universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Assessment of temperature trends of atmospheric seasonal forecasts of the 20$^{\text{th}}$ century“

verfasst von / submitted by

Markus Rosenberger, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066614
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Meteorologie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Leopold Haimberger

Acknowledgements

Apart from my supervisor Univ.-Prof. Mag. Dr. Leopold Haimberger, I want to thank Dr. Michael Mayer for guiding me through this work and being patient with me during the whole process.

Abstract

In this work the performance and utility of a centennial seasonal hindcast data set, namely ASF-20C, which consists of 51 ensemble members and provides monthly averages of several quantities for a lead time up to 4 months during the period 1901–2010, is tested. Precisely, in this work only 2 m temperature values of the months November–February and May–August are considered. Re-forecast data is compared to two different reanalyses: ERA-20C, a global atmosphere-only model that was used to initialize ASF-20C and thus covers the same period and ERA5, a state-of-the-art global reanalysis, which provides output from 1950 onwards.

To quantify the goodness of ASF-20C output, at first different verification measures like bias, Pearson correlation and reliability are used. After proving that the hindcasts deliver proper results, especially when anomalies are considered, signal-to-noise ratios (SNRs) of the hindcast ensemble in different continental and oceanic regions around the world are calculated. Time series of SNRs show correlations above 0.7 with El Niño Southern Oscillation (ENSO) activity in oceans and even higher in continents, which confirms that ENSO is a major provider of global temperature forecast skill. Similarly, high correlations with the index of the North Atlantic Oscillation (NAO) in continents and oceans of the northern hemisphere indicate that NAO provides a significant amount of forecast skill in these regions.

A major issue concerning temperature in the recent decades are trends. Results show that ASF-20C trends during several different decades covering the whole 110-year period are very similar to reanalyses trends. While at the beginning and end of the period positive trends dominate globally, around the middle of the century trends are slightly negative. However, on more local scales deviations can be rather large.

Due to the large ensemble size of ASF-20C, investigations of probability distributions can be done very well. If corrected for inherent trends, agreement between the three data sets on these distributions as well as on their changes over time is very high in most regions and all periods. Explicitly, changes of standard deviation towards the more recent period are reproduced very well by the hindcast, especially over oceans. Since reanalyses only deliver a single value per year, distribution sizes are much smaller and thus resolution is poorer compared to the large ASF-20C ensemble. This is also the reason why changes of the distribution's extreme percentiles are much better represented by re-forecasts.

Thus, from this work it can be concluded that an atmosphere-only seasonal forecast like ASF-20C delivers very useful results even on time scales of more than one century but also that a large ensemble size is essential for this type of prediction model.

Kurzfassung

Das Ziel dieser Arbeit ist es, sowohl die Performance als auch die Nützlichkeit der Saisonprognosen des ASF-20C Datensatzes zu untersuchen. Sein Ensemble besteht aus 51 Mitgliedern, deren monatliche Werte für eine Vielzahl an Größen für einen Vorhersagezeitraum von bis zu 4 Monaten während der Periode 1901–2010 verfügbar sind. Im Verlauf der Arbeit werden lediglich die Monate November–Februar und Mai–August betrachtet. Als Referenz dienen zwei Reanalyseprodukte: Zum einen ERA-20C, eine globale Reanalyse, die lediglich die Atmosphäre modelliert und zur Initialisierung von ASF-20C verwendet wurde. Zum anderen ERA5, eines der modernsten und am besten entwickelten globalen Reanalysesysteme. Sein Output deckt den Zeitraum von 1950 bis heute ab.

Um die Prognosegüte zu quantifizieren, werden diverse Verifikationsmaße wie Bias, Pearson Korrelation und Reliability verwendet. Sie alle zeigen, dass die Daten der saisonalen Prognosen konsistent mit jenen der Reanalysen sind, vor allem wenn Anomalien verwendet werden. Außerdem wird das Signal-zu-Rausch Verhältnis (SNR) in diversen Kontinenten und Ozeanen berechnet. SNR Zeitreihen zeigen sehr hohe Korrelationen von ca. 0.7 mit der El Niño Southern Oscillation (ENSO) Aktivität in Ozeanen und sogar noch höhere Werte über Kontinenten. Das zeigt, dass ENSO einen sehr großen Beitrag zur globalen Vorhersagequalität der 2 m Temperatur leistet. Ähnlich dazu sind Korrelationen mit dem Index der North Atlantic Oscillation (NAO) in der Nordhemisphäre im Sommer sehr hoch, dort ist also auch die NAO wichtig für die Prognosegüte.

Ein viel diskutiertes Thema rund um Temperaturen in den vergangenen Jahrzehnten sind Trends. Es zeigt sich, dass über die gesamte 110-jährige Periode die Trends in allen Datensätzen relativ ähnlich sind. Während am Anfang und gegen Ende der Periode auf der globalen Skala positive Trends dominieren, sind sie um die Mitte des Jahrhunderts herum sogar leicht negativ. Lokaler betrachtet, können die Unterschiede zwischen den Datensätzen jedoch recht groß ausfallen.

Da ASF-20C aus einem sehr großen Ensemble besteht, können Wahrscheinlichkeitsverteilungen seiner Werte sehr gut untersucht werden. Nachdem die Trends der einzelnen Datensätze angepasst wurden, zeigt sich, dass die Übereinstimmung der Wahrscheinlichkeitsverteilungen und auch deren Änderungen in allen Datensätzen für alle betrachteten Perioden sehr groß ist. Außerdem können die saisonalen Prognosen auch die Änderung der Standardabweichung der Verteilungen sehr gut abbilden, vor allem über Ozeanen. Da deterministische Reanalysen lediglich eine Realisierung pro Jahr liefern, verfügen ihre Verteilungen über wesentlich weniger Mitglieder, wodurch auch die Auflösung deutlich schlechter ist. Aus dem selben Grund sind auch die Änderungen extremer Perzentile zwichen zwei verschiedenen Perioden um einiges besser in ASF-20C dargestellt.

Das Ergebnis dieser Arbeit ist also einerseits, dass Saisonprognosen wie ASF-20C, bei denen lediglich die Atmosphäre modelliert wird, sehr nützliche Ergebnisse liefern, sogar auf Zeitskalen von über einem Jahrhundert. Andererseits aber auch, dass eine derartige Größe des Ensembles notwendig ist, um für diese Art der Prognosen vernünftige Resultate erwarten zu können.

Contents

List of Tables

List of Figures

1 Introduction

For more than two decades, several meteorological organisations around the globe are running operational seasonal forecasts. The Canadian Meteorological Centre (CMC) started their first seasonal forecasts already in September 1995 (Lin et al., 2020), at the European Centre for Medium-Range Weather Forecasts (ECMWF) real-time seasonal forecast systems are running since 1997 (Johnson et al., 2019) and in August 2004 the National Centers for Environmental Prediction (NCEP) started their operational Climate Forecast System (CFS; Saha et al., 2014). Because of a broad range of possible applications for seasonal forecasts, among others agriculture (Challinor et al., 2005) and health service (Morse et al., 2005), since then a lot of effort was put into the development of physical models and forecast systems as well as into the improvement of the initial conditions that are necessary to initialise them. Often, forecast models for seasonal time scales have a lot in common with those for the medium-range, i.e. with a lead time of the order of 10 days, except for some components concerning processes that are crucial for forecasts on extended time scales (Johnson et al., 2019).

Long-range prediction systems nowadays typically use the same or a very similar version of the physical forecast model used for medium-range forecasts (a so-called seamless approach), but the aim of a seasonal forecast is different from that of a typical weather prediction. While the latter aims at prediction of the atmospheric state on scales of hours to days, the former provides estimates of monthly and seasonal means or of deviations from a long-term climatological average (Weisheimer and Palmer, 2014). Thus, also the forcings and main processes that provide predictability, change when longer lead times are achieved. On seasonal time scales these processes are mainly based on slowly varying lower boundary forcings of the atmosphere, e.g. ocean dynamics and the hydrology of continental regions (Weisheimer et al., 2017). These lower boundary forcings include sea surface temperatures (SSTs), sea ice concentrations (SICs) and soil mositure (Sigmond et al., 2013). While currently operational seasonal forecasting systems use coupled ocean-atmosphere models to calculate quantities like SSTs and SICs, another (uncoupled) approach is to prescribe the boundary forcings during the whole integration process using e.g. reanalysis output. Contrary to these lower boundary conditions, the initial conditions within the atmosphere do not seem to contribute predictive skill for lead times of more than a few weeks except for certain slow atmospheric modes (WCRP, 2007). The obvious downside of the uncoupled approach is the lack of dynamic air-sea interaction, but on the other hand development of biases in, e.g., SSTs is avoided and atmospheric predictability in presence of unbiased boundary conditions can be investigated.

Today, predictability of El Niño Southern Oscillation (ENSO), together with its teleconnections, is considered as main source of forecast skill on seasonal time scales in most parts of the planet (Sigmond et al., 2013). Kiladis and Diaz (1989) already stated that the Southern Oscillation and its remote effects, which have been described several decades earlier by Walker and Bliss (1932), have the potential to provide skill for long range weather forecasts and give further insights in the global climate system. van Oldenborgh

et al. (2005) considered the ensemble means of ECMWF's seasonal forecast models 1 and 2 and described the spatial extents of these ENSO teleconnections. They found that predictability of 2 m temperature that is related to ENSO influence is largest in tropical oceans, over northern South America as well as in parts of Africa and North America. Due to the large distance to the ENSO region, forecast skill is significantly reduced in the extratropics compared to tropical regions. Moreover, predictability in the extratropics is also smaller because of the higher variability induced by instabilities and non-linearities of oceans and the atmosphere in these latitudes (Weisheimer et al., 2017). On the other hand, Hurrell and Van Loon (1997) mention that the North Atlantic Oscillation (NAO) influences temperature and precipitation on a seasonal time scale over the Atlantic and Eurasia and thus provides potential predictability in these regions. Under certain conditions forecast skill in these regions can also be drawn from other phenomena. For example, Ineson and Scaife (2009) used a general circulation model of the atmosphere and found that in El Niño years where also Sudden Stratospheric Warmings (SSWs) occur, the stratosphere acts as teleconnection pathway from the Pacific towards Europe and provides forecast skill on a seasonal time scale for late boreal winter months. Similarly, Sigmond et al. (2013) stated that SSWs can increase predictability in the Atlantic and northern Eurasia. Douville (2009) found that in North America and Eurasia soil moisture and snow cover contribute to predictability in boreal summer and spring, respectively.

Since November 2017 ECMWF is already running its fifth generation seasonal forecast system, SEAS5. This 51-member ensemble forecasting system uses coupled atmosphere, ocean and cryosphere components. Initial conditions of ocean and sea ice forecast models, namely NEMO version 3.4.1 (Nucleus for European Modelling of the Ocean; Madec and the NEMO team, 2016) and LIM2 (Louvain-la-Neuve sea-ice model version 2; Fichefet and Maqueda, 1997), respectively, come from OCEAN5 (Zuo et al., 2019) reanalysis and real-time analysis. Forecasts are initialised on the first day of every month and calculated for a lead time of 7 months. For four start dates per year, namely February, May, August and November integration time is even extended to 13 months for 15 of the 51 ensemble members. Johnson et al. (2019) state that among others SEAS5 provides high prediction skill of ENSO, Arctic sea ice concentrations and 2 m temperature near the ice edge. The latter can be achieved due to the interactive sea ice model used in SEAS5. Retrospective seasonal forecasts (re-forecasts) of the period 1981–2016 with 25 ensemble members were produced for calibration of SEAS5.

NCEP is running its current CFS version 2 since March 2011 with three forecasts per day with a lead time of one season, four per day with an integration time of 9 months and 9 sub-seasonal forecasts out to 45 days. Similarly to the ECMWF, NCEP ran re-forecasts for different periods and lead times for calibration and skill estimates of operational runs with an ensemble size of 24, except for November of each year where 28 members were available (Saha et al., 2014).

Both ECMWF and NCEP run re-forecasts (also called hindcasts) of periods with a length of the order of 30 years to calibrate their operational seasonal forecast models and indicate their skill. Though this may seem to be sufficiently long for a re-forecast data set, several studies showed that forecast skill can undergo severe changes when investigating during different periods. For example, Müller et al. (2005) found that forecast skill of the North Atlantic Oscillation (NAO) is significantly larger for the period 1987–2001 than for 1959–2001. Similarly, Shi et al. (2015) found that correlation of modeled and observed NAO indices is higher for the period 1980–2001 than for 1960–1979 but significant in both

periods. However, if the entire 42-year period is considered, correlations are smaller than in the single sub-periods and not significant any more.

In order to test the evolution of forecast skill over a more extended period, Weisheimer et al. (2017) presented the Atmospheric Seasonal Forecast data set ASF-20C, which covers the period 1901–2010. While the previously mentioned operational forecast systems use a coupled atmosphere-ocean model, ASF-20C is an uncoupled atmosphere-only forecast system with prescribed boundary conditions over oceans. This uncoupled approach avoids problems arising from the evolving reliability of ocean data available for initialization due to the lack of global sub-surface ocean data in the first half of the century. With this approach, lower boundary forcing such as that arising from ENSO can be assumed to be perfect when prescribed instead of predicted SSTs are used. Different studies were made in recent years using this data set to investigate changes of atmospheric forecast skill over the course of the 20^{th} century. They found that not only forecast skill of NAO but also of the Pacific/North American index (PNA) shows reduced values around the middle of the century and ENSO predictability is smaller during the 1930s–1950s (e.g. O'Reilly et al., 2017; Parker et al., 2019; Weisheimer et al., 2017, 2022). All of the mentioned studies also agree that these skill changes are not a consequence of the improving observational coverage towards more recent decades since predictability is higher at the beginning of the century compared to the respective periods with lower forecast skill.

Since two of the most important providers of predictability on seasonal time scales, namely ENSO and NAO, exhibit quite large changes during the 20^{th} century, the aim of this work is, among others, to investigate if 2 m temperature forecast skill is affected by these variations. Moreover, the unprecedented size of the hindcast ensemble with 51 members over a period of 110 years also allows one to address changes of 2 m temperature probability distributions and extreme event probabilities over the course of several decades. Compared to a single deterministic forecast, the ASF-20C ensemble provides 51 times more realizations, which are both possible and plausible considering the given initial conditions and the variability of the atmosphere. Thus, Weisheimer and Palmer (2014) stated that due to the atmosphere's chaotic nature any deterministic seasonal forecast is unreliable and hence also untrustworthy and therefore the aim of this work is also to show the advantages of an ensemble forecast model over a single deterministic run.

The structure of this work is as follows: in Sect. 2 some general remarks about the different considered regions are provided together with details about the forcing due to volcanic activity and ENSO, which is present in the hindcast data. In Sect. 3 the different data sets used throughout this work, namely the ASF-20C hindcast ensemble, and the reanalyses ERA-20C and ERA5, are described, followed by Sect. 4 where different verification measures are used to compare them. Sect. 5 is about 2 m temperature trends in the hindcast and reanalyses and in Sect. 6 probability distributions as well as their changes are investigated. Sect. 7 concludes this work with a summary of the achieved results.

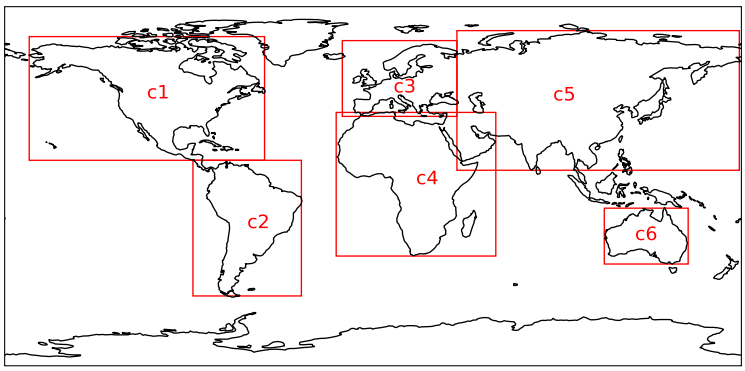
2 General remarks

2.1 Considered regions

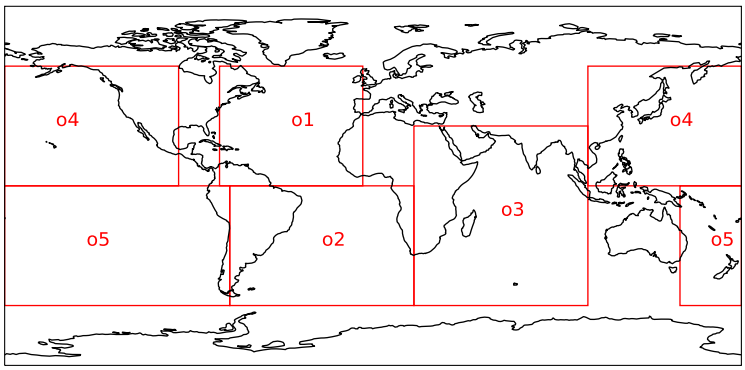
During this work, 2 m temperature data is going to be investigated not only on a global but also on a continental or sub-continental scale. Fig. 2.1 summarizes the defined boxes for each continent and ocean and Tab. 2.1 explicitly shows the border coordinates of these and also the sub-continental regions that are not shown in the maps. Within each of these regions a land-sea mask from the Copernicus Climate Change Service (C3S; Thépaut et al., 2018) Climate Data Store (Raoult et al., 2017) is applied. I define the border between land and sea as 50% ratio of land in a given grid box, as done in Hersbach et al. (2020).

Table 2.1: Border coordinates of (sub-)continental and ocean boxes as defined in this work.

Name	Longitude range	Latitude range	Label in Fig. 2.1
North America	168°W–53°W	13°N–75°N	c1
South America	88°W–35°W	55°S–13°N	c2
Europe	15°W–41°E	35°N–73°N	c3
Africa	18°W–60°E	35°S–37°N	c4
Asia	41°E–179°E	8°N–78°N	c5
Australia	113°E–154°E	39°S–11°S	c6
Antarctic	180°W–180°E	90°S–66.6°S	
Greater Alpine Region (GAR)	4°E–19°E	43°N–49°N	
Extrapolar	180°W–180°E	60°S–60°N	
Tropics	180°W–180°E	30°S–30°N	
North Atlantic Ocean	75°W–5°W	0°N–60°N	o1
South Atlantic Ocean	70°W-20°E	60°S–0°N	o2
Indian Ocean	20°E–105°E	60°S–30°N	o3
North Pacific Ocean	105°E–95°W	0°N–60°N	o4
South Pacific Ocean	150°E–70°W	60°S–0°N	o5
Tropical Atlantic Ocean	69°W–20°E	20°S–20°N	
North Atlantic Extratropical (ET)	75°W–5°W	20°N–60°N	
South Atlantic ET	69°W–20°E	60°S–20°S	
Tropical Indian Ocean	20°E–105°E	20°S–20°N	
Indian Ocean ET	20°E–125°E	60°S–20°S	
Tropical Pacific Ocean	125°E–70°W	20°S–20°N	
North Pacific Ocean ET	110°E–100°W	20°N–60°N	
South Pacific Ocean ET	125°E–70°W	60°S–20°S	
Arctic	180°W–180°E	66.6°N–90°N	



(a)



(b)

Figure 2.1: (a) Overview of the boxes defined on the continental scale. Names of the regions throughout this work are *North America* (c1), *South America* (c2), *Europe* (c3), *Africa* (c4), *Asia* (c5) and *Australia* (c6). (b) Same as in (a) but for oceans. Names are *North Atlantic* (o1), *South Atlantic* (o2), *Indian Ocean* (o3), *North Pacific* (o4) and *South Pacific* (o5). Additional information and further regions are presented in Tab. 2.1.

2.2 Volcanic eruptions

Hersbach et al. (2015) mention 6 major volcanic eruptions between 1901 and 2010 that can be seen in volcanic aerosol data of the Goddard Institute for Space Studies (GISS)[1]. Namely, these 6 eruptions were Santa María (Guatemala) in 1902, Novarupta (Alaska, United States) in 1912, Mount Agung (Indonesia) in 1963, La Cumbre (Galápagos Islands, Ecuador) in 1968, El Chicón (Mexico) in 1982 and Mount Pinatubo (Philippines) in 1991. Each of them influenced the zonal mean of stratospheric aerosol optical depth at 550 nm for at least a few years after the incident. Especially the eruptions of Santa María and Mount Pinatubo had a huge impact on the optical depth at all latitudes. The same is valid for the La Cumbre incident, though the effect was much smaller, probably because it was the weakest of all mentioned eruptions[2]. The eruptions of Mt. Agung and El Chicón had

[1] see `http://data.giss.nasa.gov/modelforce/strataer/` for figures and data

[2] see `https://de.wikipedia.org/wiki/Liste_grosser_historischer_Vulkanausbrüche` for a list of volcanic eruptions and details

larger effects on the optical depth in the Southern and Northern Hemisphere, respectively and smaller but still significant impacts on the stratosphere of the other hemisphere. In contrast to this, the Novarupta incident only affected the optical depth in the stratosphere of the Northern Hemisphere.

2.3 El Niño Southern Oscillation

The major source of interannual climate variability all over the world but especially in tropical regions is El Niño-Southern Oscillation (ENSO), where El Niño and Southern Oscillation are the oceanic and atmospheric components of this phenomenon, respectively. While it is said that Peruvian fisherman named El Niño after its occurrence around Christmas, the Southern Oscillation was first described by Sir Gilbert T. Walker (Walker and Bliss, 1932; Walker, 1933). It was Jacob Bjerknes more than 30 years later who linked these two phenomena (Bjerknes, 1969). He also introduced the *Walker Circulation*, which covers the tropical circumference of the whole globe, and which varies in association with the SO. The Walker Circulation in the tropical Pacific is characterized by a westward pressure gradient, which is driven by and in balance with SSTs across the equatorial Pacific Ocean, with warmer SSTs towards the Indo-Pacific Warm Pool. Thus Bjerknes (1969) considered it a thermal circulation. Very cold upwelling water near the coast of South America causes high surface pressure and prevents the air from ascending, whereas much warmer surface water in the western part of the Pacific Ocean, near Indonesia and Australia, moistens and warms the air above the ocean and leads to lower surface pressure and ascending air. The resulting pressure gradient force towards the west leads to easterly winds transporting air and water towards the western Pacific near the surface and also maintains the cool upwelling in the eastern Pacific. This motion is balanced by an eastward transport of air in the upper troposphere. These average conditions of ENSO are called *neutral phase*.

However, if the Walker Circulation weakens or even reverses, the westward surface transport and therefore also the cold upwelling of water near the coast of Peru do likewise, which leads to higher SSTs in the eastern Pacific Ocean. Associated with this, also surface pressure conditions change and the center of convective activity shifts towards the central Pacific. If these conditions are maintained over a well-defined amount of time it is called an El Niño event (see below for statistical definitions of El Niño). On the other hand, if the Walker Circulation is more pronounced than during a neutral phase, upwelling in the eastern equatorial Pacific is intensified, which leads to lower than normal SSTs and a strengthening of the normal conditions, ultimately resulting in a La Niña event.

In addition to the impact on the Walker circulation, ENSO also affects other parts of the planet. Kiladis and Diaz (1989) compared surface temperature and SST observations during warm (El Niño) and cold (La Niña) events and found positive land surface temperature anomalies during and after El Niño events around the Indian Ocean (i.e. in Indonesia, India, central and southern Africa) as well as in northern and western South America and along the Pacific coast of North America. During boreal winter and spring warm anomalies were even found in whole southern Canada towards the Atlantic coast. On the other hand, at this time surface temperatures are below normal in the southeastern states of the USA. Kiladis and Diaz (1989) found that the largest extent of positive SST anomalies during boreal winter and spring occurs over the equatorial Pacific (though they start to decrease

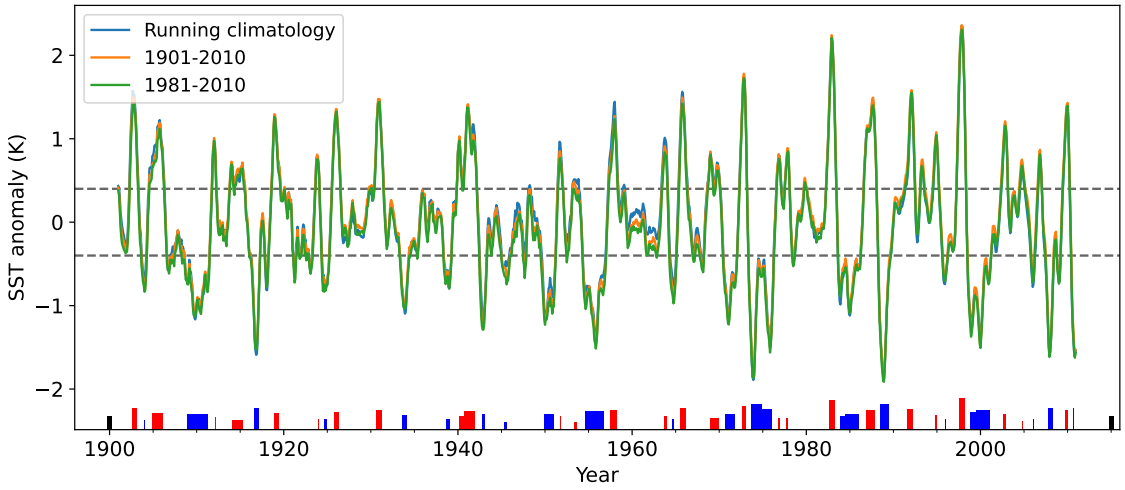


Figure 2.2: 5-month running mean of SST anomalies averaged over the Niño3.4 region using three different climatology periods, namely a 30-year running average (blue), the whole period 1901–2010 (orange), and the most recent period 1981–2010 (green). Horizontal dashed lines indicate SST anomalies of ±0.4 K. Red and blue bars indicate El Niño and La Niña events, respectively, defined as 5-month averages being for 6 consecutive months above/below these dashed lines. Height and width of bars stand for the intensity and duration of such an event, respectively. The former is defined as the highest 5-month average SST anomaly during an event normalized to the most extreme El Niño event in 1997/98. Black bars represent an event with a duration of 6 months and a maximum anomaly of 1 K.

in spring in the eastern Pacific) and Indian Ocean, the southern North Atlantic, and along the Pacific coast of North America until Alaska. Contrary to this, anomalies are negative in the central North and South Pacific towards the mid-latitudes. Kiladis and Diaz (1989) also state that the mentioned signs reverse in the case of a La Niña event in the affected regions and that surface temperatures and SSTs observed at numerous stations throughout the tropics lag SST anomalies in the eastern Pacific by about 1–2 seasons dependent on the observed region. More recently, Huang and Huang (2009) and Bosilovich et al. (2020) state that the mean tropical land temperatures lag Pacific SST anomalies by only about 3–4 months. Contrary to this, Pan and Oort (1983) showed that temperature in the tropical free atmosphere does not lag eastern Pacific SSTs at all.

Several indices were defined to describe both ENSO components using different quantities in different regions (for overviews see e.g. Trenberth, 1997; Barnston, 2015; Trenberth, 2018). Due to the coupled nature of ENSO, most of them use SST anomalies in the equatorial Pacific or pressure differences between two land stations (e.g. Darwin (Australia) and Tahiti for the Southern Oscillation Index (SOI)). In this work, I use the Niño3.4 index as defined in Trenberth (1997): If the 5-month running mean of area averaged SST anomalies between 5°S–5°N and 170°W–120°W is above/below +/-0.4 K for 6 or more consecutive months, it is considered as El Niño/La Niña event. Monthly SST data averaged over this

area is used from the Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) 1 dataset (Rayner et al., 2003; Smith, 2018).

Since a very long period of more than a century is considered in this work and conditions were not stationary throughout this period, e.g. due to global warming trends in the most recent decades, the choice of a suitable climatological period can affect results. To illustrate this, Fig. 2.2 shows 5-month averages of SST anomalies averaged over the Niño3.4 region from 1901 until 2010 using three different periods to calculate climatology. For the blue line a 30-year running average centered at the considered year was used, except for the last 5 years where the same period 1992–2021 was used because SST data was only available until January 2022 at the time of calculation. For the orange and green curves constant climatology periods 1901–2010 and 1981–2010 were used, respectively. It can be seen that the differences between anomalies are negligible most of the time. The largest differences occur around 1960 because SSTs in the eastern Pacific were lower around the middle of the century for about two decades. Moreover, anomalies calculated using the most recent period 1981–2010 as climatology tend to be the smallest during the whole century because of the recent warming trend leading to higher monthly averages.

Following the definition of the Niño3.4 index by Trenberth (1997), the time series shown in Fig. 2.2 have to be above/below the horizontal dashed lines, which indicate SST anomalies of ± 0.4 K, for 6 consecutive months for an El Niño/La Niña event to occur. These events are indicated by red and blue bars at the bottom of the panel, respectively. Bar width and height indicate duration and strength of the events, respectively. I defined the strength as the maximal 5-month average SST anomaly during an event and the corresponding bar height is normalized to the strongest El Niño event in 1997/98. The two black bars in the lower corners of the figure represent an event with a duration of 6 months and a maximum SST anomaly of 1 K. Years of the ENSO events derived in this way, are in good agreement with previous studies (e.g.: Kiladis and Diaz, 1989; Trenberth, 1997; Bosilovich et al., 2020). It has to be emphasized that since ASF-20C re-forecasts use prescribed SSTs as lower boundary conditions, one cannot investigate ENSO forecast skill from its output. However, it is possible to investigate the influence of ENSO events on 2 m temperatures in this data set, which will be done later in this work.

3 Data

As already mentioned at the beginning, the aim of this work is to investigate the performance of the seasonal forecast ensemble ASF-20C. Hence, reference data sets are needed for verification. One of them is ERA-20C, a centennial reanalysis that was used to initialize ASF-20C forecasts and thus covers the same period. On the other hand, the state-of-the-art reanalysis product ERA5, provided through Copernicus Climate Change Service (C3S), is used. Since it covers only the period from 1950 onwards, it will be used to verify ASF-20C output in more recent decades. Details regarding the dynamical model as well as the forcing data of the three data sets are summarized in Tab. 3.1.

3.1 ERA-20C

Poli et al. (2016) presented the ECMWF twentieth century atmospheric reanalysis *ERA-20C*, a global climate reanalysis which provides atmospheric data for the period 1900–2010. It is based on version cy38r1 of ECMWF's Integrated Forecast System (IFS; ECMWF, 2013). The IFS itself contains an Atmospheric General Circulation Model (AGCM) and a variational analysis scheme. The ability of the AGCM to correctly simulate the observed changes during the twentieth century was tested by running a 10 member model-only version of the reanalysis called ERA-20CM (Hersbach et al., 2015). The results showed that the IFS AGCM is indeed able to represent the variability of a number of atmospheric quantities, e.g. 2 m temperature and net outward top-of-atmosphere energy flux, throughout the whole century, if a suitable set of boundary conditions and radiative forcing data is used. To create the ERA-20C data set the variation of a number of forcing parameters was prescribed. Specifically, sea ice concentration and sea surface temperature data was used from HadISST version 2.1.0.0 (Titchner and Rayner, 2014) and greenhouse gases, ozone, tropospheric and stratospheric aerosols, as well as solar radiation data was used from the same sources as for CMIP5 (Taylor et al., 2012)[1]. The physical model is integrated on 91 vertical levels from the surface up to 0.01 hPa, which corresponds to an altitude of approximately 80 km. For the horizontal grid spectral triangular truncation T159 is used, leading to a resolution of about 125 km and the model time step is 30 minutes.

Apart from a dynamical model, the second component needed for a reanalysis are observations. ERA20C only assimilated marine wind observations from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; Woodruff et al., 2011) version 2.5.1 together with surface pressure observations from ICOADS and the International Surface Pressure Databank (ISPD; Cram et al., 2015). A set of quality control mechanisms ensures that no duplicate or suspicious observations are included. In 1900 about 30 000 surface pressure observations were assimilated per month and until 2010 this number increased to 3.6 million observations per month. The two components, model and observations, are combined using four-dimensional variational (4D Var) analysis. Thus every day at 0900

[1] A brief overview of these sources is shown in `https://pcmdi.llnl.gov/mips/cmip5/forcing.html`

UTC the model state is adjusted in a way that the trajectory for the following 24 hours is a best fit to the available observations. In order to adjust the trajectory as good as possible, an estimate of the temporary and spatially varying background error is given by the output of a 10-member ensemble of previously produced reanalysis experiments (Poli et al., 2013). For production, the 110 year period was divided into overlapping 6-year segments, where the first year of each segment was used as spin-up for the model and therefore not used in the final product. ERA-20C output consists of 3-hourly fields and monthly averages of all variables for the already mentioned period 1900–2010. However, in this work only monthly averages of 2 m temperature are considered.

3.1.1 Performance of ERA-20C

Poli et al. (2016) compared nighttime air temperature observations over oceans to ERA-20C output and found a steady cold bias of about 1 K of the reanalysis values. However, around World War II this bias appeared to be stronger, though this was assumed to be due to an unresolved warm anomaly in ship measurements. Concerning the water cycle, model output rainfall shows rather good agreement with rain gauge data over Europe, North America, Japan and Australia, especially from the middle of the century onward. However, it has to be mentioned that rain gauge data over the course of the century is very sparse in most parts of the world. Also global averages of precipitation minus evaporation are more stable in ERA-20C output than in other reanalyses products, although ERA-20C total column water vapor shows a dry bias compared to different observational products.

It is very important for a centennial reanalysis like ERA-20C to reproduce low-frequency fluctuations of climate indices during the whole period. Thus Poli et al. (2016) also investigated the agreement between ERA-20C and other reanalyses like JRA-55 (from Japan Meteorological Agency (JMA); Kobayashi et al., 2015) and ERA-Interim (from ECMWF; Dee et al., 2011) among others. One major climate index which was investigated is El Niño-Southern Oscillation (ENSO). In fact, Poli et al. (2016) checked both ENSO components, ocean and atmosphere, separately and showed that ERA-20C output shows a rather good agreement with other reanalyses products, though the atmospheric component showed some deviations in the first half of the 20^{th} century. Behaviour is very similar for NAO and PNA indices, where again ERA-20C shows some deviations until 1950 but afterwards the performance is very good. On the other hand, the model-only simulation ERA-20CM does not perform as well as all the other reanalyses, which points out the importance of the assimilated observations.

3.2 ASF-20C

Weisheimer et al. (2017) presented the global Atmospheric Seasonal Forecast data set (ASF-20C) using ERA-20C data for initialisation and verification. ASF-20C consists of 51 ensemble members and covers the period 1901–2010. Seasonal re-forecasts with a lead time up to 4 months are started on 1 February, May and August of each year from 1901 until 2010 and on 1 November from 1901–2009. For each run only information, which is available at the starting date is used. The atmospheric model of ASF-20C is also the atmospheric component of the IFS but a more recent version than used in ERA-20C, namely cy41r1. The vertical resolution is the same as for ERA-20C with 91 vertical levels up to 0.01 hPa but the horizontal resolution is higher than in the reanalysis with a spectral truncation of

T255, corresponding to a grid size of about 80 km. Since ASF-20C is an atmosphere-only model, prescribed SSTs and SICs are used as a lower boundary and perfect forcing is assumed. Weisheimer et al. (2017) stated that dynamical atmospheric processes between extratropical and tropical regions do not necessarily need an atmosphere-ocean coupling and therefore phenomena like NAO can still be reproduced well enough. Further forcings are time-varying greenhouse gas concentrations, a time-varying solar cycle and volcanic aerosols as well as soil moisture and snow cover at the land surface. All of them are implemented in the same way as in ECMWF's seasonal forecast system 4 (System4; Molteni et al., 2011), which was ECMWF's operational system from 2011–2017 (Weisheimer et al., 2019).

The 51 re-forecast ensemble members are generated using stochastic parameterizations of subgrid-scale atmospheric processes. The archived output of this hindcast are global fields of monthly means of several quantities like 2 m temperature or precipitation on a $1^\circ \times 1^\circ$ grid for 1–4 months lead time. In this work, only 2 m temperature fields for winter and summer, covering the months November, December, January, and February as well as May, June, July, and August are going to be used. To clarify, the expression *DJF average 1996/97* denotes the average of December 1996, January 1997 and February 1997.

3.2.1 Performance of ASF-20C

In order to investigate the quality of ASF-20C output, Weisheimer et al. (2017) compared global average DJF-mean 2 m temperatures of ASF-20C and ERA-20C data sets. They found that, although the global cooling period between the 1950s and 1980s is underestimated in the hindcasts, both multi-decadal variability throughout the 110-year period and also the warming trend in the most recent period are covered well in ASF-20C. The underestimation of the cooling period is also present, but weaker, over sea, although forcing via prescribed SSTs should lead to more accurate results there. Weisheimer et al. (2017) also investigated NAO forecast skill of ASF-20C using the anomaly correlation coefficient (ACC) in moving 30-year windows, which is defined as in eq. 4.3, though climatological averages instead of plain mean values are taken. They found high skill for all positive and strong negative NAO events throughout the 20^{th} century but less skill for weak negative NAO events especially around the middle of the century. More generally, correlation skill of 500 hPa geopotential height anomalies in this period is reduced over the North Pacific and vanishes completely over North America and the North Atlantic (O'Reilly et al., 2020). O'Reilly et al. (2017) found a very similar result regarding forecast skill of the PNA index. For this index forecast skill is even higher than for the NAO index during the whole hindcast period but again it exhibits a pronounced drop around the 1950s, which seems to be a result of less predictable negative PNA events at this time. Reduced predictability comes from a weaker tropical SST forcing and thus the mid-century drop of predictability of NAO and PNA indices may be related. On the other hand, Parker et al. (2019) claim that forecast skill of the NAO index is dominated by skill in predicting interannual variations in jet stream latitude.

3.3 ERA5

An additional global reanalysis data set used in this work is ECMWF's ERA5 (Hersbach et al., 2020). Initially, it covered the period from 1979 onwards but it was recently extended

backwards until 1950 (Bell et al., 2021). The IFS model version is cy41r2, which is very similar to the one that was used for the integration of ASF-20C. What is outstanding about this reanalysis is its very high resolution compared to similar products. The horizontal spectral resolution is TL639 (approximately 31 km grid box size) and there are 137 vertical levels up to 0.01 hPa. As for ERA-20C, for data assimilation a 4D Var system is used but this time the length of the analysis window is only 12 h (0900 UTC–2100 UTC and 2100 UTC–0900 UTC) instead of 24 h. Observations, which are used within these 12 h windows, are a combination of both satellite and in situ data. In total, more than 200 satellite instruments and other conventional data sources are used[2]. The number of observations increased from an average of about 53,000 per day in January 1950 (only conventional sources) to approximately 750,000 per day in 1979 and to around 24 million per day in January 2019, which results in a total of 94.6 billion actively assimilated observations in 4D Var during the whole period.

Concerning radiation forcings like ozone, greenhouse gases, and aerosols until 2005 the same data as for CMIP5 and afterwards forcings from Representative Concentration Pathways (RCP) 2.6 are used (Loeb et al., 2022). The same approach was taken in ERA-20CM and ERA-20C (Hersbach et al., 2015; Taylor et al., 2012). SSTs and sea ice information again come from the HADISST2 product together with data from the Climate Change Initiative (ESA CCI) SST v1.1 (Merchant et al., 2014), the Met Office OSTIA product (Donlon et al., 2012) and the EUMETSAT OSI SAF reanalysis product (v409a; Eastwood et al., 2014). Details on SST and SIC forcing data are presented in Tab. 3.1. Background errors, which are needed to combine observation and model components of the data assimilation, are estimated using an ensemble of one control member with full resolution and nine perturbed members with halved resolution (EDA; Isaksen et al., 2010). Thus, background errors in ERA5 are flow-dependent. The perturbed ensemble members are generated by adding random perturbations to observations and physical tendencies, but also SSTs and SICs are perturbed, though in a different way (Hirahara and Hersbach, 2016). A large number of oceanic, atmospheric and land variables is produced as output every hour, most of them from the high resolution run on a $0.25° \times 0.25°$ grid. In this work, only monthly averaged 2 m temperature data on a $1° \times 1°$ grid spanning the whole globe for the period 1950–2010 will be used. Because of its high spatial resolution and the enormous amount of assimilated observations, ERA5 is going to be used for verification or as reference of the other data sets most of the time.

3.3.1 Performance of ERA5

Bell et al. (2021) stated that the linear trend in global mean 2 m temperature in ERA5 for 1979–2018 is very similar to several other data sets like JRA-55 and NASA's GISTEMP v4 (Hansen et al., 2010; Lenssen et al., 2019). However, for the period 1991–2020 the ERA5 trend of about 0.24 K per decade is up to 0.03 K per decade higher than in the other products. Simmons (2022) added that for the period 1979–2022 not only trend magnitude but also their confidence intervals are in good agreement with the same reference data sets. In the same work it was also stated that global land-only temperature trend in ERA5 data is 60% higher than if both land and sea grid points were considered. Moreover, Simmons (2022) state that trends (and also confidence intervals) in Europe are larger than in any

[2] see `https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Observations` for a full list

Table 3.1: Comparison of the data sets used in this work in terms of model setup and forcings.

	ASF-20C	ERA-20C	ERA5
Coupling	No	No	Yes
IFS cycle	cy41r1	cy38r1	cy41r2
Horizontal resolution	ca. 80 km	ca. 125 km	ca. 31 km
Radiative forcing	Time-varying GHGs and solar cycle; Volcanic aerosols	Time-varying GHGs, solar radiation, ozone, tropo- and stratospheric aerosols	Time-varying GHGs, solar radiation, ozone, tropo- and stratospheric aerosols
SST data source	HadISST 2.1 monthly	HadISST 2.1 monthly	HadISST 2.1 monthly, until Dec 1960 HadISST 2.1 daily, Dec 1960 – Aug 2007 OSTIA afterwards
SIC data source	HadISST 2.1 monthly	HadISST 2.1 monthly	HadISST 2.0 monthly, until Dec 1978 OSI SAF daily, afterwards

other continent, except for the Arctic, and that the reason is the small size of the continent (increasing variability of the time series) as well as its position at the end of the North Atlantic storm track.

Bell et al. (2021) also compared 12 month running mean 2 m temperature anomalies of ERA5, JRA55, HadCrut5 (Morice et al., 2021) and GISTEMP with observations over European and Australian land areas. Over Europe there was a good agreement between ERA5 and the reference data for the whole period. On the other hand, over Australia ERA5 and JRA-55 show quite large positive deviations from the other datasets until 1970 and small negative deviations in a few years afterwards. The latter is an effect of above average rainfall, often linked to La Niña events. In other regions ERA5 also has some problems with realistic representation of quantities like 2 m temperature or surface pressure in times when little or no observational data is available (e.g. in China prior to August 1956).

Simmons et al. (2021) showed that there are some deviations in twelve-month running mean temperature anomalies between ERA5 and several other data sets (e.g. ERA-Interim, JRA-55) from the 1950s until the 1970s over tropical and mid-latitude sea as well as over the Antarctic. While differences in the former location come from different SST analyses used in the products, over the Antarctic the reason is simply the lack of long-term observations. Over land areas, ERA5 is slightly colder over Europe and North America in the 1950s and 1960s but warmer over Australia from 1950 to the mid-1970s. While in the latter region the warm bias is more than 1 K at the beginning of this period, it gradually decreases until about 1975. Simmons et al. (2021) also discussed three local warm biases, which are located in and south of the Congo basin from 1950–1952, over the region from Nigeria to Ethiopia in 1965 and 1966, and over Brazil in 1961. In each of these cases ERA5 showed a warm bias of several K compared to GISTEMP. In addition, in the first case there seem to be strong cold conditions to the east and north-east of the warm pattern and a smaller cold signal over the other parts of the African continent. While the latter conditions are in good agreement with GISTEMP data, the former seems to be a compensation of the warm bias over the Congo basin. Other known local issues are too cold temperatures in the region of Iran and Iraq but also too high temperatures south of the Caspian Sea during the 1980s due to a lack of observational data. For the same reason, ERA5 output is too cold over southern and eastern China in 1950–1955, in 1965 and in 1966. Also, 2 m temperature over the Great Lakes, and strongest over Lake Superior,

exhibits a too large annual cycle from 1979 until 2013 (Hersbach et al., 2020).

4 Verification

In this section 2 m temperature output of the three data sets introduced above is going to be compared both qualitatively and quantitatively using different verification measures like bias, Pearson correlation and reliability. ERA5 is considered as reference data for the period 1950–2010 for ERA-20C and ASF-20C. ERA-20C is used for verification of re-forecasts during the whole period 1901–2010.

4.1 Bias

In this work, bias between 2 m temperatures of two data sets is calculated to show the average difference during a considered period (Stanski et al., 1989):

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) \tag{4.1}$$

where F_i and O_i are temperature values in year i and N is the total amount of years in the period. Initially, F and O represent forecast and observation data but in this work they stand for investigated and reference data set, respectively.

Fig. 4.1 shows differences of monthly mean 2 m temperatures in boreal summer months during the period 1973–2010 between ASF-20C and ERA-20C. The maps in the first and second row show May, Jun, Jul and Aug and the one in the lower left panel displays JJA average. The lower right panel shows differences of DJF averages during the same period. The small panel attached on the right side of each map contains the zonal average. The global mean is calculated using the cosine of latitudes as weights and is shown on the top right of each panel. It can be seen that in the global average ASF-20C output is slightly warmer than ERA-20C from lead month 1 in this period. However, locally there is considerable variability. For example, patterns indicating a cold bias evolve over India and south of the Sahara with increasing lead time, showing a maximum in June and July, respectively. A persistent cold bias also occurs along the Andes which may be due to a different representation of the topography (arising from the different spatial resolution used in ASF-20C and ERA-20C) there. On the other hand, the warm bias is largest in North America and central Asia at all lead times. Over oceans, ASF-20C bias is in general positive with the largest values in tropical oceans and around Australia. Fig. 4.2a shows that these are also the regions where ASF-20C bias with respect to ERA5 in the same period is smaller than in adjacent regions. The hindcast's bias against ERA-20C does not change much over the course of the century but there are differences between winter and summer seasons. For example, the lower right map in Fig. 4.1 indicates that warm bias patterns over northern hemispheric mid-latitudes shift northwards in boreal winter though the pattern in central Asia remains the same. Also over Africa slight cold bias dominates except for the southern tip of the continent where bias is positive. According to Fig. 4.2b,

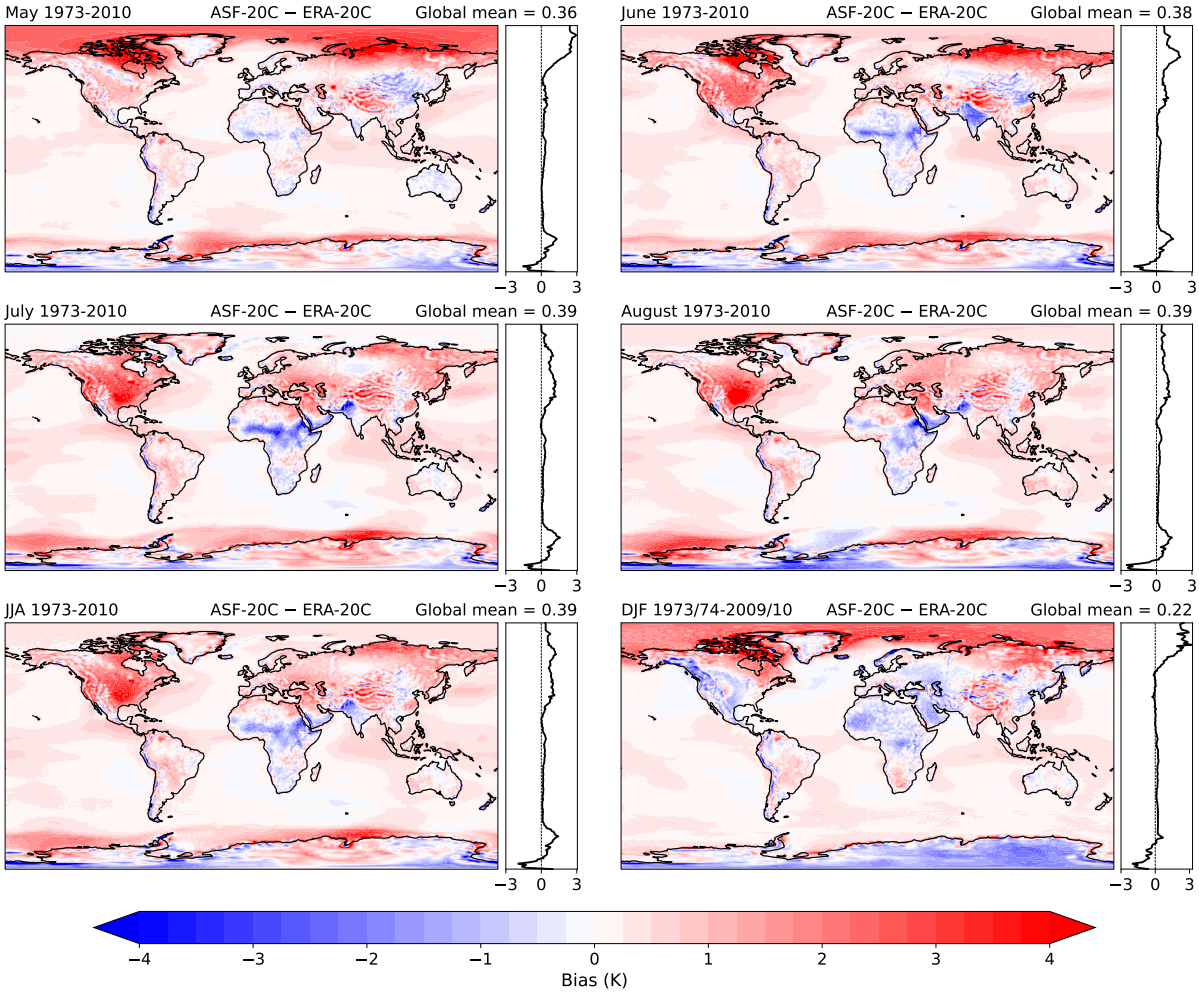


Figure 4.1: Monthly averaged 2 m temperature differences between ASF-20C and ERA-20C in the period 1973–2010 for May, June, July, August and JJA averages. The lower right plot shows the same for DJF averages during the period 1973/74–2009/10. Small panels attached to each map contain the zonal mean. Global mean values are calculated using the cosine of latitudes as weights.

ERA-20C shows a warm bias compared to ERA5 in mid-latitudes of Asia and south of the Sahara in boreal summer during the most recent decades.

ASF-20C is also compared to ERA5 in the periods 1950–1979 and 1980–2010 for boreal summer and 1950/51–1979/80 and 1980/81–2009/10 for boreal winter as is shown in Figs. 4.2c–4.2f. It can be seen that global averages of ASF-20C output are slightly colder than ERA5 in both seasons and both periods. Figs. 4.2c & 4.2d display a band of warm bias over the mid-latitudes of northern hemispheric continents during boreal summer months. On the other hand, there is a weak cold bias over the continents of the Southern Hemisphere in both seasons. For boreal winter months during both periods (Figs. 4.2e & 4.2f) a warm bias of more than 2 K evolves with increasing lead time over Asia and Canada, though it is stronger and more expanded in both areas during the earlier period. The largest cold

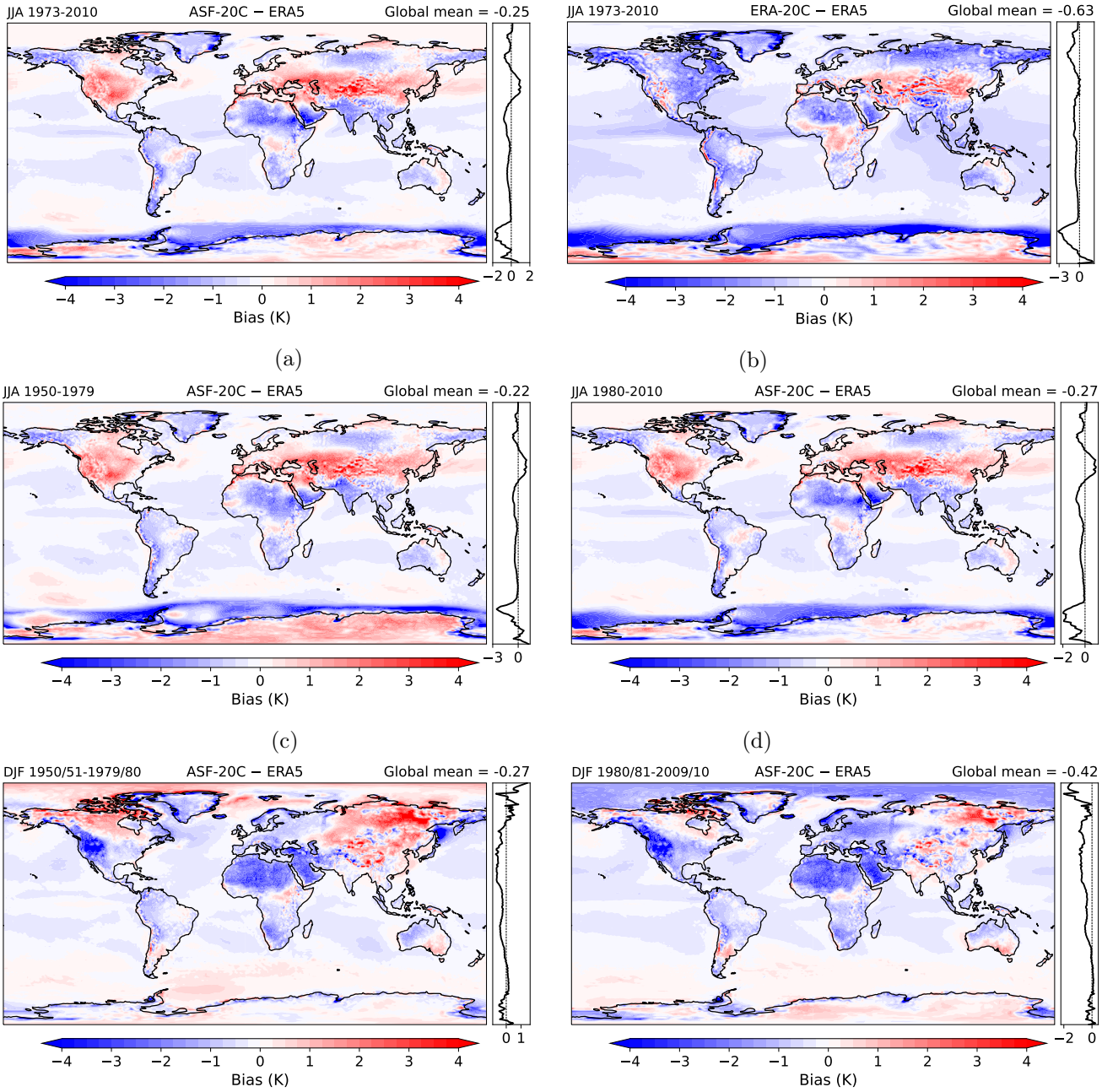


Figure 4.2: Bias of JJA average 2 m temperatures during the period 1973–2010 of (a) ASF-20C ensemble mean and (b) ERA-20C with respect to ERA5. (c) and (d) show JJA averages of ASF-20C ensemble mean bias with respect to ERA5 for the periods 1950–1979 and 1980–2010, respectively. (e) and (f) contain DJF average bias of ASF-20C ensemble mean with respect to ERA5 for the periods 1950/51–1979/80 and 1980/81–2009/10, respectively.

bias of ASF-20C compared to ERA5 during boreal winter emerges in both periods over the Sahara, the Arabian Peninsula and along the North American Pacific coast. Temperatures differ by up to 4 K in the first two regions mentioned and in the latter even more.

Over oceans of the tropics and mid-latitudes ASF-20C ensemble mean bias is negative and down to −0.5 K. Contrary to this, hindcast 2 m temperatures over the Southern Ocean and the northern North Pacific Ocean are slightly higher than ERA5 values. While there is a strong cold bias of several K around the Antarctic for JJA averages in both seasons and a smaller negative bias of 1–2 K in the Arctic during boreal winter of the more recent period, arctic temperature bias is slightly positive during the prior period. This suggests that ASF-20C has problems with properly describing boundary conditions in regions with sea ice. In fact, Titchner and Rayner (2014) mention that in the Arctic until 1953 and in the Antarctic even until 1973 sea ice concentration fields of the HadISST 2.1.0.0 data set base mostly on climatology and that in the Antarctic these climatologies rely on sparse ship observations. And even after 1978, when passive microwave satellite measurements are the primary sea ice concentration data source, errors occur due to several limitations in the measurements coming from different surface processes like melt ponds on the ice surface or roughening wind over open water. Thus, I consider the forcing data as the source of these problems in regions with sea ice rather than the re-forecast model itself.

Table 4.1: Bias between ASF-20C, ERA-20C and ERA5 for regional DJF and JJA averages in different periods and regions. Only land grid points are considered for the Antarctic and Europe.

		ASF-20C - ERA-20C	ASF-20C - ERA5	ERA-20C - ERA5
DJF 1950/51–1979/80	Global all	0.37	-0.27	-0.64
	Global land	0.22	-0.50	-0.72
	Global sea	0.43	-0.21	-0.64
	Arctic sea	3.15	0.43	-2.72
	Antarctic	-0.42	-0.26	0.16
	Europe	0.03	-1.27	-1.30
JJA 1950–1979	Global all	0.44	-0.22	-0.66
	Global land	0.21	-0.47	-0.68
	Global sea	0.41	-0.25	-0.65
	Arctic sea	0.35	0.05	-0.30
	Antarctic	1.12	0.11	-1.01
	Europe	1.12	0.77	-0.35
DJF 1980/81–2009/10	Global all	0.20	-0.42	-0.62
	Global land	-0.03	-0.77	-0.74
	Global sea	0.35	-0.27	-0.62
	Arctic sea	2.52	-0.74	-3.26
	Antarctic	-0.51	-0.22	0.29
	Europe	-0.67	-1.69	-1.02
JJA 1980–2010	Global all	0.36	-0.27	-0.63
	Global land	0.15	-0.58	-0.73
	Global sea	0.32	-0.27	-0.59
	Arctic sea	0.39	0.05	-0.34
	Antarctic	0.00	-0.75	-0.75
	Europe	0.82	0.63	-0.19

Tab. 4.1 summarizes mean biases of global (total, land-only and sea-only), Arctic, Antarctic and European grid points for seasonal averages in the periods 1950–1979 and 1980–2010 for boreal summer as well as 1950/51–1979/80 and 1980/81–2009/10 for boreal winter between the three data sets. The third column, *ASF-20C - ERA-20C* shows that in almost all considered cases hindcast temperatures are above those of ERA-20C reanalysis. The largest positive differences occur in the Arctic during winter. At the same time, bias is around −0.5 K in the Antarctic. The fourth and fifth column however indicate that both ASF-20C and ERA-20C temperatures are mostly below that of ERA5 in the mentioned regions. In Europe during winter negative biases exceed −1 K, while they are less pronounced but still negative in summer for ERA-20C and, in contrast, above +0.6 K for ASF-20C (both with respect to ERA5). Though SIC forcing data is the same in ERA-20C and ASF-20C, spatially averaged bias of polar regions with respect to ERA5 tends to be smaller in hindcast data, except for Antarctic JJA averages of the more recent period where output of both data sets is almost identical. The reason for the differences may be the newer IFS cycle used in ASF-20C. Global average bias of both land and ocean grid points is negative in all seasons and periods and more extreme in ERA-20C than in re-forecast data when compared to ERA5. However, Fig. 4.2b shows that there are also continental regions like central Asia, equatorial Africa and along the Andes where ERA-20C bias against ERA5 is positive for JJA averages of the more recent period.

4.2 Time series

Fig. 4.3 shows land-only DJF averages of 2 m temperature anomalies over South America and land-only JJA averages over GAR and globally. Anomalies of each data set are calculated with respect to their own climatological mean of 1950–1980, which can be seen as a linear bias correction (Weisheimer et al., 2019). Each member of the ASF-20C ensemble is shown as a grey line and the ensemble mean is represented by the black line. ERA-20C and ERA5 are displayed by the blue and orange lines, respectively. To represent possible external forcings, red and blue boxes mark El Niño and La Niña events and vertical grey lines indicate major volcanic eruptions (for details about both ENSO and volcanic forcing see Sect. 2). Root Mean Squared (RMS) error and Pearson correlation coefficient r of ERA-20C and ASF-20C ensemble mean are calculated with respect to ERA5 from 1950 onwards using

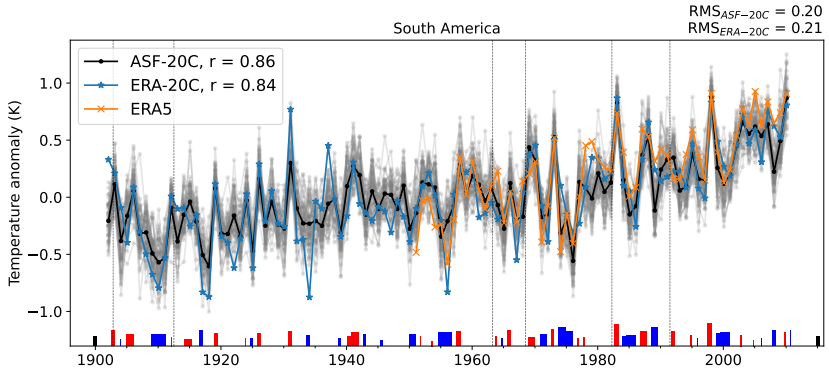
$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \tag{4.2}$$

and

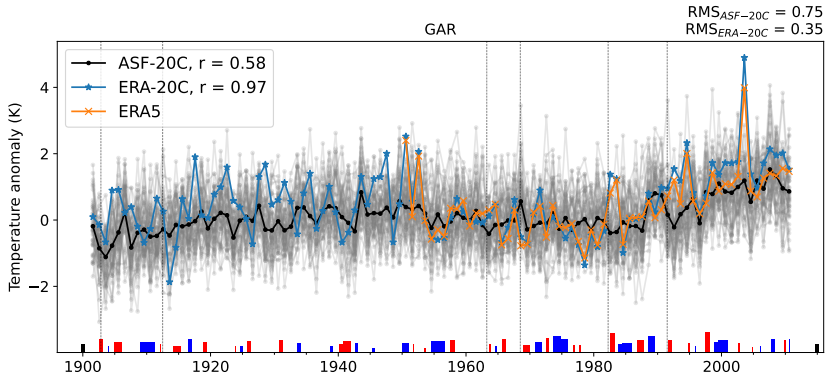
$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{4.3}$$

where x_i and y_i are values of the compared data sets in the i-th year with respective mean values $\overline{x}$ and $\overline{y}$. N is the number of considered years.

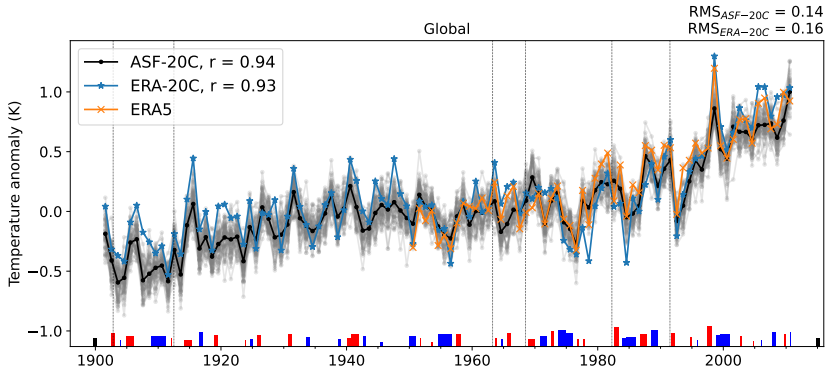
In Fig. 4.3a both measures indicate a very good agreement between ASF-20C ensemble mean/ERA-20C and ERA5 over South America. The ASF-20C ensemble covers the values of the two other datasets most of the time except for extreme anomalies in a few



(a)



(b)



(c)

Figure 4.3: (a) Land-only DJF average of 2 m temperature anomalies over South America. Vertical grey lines are major volcanic eruptions, red and blue bars show years in which El Niño and La Niña events occurred, respectively. RMS and r denote root mean squared error and Pearson correlation coefficient of the respective data set with respect to ERA5 from 1950 onwards. (b) and (c) Same as (a) but for JJA in GAR and JJA global, respectively. In each region, anomalies are calculated with respect to averages of 1950–1980 of the respective data set.

years. For example in 1933, the smallest temperature anomaly in the ensemble is about -0.60 K, whereas ERA-20C shows a value of ca. -0.88 K. A difference that may indicate

an underestimation of the response to the moderate La Niña event at the end of 1933 by the hindcasts. The behaviour is very similar in 1955, where ERA5 and ASF-20C indicate anomalies of about -0.55 K but ERA-20C shows a value of -0.83 K. The La Niña event in this year was even more pronounced both in terms of duration and strength.

In Fig. 4.3b agreement between ERA-20C and ERA5 is again very good resulting in a correlation coefficient close to one, which also raises confidence in ERA-20C data prior to 1950. Moreover, although ERA-20C anomalies tend to be slightly more extreme than ERA5 anomalies, especially after the year 2000, the extremely cold anomaly in summer 1913 of approximately -1.87 K is deemed trustworthy since it is also confirmed by the study of Luterbacher et al. (2016). I want to note that this cold anomaly does not seem to be a consequence of the Novarupta eruption in 1912 since no such temperature anomaly can be found in the data of the whole European continent in any of the two seasons. Correlation between ASF-20C ensemble mean and ERA5 is much lower, also compared to its value over South America. On the one hand, forecasts do not benefit that much from prescribed SSTs and high forecast skill due to ENSO over GAR as they do over South America and on the other hand synoptic activity can be high over central Europe during boreal summer. Both factors lead to an ensemble spread over GAR that is up to 7 times higher than over South America, which suggests increased forecast uncertainty in GAR. This large spread prevents the ensemble mean from reaching high anomalies and therefore almost no peaks of the ERA5 and ERA-20C time series are covered by the hindcast ensemble mean. For the same reason, correlation coefficients between ASF-20C and ERA5 are similar to the JJA GAR value over all northern hemispheric continents in boreal winter (correlation in GAR is even below 0.4 for DJF averages), while ERA-20C shows correlations of above 0.9. Though the ensemble mean may show large differences to reanalyses in Fig. 4.3b, still at least one ensemble member usually covers extreme anomalies of ERA-20C and ERA5 in almost all years, such as, e.g., in the coldest year 1913. However, in the years with extreme warm anomalies 1917, 1950 and 2003, not even a single ensemble member reaches reanalyses values. Concerning the extremely warm European summer in 2003, ASF-20C ensemble comes close to reanalyses values but still underestimates the observed anomaly. Explicitly, ERA-20C and ERA5 anomalies are +4.90 K and +3.99 K, respectively, while the warmest ASF-20C ensemble member output is +3.80 K.

Fig. 4.3c shows time series of JJA averages of global land-only grid points. Correlation and RMS values indicate even better agreement between data sets as in South America and GAR. Again the hindcast ensemble as well as reanalyses seem to properly respond to forcings. In the year after each volcanic eruption that is displayed, a clear decrease in temperature anomalies can be seen in each data set due to dimming effects of material that is emitted into the atmosphere. Though in 1969, the year after the eruption of La Cumbre, the decrease may be masked by the El Niño event in this year. The overlap of both forcing mechanisms may also be the reason why the large majority of ensemble members shows too high values in this year. The data sets also react properly to ENSO events with negative anomalies following a La Niña and positive anomalies following an El Niño event. The most obvious example is the most extreme El Niño event in 1997/98, which is followed by global 2 m temperature anomalies of +1.30 K and +1.20 K in ERA5 and ERA-20C, respectively. The highest ensemble value in this year is just above +1.10 K. So again, hindcasts come very close to reanalyses extreme values but do not reach them.

It should be mentioned that in some regions correlation of ASF-20C ensemble mean and ERA5 is even better than that of ERA-20C and ERA5, for example over Africa and North

America if JJA averages are considered. In the prior region $RMS = 0.23$ and $r = 0.86$ for the hindcasts and $RMS = 0.36$ and $r = 0.76$ for ERA-20C and the respective values in North America are $RMS = 0.29$ and $r = 0.79$ as well as $RMS = 0.39$ and $r = 0.66$. These results indicate uncertainties in ERA-20C in these regions. Especially for the latter region rather large deviations occur between hindcast and ERA-20C data until the middle of the 20^{th} century. Over all oceans, correlation coefficients result close to one and RMS values are almost zero for both ASF-20C and ERA-20C, which again is a result of prescribed SSTs and SICs in both data sets.

4.3 Temporal correlation on grid-point scale

Pearson correlation coefficients are not only calculated for continental averages but also on a global 1° x 1° grid according to eq. 4.3 for ASF-20C ensemble mean and ERA5 data in two different 30-year periods. Figs. 4.4 and 4.5 show the results of boreal winter and summer, respectively, for all lead months of the period from 1980 onwards. The lower right panel in both Figs. displays respective seasonal averages of the prior period until 1979. Tab. 4.2 summarizes correlations of spatial averages of these two data sets in different regions in both seasons and both periods. Asterisks mark values that are not significant on the 95% level. The results show that globally correlation over oceans is around 0.9 at almost all times but over continents values are dependent on both period and season. In general, performance over continents tends to be better during the respective summer months, especially from lead month 2 onwards. And as expected, correlation is higher and regions with statistical significance are more expanded in the more recent period, as suggested by an increase of the global mean correlation by around 0.1 in both seasons. Higher correlations from 1980 onwards are most probably because of the improved observational coverage. This can be seen from the fact that in the most recent period at lead month 1 in both seasons correlations are around 0.7 and statistically significant in many parts of the globe, while during 1950–1979 in regions like Africa, Europe and the Antarctic patterns with correlations around 0 occur already in November and May.

The only continental regions during boreal winter of the prior period that show significant correlations with values around 0.6 in all lead months are regions that are strongly influenced by ENSO teleconnections, i.e. northern South America, equatorial Africa and Australia. But as Fig. 4.4 shows, improvements are made over Europe, northern Africa, eastern and southern Asia as well as over northern and southern North America towards the most recent period. But still, correlation is decreasing with lead time over all continents, especially in the northern hemisphere. Between 1950 and 1979, correlation during boreal summer is very low from lead month 2 onwards except for northern South America, parts of central Africa and northern Europe. On the other hand, boreal summer correlations during 1980–2010, which are displayed in Fig. 4.5, are higher and more persistent with lead time over northern hemispheric continents, except for central and northern Asia and central North America. In the southern hemisphere, again ENSO-affected regions perform best.

Concerning polar regions, differences between seasons in the Antarctic are not as pronounced as over extrapolar continents. Correlations over the Arctic are much larger in boreal summer during the prior period. From 1980 onwards differences are less pronounced. Apart from the Arctic during boreal winter, major improvements from the early towards

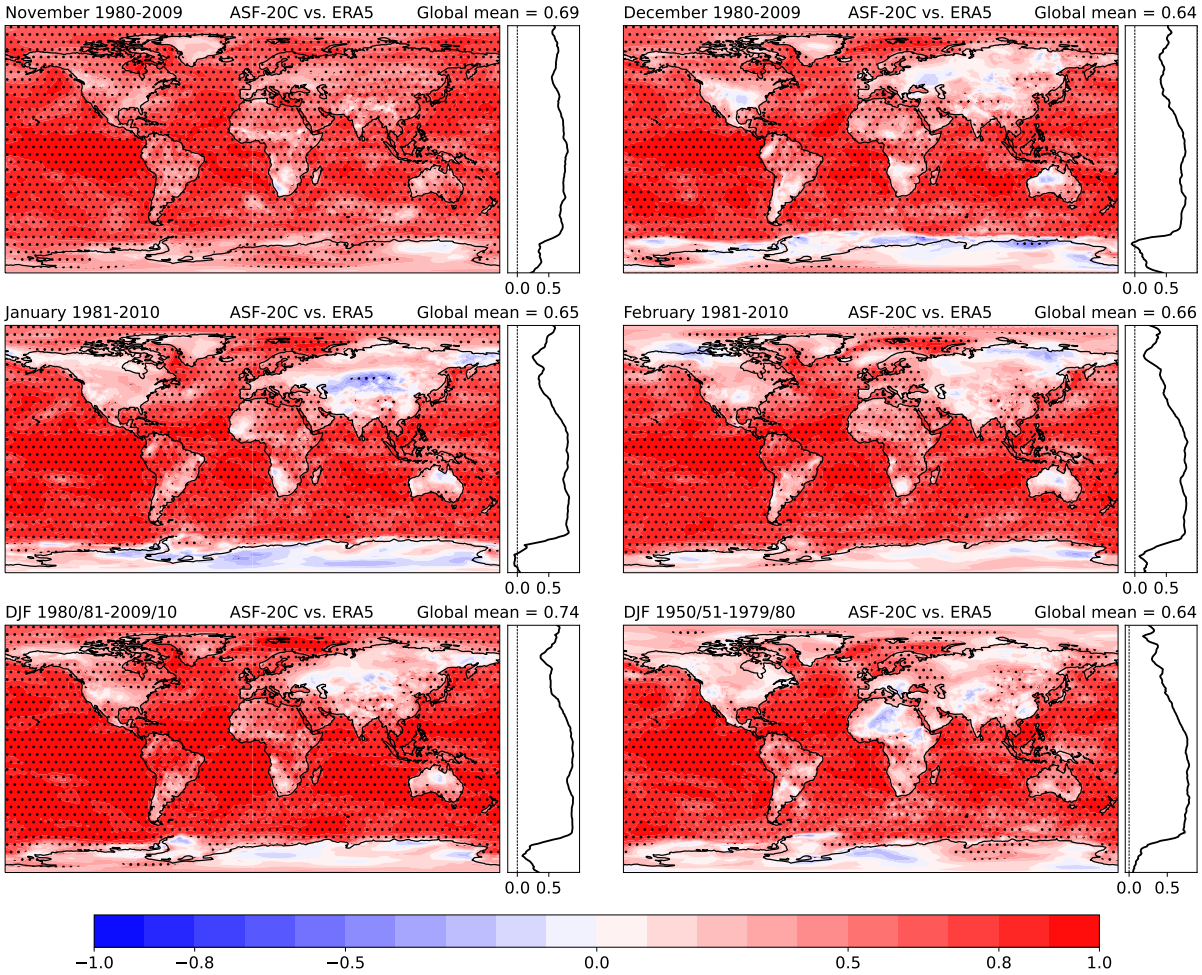


Figure 4.4: Correlation between ASF-20C ensemble mean and ERA5 on a 1° x 1° grid for boreal winter months in the period 1980/81–2009/10. Correlations of DJF averages of the period 1950/51–1979/80 are displayed in the lower right panel. Dotted areas show regions with 95% significance.

the most recent period in polar regions are visible in the Southern Ocean in both seasons and in the Antarctic during boreal summer, especially in May and for seasonal averages.

Correlation between ASF-20C and ERA-20C for DJF averages of three different periods, namely 1901/02–1936/37, 1937/38–19772/73 and 1973/74–2009/10, is shown in Figs. 4.6a–4.6c. Not many changes are visible in the first two periods. From November to February correlation decreases with lead time over most of northern hemispheric land areas, except for India and eastern Asia and is very high in ENSO affected regions of the southern hemisphere. During boreal summer, performance is slightly enhanced in the northern hemisphere. Towards the most recent period from 1973 onwards there are large improvements globally in both seasons but still regions with lower and not significant values can be found in Asia in both seasons as well as in Europe and North America during boreal winter.

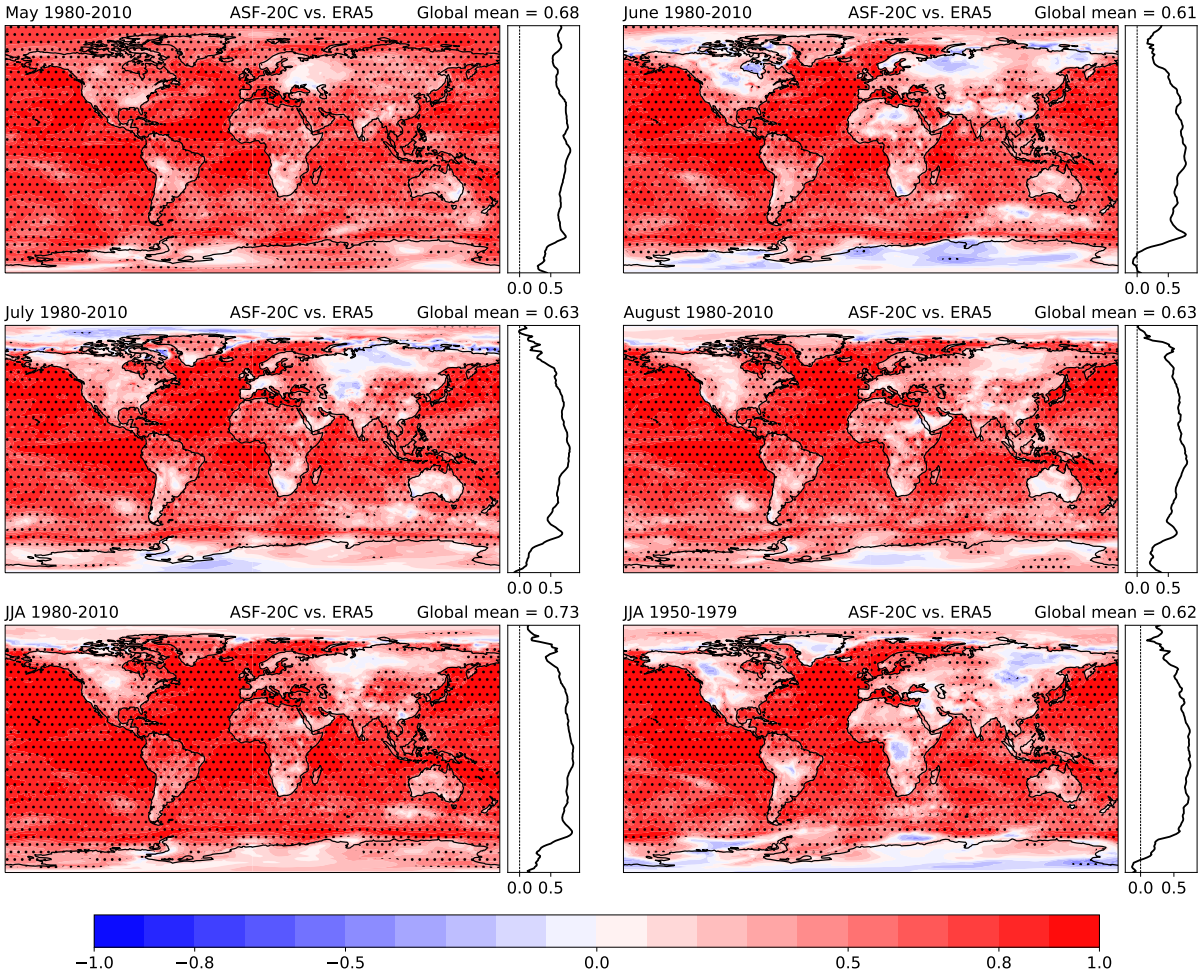


Figure 4.5: Same as Fig. 4.4 but for boreal summer months.

Considering ERA-20C reanalysis data, correlation with ERA5 of DJF averages of the period from 1973 onwards is displayed in Fig. 4.6d. Correlation coefficients are much higher and reach global mean values of almost 0.9 in both seasons in this period. In fact, correlation is everywhere close to one except for equatorial Africa and polar regions, but even there zonal mean values exceed 0.7. As for ASF-20C, values are decreased between 1950 and 1979 but still above 0.7. Very likely, this is again a result of more assimilated observations in ERA-20C towards the end of the covered period since regions with reduced correlation are those that are known to have a lack in observation data during the course of the 20$^{\text{th}}$ century, e.g. central Africa, the Southern Ocean and Antarctica. In such regions, additional observations do not only provide information at the time they are assimilated but also increase consistency of reanalysis output on scales of several months. Since re-forecasts are initialised using ERA-20C output, they are not affected as much and thus correlation may increase in these regions. Laloyaux et al. (2018) investigated confidence in 2 m temperature output in JJA 1959 of their coupled reanalysis CERA-20C and found decreased confidence in the same regions because too little data was available

Table 4.2: Correlation coefficients of spatially averaged 2 m temperatures of ASF-20C ensemble mean and ERA5 for all lead months as well as seasonal averages in different periods and regions. Only grid points over land are considered unless stated otherwise. Asterisks denote trends that are not statistically significant on the 95% level.

		Lead month 1	Lead month 2	Lead month 3	Lead month 4	Seasonal average
DJF 1950/51–1979/80	Global all	0.77	0.70	0.57	0.69	0.74
	Global land	0.52	0.34*	0.38	0.48	0.54
	Global sea	0.90	0.90	0.78	0.91	0.91
	Africa	0.54	0.32*	0.55	0.49	0.60
	Antarctic	-0.05*	0.26*	0.16*	0.00*	0.36
	Arctic sea	0.52	0.19*	0.13*	0.28*	0.25*
	Asia	0.66	0.14*	-0.10*	0.20*	0.16*
	Australia	0.46	0.43	0.56	0.47	0.61
	Europe	0.40	0.26*	0.29*	0.08*	0.35*
	GAR	0.07*	0.12*	-0.23*	0.18*	0.04*
	North America	0.52	0.22*	-0.20*	0.25*	0.21*
	South America	0.72	0.56	0.59	0.71	0.73
JJA 1950–1979	Global all	0.83	0.81	0.70	0.71	0.80
	Global land	0.43	0.49	0.56	0.53	0.71
	Global sea	0.92	0.87	0.85	0.85	0.90
	Africa	0.62	0.46	0.52	0.60	0.61
	Antarctic	0.23*	0.56	-0.07*	0.05*	0.18*
	Arctic sea	0.72	0.51	0.88	0.80	0.87
	Asia	0.44	0.31*	0.44	0.34*	0.41
	Australia	0.68	0.54	0.23*	0.24*	0.50
	Europe	0.22*	0.51	0.65	0.44	0.57
	GAR	0.38	0.26*	0.46	0.31*	0.39
	North America	0.66	0.17*	0.34*	0.60	0.49
	South America	0.34*	0.44	0.47	0.30*	0.70
DJF 1980/81–2009/10	Global all	0.95	0.90	0.85	0.82	0.92
	Global land	0.82	0.75	0.77	0.77	0.88
	Global sea	0.97	0.98	0.95	0.91	0.98
	Africa	0.68	0.66	0.64	0.83	0.83
	Antarctic	0.28*	-0.10*	0.01*	0.07*	0.03*
	Arctic sea	0.80	0.79	0.72	0.43	0.81
	Asia	0.73	0.39	0.22*	0.32*	0.45
	Australia	0.67	0.28*	0.25*	0.23*	0.41
	Europe	0.62	0.32*	0.39	0.52	0.53
	GAR	0.45	0.50	0.39	0.44	0.63
	North America	0.75	0.45	0.38	0.30*	0.58
	South America	0.57	0.50	0.87	0.65	0.88
JJA 1980–2010	Global all	0.90	0.91	0.92	0.92	0.95
	Global land	0.85	0.85	0.84	0.82	0.91
	Global sea	0.88	0.93	0.92	0.94	0.96
	Africa	0.77	0.77	0.82	0.85	0.88
	Antarctic	0.58	0.02*	0.21*	0.26*	0.41
	Arctic sea	0.80	0.72	0.86	0.94	0.92
	Asia	0.89	0.80	0.82	0.86	0.89
	Australia	0.51	0.38	0.36*	0.09*	0.48
	Europe	0.47	0.68	0.71	0.70	0.85
	GAR	0.77	0.56	0.17*	0.16*	0.50
	North America	0.69	0.59	0.73	0.74	0.77
	South America	0.46	0.53	0.60	0.56	0.76

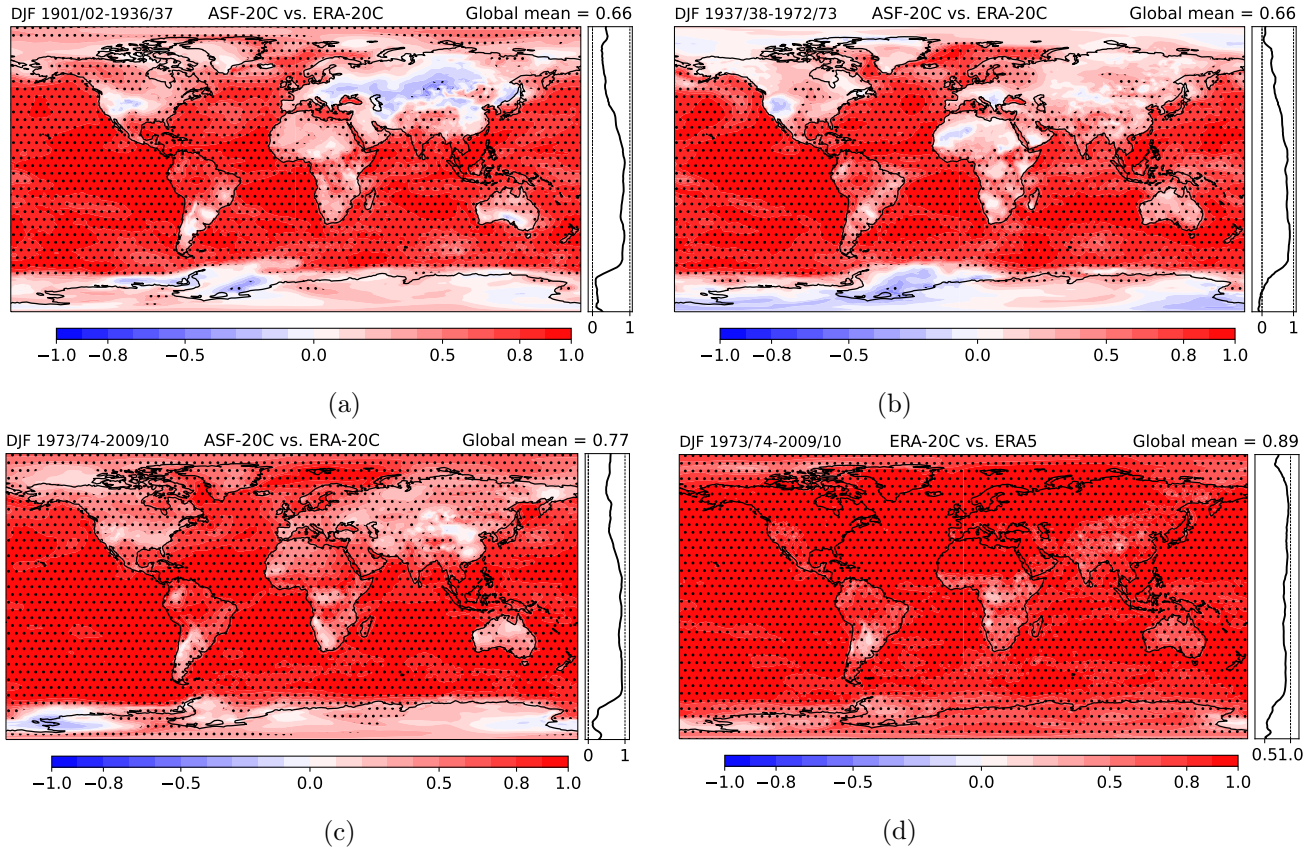


Figure 4.6: Correlation of ASF-20C ensemble mean and ERA-20C for boreal winter averages of periods (a) 1901/02–1936/37, (b) 1937/38–1972/73 and (c) 1973/74–2009/10. (d) Correlation of ERA-20C and ERA5 during the latter period. Dotted areas indicate regions where correlation coefficients are significant on the 95% level.

there, although even more observations were assimilated for CERA-20C than Poli et al. (2016) used for ERA-20C.

However, changes of correlation over the course of the 20$^{\text{th}}$ century occur not only because of improvements of the observational system. Since mostly ENSO, but in the northern hemisphere also NAO, are very important sources of forecast skill on seasonal time scales, the consequences of multi-decadal ENSO and NAO predictability variations on 2 m temperature forecasts around the world shall also be investigated. O'Reilly et al. (2017) and Weisheimer et al. (2017) stated that NAO forecast skill undergoes distinct changes during the 20$^{\text{th}}$ century with a minimum of forecast skill between the 1950s and 1970s. Weisheimer et al. (2017) conclude that the forecast model's performance for weak negative NAO events, which occurred more frequently at that time, is not as good as for strong negative and all positive NAO events. To get the NAO index, ERA-20C reanalysis mean sea level pressure values in Ponte Delgada (Portugal) and Reykjavik (Iceland) are used. Monthly anomalies are calculated in both locations using the full period 1901–2010 as climatological reference. These anomalies are normalized by dividing them by their own standard deviation over the same period. Furthermore, the linear trend of DJF average values over the period 1901–2010 is subtracted from the normalized anomalies.

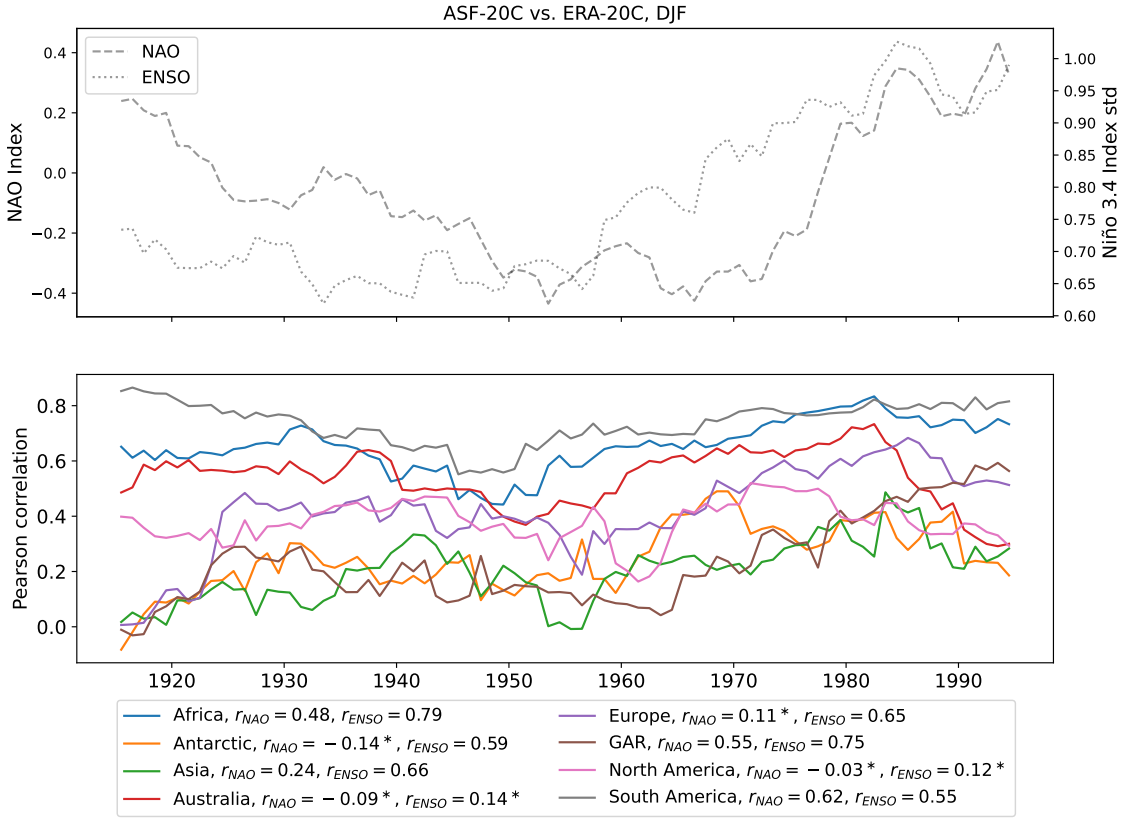
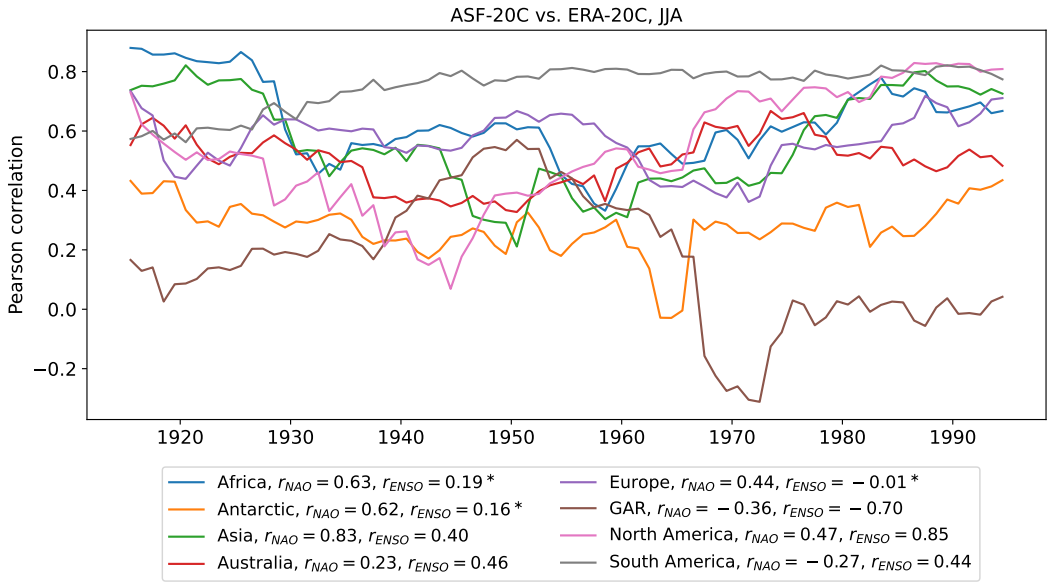


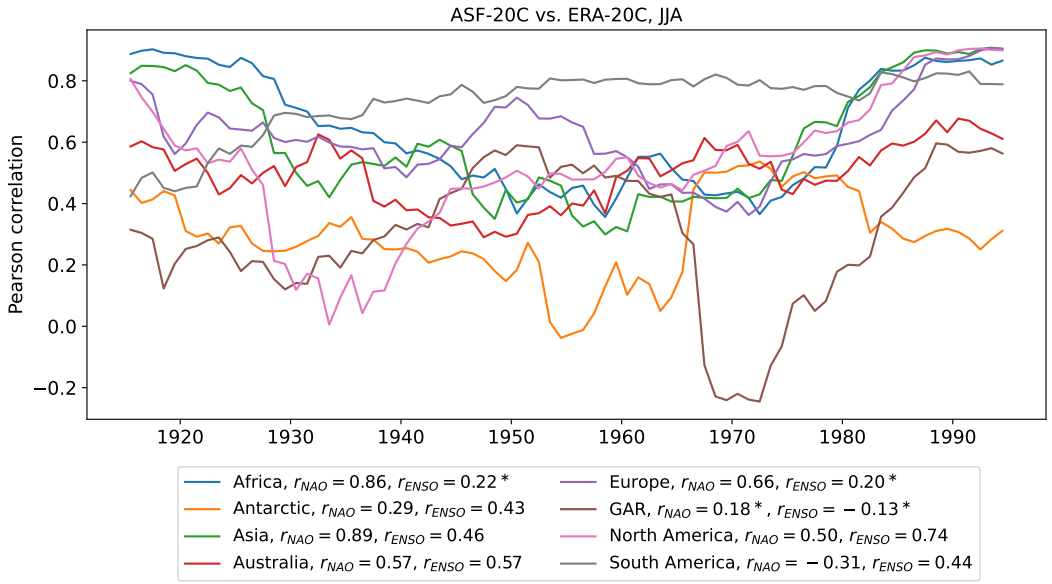
Figure 4.7: *Upper panel:* Averages of NAO index (dashed) and standard deviations of Niño3.4 index (dotted) in a moving 30-year window. *Lower panel:* Correlation of DJF averages between ASF-20C ensemble mean and ERA-20C for regional averages of land-only grid points in a moving 30-year window. Values in the legend give correlation coefficients with NAO index average (r_{NAO}) and Niño3.4 index standard deviation (r_{ENSO}) time series in the upper panel. Asterisks indicate values that are not significant on the 95% level.

The difference of the de-trended DJF averages of the normalized anomalies between Ponte Delgada and Reykjavik gives the final NAO index for every year (Hurrell, 1995; Türkeş and Erlat, 2003). Finally, averages of these values in a moving 30-year window are calculated. The result is visible as dashed line in the upper panel of Fig. 4.7. As already mentioned, Weisheimer et al. (2017) state that forecast skill is lowest for weak negative NAO events between the 1950s and 1970s. This time span corresponds very well to a period of slightly negative 30-year running averages of the NAO index calculated in this work.

Similarly to NAO forecast skill, Weisheimer et al. (2022) found that ENSO forecasts suffer from lower skill between the 1930s and 1950s and claimed that weak ENSO amplitudes during this period lead to this reduced predictability. Barnston et al. (2012) also found varying ENSO prediction skill during the period 1981–2010 due to lower variability. To quantify ENSO activity, DJF averages of the Niño3.4 index calculated as described in



(a)



(b)

Figure 4.8: (a) Same as the lower panel of Fig. 4.7 but for JJA averages. (b) Same as (a) but without detrending the ASF-20C ensemble mean and ERA-20C within each considered 30-year period.

Sect. 2 are used. In a moving 30-year window again the linear trend during each period is subtracted from the DJF time series. Of these residuals standard deviation within each window is used, since both Barnston et al. (2012) and Weisheimer et al. (2022) state that the amplitude of an ENSO event and its variability are crucial for ENSO predictability. The dotted line in the upper panel of Fig. 4.7 clearly shows that variability of the Niño3.4 index was at a minimum between the 1930s and 1950s, reproducing the results of Weisheimer

et al. (2022) very well. For both indices, ENSO and NAO, forecast skill is only reduced during the mentioned periods but higher in earlier decades, which rules out that poorer observational coverage at earlier times of the century is responsible for these skill variations.

Time series in the lower panel of Fig. 4.7 show correlations in a moving 30-year window of regional averages of ASF-20C ensemble mean and ERA-20C for DJF averages. In each 30-year window, both data sets were detrended before calculating correlation. Correlation coefficients with NAO index and Niño3.4 index standard deviation time series are given in the legend as r_{NAO} and r_{ENSO}, respectively. Asterisks denote values that are not significant on the 95% level. It can be seen that in continents like South America and Africa correlation between hindcast ensemble mean and ERA-20C reanalysis is at a minimum around the middle of the century. Respective r_{ENSO} values of +0.55 and +0.79 indicate that reduced ENSO activity at this time may be responsible for reduced correlations. r_{ENSO} is also around +0.6 in other continents like Asia, Europe and the Antarctic and even +0.75 in GAR. However, in the Antarctic changes in the SST and SIC forcing data around the 1950s may lead to increased correlations and mimic an effect of ENSO forecast skill. Improved observational coverage may also be the reason for unusually high correlation with the NAO index time series of +0.62 in South America.

Fig. 4.8a shows the same as the lower panel of Fig. 4.7 but for JJA averages. In this season, r_{ENSO} is largest in North America and around +0.4 but still statistically significant in Asia, Australia and South America. Interestingly, in GAR r_{ENSO} is also very large but negative. As supposed, r_{NAO} in JJA is large in continents of the Northern Hemisphere, though it is negative in GAR. In this season r_{NAO} is suspiciously large in the Antarctic. This time, changes in forcing data induced by satellite observations starting in the 1970s may be the reason.

As already mentioned previously, ASF-20C and ERA-20C data were detrended in each 30-year window before correlations were calculated for the time series in Figs. 4.7 and 4.8a. However, for the time series in Fig. 4.8b detrending was skipped. Differences are rather small until about the middle of the 1970s. From this time on a steep increase in correlation is visible in most continents in Fig. 4.8b but not at all in Fig. 4.8a. Thus, one reason for higher correlation in the most recent decades is most probably a consistent representation of temperature trends at that time an both data sets. More information on this issue will be given in Sect. 5.

4.4 Reliability and resolution

How reliable a probability forecast is, can be assessed qualitatively using a reliability diagram (Stanski et al., 1989). In such a diagram the observed frequency of a specific event is plotted against the forecast probability. More explicitly, for a specific event, e.g. monthly mean 2 m temperature being above a given threshold, forecast probability given by the ensemble is calculated and an observation/reanalysis product is used to verify if the event actually occurred. After doing this for a large number of events, forecast probabilities are binned and for each bin the ratio of occurrence of the corresponding events is calculated. Mason and Stephenson (2008) state that it is standard for seasonal forecasts to center probability bins around each 10%, i.e. 0%–5%, 5%–15%, etc., leading to 11 data points. Hsu and Murphy (1986) presented the attributes diagram, an extension of the reliability diagram. It includes additional reference lines from which attributes like resolution and

skill can be assessed. In the following, a way to quantify these measures will be shown.

In general, accuracy of forecasts can be calculated using Brier Score (Brier, 1950):

$$BS = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2 \tag{4.4}$$

where N is the total number of considered events and F_i is the forecast probability. For the case of a perfect forecast system, all members agree on their output, leading to either $F_i = 0$ if the event is not expected to happen or $F_i = 1$ if it is expected. O_i represents the observation and is equal to 0 if the event did not occur and 1 if it did. Thus a Brier Score of zero denotes the perfect forecast and 1 is the worst possible score.

For the case that T distinctive and mutually exclusive probability bins are defined, Murphy (1973) showed a way to split up Brier Score into three terms:

$$BS = \underbrace{\overline{O}(1-\overline{O})}_{\text{UNC}} + \underbrace{\frac{1}{N}\sum_{k=1}^{T} N_k(F_k - \overline{O}_k)^2}_{\text{REL}} - \underbrace{\frac{1}{N}\sum_{k=1}^{T} N_k(\overline{O}_k - \overline{O})^2}_{\text{RES}} \tag{4.5}$$

where $\overline{O}$ is the climatological probability of occurrence, F_k is the predicted probability, $\overline{O}_k$ the observed frequency of occurrence and N_k the number of cases within the k-th probability bin.

The three terms, abbreviated with *UNC*, *REL* and *RES* denote uncertainty, reliability and resolution, respectively. But only the latter two actually depend on the forecast system, whereas uncertainty is defined only by the climatological frequency of the predicted event.

Reliability measures the agreement between forecast probability and frequency of occurrence. Perfect reliability is achieved, if all data points lie on the 45° line in an attributes diagram, which means that of all events with a forecast probability of p% this specific event is actually observed in p% of the cases. The term *REL* as it is defined in eq. 4.5 is a weighted squared difference between the data points and the 45° line in an attributes diagram. Therefore $REL = 0$ corresponds to perfect reliability.

Resolution measures the degree to which the frequency of occurrence within each probability bin differs from the total frequency of occurrence. Or in other words if the forecast system can distinguish cases with high probability of the event to occur from those with low probability, and therefore high resolution also implicates high sharpness (Mason and Stephenson, 2008). Recalling the *RES* term in eq. 4.5, one can see that if $\overline{O} = \overline{O}_k$ in every bin, $RES = 0$ and therefore small values for resolution indicate a bad score. Hence, the horizontal line in the attributes diagram, which illustrates the long time average of occurrences of the event, i.e. the climatological probability of occurrence, is also called *no resolution line*. Hsu and Murphy (1986) showed that maximum or optimal resolution is represented by a step function that has the value 0, where $F_k < \overline{O}$ and 1 elsewhere. Therefore, optimal resolution is marked in an attributes diagram as a vertical line at $F_k = \overline{O}$.

Another important quantity of a forecast system is skill, i.e. if the forecast provides more information than some reference forecast. To quantify skill, one can use Brier Skill Score (Mason, 2004):

$$BSS = 1 - \frac{BS}{BS_{\text{ref}}} \tag{4.6}$$

where BS and BS_{ref} are the Brier Scores of the investigated and the reference forecast system, respectively. The aim of BSS is to quantify the improvement of the investigated forecast system compared to a reference forecast. The maximum value is 1, which denotes a perfect forecast. $BSS = 0$ means that the forecast does not provide any improvement over and negative values occur if it is even worse than the reference strategy. However, Mason (2004) stated that even if skill score is negative, the forecast system can provide useful information.

Very often, climatology is used as reference under the assumption that the climatological frequency of an event is stationary. Recalling eq. 4.5 this leads to $F_k = \overline{O}_k$ and $\overline{O}_k = \overline{O}$ and thus $REL = 0$ and $RES = 0$, leading ultimately to $BS_{\text{clim}} = UNC$. Using eq. 4.6, this leads to a formula for the Brier Skill Score when climatology is used as reference forecast (Mason, 2004):

$$BSS_{\text{clim}} = \frac{RES - REL}{UNC} \tag{4.7}$$

From this equation it can be derived easily that skill is zero if $RES = REL$ and also that a forecast system provides an improvement over climatology whenever $RES > REL$. Hsu and Murphy (1986) derived the existence of a *no skill line* which is located equidistant between the *no resolution line* and the 45° line in an attributes diagram. The area between the no skill and the optimal resolution line indicates skill of the forecast system and is therefore often shaded in attributes diagrams.

4.4.1 Attributes diagram

In the upper (lower) row of Fig. 4.9 attributes diagrams of average summer (winter) 2 m temperature anomalies over 50 randomly chosen European land grid points in the hindcast data being above (below) the 80^{th} (20^{th}) percentile in 1950–1980 and 1981–2010 are shown. Percentile values are calculated from ERA5 anomalies during the respective period and ERA5 is also used as verification. The size of the data points represents the number of values in each bin and the red line illustrates a weighted linear regression. Shading and reference lines are drawn as described above. REL, RES and UNC are calculated as shown in eq. 4.5 and BSS is calculated with respect to climatology according to eq. 4.7. Instead of using all available grid points within the defined region, only 50 randomly chosen ones are considered in order to reduce spatial dependencies within the data.

One can see in Fig. 4.9 that hindcast reliability for anomalously warm summers and cold winters is very good over Europe in both periods. On the other hand, resolution slightly improves towards the second period. Data points lie within the skillful area almost all the time and therefore also BSS indicates skill in all 4 cases, though a large improvement is visible towards the latter period. In JJA 1981–2010 (Fig. 4.9b) ASF-20C tends to over-forecast the real distribution, i.e. anomalously warm summer events occur less frequently than predicted (Mason and Stephenson, 2008). Figs. 4.9a and 4.9c indicate over-confident forecasts, i.e. forecast probabilities for warm summers/cold winters are higher than observed frequencies for forecast probabilities smaller than the climatological frequency and vice versa for forecast probabilities that are larger than the climatological frequency. In other words, low probabilities are under-forecast and high probabilities are over-forecast and the empirical curve crosses the perfect reliability line at or around the climatological probability of occurrence (Stanski et al., 1989). Mason and Stephenson

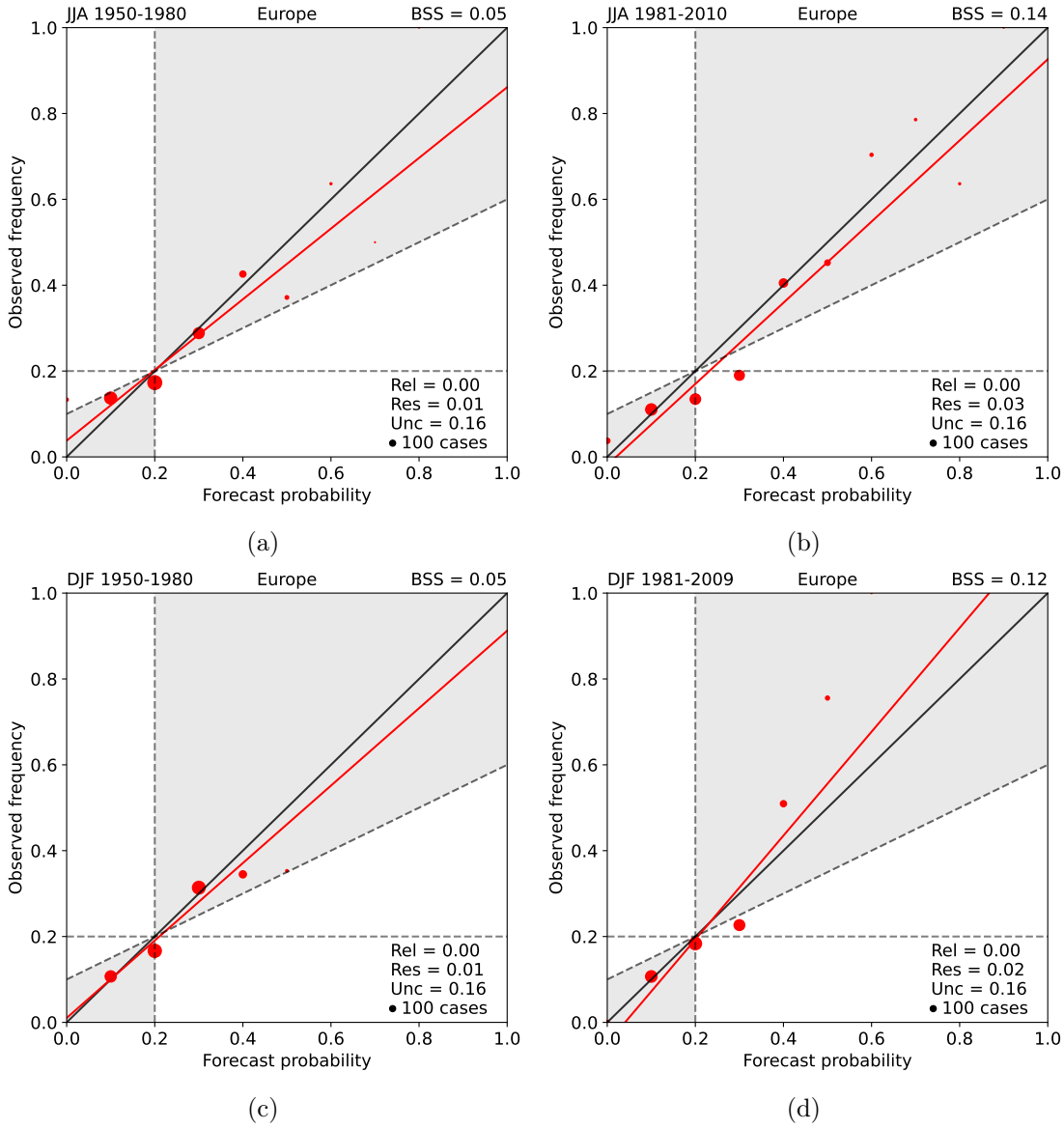
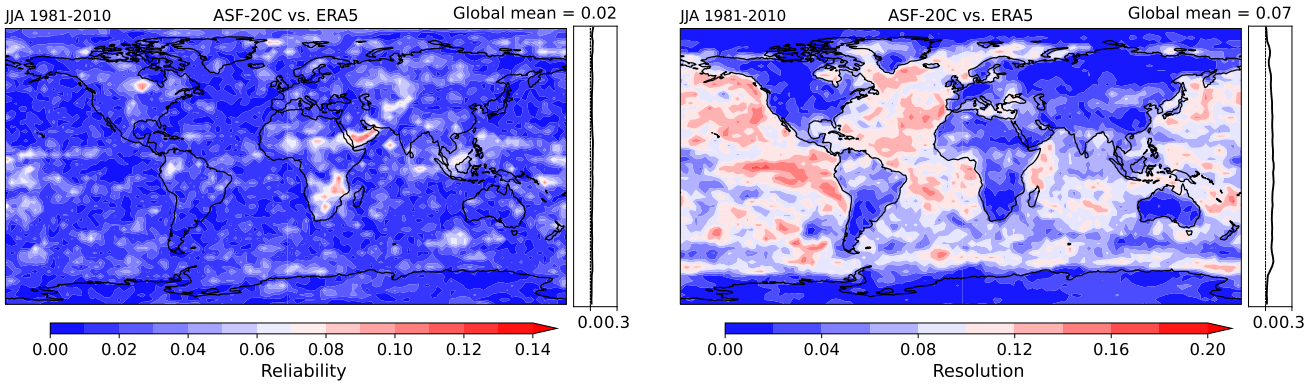


Figure 4.9: Attributes diagrams of summer/winter 2 m temperature anomalies at 50 randomly chosen grid points over European land being above/below the $80^{\text{th}}/20^{\text{th}}$ percentile in the top/bottom row for two different periods. The black and red lines indicate perfect reliability and weighted linear regression of the data, respectively. Horizontal, vertical and diagonal dashed lines show no resolution, optimal resolution, and $RES = REL$ lines, respectively. Shading indicates areas where red dots have to be located for the ensemble system to be called skillful.

(2008) state that this is the most frequent situation for seasonal climate forecasts. In Fig. 4.9d the opposite is true and thus the hindcast is said to be under-confident.



(a) (b)

Figure 4.10: Maps of (a) *REL* and (b) *RES* as defined in eq. 4.5 for JJA average 2 m temperature anomalies in the ASF-20C hindcast to be above the 80$^{\text{th}}$ percentile for the period 1981–2010. ERA5 is used for verification. Anomalies are calculated with respect to the same period and the respective data set.

4.4.2 Maps of REL, RES & BSS

In the attributes diagrams in Fig. 4.9 there were some probability bins, especially those with high forecast probabilities, without any cases during the whole period. To avoid this and to also account for possible spatial dependencies when investigating *REL*, *RES* and skill of re-forecasts globally, the number of cases in each probability bin within a 3° x 3° grid is summed before calculating the three quantities as stated above for each of these boxes. As in Fig. 4.9, the cases of DJF/JJA seasonal averages to be below/above the 20$^{\text{th}}$/80$^{\text{th}}$ percentile are considered. In the northern hemisphere this corresponds to the performance in cold winter/warm summer years and vice versa in the southern hemisphere.

The resulting maps show that *REL* of ASF-20C data compared to ERA5 is very close to zero all over the world in both seasons and in both periods. Between 1950 and 1980 exceptions are equatorial Africa in both seasons and JJA averages in the Southern Ocean. In both regions *REL* is above 0.1. Similar values emerge during the latter period in parts of southern Africa in both seasons and over the southern Arabian Peninsula during summer months of both periods as indicated by Fig. 4.10a where *REL* for JJA averages during the most recent period is shown.

Contrary to this, resolution shows high values almost only above the oceans, although even there large variations are visible. Between 1950 and 1980 performance is best over tropical oceanic regions except for the northern Indian Ocean during boreal winter months. In boreal summer during this period resolution is best over the North Atlantic, northern North Pacific, the Niño3.4 area and regions along the west coast of Africa and South America. During the more recent period regions with high resolution expand over all oceans in DJF and the northern hemispheric oceans in JJA. Fig. 4.10b shows that over the southern Atlantic and almost the whole Indian Ocean resolution is still rather small in this season. The only land areas with slightly elevated resolution are parts of northern South America, and some regions in northern and central Africa. This is valid in both seasons but only in the latter period. Very similar to correlation, resolution happens to be the best in regions that are most affected by ENSO.

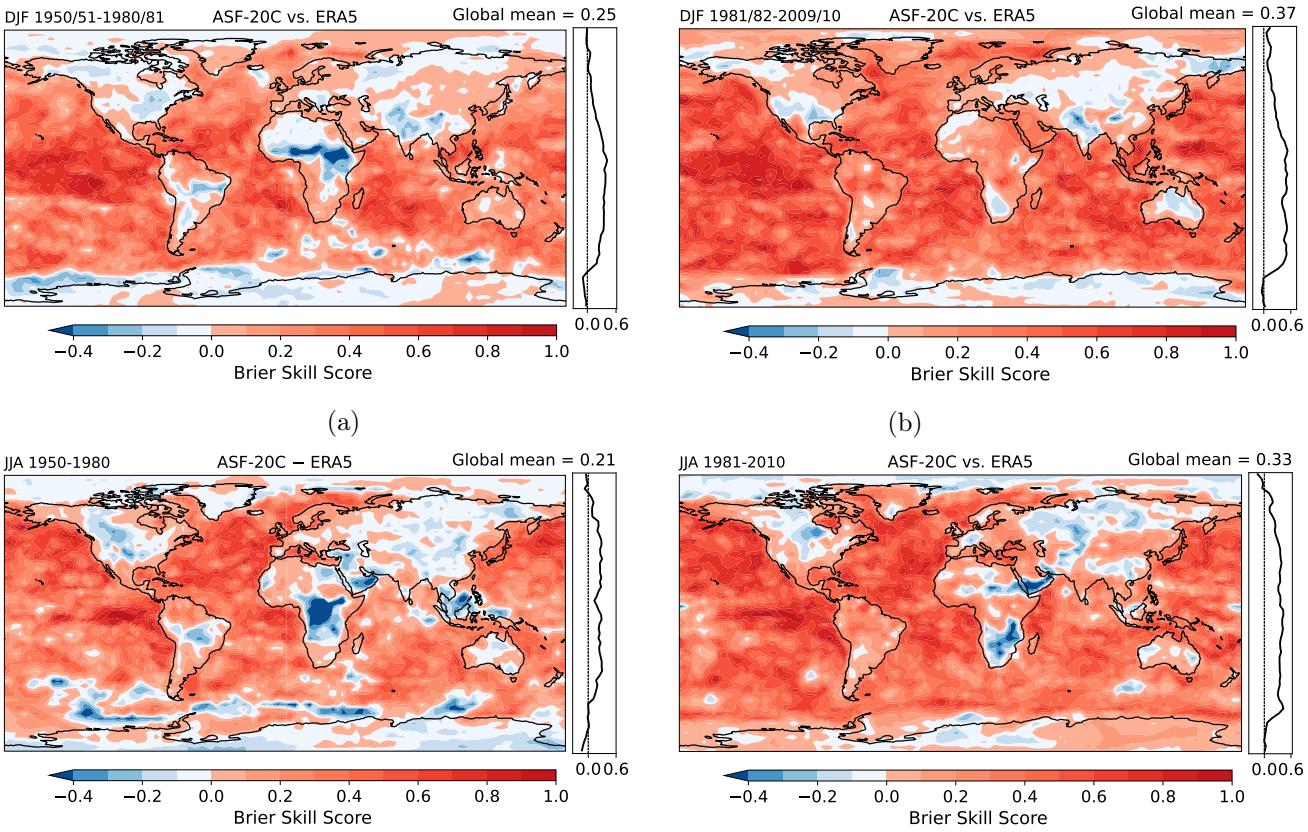


Figure 4.11: Brier Skill Score with respect to climatology of DJF average anomalies to be below the 20$^{\text{th}}$ percentile on a 3° x 3° grid during (a) 1950–1980 and (b) 1981–2009. (c) and (d) BSS with respect to climatology for JJA average anomalies to be above the 80$^{\text{th}}$ percentile during the periods 1950–1980 and 1981–2010, respectively. ERA5 is used for verification in all cases. Anomalies are calculated with respect to the respective period and data set.

BSS of DJF average anomalies to be below the 20$^{\text{th}}$ percentile is displayed in Figs. 4.11a and 4.11b for periods 1950/51–1980/81 and 1981/82–2009/10, respectively. Again climatology is used as reference forecast strategy. It can be seen that in the prior period ASF-20C has skill over all oceans except for parts of the Southern Ocean where BSS is strongly negative, which is most probably again an effect of issues with prescribed SIC data. These patterns are even more pronounced for JJA averages to be above the 80$^{\text{th}}$ percentile during the same period as can be seen in Fig. 4.11c. The situation over continents is more complex. During the first period BSS is close to zero, both positive and negative, over all continents. Very similar to reliability, performance is much worse over central Africa in both seasons which leads to strongly negative skill scores there. However, as already mentioned in Sect. 3.3, Simmons et al. (2021) stated that a warm bias occurs in ERA5 near the Congo Basin in the early 1950s and over Brazil in 1961. In both regions BSS of JJA averages is strongly negative in the first period (see again Fig. 4.11c) and partly it is also visible for DJF averages in Fig. 4.11a. Thus I do not

consider this as a shortcoming of the hindcast performance but of ERA5 output. In the most recent period, skill of DJF averages is clearly positive in South America and Africa as well is in some parts of North America and this time also in the Southern Ocean and the Arctic. Contrary to this, skill in Eurasia is very similar to that of the 1950–1980 period. During boreal summer months of the more recent period, which is displayed in Fig. 4.11d, similarities to reliability patterns in Fig. 4.10a can be obtained. This time strongly negative skill scores emerge over southern Africa and the Arabian Peninsula. BSS for warm summer anomalies is also slightly negative over northern hemispheric continents, very similar to what is visible for cold winters in Fig. 4.11b too. On the other hand, ASF-20C forecasts for warm JJA anomalies during 1981–2010 are again skillful over South America and equatorial Africa. What Fig. 4.11 also shows is that firstly over oceans BSS is higher in the summer hemisphere and secondly global averages are by 0.04 higher in DJF than in JJA. If more extreme events, e.g. 2 m temperature anomalies being above/below the $90^{\text{th}}/10^{\text{th}}$ percentile, are considered, both the number and area of regions with small or even negative BSS increase. Contrary to this global mean BSS is higher when less extreme percentiles, e.g. $66^{\text{th}}/33^{\text{rd}}$ or 50^{th}, are considered. Still, there is a significant improvement from the prior towards the most recent period in both seasons and global mean BSS reaches values close to +0.4 for these percentile values, as is also indicated by Fig. 4.11b.

4.5 Signal-to-noise ratio

The signal-to-noise ratio (SNR) tries to quantify how much useful information can be extracted from a time series. For the case of ensemble data, Weisheimer et al. (2019) define signal as the interannual variation of the ensemble mean (VAR_{signal}) and noise as the variation of all ensemble members around their mean value (VAR_{noise}). Both are calculated for a well-defined temporal period. In this work SNR of seasonal averages is defined as

$$SNR = \frac{VAR_{\text{signal}}}{VAR_{\text{noise}}} \tag{4.8}$$

which is the same as the predictable component in Polyakov et al. (2022) but slightly different to SNR in Weisheimer et al. (2019). There the denominator of the formula is $VAR_{\text{total}} = VAR_{\text{signal}} + VAR_{\text{noise}}$. However, the advantage of the definition in eq. 4.8 is that SNR adopts values between 0 and 1. The reason is that both VAR_{signal} and VAR_{noise} have to be $\geqslant 0$ and if there are no variations at all within the ensemble in each season during the whole period $VAR_{\text{signal}} = VAR_{\text{noise}}$ and thus $SNR = 1$. In any other case $0 < SNR < 1$ since then $VAR_{\text{noise}} > VAR_{\text{signal}}$ has to be valid.

The two components are defined as

$$VAR_{\text{signal}} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \overline{\tilde{x}})^2 \tag{4.9}$$

$$VAR_{\text{noise}} = \frac{1}{n \cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i,j} - \overline{\tilde{x}})^2 \tag{4.10}$$

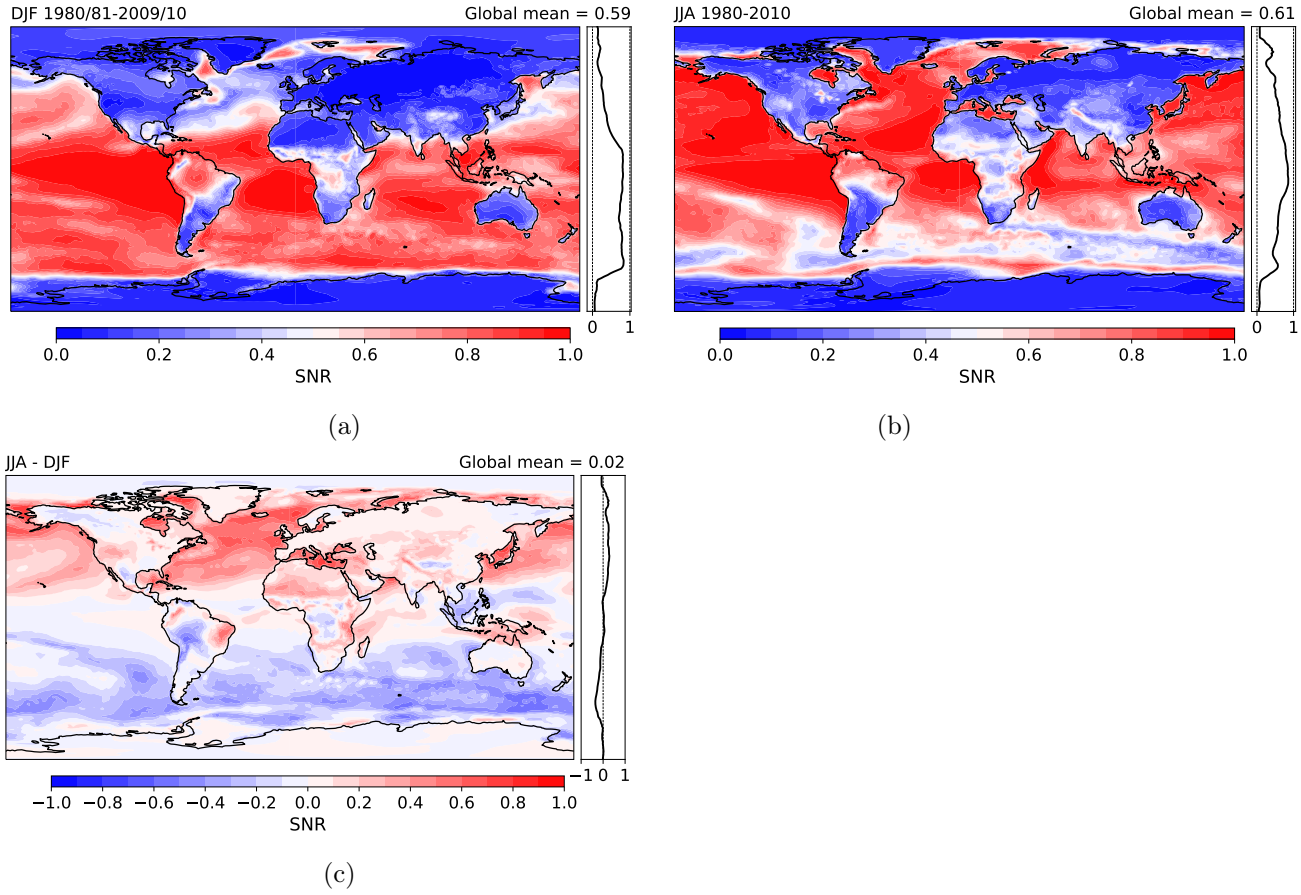


Figure 4.12: ASF-20C signal-to-noise ratio on grid-point scale of (a) DJF averages for the period 1980/81–2009/10 and (b) JJA averages for the period 1980–2010. (c) Differences of SNRs in (a) and (b).

where $x_{i,j}$ represents the j-th ensemble member in the i-th year, $\tilde{x}_i$ denotes the ensemble mean in the same year and $\overline{\tilde{x}}$ is the temporal average of all ensemble means during the considered period. If all ensemble members have the same value in one season $\tilde{x}_i = x_{i,j}$ has to be valid. Inserting this in eq. 4.10 and rearranging gives

$$\begin{aligned} VAR_{\text{noise}} &= \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} (\tilde{x}_i - \overline{\tilde{x}})^2 = \frac{1}{n \cdot m} \left[\sum_{i,j}^{n,m} \tilde{x}_i^2 - 2\overline{\tilde{x}} \sum_{i,j}^{n,m} \tilde{x}_i + \sum_{i,j}^{n,m} \overline{\tilde{x}}^2 \right] = \\ &= \frac{1}{n \cdot m} \left[m \sum_{i}^{n} \tilde{x}_i^2 - 2m\overline{\tilde{x}} \sum_{i}^{n} \tilde{x}_i + m \sum_{i}^{n} \overline{\tilde{x}}^2 \right] = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - \overline{\tilde{x}})^2 = \\ &= VAR_{\text{signal}} \end{aligned} \tag{4.11}$$

which leads to the already mentioned case of $SNR = 1$.

In general, signal-to-noise ratio of seasonal forecasts is smaller than unity and smaller in the extratropics than in tropical regions because the main source of predictability on these time-scales, ENSO, mostly affects the tropical Pacific. And on its way around the

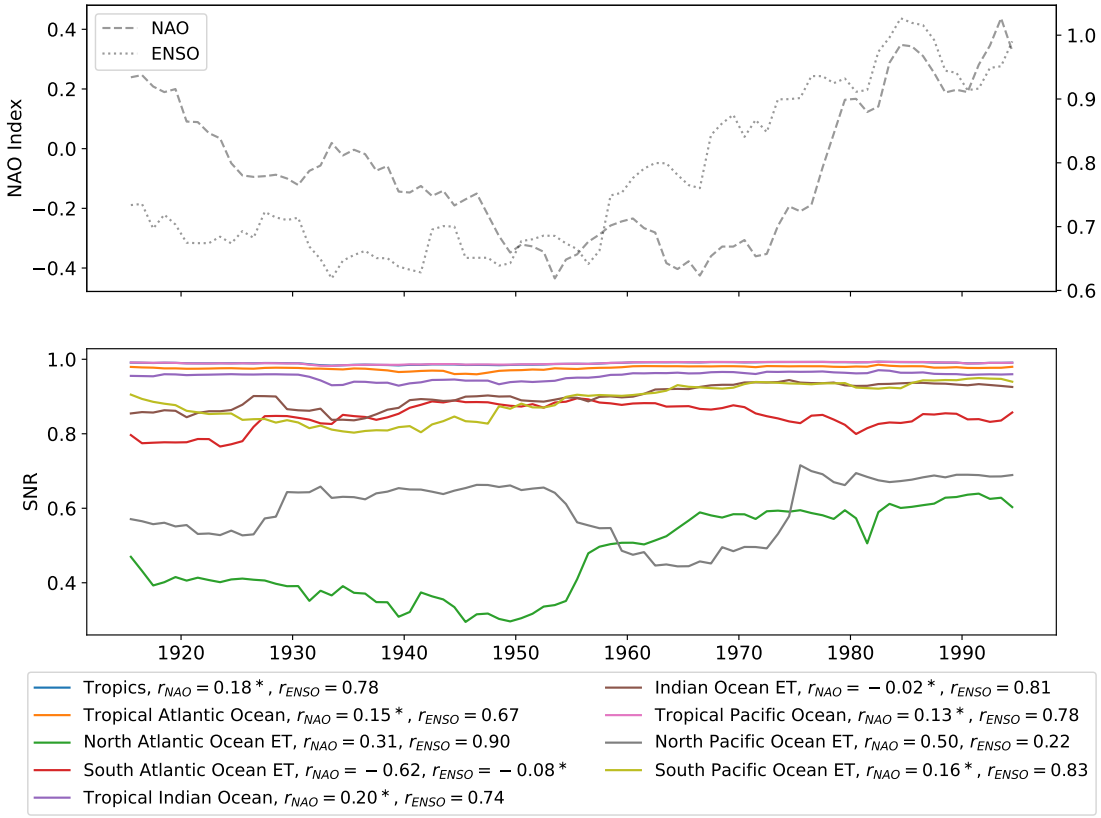


Figure 4.13: *Upper panel:* Averages of NAO index (dashed) and standard deviations of Niño3.4 index (dotted) in a moving 30-year window as described in Sect. 4.3. *Lower panel:* Signal-to-noise ratio of detrended sea-only DJF averages in different ocean regions calculated in a moving 30-year window. Correlations of each time series with NAO index and ENSO activity time series from the upper panel are given in the legend as r_{NAO} and r_{ENSO}, respectively. Asterisks denote values that are not significant on the 95% level.

globe the signal weakens due to wave disturbances and interactions with the mean flow (Weisheimer et al., 2019). SNR of detrended DJF averages of ASF-20C hindcast for the period 1980/81–2009/10 is shown in Fig. 4.12a. It can be seen that the highest values occur over all oceans, which is not surprising as 2 m temperature is strongly coupled to SSTs which are prescribed as lower boundary condition. Exceptions are in mid-latitudes of the North Atlantic and around Japan. These regions very much correspond to the Gulf Stream and Kuroshio, respectively. This leads to the assumption that highly variable surface fluxes (in dependence of weather conditions) in these regions lead to a larger ensemble spread than in other ocean regions. However, this behaviour is only visible for DJF averages when the storm tracks in these regions are most active. As Fig. 4.12b indicates, during JJA signal-to-noise ratio is only slightly reduced in the western Atlantic Ocean where the Gulf Stream originates but not in any other part of the Atlantic Ocean or near Japan. Apart from these two regions, SNR also exhibits notable seasonal changes

in the Southern Ocean. While values almost reach unity in DJF, they are below 0.5 when JJA averages are considered, for similar reasons as in the Northern Hemisphere storm track regions. Fig. 4.12c shows SNR of JJA-DJF. It can be seen that outside of tropical oceans performance in the summer hemisphere is much better. In most continental regions SNR is slightly higher in JJA. Exceptions are for example central Africa, a part of northern South America as well as Malaysia and Indonesia. These are regions where ASF-20C forecasts strongly benefit from ENSO influence. Figs. 4.12a and 4.12b show that there SNR is enhanced in both seasons but larger in DJF.

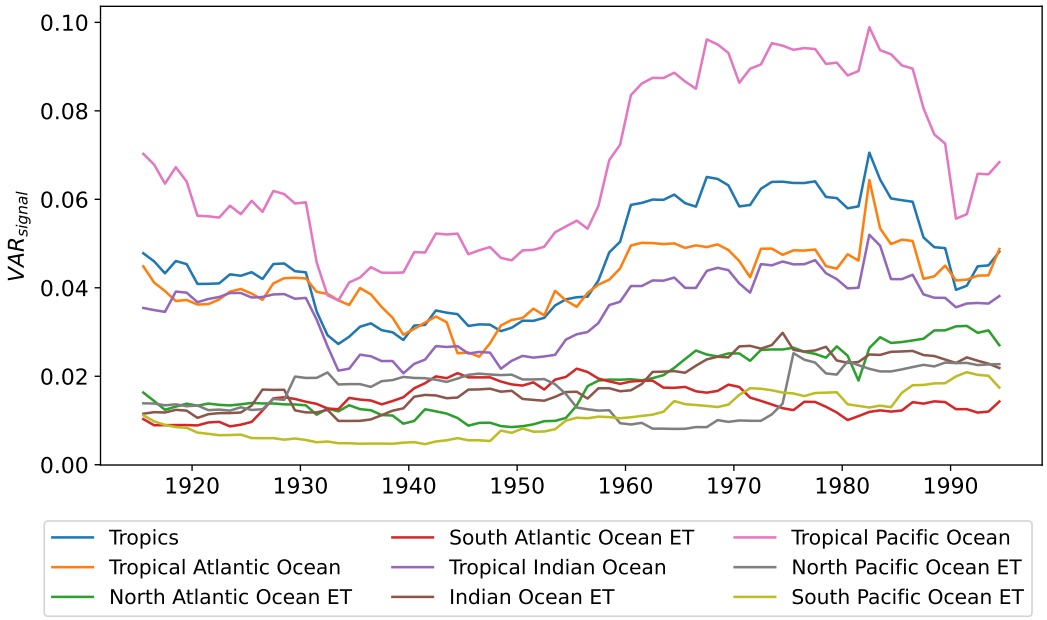
Signal-to-noise ratio is very small in almost all remaining continental regions and the reason was already visible in Fig. 4.3. Variability of the ensemble members around their mean value, i.e. VAR_{noise}, is much larger in many continental regions, like GAR, than the variance of the ensemble mean over the course of 30 years, i.e. VAR_{signal}. But the difference is way smaller over South America and at least reduced over Africa. This behaviour as well as the generally very small SNR over all other continents is almost independent of the considered season.

There are also small changes of the signal-to-noise ratio over the course of the 20^{th} century, but again not only because of improvements of the observational system. As described for correlation in Sect. 4.3, again the influence of multi-decadal variations of NAO index and ENSO activity on SNR is investigated. Thus, the upper panel of Fig. 4.13 is identical to the upper panel of Fig. 4.7.

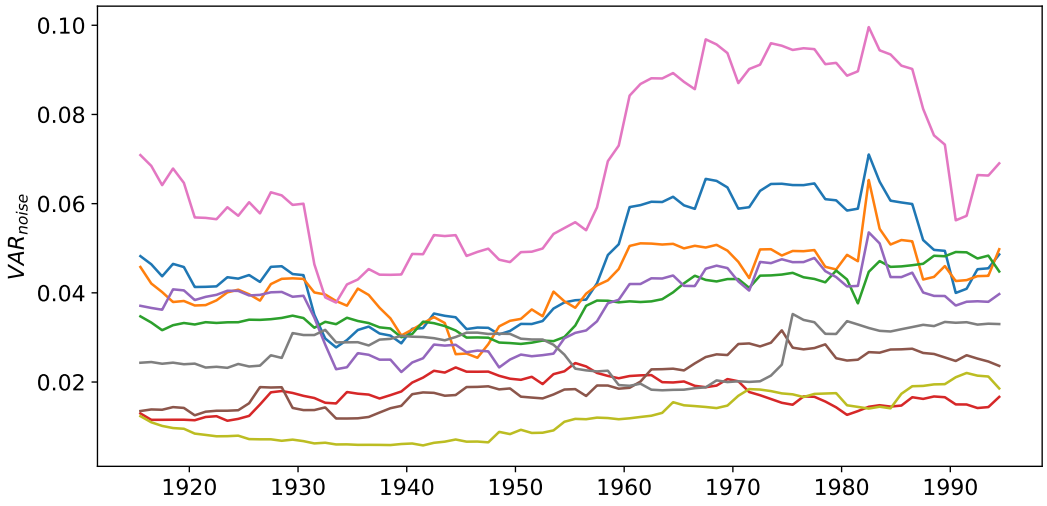
The lower panel of Fig. 4.13 contains signal-to-noise ratios of sea-only DJF 2 m temperature averages of different ocean regions calculated in a moving 30-year window. r_{NAO} and r_{ENSO} in the label box denote correlation coefficients of SNR time series with NAO index and ENSO activity curves in the upper panel, respectively. Asterisks indicate which values are not statistically significant on a 95% level. Before calculating SNR, again the ensemble mean trend is subtracted from each ensemble member in each 30-year window. This is necessary because, as already stated at the end of Sect. 4.3, otherwise temperature trends influence SNR and thus lead to spurious correlation. Note that in Figs. 4.15b and 4.15c the term *Tropics* refers to land-only grid points between 30°S and 30°N, as defined in Tab. 2.1, whereas for the investigation of SNR over ocean regions, border latitudes of the *Tropics* region are 20°N/S and sea-only grid points are considered.

In almost all considered regions SNR is around or even above 0.8, exceptions are North Atlantic extratropical (ET) and North Pacific Ocean ET. The largest variations on decadal time scales are also found in these regions. While in the North Atlantic ET SNR is mostly below 0.4 in the first half of the century it increases to values around 0.6 towards more recent decades. The shift to higher values is almost at the same time when also ENSO forecast skill enhances strongly. A correlation coefficient with ENSO of +0.90 confirms the link. Apart from South Atlantic ET and North Pacific ET, correlation with ENSO activity is very high in all ocean regions. However, in the latter region even the small correlation of around +0.22 is a significant result. In North Pacific ET, SNR is quite uniform around 0.6 during the whole century except for the period 1955–1975 where values descend to below 0.5. Correlation with the averaged NAO index is +0.50 and by far larger than in any other ocean region, except for the rather strong negative correlation in South Atlantic ET. In North Atlantic ET correlation with NAO index is only +0.31 but still significant, indicating that there the sign of NAO is at least partly the reason for reduced 2 m temperature forecast skill.

Time series of VAR_{signal} and VAR_{noise} for the same regions are shown in Fig. 4.14. In



(a)

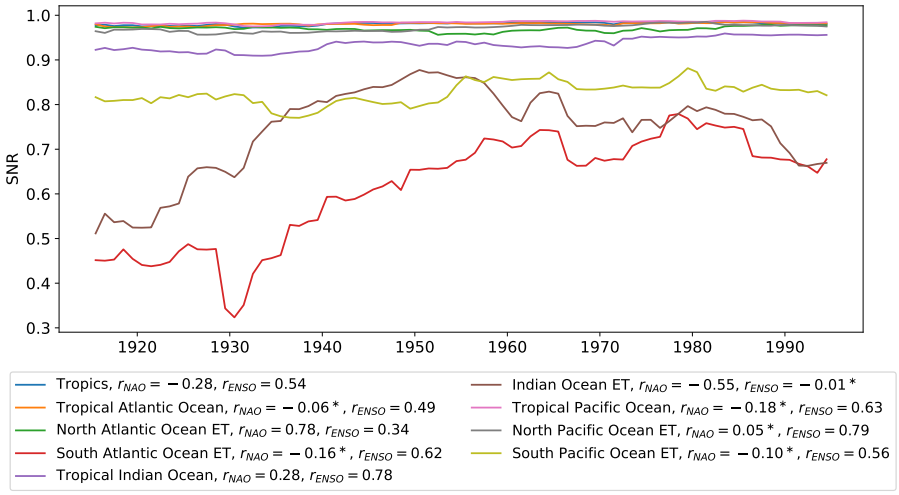


(b)

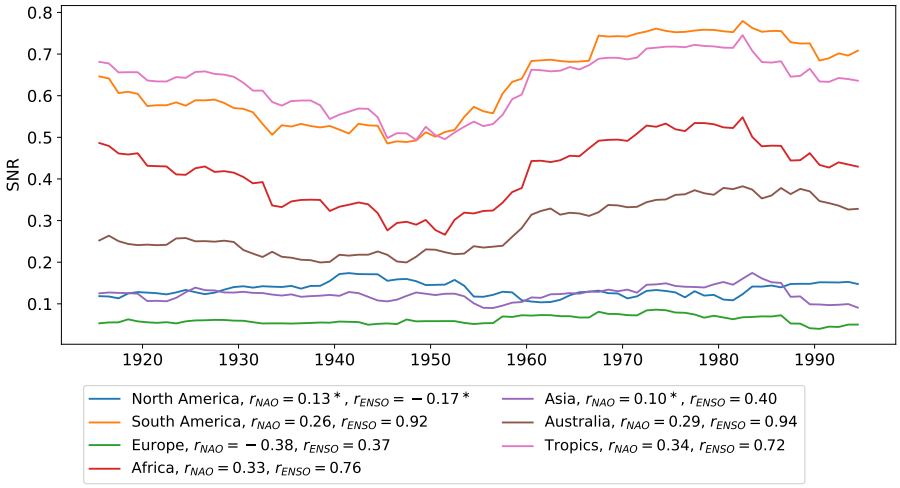
Figure 4.14: (a) VAR_{signal} and (b) VAR_{noise} of ocean regions calculated according to eqs. 4.9 and 4.10 in a moving 30-year window. Color coding in (b) is identical to (a).

some regions like the Tropical Pacific Ocean and the Tropics variations of both quantities are rather large over the course of the century. Again in many regions smallest values of both VAR_{signal} and VAR_{noise} occur between the 1930s and 1950s, confirming that lower ENSO predictability at that time may be responsible. But since signal and noise curves almost run parallel, these variations are not visible in the SNR time series of these regions.

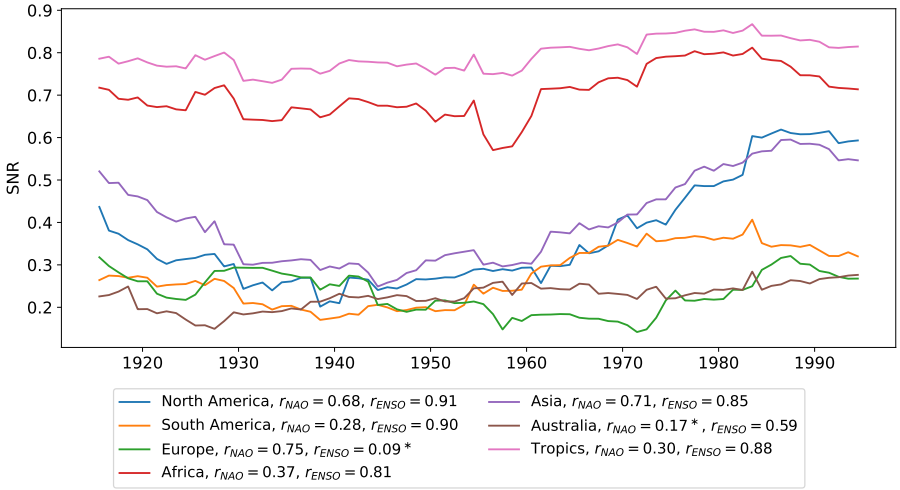
As indicated by Fig. 4.15a, SNR of JJA averages in the same ocean regions show several differences to boreal winter values. Though again values are above 0.8 in the most regions,



(a)



(b)



(c)

Figure 4.15: Same as the lower panel of Fig. 4.13 but for (a) JJA averages over oceans, (b) DJF averages over continents and (c) JJA averages over continents.

this time Indian Ocean ET shows values between 0.5 and 0.7 prior to the 1930s and in South Atlantic Ocean ET SNR does not reach 0.7 until 1960. Also correlation coefficients differ strongly from DJF values. Correlations with ENSO activity are smaller (between +0.34 and +0.63) but still significant in Tropics, Tropical Atlantic, North Atlantic ET, Tropical Pacific and South Pacific ET. On the other hand in Tropical Indian Ocean, South Atlantic ET and North Pacific Ocean ET correlations are higher than in DJF with values of about +0.78, +0.62 and +0.79, respectively. Especially in the latter two regions, where differences between both seasons are the largest, forecast skill of 2 m temperature seems to lag by several months because of ENSO teleconnections that transport the signal to remote regions. Interestingly, the only region that shows no correlation at all with ENSO activity in JJA is Indian Ocean ET, where in DJF correlation was above +0.80. The highest correlation with the NAO index in this season occurs by far in North Atlantic ET with a value of approximately +0.78. In all other regions correlation coefficients are either small (Tropical Indian Ocean, $r_{\mathrm{NAO}} \approx +0.28$), negative (Tropics and Indian Ocean ET, $r_{\mathrm{NAO}} \approx$ -0.28 and -0.55, respectively) or not statistically significant at all.

As already indicated by Fig. 4.12, over continents SNR is in general much lower than over oceans. Boreal winter SNR is displayed in Fig. 4.15b and is above 0.6 only in South America, except for the 1920s–1950s, and in the Tropics except for the 1930s–1950s. In JJA however, SNR is only in Africa and the Tropics above 0.6, as can be seen in Fig. 4.15c. The time series of these two regions run almost parallel with the latter being approximately 0.1 higher. In both seasons SNR exhibits strong variations on multi-decadal time scales in several continents like North America and Asia in JJA and Africa in DJF. Correlation of ENSO activity with SNR in DJF in South America and Australia reaches values of +0.92 and +0.94, respectively, which is even higher than over oceans. Also in the Tropics and Africa correlation is above +0.70. r_{ENSO} of JJA average SNR is close to zero in Europe but almost +0.60 in Australia and between +0.81 and +0.91 for all other continents with highest values in the Tropics, North America and South America. This again underlines the global importance of ENSO for seasonal forecast systems even in regions with large spatial distances to the tropical Pacific. On the other hand, r_{NAO} for DJF averages is either not statistically significant (North America, Asia), small and negative (-0.38 for Europe) or positive and around 0.30 (all other continents). Similarly to North Atlantic ET, correlations with the NAO index are a lot higher in boreal summer with values of +0.68, +0.71 and +0.75 in North America, Asia and Europe, respectively. Correlations in South America, Africa and the Tropics are again around +0.30 and in Australia even smaller and not significant. Thus, NAO is also a very important factor for forecast skill in the northern hemisphere during summer but only plays a minor role elsewhere. However, the fact that the influence of NAO seems to be more present during JJA is to some extent surprising since NAO is actually defined for boreal winter months.

5 Temperature trends

In the previous section, it was shown that the ASF-20C hindcast ensemble output is, among other properties, reliable and skillful, even if it is compared to ERA5, which is a state-of-the-art reanalysis product and not used for initialization. In this section, the agreement of temperature trends within the these two data sets as well as ERA-20C will be investigated and it will be shown that ASF-20C is able to reproduce low-frequency temperature changes over the course of the 20$^{\text{th}}$ century.

5.1 Ensemble mean trend

To investigate how good ASF-20C re-forecasts reproduce 2 m temperature trends over the course of the 20$^{\text{th}}$ century, linear trends of the ensemble mean are compared to ERA-20C trends in three and to ERA5 trends in two different and non-overlapping periods. The linear regression coefficients are calculated during each of these periods on a 1° x 1° grid and will be shown for both hindcast and reanalyses. To find regions where differences between the data sets are statistically significant, a z-test is applied using (Clogg et al., 1995; Paternoster et al., 1998):

$$z = \frac{b_1 - b_2}{\sqrt{s_1^2 + s_2^2}} \tag{5.1}$$

where b_1, b_2 are the linear regression coefficients of hindcast and reanalysis data, and s_1, s_2 are their respective standard errors.

Fig. 5.1 shows JJA averages of 2 m temperature trends in K per decade during the periods 1901–1936 (upper row), 1937–1972 (middle row) and 1973–2010 (lower row) of ASF-20C ensemble mean and ERA-20C in the left and right column, respectively. Black dots indicate regions where differences are significant on the 95% level. During the beginning of the century (Figs. 5.1a and 5.1b) ASF-20C indicates slightly positive trends over North America, while ERA-20C trends are up to +1 K per decade in the northern part and negative with values around −0.5 K per decade in the southern part of the continent. Over central and southern Africa, as well as the southern tip of South America no or only small positive trends are found in the hindcast data, whereas ERA-20C reanalysis shows values down to −0.7 K per decade in these regions. Differences are statistically significant in at least some parts of these regions in all lead months except May.

During the second period, which is displayed in Figs. 5.1c and 5.1d hindcast temperature trends over continents are slightly negative over central and northeastern Asia as well as central and western USA and close to zero everywhere else. ERA-20C trends differ significantly from that with values well above +1 K per decade in southeast Asia and central Africa and of the same magnitude but negative over the USA. Over the Antarctic, ERA-20C trends indicate strong cooling but re-forecasts show weak warming trends. Interestingly, both data sets agree with very strong positive and negative trends around the Antarctic and only show small spatial shifts.

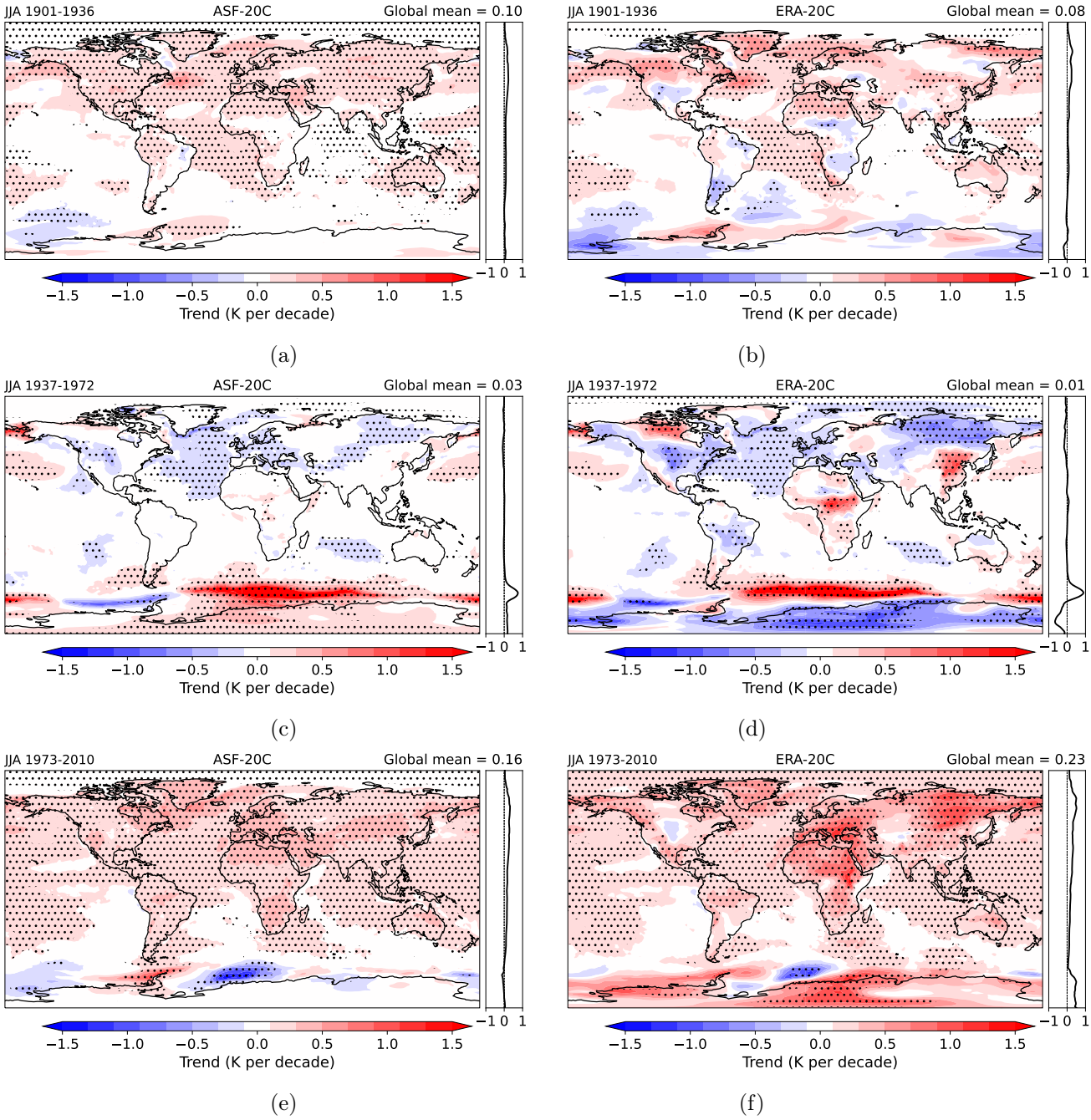


Figure 5.1: Linear trend of JJA average 2 m temperatures during the periods 1901–1936 (upper row), 1937–1972 (middle row) and 1973–2010 (lower row). ASF-20C ensemble mean and ERA-20C are shown in the left and right column, respectively. Dotted areas indicate regions where trends are significant on the 95% level.

Over the course of the most recent period (Figs. 5.1e and 5.1f), ASF-20C ensemble mean shows positive trends up to +0.3 K per decade over most continental and oceanic

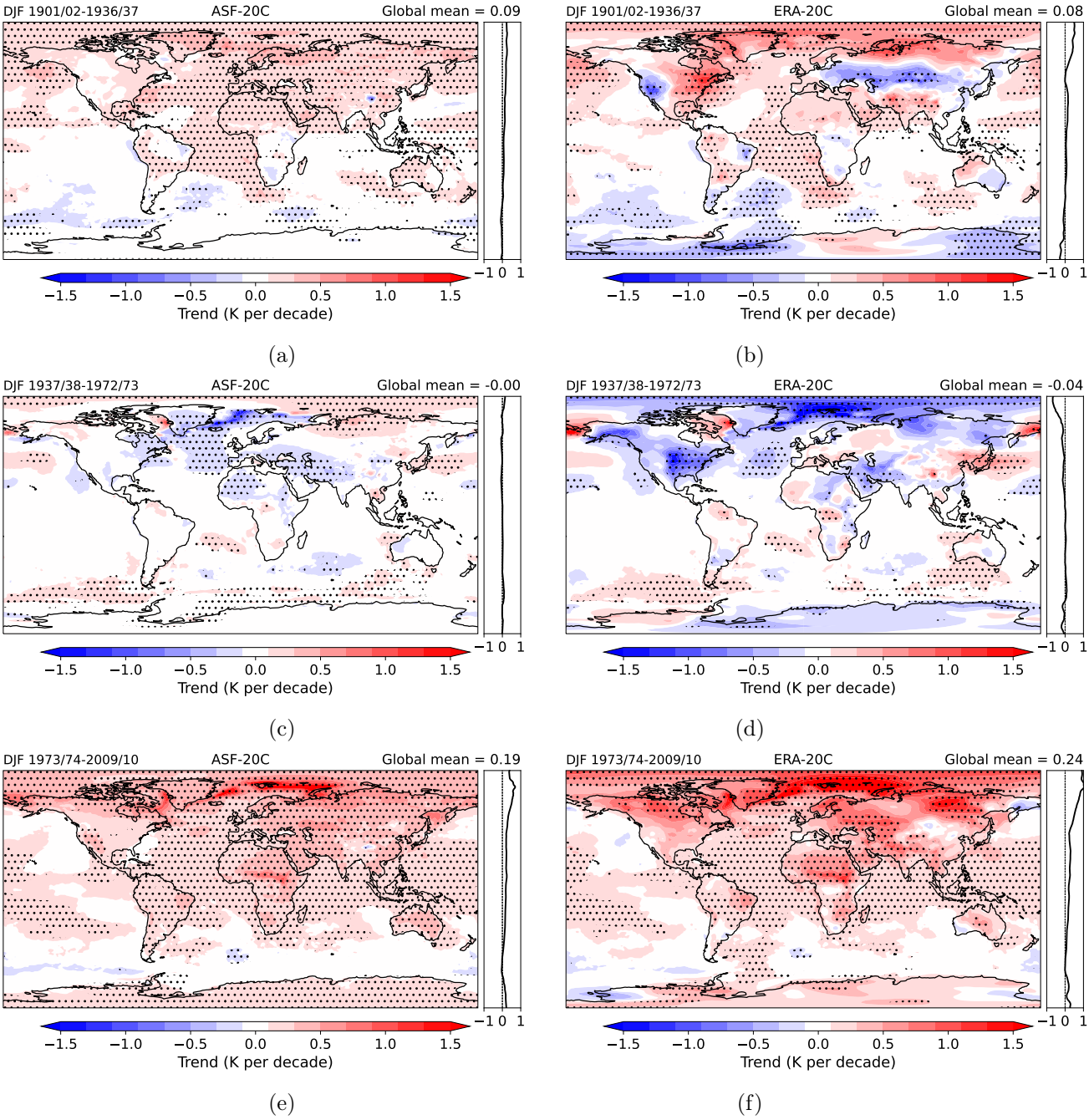


Figure 5.2: Same as Fig. 5.1 but for DJF averages of 1901/02–1936/37, 1937/38–1972/73 and 1973/74–2009/10 in the upper, middle and lower row, respectively.

regions at all lead times. Slightly higher values can be found in northern and southern Africa, Europe and central Asia. Trends are close to zero over polar regions, the South Atlantic and the eastern South Pacific Ocean. Some spurious patterns again emerge in the Southern Ocean around the Antarctic. ERA-20C trends are almost identical over oceans but in some continental regions by far larger, e.g. in northeastern Asia, northern Africa

and southern Europe values exceed +1 K per decade in every month.

As indicated by Fig. 5.2, in boreal winter the largest differences between ASF-20C ensemble mean and ERA-20C trends during the period 1901/02–1936/37 appear over northern hemispheric mid-latitudes. Hindcast trends are slightly positive there with values up to +0.3 K per decade but reanalysis trends exceed −1.5 K per decade in Asia and the western USA and +1 K per decade in eastern North America. These differences are statistically significant on large spatial scales. Other significant differences can be found over the Antarctic and the Southern Ocean.

During the middle period, shown in Figs. 5.2c and 5.2d, ensemble mean temperature trends over the Arctic and northern Russia are up +0.3 K per decade but negative ERA-20C trends are significantly stronger with values down to −1.5 K per decade. Also in eastern North America and around the Persian Gulf ERA-20C indicates large negative trends but the ensemble mean trend is either close to zero or only slightly negative in both regions. Smaller but still significant trend differences appear over northwestern and central Africa, where ASF-20C trends are slightly negative and positive, respectively. ERA-20C data shows positive trends around +0.5 K per decade in both regions.

From 1973 onwards, there are almost no locations with significant differences between ASF-20C ensemble mean and ERA-20C trends. As Figs. 5.2e and 5.2f indicate, both data sets show no or weak positive trends over all oceans, except for some small regions in the Southern Ocean, and values well above +1 K per decade over some continental regions, though the trend magnitude decreases with lead time in the hindcast data and reanalysis trends tend to be larger especially over northern hemispheric continents and the Arctic. The reason for less extreme hindcast trends than in the reanalysis throughout the whole period and in both seasons is most probably that only the ensemble mean and no single run was considered so far in this section. Thus, even if some single runs showed similarly large trends as ERA-20C, it would not be visible in any of the maps due to the averaging process.

ASF-20C ensemble mean and ERA5 trends are compared during two 30-year periods. Fig. 5.3 displays these trends for JJA averages of 1950–1979 (upper row) and 1980–2010 (lower row). During the prior period, agreement between both data sets is high over oceans with slightly negative trends over the North Atlantic and parts of the North Pacific Ocean and small positive values in the South Atlantic as well as in parts of the Indian and the South Pacific Ocean. Differences are within a range of ±0.3 K per decade, in most parts even smaller than ±0.1 K per decade. The largest trends occur over and around the Antarctic, a region that is already known for showing rather questionable patterns in all data sets (see e.g. Fig. 4.1 and related description in the text). Besides this region, differences of up to and above +1 K per decade are also found for example over northern and central Africa, Brazil and central Asia. In all of these regions differences are significant. In central Africa and Brazil a warm bias in the reanalysis data, which was already mentioned previously, leads to negative trends in both regions that are stronger than −1 K per decade. The corresponding trends in ASF-20C are around $+/-0.2$ K per decade over the Congo Basin/Brazil.

Trend differences are not that pronounced in boreal summer of the more recent period as can be seen in Figs. 5.3c and 5.3d. Significant positive differences of JJA averages occur over parts of central Asia and southern Africa as well as over northwestern Australia and Canada. In all of these regions ASF-20C ensemble mean trend is up to +0.5 K per decade but ERA5 shows no or slightly negative trends. Contrary to this, hindcast trends

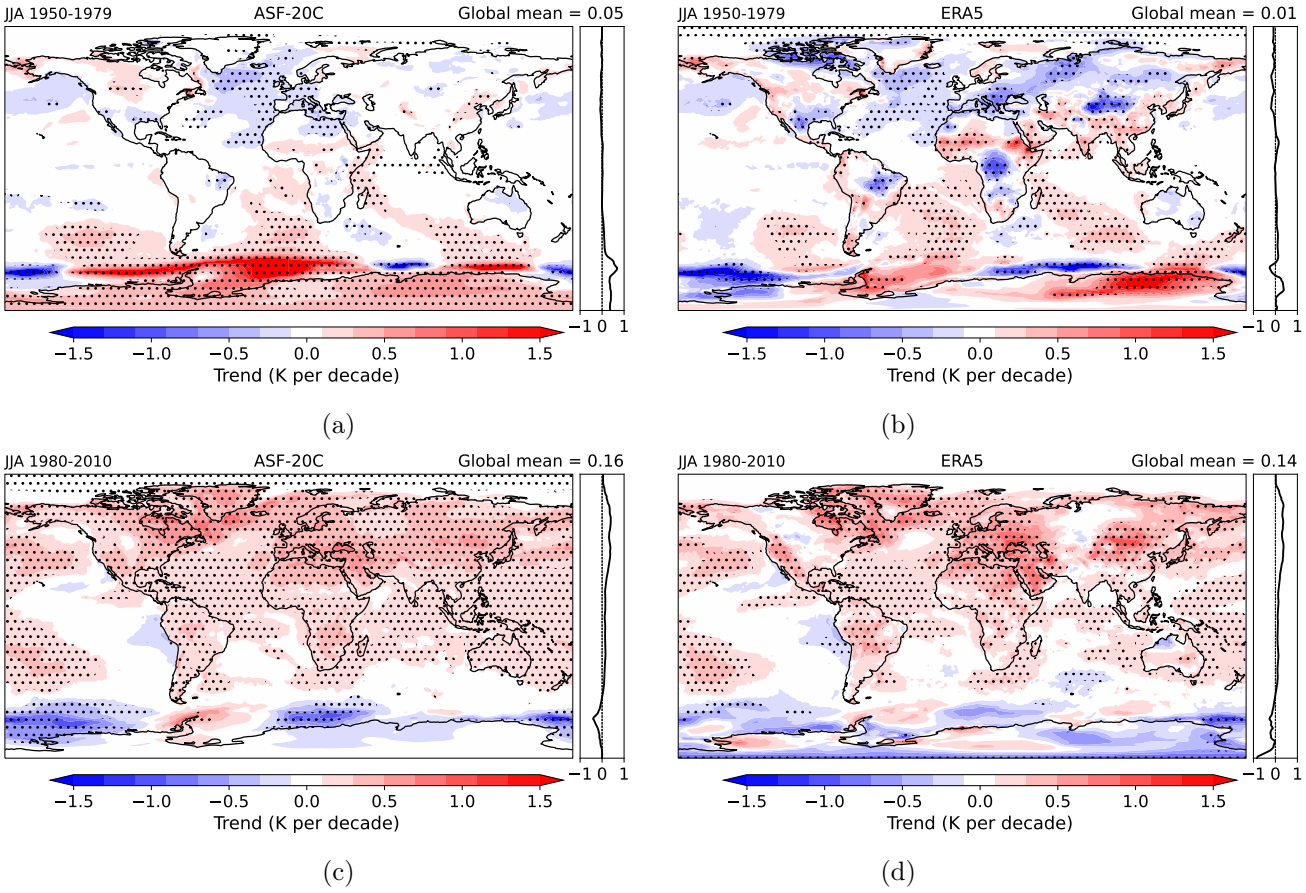


Figure 5.3: Linear trend of JJA average 2 m temperatures during the periods 1950–1979 (upper row) and 1980–2010 (lower row). ASF-20C ensemble mean and ERA5 are shown in the left and right column, respectively. Dotted regions indicate where trends are significant on the 95% level.

are smaller than ERA5 warming trends in central South America, eastern Europe, the Arabian Peninsula and central Asia eastward of the previously mentioned pattern. In most of these cases differences do not exceed ± 0.5 K per decade. However, as already mentioned previously, it is almost impossible for the ensemble mean to show as extreme trends as the reanalysis does. The 51 ensemble members aim at predicting different possible future evolution pathways of the atmosphere and the real outcome, which is approximated by the reanalyses in this work, does not necessarily have to be identical to the average of all these possibilities.

Results for DJF averages during the same periods are displayed in Fig. 5.4. Over oceans trends during the prior period are very similar to boreal summer at this time with negative/positive values dominating in the Northern/Southern Hemisphere. In most of the USA and parts of northern Asia ERA5 trends are well below -1 K per decade, while values in hindcast data are close to zero there. Differences are only significant over eastern USA in January, February and for DJF averages and in January in a small region over northwestern Russia. ERA5 indicates warming of more than $+1$ K per decade over northern and eastern Africa as well as over western and central South America. Consequences of the warm

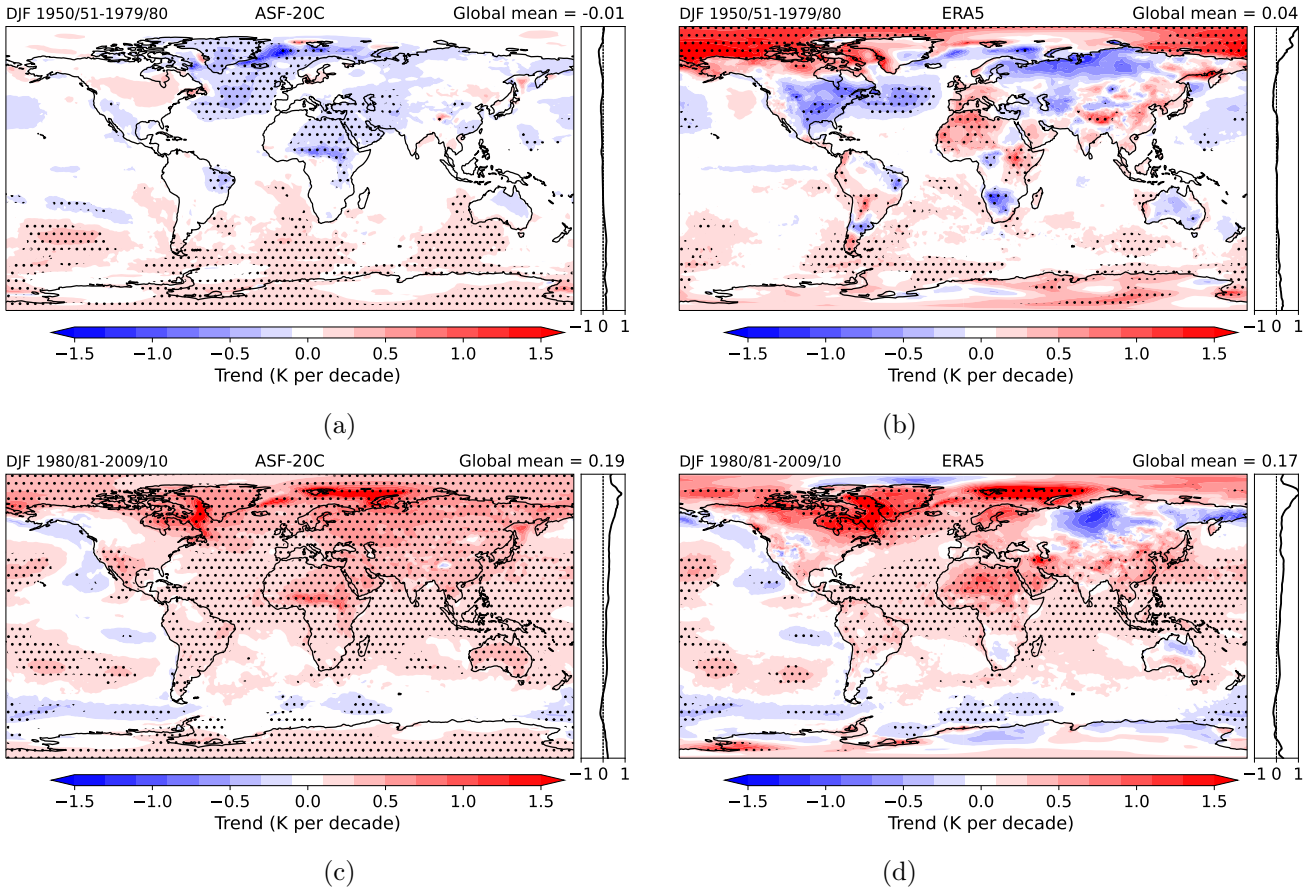


Figure 5.4: Same as Fig. 5.3 but for DJF averages of 1950/51–1979/80 and 1980/81–2009/10 in the upper and lower row, respectively.

ERA5 bias mentioned before are also present but weaker in this season in both regions. Moreover, negative trends in ERA5 are visible over southern Africa in all months. Of all these regions, re-forecasts only show a weak cooling trend in northern Africa and no trend at all in South America, except for slightly negative values over the eastern tip of the continent.

DJF averages of the more recent period from 1980 onwards in Figs. 5.4c and 5.4d show temperature trends of the ensemble mean around +0.5 K per decade over all of Asia but ERA5 indicates negative trends mainly in northern parts of the continent, though they are not statistically significant. In Africa both data sets show clearly positive trends in all months. The ensemble mean trend for the seasonal average is up to +0.5 K per decade larger than the trend in ERA5 south of the Sahara and in the western USA and vice versa over northern Africa and eastern Asia. Differences are also significant over northern Australia since both data sets indicate trends of different signs.

Tab. 5.1 lists spatially averaged trends of all continents as well as the Arctic and global values for ASF-20C ensemble mean, ERA-20C and ERA5 for both seasons and the periods 1950–1979 and 1980–2010. It can be seen that during the first period only very few of the calculated trends in any of the data sets are statistically significant. Apart from ASF-20C trends in Africa in boreal winter, trends are significant only over the Antarctic

Table 5.1: Temperature trends in K per decade of ASF-20C ensemble mean, ERA-20C and ERA5 for DJF and JJA averages in different regions during the periods 1950/51–1979/80, 1980/81–2009/10 and 1950–1979, 1980–2010, respectively. Only grid points over land are considered unless stated otherwise. Asterisks denote trends that are not statistically significant on the 95% level. Columns *Min* and *Max* show most extreme values of ASF-20C bootstrap trend distributions discussed in Sect. 5.2

		ASF-20C	Min	Max	ERA-20C	ERA5
DJF 1950/51–1979/80	Global all	-0.01*	-0.06	0.05	0.02*	0.04*
	Global land	-0.04*	-0.15	0.07	0.02*	0.07*
	Global sea	0.00*	-0.02	0.02	0.03*	0.03*
	Africa	-0.10	-0.25	0.06	-0.04*	0.08*
	Antarctic	0.10	-0.29	0.48	0.01*	0.27
	Arctic sea	-0.15	-0.61	0.28	0.11*	0.59
	Asia	-0.06*	-0.36	0.24	-0.08*	0.02*
	Australia	-0.04*	-0.39	0.29	-0.01*	-0.13*
	Europe	-0.08*	-0.79	0.64	0.07*	0.04*
	GAR	-0.08*	-0.81	0.69	0.26*	0.38*
	North America	-0.01*	-0.35	0.39	0.09*	-0.10*
	South America	0.00*	-0.11	0.09	0.05*	0.12*
JJA 1950–1979	Global all	0.05*	0.01	0.08	0.02*	0.01*
	Global land	0.03*	-0.05	0.10	-0.02*	0.06*
	Global sea	0.03*	0.01	0.05	0.03*	0.00*
	Africa	0.00*	-0.10	0.09	-0.11*	0.01*
	Antarctic	0.43	-0.21	1.18	0.19*	0.51
	Arctic sea	-0.01*	-0.06	0.05	-0.07*	-0.03*
	Asia	0.02*	-0.09	0.13	-0.04*	0.06*
	Australia	-0.05*	-0.31	0.26	-0.11*	-0.09*
	Europe	-0.07*	-0.31	0.23	-0.15*	-0.24
	GAR	-0.12	-0.63	0.58	-0.46	-0.41
	North America	-0.01*	-0.14	0.13	-0.06*	-0.06*
	South America	-0.00*	-0.20	0.17	0.03*	0.11*
DJF 1980/81–2009/10	Global all	0.19	0.13	0.25	0.23	0.17
	Global land	0.29	0.20	0.41	0.34	0.24
	Global sea	0.16	0.14	0.18	0.20	0.16
	Africa	0.32	0.13	0.50	0.40	0.33
	Antarctic	0.11	-0.37	0.43	0.21*	-0.02*
	Arctic sea	0.67	0.20	1.10	1.00	0.77
	Asia	0.42	0.11	0.74	0.37	0.24
	Australia	0.30	0.02	0.72	0.25*	0.04*
	Europe	0.42	-0.43	1.12	0.67	0.45*
	GAR	0.25	-0.59	1.20	0.63	0.46*
	North America	0.26	-0.03	0.57	0.33	0.33
	South America	0.16	0.06	0.25	0.15	0.17
JJA 1980–2010	Global all	0.16	0.12	0.19	0.23	0.14
	Global land	0.25	0.18	0.31	0.37	0.23
	Global sea	0.13	0.12	0.15	0.18	0.11
	Africa	0.30	0.20	0.38	0.47	0.28
	Antarctic	-0.08*	-0.93	0.58	0.42*	-0.10*
	Arctic sea	0.22	0.18	0.27	0.33	0.23
	Asia	0.29	0.20	0.39	0.42	0.31
	Australia	0.20	-0.08	0.52	0.25	0.05*
	Europe	0.44	0.11	0.68	0.70	0.56
	GAR	0.51	-0.04	1.03	0.78	0.59
	North America	0.29	0.16	0.43	0.32	0.25
	South America	0.20	0.05	0.35	0.20	0.22

in both seasons and the Arctic during winter in ASF-20C and ERA5. The only region with statistically significant trends in all data sets during the prior period is GAR in boreal summer. While values for the Antarctic in boreal summer are similar in ASF-20C and ERA5, the latter shows larger trends of DJF averages there. Over the Arctic the difference is in the same season almost 0.75 K per decade and even the signs of the trends are different. Again, this may indicate an issue with SIC forcing data in the hindcasts. In the period from 1980 onwards, trends are statistically significant almost everywhere. Moreover, each of the significant values is above +0.1 K per decade. Tab. 5.1 also confirms that ASF-20C ensemble mean and ERA5 trends are of very similar magnitude in the recent decades while ERA-20C trends tend to be larger. Also, these values confirm results of Simmons (2022), who states that surface temperature trends between 1979 and 2022 are largest over the Arctic and Europe. Results of this work show that between 1980 and 2010 during boreal winter 2 m temperature trends are higher in the Arctic than in Europe by about 0.2–0.3 K per decade. However, while boreal summer trends are similar to those of DJF in most regions, in the Arctic they are smaller by about 0.5 K per decade. The reason is that melting ice bounds the surface temperature to 0° C, which reduces trends in regions with sea ice in the summer hemisphere. In Europe, 2 m temperature trend in summer is between 0.44 and 0.70 K per decade in the three data sets, which are by far the largest values compared to other regions in this season and this period. Even higher values occur only in GAR, which indicates that summer warming is even more extreme in alpine regions than in the rest of the continent.

Though there are a lot of significant differences between trends of ASF-20C ensemble mean, ERA-20C and ERA5, the overall picture is very similar. ERA-20C and ASF-20C indicate warming trends at the beginning and cooling trends during the middle of the century in both seasons. ERA5 shows large patterns of both positive and negative trends in its first period, i.e. between 1950 and 1979. However, in the most recent decades, a clear global warming trend is visible over all continents and oceans in both re-forecast and reanalyses data. Over the majority of ocean regions trend differences are almost zero since SSTs and SICs are used as lower boundary forcing in the creation of ASF-20C. The only exception are polar oceans in the winter hemisphere. In boreal summer, trends in the Arctic are in general very small in all data sets because 2 m temperature is close to 0° due to melting ice at the surface (Simmons, 2022).

5.2 Bootstrap trends

Since it was just shown that the ASF-20C ensemble mean shows moderate trends even if ERA5 and ERA-20C indicate rather extreme values, the agreement between reanalyses and single member trends is investigated using the bootstrap resampling method (Efron, 1979) in this section. Explicitly, the linear trend of a time series, created by taking the temperature value of a randomly chosen ensemble member in each year for a given period, is calculated. After repeating this process 1000 times, the resulting histogram of temperature trends is compared to the "observed" values of ERA5 and ERA-20C.

Fig. 5.5 displays these histograms for trends of European DJF averages as well as for global, European and GAR JJA averages of land-only data from 1980 onwards. Dashed and dotted black lines and the red dashed line indicate trends of ERA-20C, ERA5, and the ASF-20C ensemble mean, respectively. Numbers in the legends show which percentile

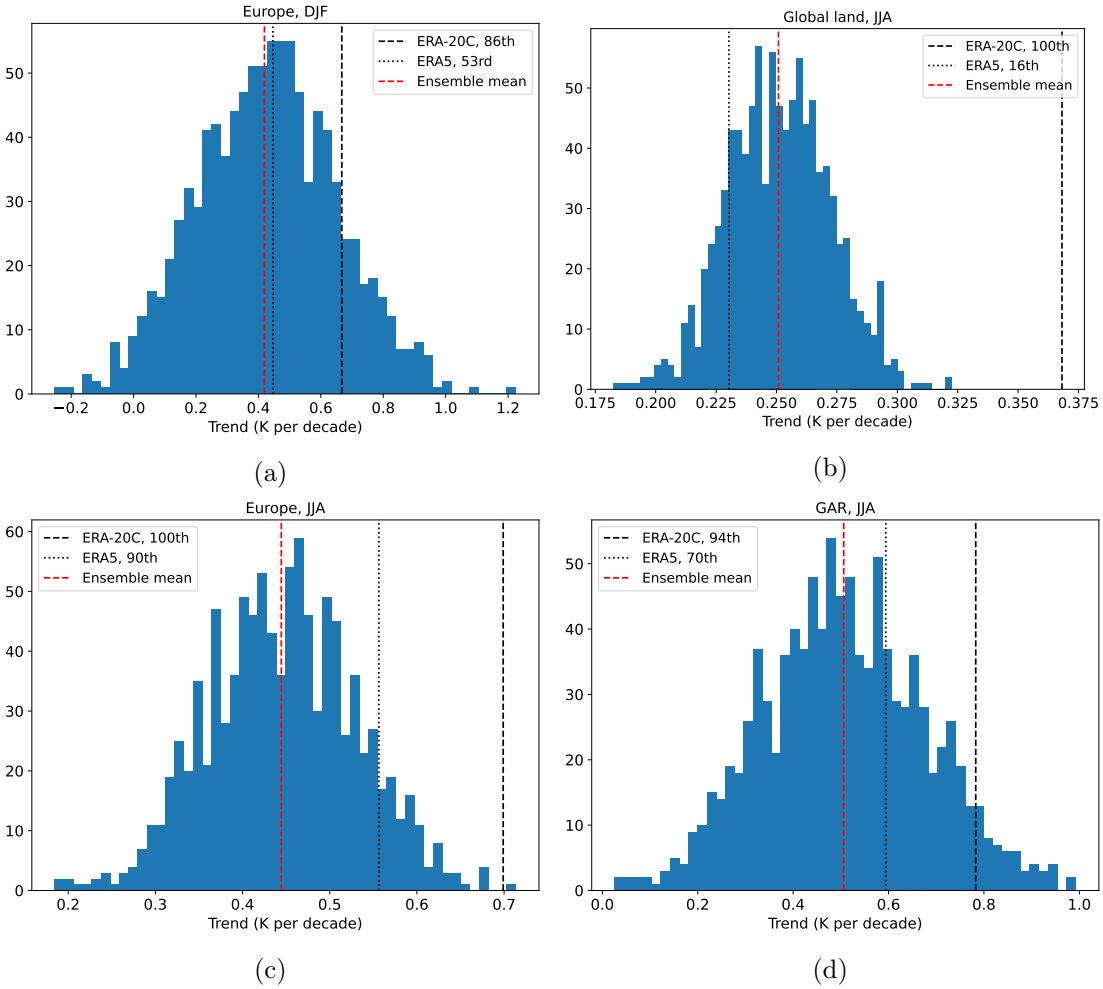


Figure 5.5: Histograms of 1000 temperature trends of time series generated by choosing a random ensemble member in each year. (a) Trends in Europe for DJF averages during the period 1980/81–2009/10. (b)–(d) Trends for JJA averages during 1980–2010 global, in Europe and in GAR, respectively. In each panel only land grid points are considered. Dashed red, dotted black and dashed black lines indicate trends of the ensemble mean, ERA5, and ERA-20C reanalyses, respectively. Numbers in the legend denote which percentile of the bootstrap trend distribution the reanalyses trends correspond to.

of the trend distribution ERA5 and ERA-20C trends correspond to. These histograms indicate that there is a wide range of possible temperature trends within the ensemble. But one can also see that the width of the trend distribution varies depending on which region and season is considered. In fact, standard deviations of the distributions of European DJF average and JJA average trends are approximately ten and four times higher than that of the global land JJA average trend distribution, respectively. This provides an important point why a large ensemble size is essential for a forecast model especially on seasonal time scales: Since every trend within this distribution is a possible outcome, given the initial conditions and natural variability, one single deterministic realization would not

be able to represent this broad range of possibilities.

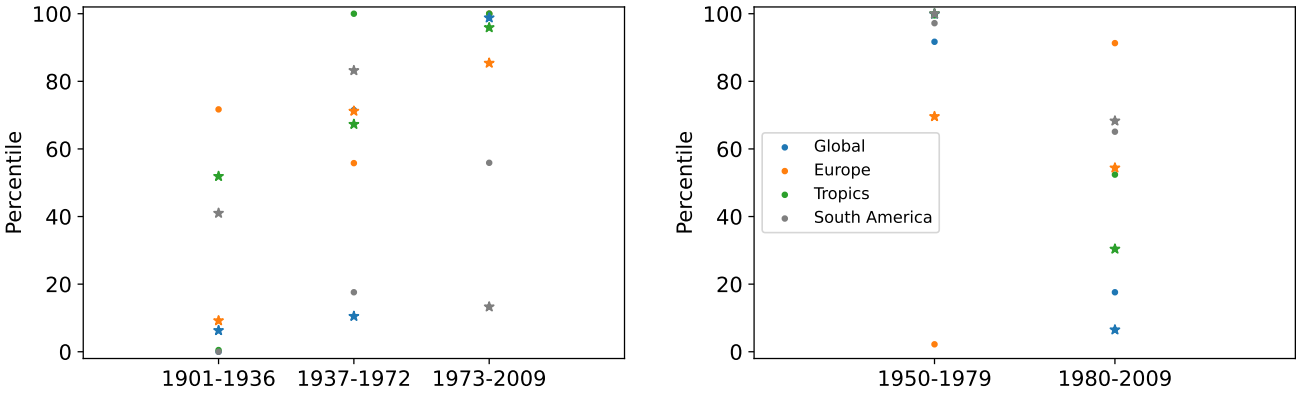


Figure 5.6: Percentile values of (a) ERA-20C and (b) ERA5 trends with respect to the ASF-20C bootstrap trends distribution considered at four different regions and during different periods. Dots and asterisks represent JJA and DJF averages, respectively. Color coding is identical in (a) and (b).

Though Fig. 5.5 implies that the ERA5 trend usually lies within the range of the ASF-20C distribution this in not true in general. In fact, both reanalyses trends are located at either extreme percentiles or even outside the minimum and maximum of the distribution in many cases, which is indicated by the dashed lines in Figs. 5.5b and 5.5c. However, as mentioned in the previous section, each ensemble member represents one possible realisation and so does each of these bootstrap time series. Thus, if a reanalysis trend is, e.g. located at a very high percentile of the distribution, this implies that in this specific region and period an extremely large of all the possible ensemble trends was observed. On the other hand, reanalyses trends outside of the distribution's range actually indicate an inconsistency between reanalyses and hindcasts.

Figs. 5.6a and 5.6b show the percentile values ERA-20C and ERA5 trends correspond to, respectively, when compared to the ASF-20C bootstrap trends distribution in four different regions. Namely, these regions are global, Europe, Tropics and South America. Again in each of them only land grid points were used. JJA and DJF averages are considered separately and they are represented by dots and asterisks, respectively. For ERA-20C periods 1901–1936, 1937–1972 and 1973–2009 are considered. For ERA5 the periods are 1950–1979 and 1980–2009. Fig. 5.6a shows that at the beginning of the 20^{th} century ASF-20C DJF trends in all regions are consistent with ERA-20C. However, during boreal summer, the ERA-20C trend is overestimated in the Tropics, in South America and also globally. This behaviour changes over the course of the century. During the middle of the century ASF-20C only underestimates the JJA trend in the Tropics. Underestimation of trends is also visible during JJA in the most recent period in all shown regions except for South America. During boreal winter, only the global land trend is underestimated by the re-forecast. Fig. 5.6b indicates that the changes of trend consistencies over time are contrary in ERA5. Though in the earlier period the hindcast model underestimates trends during DJF in all regions but Europe, hindcast and ERA5 trends are consistent in all regions and both seasons from 1980 onwards.

Columns *Min* and *Max* in Tab. 5.1 show minimum and maximum values, respectively, of the trend distributions calculated using the bootstrap method. Comparing these values for the two different periods shows that between 1950 and 1980 in both seasons and almost all regions the distribution contains both positive and negative values. On the other hand, distributions from 1980 onwards only contain positive values in the majority of regions. The only exceptions are the Antarctic and GAR in both seasons as well as Europe and North America during boreal winter and Australia in JJA. Minima and maxima of the distributions also indicate that in many regions the range between these values is higher in the respective winter hemisphere. For example, in Europe during the prior period the range is above 1.5 K per decade during winter but only around 0.5 K per decade during summer. Similarly, during the same period the range of DJF trends in South America is between −0.11 and +0.09 K per decade but almost twice as large for JJA averages.

As already mentioned previously, ASF-20C trends can be said to be consistent with ERA-20C and ERA5 trends, if the bootstrap distribution contains the reanalyses values. From Tab. 5.1 it can be concluded that this is true in many regions in both periods and both seasons. Interestingly, the bootstrap distributions do not always contain the reanalyses trends for globally averaged sea grid points. Since hindcasts use the same prescribed SSTs as ERA-20C, this inconsistency is to some extent surprising. However, it can be seen that the distribution's range is <0.05 K per decade in this region in all cases, which is very narrow, and moreover the JJA and DJF trends in the more recent period in ERA-20C are the only reanalysis trends that extend the distribution's maximum/minimum by more than 0.01 K per decade. Considering other regions, ERA5 trends are located outside of the distribution's range only in Africa, the Arctic and South America during boreal winter 1950/51–1979/80. Of these regions, by far the largest inconsistency occurs in the Arctic where the ERA5 trend is by more than 0.3 K per decade larger than the distribution's maximum. During the prior period, the only inconsistencies between ERA-20C and ASF-20C can be found in Africa and the Arctic during boreal summer, though in both cases ERA-20C trends are only 0.01 K per decade smaller than the distribution's minimum. Contrary to this, ERA-20C JJA average trend of the period 1980–2010 is in the majority of regions larger than any of the trends indicated by the hindcast ensemble. JJA trends of these two data sets during this period are only consistent in the Antarctic, Australia, GAR and the Americas.

Thus, it can be concluded that though the ensemble mean shows less extreme trends than reanalyses do, if single runs are considered, ASF-20C trends are still consistent with both ERA-20C and ERA5 in the majority of regions in both seasons and both periods. Only during boreal summer of the period 1980–2010 ERA-20C trends are higher than any trend of the bootstrap distribution in several regions in both hemispheres.

6 Changes of probabilities

6.1 Probability distributions

In the previous section, it was shown that the ensemble mean of ASF-20C hindcast data produces 2 m temperature trends that are similar to, though less extreme than, ERA-20C and ERA5 trends over the course of the 20^{th} century. However, trends of ASF-20C single runs are consistent with reanalyses values most of the time. If one considers each monthly average temperature as one realization of a given probability distribution function (pdf), i.e. a Gaussian distribution in the case of temperatures, these trends only represent shifts of mean values of the underlying distribution. But there are also other parameters of the pdf like variability and skewness that can change over time. Especially, the occurrence of extreme events is influenced by changes of the pdf's mean and variance (IPCC, 2013).

To investigate the probability distribution of a given region, its changes over time and the level of agreement with the reanalysis pdf, the influence of trends on the data can be removed in order to at least reduce possible effects coming from trend inconsistencies. This can be achieved by at first calculating anomalies of both data sets and afterwards calculating the linear trend of the hindcast ensemble mean and the reference reanalysis over the whole considered period. Then the ensemble mean trend is subtracted from each ensemble member and the reanalysis trend is added, which leads to equal mean values of both distributions. The final step is detrending the data in the specific period that is considered. Fig. 6.1 shows the ASF-20C ensemble histogram of JJA average temperatures between 1980 and 2010 of European land-only grid points in red. Anomalies are calculated with respect to the period 1950–1979. The green bars show the hindcast data after subtracting the ensemble mean trend of the period 1950–2010 from each ensemble member and adding the ERA5 trend of the same period. The blue bars result from subtracting the linear trend of the period 1980–2010 from the data represented by the green histogram. The curves represent kernel density estimators (kdes) assuming a normal distribution of the data. The left and right y-axis belong to curves and bars, respectively. One can see a small shift between the red and the green distribution coming from trend differences between ASF-20C ensemble mean and ERA5. However, the pdf changes a lot more after subtracting the linear trend of 1980–2010. As supposed, less extreme values occur because they are caused mainly by the inherent trend which widens the distribution.

Fig. 6.2 shows probability distributions in different regions of ERA5 (reddish) and ASF-20C (bluish) data for the periods 1950–1979 and 1980–2010. The ensemble mean trend of the period 1950–2010 was subtracted from each member and the ERA5 trend during the same period was added. In the end, from each data set and each period the trend of the respective data set during the respective period was subtracted as described above. Again, anomalies are calculated with respect to the prior period. The left y-axis corresponds to all kde curves, the inner right one to ASF-20C and the outermost right y-axis to ERA5 data. The first, second and third row show Europe, GAR and North America, respectively. JJA averages are shown in the left and DJF averages in the right

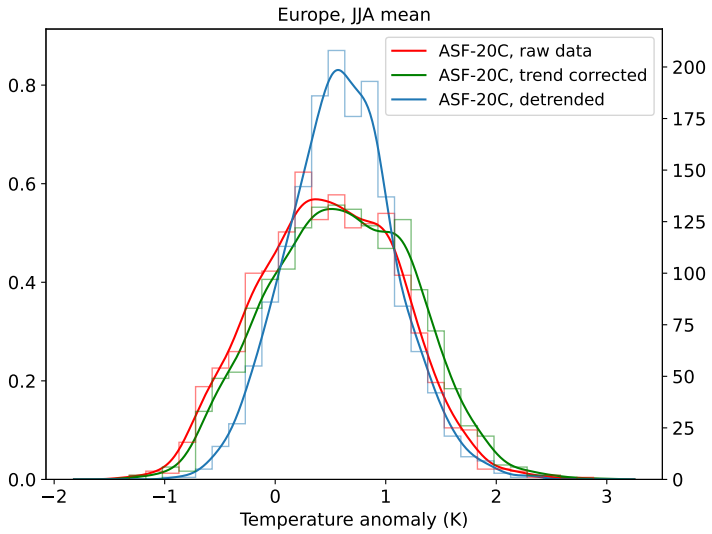


Figure 6.1: ASF-20C probability distribution of raw European land-only JJA averages of 1980–2010 (red). Green bars show the hindcast data after subtracting the ensemble mean trend of the period 1950–2010 from each ensemble member and adding the ERA5 trend of the same period. Blue bars result from subtracting its linear trend of the period 1980–2010 from the trend corrected data. The curves represent kernel density estimators assuming a normal distribution of the data. Temperature anomalies are calculated with respect to the period 1950–1979. The left and right y-axis belong to kdes and histograms, respectively.

column.

Apart from the ERA5 warming trend, which is represented by the shift of pdf mean values between two periods, Fig. 6.2 also indicates that the shape of the pdf can change over time quite a lot. How large these changes are, is dependent on both season and continent. For example, Figs. 6.2a and 6.2b indicate that while in boreal winter GAR distributions exhibit almost no changes between both periods, differences of ERA5 JJA average distributions are quite large in this region. If the whole European continent is considered, i.e. Figs. 6.2c and 6.2d, changes are not as extreme as in GAR summers. Also notable considering European and GAR JJA averages are the long tails of the hindcast distributions in the more recent period, displaying the possibility of very warm summer extremes, which is also indicated by the single outlier of +3.49 K in the year 2003 in ERA5 data in GAR, and that ASF-20C is very well capable of covering them. Moreover, also the ability of ASF-20C to reproduce changes of the pdf is strongly dependent on the considered region. While in GAR for JJA averages differences between the shapes of re-forecast and reanalysis distributions are very large, the hindcast model reproduces changes in North American summers very well. Their standard deviations, i.e. the widths of ASF-20C and ERA5 distributions, increase by about 32% and 25%, respectively. But still, standard deviations differ only by about 1% in the first and slightly above 4% in the second period. In general extreme values are reproduced pretty good by the hindcast but skewness can be slightly different. Though also in other regions differences between kdes

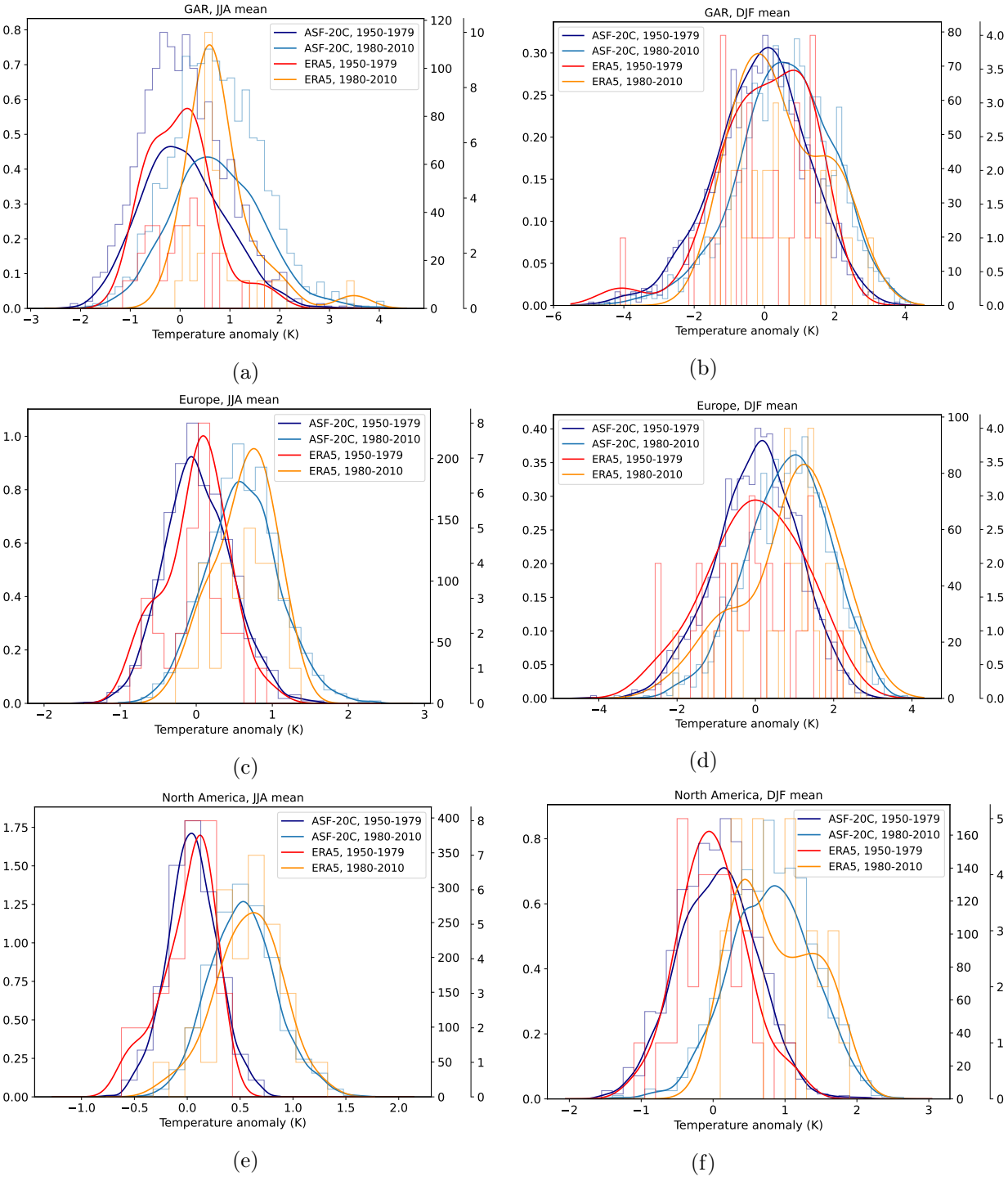


Figure 6.2: Histograms and approximated probability distributions (kdes) of different continents showing land-only temperature anomalies of ERA5 (reddish) and ASF-20C (bluish). The period 1950–1979 of the respective data set is used as climatology and data of each period is detrended as described in the text. (a) and (b) Europe, (c) and (d) GAR, (e) and (f) North America. JJA averages and DJF averages are displayed in the left and right column, respectively.

may look quite large, this is mostly just a consequence of a different coverage of extreme events. The same is valid if hindcast and ERA-20C output from the beginning of the 20$^{\text{th}}$ century are considered. Thus one can conclude that ASF-20C reproduces the shape of the ERA5 distribution very well from 1950 onwards and of ERA-20C distributions throughout the whole period, except for some inaccuracies concerning the coverage of extreme values and the skewness of the distribution.

Changes of the distribution's standard deviation over time were already visible in Fig. 6.2 and shortly discussed previously. One can see that a wider histogram is equivalent to a raised probability of both hot and cold extreme temperature anomalies to occur (see also Fig. 1.8 of IPCC, 2013). Fig. 6.3a shows a map of standard deviation ratios of ASF-20C JJA averages of 1980–2010 with respect to 1950–1979 to investigate these changes on a global scale. Data within each period is treated in the same way as for previously shown histograms. Though changes are not very large over continents, in oceanic regions ratios are up to and above 1.5 in the North Atlantic and tropical Indian Ocean. During summer, higher variances may indicate a shallower mixed layer leading to reduced heat capacity. In fact, in the second period 5% of all ensemble members show temperature values in the tropical Indian Ocean that are above the maximum anomaly of the first period even if the change of the distribution's mean between both periods is already considered. Similarly there are also a few ensemble members that give temperatures below the minimum of the prior period again taking into account the warming trend between both periods. On the other hand, standard deviation decreases in large patterns in oceans of the Southern Hemisphere. Though this is probably because during the prior period observational data was very sparse there leading to larger uncertainties and thus a broader ensemble pdf. The improvement of the observational coverage may have increased confidence and narrowed the distribution.

Fig. 6.3b shows the same as Fig. 6.3a but without trend correction and detrending as described previously. It can be seen that ratios are more extreme and larger regions are affected. The latter is especially true for continents, e.g. central Africa and northern South America. The reason for these changes is that in the more recent period trends are larger than in the prior period in many regions and therefore distributions widen. Basically, the same effect that was already visible in Fig. 6.1. However, it is notable that the majority of patterns can still be seen in Fig. 6.3a after detrending.

JJA averages of ERA5 are displayed in Fig. 6.3c and indicate very similar patterns over oceans, except for the Southern Ocean. On the other hand, differences between ASF-20C and ERA5 data are more pronounced over continents. The largest differences appear over central Africa, Brazil and eastern Asia. In each of these regions ratios decrease to values below 0.5 in ERA5. However, Simmons et al. (2021) mention temporary lacks of observation data in each of these regions between 1950 and 1979, which presumably leads to inconsistent temperature values and thus increased variance in the prior period. Considering DJF averages, which are shown for ASF-20C in Fig. 6.3d, agreement between re-forecasts and ERA5 over oceans is again very high. This time patterns in the Southern Ocean are very similar but there are small deviations in the North Pacific Ocean. The highest ratios (around 1.5) over oceans occur in the equatorial Pacific and the Southern Ocean. Over continents ratios are again close to 1 in hindcast data but smaller in many parts of Asia, Africa and South America in ERA5.

In general, reanalysis shows more extreme ratios almost everywhere. However, this may be because the ensemble distribution is a lot smoother and thus also has a higher

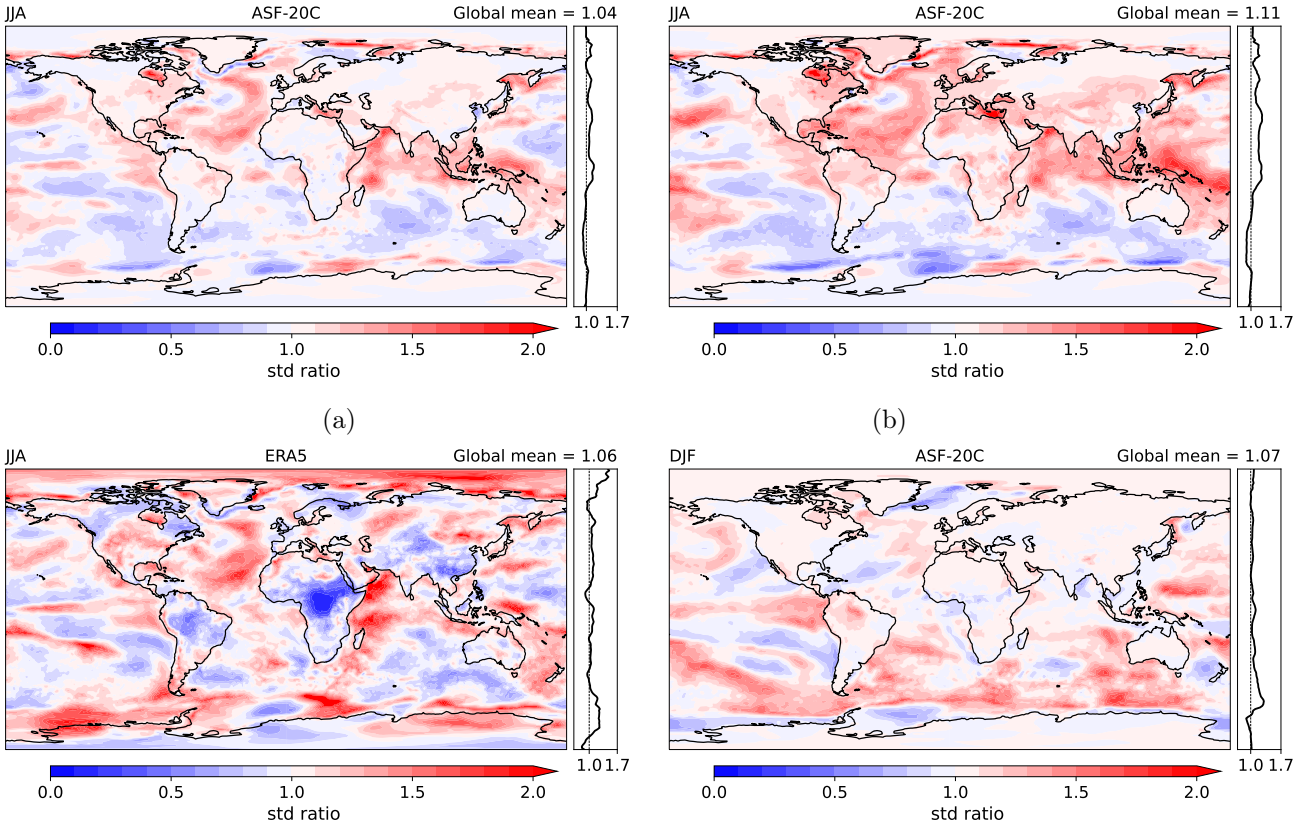


Figure 6.3: (a) Ratio of standard deviations of ASF-20C JJA average distributions of periods 1980–2010 versus 1950–1979. Data within each period is treated in the same way as for histograms in Fig. 6.2. (b) Same as (a) but using raw ASF-20C data without trend correction and detrending. (c) and (d) Same as (a) but for JJA averages of ERA5 and DJF averages of ASF-20C, respectively.

resolution compared to ERA5 which provides only 30 values per period. The small sample size of the reanalysis distribution makes measures like standard deviation very dependent on the existence of single extreme values. Contrary to this, a distribution as large as 30 years of ASF-20C ensemble output with a total of more than 1500 members is a lot more robust against the existence of outliers.

6.2 Percentiles

Another possibility to quantify changes in the occurrence of extreme values are percentiles. More specifically how the number of monthly or seasonal average temperatures above or below a certain threshold changes over time. In the following, these changes are going to be investigated for the 5^{th} and 95^{th} percentile. Thus, Fig. 6.4a displays the 95^{th} percentile of the ASF-20C distribution of JJA averages for the period 1950–1979.

Figs. 6.4b and 6.4c show the proportion of JJA average temperatures between 1980 and 2010 that are above the 95^{th} percentile of the 1950–1979 period for ERA5 and ASF-20C,

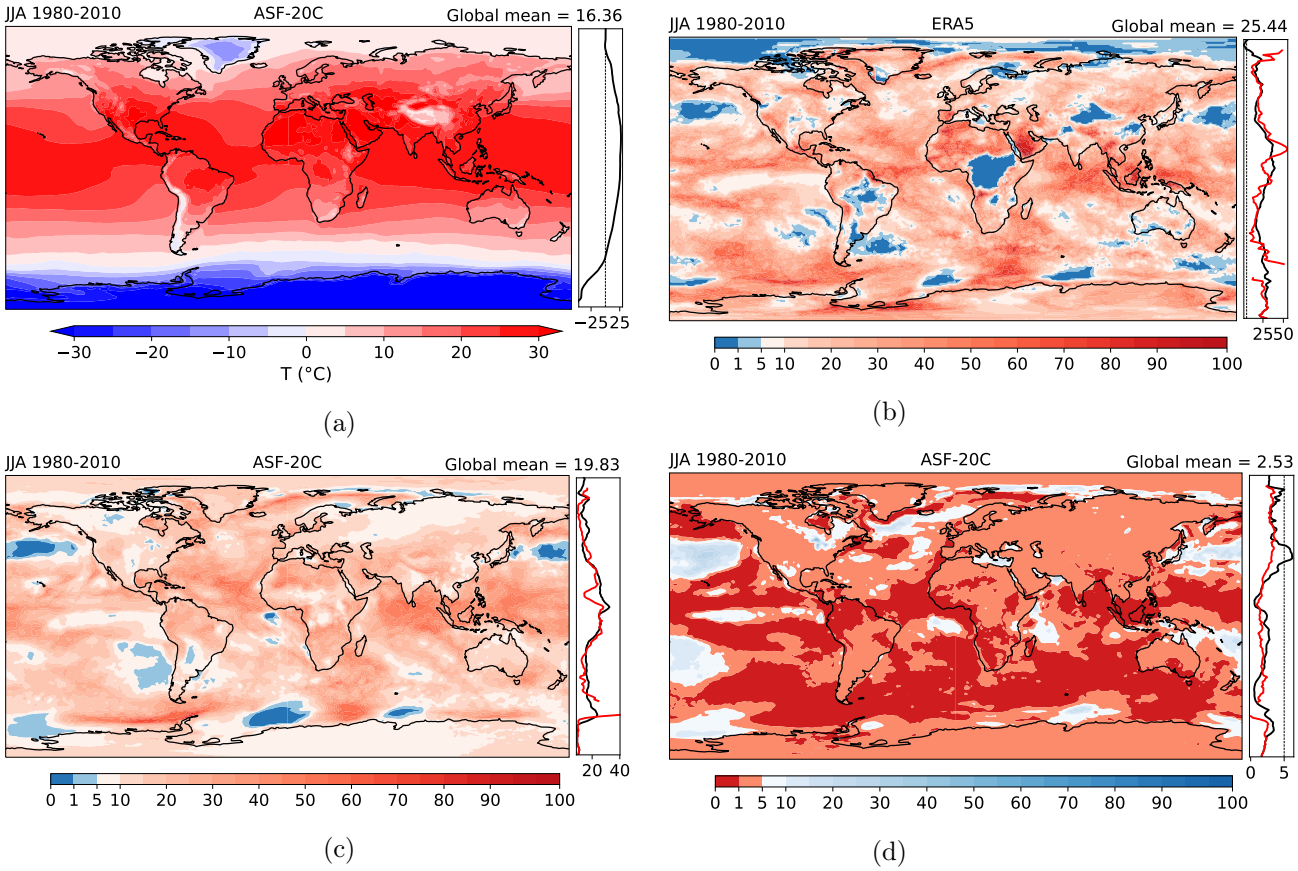


Figure 6.4: (a) 95$^{\text{th}}$ percentile of the ASF-20C distribution of JJA averages for the period 1950–1979. Proportion of JJA averages during 1980–2010 of (b) ERA5 and (c) ASF-20C that are above the 95$^{\text{th}}$ percentile and of (d) ASF-20C below the 5$^{\text{th}}$ percentile of the respective data set for the period 1950–1979. Red lines in the vertical boxes attached to each map in (b)–(d) represent zonal averages of land-only grid points.

respectively, and the proportion of ASF-20C values below the 5$^{\text{th}}$ percentile is shown in Fig. 6.4d. Percentile values are calculated from each respective data set. Red lines in the boxes attached to each map in Figs. 6.4b–6.4d represent zonal averages of land-only grid points. If there would have been no changes between both periods, one would expect a value of 5% everywhere in each of these three panels. However, all maps indicate that proportions strongly differ from this theoretical value in many regions around the globe.

Dark and light blue patterns in Figs. 6.4b and 6.4c indicate regions where $\leqslant 5\%$ of values are above this grid point specific threshold. Agreement is very high for these patterns over oceans, for example in the northern hemispheric mid-latitudes of the Pacific Ocean and near the South American Pacific coast. Agreement is also high around the Antarctic, though the reason there is most probably the already known issue with sea ice concentration data. However, there are also a number of blue patterns over continental regions in ERA5 data that are not present at all in the ensemble data. For example over central Africa, parts of Asia and Brazil, though this is not considered as a performance

problem of the hindcast model but rather as a resolution problem of the reanalysis mixed with known data inconsistencies in some regions during the first period. Since ERA5 only provides 30 values in each period, extreme percentile values like those considered in Fig. 6.4 are not clearly defined and therefore interpolation is needed. On the other hand, 5% of the hindcast data in a 30-year period correspond to at least 76 data points and therefore even the most extreme percentiles are resolved very well. This displays a large advantage of such a large ensemble model over a single deterministic run. It has to be noted however, that ERA5 also relies on a 10-member ensemble but the main goal of this ensemble is to represent the data uncertainty of the resulting deterministic high-resolution run. Thus, the ERA5 ensemble is not further considered in this work. Contrary to this, in the hindcast each ensemble member represents a both plausible and possible evolution of Earth's atmosphere and can therefore be interpreted as one realization in the pdf.

Reddish colours in Figs. 6.4b and 6.4c point out where $> 5\%$ of temperature values in the most recent period exceed the 95^{th} percentile of 1950–1979. In both panels this applies to the vast majority of grid points, regardless if they are over oceans or continents. Patterns that indicate the highest values are very similar in both reanalysis and re-forecast, e.g. in large parts of equatorial oceans, over the Sahara and in western USA. However, proportions indicated by ERA5 are much more extreme in many regions, e.g. Arabian Peninsula and southeastern Asia besides those previously mentioned, than ASF-20C values. Again this is a consequence of the sample size, since the hindcast distribution is a lot smoother and therefore does not exhibit such extreme values. This also explains why zonal and global averages are higher in Fig. 6.4b, although ERA5 also shows a lot more very small values than ASF-20C does. Land-only zonal averages are very similar to zonal means of all grid points. Apart from regions like the Arctic in ERA5 and the Southern Ocean in both data sets, where most probably SIC data inconsistencies lead to extreme values, the only large differences appear in northern hemispheric Tropics. In ERA5 very high values over the Sahara and the Arabian Peninsula raise zonal mean values, while in ASF-20C southern of the Sahara values are smaller than in ocean regions of the same latitude.

Fig. 6.4d shows the proportion of values below the 5^{th} percentile for the same periods. Colour coding is reversed compared to the upper two panels. This time in most regions less than 5% of all values are smaller than the threshold, which indicates less cold summers/winters in the northern/southern hemisphere and thus represents the warming trend in the most recent decades. However, zonal and global averages indicate that changes between both periods are larger for warm anomalies. In ASF-20C the number of values above the 95^{th} percentile quadrupled towards the most recent period while proportions below the 5^{th} percentile halved. This time large deviations between land-only and total zonal averages only occur in northern hemispheric mid-latitudes because of the large pattern in the North Pacific Ocean, which indicates that there the number of values below the 5^{th} percentile drastically increased. However, no such patterns can be found over continents in these latitudes except for a small region above the Great Lakes in North America.

Reconsidering the regions in Fig. 6.3 which show higher standard deviations of JJA averages between 1980 and 2010 than in 1950–1979, more precisely the North Atlantic, parts of the equatorial Pacific Ocean, the western Indian Ocean along the African coast northwards of Madagascar, and the eastern Mediterranean Sea, one can see that in each of these regions ASF-20C data shows higher proportions both above the 95^{th} and also below the 5^{th} percentile. Therefore these maps do not only represent the recent warming trend

but also the fact that variances of probability distributions increase in some regions leading to an increased number of both warm and cold extreme events. One reason, why most of these regions are over oceans may be that there distributions are initially rather narrow due to prescribed SSTs and the majority of variability comes from temperature trends. Contrary to this, over continents uncertainties are larger leading to wider distributions and thus effects of trends are damped.

7 Conclusion

The aim of this work was to investigate the performance of ASF-20C, a seasonal re-forecast data set with 51 ensemble members that covers the period 1901–2010. Two reanalyses were used as reference: ERA-20C, an atmosphere-only reanalysis, which was used to initialize ASF-20C and covers the same period and ERA5, a state-of-the-art reanalysis product. To quantify the performance of the hindcast data, monthly mean values of the three data sets for November–February and May–August as well as seasonal averages are compared.

It was shown that the ASF-20C ensemble mean has a small negative bias of less than -0.5 K against ERA5 over oceans, except for the Southern Ocean where bias is of the same magnitude but positive. In most continents bias of the hindcast is less extreme than -1 K. However along the Pacific coast of North America as well as over the Sahara negative deviations in boreal winter can reach a few K. On the other hand, over northern hemispheric continents a warm bias of several K emerges with lead time, especially in the mid-latitudes during boreal summer months and towards the northern continental regions during boreal winter. Similarly to the latter point, a strong cold bias occurs in polar regions of the winter hemisphere. However, if anomalies are considered, agreement between ASF-20C and the reanalyses is rather high. Time series of land-only averages of continents like South America and Africa show correlation coefficients of ASF-20C ensemble mean and ERA5 of $+0.8$ and above, which exceeds in some cases even correlation between the two reanalysis products. Results of this work also showed that in regions like GAR, where ENSO teleconnections do not have much influence or ensemble spread is enhanced due to other reasons like higher internal variability of the atmosphere, correlations are significantly reduced but still above $+0.5$. Moreover, even in these regions single ensemble members are able to cover the most extreme anomalies that occur throughout the 20^{th} century. Examples are extremely cold and warm JJA averages in GAR in 1913 and 2003, respectively. Over almost all ocean regions, correlation between re-forecasts and ERA5 is close to unity and RMS values are almost zero due to the fact that ASF-20C uses prescribed SSTs as lower boundary condition.

Maps of correlation in 30-year periods between ASF-20C and ERA5 indicate that over oceans correlation coefficients are close to one all the time. Contrary, over continents in both seasons during 1950–1979 the only regions with correlations around $+0.6$ in all lead months are those most affected by ENSO, namely northern South America, equatorial Africa and partly also Australia. However, towards the more recent period 1980–2010 large improvements over northern hemispheric continents and the Southern Ocean are visible. This is at least partly an effect of improved observational coverage. Another factor contributing to this improvement is higher predictability of ENSO and NAO in recent decades. It was shown that correlation of seasonal averages of ASF-20C and ERA-20C exhibits multi-decadal variations over the course of the 20^{th} century. The smallest correlation coefficients of DJF averages in South America and of JJA averages in North America and Australia occur between the 1930s and 1950s when ENSO activity was at a minimum. The same is valid for JJA averages in GAR and DJF averages in North America

and Europe between the 1950s and 1970s when NAO forecast skill was lower than during the rest of the century. The third reason for enhanced correlation in most recent decades is the high agreement on temperature trends in all three data sets. It was shown that not detrending temperature data before calculating correlation in 30-year periods leads to correlation coefficients that are much higher than when detrending is done. Explicitly, correlation of JJA averages of ASF-20C and ERA-20C in the period 1980–2010 in GAR is above +0.5 without detrending but close to 0 when the ensemble mean trend is subtracted. In other continents improvement is around 0.1 when trends are not considered. Also, in general correlations are slightly higher over continents of the summer hemisphere.

Two very important properties of an ensemble forecast system are reliability and resolution, i.e. how good predicted probabilities and measured frequencies of occurrence agree and how good the forecast system can discriminate events with high from those with low probability of occurrence, respectively. While reliability is very good in almost all regions, except for parts of Africa and the Southern Ocean between 1950–1980, resolution happens to be the best over oceans and again in continental regions that are affected most by ENSO. Hence, Brier Skill Score (BSS) calculated with respect to climatology indicates that seasonal forecasts are skillful between 1950 and 1980 almost only over oceans, except for the Southern Ocean, as well as over small regions over continents. However, again improvements towards the more recent periods can be found, especially in ENSO affected regions and where observational coverage improved significantly. Similarly to correlations, BSS also tends to be higher in the summer hemisphere.

Another way to indicate the amount of useful information that can be extracted from an ensemble forecast system, is the signal-to-noise ratio (SNR), which compares the variation of the ensemble mean to the variance of all ensemble members within a given period. Hence, this measure is considered to quantify potential predictability and $SNR = 1$ denotes the best possible value. Results indicated that in most ocean regions SNR is above 0.8 in both seasons throughout the 20^{th} century, exceptions are extratropical (ET) regions of northern hemispheric oceans in DJF and South Atlantic ET and Indian Ocean ET in JJA. In general, SNR over continents is much smaller and only reaches values above 0.6 in the Tropics in both seasons, in Africa in JJA and in South America in DJF. Many continental and oceanic regions have in common that SNR exhibits large multi-decadal variations just as correlation does. Again, variations are not always due to an improvement of observational coverage over the course of the 20^{th} century but may base on changing ENSO and NAO predictability as well as on the agreement on temperature trends. Thus, it is important to consider more than just the most recent decades, where SNR is generally at a relatively high level, for skill investigations of forecast systems. Very high correlation coefficients between the standard deviation of the Niño3.4 index and SNR time series reaching values above +0.90 in both seasons over different continents and oceans, e.g. South America in DJF and JJA, Africa in DJF and JJA, North Atlantic Ocean in DJF and Tropical Indian Ocean in JJA, confirm that these multi-decadal forecast skill variations base on varying ENSO activity. Hence, ENSO can be considered to be an important contributor to global 2 m temperature forecast skill in both oceans and continents. Contrary to this, correlation with the NAO index is high only in the northern hemisphere during boreal summer. In North America, Europe and North Atlantic ET approximate correlation coefficients are +0.68, +0.75 and +0.78, respectively. Therefore, NAO only seems to have a large impact on 2 m temperature forecast skill in these regions. Grid point wise calculation of SNR also showed that the areas above the Gulf Stream and Kuroshio suffer from reduced SNR

compared to the rest of the respective oceans in boreal winter when the storm tracks are most active. Therefore, it seems that ASF-20C has difficulties to exactly describe surface fluxes, dependent on weather conditions, in these regions although SSTs are prescribed.

The level of agreement on 2 m temperature trends of the three data sets is also investigated in this work. While over oceans trends do not vary by more than ± 0.3 K per decade, deviations over continents can be by far larger, though these differences are mostly not statistically significant. Also they are dependent on season, time period and the exact region. Deviations in trends in the data sets may have different reasons. On the one hand, when calculating grid point wise trends, only the trend of the ensemble mean is considered which smoothes out possible extreme values within the ensemble. Therefore ASF-20C trends showed in maps in this work are in general less extreme than reanalyses trends. Moreover, in some cases data inconsistencies in reanalyses like ERA5 warm bias in Africa and Brazil in the 1950s and 1960s distort trends and thus lead to differences to ASF-20C trends. Thirdly, there are also uncertainties in the re-forecasts, sometimes evolving with ongoing lead time. But still all data sets agree on the global picture of trends in both seasons and their changes throughout the 20^{th} century: while at the beginning of the century a small warming trend is found in most oceans and almost all continents, slight cooling dominates the period 1937–1972 in ASF-20C and ERA-20C output, especially in the North Atlantic Ocean, North America and parts of Asia. In the most recent 3–4 decades a strong warming trend is visible and statistically significant all around the globe in each data set. Though the amplitude varies, it is larger than 0.1 K per decade in each continent. The most extreme warming from 1980 onwards occurs during summer in Europe, when trends are 0.44, 0.70 and 0.56 K per decade in ASF-20C, ERA-20C and ERA5, respectively. Trends are even slightly larger in GAR, indicating that summer warming in alpine regions is even stronger. Moreover, results in this work confirm the statement of Simmons (2022) that surface temperature trends in the period 1979–2022 are largest over the Arctic and Europe.

Since ensemble mean trends smooth out possible outliers, also the bootstrap method is used to calculate 1000 trends in different regions from time series with randomly chosen ensemble members in each year. The width of the resulting distributions is dependent on the considered region and period. For example it is much larger in Europe than for global averages and also tends to be larger during respective winter months. In general, ASF-20C trends are consistent with reanalyses values, i.e. the range of the distribution contains values of ERA5 and ERA-20C. However, this is not true for ERA5 trends during boreal winter for the period 1950/51–1979/80 in Africa, South America and the Arctic. In the latter region, the reanalysis trend is more than 0.3 K per decade larger than the highest value of the bootstrap distribution. Concerning ERA-20C, JJA trends for the period 1980–2010 are only consistent in the Antarctic, Australia, GAR and the Americas but lie above the distribution's maximum everywhere else.

Towards the end of this work, probability distributions of 2 m temperatures during 30-year periods are investigated. It was shown that hindcast ensemble pdfs are very similar to those of the reanalyses after a trend adjustment and detrending process, except for small differences at the distribution's tails. Agreement is also very high throughout the whole period and not only in the most recent decades. Since climate change is known to not only change the mean but also the width of such distributions, changes of pdf shapes are also examined grid point wise. Results showed that their standard deviation changed by a factor of 1.5 between 1950–1979 and 1980–2009/10 in the North Atlantic and the

equatorial Indian Ocean in JJA and in several ocean regions in the southern hemisphere in DJF (even when decadal trends are removed). Agreement over oceans is very high between ERA5 and ASF-20C but over continents results are different. Though the tendency of the change, i.e. if standard deviation increases or decreases, is very often similar in both datasets, the amplitude of the change sometimes differs strongly. The reason is that the ASF-20C distribution size is more than 50 times larger in each period leading to much smoother distributions, which are not that affected by single outliers.

The changes of extreme percentiles between 1950–1979 and 1980–2009/10 were also addressed. Results indicate that almost globally the ratio of values during the latter period that is higher than the 95^{th} percentile of the former period is above the theoretical value of 5%. Values around 40% are found for JJA averages of the hindcast in tropical oceans. Similarly high values occur in the North Atlantic and over northern parts of Africa and South America. Qualitative agreement with ERA5 is very high over both continents and oceans but in the reanalysis values are a lot more extreme reaching and even extending 70%. As before, this is considered to be a weakness of deterministic runs, since such extreme percentiles are not resolved properly due to the small distribution size, which makes interpolation necessary. Moreover, regions that showed an increased standard deviation towards the more recent period also display more values above and below the 95^{th} and 5^{th} percentile, respectively. This confirms that higher standard deviations of the pdf indeed increase the probability of both hot and cold extreme events.

To sum it up, the results of this work showed that the centennial seasonal re-forecast data set ASF-20C performs reasonably well in terms of biases and trends compared to two different reanalyses. Even more, when probability distributions or extreme events are considered. The unprecedented size of 51 ensemble members for a 110-year period can bring large advantages in terms of resolution and robustness compared to single deterministic runs if the data is properly corrected for e.g. biases. Since both ERA-20C and ASF-20C are atmosphere only models, a follow up study may concern the performance of their counterparts CERA-20C and CSF-20C where atmosphere and ocean models are coupled. Especially over oceans, differences between ASF-20C and CSF-20C are supposed to be present. Similarly, comparisons to the output of other seasonal forecast data sets, like operational ones, could be done. Also considerations of the performance of single ASF-20C runs instead of just the ensemble mean as well as of smaller ensemble sizes compared to the full size may be done in the future. Moreover, other quantities like precipitation could also be in the focus of future studies. Another aspect that may be investigated is, if the use of machine learning for calibration of seasonal forecasts improves the performance of (re-)forecasts, especially when long forecast ranges are considered.

Bibliography

Barnston, A. (2015). Why are there so many ENSO indexes, instead of just one? `https://www.climate.gov/news-features/blogs/enso/why-are-there-so-many-enso-indexes-instead-just-one`. Accessed: 2022-04-04.

Barnston, A., Tippett, M., L'Heureux, M., Li, S., and DeWitt, D. (2012). Skill of Real-Time Seasonal ENSO Model Predictions During 2002-11: Is Our Capability Increasing? *Bulletin of the American Meteorological Society*, 93:48–.

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Radu, R., Schepers, D., Soci, C., Villaume, S., Bidlot, J.-R., Haimberger, L., Woollen, J., Buontempo, C., and Thépaut, J.-N. (2021). The ERA5 global reanalysis: Preliminary extension to 1950. *Quarterly Journal of the Royal Meteorological Society*, 147(741):4186–4227.

Bjerknes, J. (1969). ATMOSPHERIC TELECONNECTIONS FROM THE EQUATORIAL PACIFIC. *Monthly Weather Review*, 97(3):163 – 172.

Bosilovich, M. G., Robertson, F. R., and Stackhouse, P. W. (2020). El Niño-Related Tropical Land Surface Water and Energy Response in MERRA-2. *Journal of Climate*, 33(3):1155 – 1176.

Brier, G. W. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1 – 3.

Challinor, A., Slingo, J., Wheeler, T., and Doblas-Reyes, F. (2005). Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3):498–512.

Clogg, C. C., Petkova, E., and Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, 100(5):1261–1293.

Cram, T., Compo, G., Yin, X., Allan, R., Mccoll, C., Vose, R., Whitaker, J., Matsui, N., Ashcroft, L., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, B., Groisman, P., Hersbach, H., and Worley, S. (2015). The International Surface Pressure Databank version 2. *Geoscience Data Journal*, 2:31–46.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011).

The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W. (2012). The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sensing of Environment*, 116:140–158. Advanced Along Track Scanning Radiometer(AATSR) Special Issue.

Douville, H. (2009). Relative contribution of soil moisture and snow mass to seasonal climate predictability: a Pilot Study. *Climate Dynamics*, 34:797–818.

Eastwood, S., Lavergne, T., and Tonboe, R. (2014). Algorithm Theoretical Basis Document for the OSI SAF Global Reprocessed Sea Ice Concentration Product, version 1.1.

ECMWF (2013). IFS documentation. `https://www.ecmwf.int/en/publications/ifs-documentation`. Accessed: 2022-03-12.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Fichefet, T. and Maqueda, M. A. M. (1997). Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. *Journal of Geophysical Research: Oceans*, 102(C6):12609–12646.

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). GLOBAL SURFACE TEMPERATURE CHANGE. *Reviews of Geophysics*, 48(4).

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hersbach, H., Peubey, C., Simmons, A., Berrisford, P., Poli, P., and Dee, D. (2015). ERA-20CM: a twentieth-century atmospheric model ensemble. *Quarterly Journal of the Royal Meteorological Society*, 141(691):2350–2375.

Hirahara, S. and Hersbach, H. (2016). Sea Surface Temperature and Sea Ice Concentration for ERA 5. In *ERA Report Series 26*.

Hsu, W.-R. and Murphy, A. H. (1986). The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2(3):285–293.

Huang, P. and Huang, R. (2009). Delayed atmospheric temperature response to ENSO SST: Role of high SST and the western Pacific. *Advances in Atmospheric Sciences*, 26:343–351.

Hurrell, J. (1995). Decadal trends in the north atlantic oscillation. *Science (New York, N.Y.)*, 269:676–9.

Hurrell, J. and Van Loon, H. (1997). Decadal Variations in Climate Associated with the North Atlantic Oscillation. *Climatic Change*, 36:301–326.

Ineson, S. and Scaife, A. A. (2009). The role of the stratosphere in the European climate response to El Niño. *Nature Geoscience*, 2(1):32–36.

IPCC (2013). *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press.

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., and Raynaud, L. (2010). Ensemble of data assimilations at ECMWF. Technical Memorandum 636, ECMWF.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., and Monge-Sanz, B. M. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117.

Kiladis, G. N. and Diaz, H. F. (1989). Global Climatic Anomalies Associated with Extremes in the Southern Oscillation. *Journal of Climate*, 2(9):1069 – 1090.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K. (2015). The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93(1):5–48.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach, H., Kosaka, Y., Martin, M., Poli, P., Rayner, N., Rustemeier, E., and Schepers, D. (2018). CERA-20C: A Coupled Reanalysis of the Twentieth Century. *Journal of Advances in Modeling Earth Systems*, 10(5):1172–1195.

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., and Zyss, D. (2019). Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres*, 124(12):6307–6326.

Lin, H., Merryfield, W. J., Muncaster, R., Smith, G. C., Markovic, M., Dupont, F., Roy, F., Lemieux, J.-F., Dirkson, A., Kharin, V. V., Lee, W.-S., Charron, M., and Erfani, A. (2020). The Canadian Seasonal to Interannual Prediction System Version 2 (CanSIPSv2). *Weather and Forecasting*, 35(4):1317 – 1343.

Loeb, N. G., Mayer, M., Kato, S., Fasullo, J. T., Zuo, H., Senan, R., Lyman, J. M., Johnson, G. C., and Balmaseda, M. (2022). Evaluating Twenty-Year Trends in Earth's Energy Flows From Observations and Reanalyses. *Journal of Geophysical Research: Atmospheres*, 127(12):18.

Luterbacher, J., Werner, J. P., Smerdon, J. E., Fernández-Donado, L., González-Rouco, F. J., Barriopedro, D., Ljungqvist, F. C., Büntgen, U., Zorita, E., Wagner, S., Esper, J., McCarroll, D., Toreti, A., Frank, D., Jungclaus, J. H., Barriendos, M., Bertolin, C., Bothe, O., Brázdil, R., Camuffo, D., Dobrovolný, P., Gagen, M., García-Bustamante, E., Ge, Q., Gómez-Navarro, J. J., Guiot, J., Hao, Z., Hegerl, G. C., Holmgren, K., Klimenko, V. V., Martín-Chivelet, J., Pfister, C., Roberts, N., Schindler, A., Schurer, A., Solomina, O., von Gunten, L., Wahl, E., Wanner, H., Wetter, O., Xoplaki, E., Yuan, N., Zanchettin, D., Zhang, H., and Zerefos, C. (2016). European summer temperatures since Roman times. *Environmental Research Letters*, 11(2):024001.

Madec, G. and the NEMO team (2016). NEMO ocean engine. `https://www.nemo-ocean.eu/wp-content/uploads/NEMO_book.pdf`. Accessed: 2022-08-09.

Mason, S. and Stephenson, D. (2008). How Do We Know Whether Seasonal Climate Forecasts are Any Good? In *Seasonal Climate: Forecasting and Managing Risk*, pages 259–289. Springer Dordrecht.

Mason, S. J. (2004). On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores. *Monthly Weather Review*, 132(7):1891 – 1895.

Merchant, C. J., Embury, O., Roberts-Jones, J., Fiedler, E., Bulgin, C. E., Corlett, G. K., Good, S., McLaren, A., Rayner, N., Morak-Bozzo, S., and Donlon, C. (2014). Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geoscience Data Journal*, 1(2):179–191.

Molteni, F., Stockdale, T., Alonso-Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F. (2011). The new ECMWF seasonal forecast system (System 4). Technical Memorandum 656, ECMWF.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R. (2021). An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres*, 126(3):e2019JD032361.

Morse, A., Doblas-Reyes, F., Hoshen, M., Hagedorn, R., and Palmer, T. (2005). A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system using a malaria model. *Tellus A*, 57:464–475.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology*, 12(4):595 – 600.

Müller, W., CA, A., and Schär, C. (2005). Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Climate Dynamics*, 24:213–226.

O'Reilly, C. H., Weisheimer, A., MacLeod, D., Befort, D. J., and Palmer, T. (2020). Assessing the robustness of multidecadal variability in Northern Hemisphere winter-time seasonal forecast skill. *Quarterly Journal of the Royal Meteorological Society*, 146(733):4055–4066.

O'Reilly, C., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T., Schaller, N., and Woollings, T. (2017). Variability in seasonal forecast skill of northern hemisphere winters over the 20 th century: Variability in seasonal forecast skill. *Geophysical Research Letters*, 44.

Pan, Y. H. and Oort, A. H. (1983). Global Climate Variations Connected with Sea Surface Temperature Anomalies in the Eastern Equatorial Pacific Ocean for the 1958–73 Period. *Monthly Weather Review*, 111:1244–1258.

Parker, T., Woollings, T., Weisheimer, A., O'Reilly, C., Baker, L., and Shaffrey, L. (2019). Seasonal predictability of the winter north atlantic oscillation from a jet stream perspective. *Geophysical Research Letters*, 46.

Paternoster, R., BRAME, R., Mazerolle, P., and Piquero, A. (1998). Using the Correct Statistical Test for Equality of Regression Coefficients. *Criminology*, 36:859 – 866.

Poli, P., Hans, H., David, T., Dick, D., Jean-Noël, T., Adrian, S., Carole, P., Patrick, L., Takuya, K., Paul, B., Rossana, D., Yannick, T., Elias, H., Massimo, B., Lars, I., and Mike, F. (2013). The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C). *ERA Rep. Ser.*, 14.

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., Laloyaux, P., Tan, D. G. H., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L., and Fisher, M. (2016). ERA-20C: An Atmospheric Reanalysis of the Twentieth Century. *Journal of Climate*, 29(11):4083 – 4097.

Polyakov, I. V., Mayer, M., Tietsche, S., and Karpechko, A. Y. (2022). Climate Change Fosters Competing Effects of Dynamics and Thermodynamics in Seasonal Predictability of Arctic Sea Ice. *Journal of Climate*, 35(9):2849 – 2865.

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

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, 108(D14).

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, 27(6):2185 – 2208.

Shi, W., Schaller, N., Macleod, D., Palmer, T., and Weisheimer, A. (2015). Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters*, 42:1554–1559.

Sigmond, M., Scinocca, J., Kharin, V., and Shepherd, T. (2013). Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience*, 6:98–102.

Simmons, A., Hersbach, H., Munoz-Sabater, J., Nicolas, J., Vamborg, F., Berrisford, P., de Rosnay, P., Willett, K., and Woollen, J. (2021). Low frequency variability and trends in surface air temperature and humidity from ERA5 and other datasets. Technical Memorandum 881, ECMWF.

Simmons, A. J. (2022). Trends in the tropospheric general circulation from 1979 to 2022. *Weather and Climate Dynamics*, 3(3):777–809.

Smith, C. (2018). Niño 3.4 SST Index. `https://psl.noaa.gov/gcos_wgsp/Timeseries/Nino34/`. Accessed: 2022-04-07.

Stanski, H. R., Wilson, L. J., and Burrows, W. R. (1989). *Survey of common verification methods in meteorology*. Geneva: World Meteorological Organization.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4):485 – 498.

Thépaut, J.-N., Dee, D., Engelen, R., and Pinty, B. (2018). The Copernicus Programme and its Climate Change Service. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1591–1593.

Titchner, H. A. and Rayner, N. A. (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, 119(6):2864–2889.

Trenberth, K. (2018). NINO SST INDICES (NINO 1+2, 3, 3.4, 4; ONI AND TNI). `https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni`. Accessed: 2022-03-30.

Trenberth, K. E. (1997). The Definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2778.

Türkeş, M. and Erlat, E. (2003). Precipitation changes and variability in Turkey linked to the North Atlantic Oscillation during the period 1930–2000. *International Journal of Climatology*, 23:1771 – 1796.

van Oldenborgh, G. J., Balmaseda, M. A., Ferranti, L., Stockdale, T. N., and Anderson, D. L. T. (2005). Evaluation of Atmospheric Fields from the ECMWF Seasonal Forecasts over a 15-Year Period. *Journal of Climate*, 18(16):3250 – 3269.

Walker, G. T. (1933). Seasonal Weather and its Prediction. *Nature*, 132(3343):805–808.

Walker, G. T. and Bliss, E. W. (1932). World Weather V. *Memoirs of the Royal Meteorological Society*, 4(36):53–84.

WCRP (2007). WCRP Position Paper on Seasonal Prediction.

Weisheimer, A., Balmaseda, M. A., Stockdale, T. N., Mayer, M., Sharmila, S., Hendon, H., and Alves, O. (2022). Variability of ENSO forecast skill in 2-year global reforecasts over the 20th Century. *Geophysical Research Letters*, 49(10):e2022GL097885.

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T. N., Johnson, S., and Palmer, T. N. (2019). How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society*, 145(S1):140–159.

Weisheimer, A. and Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, 11(96):10.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A., and Palmer, T. (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. *Quarterly Journal of the Royal Meteorological Society*, 143(703):917–926.

Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., Brohan, P., Kent, E. C., Reynolds, R. W., Smith, S. R., and Wilkinson, C. (2011). ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, 31(7):951–967.

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M. (2019). The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean Science*, 15(3):779–808.