 The relationship between skill and the AMV

In Fig. 3b, we showed a clear pattern of ACC skill score dependence on P for the whole European region, with higher prediction skills occurring for periods starting outside of the early 1970s. Such skill modulation along P appears more marked when M belongs to the central part of the year. This suggests a possible link with the predictability of the AMV, as the latter has been shown to have its largest impact on European temperature during the summer (Sutton and Hudson, 2005). In Fig. 5 we compare the skills in predicting the AMV over the different combinations of 26-yr periods P with the pattern of S=f(P,M) for all the 7 sub-regions for a common fixed LT, i.e. LT=1-5 years. Here we defined the AMV signal as the 5-year low-pass filtered annual mean temperature averaged over the North Atlantic basin (80W-0W, 0N-65N). Note that we used a 5-year low-pass filter for a direct comparison with temperature predictions over Europe with LT=1-5 years. In this framework, we also analyse the modelled and observed AMV standard deviation over the different periods, which have been shown to characterise both the predictability of the AMV and its teleconnections with the summer temperature over Europe (Qasmi et al., 2017). We find that the predictability of the AMV is phased with the observed AMV variance (Fig. 5a). Indeed, the highest AMV skill occurs for those periods in which the observed AMV standard deviation (blue curve in Fig. 5a) is greatest, while lower skill corresponds to periods with a less variable AMV. It is also worth noting that individual members of DCP produce a less skilful AMV than the ensemble mean, thereby confirming that averaging more realisations reduces the unpredictable noise (Mignot, et al., 2016; Smith et al., 2019). The simulated AMV (red curves in Fig. 5a) is characterised by an underestimated variance, which is not merely due to the ensemble mean effect, as individual model members also show lower AMV variance than observed. In addition, the AMV variance in the DCP does not exactly phase with the observed AMV variance, although the peaks of maximum AMV variance in the model coincides with those in observations.

We aim to understand whether these features are linked to the predictability of air temperature over Europe and over its different sub-regions. The comparison between Fig. 5a and Fig. 3b shows that the periods of best ACC skill for air temperature over Europe (Fig. 3b) coincide with the peaks of maximum AMV predictability (Fig. 5a). Nevertheless, low temperature prediction skill for the 26-yr periods P starting around the 1970s appear to be phased with the lower AMV standard deviation in the model, rather than with the AMV predictability itself. Therefore, in agreement with what was already found by Qasmi et al. (2017), the teleconnection between the AMV and European summer temperature in the model appears to be linked with the simulated AMV variance. This demonstrates the possibility to consider the variance of the predicted AMV as a potential indicator of future windows of opportunity in the decadal prediction of air temperature over Europe. As an example, we can estimate that a large shift in predicted AMV, as has recently been suggested for the coming years (Robson et al. 2016), might lead to enhanced predictability.

That is, real-world decadal predictions of the coming decade may be more accurate than what overview metrics (cf. Figure 1) might imply.

The patterns of S=f(P,M) within the 7 sub-regions (Figs. 5b-g) evidence that southernmost sectors, e.g. Iberian Peninsula, Mediterranean sector, show the highest skill for predictions after the 1970s, contrary to the northernmost regions, e.g. Scandinavia, Central Europe, for which the highest skill is found for P starting prior to the 1970s. Also, the skill variability on the P-axis is in general weaker for the Eastern regions, suggesting a lower impact of AMV variations there. Despite these regional differences, the best ACC skill scores in all the 7 sub-regions mainly concern the temperature prediction of the extended summer seasons for 26-yr initialisation periods starting around 1965 or 1980, when the AMV predictability is maximum. Furthermore, none of the regions is characterised by good temperature prediction skills for 26-yr periods P starting around the 1970s, when the simulated AMV show the minimum standard deviation values.

3.4 The pattern of initialisation added value and improvements due to de-biasing

Following the systematic approach adopted so far, we focus in the last part of this study on the potential skill improvement due to both the initialisation and de-biasing procedure for varying P, LT and M. For the reference context, Fig. 1 already demonstrated a progressive skill improvement in predicting air temperature over Europe due, successively, to initialisation (Figs. 1b, 1e) and to the de-biasing (Figs. 1c, 1f). By producing a 3-dimensional matrix of skill anomalies for air temperature predictions over Europe, both for ACC and RMSE, we now explore if the added values found for the reference context also hold for different combinations of P, LT and M. In other words, we study the function DS=f(P,LT,M), where DS is $S_{RAW} - S_{HIS}$ (Fig. 6) and $S_{DEB} - S_{RAW}$ (Fig. 7), respectively.

Initialisation leads to an overall improvement of the skill in predicting air temperature over Europe (Fig 6). Added value can be seen for most of the combinations of M and LT (Figs. 6a, 6d). Largest skill increases are found, in general, for predictions of spring and summer seasons (Figs. 6a, 6b, 6d, 6e), although the largest ACC increases do not always correspond to the largest RMSE decreases (e.g. Figs. 6b, 6e). Nevertheless, for both skills, the improvement is statistically significant for windows implying short lead times, e.g. LT=1-3 years or for relatively long forecast periods, e.g. LT=2-8 years (Figs. 6a, 6c, 6d, 6f). Finally, the features on the P-axis appear to be the most heterogeneous. Indeed, while skill improvements uniformly involve most of the combinations of P (Figs. 6b, 6c, 6e, 6f), the largest ACC increases are related to the 26-yr initialisation periods starting around the 1960s (Figs. 6b, 6c). For these periods, the ACC increase is statistically significant, while for initialisation periods starting after 1970s the skill improvement mainly concerns lead-time years averaged over long periods, i.e. more than 7 years (Fig. 6c). At the same time, significant RMSE decreases do not appear to prefer a particular period of initialisation, being uniformly distributed over the P axis (Figs. 6e, 6f).

The effect of the de-biasing procedure is strongly dependent on the specific time window analysed (Fig. 7). It produces an overall skill improvement for the period P=1960-1985 (Figs. 7a, 7d), and, in general, for

about the first five combinations of 26-yr initialisation periods P starting after 1960 (Figs. 7b, 7e). This mainly concerns the prediction of autumn months. On the opposite, for the time windows including the first months of the year, de-biasing produces just a slight skill improvement or even no skill improvement. This pattern is completely reversed for 26-yr initialisation periods starting after around 1965 (Figs. 7b, 7e), for which de-biasing appears more beneficial for the prediction of the first months of the year. These improvements are, for most of the cases with LT=1-5 yr, statistically significant when considering the RMSE metric (Fig. 7e). Yet, de-biased predictions appear to be less skilled than raw predictions when simulating the last months of the year (Figs. 7b, 7e). In total, the de-biasing implies a statistically significant ACC improvement of just less than 2% of the 120,120 time windows detected, while a similar amount of time windows are characterised by a significant ACC decrease. Concurrently, the RMSE skill is significantly improved for slightly more than 2% of the time windows, while no significant RMSE degradation has been reported. Therefore, in general it is not possible to establish *a priori* whether the data adjustment has a beneficial effect on prediction skill: this strongly depends on the time window considered, thus making the systematic detection of skill opportunities a necessary step before the application of DCP to the analysis of the impact of the near-term climate variations.

Conclusions And Discussion

In this work we have used hindcasts of the IPSL-CM5A-LR DCP system to systematically assess its skill in predicting air temperature over Europe. We explored the degree of predictability for all 78 combinations of consecutive predicted months of the year, and of all the 55 combinations of lead-time years. We also investigated the potential existence of windows of opportunity by evaluating these skills for all the 28 combinations of 26-year moving periods initialisation starting from 1960. Such a systematic approach can be easily adopted for evaluating the skill of different DCP systems, and for different climatic variables, thus representing a prototype for a comprehensive exploration of the potential exhibited by a DCP system.

We found that temperature prediction skill over Europe is generally larger for the simulation of late spring, summer and early autumn seasons. The systematic evaluation of the hindcasts show peaks of predictability for boreal summer, independently of the considered epoch and lead-time, in agreement with what was found in Yeager at al. (2018). The length in years of the forecast period has also been found to be a factor influencing the skill score. In general, the longer this length is, the better the predictability that results, likely because most of the predictable signal might be found at low frequency due to oceanic processes. This might be compared with the principle for which increasing the model members of a prediction improves its performance (Bellucci et al., 2015; Smith et al., 2019), although this is likely related to a better estimate of the signal and not of its physical characteristics. Furthermore, we identified two main windows of opportunity, namely for the 26-year periods starting around 1965 and around 1980, which we found to be characterised by the largest skill scores. We suggested that, consistent with previous studies, the source of temperature predictability over Europe for these different windows of time is linked with the AMV. Indeed, the peaks of summer temperature predictability over Europe coincide with the peaks of AMV predictability, which is in turn is correlated with the peaks of observed AMV variance.

The identification of the sources of skill is an important issue for reducing uncertainties in forecasts and improving the next generation of DCPs. Since IPSL-CM5A-LR underestimates the amplitude of the AMV variability, similarly to state-of-the-art climate models (Qasmi et al., 2017), it is possible that a better reproduction of the AMV variability in the model can produce more skilful air-temperature predictions over Europe. In this context, the possible on-going shift of the AMV phase might lead to an increase in variance of the AMV, which may therefore open a good window of opportunity for predicting European air temperature in the coming decades. Nevertheless, the effect associated with the simulated AMV is just one of the possible factors influencing temperature predictions over the continent, as other factors not analysed here may also explain it, e.g. the atmospheric circulation (Smith et al., 2016).

An important finding of this study is that, when simulating air temperature over Europe, hindcasts are significantly more skilful than historical simulations for most of the contexts analysed, thereby demonstrating the general added value due to initialisation. Furthermore, for certain time windows we have shown a possible additional significant improvement of the prediction skill coming from DCP de-biasing. However, this beneficial effect is strongly dependent on the context and on the skill metric considered, and it is overall limited to less than 2% of the time windows analysed. Qualitatively similar results have been found by using adjusted data with a different de-biasing method, yet the skill changes (both improvement and degradation) were much smaller than those presented in this study (not shown here). It is therefore likely that the specific bias-adjustment method (e.g. the calibration period used, the way in which the transfer function is calculated, the observation-based data used as reference, etc.) is essential in affecting the eventual prediction skill.

Concerning the impact analyses that can possibly follow this pilot study, it is worth stressing that the adjusted data also exist at higher spatial resolution, thus potentially representing a better dataset for applications to targeted studies. These higher resolution data have been obtained for the predictions with IPSL-CM5A-LR model by calculating, for instance, the transfer function using WFDEI (WATCH Forcing Data methodology applied to ERA-Interim data) reanalysis data after the projection of model predictions on the 0.5^{o}x0.5^{o} WFDEI grid, thus making the de-biasing procedure both a data correction and a spatial downscaling. The same procedure is planned to be carried out for the IPSL-CM6-LR DCP system. Our results are therefore rather promising in light of the fact that the next generation of DCP is expected to perform even better than the previous one. Indeed, the IPSL-CM6-LR DCP system, which was released recently, relies on a higher resolution, on a better estimation of the external forcing, on more model members, i.e. 10 members against 3 for the IPSL-CM5A-LR, and on a more sophisticated initialisation method, i.e. the data assimilation involves also the Sea Surface Salinity and not only the SST. This new version of the IPSL model, along with the new generation of DCPs contributing to the CMIP6 (Boer et al., 2016), have already shown to generally produce more accurate predictions of the North Atlantic SST than the previous versions (Borchert et al., 2020), although most of the improvement was associated to a more accurate response to the external forcing.

Finally, we suggest that a systematic analysis, such as the one presented here, provides relevant information for the development of climate services based on DCP (Bruno Soares and Buontempo, 2019).

For example, the fact that temperature predictions appear to be particularly skilful over spring and summer seasons represents a promising base for the development of a reliable climate service for agriculture sector based on DCP. As a specific example, assume we want to use the IPSL-CM5-LR DCP system for impact analyses in viticulture. The climatic impact on grapevine yields is classically studied by using simulated temperature data to force phenological models, e.g. Sgubin et al. (2018, 2019), which mainly operates over the growing season of the year. While the temperature during the grapevine dormancy occurring between autumn and winter influences the grapevine budburst (Garcia de Cortazar-Atauri et al., 2011), the main following phenological phases (flowering, veraison and maturity) are predominately determined by the spring and summer temperature (Parker et al., 2013). Therefore, the high skill scores found for decadal prediction of temperature over these seasons promises high confidence in DCP for a near-term impact analysis on viticulture. A deeper analysis of the skill in predicting the 1-10 years phenological stages for different grapevine varieties over Europe is the object of an on-going study aimed at testing the effective usability of DCP in the development of a prototype service for viticulture targeting the time horizon of years to decades ahead. In this framework, the use of a new generation of DCP models, along with the de-biasing adjustments proposed in the present study, will enable higher spatial resolution data, which might be an essential factor for a reliable integration of phenological models for near-term predictions of grapevine growing.

Appendix

Table 1: Correspondence between the number of combination N and the time windows identified by M, LT and P. The higlighted values on M and LT columns correspond to the principal time-series.

Combination number (N)	Months (M)	Lead-Time years (LT)	Period of initialisation (P)
1	Jan	1	1960-1985
2	Jan-Feb	1-2	1961-1986
3	Jan-Mar	1-3	1962-1987
4	Jan-Apr	1-4	1963-1988
5	Jan-May	1-5	1964-1989
6	Jan-Jun	1-6	1965-1990
7	Jan-Jul	1-7	1966-1991
8	Jan-Aug	1-8	1967-1992
9	Jan-Sep	1-9	1968-1993
10	Jan-Oct	1-10	1969-1994
11	Jan-Nov	2	1970-1995
12	Jan-Dec	2-3	1971-1996
13	Feb	2-4	1972-1997
14	Feb-Mar	2-5	1973-1998
15	Feb-Apr	2-6	1974-1999
16	Feb-May	2-7	1975-2000
17	Feb-Jun	2-8	1976-2001
18	Feb-Jul	2-9	1977-2002
19	Feb-Aug	2-10	1978-2003
20	Feb-Sep	3	1979-2004
21	Feb-Oct	3-4	1980-2005
22	Feb-Nov	3-5	1981-2006
23	Feb-Dec	3-6	1982-2007
24	Mar	3-7	1983-2008
25	Mar-Apr	3-8	1984-2009
26	Mar-May	3-9	1985-2010
27	Mar-Jun	3-10	1986-2011
28	Mar-Jul	4	1987-2012

29	Mar-Aug	4-5
30	Mar-Sep	4-6
31	Mar-Oct	4-7
32	Mar-Nov	4-8
33	Mar-Dec	4-9
34	Apr	4-10
35	Apr-May	5
36	Apr-Jun	5-6
37	Apr-Jul	5-7
38	Apr-Aug	5-8
39	Apr-Sep	5-9
40	Apr-Oct	5-10
41	Apr-Nov	6
42	Apr-Dec	6-7
43	May	6-8
44	May-Jun	6-9
45	May-Jul	6-10
46	May-Aug	7
47	May-Sep	7-8
48	May-Oct	7-9
49	May-Nov	7-10
50	May-Dec	8
51	Jun	8-9
52	Jun-Jul	8-10
53	Jun-Aug	9
54	Jun-Sep	9-10
55	Jun-Oct	10
56	Jun-Nov	
57	Jun-Dec	
58	Jul	

59	Jul-Aug
60	Jul-Sep
61	Jul-Oct
62	Jul-Nov
63	Jul-Dec
64	Aug
65	Aug-Sep
66	Aug-Oct
67	Aug-Nov
68	Aug-Dec
69	Sep
70	Sep-Oct
71	Sep-Nov
72	Sep-Dec
73	Oct
74	Oct-Nov
75	Oct-Dec
76	Nov
77	Nov-Dec
78	Dec

Declarations

Acknowledgments

This study was supported by the EUCP project funded by the European Union's Horizon 2020 program, grant agreement number 776613. The authors are grateful to Marion Devilliers, Simona Flavoni, Cassien Diabe Ndiaye, Victor Estella-Perez and Brady Ferster for the stimulating discussions around this work.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Aumont O, Bopp L (2006) Globalizing results from ocean in situ iron fertilization studies. Glob Biogeochem Cycles. doi:10.1029/2005GB002591

Arent, DJ, Tol RSJ, Faust E et al (2014) Key economic sectors and services. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 659-708.

Bellucci A, Haarsma R, Gualdi S et al (2015) An assessment of a multi-model ensemble of decadal climate predictions. Clim Dyn 44:2787–2806. https://doi.org/10.1007/s00382-014-2164-y

Bellucci A, Mariotti A, Gualdi S, (2017) The Role of Forcings in the Twentieth-Century North Atlantic Multidecadal Variability: The 1940–75 North Atlantic Cooling Case Study. *J Clim*30:7317–7337. https://doi.org/10.1175/JCLI-D-16-0301.1.

Bilbao R, Wild S, Ortega P, 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. https://doi.org/10.5194/esd-12-173-2021.

Boer GJ (2004) Long time-scale potential predictability in an ensemble of coupled climate models. Clim Dyn 23:29-44. doi:10.1007/s00382-004-0419-8

Boer GJ (2011) Decadal potential predictability of twenty-first century climate. Clim Dyn 36: 1119–1133.

Boer GJ, Smith DM, Cassou C et al (2016) The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. Geosci Model Dev 9:3751–3777. https://doi.org/10.5194/gmd-9-3751-2016.

Booth BBB, Halloran PR, Dunstone NJ, Andrews T, Bellouin, N (2012) Aerosols implicated as a prime driver of 20th century variability within the North Atlantic. Nature 484:228-232.

Borchert LF, Düsterhus A, Brune S, Müller WA, Baehr J (2019) Forecast-oriented assessment of decadal hindcast skill for North Atlantic SST. Geophys Res Lett 46:11444-11454. https://doi.org/10.1029/2019GL084758.

Borchert L F, Menary M B, Swingedouw D, Sgubin G, Hermanson L, Mignot J (2021) Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. Geophys Res Lett 47:e2020GL091307. https://doi.org/10.1029/2020GL091307

Branstator G, Teng, H (2012) Potential impact of initialization on decadal predictions as assessed for CMIP5 models. Geophys Res Lett 39: L12703. doi:10.1029/2012GL051974.

Brune S, Düsterhus A, Pohlmann H et al. (2018) Time dependency of the prediction skill for the North Atlantic subpolar gyeare in initialized decadal hindcasts. Clim Dyn 51:1947–1970. https://doi.org/10.1007/s00382-017-3991-4

Bruno Soares M, Alexander M, Dessai S (2017) Sectoral use of climate information in Europe: A synoptic overview. Clim Serv 9: 10.1016/j.cliser.2017.06.001.

Bruno Soares M, Buontempo C (2019) Challenges to the sustainability of climate services in Europe. WIREs Clim Change 10:e587. https://doi.org/10.1002/wcc.587

Buontempo C, Hewitt, CD, Doblas-Reyes FJ, Dessai, S (2014) Climate service development, delivery and use in Europe at monthly to inter-annual timescales. Clim Risk Manag 6:1-5. https://doi.org/10.1016/j.crm.2014.10.002.

Collins M, Botzet A, Carril AF et al (2006) Interannual to decadal climate predictability in the North Atlantic: a multi model-ensemble study. J Clim 19:1195–1203.

Dessai S, Bruno Soares M (2013) Climate services providers and users' needs - workshop report. European Provision Of Regional Impact Assessment on a Seasonal-to-decadal timescale, Deliverable D12.2. University of Leeds. Available at: www.euporias.eu

Dessai S, Bruno Soares M (2015) Report Summarising Users' Needs for Seasonal to Decadal Climate Predictions. European Provision of Regional Impact Assessment on a Seasonal-to-decadal timescale. Deliverable D12.3. Leeds University. Accessible at: http://www.euporias.eu/system/files/D12.3_Final.pdf.

Dima M, Lohmann G (2007) A hemispheric mechanism for the Atlantic multidecadal oscillation. *J Clim* 20:2706–2719. doi:10.1175/JCLI4174.1

Doblas-Reyes FJ, Andreu-Burillo I, Chikamoto Y et al. (2013) Initialized near-term regional climate change prediction. Nat Commun 4:1715. https://doi.org/10.1038/ncomms2704

Dufresne J-L., Foujols M, Denvil S et al. (2013) Climate change projections using the IPSLCM5 Earth System model: from CMIP3 to CMIP5. Clim Dyn 40:2123–2165.

Dutton JA (2002) Opportunities and priorities in a new era for weather and climate services. Bull Am Meteorol Soc 83:1303–1312.

Estella-Perez V, Mignot J, Guilyardi E, Swingedouw D, Reverdin G (2020). Advances in reconstructing the AMOC using sea surface observations of salinity. Climate Dynamics. 10.1007/s00382-020-05304-4.

Famien AM, Janicot S, Ochou AD, Vrac M, Defrance D, Sultan B, Noël T (2018) A bias-corrected CMIP5 dataset for Africa using the CDF-t method – a contribution to agricultural impact studies. Earth Syst Dynam 9:313–338. https://doi.org/10.5194/esd-9-313-2018.

Garcia de Cortazar-Atauri, I, Brisson N, Gaudillere JP (2009) Performance of several models for predicting budburst date of grapevine (Vitis vinifera L.). Int J Biometeorol 53: 317–326.

García-Serrano JV, Guemas V, Doblas-Reyes FJ (2015) Added-value from initialization in predictions of Atlantic multi-decadal variability. Clim Dyn 44:2539-2555.

Giannini V, Bellucci A, Torresan S (2016) Sharing skills and needs between providers and users of climate information to create climate services: lessons from the Northern Adriatic case study. Earth Perspectives 3:1. https://doi.org/10.1186/s40322-016-0033-z

Goddard L (2016) From science to service. Science 353:1366–1367 (2016).

Kerr RA (2000) A north Atlantic climate pacemaker for the centuries. Science 288:1984–1985.

Knight JR, Allan RJ, Folland CK, Vellinga M, Mann, ME (2005) A signature of persistent natural thermohaline circulation cycles in observed climate. Geophys Res Lett 32:L20708. doi:10.1029/2005GL024233.

Kim H-M, Webster PJ, Curry JA (2012) Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. Geophys Res Lett 39:L10701.

Knight J, Folland CK, Scaife AA (2006) Climate impacts of the Atlantic multidecadal oscillation. Geophys Res Lett 33: L17706.

Kushnir Y, Scaife AA, Arritt R et al (2019) Towards operational predictions of the near-term climate. Nat Clim Change 9: 94–101. https://doi.org/10.1038/s41558-018-0359-7.

Haerter JO, Hagemann S, Moseley C, Piani C (2011) Climate model bias correction and the role of timescales. Hydrol Earth Syst Sci 15:1065–1079. https://doi.org/10.5194/hess-15-1065-2011.

Hourdin F, Foujols M, Codron F et al (2013) Impact of the LMDZ atmospheric grid configuration on the climate and sensitivity of the IPSL-CM5A coupled model. Clim Dyn 40: 2167–2192. https://doi.org/10.1007/s00382-012-1411-3

Hurrell, JW, Kushnir Y, Ottersen G. Visbeck M (2013) An Overview of the North Atlantic Oscillation. In: The North Atlantic Oscillation: Climatic Significance and Environmental Impact (eds J.W. Hurrell, Y. Kushnir, G. Ottersen and M. Visbeck). doi:10.1029/134GM01

IPCC (2014) Summary for policymakers. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1-32.

Lavaysse C, Vrac M, Drobinski P, Lengaigne M, Vischel, T (2012) Statistical downscaling of the French Mediterranean climate: Assessment for present and projection in an anthropogenic scenario. Nat Hazards Earth Syst Sci 12: 651–670.

Lemos MC, Kirchhoff CJ, Ramprasad V (2012) Narrowing the climate information usability gap. Nat Clim Change 2: 789–794.

Li JP, Sun C, Jin FF (2013) NAO implicated as a predictor of Northern Hemisphere mean temperature multidecadal variability. Geophys Res Lett 40:5497–5502.

Madec G (2008) NEMO ocean engine. Tech Rep 27. Institut PierreSimon Laplace, Paris.

Maraun D (2012) Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. Geophys Res Lett 39:L06706. doi:10.1029/2012GL051210.

Mariotti A, Bagget C, Barnes EA et al (2020) Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. Bull Amer Meteor Soc 101:E608–E625. https://doi.org/10.1175/BAMS-D-18-0326.1.

Matei D, Pohlmann H, Jungclaus JH, Müller W, Haak, H, Marotzke, J (2012) Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. J Clim 25:8502–8523.

Meehl GA, Godgard L, Boer G et al (2014) Decadal Climate Prediction: An Update from the Trenches. *Bull Amer Meteor Soc***95**:243–267. https://doi.org/10.1175/BAMS-D-12-00241.1.

Mehta, VM, Knutson CL, Rosenberg NJ, Olsen JR, Wall NA, Bernadt TK, Hayes MJ (2013) Decadal climate information needs of stakeholders for decision support in water and agriculture production sectors: A case study in the Missouri River Basin. Weather Clim Soc 5:27–42.

Menary MB., Hodson DLR, Robson JI, Sutton RT, Wood RA, Hunt JA (2015) Exploring the impact of CMIP5 model biases on the simulation of North Atlantic decadal variability. Geophys Res Lett 42**:** 5926–5934. https://doi.org/10.1002/2015 GL064360.

Michelangeli PA, Vrac M, Loukos H (2009) Probabilistic downscaling approaches: Application to wind cumulative distribution function. Geophys Res Lett 36:L11708. doi:10.1029/2009GL038401.

Mignot, J., Garcia-Serrano J, Swingedouw D et al (2016) Decadal prediction skill in the ocean with surface nudging in the IPSL-CM5A-LR climate model. Clim Dyn, 47:1225-1246.

Murphy J, Kattsov V, Keenlyside N, Kimoto M, Meehl G, Mehta V, Pohlmann H, Scaife A, Smith D (2010) Towards prediction of decadal climate variability and change. Procedia Environ Sci 1:287–304

Oettli P, Sultan B, Baron C, Vrac M (2011) Are regional climate models relevant for crop yield prediction in West Africa? Environ Res Lett 6:014008. https://doi.org/10.1088/1748-9326/6/1/014008

Omrani NE, Keenlyside NS, Bader J, Manzini E (2014) Stratosphere key for wintertime atmospheric response to warm Atlantic decadal conditions. Clim Dyn 42:649-663. https://doi.org/10.1007/s00382-013-1860-3

Ottera OH, Bentsen M, Drange H, Suo L (2010) External forcing as a metronome for Atlantic multidecadal variability. Nature Geosci 3: 688-694.

Oelsmann J, Borchert L, Hand R, Baehr J, Jungclaus JH (2020) Linking ocean forcing and atmospheric interactions to Atlantic multidecadal variability in MPI-ESM1.2. Geophys Res Lett 47:e2020GL087259. https://doi.org/10.1029/2020GL087259.

Qasmi S, Cassou C, Boé J (2017) Teleconnection between Atlantic Multidecadal Variability and European temperature: Diversity and evaluation of the Coupled Model Intercomparison Project phase 5 models. Geophys Res Lett 44**:**11140–11149. https://doi .org/10.1002/2017GL074886.

Parker A, Garcia de Cortá**z**ar-Atauri I, Chuine I, et al (2013) Classification of varieties for their timing of flowering and veraison using a modelling approach: A case study for the grapevine species Vitis vinifera L. Agric For Meteorol 180:249–264.

Robson J, Ortega P, Sutton R (2016) A reversal of climatic trends in the North Atlantic since 2005. Nat Geosci 9:513–517

Reynolds RW, Smith TM, Liu C, Chelton DB, Casey KS, Schlax MG (2007) Daily high-resolution-blended analyses for sea surface temperature. J Clim 20(22):5473–5496. doi:10.1175/2007JCLI1824.1

Slivinski, LC, Compo, GP, Whitaker, JS et al (2019) Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. Q J R Meteorol Soc 145: 2876–2908. https://doi.org/10.1002/qj.3598

Sgubin G, Swingedouw D, Dayon G, García de Cortá**z**ar-Atauri I, Ollat , Pagé C, Van Leeuwen C (2018) The risk of tardive frost damage in French vineyards in a changing climate. Agric For Meteorol 250–251:226–242.

Sgubin G, Swingedouw D, García de Cortázar-Atauri I, Ollat N, van Leeuwen C. (2019) The Impact of Possible Decadal-Scale Cold Waves on Viticulture over Europe in a Context of Global Warming. *Agronomy9*:397. https://doi.org/10.3390/agronomy9070397

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317:796–799.

Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. Clim Dyn 41**:**3325–3338. https://doi.org/10.1007/s00382-013-1683-2

Smith DM, Scaife AA, Eade R, Knight JR (2016) Seasonal to decadal prediction of the winter North Atlantic Oscillation: emerging capability and future prospects. QJR Meteorol Soc, 142: 611-617. doi:10.1002/qj.2479

Smith DM, Eade R, Scaife AA et al (2019). Robust skill of decadal climate predictions. npj Clim Atmos Sci 2:13. https://doi.org/10.1038/s41612-019-0071-y

Street RB (2016) Towards a leading role on climate services in Europe: a research and innovation roadmap. Clim Serv 1, 2–5.

Sutton RT, Hodson DLR (2005) Atlantic Ocean forcing of North American and European summer climate. Science 309: 115-118.

Swingedouw, D, Mignot, J, Labetoulle S, Guilyardi E, Madec G (2013) Initialisation and predictability of the AMOC over the last 50 years in a climate model. Clim Dyn 40:2381–2399. https://doi.org/10.1007/s00382-012-1516-8

Swingedouw D, Ortega P, Mignot J et al (2015) Bidecadal North Atlantic Ocean circulation variability controlled by timing of volcanic eruptions. Nat Commun 6:6545. https://doi.org/10.1038/ncomms7545

Swingedouw D, Mignot J, Ortega P, Khodri M, Menegoz M, Cassou C, Hanquiez V (2017) Impact of explosive volcanic eruptions on the main climate variability modes. Glob Planet Changes 150:24-45.

Ting MF, Kushnir Y, Seager R, C. Li CH (2009) Forced and Internal Twentieth-Century SST Trends in the North Atlantic. J Climate, 22, 1469-1481.

Taylor KE, Stouffer RJ, Meehl GA (2012) An Overview of CMIP5 and the Experiment Design. Bull Am Meteorol Soc 93:485–498.

Trenberth KE, Marquis M, Zebiak S (2016) The vital need for a climate information system. Nat Clim Change 6:1057–1059.

van Oldenborgh GJ, Doblas-Reyes FJ, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Clim Dyn 38:1263–1280.

Vaughan C, Dessai S (2014) Climate services for society: origins, institutional arrangements, and design elements for an evaluation framework. Wiley Interdiscip Rev Clim Change 5: 587–603. https://doi.org/10.1002/wcc.290.

Vautard R, Noël T, Li L, Vrac M, Martin E, Dandin P, Joussaume S (2013) Climate variability and trends in downscaled high-resolution simulations and projections over metropolitan France. Clim Dyn 41:1419–1437. https://doi.org/10.1007/s00382-012-1621-8.

Vigaud N, Vrac M, Caballero Y (2013) Probabilistic downscaling of GCM scenarios over southern India. Int J Climatol 33:1248–1263.

Vrac M, Noël T, Vautard R (2016) Bias correction of precipitation through Singularity Stochastic Removal: Because occurrences matter. J Geophys Res Atmos 121:5237– 5258. doi:10.1002/2015JD024511.

Vrac M, Friederichs P (2015) Multivariate – intervariable, spatial, and temporal – bias correction. J Clim 28:218–237.

Weedon, GP, Balsamo G, Bellouin N, Gomes S, Best MJ, Viterbo P (2014) The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. Water Resour Res 50: 7505– 7514. doi:10.1002/2014WR015638.

Wunsch, C (1999) The Interpretation of Short Climate Records, with Comments on the North Atlantic and Southern Oscillations. *Bull Amer Meteor Soc* 80:245-255.

Yeager SG, Robson JI (2017) Recent Progress in Understanding and Predicting Atlantic Decadal Climate Variability. Curr Clim Change Rep 3:112–127. https://doi.org/10.1007/s40641-017-0064-z

Yeager SG at al (2018) Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. Bull Amer Meteor 99. 10.1175/BAMS-D-17-0098.1.

Figures

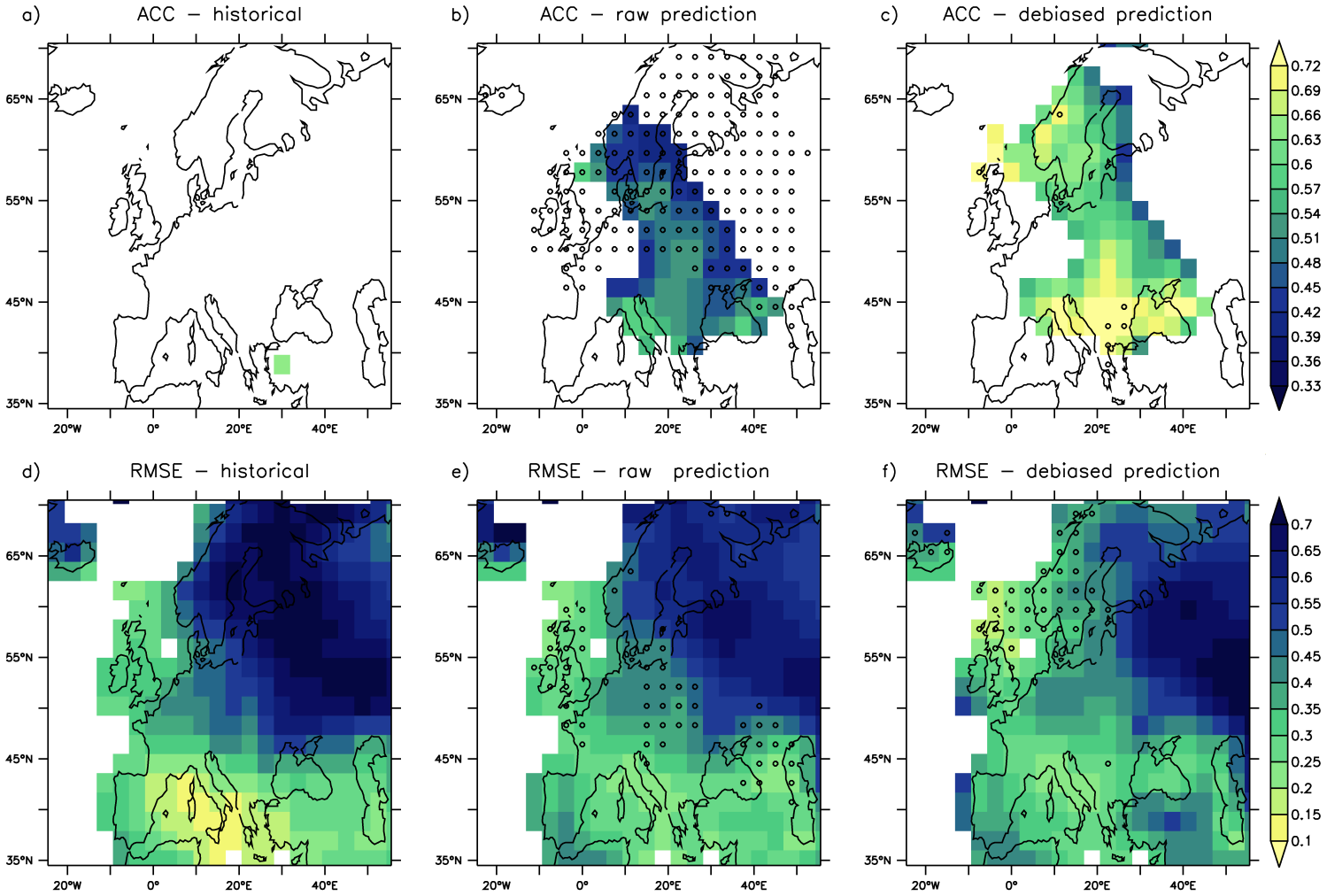

Figure 1

Skill scores exhibited by the different experiments in predicting the 1-5 years annual mean air-temperature over Europe for simulations started along the period 1960-1985. The predictability is defined by two different the skill score metrics, i.e. the ACC (upper panels) and the RMSE (lower panels). For ACC maps, only correlations that are statistically significant at the 95% confidence level have been displayed. The left panels show the skill scores for historical simulation, the middle panels show the skill scores for the raw hindcast and the right panels show the skill scores for the de-biased hindcast. The circles on the central and right panels indicate, respectively, a statistically significant skill improvement (p<0.5) due to initialisation (with respect to HIS skill) and a statistically significant skill improvement (p<0.5) due to de-biasing (with respect to raw DCP skill).

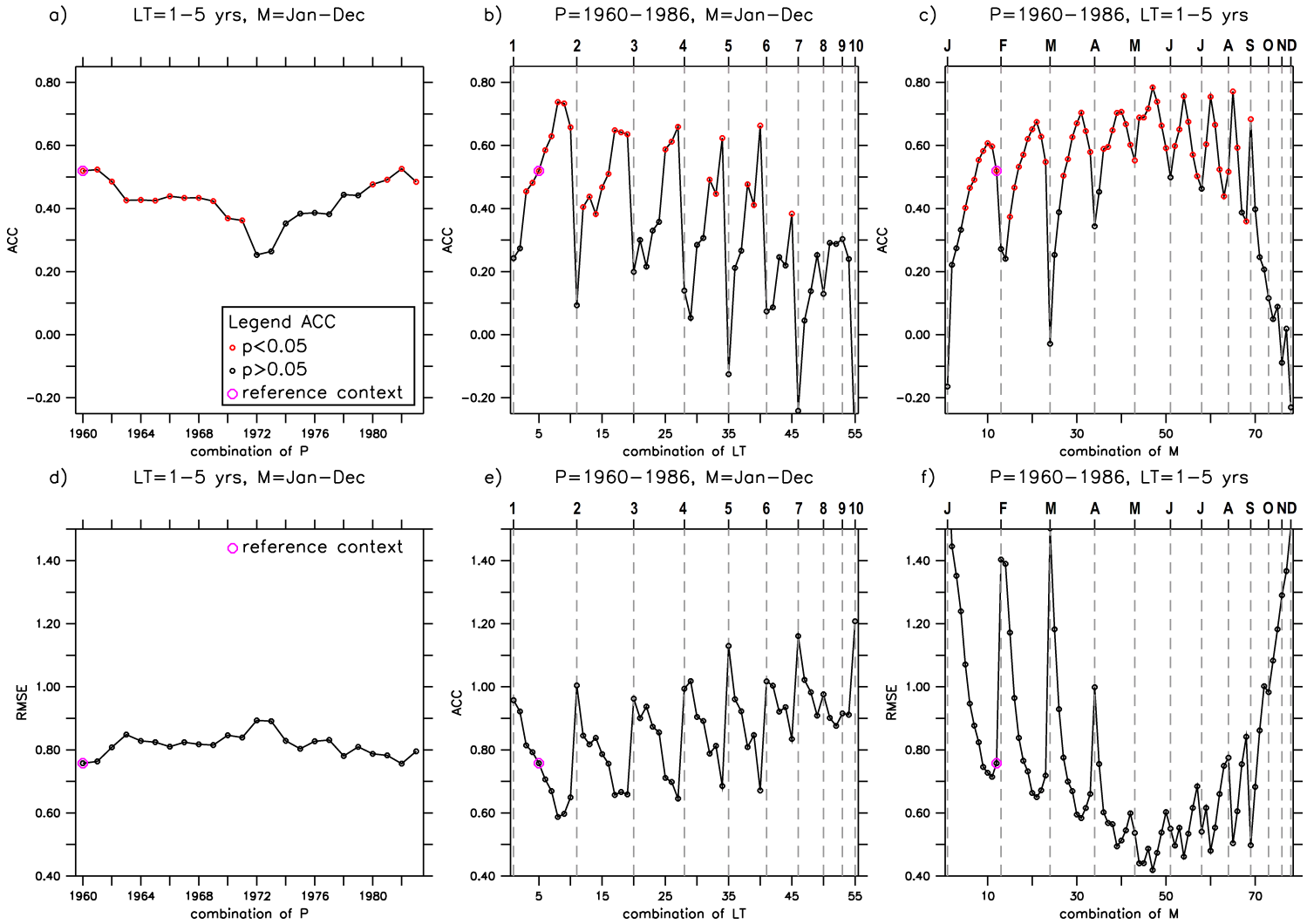

Figure 2

One-dimensional pattern of the prediction skill dependence on (left panels) the period P (for LT=1-5 years and M=Jan-Dec), (middle panels) the lead-time LT (for P=1960-1985 and M=Jan-Dec), and (right panels) the months M (for P=1960-1985 and LT=1-5 years). Prediction skills have been obtained by comparing de-biased ensemble mean temperature hindcasts averaged over Europe with the temperature observations averaged over the same region. The skill metrics are the ACC (upper panels) and the RMSE*

(lower panels). The red circles in the upper plots indicate statistically significant correlation at the 95% confidence level, while black circles indicate not statistically significant correlations. The violet circles indicate the skill score over Europe for the standard context, i.e. for P=1960-1985, for LT=1-5 years and M=Jan-Dec. The 28 different combinations of initialisation periods have been denoted on P axes with the first year of the respective 26-year period. The 55 different combinations of prediction lead-time have been sorted on LT axes such that the first combination corresponds to LT=1 year, the second combination corresponds LT=1-2 years and so on all the combinations LT=1-N years until the eleventh combination coinciding with LT=2 and followed by all the combinations LT=2-N etc. The 78 different combinations of consecutive months have been sorted on M axes such as the first combination corresponds to M=Jan, the second combination corresponds M=Jan-Feb years and so on and so on all the combinations M=Jan-N until the thirteen combination coinciding with M=Feb and followed by all the combinations M=Feb-N etc. See Table 1 in the Appendix for a more detailed explanation.

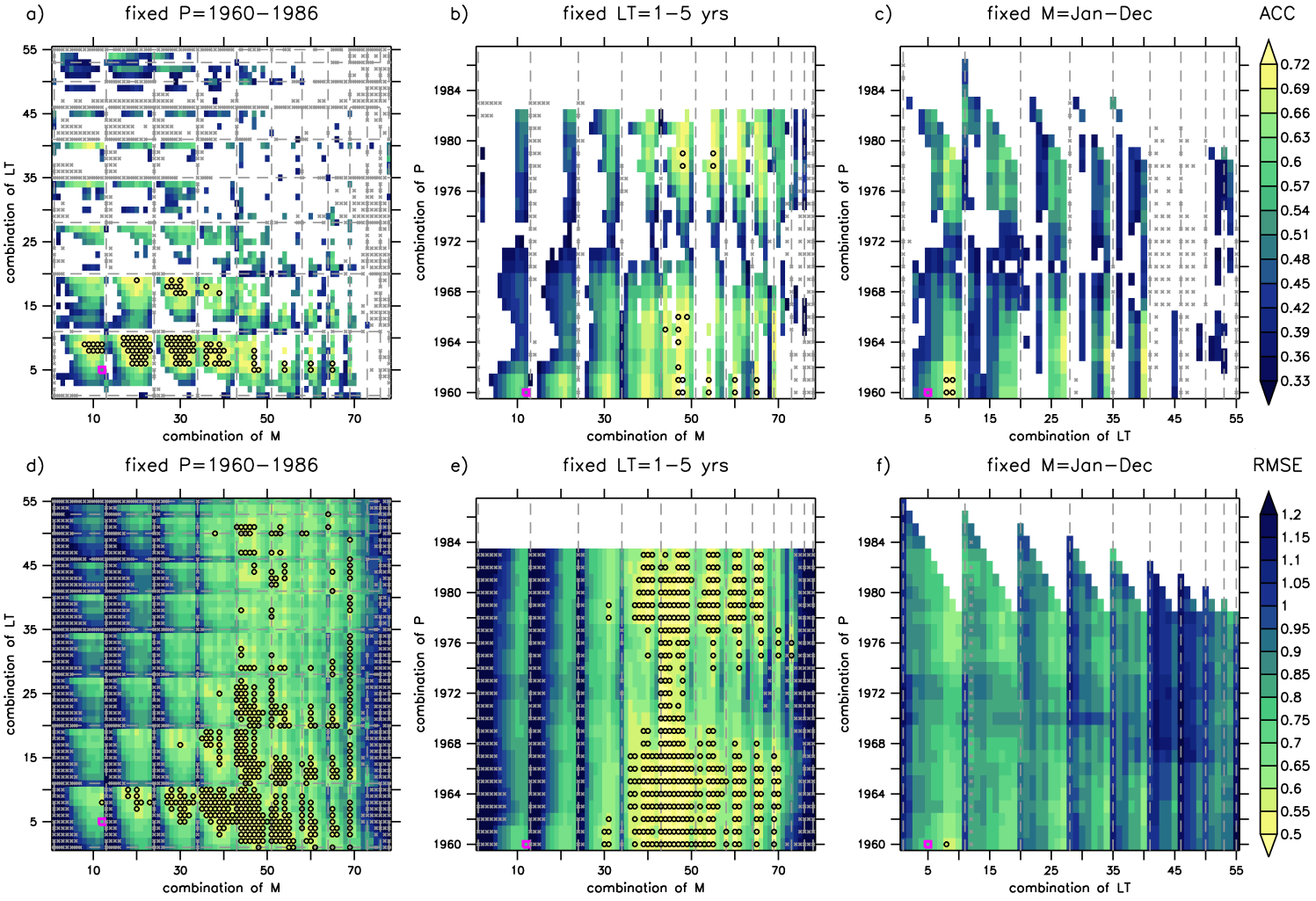

Figure 3

Two-dimensional pattern of the prediction skill dependence on (left panels) lead-time LT and months M (for P=1960-1985), (middle panels) period P and months M (for LT=1-5 years), and (right panels) period P and lead-time LT (for M=Jan-Dec). Prediction skills have been obtained by comparing de-biased

temperature hindcasts averaged over Europe with the temperature observations averaged over the same region. The skill metrics are the ACC (upper panels) and the RMSE* (lower panels). Only correlations that are statistically significant at the 95% confidence level have been displayed. The violet squares indicate the reference context. Black circles indicate that the skill is significantly (p<0.05) higher than the skill for the reference context. Grey crosses indicate that the skill is significantly (p<0.05) lower than the skill for the reference context. Dashed grey lines on M and LT axes correspond to the principal time-series (see Table 1 in the Appendix).

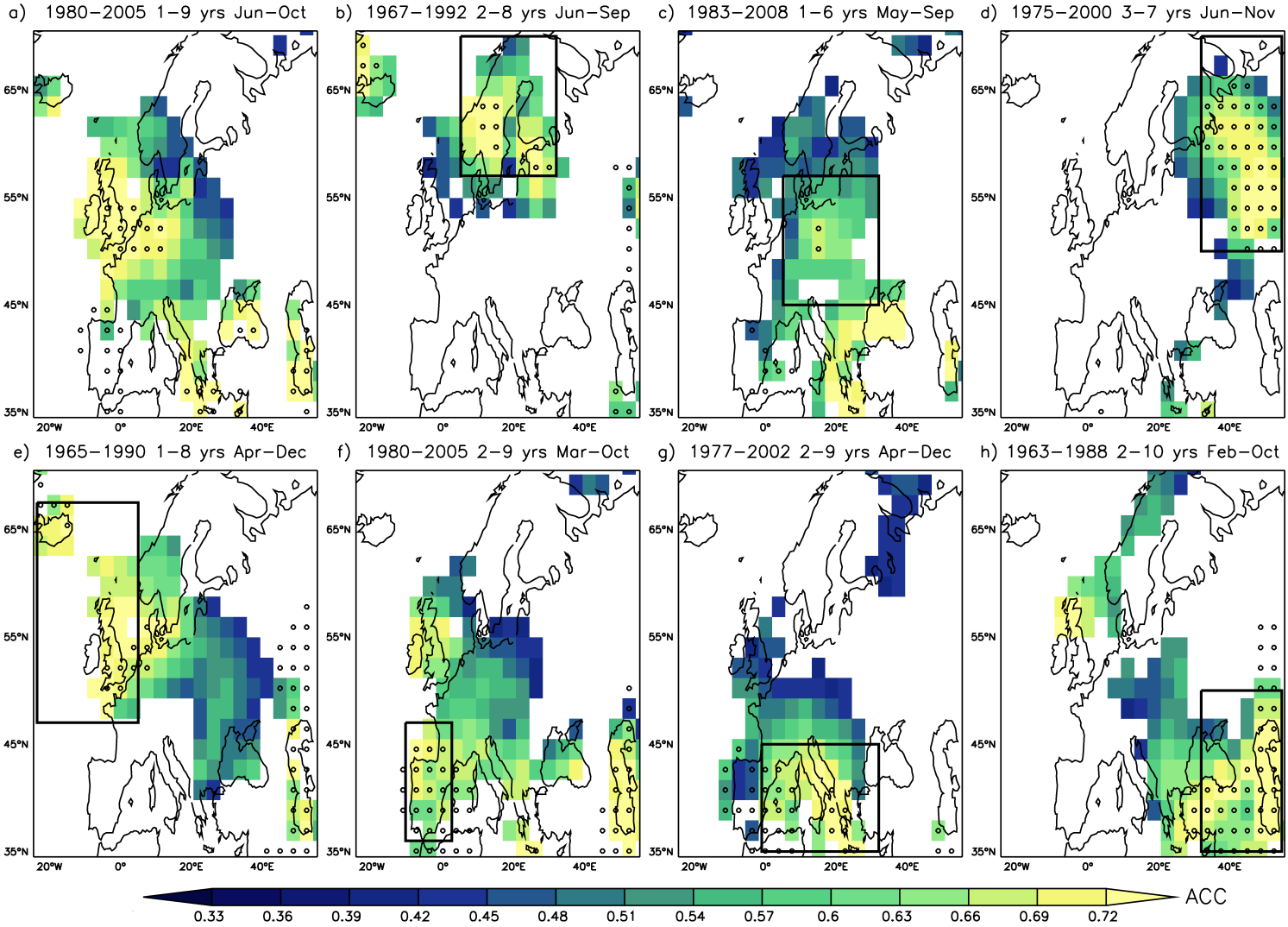

Figure 4

Pattern of ACC skill score under the conditions of best predictability for a) Europe; b) Scandinavia (23W-4E, 44N-67N); c) Central Europe; d) North-eastern Europe; e) North Atlantic sector (23W-4E, 44N-67N); f) Iberian Peninsula; g) Mediterranean ; h) South-eastern Europe. The different conditions of best predictability are reported in the heading of each panel. They have been identified following the systematic approach defined in Section 2, implying the use of ensemble mean air temperature averaged over the corresponding regions, here delimited by the black boxes. Only correlations that are statistically significant at the 95% confidence level have been displayed. The circles indicate, a statistically significant ACC increase (p<0.5) with respect to the ACC for the reference context (Fig. 1b).

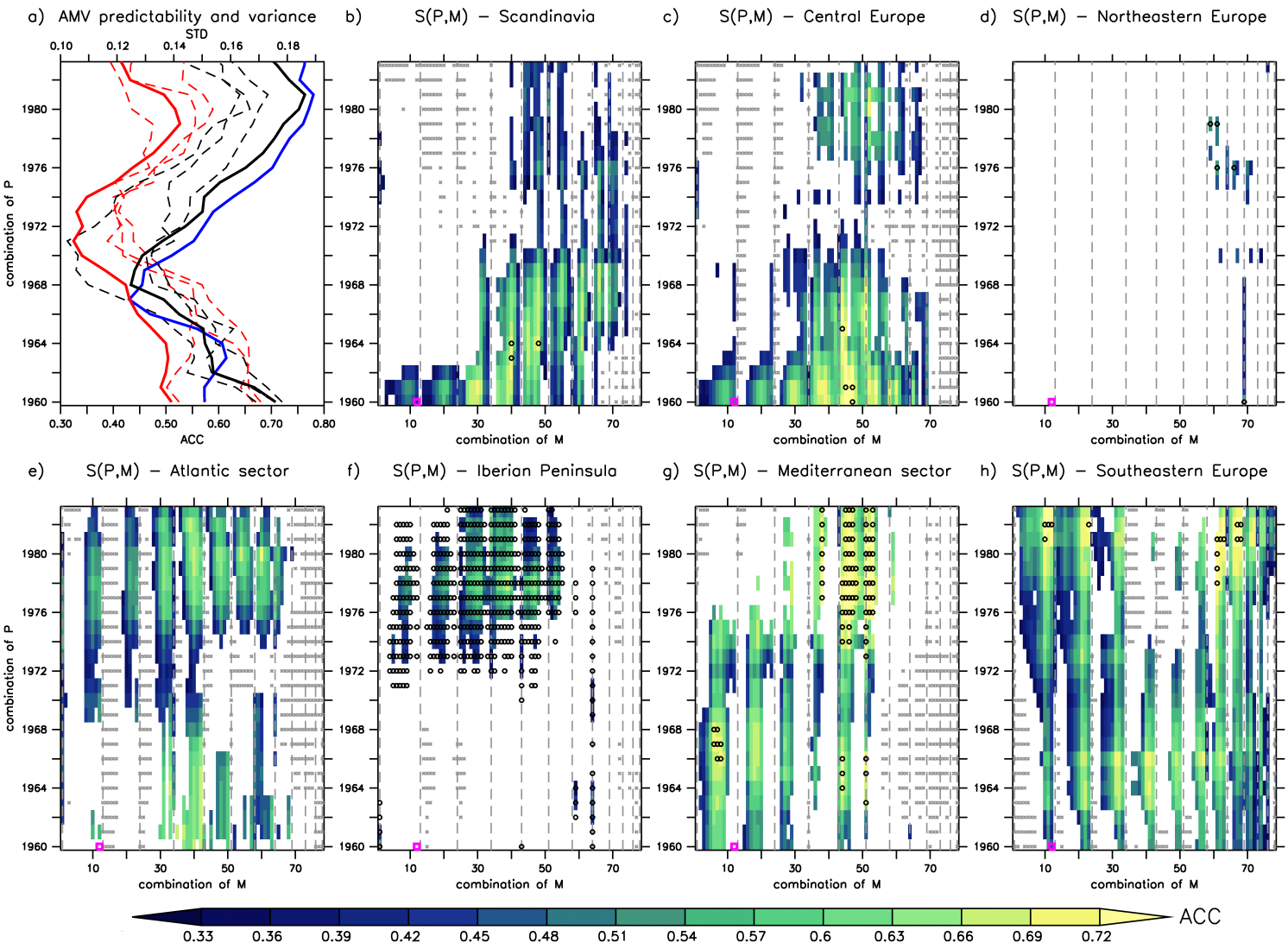

Figure 5

Comparison of (a) the ACC skill of the 1-5 years AMV index (black lines) in hindcasts, and the observed (blue lines) and modelled (red lines) AMV standard deviation over different 26-years periods with (b-h) the two-dimensional ACC skill patterns in predicting the 1-5 years mean temperature over different sub-regions of Europe (see Section 2 for their definition). Thick red and blues lines in the first panel have been calculated from ensemble mean temperature, while dashed black and blue lines have been calculated from the single members. Only correlations that are statistically significant at the 95% confidence level have been displayed. Black circles indicate that the skill is significantly ($p<0.05$) higher than the skill for the reference context. Grey crosses indicate indicate that the the skill is significantly ($p<0.05$) lower than the skill for the reference context. Dashed grey lines on M and LT axes correspond to the principal time-series (see Table 1 in the Appendix).

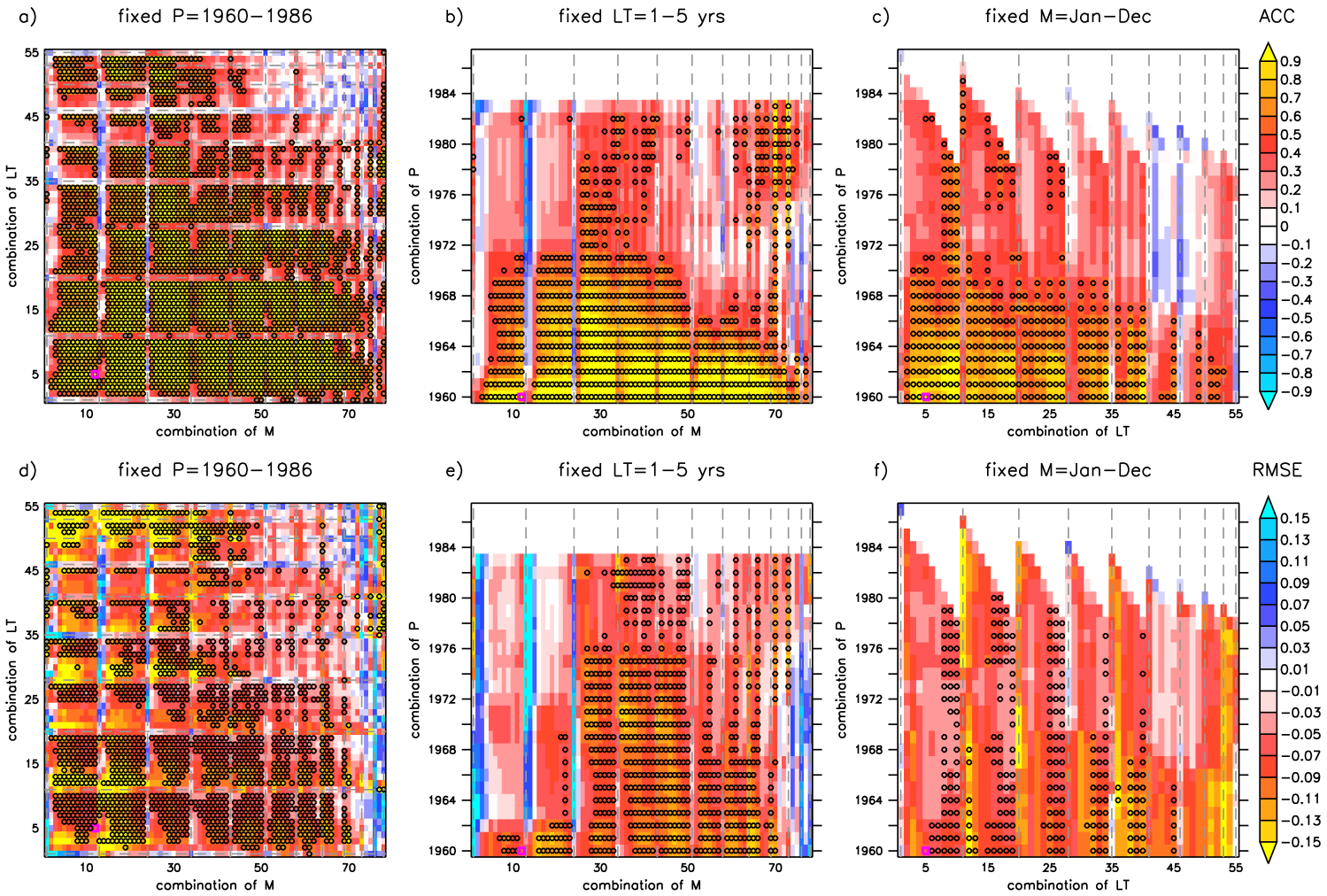

Figure 6

Two-dimensional patterns of the added value due to initialisation in predicting air temperature over Europe for (left panels) fixed P=1960-1985, (middle panels) fixed LT= 1-5 years, and (right panels) fixed M=Jan-Dec. The added values have been obtained by subtracting the skills associated with the HIS ensemble mean from the skill associated with the DCP ensemble mean (SRAW – SHIS). Prediction skills are ACC (upper panels) and RMSE (lower panels), and have been obtained by comparing simulated temperature over Europe with OBS data over the same region. Circles on upper panels indicate where the ACC for the raw DCP experiment is significantly greater ($p<0.05$) than the ACC for the HIS experiment. The violet squares indicate the reference context. Circles on the lower panels indicate where the difference between the RMSE in DCP and HIS experiments is statistically significant at the 95% confidence level according with the Welch's t-test on the mean squared errors. The violet squares indicate the reference context. Dashed grey lines on M and LT axes correspond to the principal time-series (see Table 1 in the Appendix).

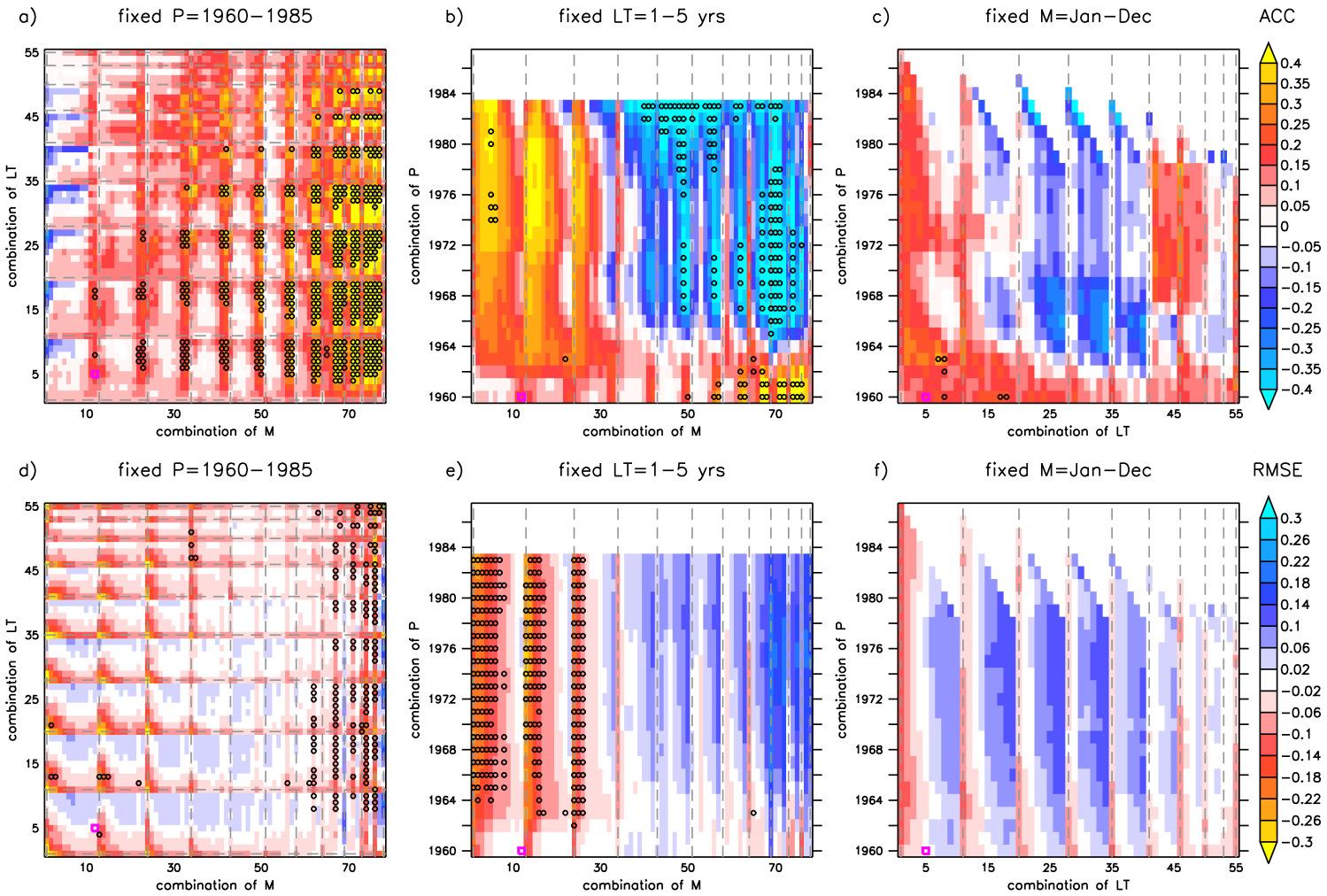

Figure 7

Comparison of skill metrics between raw and debiased DCP for (left panels) fixed P=1960-1985, (middle panels) fixed LT= 1-5 years, and (right panels) fixed M=Jan-Dec. The skill improvements have been obtained by subtracting the skills associated with the raw DCP ensemble mean from the skill associated with the de-biased DCP ensemble mean (SDEB – SRAW). Prediction skills are ACC (upper panels) and RMSE (lower panels), and have been obtained by comparing simulated temperature over Europe with OBS data over the same region. Circles on upper panels indicate where the ACC for the de-biased DCP experiment is significantly greater ($p<0.05$) than the ACC for the raw DCP experiment. The violet squares indicate the reference context. Circles on the lower panels indicate where the difference between the RMSE in DCP and HIS experiments is statistically significant at the 95% confidence level according with the Welch's t-test on the mean squared errors. The violet squares indicate the reference context. Dashed grey lines on M and LT axes correspond to the principal time-series (see Table 1 in the Appendix).