BAMS Article

Seasonal Forecasts of the Twentieth Century

Antje Weisheimer, Daniel J. Befort, Dave MacLeod,
Tim Palmer, Chris O'Reilly, and Kristian Strømmen

ABSTRACT: Forecasts of seasonal climate anomalies using physically based global circulation models are routinely made at operational meteorological centers around the world. A crucial component of any seasonal forecast system is the set of retrospective forecasts, or hindcasts, from past years that are used to estimate skill and to calibrate the forecasts. Hindcasts are usually produced over a period of around 20–30 years. However, recent studies have demonstrated that seasonal forecast skill can undergo pronounced multidecadal variations. These results imply that relatively short hindcasts are not adequate for reliably testing seasonal forecasts and that small hindcast sample sizes can potentially lead to skill estimates that are not robust. Here we present new and unprecedented 110-year-long coupled hindcasts of the next season over the period 1901–2010. Their performance for the recent period is in good agreement with those of operational forecast models. While skill for ENSO is very high during recent decades, it is markedly reduced during the 1930s–1950s. Skill at the beginning of the twentieth century is, however, as high as for recent high-skill periods. Consistent with findings in atmosphere-only hindcasts, a midcentury drop in forecast skill is found for a range of atmospheric fields, including large-scale indices such as the NAO and the PNA patterns. As with ENSO, skill scores for these indices recover in the early twentieth century, suggesting that the midcentury drop in skill is not due to a lack of good observational data. A public dissemination platform for our hindcast data is available, and we invite the scientific community to explore them.

https://doi.org/10.1175/BAMS-D-19-0019.1
Corresponding author: Antje Weisheimer, antje.weisheimer@physics.ox.ac.uk
In final form 28 February 2020

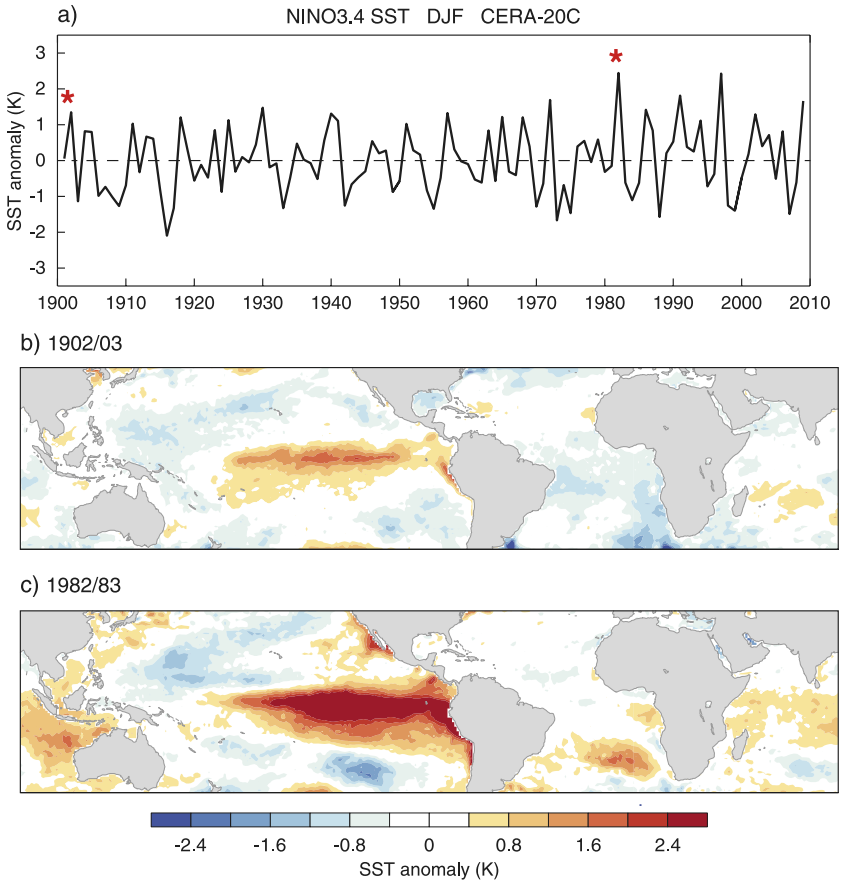
AFFILIATIONS: **Weisheimer**—National Centre for Atmospheric Science, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, and European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; **Befort, MacLeod, and Strømmen**—Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom; **Palmer and O'Reilly**—National Centre for Atmospheric Science, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom

Forecasts of seasonal climate anomalies using physically based global circulation models are routinely made at many operational meteorological centers around the world. Slow variations in the lower boundary forcing of the atmosphere due to the dynamics of the oceans and the hydrology of the landmasses, together with stratospheric variability, sea ice anomalies, and volcanic eruptions, are sources of predictability on seasonal time scales. The single most important of these factors is variations in the tropical Pacific sea surface temperatures (SSTs) linked to the El Niño–Southern Oscillation (ENSO) phenomenon, which impacts climate conditions in remote regions of the world through atmospheric teleconnections (Yeh et al. 2018) (see also sidebar "ENSO during the twentieth century").

A crucial component of any seasonal forecast system is the retrospective forecasts, or hindcasts, from past years that are used to estimate skill and to calibrate the forecasts. Operational hindcasts are usually produced over a period of 20–30 years (Saha et al. 2014; MacLachlan et al. 2015; Johnson et al. 2019). Such small hindcast periods can be problematic not only for the evaluation of sporadically occurring drivers of seasonal predictability such as El Niño events, but also for regions where the interannual variability is high, for example, the atmospheric flow over the North Atlantic/European region (Kumar 2009; Kumar and Chen 2018). In addition, short hindcast periods do not allow for a robust sampling of the variability of

ENSO during the twentieth century

The top plot (a) shows the time series of the Niño-3.4 SST index anomaly in the central tropical Pacific (5°N–5°S, 170°–120°W) during DJF from 1901 to 2009 as a measure of ENSO in the CERA-20C reanalysis. Positive anomalies correspond to El Niño states and negative anomalies to La Niña states. The twentieth century was characterized by a pronounced ENSO variability. Two examples of strong El Niño events near the beginning (1902/03) and the end (1982/83) of the twentieth century are depicted by the red stars. The maps below show the spatial structures of SST anomalies during these two El Niño example events [1902/02 in (b) and 1982/83 in (c)] in the CERA-20C reanalysis. The strong and large-scale warming of the central and eastern tropical Pacific is clearly visible. Cold anomalies develop in the western parts of the tropical Pacific.



the climate system on decadal and longer time scales (Shi et al. 2015; O'Reilly 2018). While previous seasonal hindcasts in research mode were carried out from the late 1950s onward (e.g., Palmer et al. 2004; Weisheimer et al. 2009; Huang et al. 2017), recently available climate reanalyses over the entire twentieth century up to the year 2010 (Compo et al. 2011; Poli et al. 2016; Laloyaux et al. 2018) provide ideal datasets to initialize, verify, and test the performance of seasonal forecasts over longer historical periods.

Here we introduce a new and unprecedented 110-year-long hindcast dataset to predict the next season ahead, based on simulations with ECMWF's coupled atmosphere–ocean–sea ice model over the period 1901–2010. We will refer to this dataset as Coupled Seasonal Forecasts of the 20th Century (CSF-20C). These hindcasts are shown to have comparable levels of skill to state-of-the-art operational seasonal forecasts. While predictions of ENSO during recent periods and at the beginning of the twentieth century are very skillful, the model's ability to forecast ENSO is reduced during the 1920s and 1930s. Consistent with previous findings using atmosphere-only hindcasts with prescribed SSTs (Weisheimer et al. 2017; O'Reilly et al. 2017), we demonstrate that coupled seasonal forecast skill undergoes pronounced decadal variability. A noticeable drop in forecast skill during the middle of the twentieth century is found for a range of atmospheric fields including large-scale extratropical circulation indices such as the North Atlantic Oscillation (NAO) and the Pacific–North American (PNA) pattern that cannot be simply explained by the poorer quality of observational datasets in the past.

The new long seasonal hindcast datasets provide a unique opportunity to study a wide range of predictability problems, including for the training of modern machine learning techniques. Data from these hindcasts will become publicly available through a dedicated online dissemination platform, and we invite the wider international scientific community to explore the plethora of outputs from these novel simulations.

This article is structured as follows: the second section describes the long hindcast experiments, the third section discusses some aspects of ENSO predictability, and the fourth section highlights selected key findings for extratropical seasonal forecast performance over the twentieth century. Our data dissemination activities are described in the fifth section, followed by a summary and conclusions in the final section.

The CSF-20C hindcast experiments

The new coupled hindcast experiment CSF-20C was performed with ECMWF's Integrated Forecasting System (IFS) coupled model version cycle 41r1, which includes state-of-the-art atmospheric, land surface, oceanic, and sea ice components. It is similar to ECMWF's currently operational seasonal forecasting system SEAS5 (Johnson et al. 2019) but was run with a slightly earlier model version. The atmospheric resolution for these runs is T_L255 (approximately 80 km) horizontally, with 91 vertical levels. The ocean resolution is 1° horizontally with 42 vertical levels. All model components of CSF-20C were initialized with ECMWF's first coupled reanalysis of the twentieth century, CERA-20C (Laloyaux et al. 2018), which provides data from 1901 to 2010. Only conventional surface observations in the atmosphere (i.e., surface pressure and marine winds but no satellite data) and observed subsurface temperature and salinity profiles in the ocean were assimilated in CERA-20C. The hindcasts were initialized on the first of February, May, August, and November from 1901 to 2010 and run for 4 forecast months so that they cover all seasons. They consist of ensembles of 51 members for the May and November start dates and 25 members for the February and August start dates. The ensembles were created by a combination of stochastic perturbations to the model physics in the atmosphere and the 10 members of CERA-20C.

The experiments were set up in a way to mimic ECMWF's operational forecasts as much as possible to enable a clear comparison with a real-time forecasting system when only information before the initial date is available to use. Time-varying forcings from greenhouse gases,

the solar cycle, and volcanic aerosols were prescribed.

A companion hindcast experiment using an identical model setup but with prescribed observed SSTs instead of a fully coupled ocean, called Atmospheric Seasonal Forecasts of the 20th Century (ASF-20C) (Weisheimer et al. 2017), will also be used. ASF-20C has been initialized using ERA-20C, ECMWF's first atmospheric reanalysis of the twentieth century (Poli et al. 2016). The implied perfect knowledge of the SSTs and the disabled coupling between the atmosphere and the ocean in ASF-20C represent an idealized setup that will serve as a benchmark in the comparison with CSF-20C. Figure 1 shows the evolution of global mean surface temperature in December–February (DJF) from 1901 to 2010 in reanalysis and the hindcast data. There is, in general, very good agreement of the interannual variability in the reanalysis and both model runs (the correlation of the ensemble mean with the verifying CERA-20C data are highly significant with correlation coefficients of 0.93 and 0.94 for the ASF-20C and CSF-20C, respectively). ASF-20C has a slight tendency to underestimate the overall warming trend from the beginning of the twentieth century to the end, while the coupled hindcast CSF-20C reproduces these long-term trends more realistically.

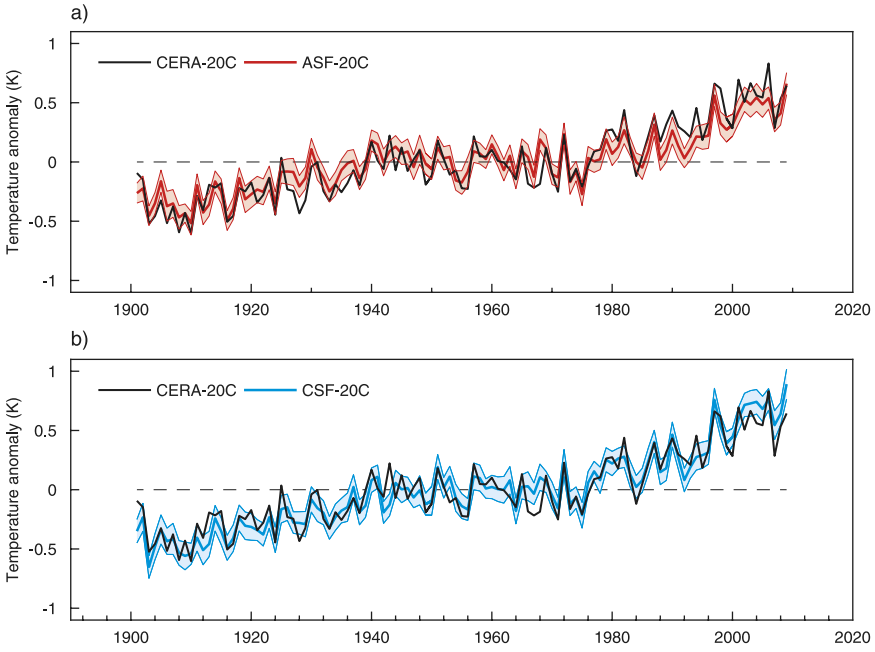


Fig. 1. Global-mean (60°N–60°S) 2-m temperature in DJF from 1901 to 2010 in CERA-20C (black) and in the seasonal hindcasts. (a) Atmosphere-only seasonal forecasts (ASF-20C) in red and (b) coupled seasonal forecasts (CSF-20C) in blue. Forecasts were initialized in November. Ensemble means are plotted with thick lines and the ensemble spread around the mean is given by the shaded band.

Forecasting ENSO

ENSO is the single largest source of predictability on seasonal time scales, and it is thus of paramount importance for any seasonal forecasting system to being able to realistically reproduce its behavior. Here, the oceanic Niño-3.4 SST index in the central tropical Pacific (average SSTs between 5°N–5°S and 170°–120°W) and the atmospheric equatorial Southern Oscillation index (SOI), defined as the standardized difference in mean sea level pressure between an equatorial western Pacific area (5°N–5°S, 90°–140°E) and an equatorial eastern Pacific area (5°N–5°S, 130°–80°W), will be used to assess model ENSO performance.

To start with, we compare the performance over the common hindcast period 1981–2009 of the coupled forecast system CSF-20C with ECMWF's operational seasonal prediction system, SEAS5 (Johnson et al. 2019). Note that SEAS5 is run at finer horizontal resolution in the atmosphere (approximately 36 km) and in the ocean (1/4°) and uses many more observed data for its initial conditions. Figure 2 shows biases (Fig. 2a) and skill (Fig. 2b) in the reforecasts of Niño-3.4 SSTs in CSF-20C, SEAS5, and SEAS5 run at lower resolution (approximately 50 km in the atmosphere and 1° in the ocean). The ECMWF model drifts toward a cold bias that varies throughout the calendar year and that is smallest for the high-resolution SEAS5. CSF-20C shows a slightly reduced cold bias compared with the low-resolution SEAS5 hindcasts which we hypothesize could be the benefit of initializing with a coupled atmosphere–ocean reanalysis

in CSF-20C, rather than using separate reanalyses for the atmosphere and the ocean as in SEAS5.

Systematic biases are normally removed from seasonal forecasts as a function of the calendar start date and lead time so that forecasts are issued in the form of seasonal-mean anomalies around the model mean state. The skill of predicting the correct interannual variations of seasonal-mean anomalies can be estimated by the correlation between the ensemble-mean forecast anomalies and the observed anomalies over the hindcast period. Figure 2b shows how the anomaly correlation decreases with lead time for several start dates throughout the year. CSF-20C achieves high levels of ENSO skill, in general comparable with the performance of high-resolution SEAS5. There is a strong seasonality of forecast skill with a pronounced drop in skill during April and May, the so-called spring barrier of predictability (Webster and Yang 1992).

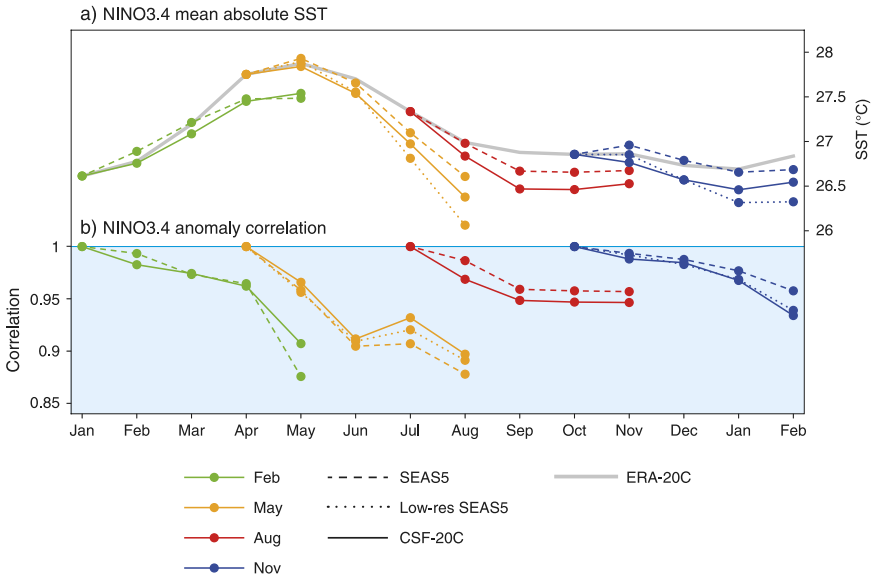


Fig. 2. (a) Drift and (b) anomaly correlation skill of predictions of the Niño-3.4 SST index for different start dates throughout the calendar year during the hindcast period 1981–2009. CSF-20C data are shown in solid colored lines, SEAS5 in dashed lines, and the lower-resolution SEAS5 in dotted lines (only available for May and November start dates). Different colors indicate different start dates of the hindcasts. The gray curve in (a) shows the climatological mean evolution of the SSTs in ERA-20C over this period.

The global distributions of SST skill during DJF in CSF-20C and SEAS5 are displayed in Fig. 3. During the common hindcast period 1981–2009 the skill in CSF-20C is, in general, well aligned with both the high- and low-resolution configurations of SEAS5. Consistent with Fig. 2b, SEAS5

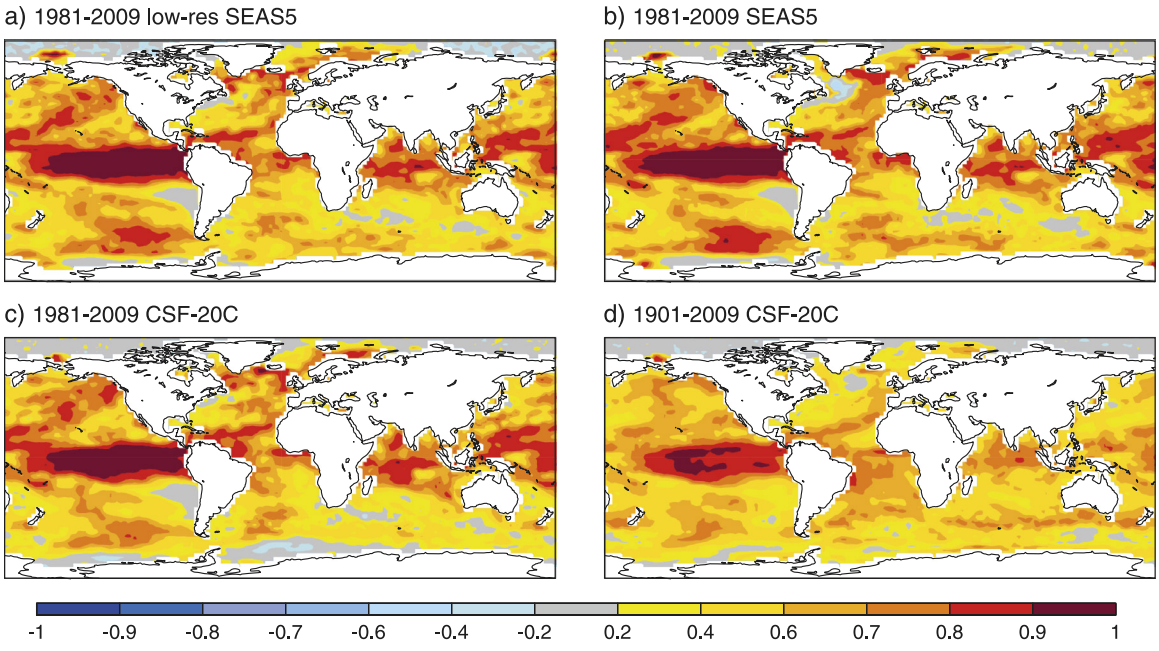


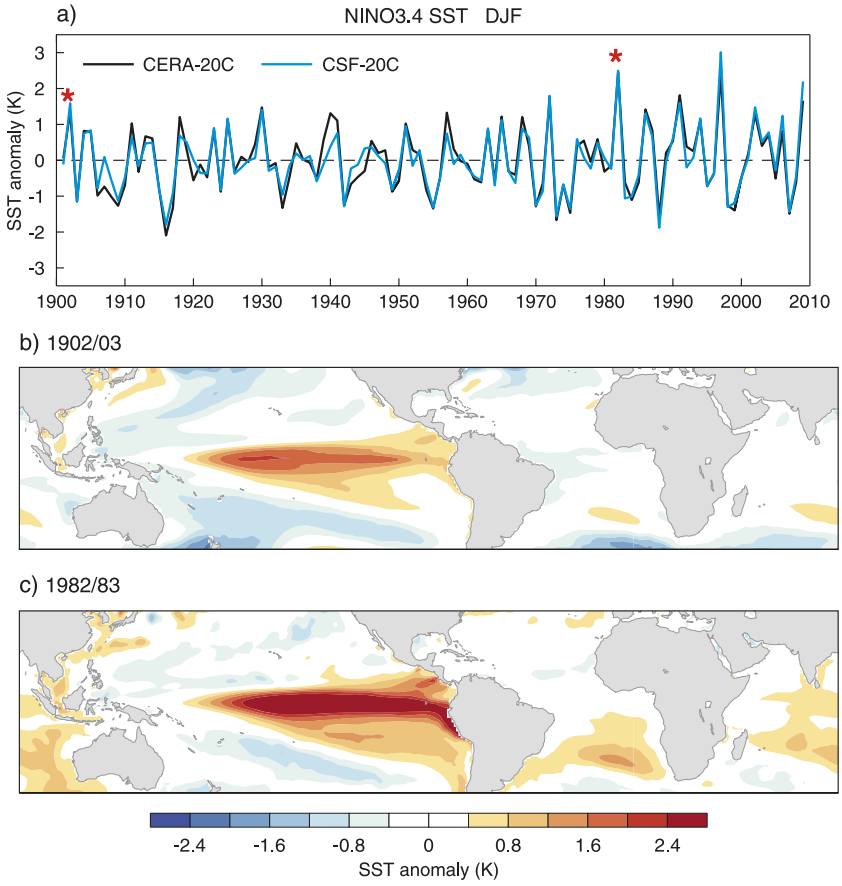
Fig. 3. Anomaly correlation skill of SST in DJF for forecasts initialized in November during the common hindcast period 1981–2009 for (a) low-resolution SEAS5, (b) SEAS5, and (c) CSF-20C and (d) for the full hindcast period 1901–2009 of CSF-20C. The verification data used are ERA-20C.

has overall slightly more skill due to the higher resolution of the ocean model. Figure 3d shows that if the full hindcast period is used, CSF-20C forecasts are still very skillful in the tropical oceans. Reductions in skill compared to recent decades are noticeable over the North Pacific and the North Atlantic.

Can our forecasting system CSF-20C uphold this excellent performance in predicting ENSO during the twentieth century? CSF-20C forecasts of Niño-3.4 SSTs in DJF initialized in November over the full hindcast period 1901–2009 have an anomaly correlation skill of $r_{\text{Niño-3.4}} = 0.95$ in CSF-20C (see also sidebar "ENSO in the CSF-20C"). The Niño-3.4 SST index is strongly coupled to the SOI, which is an atmospheric manifestation of ENSO. In the coupled reanalysis CERA-20C over the period 1901–2009, the Niño-3.4 SST and SOI in DJF are highly negatively correlated with a correlation coefficient of $r_{\text{Niño-3.4-SOI}} = -0.93$. Both the coupled CSF-20C and atmosphere-only ASF-20C hindcasts capture this strong inverse relationship between the Niño-3.4 SST and SOI indices very well with $r_{\text{Niño-3.4-SOI}} = -0.95$ for CSF-20C and $r_{\text{Niño-3.4-SOI}} = -0.92$ for ASF-20C. The predictive skill of the atmospheric SOI in terms of anomaly correlation of the ensemble mean with CERA-20C over the full 110 years is also high in both hindcasts with $r_{\text{SOI}} = 0.91$ for CSF-20C and $r_{\text{SOI}} = 0.92$ for ASF-20C. It is interesting to note that the level of skill for the atmospheric expression of ENSO is relatively similar regardless of whether the underlying SSTs were perfectly prescribed (ASF-20C) or forecast (CSF-20C). All correlations reported are highly statistically significant with $p < 0.01$.

ENSO in the CSF-20C

Similar to the sidebar "ENSO during the twentieth century," the top plot (a) shows the time series of the Niño-3.4 SST index anomaly in DJF from the CERA-20C reanalysis (black) and the ensemble mean of the CSF-20C hindcasts initialized in November (blue). The model predicts the reanalysis data overall very well with a correlation coefficient of $r = 0.95$. Two examples of strong El Niño events near the beginning (1902/03) and the end (1982/83) of the twentieth century are depicted by the red stars. The maps below show the spatial structures of SST anomalies in the CSF-20C hindcasts during these two El Niño example events in 1902/03 (b) and 1982/83 (c). A comparison with sidebar "ENSO during the twentieth century" reveals a very good agreement with the strong and large-scale observed warming of the central and eastern tropical Pacific. The observed cold anomalies in the western parts of the tropical Pacific are also visible.



Has the skill in predicting ENSO been constantly high throughout the twentieth century, or is there temporal variability in the models' ability to forecast ENSO? Figure 4 shows how ENSO correlation skill during DJF for ASF-20C (SOI) and CSF-20C (SOI and Niño-3.4) changes during the hindcast period from 1901 to 2009. Here, 30-year moving windows have

been used to calculate the temporal changes in correlations. For the period 1980–2009 the hindcast skill for all three indices is very high. The system maintained these high levels of skill for several decades in the past with a remarkable constant level of very high Niño-3.4 SST skill in CSF-20C until the early 1960s, implying, in agreement with Huang et al. (2017), that ENSO skill was largely unaffected by the decreasing number of ocean observations during these decades.

However, it becomes apparent that atmospheric ENSO skill was synchronously reduced to around less than 0.8 during the 1930s–1960s in both ASF-20C and CSF-20C. In contrast, even earlier decades at the beginning of the twentieth century clearly show higher levels of skill, indeed comparable in magnitude with the high skill of the most recent decades at the end of the century. The oceanic Niño-3.4 ENSO index also shows a drop of skill during an extended period during the first half of the century, roughly but not exactly in phase with the drop of SOI skill. Like in the atmosphere, higher SST skill is restored near the beginning of the century. The reduction in skill during the 1930s–1960s thus cannot simply be attributed to the poorer observational coverage during this period and we hypothesize instead that these changes in model predictability are linked to intrinsic changes of the coupled climate system.

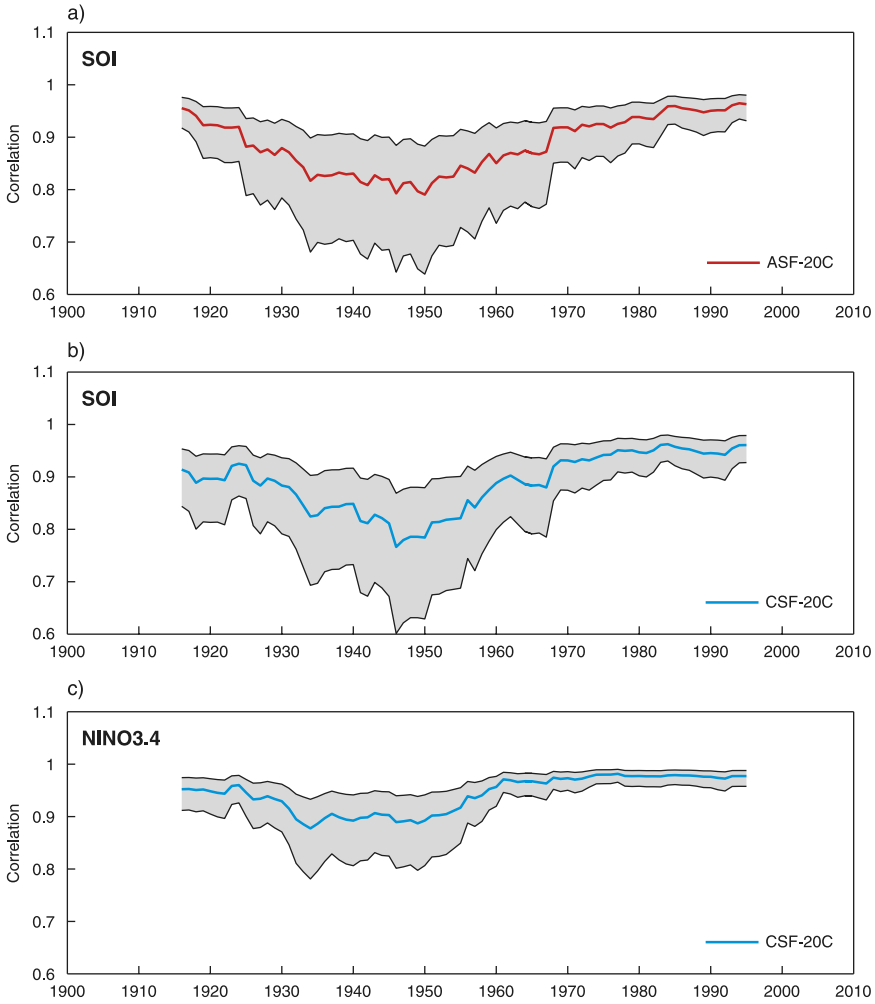


Fig. 4. Decadal variability of ENSO forecast skill in DJF from forecasts initialized in November, with 30-year moving-window correlation coefficients of the hindcast ensemble mean with CERA-20C. (a) SOI index for atmosphere-only seasonal forecasts (ASF-20C) in red. (b) SOI index for coupled seasonal forecasts (CSF-20C) in blue. (c) Niño-3.4 index for CSF-20C in blue. The gray-shaded bands indicate the 5%–95% confidence intervals. Correlations for each 30-year window are plotted at the central year.

Extratropical predictions

How well do the century-long coupled seasonal hindcasts perform for the extratropics?

In this section we are going to discuss biases and skill in predicting the mid-tropospheric flow in the extratropics with a special emphasis on the NAO, including its signal-to-noise paradox, and the eddy-driven jet regimes over the North Atlantic.

Maps of the global bias of predicting geopotential height at 500 hPa (Z500) during DJF are shown in Figs. 5a–c for SEAS5, ASF-20C, and CSF-20C computed for the common hindcast period 1981–2009 and for ASF-20C and CSF-20C for the complete hindcast periods of the twentieth century, 1901–2009 (Figs. 5d,e). ERA-20C has been used as reference data (results would not be different when using CERA-20C instead). The atmosphere-only and coupled

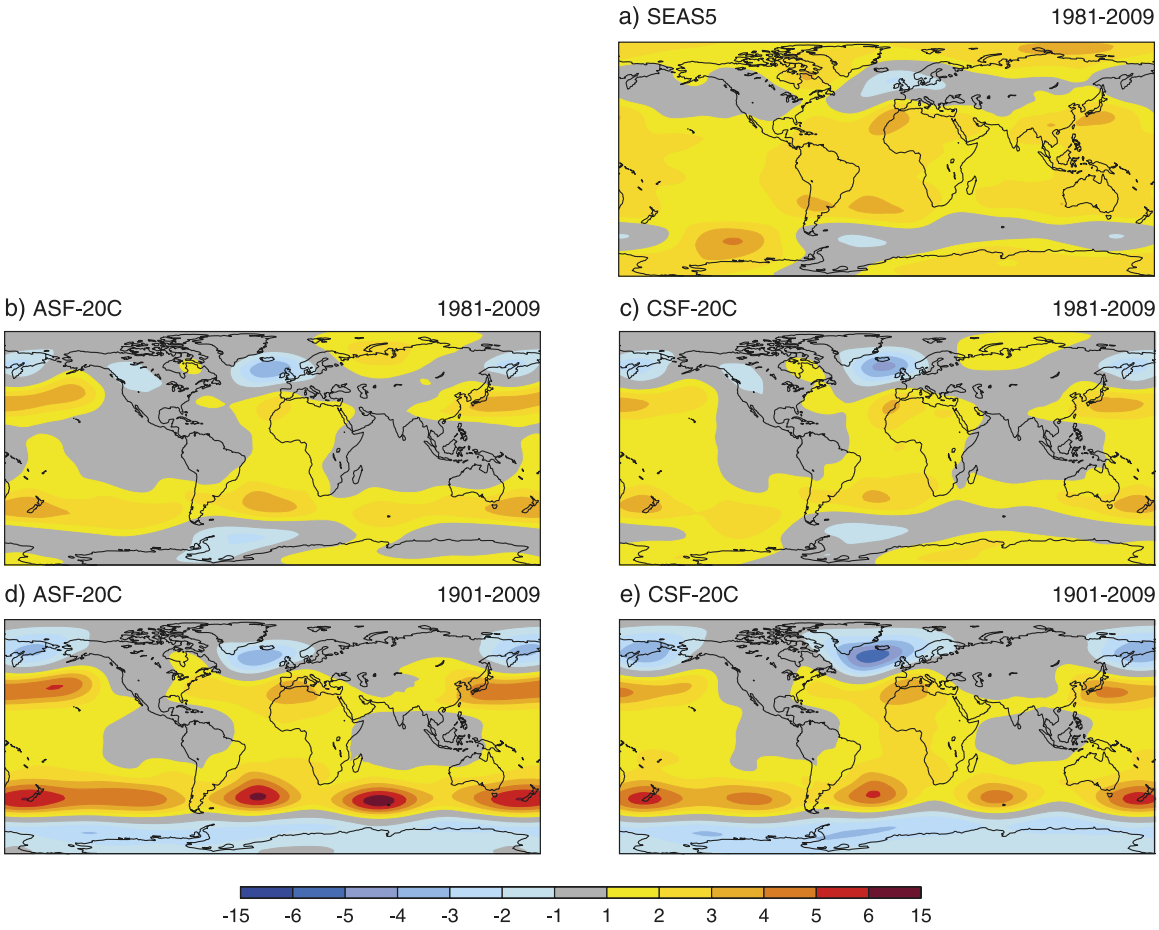


Fig. 5. Bias of geopotential height at 500 hPa (in geopotential meters) in DJF from forecasts initialized in November during the common hindcast period 1981–2009 for (a) SEAS5, (b) atmosphere-only seasonal forecasts ASF-20C, and (c) coupled seasonal forecasts CSF-20C and for the full hindcast period 1901–2009 for (d) ASF-20C and (e) CSF-20C. The verification data used are ERA-20C.

hindcasts have remarkably similar midtropospheric circulation biases which are, in terms of spatial distribution and magnitude, comparable with ECMWF's operational hindcasts SEAS5. The small differences between ASF-20C/CSF-20C and SEAS5 can be attributed to the differences in the model cycle and the finer horizontal resolution used. All model runs have a bias that projects positively onto the spatial pattern of the NAO with a negative Z500 bias over the North Atlantic near Iceland and a positive bias over the subtropical North Atlantic, reflecting a too strong westerly flow. ASF-20C and CSF-20C also have similar biases over the Aleutian Islands and the North Pacific, although the positive North Pacific bias seems smaller in the coupled hindcasts.

These Northern Hemisphere biases are fairly consistent when computed over the full 110-year hindcast period. An overall increase of the magnitude compared to the recent period is, however, noted. Larger increases in bias can be seen for the Southern Hemisphere which might be a consequence of the very poor data coverage there.

Figure 6 shows maps of the anomaly correlation skill for Z500 in DJF for the atmosphere-only ASF-20C and coupled CSF-20C hindcasts as well as for ECMWF's operational hindcasts SEAS5. For the common hindcast period of all three systems 1981–2009 (Figs. 6a–c) skill is highest in SEAS5, particularly so in the tropics. In the Northern Hemisphere extratropics, all three forecasting systems show similar areas of higher skill over parts of the northeast Pacific, North America, and Greenland. The skill is noticeably lower in all systems over the northeast Atlantic, Europe, and northern Asia.

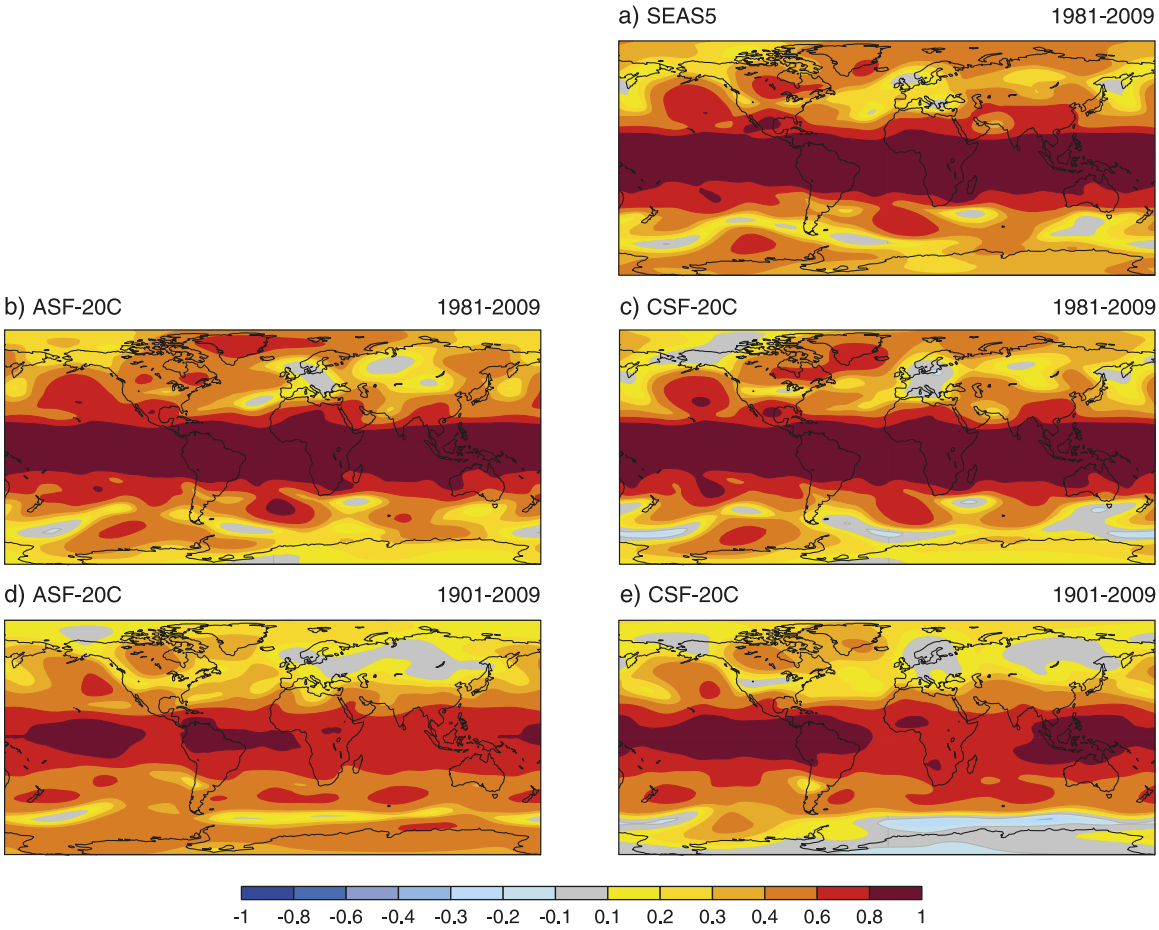


Fig. 6. Anomaly correlation skill of geopotential height at 500 hPa in DJF from forecasts initialized in November. During the common hindcast period 1981–2009 for (a) SEAS5, (b) atmosphere-only seasonal forecasts ASF-20C, and (c) coupled seasonal forecasts CSF-20C. For the full hindcast period 1901–2009 for (d) ASF-20C and (e) CSF-20C. The verification data used are ERA-20C.

The skill maps for the long hindcast period 1901–2009 (Figs. 6d,e) for both ASF-20C and CSF-20C are, broadly speaking, similar to those over the shorter recent hindcast period. In particular the spatial structures of skillful areas are quite similar. The magnitude of skill over the longer period seems slightly reduced almost everywhere, which implies that skill was overall lower before the 1980s.

In previous studies of the ASF-20C hindcasts it was found that prediction skill for two important circulation patterns of the Northern Hemisphere extratropics, the NAO index and the PNA pattern index, has undergone distinct multidecadal variability, with a decrease in skill during a midcentury period and higher skill before and after that central period (Weisheimer et al. 2017; O'Reilly et al. 2017; MacLeod et al. 2018). Figure 7a shows the NAO correlation skill estimated for moving 30-year hindcast windows from 1901–30 to 1980–2009 for ASF-20C in red and for CSF-20C in blue. The skill estimates over the full 110-year period are shown on the right of the plot, indicating that correlations for both hindcasts are highly significant ($r = 0.3$).

Interestingly, the NAO skill for the latest 30-year period is slightly higher in the coupled CSF-20C hindcasts ($r = 0.55$) than the atmosphere-only ASF-20C hindcasts ($r = 0.44$). In both hindcasts sets the skill decreases for earlier periods with a distinct midcentury minimum in skill period the 1950s to the 1970s. It is interesting to note that both hindcasts show an abrupt ~0.2 drop in correlation skill centered around the year 1972. The decrease of skill during midcentury periods cannot simply be explained by the potentially lower quality of the reanalyses (due to fewer observations) in earlier decades used to initialize the forecasts because the

skill in both hindcast datasets increases again before the midcentury low-skill period. Indeed, the skill reaches levels during the early decades that are comparable with the skill in recent decades. It should be stressed that the reanalyses ERA-20C and CERA-20C do not assimilate any satellite data and are thus not affected by the much fewer observations in the presatellite era. Uncertainty around the correlation skill estimates for the NAO over 30-year periods is high, as indicated by the shading in Fig. 7.

The PNA index is, in general, more predictable than the NAO index due to its stronger direct links to ENSO (see Fig. 7b) (Shukla et al. 2000). Over the 110-year period the correlation skill in both ASF-20C and CSF-20C hindcasts is around 0.5. Similar to the NAO, the skill to predict the PNA is also nonstationary and shows clear multidecadal variations. Interestingly, the overall temporal skill behavior follows closely the evolution of NAO skill: very skillful periods are the recent hindcast periods and toward the beginning of the twentieth century, while the midcentury decades are characterized by a pronounced drop in skill. These multidecadal fluctuations in predictability are very similar in the atmosphere-only and fully coupled hindcasts. O'Reilly et al. (2017) explained the midcentury skill drop in the ASF-20C hindcasts by negative PNA events, which were not forced in a predictable manner by tropical Pacific SST anomalies. In particular, it was found that the correlation between tropical east Pacific SSTs and the PNA index in reanalyses strongly covaries with the PNA skill. We hypothesize that a similar mechanism is active in the coupled CSF-20C hindcasts.

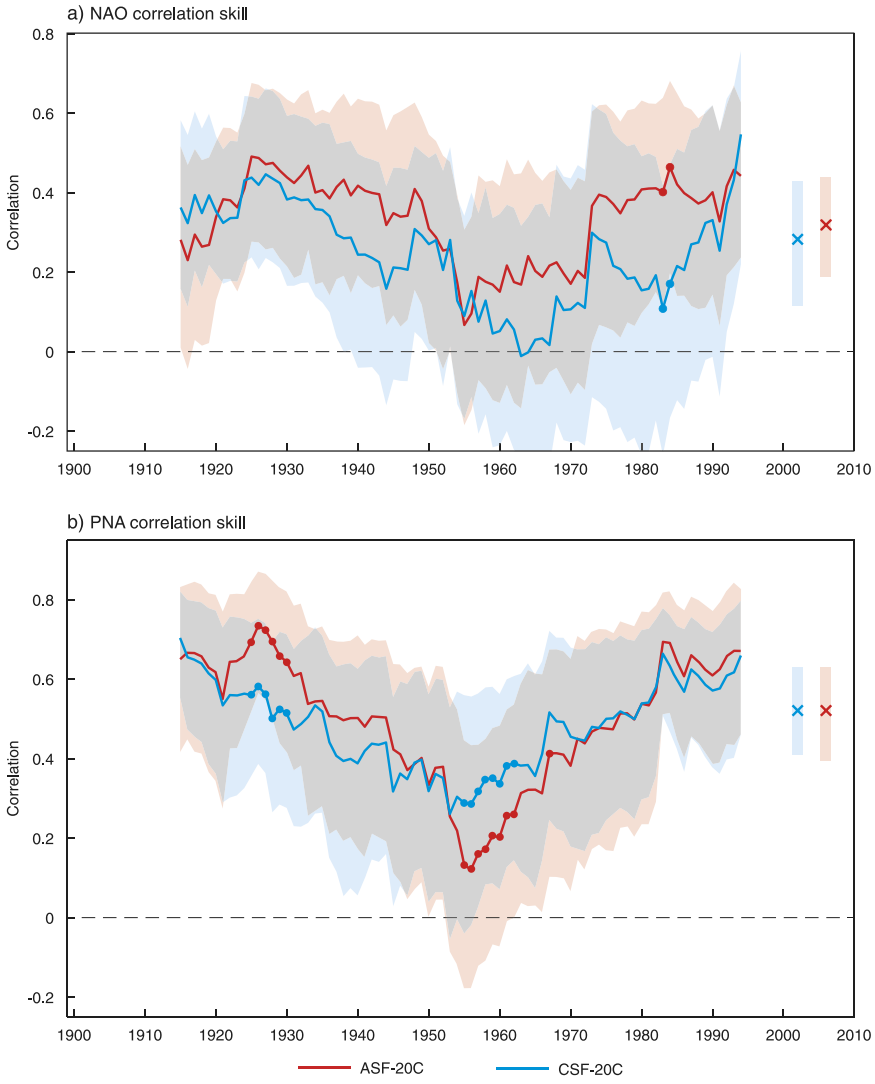


Fig. 7. Decadal variability of extratropical forecast skill of the (a) NAO and (b) PNA in DJF from forecasts initialized in November with 30-year moving-window correlation coefficients of the hindcast ensemble mean with CERA-20C. Atmosphere-only seasonal forecasts ASF-20C are in red and coupled seasonal forecasts CSF-20C are in blue. Shading indicates the 5%–95% confidence intervals. Correlations for each 30-year window are plotted at the central year. Bars on the right show correlations and confidence intervals over the full hindcast period 1901–2009. The circles on the lines show where the difference in skill between ASF-20C and CSF-20C is significant at the 95% level. The significance was calculated using a Monte Carlo significance test in which the two ensembles are shuffled and split into two random ensembles; the difference in the correlation of these two random ensembles is then calculated and this is repeated 10,000 times to give the final significance levels.

An ongoing debate within the seasonal forecasting community concerns the level of real-world predictability versus perfect-model predictability estimates, sometimes referred to as signal-to-noise paradox. It has been argued (Eade et al. 2014; Scaife and Smith 2018) that the

low strength of the predictable signal in forecast models is inconsistent with the relatively high level of correlation skill with reanalysis datasets. A measure often used to diagnose the consistency of the predictable components in the model and the real world is the ratio of predictable components [RCP; see Eade et al. (2014) for a definition]. If the real world had the same level of predictability as a well-calibrated model, the RPC would be 1. If the real world was less (more) predictable than the model world, RPC would be smaller (larger) than 1, implying an overconfident (underconfident) forecasting model where the ensemble is underdispersive (overdispersive).

Weisheimer et al. (2019) showed that while the ASF-20C hindcasts for the recent period indeed are underconfident (RPC > 1), if estimated over the longer 110-year hindcast period the RPC is almost exactly 1. Here we present in Fig. 8 the temporal evolution of the RPC for both the ASF-20C and the CSF-20C hindcasts. The RPC for CSF-20C over the 110-year period suggests a weakly underconfident system (RPC ~ 1.2). Similarly to ASF-20C, the RPC for CSF-20C undergoes extended periods of time with values above, close to and below 1. The reported signal-to-noise paradox thus is strongly dependent on the specific hindcast period analyzed and does not appear to be a general characteristic of our forecasting systems. Noticeably, for the midcentury period of reduced skill the forecasts are overconfident (RPC < 1), which means that the signal-to-noise paradox is absent. The main difference between ASF-20C and CSF-20C is a slightly larger RPC over the full hindcast period for CSF-20C, which seems to stem from the early decades of the twentieth century where in RPC values in CSF-20C are larger than 1 and reach up to 2 for specific periods. It should be noted that the confidence intervals for the RPC estimates, as given by the shading in Fig. 8, indicate a rather large sampling uncertainty of the RPC. For ASF-20C the ideal situation of RPC = 1 is within the confidence interval for every 30-year period in the full hindcasts. For CSF-20C this is also the case with the exception of one period centered around the year 1927.

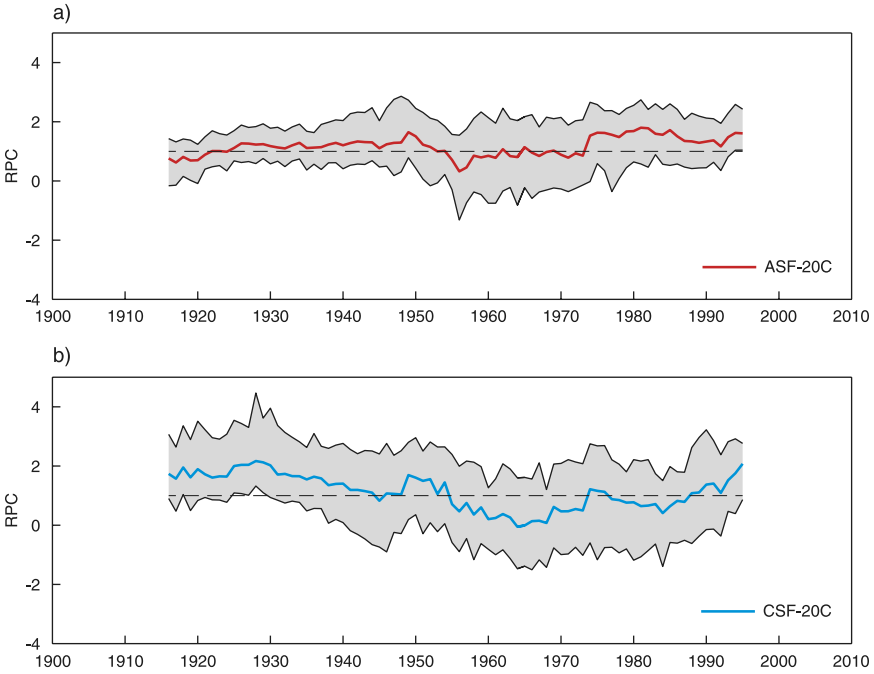


Fig. 8. Decadal variability of the RPC of the NAO in DJF forecasts initialized in November for 30-year moving windows. (a) Atmosphere-only seasonal forecasts ASF-20C are in red and (b) coupled seasonal forecasts CSF-20C are in blue. Shading indicates the 5%–95% confidence intervals. The black dashed lines show the perfect RPC = 1. Values for each 30-year window are plotted at the central year. The RPCs over the full hindcast period 1901–2010 are 1.05 for ASF-20C and 1.19 for CSF-20C.

The underlying dynamics of the NAO are closely related to variability of the Atlantic eddy-driven component of the zonal flow. Based on diagnostics of low-level wind fields, Woollings et al. (2010) suggested that there are three preferred latitudinal positions of the North Atlantic eddy-driven jet stream in winter that relate to patterns of variability like the NAO or the east Atlantic pattern. Recently, Parker et al. (2019) diagnosed how well the trimodal distribution of jet latitude is represented in the ASF-20C hindcasts. They concluded that the jet latitude regimes were simulated reasonably well and showed that the NAO skill is largely related to interannual variations in jet latitude. We extended the results of Parker et al. (2019) to include the coupled hindcasts

CSF-20C and similar conclusions can be drawn (see Fig. 9). It was found that the locations of the peaks are fairly stable across the twentieth century, suggesting that decadal variability of the jet are more related to changes in the frequency of occurrence of each latitude regime.

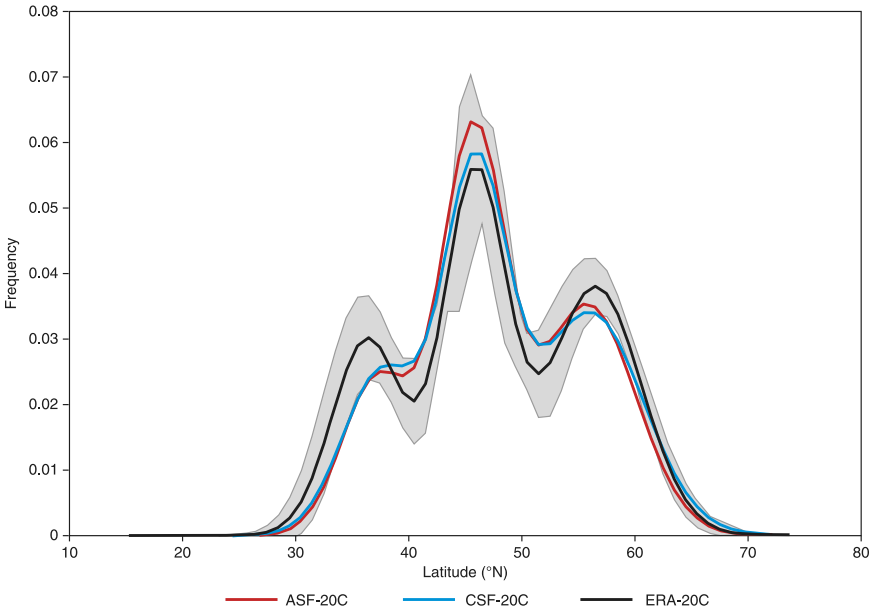


Fig. 9. Frequency distribution of North Atlantic eddy-driven jet stream latitude for all days during DJF for forecasts initialized in November in 1901–2009 for ERA-20C (black line and gray uncertainty range), ASF-20C (red), and CSF-20C (blue). Uncertainty estimates for ERA-20C (5%–95% shaded confidence intervals) were obtained via a Monte Carlo resampling method.

Dissemination of data

The ASF-20C and CSF-20C hindcasts provide ideal datasets to study a wide range of predictability problems, some of which we have briefly outlined above. We invite the scientific community to explore the datasets for their specific research interests, which could include regional assessments of predictability, physical mechanisms in large-scale variability, or the use of modern machine learning techniques.

Data from ASF-20C and CSF-20C have become publicly available through a dedicated online dissemination platform hosted by the CEDA archive at https://catalogue.ceda.ac.uk/uuid/6e1c3df49f644a0f812818080bed5e45. A set of standard monthly mean atmospheric variables including temperature, precipitation, mean sea level pressure, geopotential height, wind, and thermal and radiative fluxes have been provided as global gridded data in netCDF format.

Summary and conclusions

A crucial component of any seasonal forecast system is the retrospective forecasts, or hindcasts, from past years that are used to estimate skill and to calibrate the forecasts. Hindcasts for operational systems are usually produced over a period of around 20–30 years (Saha et al. 2014; MacLachlan et al. 2015; Johnson et al. 2019). However, several studies (Derome et al. 2005; Kumar 2009; Shi et al. 2015; Weisheimer et al. 2017; O'Reilly et al. 2017; Weisheimer et al. 2019; MacLeod et al. 2018; O'Reilly et al. 2019) found that seasonal forecast skill is nonstationary and can undergo pronounced variations at multidecadal time scales. These results imply that relatively short hindcasts are not representative for the longer-term behavior of seasonal forecasts and that small hindcast sample sizes can lead to skill estimates that are not robust.

To overcome these limitations, a new and unprecedented set of 110-year-long seasonal hindcasts based on ensemble simulations with ECMWF's coupled atmosphere–ocean–sea ice model over the period 1901 to 2010 was created using data from the coupled reanalysis CERA-20C (Laloyaux et al. 2018) to initialize the hindcasts. The purpose of this article is to (i) introduce these data to the wider scientific community, (ii) demonstrate state-of-the-art levels of forecast performance, (iii) discuss a few selected aspects of multidecadal variability in forecast skill, and (iv) advertise their public availability through a dedicated dissemination server. We mainly focus here on the DJF season but hindcast data for the other main seasons are also available. Retrospective predictions are made for only one season ahead due to computational constraints. The methodology to initialize forecasts with reanalyses of

the twentieth century is, however, not restricted to seasonal time scales. Multiyear forecasts with a specific focus on ENSO are currently being explored at ECMWF and results will be published in due course.

We have shown that the model forecast quality is, in general, in good agreement with the performance of operational models for recent hindcast periods. We have also demonstrated the coupled hindcasts were able to reproduce the correct temporal evolution and variability of the global-mean surface temperature in boreal winter from 1901 to 2009. Consistent with previous findings based on atmosphere-only hindcasts using prescribed SSTs (Weisheimer et al. 2017; O'Reilly et al. 2017; Weisheimer et al. 2019; MacLeod et al. 2018; O'Reilly et al. 2019), a pronounced drop in forecast skill during earlier periods of the twentieth century is found for a range of variability modes including ENSO (both in terms of SST as well as atmospheric pressure gradient manifestations), the North Atlantic Oscillation and the Pacific–North American pattern. For all of these forecasts the skill recovers near the beginning of the century to similarly high values as for the most recent periods. It is intriguing to note that purely statistical predictions of the winter NAO from 1900 to 2001 reveal a very similar temporal evolution of forecast skill (Fletcher and Saunders 2006).

The reduction in skill during early and midcentury periods thus cannot simply be attributed to the poorer observational coverage during this period, and we hypothesize instead that these changes in model predictability are linked to intrinsic changes of the coupled climate system. These conclusions are further supported by a skill analysis of forced climate simulations over the twentieth century with observed SSTs (AMIP-type) that shows synchronous periods of reduced extratropical skill in midcentury periods and higher skill toward the beginning of the century (O'Reilly et al. 2020). How much these changes of predictive skill of extratropical teleconnections are linked to changes in the tropics remains an open question, but there is increasing evidence that this is linked to the multidecadal variability in the strength of ENSO teleconnections to the extratropics (O'Reilly et al. 2017). For example, Huang et al. (2018) report an interesting midcentury change in the observed relationship between quasi-decadal SST oscillations in the tropical west Pacific and the western subtropical North Pacific, a region of great importance for Rossby wave sources and the teleconnections into the Northern Hemisphere.

The new ASF-20C and CSF-20C datasets, with their large samples both in terms of hindcast years and ensemble size, provide a unique opportunity to study a wide range of predictability problems with a higher level of robustness and confidence than otherwise possible. In addition to the nonstationary forecast skill behavior discussed in this manuscript, such prospective studies could, for example, involve the assessment of distinct types of climate events including extremes, exploit a range of applications of seasonal forecasts, or train modern machine learning techniques. Data from our hindcast experiments that include simulations for all four seasons are publicly available to download from our dissemination server, and we welcome a wide exploration of these datasets by the community.

Acknowledgments. The computing resources for ASF-20C and CSF-20C were kindly provided through the ECMWF special projects "Seasonal forecasts of the 20th century: Reliability, attribution and the impact of stochastic perturbations" and "Coupled seasonal forecasts of the 20th century." We thank ECMWF for the provision of the ERA-20C and CERA-20C reanalysis data and the long-range prediction team at ECMWF for discussions. AW, DB, CO, and TP acknowledge support from the EU H2020 project EUCP (Grant Agreement 776613). KJS acknowledges funding from the European Commission under Grant Agreement 641727 of the Horizon 2020 research programme. We thank one anonymous reviewer, Matt Newman, and Ben Kirtman for their support and constructive criticism which helped improve the paper.

References

Compo, G., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, https://doi.org/10.1002/qj.776.

Derome, J., H. Lin, and G. Brunet, 2005: Seasonal forecasting with a simple general circulation model: Predictive skill in the AO and PNA. *J. Climate*, **18**, 597–609, https://doi.org/10.1175/JCLI-3289.1.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, https://doi.org/10.1002/2014GL061146.

Fletcher, C., and M. A. Saunders, 2006: Winter North Atlantic hindcast skill: 1900–2001. *J. Climate*, **19**, 5762–5776, https://doi.org/10.1175/JCLI3949.1.

Huang, B., C.-S. Shin, J. Shukla, L. Marx, M. A. Balmaseda, S. Halder, P. Dirmeyer, and J. L. Kinter III, 2017: Reforecasting the ENSO events in the past 57 years (1958–2014). *J. Climate*, **30**, 7669–7693, https://doi.org/10.1175/JCLI-D-16-0642.1.

Huang, W. R., S.-Y. Simon Wang, and B. T. Guan, 2018: Decadal fluctuations in the western Pacific recorded by long precipitation records in Taiwan. *Climate Dyn.*, **50**, 1597–1608, https://doi.org/10.1007/s00382-017-3707-9.

Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.,* **12,** 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Kumar, A., 2009: Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Mon. Wea. Rev.*, **137**, 2622–2631, https://doi.org/10.1175/2009MWR2814.1.

——, and M. Chen, 2018: Causes of skill in seasonal predictions of the Arctic Oscillation. *Climate Dyn.*, **51**, 2397–2411, https://doi.org/10.1007/s00382-017-4019-9.

Laloyaux, P., and Coauthors, 2018: CERA-20C: A coupled reanalysis of the twentieth century. *J. Adv. Model. Earth Syst.*, **10**, 1172–1195, https://doi.org/10.1029/2018MS001273.

MacLachlan, C., and Coauthors, 2015: Global Seasonal Forecast System version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, https://doi.org/10.1002/qj.2396.

MacLeod, D., C. O'Reilly, T. N. Palmer, and A. Weisheimer, 2018: Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmos. Sci. Lett.*, **19**, e815, https://doi.org/10.1002/asl.815.

O'Reilly, C. H., 2018: Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dyn.*, **51**, 3333–3350, https://doi.org/10.1007/s00382-018-4081-y.

——, J. Heatley, D. MacLeod, A. Weisheimer, T. N. Palmer, N. Schaller, and T. Woollings, 2017: Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophys. Res. Lett.*, **44**, 5729–5738, https://doi.org/10.1002/2017GL073736.

——, T. Woollings, L. Zanna, and A. Weisheimer, 2019: An interdecadal shift of the extratropical teleconnection from the tropical Pacific during boreal summer. *Geophys. Res. Lett.*, **46**, 13 379–13 388, https://doi.org/10.1029/2019GL084079.

——, A. Weisheimer, D. MacLeod, D. J. Befort, T. Palmer, 2020: Assessing the robustness of multidecadal variability in Northern Hemisphere wintertime seasonal forecast skill. *Quart J. Roy. Meteor. Soc.*, https://doi.org/10.1002/qj.3890, in press.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-To-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, https://doi.org/10.1175/BAMS-85-6-853.

Parker, T., T. Woollings, A. Weisheimer, C. O'Reilly, L. Baker, and L. Shaffrey, 2019: Seasonal predictability of the winter North Atlantic Oscillation from a jet stream perspective. *Geophys. Res. Lett.*, **46**, 10 159–10 167, https://doi.org/10.1029/2019GL084402.

Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *J. Climate*, **29**, 4083–4097, https://doi.org/10.1175/JCLI-D-15-0556.1.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Scaife, A. A., and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate Atmos. Sci.*, **1**, 28, https://doi.org/10.1038/s41612-018-0038-4.

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.*, **42**, 1554–1559, https://doi.org/10.1002/2014GL062829.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.*, **81**, 2593–2606, https://doi.org/10.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2.

Webster, P. J., and S. Yang, 1992: Monsoon and ENSO: Selectively interactive systems. *Quart. J. Roy. Meteor. Soc.*, **118**, 877–926, https://doi.org/10.1002/qj.49711850705.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, https://doi.org/10.1029/2009GL040896.

——, N. Schaller, C. O'Reilly, D. MacLeod, and T. N. Palmer, 2017: Atmospheric seasonal forecasts of the twentieth century: Multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quart. J. Roy. Meteor. Soc.*, **143**, 917–926, https://doi.org/10.1002/qj.2976.

——, D. Decremer, D. MacLeod, C. O'Reilly, T. Stockdale, S. Johnson, and T. N. Palmer, 2019: How confident are predictability estimates of the winter North Atlantic Oscillation? *Quart. J. Roy. Meteor. Soc.*, **145**, 140–159, https://doi.org/10.1002/qj.3446.

Woollings, T., A. Hannachi, and B. Hoskins, 2010: Variability of the North Atlantic eddy-driven jet stream. *Quart. J. Roy. Meteor. Soc.*, **136**, 856–868, https://doi.org/10.1002/qj.625.

Yeh, S.-W., and Coauthors, 2018: ENSO atmospheric teleconnections and their response to greenhouse gas forcing. *Rev. Geophys.*, **56**, 185–206, https://doi.org/10.1002/2017RG000568.