DISS. ETH NO. 27503

Subseasonal Prediction and Predictability of European Temperatures

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES OF ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

CHRISTOPH OLE WILHELM WULFF

MSc in Climate Physics: Meteorology and Physical Oceanography, Kiel University

born on 25.10.1989

citizen of Germany

accepted on the recommendation of:

Prof. Dr. Daniela Domeisen, examiner

Dr. Antje Weisheimer, co-examiner

Prof. Dr. Christof Appenzeller, co-examiner

2021

"Det er svært at spå — især om fremtiden."[1]

Danish proverb, author unknown

[1] It is difficult to predict — especially the future.

Abstract

In the last decade, increasing efforts have been made to improve atmospheric forecasts on time scales of weeks to months. The aim of these so-called subseasonal predictions is to push the limits of traditional weather forecasting and bridge the gap to seasonal predictions as forecast users from many sectors (including agriculture, health, humanitarian sector, energy trading, re-insurance) could strongly benefit from improved subseasonal forecasts. However, subseasonal prediction skill is generally low due to the chaotic nature of the atmosphere. It is thus crucial to distinguish between situations for which the skill is enhanced and those for which subseasonal forecasts cannot provide useful information. Knowledge about the temporal and spatial variations of predictability is valuable for users of subseasonal predictions and can improve their confidence in the forecasts.

Despite a variety of suggested sources of predictability, subseasonal prediction skill is particularly poor over Europe. In this thesis, we analyze the variations of subseasonal European near-surface temperature predictability. We use operational and retrospective forecasts out to two months lead time that are generated with numerical prediction models, the main tool used for making subseasonal predictions. We show that when evaluating the skill of subseasonal forecasts, deviations from stationarity in the climatology, such as trends, have to be carefully accounted for. When neglecting long-term warming trends during the reference period, estimates of temperature skill tend to be inflated compared to the actual skill. Only when the non-stationary components of the climatology are correctly taken into account, it is possible to make inferences about the temporal variability of predictability. We then use these results in order to analyze the month-to-month variations in subseasonal forecast skill. Using 20 years of retrospective forecasts, we show that there is a distinct seasonal cycle in subseasonal forecast skill: At lead times between 10 and 20 days, European land temperatures in winter, especially in February and March, are predicted with smaller errors and higher certainty than in any other season. The situations that are predicted best during these months feature a large-scale atmospheric flow pattern of increased westerly flow over the North Atlantic and concurrent higher temperatures over Europe, central Asia and Siberia. This zonal flow is accompanied by distinct anomalies in the tropics and polar stratosphere during forecast initialization that can serve as predictors of forecast skill. Especially cold anomalies in the lower stratosphere during initialization are useful indicators of enhanced subseasonal forecast skill for European land temperatures in winter. While subseasonal forecast skill in summer is lower than in winter, it is also possible to distinguish between better and poorer forecasts in summer based on the type of forecast event. Particularly in central to eastern Europe, summer warm extremes are overall predicted better than cold extremes and average temperatures indicating an asymmetry in skill for temperatures in different parts of the climatological distribution. This has potential implications for prediction skill in a future warmer climate.

The results of this thesis can allow stakeholders to make more informed forecast-based decisions and can thus help to increase confidence in the products provided by forecasting centers.

Zusammenfassung

Im letzten Jahrzehnt wurden zunehmend Anstrengungen unternommen, meteorologische Vorhersagen mit Vorhersagehorizonten von Wochen bis Monaten zu verbessern. Ziel dieser so genannten sub-saisonalen Vorhersagen ist, die Grenzen der traditionellen Wettervorhersage zu erweitern und die Lücke zu saisonalen Prognosen zu schliessen, da Vorhersageanwender aus einer Vielzahl von Sektoren (u. a. Landwirtschaft, Gesundheitswesen, humanitärer Sektor, Energiehandel, Rückversicherung) stark von verbesserten sub-saisonalen Vorhersagen profitieren könnten. Allerdings ist die Vorhersagefähigkeit (im Folgenden "Skill") auf sub-saisonalen Zeitskalen aufgrund der chaotischen Natur der Atmosphäre generell gering. Es ist daher entscheidend, zwischen Situationen unterscheiden zu können, für die der Skill besser ist, und solchen, für die sub-saisonale Prognosen keine nützlichen Informationen liefern können. Die Kenntnis der zeitlichen und räumlichen Variationen der Vorhersagbarkeit sind wertvoll für die Nutzer von sub-saisonalen Vorhersagen und können ihr Vertrauen in die Vorhersagen vergrössern.

Trotz einer Vielzahl möglicher Ursachen der Vorhersagbarkeit ist der tatsächliche sub-saisonale Skill derzeit besonders schwach über Europa. In dieser Arbeit analysieren wir die Variabilität der sub-saisonalen Vorhersagbarkeit von europäischen bodennahen Temperaturen. Zu diesem Zweck verwenden wir operationelle und retrospektive Vorhersagen mit einer Vorhersagezeit von bis zu zwei Monaten, die mit numerischen Vorhersagemodellen, dem Hauptinstrument für sub-saisonale Vorhersagen, erstellt wurden. Wir zeigen, dass es bei der Bewertung des Skills von sub-saisonalen Vorhersagen unerlässlich ist, Abweichungen der Klimatologie von der Stationarität, wie z.B. Trends, einzuberechnen. Werden die langfristigen Erwärmungstrends während des Referenzzeitraums vernachlässigt, können die Schätzungen des Temperatur-Skills gegenüber dem tatsächlichen Skill künstlich erhöht werden. Nur wenn die Nicht-Stationaritäten korrekt berücksichtigt werden, ist es möglich, Rückschlüsse auf die zeitliche Variabilität des Skills zu ziehen. In einem nächsten Schritt nutzen wir diese Ergebnisse, um die jahreszeitlichen Schwankungen im sub-saisonalen Skill analysieren zu können. Anhand von 20 Jahren retrospektiver Vorhersagen zeigen wir, dass es einen klaren saisonalen Zyklus in der sub-saisonalen Vorhersagbarkeit gibt. Bei Vorhersagezeiten zwischen 10 und 20 Tagen werden europäische Landtemperaturen im Winter, insbesondere im Spätwinter (Februar/März), mit geringeren Fehlern und mit geringerer Unsicherheit vorhergesagt als in jeder anderen Jahreszeit. Die Situationen, die in diesen vorhersagbareren Monaten am besten vorhergesagt werden, zeigen ein grossräumiges atmosphärisches Strömungsmuster mit verstärkter Strömung von Westen über dem Nordatlantik und simultan höheren Temperaturen über Europa, Zentralasien und Sibirien. Diese zonale Strömung ist mit deutlichen Anomalien in den Tropen und der polaren Stratosphäre während der Vorhersageinitialisierung verbunden, die als Prädiktoren für den Skill dienen können. Besonders kalte Anomalien in der unteren Stratosphäre während der Initialisierung sind nützliche Indikatoren für einen verbesserten sub-saisonalen Skill für europäische Landtemperaturen. Obwohl der sub-saisonale Skill im Sommer geringer ist als im Winter, ist es möglich, auch im Sommer zwischen besseren und schlechteren Vorhersagen zu unterscheiden, basierend auf der Art des vorherzusagenden Ereignisses. Besonders in Mittel- und Osteuropa werden warme Sommerextreme insgesamt besser vorhergesagt als kalte Extreme und Durchschnittstemperaturen, was auf eine Asymmetrie im Skill für Temperaturen in verschiedenen Bereichen der klimatologischen Verteilung hindeutet. Dies hat potenzielle Auswirkungen auf den Vorhersage-Skill in einem zukünftigen, wärmeren Klima.

Die Ergebnisse dieser Arbeit haben das Potenzial, Entscheidungsträgern zu ermöglichen, fundierte Prognose-basierte Entscheidungen zu treffen, und können somit dazu beitragen, das Vertrauen in die von den Vorhersagezentren bereitgestellten Produkte zu erhöhen.

Contents

Chapter 1

Introduction

1.1 Subseasonal Prediction and Predictability

Subseasonal prediction aims at bridging the gap between weather forecasting and climate prediction (Robertson and Vitart, 2019) by providing forecasts[1] of the state of the atmosphere[2] from two weeks up to two months ahead (Figure 1.1). The need for forecasts with these *lead times* is vast, and stakeholders in a multitude of areas, such as agriculture, energy, water management, health and the humanitarian sector, to name but a few examples, can benefit from useful subseasonal forecasts (Perez et al., 2015; White et al., 2017; Guimarães Nobre et al., 2019; Lopez et al., 2020). Despite the need, research on subseasonal forecasting has begun to flourish only rather recently. Part of the reason for this is that the subseasonal time scale was long considered a "predictability desert" (Robertson et al., 2020). Thus, prediction was separated into forecasting the day-to-day weather up to a maximum of one or two weeks ahead and predicting the expected climate over a certain season a month or more in advance (see the introduction to Vitart, 2004, for a brief history of monthly forecasting before 2004). However, since atmospheric variability exists on a continuum of time scales, it is now widely accepted that the problem of predicting weather and climate should be considered "seamless" (Hoskins, 2013), i.e. without conceptually separating the continuum of time scales of atmospheric variability. In the following, we will briefly review the reasons for this initial separation of the problems into weather forecasting and climate prediction and discuss why there are reasons to also expect predictability on the subseasonal time scale.

Broadly speaking, the evolution of the atmosphere is given by its dynamics, which, in general, is expressed by the equations of motion. To make a forecast, in addition to knowing the dynamics of the atmosphere, its initial state and the boundary conditions (e.g. the state of the ocean and the land surface) need to be known. Forecasting the weather is considered as being a pure initial value problem, since the theoretical capability to predict the weather, or weather *predictability*, arises almost entirely from information inherent in the initial conditions (predictability of the first kind). To produce a deterministic weather forecast, the equations of motion are integrated forward in time from the best estimate of the initial state of the atmosphere. The equations of motion are entirely deterministic,

[1] Throughout this thesis, the words "forecast" and "prediction" will be used synonymously.

[2] Predictions of other parts of the climate system, e.g. the ocean, can also be of interest but here we mainly focus on the atmospheric part unless stated otherwise.

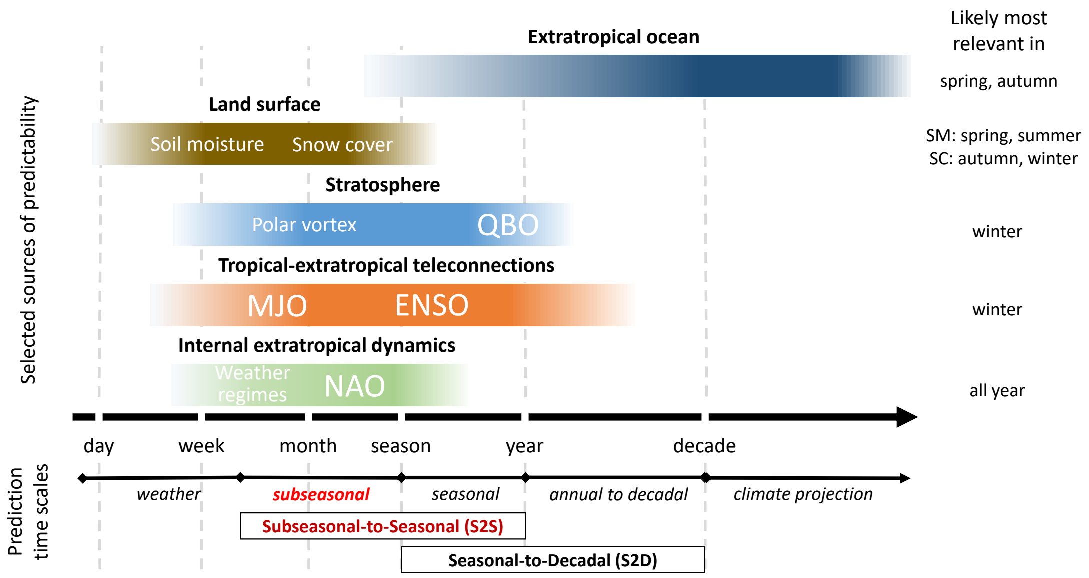

FIGURE 1.1: Overview of time scales and selected sources of predictability discussed in this Chapter. Prediction time scales are shown below the thick arrow. In subseasonal prediction (highlighted in red), we are concerned with forecasts of the atmosphere two weeks up to a season ahead. A selection of sources of extratropical predictability relevant for our discussion is shown above the arrow. The seasons in which each of the sources is most relevant for *extratropical subseasonal* prediction are indicated to the right of the figure. Adapted from Merryfield et al. (2020). QBO: Quasi-Biennial Oscillation, MJO: Madden-Julian Oscillation, ENSO: El Niño-Southern Oscillation, NAO: North Atlantic Oscillation.

meaning that for the exact same set of initial and boundary conditions, they always give the same forecast. However, Lorenz (1969) showed that even minute deviations between different sets of initial conditions several orders of magnitude smaller than the variable of interest can lead to vastly different forecasts. This means that if knowledge of the true initial conditions is less than perfect, the associated errors grow exponentially throughout the forecast due to the non-linear nature of the underlying deterministic equations of motion. Since the best guess of the initial state will in practice always be imperfect, this "deterministic chaos" sets a finite limit to the predictability of the weather, which is often thought to lie, on average, at lead times of approximately 10 – 14 days. Therefore, weather forecasts are typically not extended beyond this time scale. While this forecast horizon was not practically achievable at the time of Lorenz' study, numerical weather predictions have continuously improved since his discovery (Bauer et al., 2015). This "quiet revolution" (Bauer et al., 2015) is driven by advances in the understanding of the atmosphere and technical progress such as the tremendous increase in computational power. However, based on Lorenz' work, there should be a finite limit to the forecast horizon a weather model can achieve. Thus, if we treat forecasting as a pure initial value problem of the atmosphere, can we expect to make forecasts on the subseasonal time scale? First of all, the aforementioned predictability limit is state-dependent, which means that the predictability itself is a function of the initial conditions. In some situations, predictability from the initial conditions might be significantly longer than the average limit (implying of course, that in other situations it will be shorter) purely due to the dynamics of the atmosphere, which leads to flow-dependent or conditional predictability. Identifying theses states of extended predictability and their associated initial conditions is one prospect for providing useful forecasts in the subseasonal range. Furthermore, even the seemingly chaotic variability of the atmosphere exhibits some more or less regular variations on time scales longer than those considered in weather prediction. For instance, zonal winds in the tropical stratosphere exhibit strong variability with an average periodicity of 28 months, commonly known as the Quasi-Biennial Oscillation (QBO, see Baldwin et al., 2001). The QBO is an internal mode of atmospheric variability, meaning that it arises purely from the dynamics of the atmosphere without the need of an influence from the boundary conditions. This indicates that part of the atmospheric variability can indeed have predictability beyond the typical lead times of weather prediction. This lower frequency variability can easily be masked by the higher amplitude weather noise when considering for instance daily values. The predictable signal can be enhanced using statistical techniques of spatio-temporal aggregation (Toth and Buizza, 2019), e.g. by filtering out the variability on weather time scales.

As opposed to weather forecasting, seasonal prediction is considered mostly a boundary value problem. The premise of seasonal prediction is that atmospheric prediction is not a pure initial condition problem since the atmosphere alone cannot be considered a closed system. As such, it must be subject to boundary conditions (for instance, at the interface between the Earth and the atmosphere) that can impose an additional forcing. The most important of these boundary conditions is the solar radiation entering the atmosphere. In combination with Earth's near-spherical shape and the tilt of its axis relative to the axis of rotation around the Sun, it allows us to make the most basic seasonal forecast – that (in extratropical latitudes) summer will be warmer than winter. This simple prediction can conceptually be regarded as a result of predictability from external (or boundary) forcings on the atmosphere. At the lower boundary of the atmosphere, the ocean, land surface, sea ice, snow and vegetation constitute the boundary conditions. All these components vary on time scales longer than typical weather phenomena, and their state — most prominently that of the tropical ocean — can provide predictability of the second kind (i.e. predictability arising due to the boundary conditions). The fact that there is variability in the boundary conditions on subseasonal time scales (e.g. in the land surface conditions, Dirmeyer et al., 2019) can give rise to subseasonal predictability.

Subseasonal prediction can be considered a hybrid problem since on subseasonal time scales, predictability of the first *and* the second kind is likely important. While the separation of predictability into first and second kind can be conceptually useful, it is not always straight-forward to make this distinction in practice. The reason is that the lower boundaries[3] of the atmosphere are not truly external — depending on the time scale under consideration, their state itself depends on the atmospheric conditions. Thus, in many forecasting contexts, the ocean, sea ice and land surface have to be explicitly modelled to

[3] Note that also other components of the atmosphere are treated as boundary conditions despite being (physically) part of it, such as the concentrations of greenhouse gases or more generally chemical processes.

capture these interactions. This has led to the notion to consider seamless prediction as an initial value problem of the entire climate system (Hoskins, 2013).

In practice, there are of course technical limitations to the idea of seamlessness, and decisions have to be made on which components are most essential to simulate the variability at a particular time scale of interest. It is not entirely clear yet which components of the climate system have to be explicitly simulated for subseasonal prediction. As a consequence, forecasting models used for this purpose vary greatly in their set-up (Vitart et al., 2017). Nevertheless, many studies have addressed the known sources of subseasonal predictability and shown that there exists a multitude of predictors. However, as opposed to seasonal prediction, where variability in the tropical Pacific region is clearly the dominant source of predictability, on subseasonal time scales the predictors are weaker and vary depending on the region of interest. Some crucial sources of subseasonal predictability, with particular focus on the extratropics and Europe, are shown in Figure 1.1 and discussed in more detail in Section 1.2.

All the above arguments for potential subseasonal predictability could be made from a point of view of deterministic prediction, where we integrate only our best guess of the initial conditions forward in time using the equations of motion subject to the boundary conditions. However, subseasonal prediction has really only been made practically possible by the advancement in ensemble forecasting[4]. Simply put, in ensemble forecasting, the forecast model is integrated forward in time from a finite number (ensemble) of initial conditions, which are designed to sample the distribution of initial conditions that would be possible given the uncertainty in their estimation. At any point in the forecast, the ensemble of simulated states is a sample of the possible states that the climate system could attain given the distribution of initial conditions. This results in quasi-probabilistic forecasts that allow us to infer, for instance, the most likely outcome of a forecast and, even more importantly, the forecast uncertainty. Other than initial condition uncertainty, uncertainties due to the model formulation can also be addressed with ensemble forecasts. This ability to account for the two main sources of errors makes ensemble forecasts an important tool to produce reliable probabilistic forecasts (Weisheimer and Palmer, 2014). Since we make extensive use of ensemble forecasts in this thesis, we briefly review the concept of ensemble forecasting in Section 1.3.

Despite the discussed prospects for subseasonal atmospheric predictability, the actual skill of these forecasts is rather limited to this day. Understanding their potential value and documenting the improvement requires continuous verification against observations. This is no trivial task and there is no single measure that can completely quantify the quality of a forecast. Instead, different attributes of the forecasts can be measured and used to assess their quality and — depending on the requirements of the user — their value. We give a more detailed introduction to forecast verification in Section 1.4, where we also introduce the concept of *skill*. Until explicitly defined in Section 1.4.3, we use this term to rather vaguely refer to the ability of a forecast (or forecast system) to make a useful prediction.

[4] While the focus here is on subseasonal prediction, it should be noted that predictions on any time scale have benefited from this development (e.g. Buizza and Leutbecher, 2015).

1.2 Sources of Subseasonal Predictability

Despite subseasonal forecasting being a relatively young field, many studies have demonstrated skillful predictions, both with statistical and numerical models, proving the existence of subseasonal predictability for many regions of the world (e.g. DelSole et al., 2017; Mundhenk et al., 2018; Xiang et al., 2019; Camargo et al., 2019; Robertson et al., 2020). To illustrate the potential skill of subseasonal forecasts, we show the correlation between a reanalysis and forecasts of near-surface temperatures as well as geopotential height at 500 hPa at lead times of 3 and 4 weeks in Figure 1.2. Clearly, subseasonal forecast skill in the tropics is exceptionally high compared to the rest of the world with correlations exceeding 0.8 in some regions even at 4 weeks lead time. The reason for the higher tropical skill is the strong coupling between the tropical atmosphere and the ocean, which has much longer persistence. Furthermore, there are two main modes of variability in the tropics that act as sources of subseasonal predictability (El Niño-Southern Oscillation and the Madden-Julian Oscillation, see Section 1.2.2). In contrast, the extratropics and continental Europe in particular are areas of relatively low subseasonal prediction skill. The more dynamical climate in these regions significantly lowers extratropical predictability. Nevertheless, subseasonal skill has been shown to exist outside of the tropics (e.g. Weigel et al., 2008b; Vitart, 2014), particularly for the prediction of extreme temperature events (e.g. Hudson and Marshall, 2016; Osman and Alvarez, 2018; Ardilouze et al., 2017b; Lavaysse et al., 2018; Wulff and Domeisen, 2019), which are the focus of Chapter 2. Within the extratropics, Europe in particular stands out as a region of low skill. This is not due to a lack of potential sources of subseasonal predictability. In fact, many processes are thought to influence the subseasonal variability of the atmosphere in the North Atlantic-European (NAE) sector. Their influence, however, varies seasonally and regionally, and multiple processes might interact in complex ways. This makes Europe one of the most challenging, but at the same time, one of the most interesting areas for subseasonal prediction, and it is the main region of interest in this thesis. In the following subsections we discuss what we consider to be the most important sources of subseasonal predictability in NAE sector.

1.2.1 Extratropical Modes of Variability

Atmospheric variability in the mid-latitudes is generally dominated by baroclinic instabilities or eddies, i.e. the high- and low-pressure systems responsible for most of the daily weather (Holton, 2013). A single one of these weather systems is not predictable beyond a few days. However, their pressure distributions over multiple days are characterized by certain dominant states that can be described by weather regimes. These weather regimes are strongly related to changes in the background flow of the atmosphere and the breaking of atmospheric waves (Franzke et al., 2004; Woollings et al., 2010; Madonna et al., 2017). The regimes are accompanied by changed likelihoods for certain weather situations and thus have an imprint on other surface variables, such as temperature and precipitation (Madonna et al., 2021). They are also linked to the occurrence of blocking anticyclones,

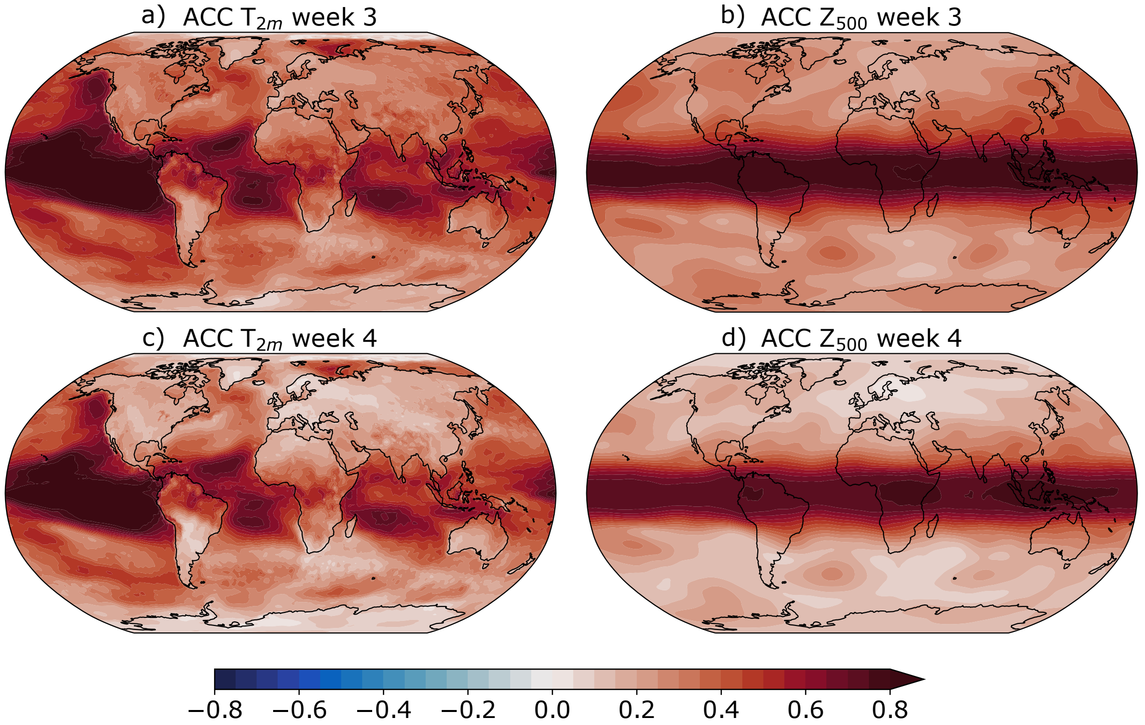

FIGURE 1.2: Global deterministic skill at subseasonal lead times. The shading shows the week 3 (a and b) and 4 (c and d) correlation skill of 7-day averages of standardized 2-m temperature (a and c) and 500hPa geopotential (b and d) anomalies for ensemble mean hindcasts from the ECMWF extended-range forecasting system verified against ERA-Interim over all initializations in the 20 year period from 1998 – 2017 (105 initializations per year). The shading interval is 0.05.

which are long-lasting phases of blocked zonal flow and are related to many surface extreme events in the mid-latitudes (Buehler et al., 2011; Pfahl and Wernli, 2012). Regimes can persist over significantly longer time periods than a single weather system, meaning that they could be predictable on subseasonal and seasonal time scales. In the NAE region, the seasonal regime behaviour is most prominently characterized by the North Atlantic Oscillation (NAO, e.g. Hurrell et al., 2003). In fact, when analysing the winter mean sea level pressure (SLP), about 35% of the year-to-year variance can be explained solely by this single pattern of variability (Hurrell and Deser, 2009). The NAO is a so-called low-frequency mode of variability because it represents the slow (compared to the time scale of the weather systems) variations of a large part of the atmospheric variability in a region (the NAE region). Other modes of variability can be identified in the NAE region but also in other parts of the extratropics (Wallace and Gutzler, 1981), all of which are connected to changes in the atmospheric background flow. It is furthermore possible to define a number of different weather regimes in the NAE region that exhibit significant variability on subseasonal time scales (e.g. Michelangeli et al., 1995; Dawson et al., 2012; Grams et al., 2017), some of which are closely related to the NAO. All of these exert some control on the surface variability in different regions of the NAE sector. Thus, if some of the subseasonal variations of these regimes could be predicted, it would likely imply some subseasonal predictability for surface variables. While the extratropical regimes may result purely from internal extratropical dynamics, there is evidence that they can be influenced by variability in other parts of the climate system. This includes the sources of predictability discussed in the following sections.

1.2.2 Tropical-Extratropical Teleconnections

Variability in the tropics can exert a strong influence on the circulation in the extratropics. This influence is referred to as atmospheric tropical-extratropical *teleconnections* (Liu and Alexander, 2007; Stan et al., 2017). The word teleconnection describes the co-variability of variables in two remote places of the globe. These teleconnections are communicated through the atmosphere by quasi-stationary planetary-scale Rossby waves. The Rossby waves themselves are generally excited in the tropics by anomalous upper-level divergent flow that acts as a Rossby wave source (Trenberth et al., 1998). The anomalous upper-level divergence is a result of diabatic heating from large-scale convective activity in the tropics, which can, for instance, be forced by anomalies in tropical sea surface temperatures (SST). Some of these tropical anomalies in large-scale diabatic heating are predictable on subseasonal time scales (Vitart, 2014; Lim et al., 2018). Jin and Hoskins (1995) show that the response to anomalous diabatic heating in the tropics takes about one to two weeks to be established in the mid-latitudes. As a result, atmospheric teleconnections from the tropics can act as a source of subseasonal predictability in the extratropics. Additionally, some tropical modes of variability tend to remain in the same phase for multiple days to months and can thus impose a quasi-permanent forcing on the subseasonal variability in parts of the extratropics. However, since the propagation of Rossby waves relies strongly on the presence of sufficient meridional gradients of potential vorticity (PV) in the background flow of the atmosphere as a restoring force (Palmer and Anderson, 1994), there are marked seasonal variations in the teleconnections. Generally, tropical-extratropical teleconnections are stronger in the winter hemisphere due to enhanced meridional PV gradients, which act as wave guides and are a result of stronger latitudinal variations in heating in the winter hemisphere (Branstator, 2002). Nevertheless, several teleconnections during the boreal summer season have been identified (Ding et al., 2011; Wulff et al., 2017; O'Reilly et al., 2018), but ultimately, the effects of teleconnections on extratropical predictability are expected to be stronger in winter. It should also be noted that the tropical and extratropical circulation are coupled and thus any change in the extratropics necessarily affects the tropics. Despite this, the teleconnections to the NAE region that are described in the next two paragraphs are often treated as one-way influences and this assumption holds in good approximation in the cases considered in this thesis.

On seasonal time scales, the most prominent teleconnections are related to diabatic heating anomalies in connection with the El Niño-Southern Oscillation (ENSO), which is a coupled atmosphere-ocean phenomenon in the tropical Pacific that controls large parts of the atmospheric variability in this region but, importantly, also globally. For instance, the atmospheric circulation over the extratropical North Pacific exhibits strong ENSO-controlled variability of the Pacific/North American (PNA) pattern (Wallace and Gutzler, 1981), which is the dominant mode of atmospheric variability in this region. To a lesser degree, ENSO also impacts the European climate (Brönnimann, 2007). Furthermore, ENSO is associated with distinct patterns of subseasonal variability in the extratropics (Compo et al., 2001).

The dominant pattern of large-scale atmospheric variability on subseasonal time scales is the Madden-Julian Oscillation (MJO, Xie et al., 1963; Madden and Julian, 1971). The MJO is characterized by an eastward-propagating zonal dipole of suppressed and enhanced organized convection (Zhang, 2005). Depending on the location of this dipole the MJO is said to be in one of eight phases. The MJO exhibits irregular frequencies between 30 and 60 days and is thus a truly subseasonal mode of variability. While of tremendous importance for weather in the tropics, the MJO also displays strong subseasonal teleconnections to the extratropics in both hemispheres (Stan et al., 2017, and references therein). Most interestingly for our discussion, the MJO has been shown to influence subseasonal variability in the NAE region by modulating the occurrence of the North Atlantic circulation regimes (Ferranti et al., 1990; Cassou, 2008), such that some regimes display a different frequency of occurrence after certain phases of the MJO. These changes show a significant dependence on the prevailing ENSO conditions (Lee et al., 2019), the state of the stratosphere (Garfinkel and Schwartz, 2017) and the propagation speed of the convection anomalies (Yadav and Straus, 2017). The change of regime occurrence reflects the fact that the MJO-induced teleconnections interact with the jet stream. As a consequence, these teleconnections can have an impact on blocking events and the occurrence of atmospheric rivers (Stan et al., 2017), large structures of high moisture content in the atmosphere that cause many precipitation-related extremes in the extratropics (Ralph et al., 2006; Dettinger, 2011). The MJO is thus crucial in the discussion of extratropical subseasonal predictability, especially given its time scales and role in the modulation of extratropical variability. The degree to which subseasonal forecasts can benefit from the MJO as a source of predictability depends on how well the models simulate its variability and the related tropical-extratropical teleconnections. Importantly, skill at forecasting the MJO has continuously increased over the last two decades (Vitart, 2014). Furthermore, Lin et al. (2010) and Vitart and Molteni (2010) show that subseasonal prediction skill for Europe is enhanced during phases of strong MJO activity indicating that forecast models capture at least part of the response in the NAE region to tropical MJO forcings. However, correctly predicting the teleconnections related to the individual MJO phases is still a challenge for forecast models, which is mainly due to model biases in the background flow that alter the propagation of Rossby waves in the extratropics (Vitart and Balmaseda, 2018; Lin, 2020).

Although teleconnections related to the MJO and ENSO are most important for the subseasonal predictability in the extratropics, other tropical-extratropical teleconnections exist. For instance, upper-level divergence in relation to the Indian Summer Monsoon can interact with the mid-latitude circulation on subseasonal time scales through the circumglobal teleconnection (Ding and Wang, 2005; Di Capua et al., 2020). While the summer monsoon systems are potential sources of subseasonal variability, many studies indicate that the interaction between the monsoon and the mid-latitude circulation cannot be considered a one-way influence from one onto the other. Instead one needs to account for the coupled nature of the phenomenon (Bollasina and Messori, 2018). Thus, the monsoons' influence on mid-latitude subseasonal variability is likely more variable and complex and not easily quantified.

1.2.3 Stratosphere

For mid-latitude winter climate variability, particularly in the NAE region, the polar stratosphere has been recognized as a major source of predictability. Its ability to provide subseasonal predictability has so far been only considered for the winter season since summer stratospheric variability is extremely low due to the inhibition of wave propagation into the stratosphere by the prevailing easterly flow (Charney and Drazin, 1961). During winter however, the polar stratosphere is dominated by westerly winds, which form the so-called stratospheric polar vortex (SPV). These prevailing westerly winds allow for the propagation of planetary-scale waves into the the polar stratosphere (Charney and Drazin, 1961), where they can break and deposit momentum (Baldwin and Holton, 1988). As a result, stratospheric variability is strongly enhanced during winter and the SPV can experience disruptions that cause it to be split or displaced from the pole. This results in anomalously warm conditions over the polar cap (Labitzke, 1977), which is why these events are referred to as Sudden Stratospheric Warming (SSW) events (Baldwin et al., 2021). Importantly, during these events, but also during phases of an enhanced SPV, the stratosphere couples with the troposphere and is thus able to alter the tropospheric circulation. SSW events are often followed by equatorward shifts of the tropospheric jet stream that persist up to several weeks after the stratospheric event. Phases of anomalously strong westerly winds in the SPV on the other hand are associated with a poleward shift of the jet stream on similar time scales. These shifts are seen in observations (Butler et al., 2019) and have been shown to exist in models of different complexity (Polvani and Kushner, 2002; Scaife et al., 2005). While this response in the jet stream is evident in the zonal mean circulation, it is strongest in the North Atlantic sector (e.g. Greatbatch et al., 2012) where it manifests as a modulation of the occurrence of weather regimes (Charlton-Perez et al., 2018; Beerli and Grams, 2019). As a result, surface winter weather in the NAE sector can be strongly affected during these phases of stratosphere-troposphere coupling. Domeisen and Butler (2020) specifically point out the role of the stratosphere in driving surface extreme events. The effect of the stratosphere onto the troposphere has been demonstrated in models of varying complexity (Polvani and Kushner, 2002; Hitchcock and Simpson, 2014). While disruptions of the SPV happen abruptly (especially for SSWs), the recovery and downward impact of the SPV take place on subseasonal time scales. In addition, the stratosphere exhibits predictability on longer time scales than the troposphere (Domeisen et al., 2020c). This longer stratospheric predictability in combination with the influence of the stratosphere on subseasonal tropospheric variability makes the stratosphere an important source of subseasonal predictability in the NAE region during boreal winter.

Figure 1.3 shows the development of the ensemble spread and probabilistic error of zonal mean zonal wind at 60°N in the stratosphere (10 hPa), which serves as an indicator of the strength of the SPV. By pooling the forecasts depending on the initial state of the SPV (terciles of weak, normal and strong vortex) it can be seen that both the error and the ensemble uncertainty on subseasonal lead times are substantially reduced after weak SPV states (red line) indicating flow-dependent predictability. However, it remains elusive how much this extended stratospheric predictability translates into tropospheric skill (Domeisen et al., 2020b). If the state of the stratosphere can be reliably predicted with lead times of one to two weeks and the stratosphere further exhibits sufficient coupling to the troposphere, this would imply subseasonal predictability for the winter climate at the surface.

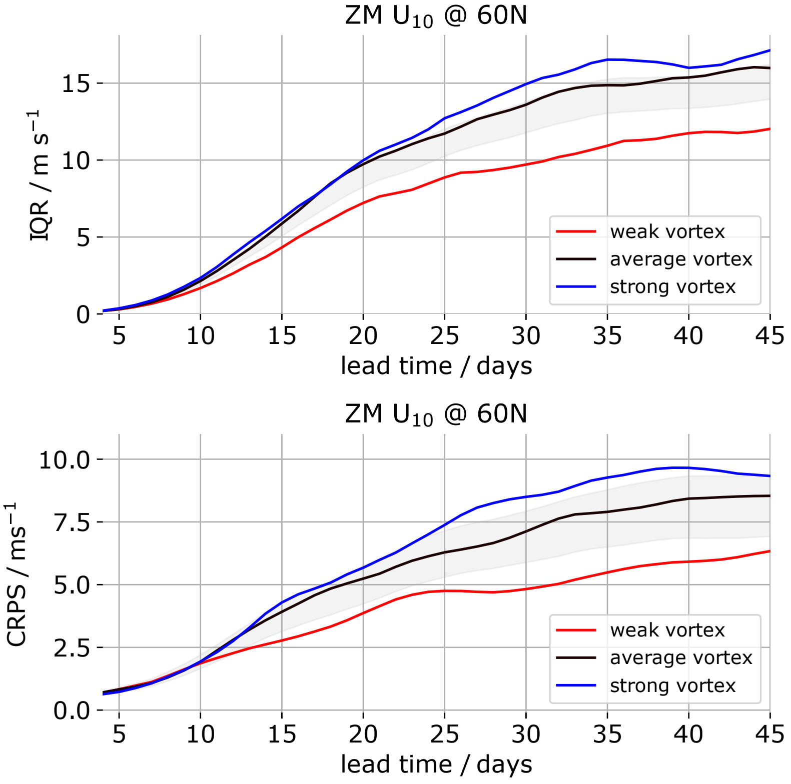

FIGURE 1.3: Flow-dependent development of ensemble spread and probabilistic skill in forecasts of stratospheric zonal mean zonal wind (U_{10}) at 60°N during winter (December – March) for ECMWF hindcasts initialized between 1998 and 2017. The top panel shows the interquartile range (IQR) of the ensemble forecasts as a function of lead time. The differently colored lines show averages over forecasts that were initialized under different stratospheric initial conditions as indicated in the labels. The binning of the initial conditions is done based on terciles of U_{10} at 60°N. The grey shading shows the 95% interval of a bootstrap distribution generated by 1000 averages of random draws of a third of the initializations in the season. The bottom panel is analogous to the top one but showing the continuous ranked probability score (CRPS), which can be interpreted as a measure of the error of a probabilistic forecast (for details, see Section 1.4).

As described above, there is strong support for the ability of stratospheric anomalies to cause extended predictability. This ability has been demonstrated in a number of studies. For instance, the added value of perfectly knowing the state of the stratosphere for seasonal predictions of the NAO and North Atlantic winter storms has been shown by

Hansen et al. (2017) and Hansen et al. (2019), respectively, using a relaxation technique to mimic a perfect prediction of the stratosphere. Including the stratosphere as a predictor can also enhance the skill of statistical seasonal winter forecasts of the North Atlantic surface climate (Karpechko, 2015; Dobrynin et al., 2018), and Beerli et al. (2017) successfully use the stratosphere as a predictor of wind electricity generation in Europe on monthly time scales. While there clearly is subseasonal skill from the stratospheric conditions, the largest errors in the state of North Atlantic climate also tend to coincide with significant anomalies in the SPV's initial state (Kolstad et al., 2020). Especially after SSWs, these errors have the potential to negatively affect surface skill over Europe (Domeisen et al., 2020b; Büeler et al., 2020). Thus, although the potential of the stratosphere to provide subseasonal predictability is great, there are indications that dynamical forecast models do not yet fully realize the extended predictability over the North Atlantic from the stratosphere, likely due to errors in the representation of stratosphere-troposphere coupling (Domeisen et al., 2020b).

1.2.4 Land Surface Processes

The land surface plays an important role for prediction mostly due to its interaction with the atmosphere as part of the global water cycle (Seneviratne et al., 2010). Dirmeyer et al. (2015) define three necessary conditions that need to be met for the land surface to significantly affect the atmosphere on subseasonal time scales. Firstly, there needs to be ample sensitivity (or coupling) of the atmosphere to the land surface conditions. Only when the surface fluxes of energy and water are mostly controlled by the land surface state they can affect the atmosphere and be considered a forcing. Secondly, the variability in the land surface conditions has to be sufficiently large to have a noticeable effect on the atmosphere. The first two conditions are independent of the time scale under consideration and are best illustrated by the regime behaviour of the surface turbulent heat fluxes (Seneviratne et al., 2010). Figure 1.4 shows the concept of how the evaporative fraction (EF, the part of the surface net radiation that is used for evaporation/latent heat flux) changes with soil moisture content θ. Inside the energy limited regime, EF is at its maximum and is independent of the soil moisture content of the land. Here, evapotranspiration is solely determined by the vapour pressure deficit of the atmosphere above the land surface, and thus there is no sensitivity to the land surface conditions. In the soil moisture limited regime however, latent heat flux depends strongly on the soil moisture that is available for evaporation. Here, the net radiation is partitioned into sensible and latent heat flux with the former increasing as less soil moisture becomes available. When soil moisture drops below a certain threshold (θ_{WILT}), all turbulent heat flux will be in the form of sensible heat. In this soil moisture limited regime, the sensitivity of the atmosphere to changes in the land surface conditions is high. However, for soil moisture to significantly influence the overlying atmosphere, its variations need to be large enough. Thus, mainly in the transitional part between the dry and wet regimes the land surface can be considered to force part of the atmospheric variability close to the surface. The third condition for subseasonal predictability from land surface conditions is their capability to act as a memory of the atmospheric conditions

(Dirmeyer et al., 2015). Soil moisture for instance is strongly controlled by precipitation. Since the land surface has a certain capacity to store water, soil moisture, especially in the deep layers, can in some cases be considered a summation of precipitation events (or the lack thereof) over multiple days. This information can then be released back to the atmosphere in the form of heat fluxes at the land-atmosphere interface during phases of coupling. Only where the land surface shows memory beyond the typical weather time scales it can be expected to influence subseasonal predictability. All three of the previously described conditions for subseasonal predictability from the land surface vary throughout the globe, and not all climatic zones are likely to exhibit subseasonal predictability from land surface conditions (Dirmeyer et al., 2019).

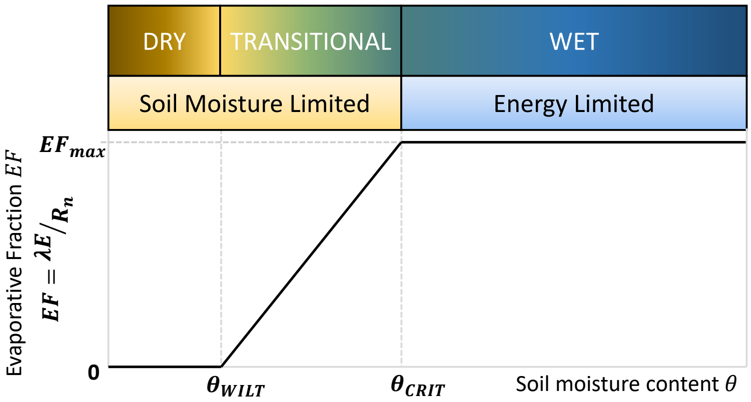

FIGURE 1.4: Illustration of evapotranspiration regimes. See text for details. Adapted from Seneviratne et al. (2010).

Many studies have shown the potential of the land surface conditions to cause subseasonal and seasonal predictability. Mid-latitude heat waves for instance can be strongly influenced by interactions between the land surface and the atmosphere (Seneviratne et al., 2006; Fischer et al., 2007b; Fischer et al., 2007a). Weisheimer et al. (2011) showed that temperatures leading to the 2003 European heat wave reached highly anomalous level partly due to the preconditioning of the soil, which was anomalously dry already in the preceding spring. This extreme dryness was maintained into summer by local circulation patterns. Using an improved land surface scheme in combination with updated parameterizations for radiation and convection they were able to produce a skillful seasonal forecast of this extreme event, highlighting the relevance of the land-atmosphere interactions as a source of predictability. Bunzel et al. (2018) further show that replacing a simple bucket model with a more elaborate five-layer hydrology scheme in the Max Planck Institute Earth System Model significantly improves summer seasonal forecasts of near-surface temperatures generally and not exclusively for extremes. They attribute the improvement to a better representation of soil moisture-temperature feedbacks. But soil moisture is not only important for seasonal predictability. Prodhomme et al. (2016) stress the relevance of initializing soil moisture in subseasonal forecasts in summer with realistic rather than climatological conditions. Their realistically initialized forecasts show better overall temperature and precipitation skill and also simulate the extreme conditions of the 2010

Russian heat wave more successfully. However, the benefits of more realistic soil moisture initialization can be hampered if the model does not produce the correct land-atmosphere interactions, leading to a decrease in subseasonal prediction skill (Ardilouze et al., 2017a).

While the focus is often on soil moisture, other land surface variables can have an impact on subseasonal and seasonal prediction skill, most prominently snow cover (Sobolowski et al., 2010; Douville, 2010; Xiang et al., 2020). Thomas et al. (2016) indicate that subseasonal forecasts in spring benefit only moderately from realistic land surface initialization. They attribute the additional skill in this season to more realistic variability of the snow water equivalent, the influence being mostly a result of the snow-albedo feedback (Jeong et al., 2013). Kolstad et al. (2017) also suggest that enhanced persistence of near-surface temperatures can be caused by soil temperature and snow depth, which implies the possibility that these variables can act as predictors of subseasonal temperature variability.

Apart from the more local influence on near-surface temperature, land surface anomalies (mainly snow cover and soil moisture) are suggested to influence the large-scale circulation by inducing anomalous wave propagation (Cohen and Entekhabi, 1999; Palmer, 1993). For instance, under the right atmospheric preconditioning, such as in relation to a blocking situation, soil moisture anomalies can modify the circulation through feedbacks, possibly amplifying the concurrent large-scale atmospheric conditions (Pal and Eltahir, 2003; Fischer et al., 2007a). These effects are likely limited to extreme conditions but for these cases, they offer an important source of subseasonal predictability.

In summary, there is ample evidence that land surface processes can act as a source of subseasonal predictability. Both subseasonal and seasonal skill at predicting temperature and to a lesser degree also precipitation have been shown to benefit from better initialization of the land surface, as well as more accurate representations of the processes within the upper surface layers and their interaction with the atmosphere. For Europe specifically, land-atmosphere interactions are strongest during the summer months and they could be especially relevant for the prediction of heat waves. Under extreme conditions, soil moisture-atmosphere coupling even has the potential to influence the persistence of the atmospheric circulation.

1.2.5 Extratropical Oceans

The ocean has larger thermal inertia than the atmosphere and consequently exhibits more low-frequency variability. It can thus provide extended predictability, especially in the tropics. The extratropical ocean, however, couples only weakly to the atmosphere (Kushnir et al., 2002) such that in the mid-latitudes the atmosphere forces the ocean much more strongly than the other way around. In recent years however, evidence has emerged that even this weak influence of the ocean onto the atmosphere can provide seasonal predictability by moderately shifting the probability distribution of the large-scale atmospheric circulation regimes (Gastineau and Frankignoul, 2015; Ossó et al., 2017). Especially during autumn and summer when daily SST variability is largest, the extratropical ocean-atmosphere coupling can be sufficiently strong to significantly alter the large-scale circulation (Nie et al., 2019; Ossó et al., 2020). During these seasons, the location of the eddy-driven jet frequently coincides with the regions of strongest SST variability close to the western boundary currents. Anomalous SST gradients in the regions where the jet is located can alter the baroclinicity in the lower troposphere and thus lead to shifts in the eddy-driven jet stream (Nie et al., 2019). The effects of this jet shift can significantly alter the summer climate over the NAE region. Ossó et al. (2017) showed that the ocean-atmosphere feedback described above can enhance seasonal predictability for the NAE summer climate. This feedback could further be the reason for the documented link between the warm European summer of 2015 and the simultaneous cold SST anomalies in the North Atlantic (Duchez et al., 2016). For the North American continent, McKinnon et al. (2016) present evidence that hot spells in eastern North America are preceded by a distinct SST anomaly pattern in the North Pacific that allows for skillful predictions of warm conditions up to 50 days ahead. They show that ocean-atmosphere interactions in the North Pacific force persistent anti-cyclonic anomalies over the eastern U.S., leading to anomalously warm conditions at the surface. In summary, the influence of the extratropical ocean on subseasonal atmospheric variability is likely limited to certain seasons. Nevertheless, the ocean's ability to both influence the local boundary layer conditions and the large-scale atmospheric circulation makes it another important source of subseasonal predictability.

1.2.6 Further Aspects

We focused the above discussion mostly on what we consider the dominant sources of subseasonal extratropical predictability for the mid-latitudes and the NAE sector in particular. Some of the discussed sources of subseasonal predictability are not exclusive to these regions. In addition, other sources could bear further potential for subseasonal predictability. Sea-ice for instance has the ability to provide subseasonal predictability but its potential outside of the Arctic and Antarctic regions is still under debate (Chevallier et al., 2019) and may be additionally modified by Arctic warming (Screen, 2014). Furthermore, aerosols can have profound influences on the radiation balance and integrating interactive aerosols into a subseasonal prediction model can have a positive effect on the monthly forecast skill in spring and summer (Benedetti and Vitart, 2018). More sources of predictability beyond the ones described in the above sections likely exist but are not further discussed here.

1.3 Ensemble Forecasting

To a significant degree, progress in subseasonal forecasting was made possible by improvements in the prediction models (Bauer et al., 2015). Increased horizontal and vertical resolution as well as more sophisticated parameterizations have resulted in better representations of many of the aforementioned processes (Section 1.2) in the models. Especially forecasts at time scales longer than multiple days have also benefited from the introduction of ensemble forecasting techniques (Slingo and Palmer, 2011; Buizza and Leutbecher, 2015). Rather than just giving deterministic forecasts (which also an ensemble forecast can be easily transformed into) these provide quasi-probabilistic forecasts that sample the forecast uncertainty and thus add relevant information to the predictions.

For the theory underlying ensemble prediction it is useful to think of a model of the atmosphere as a dynamical system in phase space (as opposed to the physical space, which is spanned by the 3 spatial dimensions). In this abstract phase space, the state of the atmosphere at one point in time can be described by a single combination of realizations of all of its variables (i.e. one specific value for every parameter at every single point of the atmosphere). The evolution of the dynamical system with time is described by its trajectory, i.e. the path of the point in phase-space that describes the state. This trajectory evolves in time as a result of the dynamics of the system. Depending on the dynamics, only a subset of all possible points in phase-space can be reached by the system and this subset is called the attractor. For a deterministic forecast, it is assumed that the initial conditions of the atmosphere are represented by a single point on this attractor. However, it is also possible to describe a probability density function (pdf) on the attractor in phase-space instead of a single point. The evolution of this pdf in phase-space, given the dynamics of the climate system, can be described with a prognostic equation, the Liouville equation[5] (Ehrendorfer, 1994). Knowing that our best estimate of the initial conditions is necessarily subject to errors, we could try to quantify the uncertainty and thus hope to better describe the possible initial states by using a pdf instead of one point. If we could solve the Liouville equation, this would allow for a truly probabilistic forecast, in which the state of the atmosphere at any given time is given by a pdf in phase space. Although the predictability of the system is completely characterized by the Liouville equation (Ehrendorfer, 2006), in practice it cannot be solved due to the high dimensionality of the problem. Instead, the concept of ensemble forecasts is used. To produce an ensemble forecast, the pdf of initial conditions, PDF(0), that represents the uncertainty in the estimate of the initial conditions is sampled by introducing small perturbations to the best guess. Generating these perturbations is a highly complex process since the perturbed states need to be consistent with the dynamics of the system (e.g. Kalnay et al., 2006). From each initial state, a forecast is then generated by integrating the equations of motion forward in time. The result is an ensemble forecast that aims to represent a finite sample of the distribution of the system's trajectories in phase-space. The principle of ensemble forecasting is illustrated in Figure 1.5.

Great benefit has been demonstrated from combining ensembles of different forecasting models into so-called multi-model ensembles (MME, Hagedorn et al., 2005). These MME consistently outperform any single MME member (which itself can be an ensemble system, Hagedorn et al., 2005). The reason for this is that the MME accounts for uncertainties in the model physics. While any single ensemble system is run with a fixed set of parameterizations, which are essentially empirical estimates of complex physical processes, in an MME the models differ in their physical parameterizations and thus the MME accounts for some of the uncertainties in the formulation of these processes. One effect of

[5] A conservation equation for the probability in phase-space analogous to the conservation of mass in physical space.

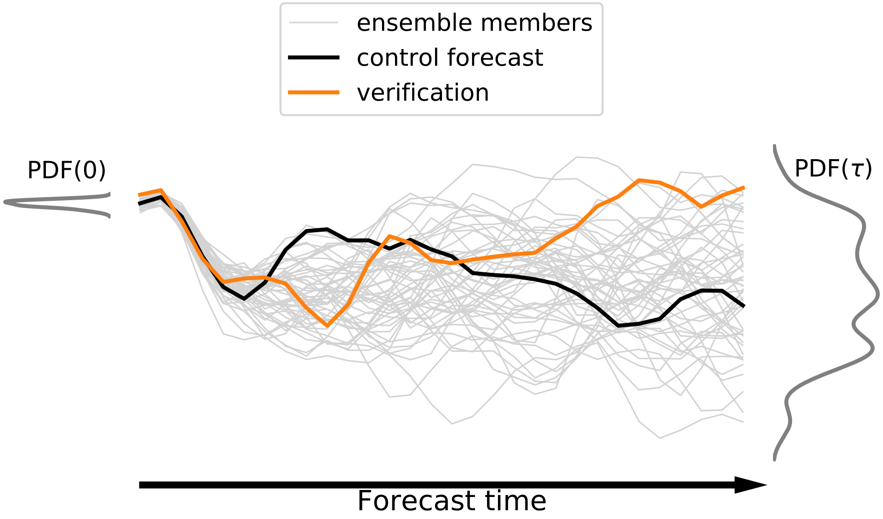

FIGURE 1.5: Illustration of the principle of deterministic versus ensemble forecasting. The true time evolution of the system is shown by the orange line. A deterministic forecast can be generated from the best guess of the initial conditions (control forecast, thick black line). However, given the uncertainty in the estimate of the initial guess, the true initial conditions are better represented by PDF(0), the probability density function of the initial states at time 0. An ensemble of forecasts (grey lines) is generated by integrating the model from a finite sample of initial conditions from the PDF(0). The ensemble at forecast time τ can then be considered a sample of PDF(τ), the distribution of states at time τ. Compare to Toth and Buizza (2019).

this unaccounted uncertainty in single model ensembles is that forecasts tend to be underdispersive, i.e. the ensemble members do not develop a spread that matches the ensemble mean error with respect to the observed variability (Weigel et al., 2008a). Combining single ensembles into an MME reduces this overconfidence of the forecasts, which results in higher skill (Weigel et al., 2008a). Because of the benefits of combining multiple ensemble systems, the two major international projects on subseasonal predictability, namely the Subseasonal Experiment (SubX, Pegion et al., 2019) and the S2S prediction project (Vitart et al., 2015, see also Infobox 1), have set up a database of forecasts from multiple forecasting centers that can in principle be combined into an MME. There is evidence that a single model ensemble could equally well account for uncertainties in the model physics and thus — among many other benefits — alleviate the overconfidence. This can be done by introducing stochastic perturbations to the model parameterizations. This approach was introduced as early as 1999 (Buizza et al., 1999). Nowadays, a stochastic physics scheme is successfully implemented into the current operational forecasting model of the European Centre for Medium-Range Weather Forecasting (ECMWF, Leutbecher et al., 2017). Apart from addressing overconfidence, stochastic parameterizations alleviate the problems stemming from the truncation and scale separation between resolved and unresolved processes in the models (Palmer, 2019).

To summarize, using ensemble predictions has greatly aided in improving forecasts because it addresses the two main sources of errors in the models (Palmer, 2006). While perturbations to the initial state address the error in initial conditions of the forecasts, the use of MMEs addresses the uncertainties in the model formulation. Ensemble techniques allow for the quantification of the uncertainty of the forecast itself, which is of tremendous use in assessing climate risks and for making decisions that crucially depend on weather and climate.

Infobox 1: The Subseasonal-to-Seasonal (S2S) prediction project

Due to the growing need for forecasts on subseasonal to seasonal (S2S) time scales (White et al., 2017), the World Climate Research Programme (WCRP) and the World Weather Research Programme (WWRP) jointly initiated the S2S Prediction Project (`http://www.s2sprediction.net/`, Vitart et al. (2015)). In the spirit of seamless prediction, this project is aimed at bringing together the weather forecasting and climate prediction communities to advance the understanding of predictability on subseasonal time scales. A number of different challenges are identified and specifically addressed in sub-projects, and a comprehensive overview of the advances and challenges in S2S forecasting has been published recently by Robertson and Vitart (2019). Another essential part of this project is the coordination of a publicly available database of subseasonal forecasts (Vitart et al., 2017), which we make extensive use of for the work in this thesis. This database consists of operational forecasts and corresponding retrospective forecasts of 11 different modelling centers around the world and is hosted at the China Meteorological Administration (CMA), the European Centre for Medium-Range Weather Forecasts (ECMWF), and the International Research Institute for Climate and Society (IRI).

1.4 Forecast Verification

Since all analyses in this thesis involve the verification of subseasonal ensemble forecasts, we briefly introduce the main concepts of (probabilistic) forecast verification. Apart from the main attributes characterizing the quality of a forecast, we discuss the concepts of value and skill, and how they relate to each other given the verification metrics that we use in this thesis. Note that in the title of this section, the word "verification" describes the process of verifying the forecasts. In the remainder of this section, we also refer to the observations that the forecasts are verified against as the "verification". The meaning in each case should be unambiguous from the context.

1.4.1 Forecast Attributes

Forecast verification aims at quantitatively assessing the quality of a forecast. This requires a comparison of the forecasts with a verification[6]. More specifically, in forecast verification we make inferences about the quality of the forecasts from the statistics of the joint

[6] This verification represents the observed evolution. In weather and climate prediction, observational or reanalysis data most commonly serve this purpose. However, these are also subject to errors in the measurements and/or assimilation of the data. The effect of these uncertainties on the process of forecast verification is not further addressed here.

forecast-verification distribution. This is the bivariate distribution defined by the pairs of predicted and observed values. Several attributes of forecast quality can be identified, and all of them are related to this distribution and its factorizations (Potts, 2012). We describe a selection of forecast attributes based on the list given by Wilks (2019a, Chapter 9) in the following. Our descriptions of these attributes are made mostly from a standpoint of probabilistic forecasts of a binary predictand ("event/no event") but the concepts can be generalized to any type of forecast and predictand.

Let this predictand indicate whether — at some hypothetical location with a stationary climate — a heat wave occurred or not. Let us assume that the threshold for whether there is a heat wave or not is 23°C and that this threshold corresponds to the 75th percentile of the climatological temperature distribution. Exceeding this threshold is therefore expected to happen 25% of the time (also called the base rate b) over a sufficiently larger number N of instances. For each of these N instances, our hypothetical probabilistic forecast system predicts a probability p_i for the occurrence of a heat wave between 0 and 100% with $i \in \{1, 2, 3, 4, 5\}$[7]. In Figure 1.6, different hypothetical forecast systems are depicted with different colors in the diagrams. In these so-called reliability diagrams (Wilks, 2019a) the observed frequency of occurrence f is displayed as a function of the predicted probabilities p_i. The diagrams illustrate some of the attributes described below.

Bias The (unconditional) bias of a forecast system refers to the deviation of the average forecast event frequency from the observed base rate. If our hypothetical system predicts a heat wave in 35% of the cases averaged over the verification period, it has an unconditional (frequency) bias since the base rate of the event is only 25%. In Figure 1.6a the orange dots illustrate a system that is unconditionally biased because for any probability p_i it predicts, the event in reality occurs less often. It is said to overforecast the event, corresponding to a warm bias if the event was a heat wave. The blue dots show a system without unconditional bias because on average over all forecasts, it predicts the correct event frequency (a line fit to the blue dots intersects the point $f(p = b) = b$). A good quality forecast system should have a low unconditional bias. An unconditional bias can be corrected. In the example from above it could be corrected by using the 75th percentile of the *forecast* temperature distribution as a threshold value instead of the absolute value of 23°C.

Reliability Reliability (also referred to as calibration) refers to the consistency between predicted probabilities and observed frequencies. If our hypothetical forecast system is reliable, for any probability p_i it predicts, the event on average occurred with the same frequency, i.e. $f(p_i) = p_i$. A reliable forecast system is illustrated by the orange dots in Figure 1.6b. In case of an unreliable system, it could be that whenever a probability of, say, 75% is predicted by the forecasts, a heat wave in reality occurred in only 40% of the cases. A forecast system that deviates from perfect reliability is shown by the blue

[7] A fixed number of 5 is only chosen for clarity. In practice, the maximum number of probability categories, C_{max}, is only limited by the number of ensemble members M, i.e. $C_{max} = M + 1$.

dots in Figure 1.6b. The reliability of a forecast system can be partially corrected for by calibrating the forecast probabilities to the observed frequencies. For instance, if we know from past experience that a forecast probability of $p_i = 75\%$ corresponds to an observed frequency of $f(p_i) = 40\%$, every time the forecast predicts 75%, we could predict the observed frequency of 40% instead. However, this requires knowledge about the past performance of the forecast. Note that forecasting the observed event base rate at any forecast instance (forecasting "climatology") has perfect reliability but no sharpness (see below).

Resolution Resolution describes how much the observed event frequencies $f(p_i)$ differ between different predicted probabilities p_i. For instance, if the system has resolution (as the orange and blue forecasts in Figure 1.6b), the observed average frequency $f(p_4)$ for a predicted chance $p_4 = 75\%$ of a heat wave will be different from $f(p_2)$ for a predicted chance $p_2 = 25\%$. If the system has no resolution at all (green forecasts in Figure 1.6b), the observed frequencies are the same regardless of the predicted probability, i.e. $f(p_i)$ is the same for all i. Resolution is a desirable property of a forecasting system since it means that the forecasts are able to sort the observed events into bins with different observed frequencies.

Discrimination Discrimination describes the system's ability to provide different forecasts for different outcomes. In the example, there are only two outcomes (heat wave or no heat wave). If a system can discriminate between these events, it should on average predict different probabilities depending on the observed outcome. A system that forecasts on average a 25% chance of a heat wave occurring, both for all cases when a heat wave was observed and when it was not, has no discrimination. A system that has discrimination on the other hand would forecast an average chance higher than 25% in cases when a heat wave happened and an average chance lower than 25% in those when it did not. Discrimination is complementary to resolution but cannot be seen in the diagrams of Figure 1.6 since this would require sorting the forecasts by whether an event occurred or not, which is opposite to the sorting applied for the reliability diagrams.

Sharpness Sharpness is a property of the forecasts alone and is independent of the verification. This attribute refers to the ability of the system to produce forecast distributions that are different from its own climatological distribution. Assume our hypothetical forecast system has no frequency bias and thus, on average, predicts a heat wave base rate of 25%. If the system has some sharpness, the forecast probability for different forecast instances will vary between 0 and 100% (green forecasts in 1.6c). In the most extreme case it will only forecast 0 or 100% (orange dots in 1.6c). In the case of no sharpness, it will forecast a probability of 25% for any given forecast instance (blue dot in 1.6c), i.e. a climatological forecast has no sharpness (but is perfectly reliable, see above). Since this attribute is independent of the verification, sharpness is a necessary but not a sufficient condition for a forecast to have good quality.

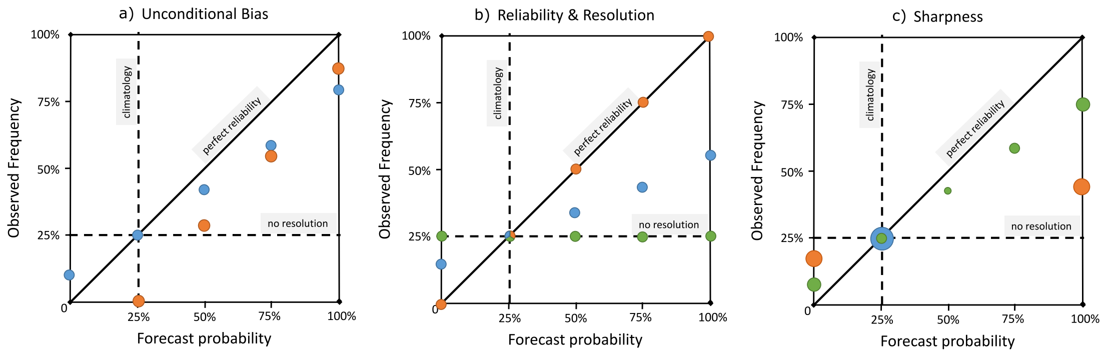

FIGURE 1.6: Illustration of selected attributes of a forecasting system discussed in the text. In the shown reliability diagrams, the forecast probabilities p_i are on the x-axis. The frequency of occurrence $f(p_i)$ of an event when the forecast predicted one of these probabilities p_i is on the y-axis. The size of a dot indicates how many of the N forecasts in the forecast period predicted a probability p_i (larger dots meaning more forecasts). Different colors indicate different forecasting systems. In a), the orange forecasts have an unconditional bias, the blue forecasts have no unconditional bias. In b), the orange forecasts are perfectly reliable, the green forecasts have no resolution. The blue forecasts show a system that deviates from perfect reliability but does have some resolution. In c) the concept of sharpness is shown. The blue forecasts always predict a probability of 25%, which is the base rate of the event. Though they are perfectly reliable, they have no sharpness. The orange forecasts only predict probabilities of 0 and 100%, thus they have infinite sharpness. The green system has some sharpness. It provides probabilities of 0 and 100% more often than intermediate probabilities.

Which of these aspects is the most important to measure depends on the requirements of the forecast user. However, for any user that only has the information of the forecast available and has to take it at face value, the most critical property is its reliability (Weisheimer and Palmer, 2014; Palmer, 2019). A forecast that is not reliable can, in the worst case, mislead the user such that it would be preferable for her to take no action at all rather than base a decision on the forecast.

1.4.2 Value

The usefulness of a forecast to a user is described by its value. The value of a forecast strongly depends on the requirements of the user. In many cases it is useful to estimate the economic value of a forecast by considering a hypothetical user that faces a decision depending on the occurrence of a certain weather event (Richardson, 2012). The user can take preventive measures to mitigate the consequences of the event for a cost C. If the event occurs and the user does not take any measures, she faces a loss L. The value is defined as the cost for taking measures based on the forecast relative to the cost for taking measures based on a hypothetical perfect forecast (that knows with certainty when to take action). In the case of a low cost/loss-ratio (C/L) it makes sense for the user to take measures even when the forecast probability for the occurrence of the event is low. A user with a high C/L on the other hand would only take measures if the forecast probability was high. A natural choice for a decision threshold p for any user would be $p = \frac{C}{L}$ (Palmer and Richardson, 2014). If the forecasts are reliable, this approach maximizes the value for the forecast user (Richardson, 2012). This illustrates the value of probabilistic forecasts and their advantage over deterministic forecasts. Deterministic forecasts do not account for the forecast uncertainty and, in this example, would only give probabilities of either 0 or 100% depending on an arbitrary threshold (e.g. a heat wave is forecast when the deterministic forecast exceeds a temperature T). The optimal threshold value will vary depending on the C/L of the user. Probabilistic forecasts provide an optimal decision threshold for any C/L, and thus have higher value than deterministic forecasts from an equivalent forecast model (Palmer and Richardson, 2014).

The value of a forecast can be defined in different ways. Dorrington et al. (2020) for instance evaluate the financial value of subseasonal forecasts based on their ability to inform about the best trading strategy for electricity in France. They demonstrate that assessment of the value of long-range predictions based only on a cost/loss-ratio as described above can underestimate the actual value of the forecasts. Different strategies for assessment of the forecast value will be appropriate in different contexts (see also Lopez et al., 2020) but we do not elaborate further on these alternative methods.

1.4.3 Skill

A common approach to assessing the quality of a forecast is by considering its skill relative to some reference system, e.g. a reference that predicts the base rate of an event for any forecasting instance ("climatology forecast"). A skill evaluation is done by computing a score, which summarizes one or more of the attributes mentioned above. The score is computed for the forecasting system (S) and the reference (S_{ref}), and their difference is considered relative to the score of a perfect forecast (S_{perf}) to form the skill score SKS (Wilks, 2019a):

$$\text{SKS} = \frac{S - S_{ref}}{S_{perf} - S_{ref}} \tag{1.1}$$

A skill score of this form is thus a normalized measure of the forecast performance and can attain values between $-\infty$ and 1. A skill score of 0 would indicate a forecast that does not perform any better than a reference forecast, while positive (negative) skill scores are obtained for a forecast that performs better (worse) than the reference. The maximum skill score of SKS = 1 is attained when $S = S_{perf}$.

In the following, we describe the main type of score (and the associated skill scores) that we use in this thesis. For all of these scores our forecasts are in the form of probabilities, which we estimate from the ensemble system. Assume again that we have a binary predictand (1 or 0 at any given day, e.g. referring to "heat wave" or "no heat wave", respectively) and a forecast of probabilities of these events happening. For illustrative purposes, let us assume only one forecast instance n. In this instance, the forecast predicts a 60% chance of a heat wave happening (implying that it forecasts a probability of 40% of no heat wave occurring) and in fact, for this forecast instance n a heat wave occurred. We can measure the error of the probabilistic forecast with the Brier Score (BS):

$$\text{BS} = \frac{1}{N}\sum_{n=1}^{N}(O_n - Y_n)^2 \tag{1.2}$$

where O_n is the predictand (1 in this case) and Y_n the probability predicted by the forecast (0.6) for forecast instance n. For our example ($N = 1$), this results in $\text{BS} = (1 - 0.6)^2 = 0.16$. A perfect forecast would predict the event with certainty in this instance and the optimal score is thus $\text{BS}_{perf} = 0$. Taking a heat wave base rate of 25% as above, we can also compute the BS of a reference that always forecasts the climatological event frequency. For the instance when a heat wave occurred, its score is $\text{BS}_{ref} = (1 - 0.25)^2 = 0.5625$ and thus the error of this reference forecast is substantially higher than that of the forecast. Using the BS as the score S in Equation 1.1 we get a Brier skill score of BSS = 0.72. Note that usually, the scores are averaged over a number $N \gg 1$ of forecast instances.

The BS is a measure of the squared probabilistic error and essentially measures the average squared difference between the observed and the forecast's (discrete) cumulative density function (cdf). This concept can be generalized to analogously define scores for different types of predictands. If the predictand is a categorical variable with $K > 2$ categories (for instance, temperature falling into the upper, middle or lower tercile of the temperature distribution), a score analogous to the BS can be defined by computing the BS for binary forecasts of each of the $K - 1$ categories and taking their sum over all k (weighted by the base rate of the respective event in category k, Bradley and Schwartz, 2011). This score is called the ranked probability score (RPS) and its respective skill score is the RPSS.

The BS can be further generalized to the case of infinitely many categories or, equivalently, continuous forecast distributions by taking the integral of the squared difference between the cdf of the forecast ($F(y)$) and the cdf of the verification ($F_o(y)$) over all possible values of the forecast variable y. This score is called the continuous ranked probability score (CRPS) and is defined as (Wilks, 2019a):

$$\text{CRPS} = \int_{-\infty}^{\infty}[F(y) - F_o(y)]^2 dy \tag{1.3}$$

Its respective skill score is the CRPSS. This score can be easily estimated for ensemble systems, which only provide a finite sample of the actual forecast cdf $F(y)$ (e.g. Hersbach, 2000).

We mostly focus our attention on these scores in this thesis since their respective skill scores have been shown to further be summary measures of forecast value. For instance, the BSS can be interpreted as the average value of a forecasting system over a distribution of users uniformly spread over all cost/loss ratios between 0 and 1 (Richardson, 2012). The CRPSS generalizes this and can be considered an average value over all users and all possible events (Palmer and Richardson, 2014).

As mentioned above, to characterize the quality of the forecasts, their scores are considered relative to a reference. A forecast is said to have skill (over the reference) if it attains a better score than the reference. The choice of the reference is of great importance for the evaluation of whether skill is attributed to a forecast (Wilks, 2019a; Hamill and Juras, 2006). Generally, the reference is chosen to be a benchmark forecast that we would not ascribe any skill to and that our forecast should outperform. In most cases this means using a reference that can be more or less trivially constructed. In weather forecasting, persistence is a competitive benchmark. In this reference, the state of the initial conditions is predicted for any given lead time (the initial conditions could also be gradually relaxed towards climatological conditions with increasing lead time). In long-range forecasting, a better benchmark is usually climatology. In a probabilistic sense, this reference forecasts the climatological distribution for any given forecast instance and thus has no resolution but perfect reliability (see Section 1.4.1). Usually, the climatological distribution is estimated from past data. In the estimation, it is essential to account for the differences in the climatologies of sub-samples of the forecast-verification pairs in order to not overestimate the skill (Hamill and Juras, 2006). In subseasonal prediction for instance, the climatological seasonal cycle needs to be accurately represented (Manrique-Suñén et al., 2020). The details of the estimation of the reference climatology are of major concern in the analyses of Chapter 3 where we show how non-stationary components in the climatology of the prediction system can inflate the skill of subseasonal forecasts.

Using an informative measure of skill is critical in communicating the quality of a prediction system. Skill scores such as the BSS, RPSS and CRPSS summarize the main attributes of a forecast system and additionally can be interpreted as measures of the system's value averaged over all users (and decision thresholds in case of the CRPSS). The reference used to assess the skill is of great importance. In subseasonal prediction, this reference is commonly a climatology forecast. While mis-estimations of the climatology can lead to inflated skill estimates, skill can also be under-estimated when using an unfair reference (Weigel et al., 2007; Ferro et al., 2008). This is especially an issue for subseasonal forecasts where skill is generally low and the potential value of the forecasts should not be further masked by using an unrealistic benchmark. Only when these aspects are correctly accounted for we can inform about when and where a forecasting system performs better or worse. These concepts will be important when we analyze the seasonal cycle of subseasonal skill in Chapter 4.

1.5 Probabilistic Subseasonal Forecast Skill in an Operational Prediction System

Having reviewed the theoretical background of subseasonal predictability and relevant concepts in the verification of probabilistic forecasts, we now briefly consider the skill of an actual forecasting system. Note that throughout this thesis we refer to subseasonal forecasts as forecasts of some temporally averaged variable verified at lead times greater than 10 and up to 30 days. For the temporal averaging we mainly use periods of 5 (Chapters 2 and 4) and 7 (Chapter 3) days. Our main variable of interest is near-surface temperature. The reasons for this are that near-surface variables are the most relevant for forecast users and that out of the relevant surface variables, temperature shows the largest forecast skill. This is related to the fact that all sources of subseasonal forecast skill summarized in Section 1.2 can control near-surface temperature variability to some degree.

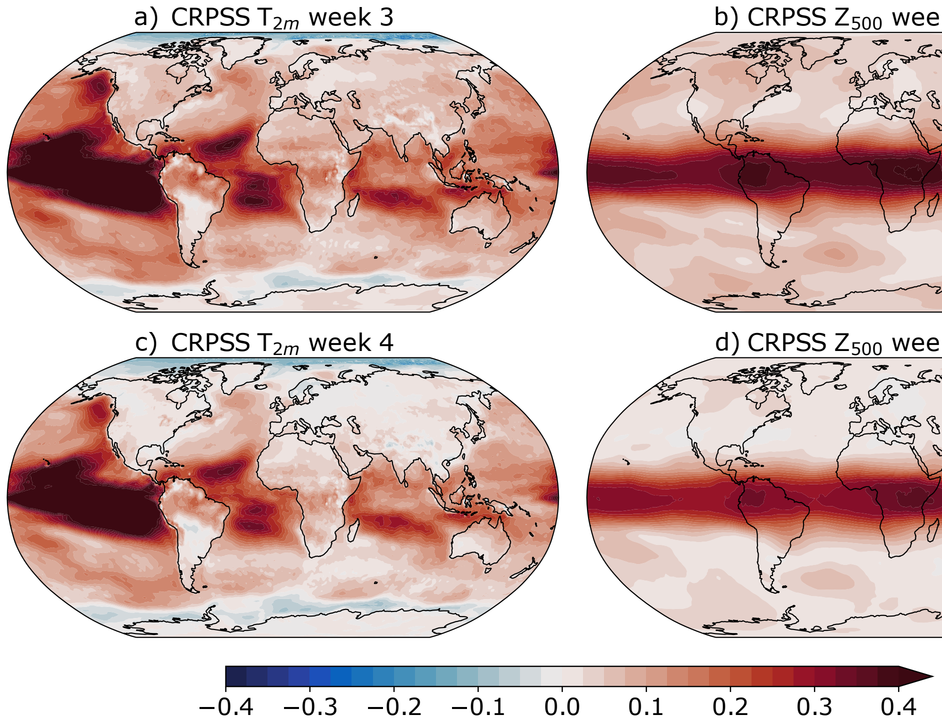

FIGURE 1.7: Global probabilistic skill at subseasonal lead times. Shading shows the week-3 (a and b) and week-4 (c and d) CRPSS of 7-day averages of standardized 2-m temperature (a and c) and 500hPa geopotential (b and d) anomalies for hindcasts from the ECMWF extended-range forecasting system verified against ERA-Interim over all initializations in the 20 year period from 1998 – 2017 (105 initializations per year). The reference forecast for computing the CRPSS is a forecast predicting the climatological distribution (a standard normal distribution since both variables were standardized) at every instance.

We now look at twice-weekly hindcasts of 7-day means of 2-m land temperatures (T2m) and geopotential height at 500 hPa (Z500) from the extended-range ensemble hindcasts from the ECMWF. A hindcast (or re-forecast) is a forecast started from initial conditions that lie in the past and that can immediately be verified against a verification. In operational forecasting, hindcasts are mainly used for calibration of the real-time forecasts. Here, we use these hindcasts and verify them against the ERA-Interim reanalysis (Dee et al., 2011) with the CRPSS over all initialization dates in the period 1998 to 2017. Both hindcasts and verification were standardized using a low-pass filtered seasonal cycle of the mean and standard deviation, and were subsequently detrended. Both procedures are performed for each lead time separately. As a consequence, the reference climatological distribution can be reduced to a standard normal distribution. The reasons for these processing steps will be become clear in Chapters 3 and 4. The CRPSS has further been adjusted according to Leutbecher (2019) to account for the effects of the limited ensemble size (11 members). The skill is shown in Figure 1.7 for hindcasts with lead times of 3 and 4 weeks (i.e. verifying averages over days 15 – 21 and days 22 – 28 of the hindcasts, respectively). We can readily see that the general pattern of the skill is similar to the deterministic skill shown in Figure 1.2 albeit with much lower overall skill. Indeed, temperature skill even becomes negative for many parts of the Arctic Ocean and the Southern Ocean around Antarctica. Note however, that we have not applied any sophisticated calibration to the hindcasts. Doing so could enhance the reliability and thus the skill. It should also be noted here that the generally lower magnitude of probabilistic skill compared the deterministic skill is by no means a sign for a better performance of the deterministic forecasts. Rather, the correlation skill score shown in Figure 1.2 is a measure of the potential skill since it neglects any unconditional bias and lack of reliability in the forecast distribution. Figure 1.7 confirms that probabilistic prediction skill is much lower in the extratropics than in the tropics. This is the case for both T2m and Z500. The reason for this is the strong atmosphere-ocean coupling in the tropics, especially in the main area of action of the ENSO, the tropical Pacific, where T2m prediction skill is extremely high, exceeding values of 0.4 in large parts of the Pacific. In the extratropics on the other hand, the skill for either variable rarely exceeds 0.1, although there are some prominent exceptions to this over the eastern Pacific. The skill over the continents is consistently lower than over the oceans for T2m, indicating an influence of the underlying ocean conditions. Over Europe in particular, the skill is very close to 0 but still slightly positive. This highlights the fact that this region is particularly challenging to forecast, but due to the aforementioned sources of subseasonal skill there is also large potential for improving predictions in this region.

1.5.1 Near-Surface Temperature Skill over Europe Illustrated Using Reliability Diagrams

To focus more closely on Europe, we next look at the probabilistic subseasonal skill for binary forecasts of European average temperatures. We first average 7-day mean temperatures over a large European region (EUR, 10°W – 50°E, 35°N – 75°N) and then standardize using the same approach as described above. The verifications are then sorted according to whether they lie above or below the median of the climatological temperature distribution, thus transforming them to a binary predictand. The ensemble forecasts are transformed into probabilities by using the fraction of ensemble members for each initialization that predict temperatures above or below the median over the entire period (all seasons). This allows for a direct computation of the BSS described in Section 1.4.3. The reference forecast is climatology, which forecasts a chance of 50% of exceeding the median at each forecast instance. The BSS is further adjusted according to Weigel et al. (2007) to account for the limited ensemble size (11 members). While the BSS is low but still positive for week 3 of the forecast (BSS = 0.06), it drops to 0 and even slightly below in weeks 4, 5 and 6. The skill for the spatially averaged temperatures thus behaves similarly as the grid-point based skill over Europe in that it is only slightly positive for a lead time of 3 weeks and almost indistinguishable from zero at longer leads.

We now examine the skill of the probabilistic forecasts of European averages of temperature further by showing them in reliability diagrams (e.g. Weisheimer and Palmer, 2014; Wilks, 2019a, and Section 1.4.1). In the reliability diagram, for each forecast probability bin p_i (here, 12 bins between 0 and 100% for a forecast ensemble with 11 members)

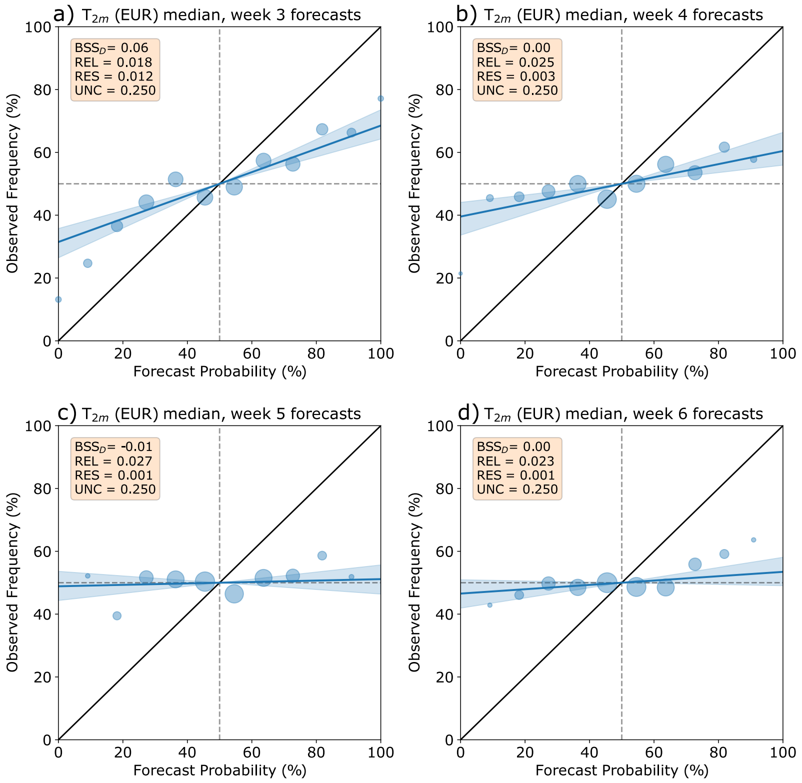

FIGURE 1.8: Reliability diagrams of binary forecasts of T2m averages over the EUR region at forecast weeks 3 – 6. The forecasts are for temperatures lying above or below the median. The forecasts are the same as used in Figure 1.7. See text for details.

we compute the frequency of event occurrence when a forecast predicted a probability p_i and plot this frequency for all values of p_i. The number of forecasts in a probability bin is illustrated by the size of the point. By plotting the forecast-verification pairs in this way, all attributes of the forecast system described in Section 1.4.1 except for the discrimination can be directly recognized from the reliability diagram (it is thus also referred to as the attributes diagram). The sharpness of the forecast can be seen by the size of the points. The larger the points of the low and high probability bins are, the more sharpness it has. A forecast with little sharpness would have the largest points in the central probability bins. In terms of reliability, deviations of the points from the diagonal indicate imperfect reliability. For a forecast to have resolution on the other hand, the points must deviate from the horizontal (grey dashed line). Note that a climatological forecast would be represented by a single point at the intersection of the two grey dashed lines which indicate the event base rate (50% for median events). Though perfectly reliable, it has neither sharpness nor resolution. We can further fit a line to the points using a weighted linear regression (blue line in Figure 1.8) and quantify the uncertainty in this estimate (blue shading in Figure 1.8 showing the 95% uncertainty interval). This line should intersect the diagonal at the base rate unless the forecasts have an unconditional bias.

For the hindcasts considered here, we see that both the resolution and the reliability become worse for longer lead times, deduced from the fact that the fitted line moves from lying closer to the diagonal to lying closer to the horizontal. We can also see that for forecast week 3, the forecasts are more evenly spread over the probability bins while they are somewhat more concentrated in the central bins for longer lead times. This indicates that also the sharpness of the forecasts decreases. The reliability and resolution can also be directly computed (e.g. Hersbach, 2000) and their values REL and RES are shown in the orange boxes in the panels of Figure 1.8. Note that while a high value of RES indicates good resolution, REL is negatively oriented and its optimal value is 0, i.e. a forecast with less than perfect reliability shows non-zero, positive values of REL. The values confirm the notion we got from considering the lines in the diagram. While REL increases, RES decreases for longer lead times. This shows again that temperature skill is low over Europe but the non-zero reliability also reveals that there is potential for improving the skill by calibrating the forecasts, i.e. improving their reliability component. Note however, that calibration always comes at the cost of reduced sharpness.

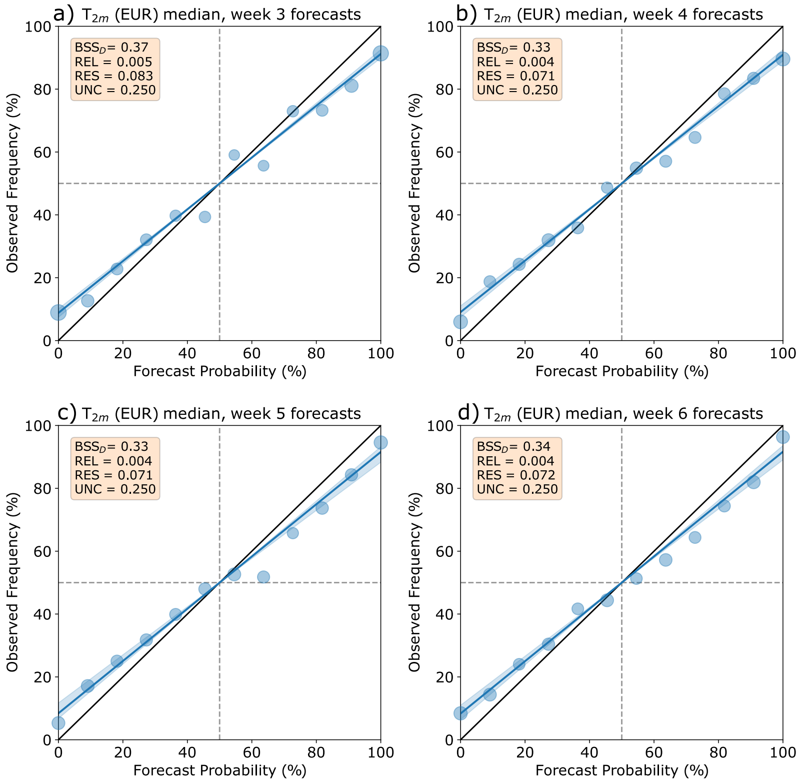

FIGURE 1.9: The effect of mis-estimating the climatology. Reliability diagrams as in Figure 1.8 but transforming the ensemble forecasts into probabilities without accounting for the seasonal cycle. See text for details.

The effect of incorrectly accounting for the climatology can be nicely illustrated in the reliability diagrams by extending the above example. Note that above, by eliminating the seasonal cycle from the hindcasts first, we have ensured that the actual probability of the (standardized) temperatures to lie above the median is close to 50% for any forecasting instance. If we repeat the above analysis but use absolute temperatures and transform the forecasts into probabilities based on the median over the *entire* period, we get the reliability diagrams in Figure 1.9. These forecasts have skill up to 0.34, even at lead times of 6 weeks, which might be expected for a tropical region but certainly not for Europe. The reason of course is that by using the median of the absolute temperatures over the entire verification period we do not account for the seasonal cycle. Thus, summer (winter) temperatures all lie above (below) the median and the model can correctly predict this seasonal cycle even without having any skill at predicting the actual subseasonal variability. While this admittedly naive approach clearly exaggerates the effect that would result from more subtle mis-estimations, it nicely illustrates the issue of falsely attributing skill to a system due to a false representation of its climatology.

1.6 Objectives and Outline of the Thesis

The current subseasonal prediction skill over the European continent is low, but skillful probabilistic forecasts on this time scale would be beneficial to a multitude of users in different sectors. The general objective of this thesis is to contribute to the understanding of how predictability and prediction skill for temperatures in Europe vary.

Due to their potentially harmful impacts, extreme events in particular are of concern in subseasonal forecasting and consequently, many studies have focused on extremes. It is often assumed that these large amplitude events are a result of large signals in their predictors. From a perspective of predictability however, there is not much justification for considering the prediction of extreme events rather than the prediction of more average conditions. One main question we would like to answer in this thesis is how the prediction skill for extreme temperature events over Europe differs from the skill for average conditions. We address this in Chapter 2 where we consider extreme temperature events in the summer season. We further answer the question of where in Europe summer heat extremes are better predicted at subseasonal lead times than average conditions and provide a comparison of the skill of four different models. We further address the role of persistence in the predictability of extreme events in Chapter 2.

As discussed in Section 1.4, the choice of a reference forecast for assessing the skill of a prediction system can have a significant effect on the estimated skill. In order to correctly report the skill, it is essential that the reference forecast accounts for all differences in the climatology of sub-samples of the forecast, which — as we saw in Section 1.5.1— implies the need to account for the seasonal cycle. A trend is another non-stationary component of the climate and it is often neglected in the reference for subseasonal forecasts. However, trends in relation to climate change can account for significant parts of the variability over a typical climatological period, particularly for near-surface temperatures. The second main objective of this thesis is therefore to quantify the effect of a long-term trend in the climatology on the estimates of subseasonal forecast skill. In Chapter 3 we approach this question in an idealized framework by using synthetic pairs of forecasts and verifications to conceptually explain the effect and deduce a benchmark estimate of the possible influence. As this framework strongly simplifies the properties of a real prediction system, we further assess the strength of the effect of an unaccounted trend in the climatology in an operational subseasonal forecast system. The results of this chapter can be generalized to some degree to the verification of forecasts on any time scale.

In Section 1.2 we discuss the sources of subseasonal predictability. Many of these have more potential in a certain season than in others. It is thus a third objective of this thesis to characterize the seasonal cycle of subseasonal European temperature skill. We address this objective in Chapter 4. Eliminating all seasonal variability in the temperatures that could cause differences in the prediction skill allows us to estimate the skill's month-to-month variations. This can inform users of subseasonal predictions about the time of the year when a forecast might be more useful. Our analysis further allows for the identification of some predictive processes in the months with highest skill. We use indicators of these processes during the initialization of the forecasts to identify windows of opportunity for more skillful predictions.

We conclude this thesis in Chapter 5 by summarizing all the results in the context of the aforementioned objectives and assess their implications for the understanding of subseasonal predictability and skill of European near-surface temperatures. Finally, we give an outlook for future avenues that could be taken to further advance our understanding of subseasonal predictability.

Chapter 2

Higher Subseasonal Predictability of Extreme Hot European Summer Temperatures as Compared to Average Summers

This chapter has been published as an article in Geophysical Research Letters:

Wulff, C. O., & Domeisen, D. I. V. (2019). Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophysical Research Letters*, 46, 11520 – 11529, `https://doi.org/10.1029/2019GL084314` ©2019. American Geophysical Union. Used with permission.

The Supporting Information for this chapter can be found in Appendix A.

Abstract

Summer temperatures in the last decades were increasingly characterized by persistent extremes, and there is evidence that this trend will continue in a warming climate. The exact timing of these extremes is less well known and it is therefore crucial to consider their subseasonal predictability. We compare the prediction of summer 2m-temperature extremes in Europe with the prediction of average events for four subseasonal forecasting systems. We find higher prediction skill for warm extremes as compared to average events, with some regional dependence. The same is not true for cold extremes, indicating an asymmetry in the processes causing opposite summer temperature extremes. The forecast skill is strongly increased by the most severe and persistent events in the analyzed period. We hypothesize that the enhanced warm extreme skill is related to persistent flow patterns and land-atmosphere interaction. This could have implications for potentially enhanced predictability in a warming climate.

2.1 Introduction

Recent years have seen an increased number of extreme heat waves across the Northern Hemisphere, e.g. over central Europe in 2003 (Schär and Jendritzky, 2004; Trigo et al., 2005; Fink et al., 2004; Fouillet et al., 2006), over Russia in 2010 (Dole et al., 2011; Barriopedro et al., 2011), and over Europe in 2018 (Schiermeier, 2018). It has been evident for decades (IPCC, 1990) that extreme events such as heat waves are exacerbated by climate change (Stott et al., 2004; Stott et al., 2015; Coumou and Rahmstorf, 2012; Sippel et al., 2016; Diffenbaugh et al., 2017; Mann et al., 2017; Imada et al., 2018). The time needed to prepare for an extreme event is often beyond the skillful prediction timescales of a few days that are currently available (White et al., 2017). The observed and projected increase in strength and frequency of heat waves therefore calls for reliable predictions on timescales of weeks to months. Currently available models cannot predict the onset, duration, or amplitude of a heat wave on subseasonal timescales, as e.g. for the 2010 Russian heat wave (Quandt et al., 2016).

On subseasonal timescales several potential predictors of summer near-surface temperatures have been identified. For instance, predictability could stem from persistent atmospheric flow patterns as a large fraction of the observed summer heat extremes are associated with atmospheric blocking (Pfahl and Wernli, 2012; Schaller et al., 2018; Brunner et al., 2018; Sousa et al., 2018). Further, warm and cold extremes can be related to the presence of upper-tropospheric Rossby wave packets (RWPs, Fragkoulidis et al. (2018)) and a regional stalling of the jet stream (Röthlisberger et al., 2016). While RWPs are not generally more predictable, 500 hPa geopotential height forecasts initialized in the presence of specific types of RWPs display enhanced skill up to week 3 of the forecast (Grazzini and Vitart, 2015). Furthermore, regimes favouring local persistent temperature anomalies are also influenced by modes of low-frequency and possibly remotely forced variability like the summer North Atlantic Oscillation (Folland et al., 2009; Ossó et al., 2017) or the summer East Atlantic pattern (Wulff et al., 2017; Neddermann et al., 2019). This large-scale control on surface temperatures has the potential to enhance their predictability beyond the typical weather forecasting timescales.

Other studies highlight the potential of land-atmosphere interactions for seasonal (Weisheimer et al., 2011; Prodhomme et al., 2016; Ardilouze et al., 2017a; Bunzel et al., 2018) and subseasonal prediction (Koster et al., 2010; Ardilouze et al., 2017b) by showing that temperature forecasts benefit from a realistic initialization of the land surface. Land-atmosphere interactions were especially important for the heat waves of 2003 in Europe (Ferranti and Viterbo, 2006; Fischer et al., 2007b) and 2010 in Russia (Miralles et al., 2014; Hauser et al., 2015). During these events the successive drying of the soil under the persistent atmospheric forcing led soil moisture to drop below a critical value, triggering a positive feedback between soil dryness and near surface temperatures (Seneviratne et al., 2010).

Many studies consider predictors of warm summer temperatures and heat waves only (e.g. Cassou et al. (2005) and Ardilouze et al. (2017b)), but the aforementioned mechanisms do not work equally for hot and cold temperature extremes and thus their predictability could also differ (Quesada et al., 2012).

Motivated by the importance of predicting summer extreme events on subseasonal timescales, we test if near-surface extreme temperatures are more predictable than average temperatures. We evaluate the hindcast skill of different subseasonal forecasting systems for the near-surface temperature evolution on timescales of several weeks in Europe and specifically focus on the comparison between the skill of predicting warm and cold extremes in comparison to average temperatures. The definition of these events, the metrics applied to verify the hindcasts and the data used are described in Section 2.2. The model skill is described in Section 2.3, which also shows the comparison of the skill for the different event types. Sensitivity analyses are shown in the Supporting Information (SI, Appendix A). The results are summarized and discussed in Section 2.4.

2.2 Data and Methods

2.2.1 Hindcasts and Verification Data

We consider hindcasts from four subseasonal forecasting systems provided by the Australian Bureau of Meteorology (BoM), the Chinese Meteorological Agency (CMA), the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP). All data are made available through the subseasonal to seasonal (S2S) prediction project (Vitart et al., 2017). The systems differ in many aspects of their forecasting strategies, which are described in detail in the SI (Table A.1). In this study, we focus mainly on the ECMWF system, which is initialized twice per week. The hindcasts were verified using the ECMWF's Interim reanalysis (ERA-Interim, Dee et al. (2011)).

The data considered in this study are summer (June, July, August) 2-m temperature anomalies (T_{2m}) in the 12-year period 1999 – 2010, which is the longest common period that hindcasts from all four systems cover. Anomalies were computed with respect to each forecasting model's lead time dependent climatological seasonal cycle, which was computed from the first four harmonics of the daily 1999 – 2010 climatology. The lead time dependence of the climatology removes effects of drift in the models' climatologies with increasing lead time. Furthermore, as forecasts on subseasonal lead times are not expected to reproduce small-scale day-to-day variability, we applied a 5-day moving average to the daily temperature anomalies (T_{2m}^{5d}). Additionally, we averaged T_{2m}^{5d} over six regions in Europe (Figure 2.1b): Scandinavia (SC), Western and Eastern Europe (WEU and EEU, respectively), Russia and Ukraine (RUK), and the Western and Eastern Mediterranean (WMED and EMED, respectively). All analyses are based on hindcasts of the averages of T_{2m}^{5d} over one of these regions R ($\langle T_{2m}^{5d} \rangle_R$), where $\langle . \rangle$ indicates the spatial average using area weighting.

As a reference for the forecast skill, we further designed persistence hindcasts with a set-up mimicking the dynamical hindcasts. These were created by keeping $\langle T_{2m}^{5d} \rangle_R$ computed from ERA-Interim at each initialization date of the BoM system constant for 62 days (length of the BoM hindcasts). Note that both for the persistence forecast and for the verification we use anomalies with respect to the seasonal cycle.

2.2.2 Evaluation of Skill

The skill at reproducing the observed patterns of T_{2m}^{5d} is evaluated using the centered Anomaly Correlation Coefficient (ACC). The ACC for a given initialization time i is defined following Wilks (2011, Chapter 8) as:

$$\mathrm{ACC}_i = \frac{\sum_{k=1}^{K} (y'_{ik} - \langle y' \rangle)(o'_{ik} - \langle o' \rangle)}{\sqrt{\sum_{k=1}^{K} (y'_{ik} - \langle y' \rangle)^2 \sum_{k=1}^{K} (o'_{ik} - \langle o' \rangle)^2}} \tag{2.1}$$

where k is an index for the grid point, y indicates the forecast and o the reanalysis. $\langle . \rangle$ indicates averaging over all K grid points within the chosen region, primes indicate anomalies with respect to the climatology.

When analyzing the skill of the hindcasts for a certain event type, we treat each individual ensemble member's hindcast as a deterministic binary forecast. We define an average event as a day on which $\langle T_{2m}^{5d} \rangle_R$ lies between the monthly 25th and 75th percentile of the distribution. An extreme warm (cold) event is detected when $\langle T_{2m}^{5d} \rangle_R$ exceeds (falls below) the 95th (5th) percentile. Thus, extreme (average) events in our analyses have a base rate of $p_x = 5\%$ ($p_a = 50\%$) Note that by using percentiles of the distribution of anomalies we eliminate contributions to the forecast skill resulting from a successful reproduction of the seasonal cycle. Furthermore, by defining an event based on the model's own climatological distribution, we eliminate the frequency bias in the hindcasts, i.e. each system's hindcast ensemble predicts an extreme (average) event on 5% (50%) of the considered days at any given lead time by design.

In order to be able to compare the skill for events with different base rates, we apply a skill measure that is base rate independent and does not degenerate when the base rate decreases (Hogan and Mason, 2012). The extremal dependence index (EDI, Ferro and Stephenson (2011)) fulfills these requirements and is defined as:

$$\mathrm{EDI} = \frac{\log F - \log H}{\log F + \log H} \tag{2.2}$$

where H is the hit rate, i.e. the number of hits divided by the number of observed events, and F the false alarm rate, i.e. the number of false alarms divided by the number of observed non-events. The EDI varies between -1 and 1 where 0 indicates no skill and 1 is the skill of a perfect forecast. The standard error of the EDI is given as (Ferro and Stephenson, 2011):

$$s_{\mathrm{EDI}} = \frac{2|\log F + \frac{H}{1-H} \log H|}{H(\log F + \log H)^2} \sqrt{\frac{H(1-H)}{np}} \tag{2.3}$$

with sample size n, i.e. the total number of days considered and base rate p. The EDI is always shown with an interval of 2 standard errors covering approximately the 95% uncertainty interval. The skill is here considered significant if this interval does not encompass 0.

We will analyze the skill of the forecasting systems in terms of their ACC and EDI. These scores measure different aspects of forecast skill. While the ACC quantifies how well the anomaly pattern of the verification field is reproduced, the EDI measures the deterministic skill at predicting a certain event type. We restrict ourselves to comparing the EDI between different forecasting systems, regions and event definitions. The EDI for extreme and average events will be referred to as xEDI and aEDI, respectively.

2.3 Results

2.3.1 Summer Temperature Skill on Subseasonal Timescales

First, we assess the prediction skill for summer surface temperatures over European land areas. As a reference for the general temperature skill in the six European regions, we compute the ACC of the ensemble mean hindcasts (Figure 2.1a) in the four analyzed systems as well as that of a persistence forecast as defined in Section 2.2.1.

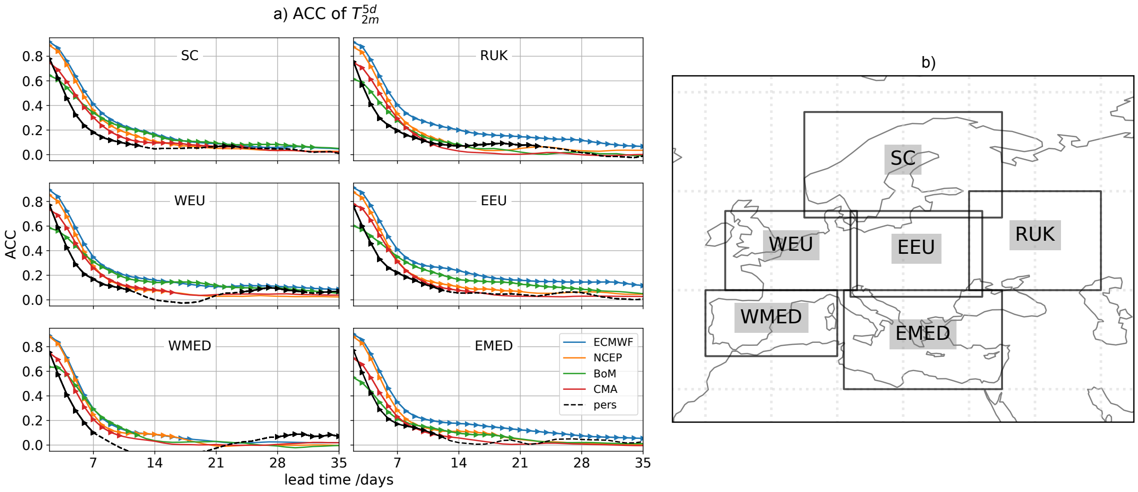

FIGURE 2.1: a) ACC of summer $\langle T_{2m}^{5d} \rangle_R$ as a function of lead time for the six European regions indicated in b). The ACC is shown for each of the four forecasting systems used in this study (see legend) and a persistence forecast based on ERA-Interim (black dashed line). Verification is with respect to ERA-Interim. Triangles indicate where the ACC is significantly different from zero. Note that the lead time corresponds to the central day of a 5 day running mean, thus the shortest lead time is two days.

All models show a fast decrease in ACC in week 1 (up to day 7) and their skill tends strongly toward that of persistence by week 4 (lead time 21 – 28 days) at the latest. The ECMWF system's ACC curve lies above all other models for all considered regions and lead times with only few exceptions. For lead times up to 7 days only the NCEP model compares well but is consistently slightly below. For lead times longer than 7 days, the NCEP system's skill deteriorates more quickly and becomes indistinguishable from the

ACC of the persistence forecast within week 2 (lead time 7 – 14 days) of the forecast. The ACC of the CMA system and the persistence forecast start at a similar value but the CMA outperforms the persistence forecast generally until the middle of week 2 and exhibits a comparable skill after. The BoM system shows a slightly different behavior from the other systems as its skill is lower than the persistence skill up to approximately 4 days lead time. However, the decrease in ACC of the BoM system is much more gradual than that of all other systems. In fact, despite having the lowest skill in week 1, the BoM outperforms all other systems except the ECMWF starting in week 2 in the regions SC, WEU and EEU, where it exhibits a skill comparable to the ECMWF system.

The ACC furthermore shows some dependence on the region under consideration. Focusing on the skill of the best performing model, the ACC in the SC, RUK and EEU regions drops below 0.4 approximately one day later than in the other regions. In RUK and EEU it remains above 0.2 until the end of week 2 whereas in SC it already falls below that threshold during week 2. Despite the slower decrease of the ACC in the first two weeks, in week 3 the skill is lowest in SC along with the WMED region and effectively drops to zero, while it remains above 0.1 for the other regions. For the RUK and EMED regions however, this is only the case for the ECMWF system. The ACC in the other systems is zero in these regions in week 3 as well. In WEU and EEU, both ECMWF and BoM keep an ACC different from zero until week 4. In the EEU region the ECMWF system remains an ACC above 0.1 until week 5.

Out of the four considered forecasting systems, the ECMWF system clearly performs best in terms of ensemble mean ACC at all lead times, but it is not able to outperform a persistence forecast beyond week 4 (except in the EEU region). Despite its lower skill in week 1, the BoM system performs better than NCEP and CMA after week 2 in two thirds of the regions. For subseasonal lead times the ACC of the ensemble mean forecasts to predict pentad means of 2m-temperature is generally very low with some dependence on the region under consideration.

2.3.2 Prediction Skill for Extreme versus Average Temperatures

Next, we consider how the EDI of predicting extreme near-surface temperature events (xEDI) compares to the EDI for average events (aEDI, for event definition see Section 2.2.2). Note that for simplicity we refer to events in the lower tail of the temperature distribution as "cold events". For warm events in the ECMWF system, there is a clear tendency of the extreme event skill to significantly exceed the average event skill in all regions in forecast week 1 as indicated by the grey areas in Figure 2.2a. At lead times up to 14 days, however, in regions EEU and EMED both the xEDI and the aEDI become effectively zero. In the remaining four regions, the average event skill drops much more quickly than the warm event skill such that in week 2 the warm event skill is significantly higher. In week 3, the EDI difference drops to zero in the WMED region but stays positive in the other three regions. In week 4 the warm event skill lies above the average event skill only in the WEU and RUK regions. These regions show the largest difference between warm and average

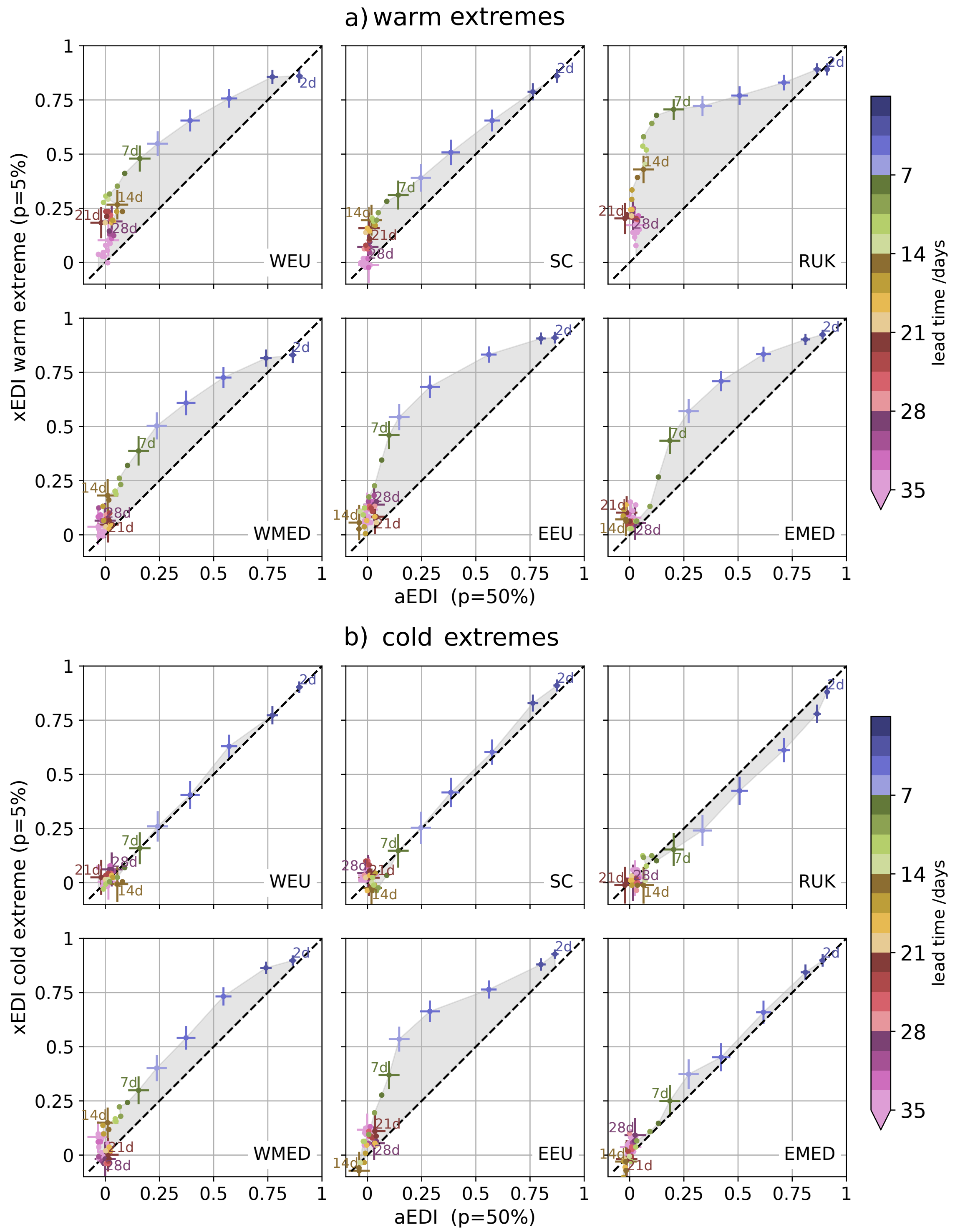

FIGURE 2.2: aEDI against xEDI for (a) extreme hot and (b) extreme cold events in the six regions for the ECMWF system. Bars indicate two standard errors around the aEDI and xEDI and are shown for lead times of 2-7, 14, 21, 28, and 35 days. Values are colored according to lead time. The black dashed diagonal indicates where aEDI = xEDI.

event skill. Notably however, the absolute warm event EDI drops much slower in the RUK than in the WEU region.

In contrast to the warm events, the xEDI for cold events does not significantly exceed the aEDI at any lead time and in any region except WMED and EEU (Figure 2.2b). In these regions, we observe significantly higher skill for extreme than for average events in week 1. However, these values decrease to zero within week 2 in both regions showing that also for those regions — despite the slower skill decrease in week 1 — the EDI difference at lead times longer than one week is effectively zero. This implies that there is no extended skill for cold extremes over average temperatures at subseasonal timescales.

Generally, in the ECMWF system summer warm extremes in Europe are better predicted than climatological events, with a strong dependence on the region considered. Especially at subseasonal lead times, warm extremes in the RUK and WEU and to some degree in the SC regions are significantly better predicted than average events. This is in contrast to the prediction skill for cold extremes in the same regions which is of the same magnitude as that for average events throughout all forecast lead times. Even though the EDI for cold extremes is larger in the WMED and EEU regions, this difference vanishes at subseasonal lead times. A sensitivity test with respect to the percentile threshold for defining extreme events using base rates p_x of 10% and 2.5% yields that the skill difference between extreme and average events is relatively insensitive to the exact percentile threshold chosen for the definition of the events (Figure A.1). Our main conclusions equally hold for an analysis of xEDI and aEDI for the ECMWF for an extended period of 20 years (1998 – 2017, see Figure A.2).

In order to compare these results between different forecasting systems, we consider the EDI differences for the same regions but restrict the analysis to subseasonal lead times only, i.e. week 3 and 4 (Figure 2.3 a and b, respectively). The xEDI and aEDI are effectively equal in the WMED, EEU and EMED regions and across all models. The only exception is the warm event skill of the BoM system in the EEU region in week 3. In the SC region, Figure 2.3 confirms that the difference between xEDI and aEDI events is not significant for other models. The warm event xEDI in the WEU region exceeds the aEDI for all models even though the differences in the other systems are less pronounced than in the ECWMF system. Especially the BoM system shows a larger uncertainty in the skill estimates which is likely due to the higher climatological spread of its ensemble. Again, the skill for cold events is not significantly different from the average event skill. Across all models, the most pronounced differences in the skill arise for warm events in the RUK region, especially for ECMWF and NCEP.

In summary, the ECMWF system performs significantly better at forecasting extreme warm near-surface temperature events than average events on subseasonal timescales in three (WEU, SC and RUK) out of six European regions. In all those regions the xEDI for warm events is significantly different from zero for all models indicating some prediction skill of extreme hot events on subseasonal timescales. The EDI difference is only robust in two (WEU and RUK) out of those regions when taking the other forecasting systems into consideration. For cold events, there is no enhanced skill over average temperature events

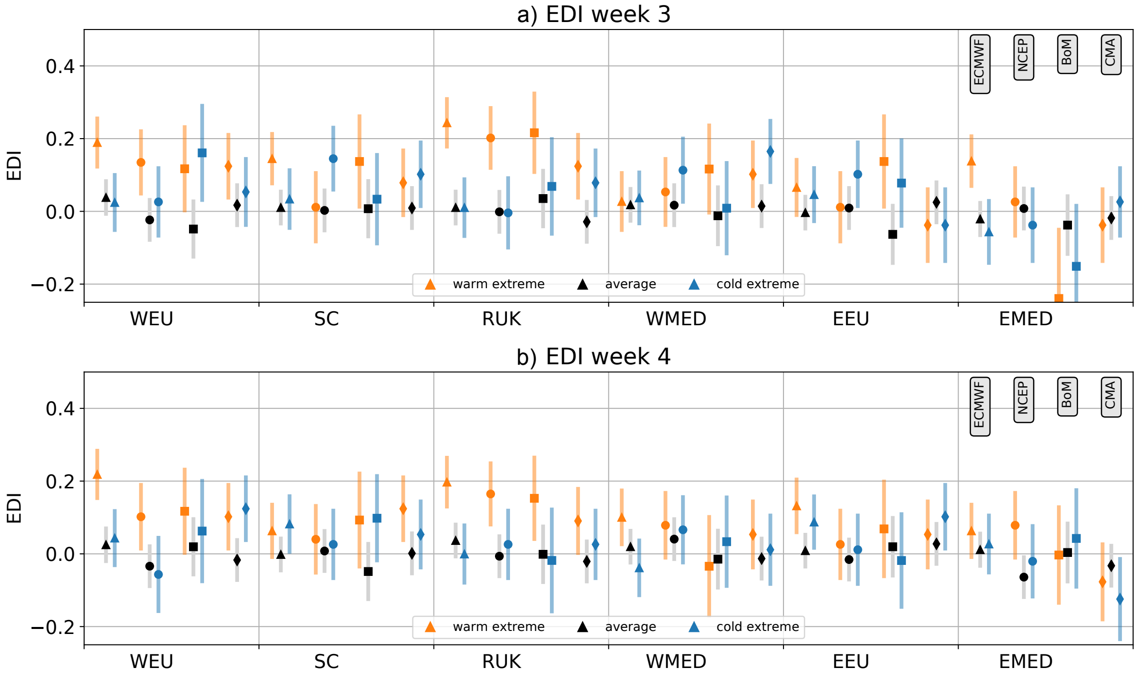

FIGURE 2.3: EDI by region and model on subseasonal lead times for (a) 3 and (b) 4 weeks. Orange, black, and blue dots and intervals indicate the EDI and two standard errors around it for extreme hot, average and extreme cold 2m-temperature events, respectively. For each region, the EDI for each of the four considered forecasting systems is shown, where triangles, circles, squares and diamonds indicate the ECMWF, NCEP, BoM and CMA system respectively.

at subseasonal lead times. This finding is consistent for all considered forecasting systems. To further test for the robustness of these results, we consider another base-rate independent skill measure. The odds ratio skill score (ORSS, Text S1 in Appendix A & Figure A.3) confirms the overall findings of Figures 2.2 and 2.3 but at subseasonal lead times of 3 and 4 weeks the ORSS is only significantly different between extreme warm and average events for the ECMWF system. It should additionally be noted that the ORSS — despite being base-rate independent — increases the rarer the event is making its interpretation more difficult (Ferro and Stephenson, 2011; Hogan and Mason, 2012).

2.3.3 Sensitivity to Most Severe Heat Waves

Due to the limitations of the hindcast period for the S2S model systems, we are only able to sample 12 summers from 1999 to 2010. Since these years contain the two most severe heat waves in two of the study regions (European heat wave 2003 dominant for WEU region, Russian heat wave 2010 dominant for RUK region) we check the sensitivity of our skill estimates to removing these events from the sample for those two regions. Removing 2003 (2010) from the xEDI computation in the WEU (RUK) region reduces the sample to three quarters (one third) of its original size (orange percentages in Figure 2.4 b and d). Indeed, there is a dependence in both regions to removing the most severe events (Figure 2.4) such that the xEDI at subseasonal lead times is strongly reduced. While in the WEU region the xEDI of the full sample remains significantly different from the aEDI out to week 5, the errorbars for aEDI and xEDI start overlapping already after 10 days when removing

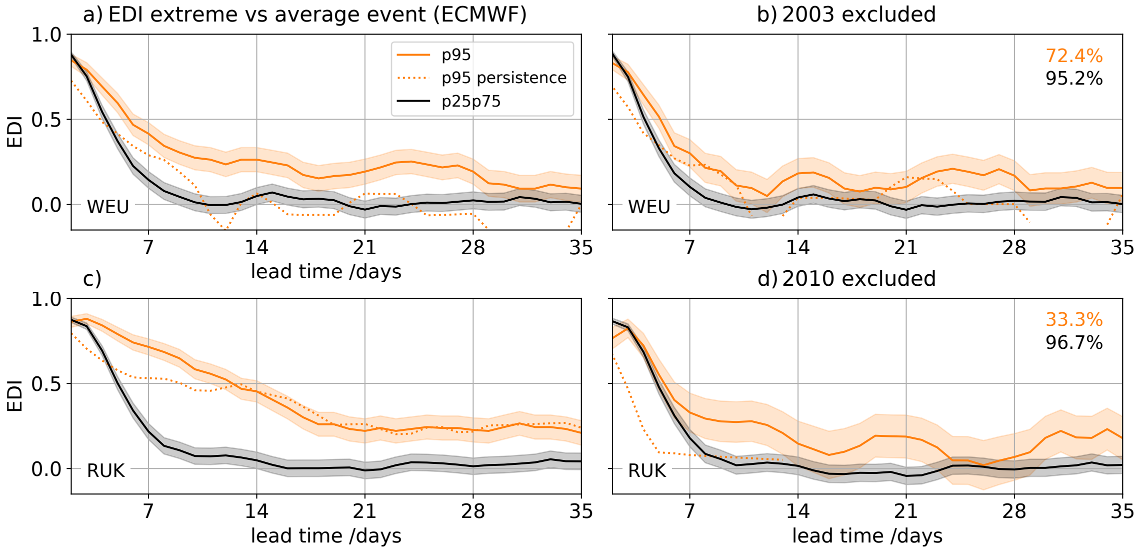

FIGURE 2.4: EDI for average (black) and extreme hot (orange) temperatures as a function of lead time taking into account ECMWF hindcasts in all available years (a,c) and excluding the most extreme years (b,d) in the respective region (2003 for WEU, 2010 for RUK). The shading around the curves indicates two standard errors around the EDI. The orange dotted line shows the xEDI of a persistence forecast. Percentages in b and d indicate the fraction of extreme (orange) and average (black) events that are retained in the estimation of the curves in the respective panel.

the summer of 2003. In general, the xEDI at lead times up to week 3 is most strongly affected by the removal of 2003 from the sample. In the RUK region the effect on the xEDI of removing the strongest event is even more pronounced but also note that the sample is more strongly reduced in the case of 2010. When 2010 is removed for the estimation in the RUK region the xEDI drops as quickly as the aEDI during the first week. It then levels off more gradually but due to the larger uncertainty it is no longer significantly different from the aEDI by the end of week 2.

These differences in the extreme event skill imply that there was larger skill at predicting the two events that were removed in Figures 2.4b and d. To answer the question why these events show extended prediction skill we also considered the xEDI of a persistence forecast (orange dotted line in Figure 2.4). The persistence xEDI is lower than that of the dynamical forecasts and rather insensitive to removing the 2003 event from the sample in the WEU region. In contrast, the persistence xEDI in the RUK region is comparable to the xEDI of the dynamical hindcasts when all years are taken into account, especially at subseasonal lead times. When removing 2010 from the sample, the xEDI of the persistence forecast drops to effectively zero already in week 1. This highlights the importance of the temperature persistence during the Russian heat wave of 2010 and the ability of the hindcast models to capture this aspect of the event.

2.4 Summary and Discussion

We investigated the prediction skill of near-surface summer temperatures in Europe from ensemble hindcasts of four subseasonal forecasting systems for the years 1999 – 2010. The skill of the ensemble mean hindcasts was evaluated using the ACC, which measures the pattern correlation between the hindcast and reanalysis fields. The ensemble mean ACC for 5-day means of 2m-temperature lies below 0.4 for lead times of more than one week. The ECMWF performs best out of the four considered forecasting systems. While the BoM system performs poorest in the first week of the forecast it outperforms the CMA and NCEP systems in multiple regions on subseasonal lead times. This could be due to the generally larger climatological spread of the BoM system, which could be an advantage in capturing a weak predictable signal on these time scales, especially for cases where the NCEP and CMA systems are overconfident. The model skill also exhibits regional dependence. The regions with largest skill beyond week 1 are Russia/Ukraine (RUK), Eastern Europe (EEU) and Western Europe (WEU). The Extremal Dependence Index (EDI) for different event types (see Section 2.2.2) was applied to evaluate how the prediction skill for extremes compares to that for average events. Our analysis shows that in most regions of Europe, the prediction of warm near-surface temperature extremes is significantly better than for events close to the mean of the temperature distribution. This holds for week 1 in all regions, but the difference between the EDI for average events (aEDI) and for extreme events (xEDI) is strongest in the RUK, EEU and WEU regions. At subseasonal lead times the xEDI is significantly higher than the aEDI only in the RUK and WEU regions. In the RUK region, this result is robust across all considered forecasting systems. Excluding the European heat wave 2003 and the Russian heat wave 2010 from the sample for the computation of the EDI yields that these events contributed most strongly to the extended skill in the WEU and RUK regions, respectively. While warm extremes clearly show a higher predictability than average events, this does not hold for cold extremes with the same base rates. The cold event xEDI is only significantly different from the aEDI in the EEU and the Western Mediterranean (WMED) regions up to approximately 10 days lead time and not distinguishable from zero in the remaining regions at all lead times.

Although the differences between systems are small we find the ECMWF to have the largest skill at subseasonal lead times. Note however that we use ERA-Interim (i.e. a product of the ECMWF's Integrated Forecasting System used to generate the hindcasts) for the verification of the hindcasts, which potentially favors the ECMWF system in the skill comparison.

The asymmetry in the prediction skill between warm and cold extremes points to different processes for these opposing events. The large-scale control on near-surface temperatures could offer one explanation for the higher skill in forecasting extremes. Although both the onset and the duration of atmospheric blocking are rather poorly predicted in forecast models (Tibaldi and Molteni, 1990; Matsueda and Palmer, 2018), Grazzini and Vitart (2015) show that subseasonal forecast skill of the synoptic-scale circulation is enhanced when long, coherent RWPs from the Pacific to the Atlantic are present in the initial conditions. Although we did not explicitly test the occurrence of such wave packets in the initial conditions of our extreme event forecasts, some of these situations are captured in the analyzed period (Fragkoulidis et al., 2018). This could explain part of the observed extended skill for extreme temperatures. Land-atmosphere interactions have also been suggested to extend the predictability of near-surface temperatures to subseasonal timescales (Koster et al., 2010). In particular, the soil moisture-temperature feedback has been shown to be crucial for temperature extremes (Seneviratne et al., 2010) and could force temperatures to rise. To the extent that land-atmosphere fluxes are correctly represented in the models, they may yield extended prediction skill for the most long lasting warm extremes. Note that the regions considered here mostly have soil moisture above the critical value (Teuling et al., 2009). Thus the process described above only applies for extreme warm temperatures and could account for part of the asymmetry in the prediction skill between warm and cold extremes. Furthermore, the feedback would likely only act under strong and persistent atmospheric forcing as was the case in 2003 (Fischer et al., 2007b) and 2010 (Miralles et al., 2014).

The results obtained here are specifically relevant in the context of a warming climate with potentially different predictability characteristics (Scher and Messori, 2019). Under global warming the temperature distribution exhibits both a shift and a broadening towards warmer values (Schär et al., 2004) and temperature extremes become more frequent (Coumou and Rahmstorf, 2012). While an increased number of heat events can have harmful impacts on ecosystems and society, the potential for a better prediction of the most severe heat events as presented here is promising.

Acknowledgments

ERA-Interim and S2S data were obtained from the ECMWF. The S2S data is provided through the S2S Prediction Project. Funding by the Swiss National Science Foundation through project PP00P2_170523 is gratefully acknowledged. We thank Sonia Seneviratne and Mathias Hauser for helpful discussions on soil moisture-atmosphere interactions. All analyses and visualizations were conducted in python.

Chapter 3

Trends Inflate Subseasonal Surface Temperature Skill

This chapter is currently under review for publication in the Quarterly Journal of the Royal Meteorological Society[1]*:*

Wulff, C. O., Vitart, F., and Domeisen, D. I. V. (2021). Trends Inflate Subseasonal Surface Temperature Skill. Submitted to *Quarterly Journal of the Royal Meteorological Society*

Abstract

Subseasonal-to-seasonal (S2S) predictions have a wide range of applications. Improving forecasts on this time scale has therefore become a major effort. To evaluate their performance, these forecasts are routinely compared to a reference that represents the climatological distribution at any given time. This distribution is commonly assumed to be stationary over the verification period. However, there are prominent deviations from this assumption, especially considering trends associated with climate change. By employing synthetic forecast-verification pairs we show that estimates of the probabilistic skill of both continuous and categorical forecasts with a constant actual level of skill increase as a function of the variance explained by the trend over the verification period. The skill of a categorical prediction can be enhanced even further when evaluated over a longer forecast period. We also show this skill enhancement due to the trend in the ECMWF extended-range ensemble prediction system. We demonstrate that the effects on the skill in an operational forecast setting are currently strongest in the tropics and mainly relevant for categorical forecasts. This shows that care needs to be taken when evaluating forecasts that are subject to non-stationarity on time scales much longer than the forecast verification window, especially for categorical forecasts. The results presented in this study are not exclusive to the S2S time scale but have wider implications on forecast verification on seasonal to decadal time scales, where the existence of trends can further enhance forecast skill.

[1] Note that British English spelling is used throughout this chapter.

3.1 Introduction

Stakeholders face decisions related to weather and climate risks on a continuum of time scales from minutes into the future to multiple decades or even centuries. In recent years, increasing efforts have been made to move towards "seamless" prediction (Hoskins, 2013) to bridge the gap between classical weather forecasting for lead times of days and climate projections for the next century. The potential for successful predictions on all time scales stems from a multitude of large-scale phenomena in the atmosphere and the ocean that evolve in different parts of this temporal spectrum. One case in point is the El Niño-Southern Oscillation — a coupled ocean-atmosphere phenomenon with a variable frequency between 2 and 7 years that gives rise to seasonal prediction skill not only in its region of occurrence but also in the mid-latitudes of the globe through atmospheric teleconnections (Shukla et al., 2000). Other sources of predictability can be found on time scales from weeks to months, the so-called subseasonal-to-seasonal (S2S) time scales (Vitart et al., 2017, for a general overview), for instance the Madden-Julian Oscillation (e.g. Lee et al., 2019), stratospheric processes (e.g. Domeisen et al., 2020b), land surface processes (e.g. Koster et al., 2011), and sea ice variability (e.g. Jung et al., 2014). The S2S time scales have furthermore been shown to be particularly relevant for decision makers (White et al., 2017; Robertson et al., 2020). We will focus on forecasts on these time scales in this study.

Despite the potential for predictability, S2S forecasts generally exhibit significantly lower skill than forecasts on weather time scales. It is thus crucial to use ensemble systems in order to sample the space of possible outcomes given the uncertainty in the estimate of the initial state (Leutbecher and Palmer, 2008). Despite the low deterministic skill on S2S time scales, these probabilistic forecasts can be useful in making decisions that depend on the weather evolution given that they quantify the uncertainty in the outcomes correctly. Probabilistic subseasonal forecasts have, in fact, been shown to be skilful in a variety of settings (e.g. Alvarez et al., 2020; Materia et al., 2020; Robertson et al., 2020).

In order to increase users' confidence in S2S forecasts, it is imperative to show under which circumstances these can provide useful information that go beyond what a "trivial" forecast can provide, such as the fact that summer will be warmer than winter in the case of seasonal outlooks. For this, the forecasts need to be carefully verified and their performance compared with that of a reference that includes all the "obvious" outcomes. Only if a forecast does better than this reference it is said to have skill (Wilks, 2019a). The reference can take different forms (e.g. persistence forecast, forecast with a previous model version), and often "climatology" is chosen as a reference forecast, i.e. an average of the variability of the system over a reference period. The definition of this climatology forecast is not unique, however and often it is limited by the available hindcast data. In a more conceptual way this was shown previously by Hamill and Juras (2006) using synthetic forecast-observation pairs. They introduced two hypothetical islands with different climatological event frequencies to illustrate how a forecasting system that always issues the climatological frequency for each respective island (and thus has no actual skill) can appear to have skill when the aggregated climatology for both islands is used as reference. While the solution to the artificial enhancement of the skill in this hypothetical example is straight-forward, in an operational setting it might not be trivial due the available sample size to estimate the climatology. Manrique-Suñén et al. (2020) used the operational extended-range predictions from the European Centre for Medium-Range Weather Forecasting (ECMWF) and found that there can be substantial differences in the estimates of the skill of subseasonal forecasts depending on the chosen climatology. They showed that computing the climatology of weekly mean temperatures from a limited hindcast period with different temporal aggregation leads to different skill estimates.

In a general sense, the aforementioned results illustrate the importance of properly accounting for the non-stationary components in the climatological distribution when assessing the skill of a forecast with respect to climatology. While the seasonal cycle is often accounted for in the reference climatology for the evaluation of forecasts, non-stationary components that act on longer time scales are commonly neglected. One prominent example of a non-stationary component in temperature is global warming (IPCC, 2013). For seasonal forecasting systems to produce realistic warming, the greenhouse gas forcing giving rise to it needs to be accounted for in the boundary conditions (Doblas-Reyes et al., 2006; Liniger et al., 2007; Boer, 2009). Though the magnitude of global warming shows large regional variability, it is manifest in temperature time series throughout the globe. In many places, a shift in the mean temperature can be detected even when considering only the recent past, e.g. the last 30 years, which is a common period for defining a climatology. Based on the arguments above, this non-stationary component of the climatology has the potential to affect the estimates of subseasonal forecast skill.

Our study aims to characterise and quantify the effect of a trend in the climatological reference period on the probabilistic skill of subseasonal forecasts. We hypothesise that there is an enhancement of skill when there is an underlying trend and that the magnitude of the improvement depends on the amount of variance of the respective time series that the trend accounts for. We introduce the forecast and verification data and the processing methods in Section 3.2 along with the probabilistic scores for evaluating their performance. In Section 3.3 we quantify the effect of a trend on different probabilistic skill scores in a set of synthetic forecast-verification pairs to separate the stationary from the non-stationary component. Here, we also assess what happens to the skill scores when there is a trend in the verification but the forecasts fail to reproduce it correctly. We then compare the behaviour of an operational prediction system (the ECMWF's extended-range ensemble prediction system) to our synthetic model in Section 3.4 and show where on Earth subseasonal forecast skill may be most strongly affected by the presence of a trend in the mean of the climatological distribution. We review the results in the context of previous literature and discuss limitations of the synthetic model in Section 3.5. Finally, we present the conclusions of our study in Section 3.6.

3.2 Data and Methods

3.2.1 Forecast Data and Verification

For assessing the trend effect in an operational ensemble prediction system, we make use of the extended-range forecasting system of the European Centre for Medium-Range Weather Forecasts (ECMWF). This extended-range forecast ensemble is operationally produced with the most recent version of the ECMWF Integrated Forecasting System (IFS) by extending the weather forecast runs twice a week out to 46 days (instead of 15). This is done by re-starting the forecast runs on day 14 at a reduced horizontal resolution (Tco319 instead of Tco639) but note that this first day of the re-start is only used as spin-up. The horizontal resolution corresponds to approximately 16 km up to day 15 and 32 km after and the model has 91 vertical levels. The ensemble in the operational setting consists of 51 members. For each forecast, hindcast ensembles with 11 members are produced by initialising the same model version from re-analysis (ERA-Interim up to and including IFS cycle 45R1, ERA5 from cycle 46R1) on the same calendar day for the previous 20 years. The atmospheric component of the IFS is coupled to an ocean and an interactive sea ice model and uses the HTESSEL land surface scheme. The IFS uses a boundary forcing with varying greenhouse gas (GHG) concentrations following the CMIP3 A1B scenario (Meehl et al., 2007). Since the GHG forcing thus does not only enter through the initial conditions, we expect the model to produce realistic trends at all lead times.

To cover a period that is sufficiently long to consider an effect of trends on the forecasts, we retrieved 20 years of hindcast data through the S2S database (Vitart et al., 2017). We downloaded daily mean 2-m temperatures from the ECMWF's extended range ensemble forecasts initialised between Jan 1, 2018 and December 31, 2018 (twice-weekly initialisation, giving 105 forecasts) as well as the corresponding hindcasts (same initialisation days within the year for the period 1998 – 2017). For each initialisation 20 years of hindcasts are produced, yielding a sample of 2100 hindcast-observation pairs for each lead time. In addition, we extended the number of samples in the forecast period by including all forecasts initialised between July 11, 2017 and May 18, 2020 resulting in 298 initialisations.

Note that in order to increase the sample size for the forecasts and hindcasts, we use varying model versions in our analysis. In particular, for this period of hindcasts and forecasts, the ECMWF changed from cycle 43R3 to 45R1 to 46R1 of the IFS (`https://confluence.ecmwf.int/display/S2S/ECMWF+Model#app-switcher`), and thus the model data considered here were generated with three versions of the IFS. There are some important differences in the IFS between versions, which impact the forecast and hindcast skill to some degree but the effect of the changes in model version used here has been mainly visible on the shorter lead times of the forecasts (e.g. Vitart et al., 2019).

As verification data for the hindcasts and forecasts, we use daily mean 2-m temperatures from ERA5 (Hersbach et al., 2020). The data were downloaded globally at $1^\circ \times 1^\circ$ resolution for the period January 1, 1997 – June 30, 2020.

3.2.2 Estimating the Seasonal Cycle

For the remainder of this study, it is useful to consider standardised temperature anomalies because it allows us to easily identify the amount of variance that different components of the time series account for in the hindcast period. By accounting for the models' own climatological mean and standard deviation, it additionally ensures a certain degree of calibration of the forecasts, but note that for a proper calibration, a more sophisticated approach would be required. For our simple calibration, we compute the seasonal temperature cycle from the hindcast period (1998 – 2017) only.

To transform to standardised anomalies, we need to define the climatological seasonal cycle of the mean and standard deviation. In order to compute these for the hindcasts, we use all 7-day mean absolute temperature values for one lead time in the hindcast period at a time. This gives 105 initialisations for each of the 20 years. Since the weekly means for each hindcast are computed (i.e. averaging over the lead time dimension) we are left with 6 lead weeks, i.e. days 1 – 7, 8 – 14, ..., 36 – 42. This results in an array with dimensions 20 (years) × 105 (initialisations) × 6 (lead times). In order to estimate the seasonal cycle in the climatological mean (for each lead time individually), we then average over each calendar day (i.e. the "years" dimension) and fit four harmonics to the resulting time series of length 105. This retains only variations with a period greater than approximately 90 days and provides us with a smooth seasonal cycle as a function of lead time. This is the climatological mean seasonal cycle. To get the seasonal cycle of the standard deviation, we first compute hindcast anomalies by subtracting the seasonal cycle of the mean from the absolute temperatures. We then compute the standard deviation for each calendar day and again - for each lead time individually - fit four harmonics to the resulting time series, thus obtaining a smooth seasonal cycle of the standard deviation, which we will use for standardisation.

3.2.3 Processing of Hindcasts and Verification

We first average the hindcast, forecast and ERA5 temperatures to 7-day averages. We then transform these time series into dimensionless standardised anomalies to minimise contributions of the seasonal cycle to the hindcast skill. For this, we first subtract the climatological mean seasonal cycle from the absolute temperatures and then standardise by dividing these absolute anomalies by the seasonal cycle of the standard deviation (see Section 3.2.2 for the estimation of the respective seasonal cycles). This procedure results in the dimensionless hindcast temperatures at every time step being approximately standard normally distributed. Note that this is not necessarily the case for the forecasts and the verification in the forecast period.

We further compute the trend over the 20 years of 7-day mean standardised anomalies by linear regression. This linear trend is subtracted to obtain the de-trended values.

Note that in the case of the forecasts, the seasonal cycle (mean and standard deviation) as well as the trend are a function of the lead time to account for possible drifts in the model climatology.

For transforming the ERA5 temperatures to standardised anomalies we follow the same approach as outlined above for the hindcasts. This means we take ERA5 temperatures for the same 20 years as the hindcasts (January 1, 1998 until December 31, 2017) averaged over 7-day periods. The seasonal cycles in mean and standard deviation are then computed as above (without the need to account for a lead time dimension) but using all 365 days of the year.

The above described estimation of the seasonally varying climatology does not result in the standardised anomalies perfectly following a standard normal distribution, even during the hindcast period. Since it is mainly their variance that deviates from unity, in a last step, we standardise the time series again by dividing by the standard deviation computed from the hindcast period. This ensures the unit empirical variance over the hindcast period for both the verification and hindcasts at all lead times.

Note that the above described processing of the data gives the most reasonable results when the temperatures are normally distributed. Especially in the high Arctic however, the near-surface temperatures in the prediction model are strongly non-Gaussian, probably due to the effect of sea ice. We thus exclude grid-points north of 81°N from the analysis.

3.2.4 Forecast Verification

Scoring

There exists a multitude of scores that can be used to assess the performance of a forecast (e.g. Jolliffe and Stephenson, 2012). Since there is no single measure that captures all aspects of the performance, usually multiple measures are applied and the choice of the scores often depends on the specific application. Here, we use a fairly general score for the evaluation of forecasts of continuous variables, the continuous ranked probability score (Wilks, 2019a, CRPS,), which summarises multiple attributes of a forecasting system. For a forecast of temperatures y at a single instance when a temperature o was observed, the CRPS is given by:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy \tag{3.1}$$

where $F(y)$ is the cumulative distribution function (CDF) of the temperatures in the forecast ensemble and $F_o(y)$ is the CDF of the observation given by:

$$F_o(y) = \begin{cases} 0, & y < o \\ 1, & y \geq o \end{cases} \tag{3.2}$$

which describes a step function with a jump from 0 to 1 at temperature $y = o$.

The CRPS describes the error of a probabilistic forecast by the integrated squared distance between the forecast and the observed CDFs. The CRPS is negatively oriented, meaning the smaller the score, the better the forecast. In all cases considered, we average the CRPS over the number of all forecast-verification pairs.

To estimate the CRPS we follow the approach described in detail in Hersbach (2000), which further allows us to decompose the CRPS into three components, namely

$$\mathrm{CRPS} = \mathrm{REL} + \mathrm{CRPS}_{pot} \tag{3.3}$$

Reliability (REL) is a measure of how well on average the forecast probability matches the frequency of occurrence of a certain temperature (interval). Note that the naming is counterintuitive: a small reliability means the forecasts are reliable (perfectly reliable when REL= 0). The reliability component is sometimes also referred to as calibration.

The potential CRPS (CRPS_{pot}) is the CRPS a prediction system would attain if it were perfectly calibrated (i.e. REL= 0). It will decrease for forecasts that on average have a lower spread but will increase with more of the verification values lying outside of the range of ensemble members (Hersbach, 2000).

We also use a categorical score to evaluate the performance of the forecasts. For this, we use the RPS as defined by Wilks (2019a):

$$\mathrm{RPS} = \sum_{m=1}^{J} (Y_m - O_m)^2 \tag{3.4}$$

where J is the number of categories that verifications and forecasts are sorted into. The observation O_m attains a value of 0 or 1 if the verification at the considered time step lay outside of or inside category m, respectively. The forecast probability Y_m can have values between 0 and 1 and indicates how many ensemble members were in category m. Choosing $J = 3$ equiprobable categories makes the RPS the natural choice for evaluating tercile forecasts. This requires defining the tercile thresholds. In the case of the synthetic model (Section 3.3.1) we know the exact values for these thresholds at any time step. In reality however, the thresholds have to be estimated from the climatology. Since we standardise the verification and predictions in the hindcast period to have mean 0 and unit variance, we use the tercile thresholds of the standard normal distribution ([-0.431, 0.431]) to define the categories. Note that these values would also be obtained when estimating the tercile thresholds from the undetrended climatological distribution of the hindcast period.

Skill

In the following sections, we further evaluate the skill of the forecast by considering the relative improvement in its score S over the score S_{ref} of a reference forecast. For this we define a skill score SKS, which can vary between $-\infty$ and 1, where 0 indicates no improvement over the reference and 1 means that the score S attains its optimal value $\mathrm{S}_{opt} = 0$. The skill score SKS is thus given by:

$$\mathrm{SKS} = \frac{\overline{\mathrm{S}} - \overline{\mathrm{S}_{ref}}}{\overline{\mathrm{S}_{opt}} - \overline{\mathrm{S}_{ref}}} = 1 - \frac{\overline{\mathrm{S}}}{\overline{\mathrm{S}}_{ref}} \tag{3.5}$$

Here, the overbar denotes the average over all forecast-verification pairs. Since we only consider averages over the entire sample (either hindcasts or forecasts), it will be dropped in the following. In our case, S will either be the CRPS or the RPS and SKS will be the CRPSS or RPSS, respectively. For a reference score S_{ref}, we always consider the score of the detrended hindcasts at predicting the detrended verification.

In the ECMWF ensemble prediction system, the hindcast ensemble has only 11 members compared to 51 in the forecasts. We also follow this set-up in our synthetic model. However, several studies have shown that both continuous and discrete ranked probability skill scores depend on the size of the ensemble (e.g. Richardson, 2001; Müller et al., 2005; Weigel et al., 2007). In order to compare the scores for forecasts and hindcasts (as will be done when computing skill scores of the forecasts), we thus need to account for the effect of the different ensemble size by using a fair score. Ferro et al. (2008) derive estimates for a fair RPS and CRPS for forecasting systems assuming perfect reliability and independent ensemble members. Under these conditions, the ratio D of the unadjusted score S_M for an ensemble of size M to the fair or adjusted score S_{fair} is $D = (1 - \frac{1}{M})$. Thus, we multiply our estimates of the CRPS and RPS by D^{-1} to obtain the fair scores. The assumptions of perfect reliability and independent ensembles hold fairly well for our synthetic forecasts. For the real forecasting system both assumptions are violated. Figure 5 in Leutbecher (2019) shows that in that case the actual ratio D tends to be overestimated by using $D = (1 - \frac{1}{M})$ and more strongly so for smaller ensemble sizes. This implies that by using this simple adjustment factor we are likely to underestimate the fair score. The underestimation is likely to be stronger for the hindcasts due to their smaller ensemble size. Thus, the reference score for the CRPSS will be better than it actually is, leading to the CRPSS giving rather a too low estimate of the skill of the forecasts over the reference. We use the simple scaling of the unadjusted scores throughout but we do not refer to these as fair CRPS and fair RPS in the following.

3.3 Dependence of the Skill on Underlying Trends: Synthetic Ensemble System

In the following, we consider a linear trend in a verification time series to test how the probabilistic skill of a simple hypothetical ensemble prediction system changes as a function of the magnitude of the trend. We define a set of synthetic forecast-verification pairs to be able to cleanly separate the stationary random and predictable parts of the time series from the non-stationary component (the linear trend). Since any real forecasting system is also subject to errors, we relax the assumption of a perfect simulation of the trend in the forecast and employ the toy forecast to assess how such an imperfect estimation of the trend can further affect the skill. This assessment provides a benchmark for the potential magnitude of the effect that a trend in the forecast period can have on the skill of the forecast ensemble depending on the trend magnitude and the mis-estimation.

3.3.1 Set-up of the Artificial Forecast–Verification Pairs

In the next sections, we describe the synthetic ensemble forecast–verification pairs that we generate in order to answer the questions posed above. We follow the approach of Weigel et al. (2008a) who used a very similar toy forecast (without a trend but with a parameter controlling the forecast dispersion) to study why multi-model ensembles can outperform the single best models. We will use the notation $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2.

Verification Time Series

We first generate an artificial verification time series v_t consisting of a random component x_t and a linear trend T_t:

$$v_t = x_t + T_t \tag{3.6}$$

with index $t \in \{0, 1, ..., L_{hc}, L_{hc} + 1, ...L\}$ indicating the time step and L the length of the entire time series. The first interval $[0, L_{hc} - 1]$ has length L_{hc} and is referred to as the hindcast period. The interval $[L_{hc}, L]$ of length $L_{fc} = L - L_{hc} + 1$ is the forecast period.

T_t is a linear trend defined as:

$$T_t = s\left(t - \frac{L_{hc}}{2}\right) \tag{3.7}$$

with s being the slope of the trend. By definition the trend line has mean $\mu_T = 0$ in the hindcast period $[0, L_{hc} - 1]$. Since T_t follows a uniform distribution, the variance of the trend line in the hindcast period is:

$$\sigma_T^2 = \frac{(sL_{hc})^2}{12} \tag{3.8}$$

Lastly, we let the random component of the verification x_t be white noise[2] following a Gaussian distribution given by $\mathcal{N}(0, \sigma_x^2)$. To ensure that v_t has unit variance in the hindcast period, we set $\sigma_x^2 = 1 - \frac{(sL_{hc})^2}{12}$. Note that holding the variance of the verification time series constant in the hindcast period results in a constant uncertainty component (UNC) of the CRPS for the hindcasts (see Section 3.2.4) for all values of s.

Ensemble Prediction System

We generate a prediction ensemble with M members at a time step t in a similar fashion to Weigel et al. (2008a). In the following, we will refer to the predictions for the hindcast

[2] Arguably, a more realistic representation of a climate process would be red noise (e.g. Pelletier, 1998). However, in our synthetic example, we could consider the auto-covariance of an AR(1)-process as the predictability of the time series. White noise on the other hand is inherently unpredictable. But since we prescribe the potential skill of the forecasts simply with a parameter α that lets one component of the forecast predict a fraction of x_t at any given time step, we do not need predictability in the verification to mimic some skill in the model. In fact, we tested the results by replacing x_t with red noise following an AR(1)-process within a range of auto-covariance — the conclusions drawn from evaluating the synthetic forecast–verification pairs remain unaltered.

period ($[0, L_{hc} - 1]$) as the hindcasts and the predictions for the forecast period ($[L_{hc}, L]$) as the forecasts. First, we let each ensemble member f_m ($m \in [1, 2, \ldots, M]$) predict a fraction α of the random component x_t of the verification. Note that this results in the predictions — on average over all L time steps — having a correlation of α with the verification. We further let the ensemble predict a certain trend τ_t and lastly, let it have a noise component $\varepsilon_m \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In summary, our prediction system is described by:

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{pmatrix}_t = \alpha x_t + \tau_t + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{pmatrix} \tag{3.9}$$

The trend τ_t of the prediction is defined as the observed trend T_t multiplied by a factor p, which represents the mis-estimation of the trend in the prediction system:

$$\tau_t = pT_t \tag{3.10}$$

To achieve some calibration of the predictions to the verification, we set the variance of the hindcasts to $\sigma_{f_m}^2 = 1$. Since the variance of each ensemble member is given by

$$\sigma_{f_m}^2 = \alpha^2\sigma_x^2 + p^2\sigma_T^2 + \sigma_\varepsilon^2, \tag{3.11}$$

we choose

$$\sigma_\varepsilon^2 = 1 - \left(p^2\frac{(sL_{hc})^2}{12} + \alpha^2\sigma_x^2\right) \tag{3.12}$$

In this way, we ensure the hindcasts to have zero mean and unit variance, as is the case for the verification in the hindcast period. Note that due to setting the variance of hindcasts and verification to 1, the variances of the different components in the verification and predictions are equivalent to the fraction of explained variance in the hindcast period of the respective time series. For instance, σ_T^2 describes the fraction of variance of the verification accounted for by the trend during the hindcast period, while $\alpha^2\sigma_x^2$ is the fraction of variance of the forecast explained by the predicted, de-trended part (αx_t).

In a last step, we have to choose the parameters of our model. Our aim is to mimic the set-up of the actual prediction system. At the same time, since the uncertainty of the scores is smaller for a larger sample size, we aim for a larger sample than in the operational system to obtain a more robust estimate. Since we suspect the trend effect to partially depend on the length of the forecast period relative to the length of the hindcast period, we keep this ratio the same by letting our synthetic prediction system have the same number of forecast and hindcast years as the real system (3 and 20, respectively), but with more initialisations in each year. Thus, we set $L_{fc} = 1050$ and $L_{hc} = \frac{20}{3} \cdot L_{fc} = 7000$. Consistent with the real system, we only use $M_{hc} = 11$ ensemble members for the ensemble hindcasts but $M_{fc} = 51$ members for the forecasts. The remaining parameters, s, α and p are varied. Setting L_{hc} and fixing the variance of the verification to 1 in the hindcast period sets a limit to the possible values of the slope of the trend s, which is readily understood when we consider σ_T^2 as the fraction of variance explained by the trend — this value cannot exceed 1. Thus, we obtain $0 \leq s < \frac{\sqrt{12}}{L_{hc}}$. Since α represents the correlation between verification and prediction, we let the prediction system vary from having no correlation skill to having nearly perfect correlation with the verification, i.e. $0 \leq \alpha < 1$. Finally, the mis-estimation factor p of the trend can be varied. The unit variance of the hindcasts allows us to vary it within the limits $0 \leq p < \frac{1}{\sigma_T}\sqrt{1-\alpha^2\sigma_x^2}$.

3.3.2 Effects of Varying Trend and Average Correlation Skill on the Probabilistic Skill

To illustrate the expected effect of a trend in the verification, consider the synthetic verification time series shown in the upper panel of Figure 3.1, which is an extension of Figure 10.2 in Livezey (1999). In this case, the trend explains 6% of the variance in the hindcast period. At each time step t, the verification is a single draw from a normal distribution $\mathcal{N}(T_t, \sigma_x^2)$. The tercile thresholds of this distribution as functions of t are shown by the green lines. However, in a real forecast situation, the tercile thresholds have to be estimated from the hindcasts. It is common to use all time steps in the hindcast period for estimating the climatological distribution. Due to the way we defined the verification, this distribution will have mean zero and unit standard deviation. The tercile thresholds of a corresponding standard normal distribution (± 0.431) are shown by the black dash-dotted lines. The lower panel of the figure illustrates what happens when these thresholds are used to define the tercile categories for the forecasts. For this, we split the time series by years (one year consisting of 350 time steps). Only in the centre of the hindcast period, approximately $\frac{1}{3}$ of values fall into each category. Towards the beginning, 50% of values end up in the lower tercile, while towards the end of the hindcast period, almost half of the values are sorted into the upper tercile. Since it lies after the hindcast period, this effect is even stronger in the forecast period where more than 50% of values are in the upper tercile even though the trend only explains 6% of the variance. If we now imagine a forecast that is able to reproduce this trend correctly and has realistic dispersion (as our synthetic predictions) but no skill at predicting any other part of the detrended variability in the verification, it will similarly sort more than a third of the forecast values into the upper tercile. When evaluated with a categorical score such as the RPS, the score for the hindcasts will inevitably be higher than if we had used the real percentiles (green lines in the upper panel of Fig. 3.1) of the verification. Note especially that the score of the forecasts will be higher than the score of the hindcasts despite the forecasts not having any more actual skill. Finally, the RPS will also increase as the forecast period that the skill is evaluated over is extended. This illustrates the problem we face when evaluating categorical forecasts over a period that is subject to a trend using empirical category thresholds estimated from the hindcast period.

To assess the effect of any underlying trend on different probabilistic scores of the toy model for different levels of correlation skill α at predicting the "non-trend component" x_t, we now vary the prescribed slope s of the trend. For the following results, we assume that the prediction system simulates the observed trend T_t perfectly, i.e. $p = 1$ and thus $\tau_t = T_t$.

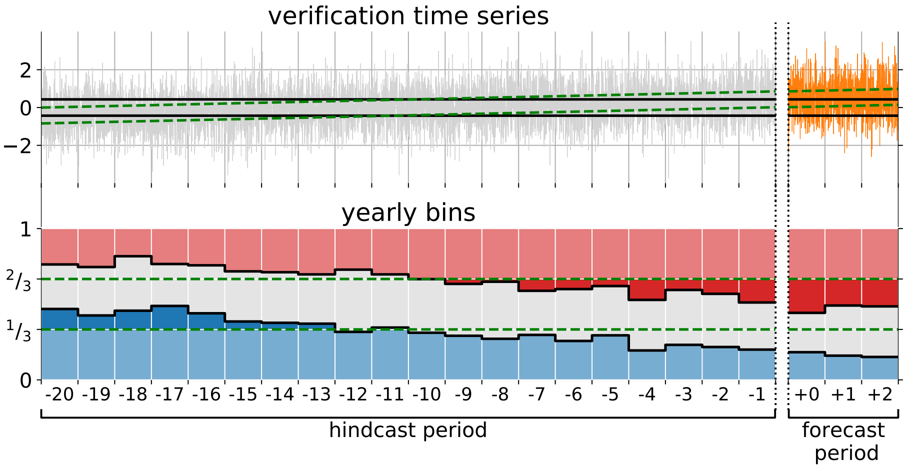

FIGURE 3.1: Illustration of trend effect on a categorical forecast. The upper panel shows a synthetic verification time series generated as described in Section 3.3.1. The grey line indicates the hindcast period and the orange part the forecast period (separated by a dotted line). The black solid lines show the 33.3 and 66.7 percentiles of the distribution of a standard normal distribution. Green dashed lines show the same percentiles but accounting for the underlying trend that here explains 6% of the variance in the hindcast period. The bottom panel shows the fraction of time steps in each year (consisting of 350 time steps) in which the verification falls into the lower (blue), middle (grey) and upper (red) tercile of a standard normal distribution based on the percentile thresholds shown by the black dash-dotted lines in the upper panel.

We then compute the average CRPS and RPS over all time steps t as described in Section 3.2.4 in both the hindcasts and forecasts. Note that we chose a limited sample size to closely resemble the actual prediction system that we consider later, which results in some uncertainty in the estimates of the (skill) scores which appear as noise in the following figures of this section.

The results are shown in Figure 3.2 where the green contour lines show the score (CRPS in a and b, RPS in c and d). All scores decrease (implying a better prediction) with increasing amounts of variance explained by the trend (along the x-axis) but also with more of the detrended variability being predicted, i.e. with higher correlation skill α (along the y-axis). This is to be expected since the fraction of variance in the unpredictable component ε decreases when moving along the 1:1 diagonal. For the CRPS of both hindcasts and forecasts (3.2a and b) as well as for the RPS of the hindcasts (3.2c), the tilt of the green contour lines turns from lying nearly parallel to the x-axis to lying exactly perpendicular to the 1:1 diagonal. This implies that the effect of increasing α outweighs the effect of increasing the trend except when the trend becomes very strong. Note that this is not the case for the RPS of forecasts (Figure 3.2d), where we now see the effect that was illustrated in Figure 3.1. In the forecast period the effect of increasing the trend almost immediately outweighs the effect of higher α. Even for trends that explain less than 70% of the variance in the hindcast period, the RPS drops to the optimal score of 0. Note that Figures 3.2a–c would look exactly the same if we had used different hindcast and/or forecast sizes L_{hc} and L_{fc}. The increase in RPS in panel d, however, depends on the length of the forecast period L_{fc} such that the colour gradient would become much stronger for larger L_{fc} and saturate (i.e. reach RPSS= 1) at a lower value of σ_T^2 (further discussed below).

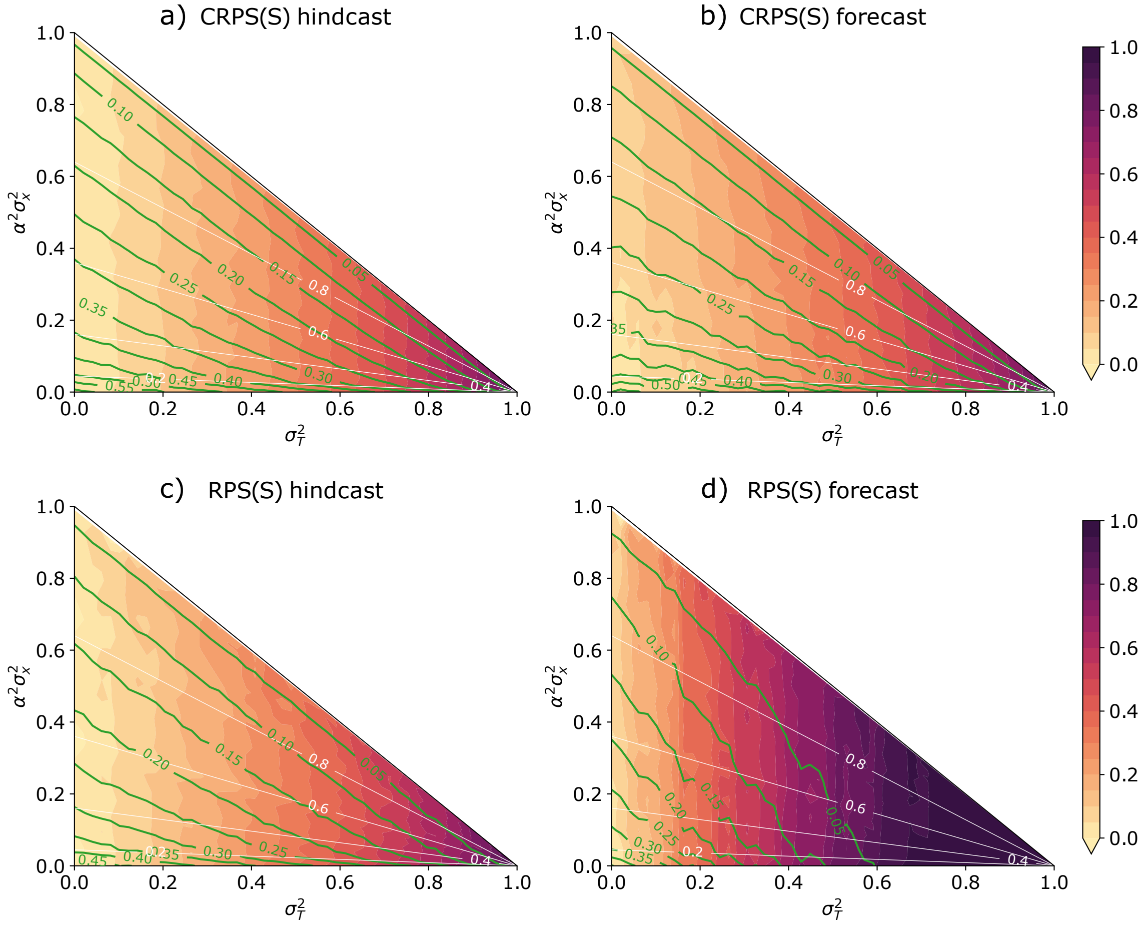

FIGURE 3.2: Scores and skill of the synthetic hindcasts (a and c) and forecasts (b and d) as a function of the variance explained by the trend σ_T^2 in the hindcast period on the x-axis and the fraction of variance in the predictable de-trended component on the y-axis. Green contour lines show the score S (CRPS in a and b, RPS in c and d). In a and c, the shading shows the skill score SKS of the hindcast with the score S of the hindcast when $\sigma_T^2 = 0$ as reference (see text for details). In b and d, the shading shows the skill score of the forecasts with the same reference. White contour lines indicate the prescribed correlation skill α of the hindcasts. Note that the triangular shape arises due to $\alpha^2\sigma_x^2 + \sigma_T^2 < 1$.

Since we are mainly interested in the improvement of the score in the presence of a trend relative to when there is no trend, we compute the skill scores described in Section 3.2.4, namely the CRPSS and RPSS (shown by the shading in Figure 3.2 a, b and c, d, respectively). Note that this means considering the relative improvement of the score with respect to the score of the hindcast at $\sigma_T^2 = 0$ (i.e. on the y-axis of Figure 3.2a for the CRPSS and on the y-axis of c for the RPSS). It can readily be seen that the increase in the skill scores is mainly a function of σ_T^2 and only depends weakly on the fraction of detrended variability that is predicted by the system (σ_ε^2). Only the forecast CRPSS (shading in Figure 3.2) shows a slight tilt with respect to the y-axis. Furthermore, for σ_T^2 being smaller than approximately 60% the increase in skill score is relatively linear with σ_T^2. This allows us to easily quantify the improvement of a score for the hindcast and forecast periods for a given amount of variance explained by the trend in the time series. In case of the CRPSS (both hindcast and forecast) this improvement is approximately 0.6 (i.e. 6% for a 10% increase in

σ_T^2). The RPSS increase of the hindcast is slightly higher at 0.7. In Figure 3.2d the forecast RPSS increase is 1.3 but note that this increase is specific to our choice of $L_{fc} = 1050$ and will be higher (lower) for longer (shorter) forecast periods. This is shown in Figure 3.3 where it can be seen that for a constant trend, the RPSS is higher when computed over a longer forecast period (i.e. moving up the y-axis). This will result in a steep increase in RPSS with σ_T^2 for longer L_{fc}.

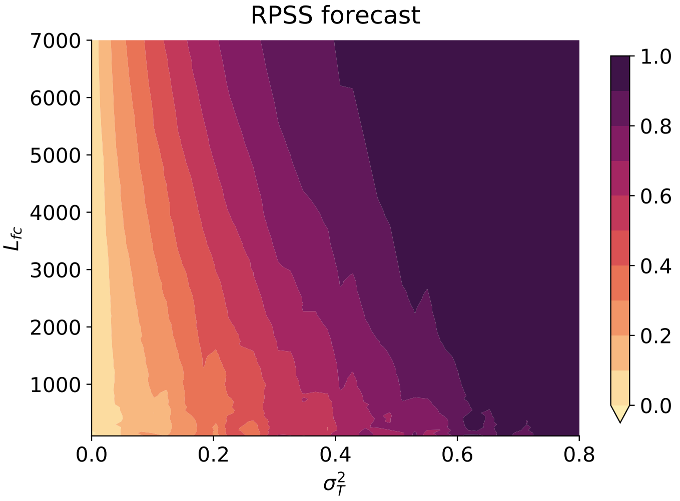

FIGURE 3.3: Forecast RPSS with the RPS of the hindcast without trend as reference as a function of the variance explained by the trend σ_T^2 in the hindcast period and the length of the forecast period L_{fc} that the score is averaged over.

For the RPS it is clear from Figure 3.1 that the presence of a trend results in an artificial increase of the RPSS and that this increase is enhanced during the forecast period. The CRPS does not rely on the definition of category thresholds and thus the increase of the CRPSS is somewhat less intuitive. Using the decomposition of the CRPS into its reliability component and the potential CRPS (equation (3.3)) we can obtain a notion for the reasons for this improvement. The reliability component of hindcasts and forecasts remains close to zero for all σ_T^2, which is a result of our prediction system being almost perfectly calibrated for any choice of the parameters. Thus, it must be a decrease in the potential CRPS that drives the improvement of the CRPS. This reflects the fact that with increasing σ_T^2 the fraction of detrended variance decreases. Since our prediction system reproduces the trend and is well-calibrated its ensemble spread decreases as well. This results in forecasts with narrower ensemble distributions at any given time step (more sharpness) while at the same time the number of outliers (verifications falling outside of the ensemble spread) remains constant, manifesting in a lower CRPS$_{pot}$.

We saw that when the prediction system reproduces the trend in the verification perfectly, hindcasts and forecasts appear to gain skill (i.e. have decreasing CRPS and RPS) the larger the trend is. This happens despite the fact that their skill at predicting the detrended variability does not change. The decisive factor for the apparent improvement is the amount of variance explained by the trend σ_T^2. For the CRPS an increase in σ_T^2 leads to an increase in the sharpness of the prediction ensemble. The effect on the CRPS is the same in the hindcasts as in the forecasts. In contrast, the RPS in the forecast period shows much stronger sensitivity to the trend than in the hindcast period when the terciles are estimated from the hindcast period without accounting for the trend. For a strong enough trend, forecasting the right tercile can become trivial and does not depend on the skill at forecasting any of the detrended variability anymore.

3.3.3 Effects of Mis-Estimation of the Trend in Forecasts

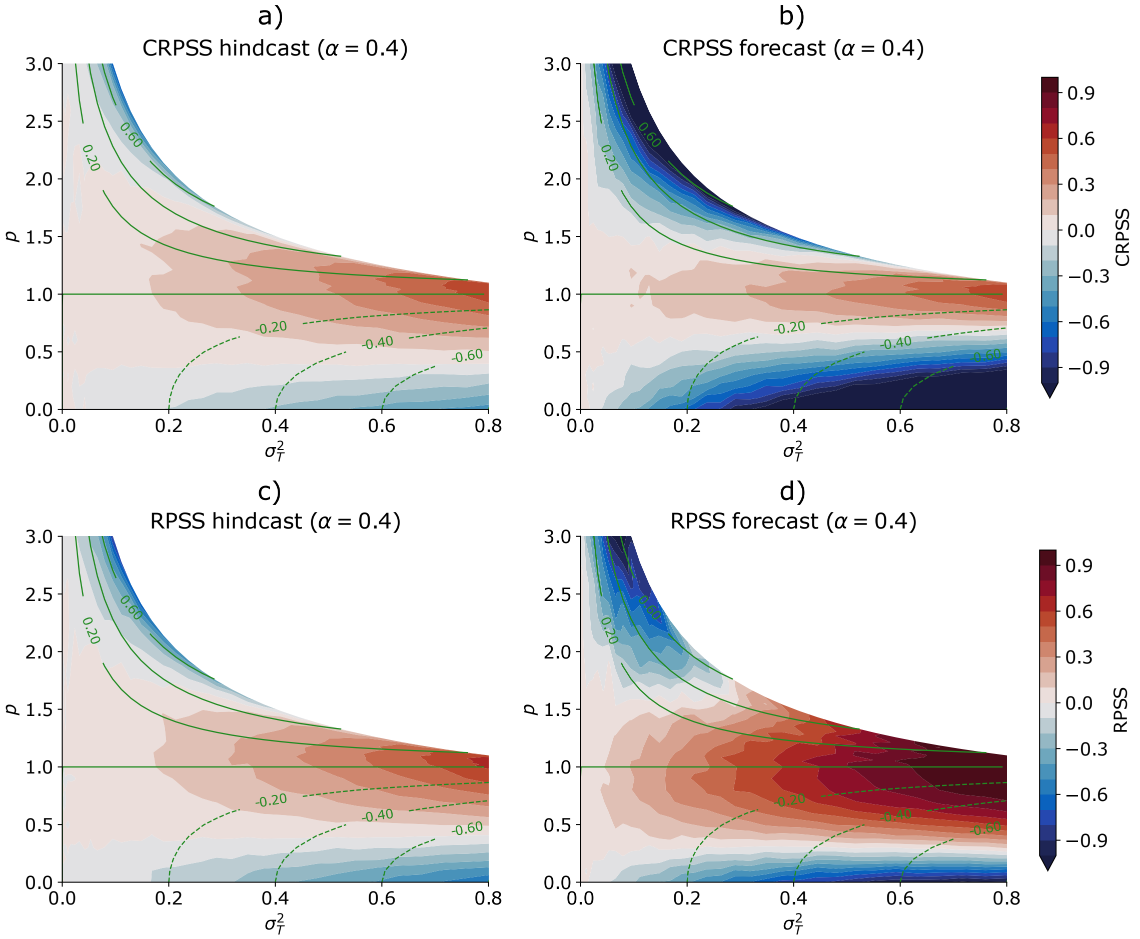

FIGURE 3.4: CRPSS (a, b) and RPSS (c, d) with respect to a hindcast with $\sigma_T^2 = 0$ as a function of variance fraction of the verification explained by the trend (σ_T^2) and the factor of the mis-estimation of the trend in the forecasts (p) for a correlation skill of the forecasts of $\alpha = 0.4$. Green contour lines show the difference (absolute error) between the hindcast's trend variance and the verification's trend variance, i.e. $\Delta_T = \sigma_\tau^2 - \sigma_T^2$. The green zero contour is equivalent to $p = 1$, i.e. where the forecasts predict the trend in the verification perfectly.

Since we do not expect an actual prediction system to perfectly reproduce the observed trend, we test the sensitivity of the skill improvement to the error in the trend. Thus, we next allow the parameter p (see Equation 3.10) of our model to vary where $p < 1$ means an underestimation of the trend in the prediction system and $p > 1$ an overestimation. The effect is shown by the shading in Figure 3.4 as the score (CRPS in a and b, RPS in c and d) of the hind- and forecasts with respect to the score of the hindcast of a time series without trend, i.e. the skill scores as we defined them in Section 3.2.4. We plot the skill scores as functions of σ_T^2 and p. Note that the choice of p as a coordinate results in lower absolute errors $\Delta_T = \sigma_\tau^2 - \sigma_T^2$ occupying a larger area in the plot, which is evident from the green contour lines that show Δ_T. For the figure, we use a fixed level of $\alpha = 0.4$, which is the approximate global average correlation skill for week 3. The figure would look similar for higher (lower) levels of α but with the red areas becoming narrower (wider) for all panels.

Along the straight green line in Figure 3.4 ($p = 1$) we see the same near-linear increase of the skill score as in the respective panels of Figure 3.2. However, even when the trend is over- or underestimated (moving up or down, respectively, relative to the horizontal green line) the skill can still benefit from the presence of a trend in the time series, although to a smaller extent. In every panel of Figure 3.4, there exists also some level of mis-estimation p for which the skill does not increase anymore with σ_T^2 but instead deteriorates (blue shading). The value of p where the skill score changes from positive to negative is similar for the hindcast CRPSS and RPSS but it varies between hindcasts and forecasts. For instance, while the hindcast CRPSS of a system that only assigns half of the actual variance to the trend ($p = 0.5$ in Figure 3.4a) does not change with a stronger trend, its forecast CRPSS (Figure 3.4b) actually decreases. The forecast RPSS in comparison shows a relatively large red area where the skill is improved despite a mis-estimation of the trend. The red area mostly extends to lower values of p ($p \approx 0.25$), i.e. stronger underestimation still allows a positive forecast RPSS. The limit for an overestimation ($p \approx 1.6$) does not change much but above this limit the negative effect on the RPS is much stronger than in the hindcasts.

In summary, this shows that the skill can be positively affected by the trend even if it is not perfectly reproduced in the forecasts. In order to improve the forecast CRPS, the trend cannot be mis-estimated too strongly. In contrast, the forecast RPS still improves for relatively strong underestimation (low p).

3.4 Trend Effect in the Hindcast Skill of a Subseasonal System

In order to compare to the synthetic forecasts from the example above (Section 3.3.3), we now analyse the behaviour of the ECMWF hindcasts. We first consider the trends of the verification (ERA5) over the hindcast period. The trends of the standardised and absolute anomalies are shown in Figure 3.5 a) and b), respectively. Over large parts of the globe the 20-year trends are generally in agreement with global warming signals computed over much longer periods (compare with IPCC, 2013, e.g. Fig. 2.21). For instance, strong absolute temperature trends (Fig. 3.5b) are found in the Arctic regions, consistent with the well-known Arctic amplification signal (Screen and Simmonds, 2010). We can further see enhanced warming over Siberia, as well as generally stronger trends over land than over the ocean. There is also a pronounced lack of warming in the North Atlantic, consistent with the North Atlantic warming hole (Drijfhout et al., 2012). Despite these consistencies there are also some differences to longer term trends. Specifically, negative trends around the Labrador Sea region and south of South America, the relatively strong warming over large parts of the tropical and eastern North Pacific ocean and some smaller scale trend patterns are not typically identified as long term temperature change signals. The reason for this is that the available hindcast period (1998 – 2017) is not long enough to average out decadal to multi-decadal variability. The warming in the Pacific in Figures 3.5 is a manifestation of the Pacific Decadal Oscillation having mainly resided in its negative phase from the late 90s until approximately 2015 and a subsequent switch to more positive conditions within the last couple of years (see Fig. 1 in Newman et al., 2016). Thus, although the linear trends computed from the hindcast period are not purely a manifestation of global warming they do represent deviations of each grid point's temperature time series from stationarity during the hindcast period and can thus be treated approximately as the trends in our synthetic model.

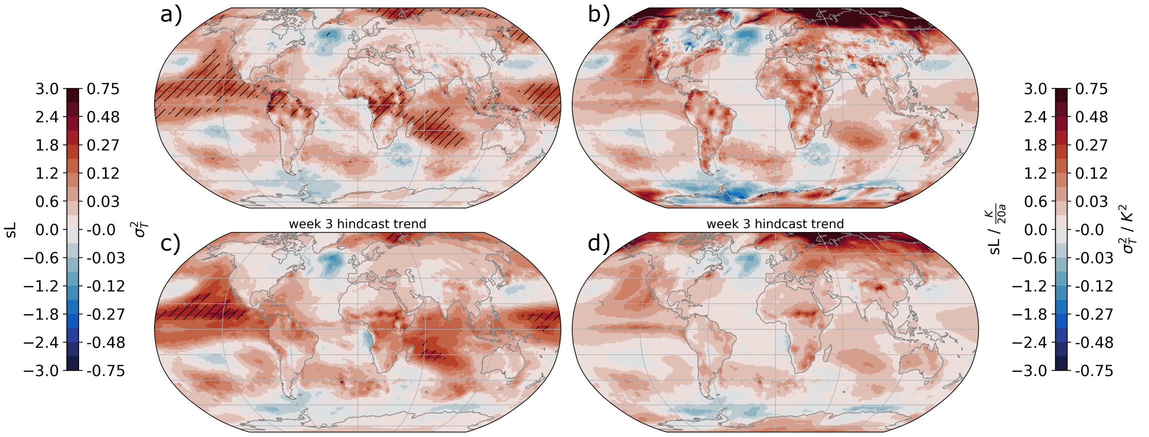

FIGURE 3.5: Trend (1998 – 2017) of the dimensionless standardised T_{2m} anomalies in ERA5 (a) and the week 3 hindcasts (c). Trends of absolute T_{2m} anomalies for ERA5 and hindcasts at 3 weeks lead time are shown in b) and d), respectively. Left colour scale refers to a) and c), right scale to b) and d). Values to the left of the colour scale give the trend as the product of the estimated slope and the length of the time series (in standard deviations per 20 years for a and c and in Kelvin per 20 years for b and d). Values to the right of the colour scale give the trend in terms of its variance over the entire period, which for the normalised trends is equivalent to the fraction of variance of the time series at each grid point that the trend accounts for. Hatching in a) and c) shows where the trend explains more than 10% of the variance of the time series.

In Section 3.3.2, we saw that the effect of the trend on the prediction skill is determined by the amount of a time series' variance that is explained by the trend. The absolute trends are thus not appropriate to identify regions where the forecast skill could be affected by the trend. Figure 3.5a shows the variance explained by the trends of the standardised T_{2m} anomalies from ERA5 in the hindcast period. Clearly, in many regions with strong absolute trends the trends are actually small in comparison to the week-to-week variance and thus are not necessarily contributing strongly to the forecast skill. This is the case for large parts of the Arctic (except for the Barents Sea and parts of the Kara Sea) and many land areas. Instead, in terms of the variance explained by the trend, the oceans and tropical belt show rather stronger signals. The regions where the trend explains more than 10% of the variance are hatched in Figure 3.5a). Following the synthetic model from the previous section, the hindcast skill could be enhanced by up to 5% (not accounting for a mis-estimation of the trend) and the forecast skill could be enhanced even more strongly. We expect the skill to be most affected by the trend in these highlighted regions.

The temperature trends estimated from ERA5 are reasonably well reproduced by the forecast model (Figures 3.5 c and d). On average, there is a weak tendency of the model to underestimate the trends, both in absolute terms and when considering the amount of variance they explain. However, there are substantial spatial variations in this mis-estimation, which is discussed further in Section 3.4.2. The fact that the mis-estimation is largely similar independent of whether we consider the absolute trends or their contribution to the variance indicates that the total variance in ERA5 is well reproduced in the forecast model (not shown). Although we only show the trend computed from the hindcast temperatures at week 3 to represent a subseasonal lead time, in any other forecast week, the fraction of variance explained by the trend is almost the same as in week 3 (not shown).

3.4.1 Comparison of the ECMWF System with the Synthetic Model

Knowing the trend patterns in the hindcast period and their differences in both the verification and the hindcasts, we now assess how the probabilistic skill of the real prediction ensemble behaves in comparison to the synthetic ensemble from the previous section. For this, we sort the grid points of the model by the variance explained by the trend in the verification σ_T^2 and the ratio p of the relative trend slopes in the hindcasts and the verification and bin the data into 100 bins in each of these dimensions (same dimensions as in Figure 3.4). We then compute the bin-average of the scores of both hindcasts and forecasts with reference to the detrended hindcasts (as defined in Section 3.2.4). These are shown for forecast week 3 in Figure 3.6. The overall pattern that the simple synthetic model shows is clearly also present in the real system's skill: the skill for both hindcasts and forecasts increases with an increasing amount of variance explained by the trend and the strength of the trend effect on the skill score tends to decrease the more p deviates from 1. The agreement with the synthetic model is much clearer in the hindcasts (3.6 a and c), mainly due to the larger sample size leading to less uncertainty in the estimates of the average scores. Even the approximate magnitude of the skill score is reproduced (note the different colour scales between Figures 3.4 and 3.6). However, the hindcast CRPSS is negative for weak trends even if the trend is perfectly reproduced. This could be due to the fact that the actual trends are not entirely linear, which could generate additional errors when detrending. Although only half of the populated bins in Figure 3.6a are blue, due to the distribution of the grid-points in the bins 83% of the points fall into these blue bins where the CRPSS on average is negative. Thus, in the hindcasts the CRPS is only improved by the presence of a trend in small parts of the globe — in most parts, it is weakly decreased. This is slightly different for the hindcast RPSS (3.6c). In the synthetic model in Figure 3.4c, the RPSS showed a very similar improvement with σ_T^2 as the CRPSS. For the ECMWF system, the RPS is improved over a somewhat broader area. Only 21% of the populated bins are negative on average and these occur mostly where the trend is very weak. Even for significantly underestimated trends ($p < 0.5$) the RPSS remains positive, which is in contrast to 3.6c. Thus, the RPSS appears to be more strongly affected by the trend in the real prediction system. Again note that the improvement does not happen over a majority of grid points — about 57% of the grid-points lie in the blue bins.

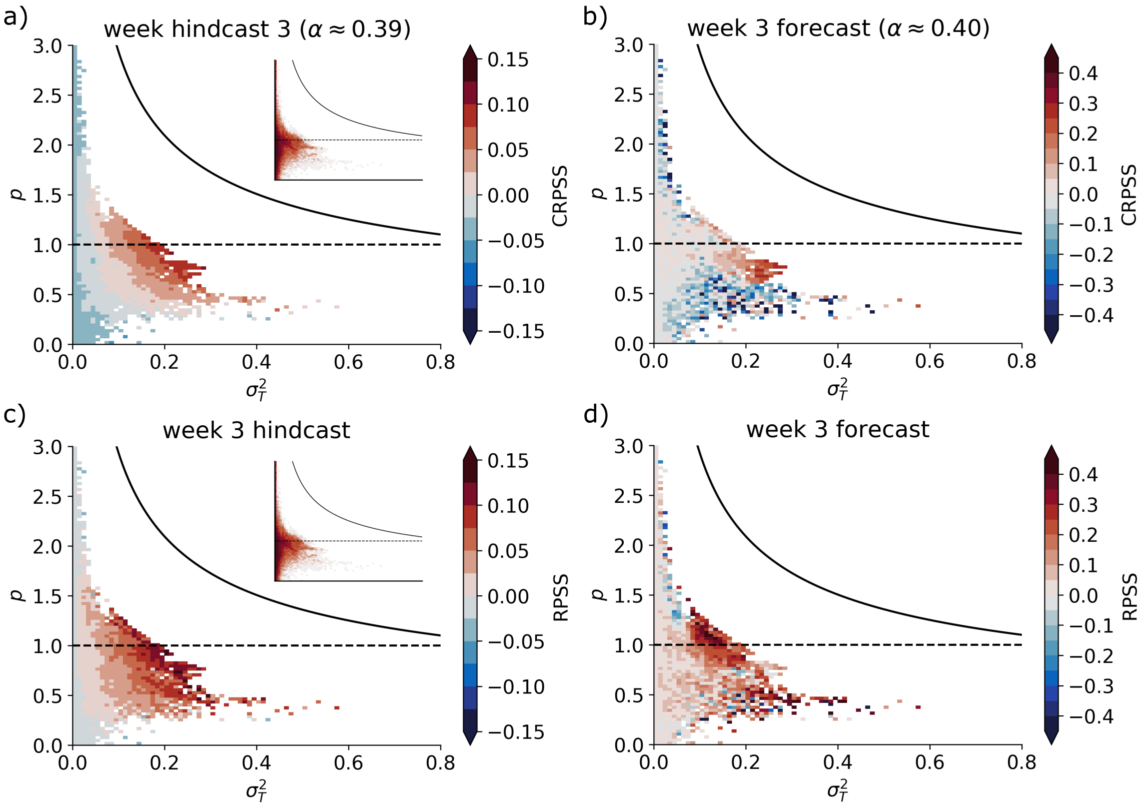

FIGURE 3.6: Skill scores (CRPSS in a and b, RPSS in c and d) of hindcasts (a and c) and forecasts (b and d) as defined in Section 3.2.4 with reference to the score of the detrended hindcasts at every grid point of the model (colours) for week 3 in the same coordinates as Figure 3.4. Every grid point of the model was sorted into the plot according to the variance explained by the trend in the verification and the factor p by which the trend is mis-estimated by the forecasts. The skill scores at the grid points are then averaged over 100 × 100 equally spaced bins between $0 \leq \sigma_T^2 < 1$ and $0 \leq p < 3$ (without weighting by grid-cell area). The inset in a shows the density of points in each bin on a logarithmic scale (valid for all panels). The black dashed line shows where the observed trend is perfectly reproduced by the forecasts. The black solid line shows the theoretical upper limit for p, which is $p_{mx} = \frac{1}{\sigma_T}\left(1 - \alpha^2\sigma_x^2\right)$ where we use the average of the correlations in the hindcasts (a and c) and forecasts (b and d) at all grid points as α.

In the forecasts, the comparison with the synthetic model is hampered by the uncertainty in the estimation of the CRPS as is evident from the noisier distribution in Figures 3.6b and d. However, some aspects of the synthetic model are still visible in the forecasts. The forecast CRPSS in the ECMWF system exhibits an average improvement with increasing trend similar in magnitude as in the synthetic model (note the different colour scale) although this is difficult to say with certainty due to the real trend variance σ_T^2 remaining below 0.2 for most grid-points. More evidently, the red area, i.e. the range around $p = 1$ where there is still an improvement in CRPS despite a wrong trend appears narrower than for the hindcasts in Figure 3.6a, which is in line with the synthetic model. In Figure 3.6b, 62% of the populated bins have a negative average CRPSS (blue bins) as opposed to 48% in the hindcasts in panel a. For the forecast RPSS, it becomes more difficult to compare with the synthetic model as the noise in the estimates is higher as would be expected for a categorical score. This manifests in the fact that only 4% of the grid-points lie in blue bins, although 38% of grid-points have a negative RPSS. The general tendency of a higher forecast RPSS with a stronger trend is still visible but the strongest improvements in RPS actually lie above the $p = 1$ line for the real prediction system. Despite these uncertainties it is clear that the improvement of the RPS in the presence of a trend is strongest for the forecasts (compare with Figure 3.6d with c) and stronger than the enhancement of the CRPS (compare to Figure 3.6b).

In total, the estimates of the average trend effect on the hindcast and forecast skill of the ECMWF model agree well with the trend effect simulated by our simple toy forecast model. We demonstrated the similarity for week 3 forecasts, but the resemblance for other weekly averaged lead times is the same and even stronger in forecast week 1. However, one main difference is an asymmetry around $p = 1$ such that for a constant σ_T^2 the scores improve more strongly for overestimated than for underestimated trends. It can also be seen from Figures 3.6b and d that a large sample size (i.e. large number of initialisations) is necessary to clearly observe the mean effect of a trend on the probabilistic skill in the full prediction system. This shows that the synthetic model can be useful for giving an estimate for the enhancement of the probabilistic skill in the presence of a trend.

3.4.2 Geographical Distribution of the Trend Effect in Forecasts

Having seen that the trend in fact has an influence on the forecast skill of the real prediction system we now consider the geographical distribution of the effect. As became clear from Figures 3.6b and d, there is no ubiquitous improvement of the skill by the presence of the trend. Instead, the skill can even be hampered by the mis-estimation of the trend. The areas where the probabilistic skill of the model increases due to the trend effect is indicated in red in Figures 3.7 (CRPSS in a, RPSS in b, both with respect to the respective score of the detrended hindcasts). The improvement of the forecast CRPS is evidently weaker than for the RPS as was discussed in the previous sections. Additionally, not all areas where the scores are enhanced/hampered agree between the CRPS and RPS. The reason for this is that the CRPS is more sensitive to the mis-estimation of the trend in the model (compare with Figure 3.4b and d). The most prominent regions with this mismatch are the tropical Atlantic and Africa south of the equator. Here, the CRPSS is clearly negative while the RPSS is strongly positive. There is a clear trend in this region, which is strong over the continent and somewhat weaker over the ocean (see Figure 3.5a). However, the trend is quite strongly mis-estimated in this region and p deviates substantially from 1 (Figure 3.7c). While the CRPS does not improve due to this mis-estimation of the trend, the RPS benefits despite of it. The effect of the relatively strong trend outweighs the effect of the mis-estimation. The same is true for other regions where CRPSS and RPSS disagree in sign.

In summary, it can be seen from the Figure 3.7 that the effect of the trend on the forecast CRPS is small globally but can become substantial regionally (e.g. tropical East Pacific south of the equator). At the same time, the mis-estimation of the trend in the prediction system can also hamper the skill locally (e.g. off the coast of West Africa south of the equator). A categorical score like the RPS is enhanced much more strongly than the CRPS.

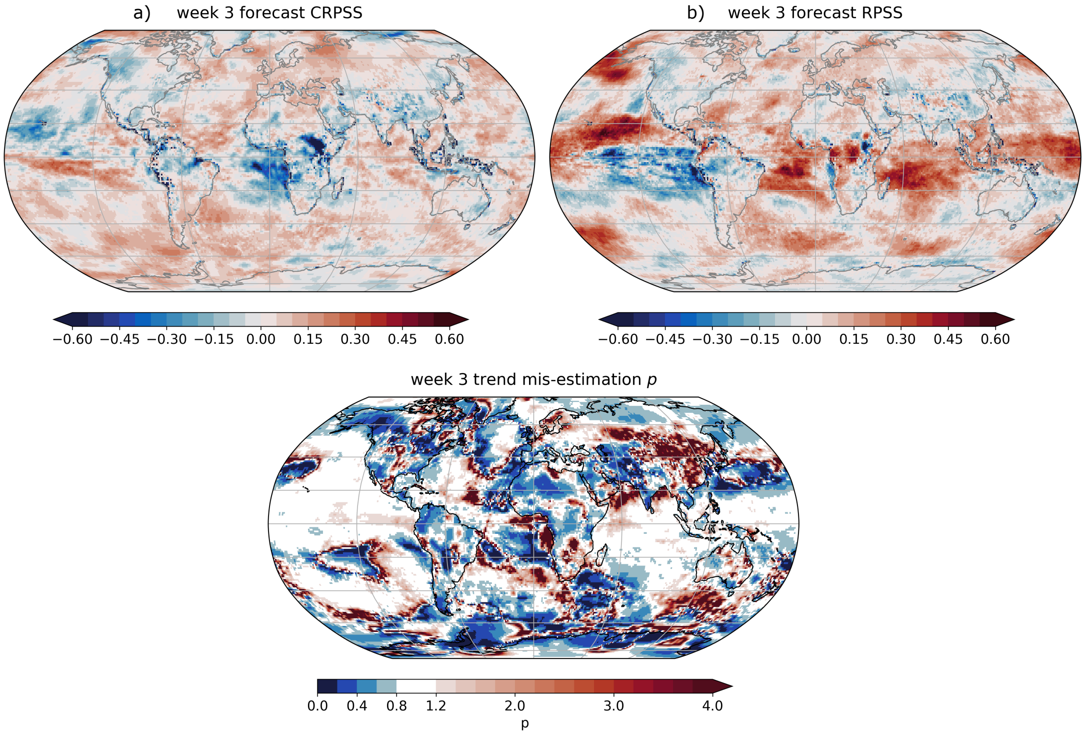

FIGURE 3.7: CRPSS of the ECMWF week 3 forecasts with respect to the detrended week 3 hindcasts (a) at every grid-point (excluding latitudes higher than 81°N). Panel b as a but for the RPSS. Panel c shows the mis-estimation factor p between ERA5 and the model in the hindcast period.

Although it can also be hindered by the mis-estimation (e.g. eastern tropical Pacific), especially in the tropical belt and the North Pacific, the RPS can be improved significantly in the presence of the current trends. The effect is less visible in the Arctic where absolute temperature trends are strong but are small compared to the high week-to-week variability in these regions.

3.5 Discussion

We have used a simple synthetic forecast model to show that probabilistic forecast skill can be enhanced in the presence of a trend. While a prediction system is correct in forecasting warmer temperatures more frequently when there is a positive temperature trend, in a strict sense this is not any more skill than forecasting climatology is. In fact, a trend can be understood as a shift or non-stationarity in the mean of the climatological distribution and can be estimated reasonably well from a sufficiently long past period. Especially in the case of categorical forecasts, it is readily visible from a simple illustration like in Figure 3.1 why a trend in the verification period can have an effect on the predicted category and thus on the skill score: if the categories are defined as percentiles of the climatological distribution and estimated under the assumption of a stationary climate in the hindcast period, for a positive trend, the upper category will be more likely to occur in the forecast period. A prediction system that knows only this trend but has no actual skill will appear to have skill since its (random) forecasts also lie in the upper category more often. This concept was illustrated and discussed before by Livezey (1999, see specifically their Figure 10.2). Our results confirm Livezey's arguments quantitatively for a categorical skill score, namely the RPSS. We further show that there is an enhancement of the probabilistic skill for continuous forecasts (namely, the CRPSS) that do not rely on the definition of categories that are estimated from the hindcast period.

The problem of skill improvement through a failure to account for varying climatologies has also been addressed by Hamill and Juras (2006). They showed that categorical skill scores (both deterministic and probabilistic) can appear higher when locations with different climatologies are pooled to estimate the climatological reference as opposed to accounting for the different event frequencies when scoring the forecasts. Our results are in line with this finding, the difference being mainly that we do not consider spatially varying climatologies but a monotonously changing (and thus non-stationary) climatology over the prediction period. The effect of this monotonous change is especially obvious for the RPS — in case the trend is not accounted for, the average RPSS will increase when evaluating a longer forecast period, which is at odds with our understanding of skill. This problem is in fact one of the reasons why the subseasonal hindcasts at the ECMWF are computed for only 20 years instead of 30 – 40 years, which is a common hindcast period for seasonal forecasts.

With the simple statistical model we designed, we can furthermore quantify the above described effect as a function of the variance explained by the trend in the hindcast period. The estimated effect of a trend that explains 10% of the variance in the hindcast period is to improve the hindcast and forecast CRPS as well as the hindcast RPS by 6 – 7%. As discussed above, for the forecast RPS, the improvement additionally depends on the length of the forecast period L_{fc} and will be increasing with larger L_{fc}. In the selected set-up, which mimics the verification that we apply to the real prediction system (3 years of forecasts, 20 years of hindcasts), the improvement is an estimated 13% for a trend accounting for 10% of the variance.

Since we do not expect a real prediction system to perfectly reproduce the observed trend in the hindcast period at every location, we also tested the effect of mis-estimating the trend in the synthetic model (described by the ratio p of the trend variance in the forecast to the trend variance in the verification). While the RPS increases over a broad range of values of p around a value of one, for the CRPS to increase in the presence of a trend, p cannot deviate too strongly from one. The exact range for p in which there is an improvement of the skill depends on the skill of the prediction system at predicting the detrended variability in the verification.

The real prediction system on average broadly follows the behaviour of the synthetic prediction system in terms of the effect of a trend on the probabilistic skill. This is despite the fact that the toy model represents a strong simplification of the statistical properties of the real forecasts. Nevertheless, we want to point out the main limitations on the realism of the synthetic model that likely play a role in causing the differences with respect to the real prediction system. These are largely coincident with the limitations discussed by Weigel et al. (2008a) who used a similar set-up of synthetic forecast-verification pairs. For one, the verification is drawn from a normal distribution at every time step. Although this assumption might hold in reality for weekly means of near-surface standardised temperature in many regions of the globe, it will certainly be violated in places where temperatures are subject to feedbacks at certain times of the year. This could for instance be the case for ocean regions with sea ice during parts of the year, land regions where snow melt occurs or places in which land-atmosphere feedbacks tend to be relevant. This strongly non-linear behavior is the reason why we exclude grid-points in the high Arctic. Secondly, just as the synthetic verification follows a normal distribution so do the synthetic predictions. This leads to almost perfect calibration, i.e. an optimal reliability component of the CRPS ($\text{REL} = 0$). Since we apply a rather crude calibration to the real predictions, the reliability is certainly worse than optimal (i.e. $\text{REL} > 0$). This affects the considered scores (Ferro et al., 2008) and we hypothesise that it is responsible for large parts of the deviations between Figures 3.6 and 3.4. Another limitation on the realism of the synthetic model is that it only considers a linear change in the mean of the climatology. Non-stationarity can be assumed to be more complex and affect other moments of the climatological distribution (e.g. Schär et al., 2004). Related to this, the real trend itself likely exhibits seasonal variations. Despite these limitations, the synthetic model describes the average behaviour of the ECMWF prediction system quite well.

We showed that the trend effect on the forecast skill varies throughout the globe. While it is a function of the fraction of variance accounted for by the trend, it also depends on how well exactly it is reproduced by the models. In fact, the positive effect of a trend on the score of a forecast can be outweighed by the negative effect of a mis-estimation of the trend in the prediction system. This is seen particularly in the tropics for the CRPS. Globally, the CRPS of weekly subseasonal temperature forecasts is only weakly enhanced by the trend. The RPS, on the other hand, is much more strongly affected. In most regions of the tropics (except for the eastern Pacific south of the equator) the RPS is quite strongly enhanced by up to 60% locally due to the relatively large contribution of the trend to the variance of the time series in tropical locations. In the Arctic, where trends are large but internal variability is also strong, the trend enhances the skill in most regions but much less severely.

It can be argued that the estimation of a trend from a 20-year period from ERA5 is subject to large uncertainties, and we showed in Section 3.4 that our estimate does not separate the trend from multi-decadal variability. We do not see this as an issue for the current analysis since the ECMWF prediction system produces a trend pattern that is largely in agreement with ERA5. For our statistical separation of the trend effect it does not matter whether the trend is part of a multi-decadal variation or a longer term change. In an operational setting, the estimate is in any case limited by the available hindcast period and thus it is common practice to compute the trend from at most a few decades. Attributing the trend effect to anthropogenic climate change, however, would require a robust estimate of the long-term trend.

3.6 Conclusion

A trend in a time series represents a non-stationary component in the climatology. We have shown that a simple linear trend can improve the forecast scores of a toy forecast even in the absence of any actual predictive skill. This skill improvement is a function of variance explained by the trend but also of the mis-estimation of the trend p in the prediction system. The effect is stronger and less dependent on p for categorical scores (RPS) and is further enhanced when averaging the RPS over a longer forecast period. The forecast CRPS can also be enhanced but the mis-estimation of the trend in prediction systems can impair the skill and outweigh the positive effect of the trend. The effect simulated by a simple synthetic model is in good agreement with the average effect in a real prediction system (the ECMWF extended-range forecasting system). While the subseasonal CRPS is only weakly affected by the trend effect, the RPS of the subseasonal forecasts can be strongly enhanced regionally. This effect is strongest in the tropics where the trend accounts for a larger part of the variability than in other regions. Even though we focused our attention on subseasonal forecasts, our results can be generalised to some degree to forecasts at any time scale and lead time. The trend effect is a function of the signal-to-noise ratio, i.e. the trend relative to the internal variability on the considered time scale. Thus, considering for instance decadal forecasts where annual means are predicted, we would expect to see a stronger effect of the trend than for the prediction of weekly means in subseasonal forecasts, mostly because the trend will be larger in comparison to the interannual variability than to the weekly. Similarly, spatial aggregation has an effect on the signal-to-noise ratio and thus the effect on skill scores of regional averages will be different than shown here even when the same time scale is considered. Our results allow for a simple benchmark estimate of the contribution of a long-term trend on the skill of a forecast. This is important to quantify since the trend can be estimated from a period prior to the forecast time and it can be argued that this needs to be accounted for as part of the climatological forecast that serves as a reference when computing the skill. Clearly communicating where the skill of our forecasts stems from can enhance the users' confidence in these products.

Acknowledgements

We thank the ECMWF for providing extended-range forecasts from the S2S prediction project through the MARS archive (`https://apps.ecmwf.int/datasets/data/s2s/`) and Copernicus for providing the ERA5 data through the Climate Data Store (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview`).
All analyses and visualisations were carried out using the Python packages numpy, matplotlib, scipy and properscoring.
Funding from the Swiss National Science Foundation to O.W. and D.D. through project PP00P2_170523 is gratefully acknowledged.

Chapter 4

Seasonal Variations in Subseasonal European Near-Surface Temperature Prediction Skill

The Supporting Information for this chapter can be found in Appendix B.

Abstract

There is an evident need for forecasts of the atmosphere weeks to months ahead and it is crucial to inform users about the times when subseasonal forecasts are potentially more valuable. In this study, we consider the question whether there are times of the year in which subseasonal forecast skill for large-scale averages of European temperatures in the ECMWF model is enhanced compared to other times of the year. We find that there is a significant seasonal cycle in the skill that is consistent with the month-to-month variations of the ensemble spread. Skill in winter, especially in February/March, is generally highest, while it is lower in late summer and early fall, reaching its lowest point in September/October. There is furthermore a marked drop in skill from March to April. In winter, the best predicted forecast situations tend to be associated with warm temperatures over Eurasia and an associated zonal configuration of the large-scale flow. In fall, no clear average pattern of temperatures or circulation arises for the best predicted forecasts. In both seasons, the poorest temperature forecasts occur in blocked flow conditions. For the best forecasts in winter we find that there are distinct anomalies in some of the well-known predictors of the North Atlantic winter climate. The composite of the best winter forecasts on average exhibits cold conditions in the lower stratosphere during initialization and a convection dipole over the Indian Ocean and Maritime Continent that is reminiscent of phases 2 and 3 of the Madden-Julian Oscillation. Using an index for these predictors to split the set of forecasts, we show that the stratospheric state at initialization is a useful predictor of subseasonal forecast skill for large-scale European near-surface temperatures. Despite the complex nature of the tropical-extratropical teleconnections to the North Atlantic, our index also exhibits some potential as a predictor of skill, most strongly during El Niño years. Our results can inform users *a priori* about the skill of subseasonal forecasts relative to other months. Within the winter months our study can additionally provide

information on the skill relative to forecasts initialized under different conditions in the stratosphere and in the tropics.

4.1 Introduction

Subseasonal forecasts provide predictions of hydro-meteorological fields at lead times from approximately two weeks to two months. Although the potential for successful subseasonal forecasts has long been suggested (Miyakoda et al., 1983), the field of subseasonal prediction only started growing rapidly in the last years with the launch of two large projects dedicated to subseasonal prediction, namely the Subseasonal-To-Seasonal (S2S) Prediction Project (Vitart et al., 2017) and the Subseasonal Experiment (SubX, Pegion et al., 2019). The growing interest is driven to a large degree by the evident need for forecast information on S2S time scales in a number of different sectors (Perez et al., 2015; White et al., 2017; Guimarães Nobre et al., 2019). Additionally, numerical weather prediction (NWP) models have continuously improved over the last decades due to increased physical understanding and technical advances (Bauer et al., 2015) and nowadays it is possible to use NWP models to provide skillful and useful forecasts in the subseasonal time range, which had long been deemed a "predictability desert" (Robertson et al., 2020).

Although subseasonal forecasts can provide value to users in various sectors (Grams et al., 2017; Dorrington et al., 2020; Lopez et al., 2020), forecast skill on subseasonal lead times is generally considered low, especially in the extratropics and particularly over Europe (Vitart, 2004). Knowledge of when subseasonal forecasts can be expected to be better or worse can thus help to inform users about the relative value of a forecast provided at a certain time. In this context, the seasonal cycle of the subseasonal prediction skill is important information. Weigel et al. (2008b) evaluate forecasts and hindcasts from the monthly prediction system from the European Centre for Medium-Range Weather Forecasting (ECMWF) from 1994 to 2006. They find that, while there is a clear drop in skill from late winter to spring, the seasonal cycle of subseasonal weekly near-surface temperature prediction skill averaged over the Northern Hemisphere (NH) is generally weak and not significant after forecast day 12. As part of a general evaluation of choices in the verification of subseasonal forecasts, Manrique-Suñén et al. (2020) note that the subtleties of the estimation of the reference forecast used to compute the skill are critical for the reported skill (see also Hamill and Juras, 2006). Only when the reference forecasts accurately reflect the seasonal variations in the climatology, an evaluation of the seasonal cycle of skill is possible without falsely attributing skill to the model for outperforming an inaccurate reference. Manrique-Suñén et al. (2020) further show that the seasonal cycle of week 2 (forecast days 12 – 18) near-surface temperature skill averaged over the North American continent is marginal, but again with a slight tendency for better predictions in winter. These studies suggest that the seasonal cycle in large-scale averages of subseasonal temperature prediction skill is insignificant. However, skill also varies as a function of the spatio-temporal aggregation of the forecast fields (van Straaten et al., 2020). It is thus conceivable that the seasonal skill variations, too, are dependent on the level of aggregation.

In this study, we focus our attention on the prediction skill for large-scale averages of European land temperatures and its month-to-month variations.

Another prospect for providing *a priori* information on the expected trustworthiness of forecasts is through studies of flow-dependent predictability. A common approach for this is to split a set of forecasts into subsets based on differences in their initial conditions. Often, this separation is made by considering the initial large-scale flow configuration through modes of variability or flow regimes. Using the presence of certain North Atlantic (NA) weather regimes during forecast initialization, Ferranti et al. (2015) find that cold season (October – April) forecasts of geopotential height initialized in the negative phase of the North Atlantic Oscillation (NAO) exhibit enhanced prediction skill, most prominently at lead times of 9 – 15 days (see also Matsueda and Palmer, 2018). This is mainly because the forecast model manages to predict the longer persistence of the negative NAO-regime. On the other hand, the model misses transitions to the blocked flow regimes and generally shows a tendency towards favouring a more zonally oriented flow. Since these large-scale regimes have distinct imprints in the variability of surface fields, this likely has implications for the forecast skill for near-surface temperature.

In addition to the local flow conditions during initialization of the forecasts, predictors in regions external to the North Atlantic-European (NAE) sector can potentially provide some information on the relative skill of a forecast. For instance, the state of the NH stratospheric polar vortex (SPV) can exert some control on the tropospheric winter circulation in the NAE sector during phases of strong stratosphere-troposphere coupling and modulate the occurrence of certain regimes (Charlton-Perez et al., 2018; Beerli and Grams, 2019; Domeisen et al., 2020a). On average, the likelihood of a negative (positive) phase of the NAO is increased up to several weeks after weak (strong) vortex events (Baldwin and Dunkerton, 2001; Limpasuvan et al., 2005). Thus, the state of the stratosphere can provide subseasonal predictability for the troposphere (Domeisen et al., 2020b) and indeed, this manifests in better subseasonal prediction skill for the Northern Annular Mode of forecasts initialized under both weak and strong states of the SPV compared to neutral conditions (Tripathi et al., 2015). However, especially forecasts initialized under weak vortex conditions also tend to persist the negative NAO response in the troposphere too strongly, resulting in large-scale errors in the NA climate (Kolstad et al., 2020) that can also negatively impact surface prediction skill in Europe (Büeler et al., 2020). In total, though some flow-dependent predictability from the stratospheric initial conditions is expected, errors in the representation of stratosphere-troposphere coupling often hamper the potential of the stratosphere to act as a source of predictability for the surface winter climate (Domeisen et al., 2020b).

The tropical atmosphere is another source of subseasonal predictability of the large-scale extratropical circulation. The Madden-Julian Oscillation (MJO, Madden and Julian, 1971) is the strongest mode of tropical subseasonal variability and characterized by large-scale diabatic heating anomalies that can trigger Rossby waves, which influence the extratropical climate, also in the NAE region (Ferranti et al., 1990; Cassou, 2008). Specifically the MJO phases that exhibit a strong convection dipole between the Indian Ocean and the

Western Pacific (MJO phases 2, 3, 6 and 7) exert an influence on the winter NAO (Lin et al., 2010). Generally, MJO phases 2 and 3 (6 and 7) tend to be followed by positive (negative) phases of the NAO between 5 and 15 days later (Cassou, 2008) but the exact extratropical response also depends on the propagation speed of the MJO (Yadav and Straus, 2017), the concurrent state of the El Niño-Southern Oscillation (ENSO, Lee et al., 2019) and the stratosphere (Garfinkel and Schwartz, 2017). The influence of the MJO onto the winter climate in the NAE region is sufficiently strong to affect the prediction skill, such that active phases of the MJO tend to be followed by an increase in NAO prediction skill (Vitart, 2014), especially for active MJO phases 2, 3, 6 and 7, that feature a strong convection dipole over the Indian Ocean and the Western Pacific (Lin et al., 2009). Based on these studies, it is expected that there is some flow-dependent predictability of European surface temperatures from the state of the MJO.

More sources of subseasonal predictability have been suggested and could also be active in other seasons than winter. There are indications that the North Atlantic ocean exerts some control on the atmospheric circulation above and the influence is most likely seen in summer and autumn (Nie et al., 2019; Ossó et al., 2020). Eurasian snow cover has the potential to alter the state of the Arctic Oscillation by triggering a stratospheric response (Cohen et al., 2014) but this potential for winter predictability is currently not realized by subseasonal forecast models (Garfinkel et al., 2020). The state of the land surface and especially soil moisture is another source of subseasonal predictability that is likely most relevant in spring and summer (Dirmeyer et al., 2019). Soil moisture is thought to mainly affect the local climate but there are indications that strong soil moisture anomalies can also interact with the large-scale circulation, especially during heat waves when strong land-atmosphere coupling exists (Fischer et al., 2007a).

The multitude of potential predictors of subseasonal temperature variability in the NAE sector and their varying importance suggests that there could also be seasonal variations in the subseasonal skill. From the existing literature, the most promising seasons for subseasonal prediction appears to be winter. Note that the tendency for better temperature forecasts on subseasonal time scales in winter than in summer is in contrast to seasonal predictions, for which slightly higher temperature skill has been reported in summer compared to winter (Baehr et al., 2015; Johnson et al., 2019). One objective of this study is thus to assess whether there exists a seasonal cycle in the prediction skill for large-scale temperature variability over Europe. We use the ECMWF extended-range ensemble forecasts to test our hypothesis that subseasonal prediction skill is enhanced during winter. We further analyse composites of the forecasts with highest and lowest skill to identify common structures in the situations that are more skillfully predicted. An analysis of the conditions in some of the suggested predictor fields during forecast initialization then aims at identifying the potential for flow-dependent predictability. We further examine whether separating the analysed forecasts by their initial state of these indicators can also split the forecasts in terms of their subseasonal temperature skill. The subseasonal forecast model and the verification data are introduced in Section 4.2 along with the methods used to process the data and verify the forecasts. In Section 4.3 we present the results of our analysis of the seasonal cycle of subseasonal skill and the potential predictors of skill. These results are discussed and placed into context with the existing literature in Section 4.4 and the main conclusions are summarized in Section 4.5.

4.2 Data and Methods

4.2.1 Hindcast and Verification Data

To ensure a sufficiently large sample of forecasts, we make use of 20 years of subseasonal ensemble hindcasts from the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF). These are available from the ECMWF as part of the S2S Prediction Project (Vitart et al., 2017). The system is initialized twice weekly (Monday, Thursday) resulting in 105 forecasts per year. For each forecast, 20 ensemble hindcasts with 11 ensemble members are generated by initializing on the same calendar day 1 to 20 years before the actual forecast. We here focus on hindcasts from 1998-2017, which were generated with the cycles 43R3 and 45R1 of the ECMWF IFS. The model is run at Tco639 spatial resolution (approximately 16 km) up to day 14 and Tco319 (approximately 32 km) after. We retrieved daily mean fields of 2-m temperatures (T2m) and instantaneous (00:00:00 UTC) geopotential height at 500 hPa (Z500) at 1° global resolution. Note that we only use hindcasts in this study and refer to them synonymously as hindcasts and forecasts in the following.

As verification data for the hindcasts, we use the ERA-Interim reanalysis (Dee et al., 2011) for the same period downloaded at the same 1° global resolution. Apart from daily mean T2m and instantaneous Z500, we retrieved fields of lower stratospheric temperatures at 100 hPa (T100), as well as zonal and meridional wind components at 200 hPa. These were used to compute the velocity potential at 200 hPa (VP200) using the windspharm python package (Dawson, 2016).

Processing of Hindcasts and Verification

In this study, we focus our attention on 5-day mean, large-scale averaged land temperatures in the European region. The rationale behind this aggregation is that part of the unpredictable noise can be eliminated from the forecasts by averaging, potentially enhancing the predictable signal if a suitable combination of spatial and temporal aggregation is chosen (Toth and Buizza, 2019). Complying with this, we average land temperatures over a box covering a larger European domain (EUR box, shown in Figure 4.2) from 10°W – 50°E, 35°N – 75°N and average all temperatures over 5-day windows. These scales allow us to focus on the large-scale patterns of prediction skill for European temperatures. Note that, in order to maximize the skill, more optimal levels of spatio-temporal aggregation could be found (van Straaten et al., 2020). In the following, when we refer to EUR T2m, it is implied that we mean the averages described above. We use pentad prediction windows starting from day one of the hindcasts that are separated from each other, i.e. pentad 1 refers to averages over days 1 – 5, pentad 2 to days 6 – 10, etc.

Our aim is to identify differences in subseasonal forecast skill between different months of the year. However, these differences should not merely represent the fact that temperature variability is higher in winter than in summer. Since the skill of a system is defined as its score relative to a the score of a reference system, we can either account for these seasonal variations in the reference or eliminate the seasonal cycle from both hindcasts and verification. We choose the latter approach since it allows us to directly use the ensemble spread of the hindcasts without correcting for seasonal variations afterwards. Similarly, we can use the correlation between hindcasts and verification as a deterministic skill score since the correlation for standardized anomalies for a system without skill is 0.

To account for seasonal variations in the temperature distribution we estimate the seasonal cycle of the mean and standard deviation of the temperatures as outlined in Narapusetty et al. (2009). The seasonal cycle of the mean is estimated by first averaging temperatures over all years for each calendar day and low-pass filtering the resulting time series (eliminating periods shorter than 90 days). This smoothed seasonal cycle is subtracted from the absolute temperatures to form anomalies. The seasonal cycle of the standard deviation is then estimated by computing the standard deviation of these anomalies over all years for each calendar day and again applying the same low-pass filter. The anomalies are then divided by this estimate to form the standardized anomalies. In a last step, the data are detrended by fitting a linear trend to the daily standardized anomalies over all 20 years in the hindcast period since long-term trends have the potential to inflate the skill estimates (Wulff et al., 2021). In the case of the hindcasts, all the above steps are performed for each lead time individually to account for drifts in the climatologies of the forecast model. This approach is similar to a simple bias correction with a weekly climatology as described in Manrique-Suñén et al. (2020). From their analyses it can be expected that the approach does not fully correct the bias in the forecast model, which can have a negative effect on the skill. At the same time, it is a more optimal method for accounting for the seasonal cycle, which is a strongly preferable property for our analyses.

4.2.2 Forecast Performance Measures

To assess the performance of the forecasts we compute three different skill scores. A skill score SKS is generally defined as (Wilks, 2019a):

$$\mathrm{SKS} = \frac{S - S_{ref}}{S_{perf} - S_{ref}} \tag{4.1}$$

where S refers to some accuracy measure or score and the subscripts S_{perf} and S_{ref} indicate the score of a perfect and a reference forecast, respectively. Throughout the study, climatology will serve as a reference, which is a more competitive benchmark for a subseasonal forecast over the continents than persistence (Weigel et al., 2008b). As discussed above, we consider forecasts of standardized temperature anomalies with respect to a seasonal cycle. This allows us to use a very simple form of climatological forecasts, namely a Gaussian distribution with zero mean and unit standard deviation at any initialization and lead time.

There are a multitude of different scores used in the verification of forecasts that highlight different attributes of the forecasting system (Wilks, 2011). We introduce three common measures here that we use in the evaluation of the seasonal cycle of skill.

The first score we employ is the Continuous Ranked Probability Score (CRPS), which measures the integrated squared distance between the observed and the forecast cumulative distribution function (cdf) of a continuous variable y and is defined after Wilks (2019a) as:

$$\text{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy \tag{4.2}$$

where $F(y)$ is the forecast cdf for y (here: temperature) and $F_o(y)$ is the observed cdf of the same variable, which is a step function with step from 0 to 1 at the value $y = o$ with o being the observed value of the variable. The CRPS can be easily estimated using its kernel representation (Gneiting and Raftery, 2007). However, in this representation the CRPS depends on the number of members in the forecast ensemble generally resulting in higher CRPS for less members. Leutbecher (2019) thus introduce an adjusted version of the CRPS (CRPS_a) that corrects for the limited ensemble size. Since we compute the skill with reference to a forecast of the climatological distribution (i.e. assuming an infinite number of random draws) we use this adjusted score in the following. Noting that the CRPS of a perfect forecast is 0, we use the CRPSS defined as:

$$\text{CRPSS} = 1 - \frac{\text{CRPS}_a}{\text{CRPS}_{ref}} \tag{4.3}$$

When computing the average skill over a number N of forecast-verification pairs, we use averages of CRPS_a and CRPS_{ref} in Equation 4.1, such that:

$$\langle\text{CRPSS}\rangle = 1 - \frac{\langle\text{CRPS}_a\rangle}{\langle\text{CRPS}_{ref}\rangle} \tag{4.4}$$

where $\langle . \rangle$ indicates an average over N forecast-verification pairs.

To measure the discrimination of the categorical forecasts we use the Relative Operating Characteristic (ROC) diagram (Wilks, 2019a). In the ROC diagram, the hit rate of the forecasts for one specific event (here: temperatures in the upper tercile) is plotted against the false alarm rate for varying decision thresholds (i.e. the number of ensemble members showing the event for which we would issue the forecast that an event will happen). The area under the thus obtained ROC curve is referred to as A and is a scalar summary measure of how well the forecasts discriminate events from non-events. Using A as a score in Equation 4.1 and the fact that A will obtain a value of 1 and 0.5 for a perfect and a random forecast, respectively, the ROC skill score (ROCSS) is given by:

$$\text{ROCSS} = 2A - 1 \tag{4.5}$$

Note that the ROCSS should be interpreted as a measure of the potential skill only since it is insensitive to biases (both conditional and unconditional) in the forecasts (Wilks, 2019a). Furthermore, it is computed for a fixed event threshold and thus cannot be compared directly to the CRPSS.

We further use a simple deterministic score for comparison, namely a correlation between the standardized anomalies of the hindcasts and the verification over a subset of initializations, referred to as CORR. Note that, following Murphy and Epstein (1989) the correlation as a skill score — similar to the ROCSS — should be interpreted as a measure of potential skill only, since it disregards biases in the forecast distribution.

To quantify the uncertainty of the hindcast ensemble, we use the spread of the ensemble averaged over a number of different initializations. We estimate the spread by computing the standard deviation over all 11 ensemble members. Following Weigel (2012), the ensemble spread can provide information on the reliability of the forecasts when scaled to account for the limited ensemble size. The scaling factor is $\frac{M+1}{M}$ where M is the number of ensemble members. For a reliable forecasting system, this scaled ensemble standard deviation (ESTD) should equal the root mean squared error (RMSE). This implies that in the case of forecasts of standardized anomalies and no predictability (i.e. RMSE$=$ 1), the ESTD should approach one as well. Thus, similarly to using a reference score in the evaluation of skill above, we use $ESTD = 1$ as the reference spread of a system without skill.

4.2.3 Sorting by Season and Skill

For our analysis of the seasonal cycle of the skill, we form composites of forecasts for each month. These are obtained by sorting the forecasts first by the month that the first day of the considered prediction window (pentad 3 or pentad 4) falls into. Note that thus, the initialization can lie outside of the considered season since we choose a different set of forecast initializations for each lead time. For instance, a hindcast initialized on January 16 enters the January composite for pentad 3 (days 11 – 15) hindcasts but belongs to the February composite for pentad 4 (days 16 – 20) hindcasts.

In a next step, we divide the seasonal subsets of forecasts further. We sort them according to their skill as measured by their CRPSS in the respective prediction window. We use the 25% of hindcasts in a season with the highest CRPSS as the best hindcast composite and the quarter with the lowest CRPSS as the poorest hindcast ensemble (i.e. skill in upper and lower quartile). We also tested using only the 10% forecasts with the highest/lowest skill and the resulting composite averages show largely the same patterns as in Figures 4.2 – 4.5.

Note that we compute the skill (CRPSS) for each forecast-verification pair as given by Equation 4.3 and use it for the sorting. With this approach we choose the best (poorest) hindcasts as those that perform best (poorest) with respect to a climatological forecast in each individual situation. If we had used the score (CRPS) for the sorting instead, we would only account for the probabilistic error of the forecast and disregard the difficulty of predicting a situation. At the same time, using the skill (CRPSS) in each individual situation can lead to apparent forecast "busts". In these cases the skill is very low, not because the error of the model is large but because the observed situation is close to climatological conditions. Under these circumstances, the reference forecast would perform extremely well (i.e. low CRPS_{ref}) and potentially become very difficult to beat for the forecasts. Thus, when sorting by the *skill*, there is the risk that a low skill composite reflects the forecast situation rather than the forecast performance. However, we argue that our approach of sorting by the skill is still justified and does not solely reflect the forecast situation in the low skill composite. For one, scoring the forecasts with the CRPS does not only reward if the forecast average is close to the verification but also penalizes a lack of sharpness (large spread). Thus, even if the observed situation is exactly the climatological mean (i.e. zero anomaly), the reference forecast still has $\text{CRPS}_{ref} > 0$ since it has no sharpness. In that case, a forecast that has some sharpness can still outperform this reference. A second argument for our approach is that choosing a composite that includes a quarter of all forecast cases is not strongly influenced by the aforementioned "busts". We do indeed see that the low skill composite averages show nearly no anomalies when we choose only 5% for the compositing. However, as mentioned above, even choosing 10% of the forecasts for the composites results in patterns that are in strong agreement with the ones we obtain for 25%. We take this as indication that our approach is not biased towards sorting all climatological situations into the low skill composite independently of the forecast. Instead, it is able to account for both the ability of the forecasts and the ease of predicting a certain situation.

Note that only our sorting is based on the skill (CRPSS) in each individual forecast situation. Wherever we show the average CRPSS over a number of samples, such as in Section 4.3.1, we use the standard approach of computing the average skill, which is given by Equation 4.4.

Statistical Significance The significance of the composite patterns is assessed using a block bootstrap approach (Wilks, 2019b) to allow for the auto-correlation in the skill estimates. Each block is chosen to consist of two consecutive initializations (3 or 4 days apart, by design of the initialization strategy). The reason for choosing two consecutive initializations is that we expect two successive forecasts to exhibit similar skill for forecasts of 5-day means but beyond this time scale, the auto-correlation of the skill is weak. At a lag of 2 initializations (corresponding to 7 days) the auto-correlation of the CRPSS is smaller than 0.1 for all lead times considered here. Thus, we re-sample each season's initializations 1000 times, each time randomly drawing 25% of the initializations with a block length of two and averaging over the samples in this random composite. This gives an estimate of the bootstrap distribution and we consider an anomaly significant if it falls above the 95th or below the 5th percentile of this distribution, respectively.

4.3 Results

4.3.1 A Seasonal Cycle in Subseasonal Temperature Forecast Skill

We now consider the seasonal variations in subseasonal EUR T2m skill evaluated as described in Sections 4.2.2 and 4.2.3. Figure 4.1 shows the ensemble spread and three skill measures for pentad 3 (lead time 11 – 15 days) and pentad 4 (lead time 16 – 20 days) hindcasts. In the ensemble spread (ESTD, Figure 4.1a), there are pronounced seasonal differences between winter and summer for both pentads. While the ensemble spread is lower in the extended winter season (November to March) it is generally higher in summer. Especially for pentad 3, there is a sharp increase from March to April and a similarly sharp drop in spread from October to November. For pentad 4, the seasonal cycle of ESTD is smoother. Interestingly, the spread decreases towards peak summer (July, August) for both pentads, which is contrary to shorter lead times up to 5 days, where the spread attains its maximum in July (Figure B.1).

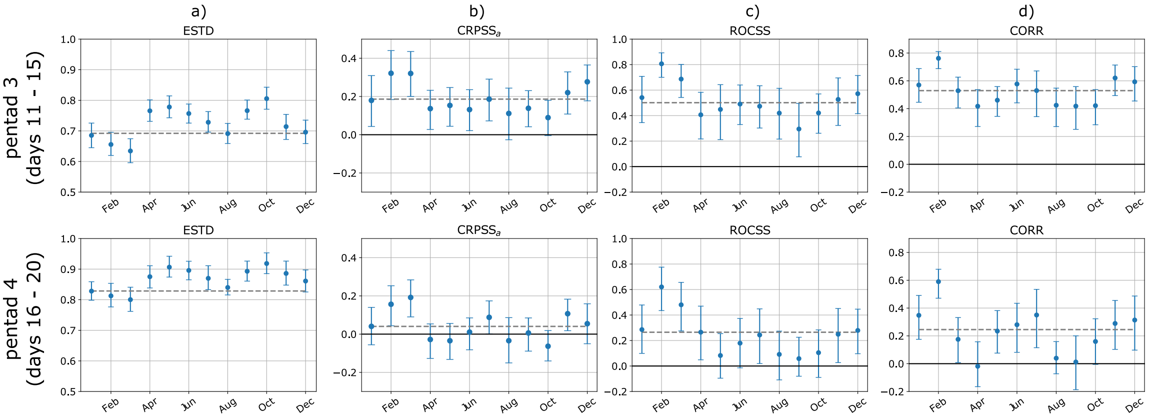

FIGURE 4.1: Hindcast spread (a) and skill (b – d) measures for forecast pentad 3 (days 11 – 15, top) and pentad 4 (days 16 – 20, bottom) as a function of the month that the first day of the verification period falls into. The error bars show the 95% uncertainty interval based on a bootstrap re-sampling with 1000 iterations. The grey dashed lines show the spread/skill score computed over all months. The black horizontal lines in b – d indicate the zero skill lines.

Between all of the skill measures for both lead times (Figure 4.1b–d), the seasonal cycle of the skill is in close agreement although, as expected, the annual average skill decreases with lead time (grey dashed lines in Figure 4.1). Additionally, the skill is strongly anticorrelated with the spread. A strong relationship between spread and skill is a desired property of a forecasting system since it implies that lower uncertainty in the ensemble corresponds to better predictions. Note however, that the relationship shows on average over a number of forecasts but does not hold well for individual forecasts (see also MacLeod et al., 2018). Generally, the skill is higher for the extended winter season from November to March. The late winter months February and March (FM) particularly stand out as having higher probabilistic skill (CRPSS and ROCSS) and significantly lower spread than other months. In fact, pentad 4 skill in FM is comparable or even higher than pentad 3 skill in many other months (e.g., October). Deterministic skill (CORR) is furthermore highest in February but displays a marked drop to annual average levels in March. Prediction skill in the summer months is overall worse but slightly increases for the months June and July, which is in agreement with the lower ensemble spread. August however, being the summer month with lowest ensemble spread does not exhibit better skill than other months. Taking all the forecast performance measures into account, in summary we can say that out of all seasons, in winter European temperatures are predicted best on subseasonal time scales with February and March being the months with highest probabilistic prediction skill. Summer temperature skill is somewhat lower and the skill in the transition seasons is lower yet. September and October (SO) are the least skillfully predicted months in terms of European temperatures. Note that this seasonal cycle is qualitatively similar to shorter lead times (pentads 1 and 2, Figure B.1).

Considering the absolute levels of prediction skill it is particularly interesting to note that the deterministic skill is still high for the lead times considered here. CORR only becomes indistinguishable from 0 for some months in pentad 4 (Figure 4.1, bottom row) and is around 0.6 for February, which is often considered to be a level of skill above which deterministic forecasts are still deemed useful (e.g. Domeisen et al., 2020c). For pentad 3, the skill is as high as 0.8 in February and above 0.4 from November to March. In terms of the probabilistic scores, the ROCSS is at a similar level as CORR and notably higher than the CRPSS. As pointed out above, the ROCSS needs to be regarded as a measure of potential skill since it does not penalize biases in the forecast distributions. However, as it measures the skill of a categorical forecast, we cannot directly compare it to the CRPSS to infer the potential gain of calibrating the forecasts. A measure more appropriate to compare to the ROCSS of the upper tercile would be the Brier skill score (BSS). We do not show the seasonal cycle of the BSS here but the variations are the same and FM and SO emerge as the seasons with highest and lowest BSS, respectively. The BSS is generally lower than the ROCSS by approximately 0.2 indicating that there is some potential for improving the skill by calibrating the forecasts.

4.3.2 Composites of Best and Poorest Forecasts in Each Season

In the previous section it became evident that there are clear seasonal variations in the subseasonal hindcast skill of European near-surface temperatures. We now analyze the differences in the meteorological situations during higher and lower skill forecasts. For this, we focus on a subset of the two two-month seasons that showed the strongest differences in subseasonal prediction skill, namely FM as the best predicted months and SO as the poorest ones. For each of these seasons, we form composites of the best and poorest quartiles of forecasts (in terms of their CRPSS for EUR T2m) as described in Section 4.2.3 and consider the composite average fields of T2m and Z500 anomalies both in the verification (ERA-Interim) as well as the hindcasts themselves during forecast pentads 3 and 4 (Figs. 4.2 – 4.4).

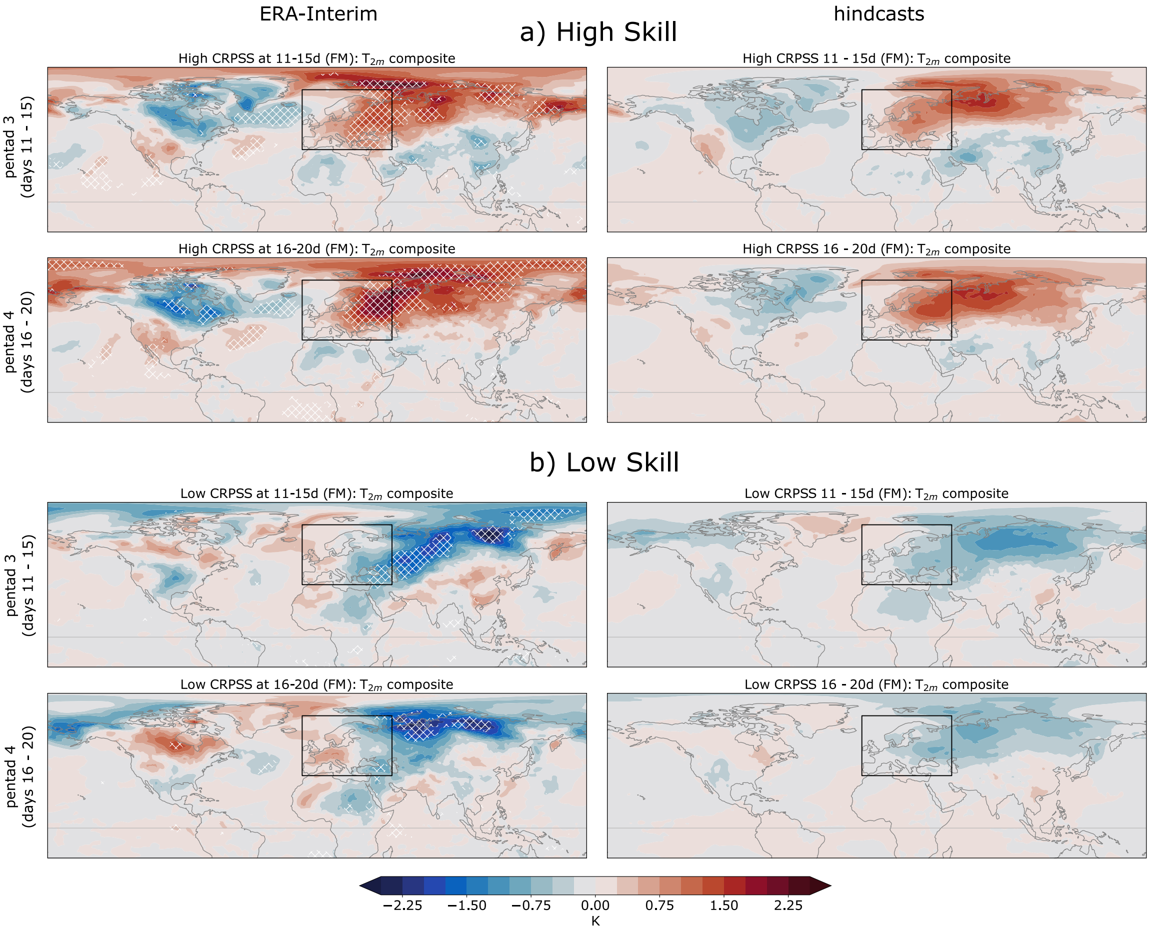
ERA-Interim
hindcasts
a) High Skill
High CRPSS at 11-15d (FM): T2m composite
High CRPSS 11 - 15d (FM): T2m composite
pentad 3 (days 11 - 15)
High CRPSS at 16-20d (FM): T2m composite
High CRPSS 16 - 20d (FM): T2m composite
pentad 4 (days 16 - 20)
b) Low Skill
Low CRPSS at 11-15d (FM): T2m composite
Low CRPSS 11 - 15d (FM): T2m composite
pentad 3 (days 11 - 15)
Low CRPSS at 16-20d (FM): T2m composite
Low CRPSS 16 - 20d (FM): T2m composite
pentad 4 (days 16 - 20)
−2.25
−1.50
−0.75
0.00
0.75
1.50
2.25
K

FIGURE 4.2: Composite mean 2-m temperature anomalies of best (a) and poorest (b) predicted situations in FM in ERA-Interim (left) and ECMWF hindcasts (right) during the target pentad for composites formed based on the EUR T2m CRPSS in pentad 3 (top rows in a and b) and 4 (bottom rows in a and b) as described in Section 4.2.3. Anomalies that are significantly different from zero at the 90% level according to a block bootstrap (Section 4.2.3) are indicated by the hatching for the ERA-Interim composites only.

Best and Poorest Forecasts in February/March

The composite pattern for FM shows that, on average for the most successfully predicted EUR temperatures, warm temperature anomalies prevail over the entire European continent as well as the Barents Sea, large parts of Siberia and, to a slightly lesser degree, the entire Arctic (Figure 4.2a). The high skill composite for the EUR box exhibits a positive temperature anomaly. At the same time, temperatures over the NA, Greenland and the eastern North American continent are anomalously low. This pattern is largely independent of whether the skill to composite the fields was computed for pentad 3 or 4 (compare top and bottom of 4.2a) although the strength of the anomalies is slightly enhanced for pentad 4. As expected for an average over predictions at subseasonal lead times, the average temperatures in the hindcasts have lower magnitude than in ERA-Interim (compare left and right panels of Figure 4.2a). Although weaker in magnitude, the hindcasts exhibit a very similar average temperature pattern as the verification, reproducing both the warm anomalies over Europe as well as the cold anomalies over eastern North America, Greenland and the NA. In fact, the hindcast anomalies are stronger for pentad 4 than pentad 3, which is in agreement with the reanalysis. This is remarkable since the average over many forecasts tends to be closer to climatology with longer lead time. Thus, we would rather expect to see the opposite, namely weaker anomalies in pentad 4 than pentad 3.

The low skill composites (Figure 4.2b) exhibit largely opposite anomalies over eastern Europe and Siberia compared to the high skill composite. Weak cold anomalies dominate the eastern part of the EUR box while the rest of the box is close to climatology. Anomalies over the rest of the Northern Hemisphere (NH) are weaker and of smaller scale than in the high skill composite with the exception of Canada and Alaska for pentad 4. Perhaps surprisingly at first sight, the low skill composites are relatively similar between ERA-Interim and the hindcasts. Both clearly show the strong cold anomaly over eastern Europe and Siberia. In the EUR region, however, the hindcasts tend to produce weak cold conditions over western and central Europe when in the verification weak warm anomalies were present. Furthermore, note that the CRPSS that we use to sort the composites measures the probabilistic error rather than the mean error that we are able to visually estimate from the depicted composites. Although the CRPSS also penalizes a large mean error it furthermore penalizes a lack of sharpness of the forecasts. Comparing the ensemble spread between the high and low skill composites confirms that the spread in the low skill situations is consistently higher in the EUR region (Figure B.2).

Since both the low and the high skill composites in FM display significant and relatively large-scale temperature anomalies, we next consider composites of Z500 for a view on the related mid-tropospheric circulation (Figure 4.3). Evidently, the warm temperatures in connection with high skill (Figure 4.3a) are accompanied by a strong meridional dipole in geopotential height that is reminiscent of the positive phase of the NAO (NAO+) or, comparing to the weather regimes of Grams et al. (2017), a mixture of the Zonal and Scandinavian Trough regimes. The related geostrophic flow over the NA has a pronounced zonal component, which leads to advection of warm air from the Atlantic towards western Europe consistent with the observed warm anomalies. The strongly increased temperatures over eastern Europe are related to the southerly component of the geostrophic flow that advects air from the Mediterranean into more northern latitudes. This northward tilt in the geostrophic flow is stronger in the pentad 4 composites possibly resulting in stronger warm anomalies. In the hindcasts, the large-scale circulation is fairly well captured but the geostrophic flow over eastern Europe is more southerly, leading to relatively warmer anomalies in the hindcasts (note again that in general the amplitude in the hindcasts is weaker due to taking the ensemble mean). For the pentad 4 composites, this more northward advection is more consistent with the verification, which is a possible reason why the average temperature patterns (Figure 4.2a) agree better between hindcast and verification for the pentad 4 composites.

In the low skill composites (Figure 4.3b) we can still see significant anomalies, indicating that even the situations that are predicted poorest in FM display some common characteristics in the circulation. However, the pattern that arises in these composites is

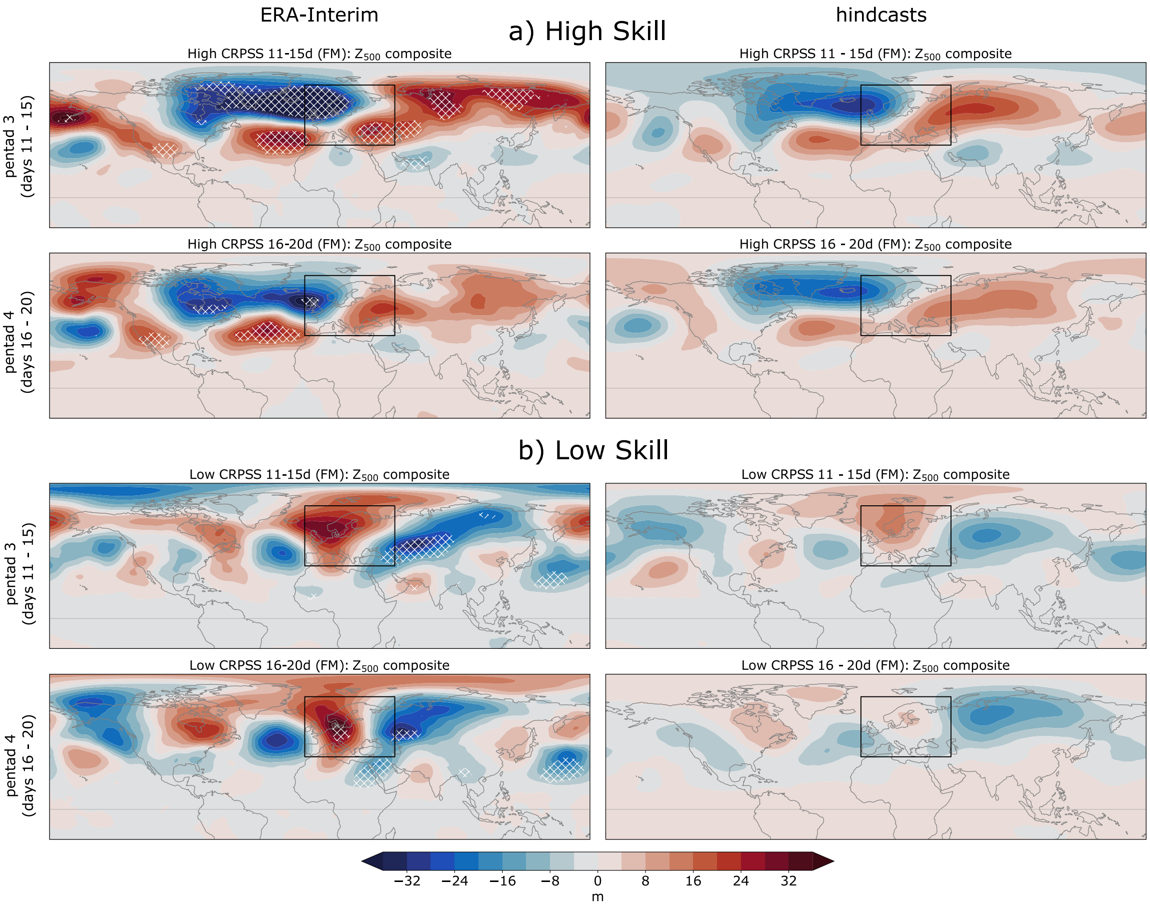
ERA-Interim
hindcasts
a) High Skill
High CRPSS 11-15d (FM): Z500 composite
High CRPSS 11 - 15d (FM): Z500 composite
pentad 3 (days 11 - 15)
High CRPSS 16-20d (FM): Z500 composite
High CRPSS 16 - 20d (FM): Z500 composite
pentad 4 (days 16 - 20)
b) Low Skill
Low CRPSS 11-15d (FM): Z500 composite
Low CRPSS 11 - 15d (FM): Z500 composite
pentad 3 (days 11 - 15)
Low CRPSS 16-20d (FM): Z500 composite
Low CRPSS 16 - 20d (FM): Z500 composite
pentad 4 (days 16 - 20)
−32 −24 −16 −8 0 8 16 24 32
m

FIGURE 4.3: As in Figure 4.2 but for composites of Z500.

very different. The geopotential height anomalies display a zonal wave train-like pattern, most pronounced over the NA but more generally over the entire NH. In the NA sector the circulation weakly resembles the negative phase of the NAO (NAO−) but corresponds more to the European Blocking (EuBl) regime (Grams et al., 2017, Supplementary Figure 1). These circulation types are again consistent with the temperature patterns from Figure 4.2b with the northerly component of the geostrophic flow advecting cold Arctic air over northern and eastern Europe. In the hindcasts, these blocked regimes also occur to some degree. It is noticeable however that the flow upstream of the NA does not match the reanalysis and the subtleties of Z500 anomalies in the EUR region are less well captured. For the pentad 3 composites in Figure 4.3b for instance, the Z500 anomaly in ERA-Interim reaches all the way into northern Africa, while in the hindcasts there is more north-westerly advection into the Mediterranean. This leads to the differences between the temperature patterns in ERA-Interim and the hindcasts. In the pentad 4 composites it can be seen that the predicted anomalies additionally become very weak. Thus, they display even weaker gradients in geopotential height, which leads to the relatively weaker cold anomalies over Europe in the hindcasts.

It is evident that for the best predicted season (FM) of the year in terms of the EUR average temperature anomalies, the more zonal flow regimes that lead to increased temperatures over Europe and especially downstream of Europe are the ones that are predicted best. On the other hand, the hindcasts are not able to capture the details of the blocked flow regimes over the NA in the low skill composite. Especially in the western part of the domain they tend to underestimate the southerly advection of warm subtropical air towards the Iberian peninsula, which leads them to predict weak cold anomalies when in reality the conditions were close to average.

Best and Poorest Forecasts in September/October

We now repeat the analyses from above for the season that exhibits the lowest skill for EUR T2m (SO). No significant large-scale temperature pattern arises in the T2m composites (Figure 4.4). The better predicted situations show a tendency for cold anomalies over Europe and central to eastern North America along with warm temperatures over central Asia and around the Baffin Bay and Labrador Sea (the latter only for pentad 4). Note that part of the reason for the overall weaker anomalies is the weaker variability in SO temperatures compared to FM (Figure B.3). This, however, cannot explain the lack of significant anomalies. In fact, even though the hindcast's ensemble mean tends to show weak anomalies for longer lead times, in Figure 4.4a the magnitudes of the hindcast anomalies are nearly as large as for the verification. This further supports the notion that the average pattern arising in the temperatures for the best predicted situations in SO is a mixture of many different forecast situations. On average, as we would expect for the better predicted composite, the hindcasts produce a pattern of temperature anomalies similar as in ERA-Interim although they tend to extend the warm anomalies over central Asia further into eastern Europe.

For the low skill composite (Figure 4.4b), slightly stronger average anomalies emerge in the verification. Although not significant, we see a clear tendency towards warmer temperatures over the EUR region accompanied by cold anomalies over Greenland and parts of central Asia. Comparing the hindcasts with the verification, we see that the predicted anomalies are much weaker than in the verification. Furthermore, they tend to produce too cold temperatures over the EUR region, especially in the south-eastern part of the domain and more so for pentad 4 hindcasts. As opposed to the FM hindcasts that performed better at predicting phases of enhanced temperatures over Europe, in SO the poorest hindcasts are on average accompanied by anomalously warm EUR temperatures.

As before, it is helpful to additionally consider the average circulation anomalies accompanying the best and poorest predicted EUR temperatures (Figures 4.5a and b, respectively). As expected from the temperature composite (Figure 4.4a), no clear pattern arises in the Z500 verification composite for the best predicted situations (Figure 4.5a). A number of weak cyclonic and anticyclonic anomalies can be identified, but these are likely artefacts of averaging over a set of situations without a common large-scale pattern. For the pentad 4 composite there is a stronger negative anomaly over Scandinavia, which causes the slightly colder conditions seen in Figure 4.4a (lower left panel). The hindcasts on average largely fail at predicting these weak Z500 anomalies for the higher skill composite. The predicted residual anomalies that remain after averaging are displaced compared to the verification, slightly more so for the pentad 4 composites.

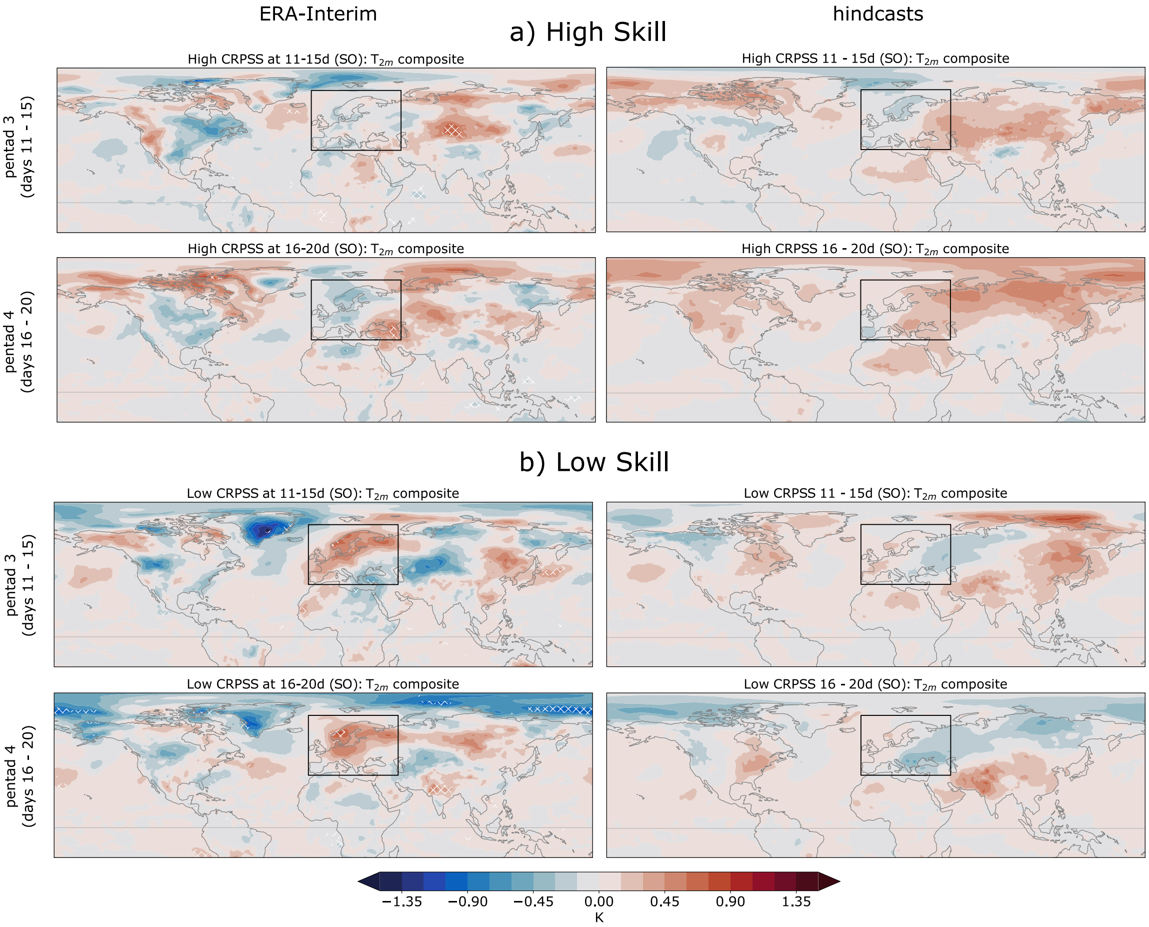
ERA-Interim
hindcasts
a) High Skill
High CRPSS at 11-15d (SO): T2m composite
High CRPSS 11 - 15d (SO): T2m composite
pentad 3 (days 11 - 15)
High CRPSS at 16-20d (SO): T2m composite
High CRPSS 16 - 20d (SO): T2m composite
pentad 4 (days 16 - 20)
b) Low Skill
Low CRPSS at 11-15d (SO): T2m composite
Low CRPSS 11 - 15d (SO): T2m composite
pentad 3 (days 11 - 15)
Low CRPSS at 16-20d (SO): T2m composite
Low CRPSS 16 - 20d (SO): T2m composite
pentad 4 (days 16 - 20)
−1.35
−0.90
−0.45
0.00
0.45
0.90
1.35
K

FIGURE 4.4: As in Figure 4.2 but for SO hindcasts. Note the different color scale compared to Figure 4.2.

The low skill Z500 composites in Figure 4.5b exhibits an anticyclonic anomaly over the EUR region, which again resembles a EuBl regime, particularly for the pentad 3 composite, although the average anomalies are not significant. For pentad 4, this anomaly over Europe is strongly reduced and instead, upstream of the NA, a pair of meridional dipoles of Z500 anomalies emerges. This could be an indication that the situations entering this composite exhibit a highly variable and meandering jet stream. While some of these situations could induce a EuBl, the exact position of the anomalies determines the regime over the NA and Europe. This would lead to a multitude of different synoptic situations entering the composite, which could explain the rather weak yet distinct pattern of average anomalies for pentad 4. Considering the associated pentad 4 hindcasts for these situations, it becomes clear that the model predicts a pattern of Z500 anomalies for these that is largely oriented along one line of latitude. This could be a result of a too zonal jet stream, specifically over the NA, which is a common model bias (e.g. Woollings and Blackburn, 2012). While this is a possible explanation for the anomalies in the pentad 4 composites, the situation is less clear-cut for pentad 3 where the anomalies upstream of the NA exhibit no clear pattern.

Comparing the composites of the best and poorest predicted situations in terms of EUR T2m, we saw that in the overall most successfully predicted season, FM, clear large-scale

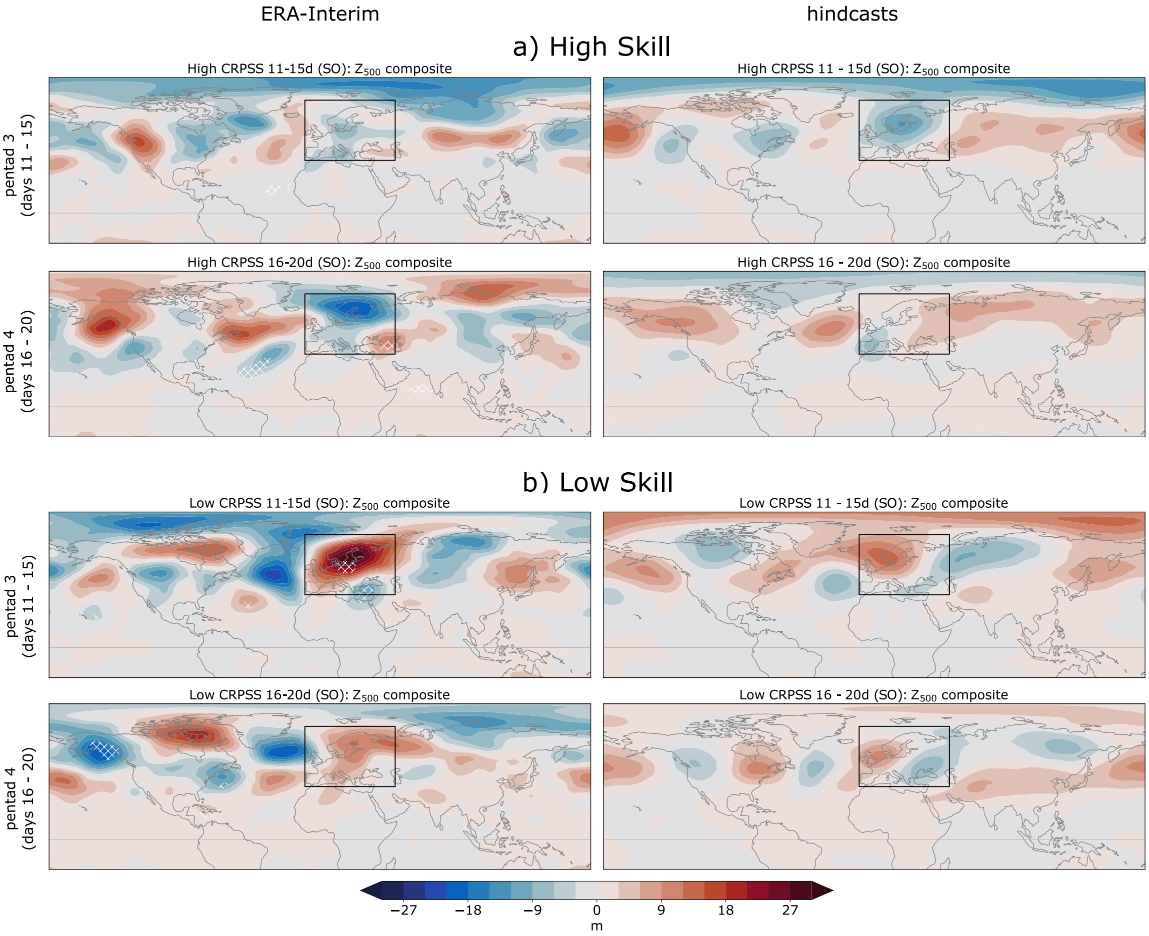
ERA-Interim
hindcasts
a) High Skill
High CRPSS 11-15d (SO): Z500 composite
High CRPSS 11 - 15d (SO): Z500 composite
pentad 3 (days 11 - 15)
High CRPSS 16-20d (SO): Z500 composite
High CRPSS 16 - 20d (SO): Z500 composite
pentad 4 (days 16 - 20)
b) Low Skill
Low CRPSS 11-15d (SO): Z500 composite
Low CRPSS 11 - 15d (SO): Z500 composite
pentad 3 (days 11 - 15)
Low CRPSS 16-20d (SO): Z500 composite
Low CRPSS 16 - 20d (SO): Z500 composite
pentad 4 (days 16 - 20)
−27 −18 −9 0 9 18 27
m

FIGURE 4.5: As in Figure 4.3 but for SO hindcasts. Note the different color scale compared to Figure 4.3.

circulation patterns arise in connection to both the best and the most poorly predicted situations. It is likely that the dominance of these large-scale regimes is part of the reason for the overall higher skill in FM. The regimes are quite successfully predicted in the hindcasts, both in pentad 3 and in pentad 4, and so is the large-scale temperature variability associated with these regimes resulting in higher skill. Quite clearly, the regimes of enhanced zonal flow across the NA (Zonal and Atlantic Trough regime) are predicted more accurately than situations of blocked flow. As a result, warm European winter temperatures are better predicted than the more variable temperatures associated with the different types of blocking. In SO, which is the season with the poorest EUR T2m predictions, the distinction between the high and low skill composites is less clear-cut. However, also in SO, situations with a tendency for blocking and a more variable jet stream lead to lower EUR T2m prediction skill. Thus, the fact that the zonal regimes occur less frequently in SO (Grams et al., 2017) likely plays a role in the lower prediction skill for EUR temperatures in the autumn months as compared to late winter. While a likely contribution to the seasonal cycle in subseasonal skill, this is by no means a complete explanation for the seasonal variations. In that case, December and January (July and August) should appear as the best (poorest) predicted months since they have the highest (lowest) frequency of occurrence of the zonal regimes. Additionally, even the low skill composites in FM on average appear to be more consistent between verification and hindcasts than for the high skill composites in SO.

4.3.3 Initial States of Best and Poorest Forecasts

In the previous section we identified the average forecast situations that occur during episodes of enhanced skill for large-scale European temperatures. This analysis indicated that part of the reason for the enhanced skill in late winter (FM) is the higher frequency of occurrence of zonal flow regimes as compared to autumn. A forecast user, however, is possibly even more interested in being informed about the forecast's potential quality (for pentad 3 and 4) at the time when the forecast is presented to them, ideally on the day the forecast was initialized. Knowing that forecasts for EUR temperature vary in skill depending on the large-scale flow regime only helps to inform about the skill if these patterns are already recognizable at initialization. However, the flow conditions for high and low skill phases observed at pentad 3 and 4 are, on average, not persistent enough to be identified in the initialization already, independent of the season (not shown). To identify phases of enhanced or reduced skill at the time of initialization, we next consider composites of potential predictor fields at the time of the hindcast initialization. As predictors of forecast skill, we mainly focus our attention on the tropical and stratospheric circulations. These have been most consistently shown to influence subseasonal predictability in the NAE region. Other potential sources of subseasonal predictability exist, but are in general more strongly debated (sea ice, see Chevallier et al., 2019), not well captured in subseasonal forecast models (snow cover, see Garfinkel et al., 2020) or are effective in other seasons than the considered ones (soil moisture Seneviratne et al., 2010). Sea surface temperatures (SST) potentially have an effect on the predictability of the NA climate in autumn (Nie et al., 2019) but we do not observe robust signals in initial SST for the SO composites. In the following, we furthermore only show composites for the late winter season (FM). We performed the same analysis as below for SO but did not find consistent signals in the initial composites in either the stratosphere or the tropics. Especially for the stratosphere, this is to be expected since stratospheric variability during SO is strongly reduced compared to winter (Plumb, 1989).

As described in Section 4.1, the stratosphere is understood to be a potential source of predictability for the European winter surface climate. To test whether we find indications of stratosphere-troposphere coupling for the winter composites, we consider T100 anomalies over the polar cap at initialization. T100 is representative of the state of the stratospheric polar vortex (SPV) and at the same time indicative of changes in the troposphere (Domeisen et al., 2020b). In the composites of the conditions during initialization of the high skill forecasts in FM (Figure 4.6, left) for both pentads, cold (but not significant) anomalies prevail in the lower stratosphere, which are displaced off the pole towards the NAE region. The negative anomalies are stronger for pentad 4 than for pentad 3. Cold stratospheric temperatures are indicative of a stronger than average SPV and indeed, zonal mean zonal winds at 60°N are enhanced in these composites (not shown). Furthermore, a strong polar vortex is consistent with a stronger and northward shifted NA jet stream,

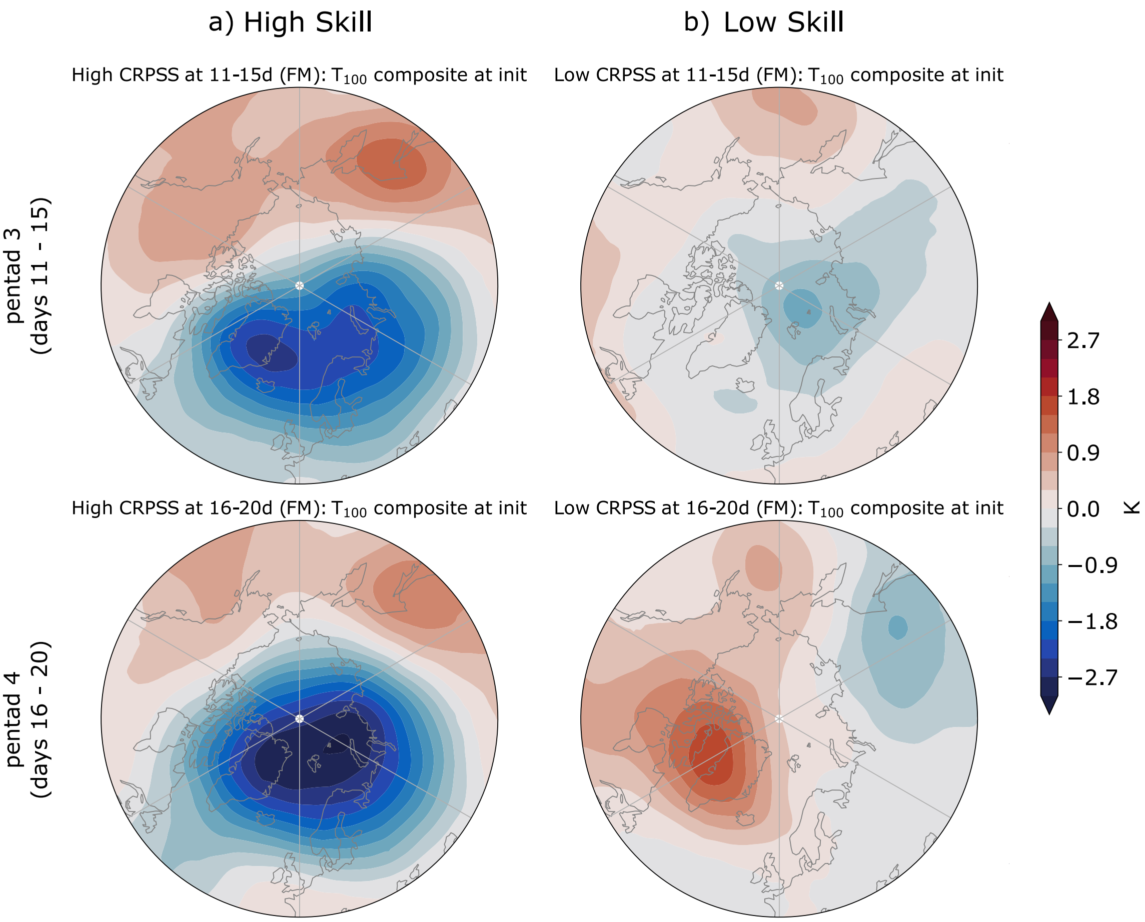
a) High Skill
b) Low Skill
High CRPSS at 11-15d (FM): T_{100} composite at init
Low CRPSS at 11-15d (FM): T_{100} composite at init
High CRPSS at 16-20d (FM): T_{100} composite at init
Low CRPSS at 16-20d (FM): T_{100} composite at init
pentad 3 (days 11 - 15)
pentad 4 (days 16 - 20)
2.7
1.8
0.9
0.0
−0.9
−1.8
−2.7
K

FIGURE 4.6: High (a) and low (b) skill composites of ERA-Interim lower stratospheric temperatures (100 hPa) during the hindcast initialization for FM hindcasts at pentad 3 (top) and 4 (bottom). Composites are computed for means over 5 days around the initialization times in the composite.

indications of which we observed in the high skill composites of Z500 for FM (Figure 4.3a). For the low skill composite on the other hand, no clear anomalies can be observed during the initialization of pentad 3 forecasts. For pentad 4 forecasts, a large-scale, weak warm anomaly emerges, indicating a slightly weakened SPV during the initialization of the poorest FM predictions. Weak vortex conditions have been shown to be associated with a higher occurrence of blocked flow regimes, which again is consistent with our observations in Figure 4.3b. However, the T100 anomalies are not significant for any of the considered composites and their usefulness as a predictor of pentad 3 and 4 forecast skill needs to be more carefully assessed. We return to this in Section 4.3.4.

Another source of subseasonal extratropical predictability are large-scale diabatic heating anomalies in the tropical atmosphere, which are often associated with distinct episodes of an active MJO (Lin et al., 2010). To test whether there are differences in the initial state of the tropical atmosphere for our skill composites, we take the same approach as above for the stratosphere but now show the tropical velocity potential at 200 hPa (VP200) as a proxy for convection during initialization for the different composites (Figure 4.7). A negative anomaly in VP200 is a result of large-scale upper-level divergent flow indicating ascent of air, which in turn suggests enhanced convection. Vice versa, positive VP200 anomalies

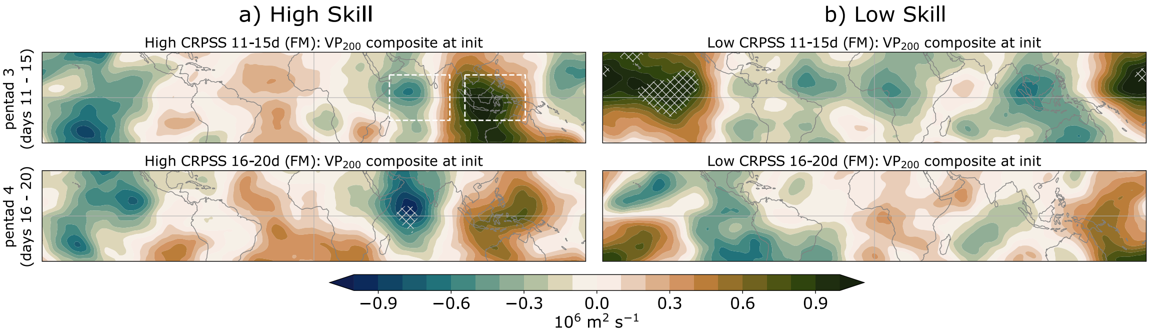
a) High Skill
b) Low Skill
High CRPSS 11-15d (FM): VP$_{200}$ composite at init
Low CRPSS 11-15d (FM): VP$_{200}$ composite at init
High CRPSS 16-20d (FM): VP$_{200}$ composite at init
Low CRPSS 16-20d (FM): VP$_{200}$ composite at init
pentad 3 (days 11 - 15)
pentad 4 (days 16 - 20)
−0.9
−0.6
−0.3
0.0
0.3
0.6
0.9
10^6 m^2 s^{-1}

FIGURE 4.7: High (a) and low (b) skill composites of ERA-Interim velocity potential at 200 hPa in the tropics (30°S to 30°N) during the hindcast initialization for FM hindcasts. Composites are computed for means over 5 days around the initialization times in the composite. White boxes outline the regions for computing the heating dipole used in Section 4.3.4.

indicate suppressed convection. Thus, VP200 can serve as a proxy for tropical large-scale convection anomalies (see also Ventrice et al., 2013). For the high skill composites in FM, weak negative and positive anomalies emerge over the central Pacific and the Atlantic, respectively. Additionally, a dipole in velocity potential between the Indian Ocean (IO) and the Maritime Continent (MC) is visible although only the negative IO anomaly at pentad 4 is significantly different from zero. This dipole is strongly reminiscent of phases 2 – 3 of the MJO. Diabatic heating anomalies in these regions have been shown to trigger a teleconnection to the NA sector that results in an NAO+-like response (Lee et al., 2019), consistent with the more zonal flow over the NA that we observe for the high skill composites (Figure 4.3). This could be an important source of predictability for the NA region and Europe, as is also supported by Lin et al. (2009). In the high skill composites, there are furthermore signs of an El Niño (EN) signal in the central Pacific, where negative VP200 anomalies indicate increased convection. EN conditions can enhance the aforementioned MJO teleconnection to the NA (Lee et al., 2019). This is further discussed in Section 4.3.4. The low skill composites (Figure 4.7b), mainly for pentad 3, exhibit negative VP200 anomalies over the central Pacific, possibly indicating a relationship with La Niña (LN). However, the ENSO-NA teleconnection is much more variable for LN than for EN resulting in a weaker signal for LN (Figure 1 in Domeisen et al., 2019). This may be due to the fact that the stratospheric pathway of the LN teleconnection is limited to strong LN events (Iza et al., 2016; Hardiman et al., 2019), while the tropospheric pathway is limited to late winter (Jiménez-Esteve and Domeisen, 2018) and exhibits saturation with respect to the forcing strength (Jiménez-Esteve and Domeisen, 2020). Thus, the fact that we do not observe an NAO+ in the low skill composites is not inconsistent with LN anomalies in the tropical Pacific at initialization.

4.3.4 Initial Stratospheric and Tropical States as Predictors of Forecast Skill

The initialization composites in the previous section give an indication of signals in some of the well-known sources of subseasonal predictability during the winter season. To consolidate their potential to serve as predictors of forecast skill, we now again compute the late winter (FM) hindcast skill for EUR T2m but this time split the set of hindcasts into three equally sized subsets. These subsets differ by the strength of the predictors that we previously tested.

For Figure 4.8a, we use the polar cap (60°N – 90°N) average anomalies of ERA-Interim T100 at the time of initialization to split the hindcasts. The CRPS (with adjustment, Section 4.2.2) of the forecasts for EUR T2m is then averaged over each subset (cold stratosphere, neutral stratosphere, warm stratosphere) and referenced to the average CRPS of the climatology over the same subset to obtain the CRPSS. Note that, since we use 5-day means of temperature, in Figure 4.8 a lead time of 11 days corresponds to an average over forecast days 11 – 16 or pentad 3. For forecasts initialized during an anomalously strong vortex corresponding to a cold lower stratosphere (blue line) the CRPSS is significantly enhanced at subseasonal lead times compared to the rest of the set. The separation is clearer at pentad 4 (day 16) and longer, consistent with the stronger anomalies in the composites of Figure 4.6. Additionally, the subset initialized under a weak vortex (warm T100) displays significantly reduced skill at some lead times showing that we can separate the EUR T2m subseasonal forecasts into better and poorer ones by splitting them based on their lower stratospheric initial conditions.

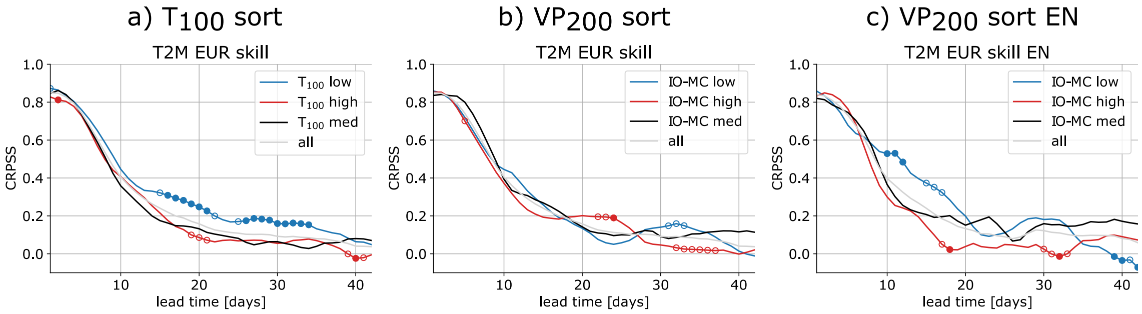

FIGURE 4.8: Hindcast skill (CRPSS) at predicting EUR T2m in FM, split by terciles of different predictors from ERA-Interim during initialization. Sorting of the hindcasts in a) by the T100 anomaly (averaged over all longitudes and 60°N – 90°N). Sorting in b) by the difference in VP200 spatially averaged anomalies between the Indian Ocean (50°E – 90°E, 15°S – 15°N) and the Maritime Continent (100°E – 140°E, 15°S – 15°N; see white boxes in Figure 4.7). c) as b) but for El Niño years only. Empty (filled) circles indicate that the difference between the CRPSS for the composite and the CRPSS for the rest ("non-composite") is significantly different from 0 at the 90% (95%) level. The significance of the difference is only tested for the high and low T100 (a) and VP200 (b, c) composites.

Motivated by the indications of a dipole in tropical diabatic heating between the IO and the MC that precedes enhanced prediction skill for the zonal regime, we use an index for the strength of this dipole to split the hindcast set. The index is defined as the difference between VP200 averaged over the boxes indicated in Figure 4.7a (IO − MC) such that a negative (positive) index indicates a dipole with enhanced (suppressed) convection over the IO and suppressed (enhanced) convection over the MC. As we can see in Figure 4.8b, separating the hindcasts using the strength of this index in the initializations does not work as well as using the lower stratosphere. In fact, while for pentads 3 and 4 the skill is slightly but not significantly enhanced for low dipole indices, it is significantly enhanced for high index values for days 21 – 23. At lead times longer than 30 days, it is again the low index initializations that have higher skill. Thus, no clear picture emerges from using the IO/MC dipole strength as a predictor of forecast skill. However, since Lee et al. (2019) stress the importance of ENSO in modulating the MJO-NA teleconnection, we also consider the dependence on the ENSO phase. Using a winter (DJF) Nino3.4 SST index, we split the hindcasts into El Niño (EN), La Niña (LN) and neutral years. Figure 4.8c shows that, as expected from the analysis of Lee et al., 2019, the hindcasts initialized under low index values in EN years show strongly enhanced EUR T2m skill. Note however, that the sample for this panel was reduced to approximately one third of the hindcasts and consequently, stronger deviations from the average skill are necessary to see a significant difference. Additionally, hindcasts initialized under a high index show reduced skill, which is somewhat surprising given that we did not see opposite VP200 anomalies arising in the low skill composites for FM in Figure 4.7b. We return to this point in Section 4.4.3.

4.4 Discussion

4.4.1 The Seasonal Cycle in Subseasonal Skill

In Section 4.3.1 we showed that there exists a seasonal cycle in the skill at predicting weekly near-surface land temperatures averaged over Europe (EUR T2m) at subseasonal lead times of 11 – 20 days. Both in terms of deterministic as well as probabilistic skill, the early months of the year (late winter) are most skilfully predicted. Independent of lead time, there is a marked drop of probabilistic skill between February and April, consistent with an increase in the ensemble spread of the models. The skill stays low during the summer months with weak improvements towards July. In late summer and early fall the skill is poorest and then improves again towards winter. The seasonal cycle of subseasonal near-surface temperature prediction skill has been previously addressed by Weigel et al. (2008b) and more recently by Manrique-Suñén et al. (2020). Although both of these studies find slightly larger prediction skill in the late winter months, the differences they identify are marginal. This apparent disagreement with our results merits further discussion. Most importantly, the considered regions differ between the studies. Weigel et al. (2008b) for instance report the skill computed over all grid points in the NH, while Manrique-Suñén et al. (2020) show the seasonal cycle of the skill averaged over a North American region. In this study, we compute the skill for a large-scale average of temperatures over Europe. Thus, we report the seasonal cycle in the skill of one variable only, while in the two other studies, the scores are computed for each grid point individually and subsequently averaged over the respective regions. It is possible that the seasonal cycle is enhanced by the spatial aggregation of temperatures that we use (see also van Straaten et al., 2020). Especially when averaging scores over the entire NH, the seasonal cycle is likely damped by averaging over multiple regions with different month-to-month variations. This is especially relevant if the seasonal cycle is already weak, as can be expected for subseasonal lead times. Furthermore, we limit our analysis to land temperatures, while in the other two studies every grid point was used. Temperature variability is damped over the oceans due to its higher heat capacity and inertia and it is conceivable that the stronger variability of land temperatures enhances the amplitude of the seasonal variations of temperature predictability. That is not to say that temperature predictability over land is higher but it is possible that the seasonal cycle in skill is less damped for temperatures over land only. Thus, we hypothesize that the stronger seasonal cycle compared to Weigel et al. (2008b) and Manrique-Suñén et al. (2020) that we observe is due to 1) using a spatial temperature average instead of a spatially averaged skill, 2) the fact that we consider land temperatures only, and 3) the exact region that is used for averaging.

Concerning the first two points, although a large-scale average of temperature might not be the most relevant forecast variable to a user, there are instances where forecasts of spatially averaged variables can be useful (Dorrington et al., 2020; Büeler et al., 2020). Furthermore, we argue that a seasonal cycle of spatially averaged skill does not provide any more useful spatial information on the skill either. In terms of land temperatures, we argue that these are more useful for the majority of users than both land and ocean temperatures combined.

Addressing the third point, in this study, we use a large European region, which likely reflects the influences of the large-scale circulation quite well. This is manifest in the Z500 composites (shown in Section 4.3.2) where, especially in late winter, large-scale flow patterns emerge that are strongly related to the observed temperature variability during these periods. The large-scale drivers of variability will be different in other regions of the globe and even in other regions of the NH. This means that the seasonal cycle may strongly differ between regions. To briefly investigate this point we compared the seasonal cycle of the skill of spatially averaged land temperatures for other regions of similar scale in the NH. There is a clear tendency towards winter being predicted better than summer in all of these regions. The seasonal cycle does however vary in its details. For instance, in Siberia the month-to-month variations are slightly enhanced compared to Europe, while for North America they appear to be more damped (not shown). It is also likely that the skill and its seasonal cycle are sensitive to the scale of the spatial aggregation, which is likely to highlight sources of predictability that are relevant for the variability in each specific region (van Straaten et al., 2020). However, choosing smaller European regions inside the considered domain does not strongly alter the overall seasonal cycle (not shown). In summary, it should be stressed that the details of the seasonal variations in subseasonal skill reported in this study are specific to the chosen region but the general tendency for higher subseasonal forecast skill in winter and lower skill in late summer/early fall can be generalized to large parts of the NH land regions.

4.4.2 Large-Scale Patterns

In FM, patterns of significant temperature and Z500 anomalies arise for both the high skill and the low skill composite. In the mid-latitudes, the high skill composite displays a clear zonal temperature dipole between eastern North America and Greenland (cold), and central to eastern Europe, Siberia and the Barents and Kara Seas (warm). This goes along with a large-scale, meridional dipole in geopotential height. This pattern is highly consistent with the second most predictable mode (MPM2) of global surface air temperature anomalies found by Xiang et al. (2019). They obtain this mode by using average predictability time analysis (DelSole and Tippett, 2009a; DelSole and Tippett, 2009b) to identify patterns that maximize subseasonal predictability of global surface air temperature anomalies in winter (DJF). This confirms the important role of the pattern that we find as a mode of predictability. Indeed, changing the boundaries of the averaging regions of the EUR box does not alter the high skill composite in FM significantly. Even when applying the skill sorting to all months of the year instead of just FM, the MPM2-like pattern emerges in the high skill composite. In this context, it is interesting to note that when considering the skill of temperatures averaged over different regions (e.g. North America, Siberia), the patterns that emerge do not resemble a single MPM as identified by Xiang et al. (2019). The fact that temperatures in Europe are strongly dominated by only one predictable pattern that explains similar amounts of variance as other MPMs (see Supporting Information of Xiang et al., 2019) could be part of the explanation why, compared to other regions in the mid-latitudes, subseasonal prediction skill over Europe is generally lower.

While we noted that the best predicted situations in FM are related to strongly zonal configurations of the large-scale flow over the NA, the most poorly predicted situations in both FM and SO show the imprint of more blocked flow regimes. Ferranti et al. (2015) investigated the flow-dependent predictability at a lead time of 10 days in the NAE region during winter in the ECMWF ensemble system. They find that the forecasts perform poorest when there is a transition to the European blocking regime during the forecast. This is in agreement with the poorest forecasts in FM. When we consider the time evolution of Z500 anomalies averaged latitudinally from the northern to the southern limit of the EUR box for the most poorly predicted composite in FM, we see indeed that the anticyclonic anomaly that dominates most of the EUR box at pentads 3 and 4 only develops during the forecast and is not present during initialization (Figure 4.9). While the model does produce the onset of the anticyclonic anomaly on average, these anomalies are not maintained in the same place and instead propagate westward (pentad 3 composite) or are soon followed by an opposite anomaly (pentad 4 composite). Our results thus corroborate the analysis of Ferranti et al. (2015). Furthermore, the reduced skill for blocked flows is in agreement with many previous studies that document the relative difficulty of predicting blocking (e.g. Pelly and Hoskins, 2003).

4.4.3 Predictors of Forecast Skill

For a user of operational forecasts it is valuable to be informed about the expected skill of a prediction already at the time of forecast initialization. The seasonal cycle studied here can inform about the relative quality of forecasts compared between different months of the year. In FM, we identified well-known patterns of NA variability related to the best and poorest predictions and there are several studies that indicate that these regimes are related to predictors outside of the extra-tropical troposphere. Charlton-Perez et al., 2018 find that the occurrence of their NAO+ regime, which bears close similarity to our FM high skill composite, is increased up to 30 days after strong vortex conditions, mainly due

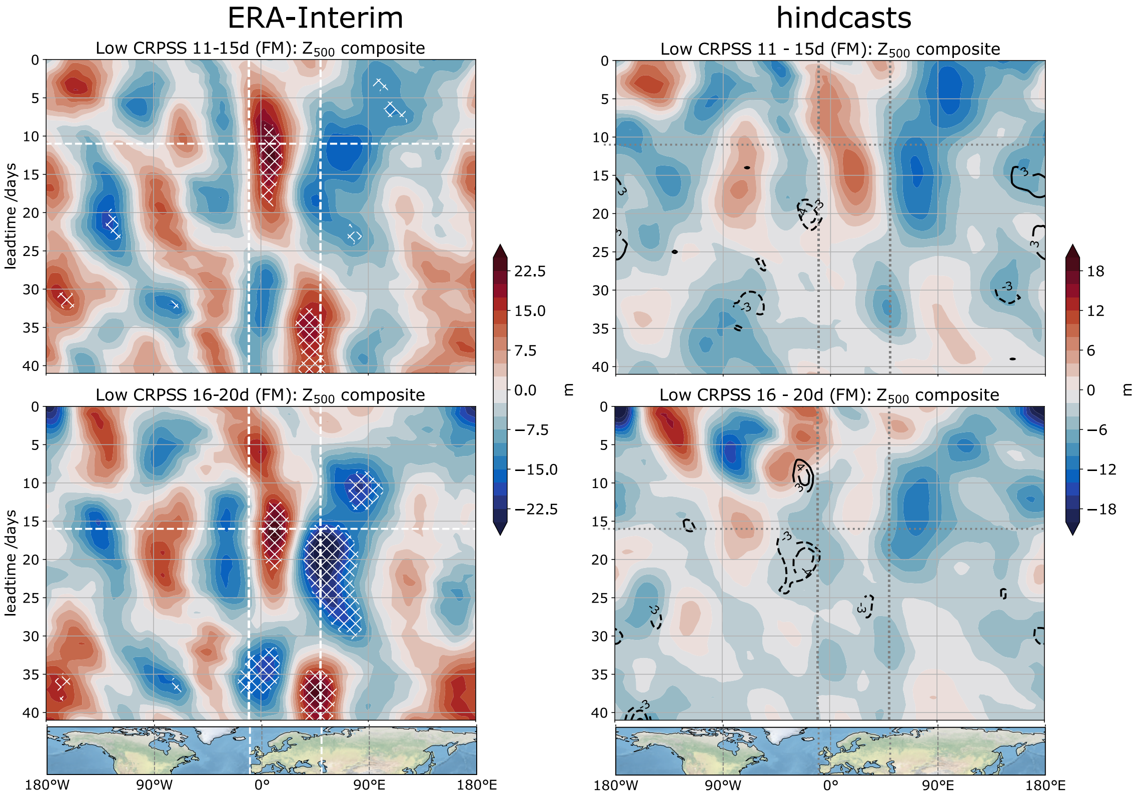

FIGURE 4.9: Composite time evolution of 5-day running mean Z500 anomalies averaged over all latitudes of the EUR box from initialization time until the end of the hindcast run for the 25% poorest hindcasts based on the CRPSS in pentad 3 (11 – 15 days, top) and pentad 4 (16 –20 days, bottom). The evolution is shown for ERA-Interim (left) where hatching indicates significant anomalies. The white dashed lines show the longitudinal limits of the EUR box. The hindcast mean evolution (right) is shown by the shading. Black contours show the composite average ensemble spread as anomalies relative to the mean ensemble spread at each lead time and each longitude. Only spread anomalies with an absolute value greater than 3 m are shown in 1-m intervals, negative anomalies are depicted by dashed and positive anomalies by solid lines. Lead times on the y-axes of the plots show the days since the initialization of the hindcast, the horizontal white dashed line shows at which day the hindcasts were evaluated for the sorting. Maps at the bottom are for reference of the longitudinal positions. Note the different color scales between the panels on the right and on the left.

to an enhanced probability of transitioning into this zonal regime. In agreement with this, for our high skill composite that exhibits pronounced zonal flow conditions, we find cold anomalies in the lower stratosphere during initialization (Figure 4.6a). The forecasts in our low skill composite, on the other hand, do not display substantial warm anomalies in the initial state of the lower stratosphere (Figure 4.6b). Nevertheless, separating the forecasts into three equally large subsets based on the polar cap temperature during initialization also separates the forecasts in terms of their subseasonal EUR T2m skill. Forecasts with a cold (strong) initial vortex display significantly higher skill than the rest (Figure 4.8a). Forecasts with a warm (weak) initial vortex tend to have lower skill but the differences are less significant. This tendency for enhanced (decreased) probabilistic forecast skill over Europe after strong (weak) vortex events is consistent with Domeisen et al. (2020b), although in our case the forecasts with weak vortex initializations depart less from the average skill. While we only considered the months FM here, it should be noted that this separation based on the initial vortex state works equally well when including January in the analysis. Note also that these results are in agreement with Xiang et al. (2019) who showed that the stratosphere is an important source of predictability of the MPM2 pattern of NH winter near-surface temperature (see Section 4.4.2), which is strongly manifest in our high skill composites.

In addition to significant stratospheric anomalies during the initialization of the high skill forecasts in FM, we found a dipole of enhanced upper-level divergence over the IO and suppressed divergence over the MC. This dipole is reminiscent of the convection associated with MJO phases 2 – 3 and also manifests in composites of 20–100 days bandpass-filtered outgoing longwave radiation (OLR, Figure B.4). Consistent with the zonal flow that we see for the high skill composites in FM, the NAO+ has been shown to occur significantly more frequently 5 – 15 days after the MJO has been in phase 3 (Cassou, 2008). Lin et al. (2010) pointed out that it is particularly the convection dipole between the IO and the western Pacific (WP) that goes along with MJO phases 2 – 3 (and 6 – 7) that is related to a modulation of NAO variability. The upper-level divergence dipole that we observe during initialization of the high skill forecasts in FM (that display mainly zonal regimes) agrees with these findings to some degree although the dipole we find appears to be shifted westward compared to Lin et al. (2009). Lin et al. (2010) further demonstrate that NAO forecast skill at subseasonal lead times in the Global Environmental Multiscale model is enhanced when a strong IO/WP convection dipole is present during initialization. Using an index of this dipole based on VP200 as a predictor of EUR T2m forecast skill however only weakly separates better from poorer forecasts at subseasonal lead times in our analyses (Figure 4.8b). It should be noted that we can slightly enhance the ability of our index to predict EUR T2m forecast skill by shifting the averaging boxes eastward by about 10°. This is clearly less consistent with the dipole we observe during initialization (Figure 4.7a) but closer to the regions that Lee et al. (2019) find to be important for triggering an MJO-NA teleconnection. It should further be stressed that, by splitting the hindcasts into three equally sized groups, we do not strictly separate them by MJO phase. In fact, our low dipole composite contains initializations under MJO phases 2 – 3 but also other phases and cases without active MJO. We further tested whether it is possible to separate the hindcasts based on the MJO phase at initialization. We used the bivariate Real-Time Multivariate MJO index (RMM, Wheeler and Hendon, 2004) computed from ERA-Interim to sort the forecasts into groups with active MJO phases 2/3, 4/5, 6/7, 8/1 and inactive MJO. With this approach we do not find significantly increased EUR T2m skill for any of the groups and in fact observe rather lower skill for forecasts initialized under MJO phases 2 and 3 (Figure B.5), opposite to the relationship suggested by our index. Consistent with this, Ferranti et al. (2018) show that the NAO+ skill is largely unaffected by the activity of the MJO at initialization, which could be another reason why our index has only weak potential for separating by subseasonal skill. Apart from the potentially weak relationship of our index to the MJO, another possible reason for the limited ability of the index to separate the forecasts by skill is that the MJO teleconnection to the NAE region is significantly modulated by ENSO (Lee et al., 2019). When considering EN years only, we indeed find that the IO/MC dipole has more potential as a predictor of subseasonal EUR T2m forecast skill (Figure 4.8c) but due to the reduced sample size these larger skill differences are not significant. The fact that the dipole in combination with EN works better at separating the forecasts by skill raises the question whether our index is more representative of an ENSO rather than an MJO signal in the initializations. Indeed, the VP200 dipole we observe bears some similarity with the precipitation dipole over the IO that was described by Abid et al. (2021). While they find that diabatic heating anomalies related to this dipole tend to cause an NAO+-like response, this pathway from the IO to the NA is mainly active during early winter rather than late winter as we observe it. Additionally, when separating the forecasts only based on the ENSO state, we do not find enhanced skill for EN years. In summary, our VP200 index is likely to mix some signals of the MJO and ENSO and disentangling the different contributions is beyond the scope of our study. Note additionally that the aforementioned teleconnections are further modified by the stratosphere (Garfinkel et al., 2014; Garfinkel and Schwartz, 2017; Lee et al., 2019) as well as the propagation speed of the tropical diabatic heating anomalies (Yadav and Straus, 2017).

4.5 Conclusions

We have shown the existence of a seasonal cycle in subseasonal forecast skill of large-scale averages of European land temperatures (EUR T2m). In the ECMWF hindcasts, winter emerges as the best predicted season. From March to April, forecast skill strongly drops consistent with overall higher ensemble spread in the model. The skill slightly but insignificantly increases towards July, is lowest in the early fall months (SO), and improves again when approaching winter. As the overall month-to-month variations are not sensitive to small changes in the boundaries of the European box, these results have the potential to inform forecast users about the expected relative quality of forecasts of European land temperatures in a specific season. We further show that there are common structures in the forecast situations that are predicted best and most poorly, especially during the winter months. While temperatures as a result of more zonal configurations of the flow over the NA are better predicted, temperatures related to more blocked flow are generally more poorly predicted. The latter also holds for the overall least successfully predicted months (SO). During the winter season, there is further potential to identify forecasts with higher subseasonal prediction skill already at the time of initialization. There is strong support for tropical diabatic heating anomalies over the Indian Ocean and Western Pacific to act as a source of predictability for the European winter climate. We found that an index based on the difference in upper-level divergence between these regions during forecast initialization has potential to separate forecasts by their EUR T2m skill, especially during El Niño years. However, due to the complex nature of the tropical-extratropical teleconnection from the IO/WP region to the NA, the signal in the forecast skill remains weak. The temperature anomaly in the lower stratosphere was found to be a more powerful predictor of forecast skill: Forecasts initialized when the stratospheric polar vortex is anomalously strong exhibit significantly enhanced skill at predicting EUR T2m at lead times between

15 and 35 days. On the other hand, forecasts initialized under weak vortex conditions do not have enhanced subseasonal skill and even lower skill than normal at lead times longer than 20 days. This asymmetry is consistent with previous studies and remains an open issue. These results could be of great benefit to forecast users since they allow for a judgement of the relative trustworthiness of a subseasonal temperature forecast during winter already at the time of initialization. Future work is needed to confirm the potential of these predictors for an early assessment of relative temperature forecast skill at smaller than continental scale.

Acknowledgements

We would like to thank Priyanka Yadav for helpful discussions on MJO teleconnections and for sharing her RMM MJO index as well as band-pass filtered OLR anomalies with us. We further thank the ECMWF for providing extended-range forecasts from the S2S prediction project (`https://apps.ecmwf.int/datasets/data/s2s/`) and the ERA-Interim reanalysis through the MARS archive (`https://apps.ecmwf.int/datasets/data/interim-full-daily/`).

All analyses and visualizations were carried out using the Python packages numpy, matplotlib, scipy, cmocean and windspharm.

Funding from the Swiss National Science Foundation to O.W. and D.D. through project PP00P2_170523 is gratefully acknowledged.

Chapter 5

Conclusions and Outlook

In this thesis, we studied the characteristics of subseasonal prediction and predictability of European near-surface temperatures. This last chapter provides a synthesis of the conclusions drawn from the previous chapters (Section 5.1) and points to possible extensions of our work for future research (Section 5.2).

5.1 Conclusions

The analyses conducted for this thesis are aimed at understanding the characteristics of European temperature predictability. To this end, forecasts and hindcasts from the S2S Prediction Project database (Vitart et al., 2017) proved to be an extremely valuable resource and all the results presented here were obtained by drawing on this extensive data pool.

Forecasts of extreme events are of major concern in subseasonal prediction. One reason for the focus on extreme events are their possible impacts. Additionally, events of large magnitude in one variable can be accompanied by persistent anomalous conditions in one or multiple other variables, potentially even before the event of interest happens. It is thus conceivable that these extreme events are inherently more predictable than more average events. However, the answer to the question of whether extremes are more predictable than average conditions is unlikely to be universal but instead depends on the region, time scale and variable of interest (Sterk et al., 2012). In Chapter 2 we addressed the issue of whether the subseasonal prediction skill for near-surface continental temperature extremes during summer over the Europe differs from the skill for more average conditions in the period 1999 – 2010, for which hindcasts were available for four different models. Overall, warm European summer extremes are more skillfully predicted than average temperatures at subseasonal lead times, which is in agreement with analyses of seasonal prediction skill (Becker, 2017). The enhanced skill for warm extremes is in contrast to low temperature extremes in summer, for which prediction skill is largely similar to average temperatures. As expected, there are regional variations in these skill differences with a stronger tendency for higher skill for summer heat waves in Central to Eastern Europe and Russia and less pronounced differences between skill for warm and average events in the Mediterranean regions. These results indicate that there is substantial spatial variability in the predictability contrast between extreme and average events. Furthermore, there exists an asymmetry in the predictability of European summer temperature extremes in the upper and lower tails of the temperature distribution. Over western Russia and the Ukraine for instance, summer heat waves are significantly better predicted than low temperature extremes. Our results further point to the importance of persistence in enhancing subseasonal prediction skill. While temperature persistence can explain large parts of the extended skill at predicting the 2010 Russian heat wave, it is not a sufficient explanation for the extended skill for the 2003 European heat wave, where other factors must have played a role.

Considering the fact that the climate is subject to a general warming trend and warm summer temperature extremes could become more frequent, our results make it conceivable that summer prediction skill in some parts of Europe could be further enhanced in the future. While climate change is thought to change the predictability in the future (see also Scher and Messori, 2019), temperature trends, or more generally, non-stationary components in the climatology can also have an effect on current prediction skill. As we showed in Chapter 3, care needs to be taken when evaluating the forecast skill over periods that are subject to long-term trends. With these trends explaining larger parts of the forecast variable's variance over the hindcast (calibration) period, the risk of an inflation of the skill increases. Though to a smaller degree, this dependence still holds when the trend is not perfectly reproduced by the forecast model. In order to avoid this artificial skill enhancement, it is necessary to account for the non-stationarity of the climatology that is estimated from the hindcast period. We showed that, while skill inflation is of minor concern for subseasonal forecasts of European temperatures in an operational system, it can significantly enhance the subseasonal skill in tropical regions, where linear trends can explain 10% or more of the temperature variance over the hindcast period. We showed that the inflation effect is especially large when evaluating categorical forecasts, which is a common form for communicating long-range forecast information. Although our focus was on the prediction skill for weekly temperature averages at subseasonal lead times, the results of this study are applicable to forecasts on any time scale. The inflation effect depends on the amount of variance of the forecast variable that is explained by the trend over the hindcast period (note that this also depends on the length of the hindcast period). While this is unlikely to be an issue for short-term weather forecasts, in seasonal-to-decadal prediction the non-stationarity of the climatology needs to be taken into account. The results of this study can in principle be generalized further to any form of non-stationary component in the climatology that can be estimated from the hindcast period and are thus a specific case of what Hamill and Juras (2006) illustrate conceptually. In particular, as pointed out recently by Manrique-Suñén et al. (2020), inaccurate estimates of the seasonal cycle can also lead to artificial skill in subseasonal forecasts.

Once the non-stationary components in the climatology are accounted for, it is possible to make inferences about the temporal variability of the skill itself. This information can be of great benefit to forecast users since it allows to judge the trustworthiness of a forecast at a certain initialization time relative to other initializations. To this end, in Chapter 4 we examined the month-to-month differences in subseasonal forecast skill of large-scale averages of European land temperatures. A distinct seasonal cycle in forecast skill between 10 and 20 days lead time can be identified which is in qualitative agreement with previous studies (Weigel et al., 2008b; Manrique-Suñén et al., 2020). Skill is generally highest during the winter months with maxima in February and March. During the rest of the year, the skill is closer to the annual average but consistently shows minima for the transition seasons in multiple skill measures, especially in September and October. In the late winter months, the forecast situations exhibiting the highest skill display a consistent large-scale structure that has previously been identified as one of the main modes of predictability of winter near-surface temperatures (Xiang et al., 2019). This mode is characterized by predominantly zonal flow conditions over the North Atlantic and concurrent warm conditions over the European continent, central Asia and Siberia. Most importantly, these conditions are connected to distinct anomalies in the stratosphere and the tropics during initialization. Especially the state of the lower stratosphere proves to be an effective predictor of subseasonal European temperature skill with a cold stratosphere generally leading to higher skill, which is consistent with previous findings (Domeisen et al., 2020b). Temperature skill in Europe furthermore tends to be increased after episodes of enhanced tropical convection over the Indian Ocean and concurrent suppressed convection over the Maritime Continent and Western Pacific, especially during El Niño years (see also Lee et al., 2019). However, the connection to the tropics proves to be more elusive and a simple separation of forecasts based on a single tropical anomaly in the initial conditions does not work as effectively as stratospheric anomalies at identifying flow-dependent subseasonal predictability. This corroborates the complex and variable nature of the tropical-extratropical teleconnections that influence predictability in the North Atlantic-European sector (e.g. Garfinkel and Schwartz, 2017; Yadav and Straus, 2017; Lee et al., 2019). In summary, the results of Chapter 4 confirm that subseasonal predictability is flow-dependent and thus that there is potential to not only predict subseasonal variability but also subseasonal forecast skill itself.

All of the studies conducted for this thesis were concerned with the spatial and temporal variability of predictability. Our results demonstrate how subseasonal forecast skill for European temperatures varies by region, season, the initial flow conditions and event type. The details of how the performance of subseasonal forecasts is evaluated are also crucial for correctly reporting the skill of subseasonal forecasts. Especially since subseasonal forecast skill is currently low, it is imperative that we identify windows of opportunity, in which subseasonal skill is significantly enhanced and carefully inform about the actual capabilities of these predictions to allow users to take well-advised, forecast-based decisions.

5.2 Outlook

Many extensions to our study of subseasonal predictability of European temperatures are conceivable and in the following we present some avenues for future research that we deem promising.

Firstly, while we did compare the performance of multiple forecasting systems from the S2S database in Chapter 2, one interesting extension of this study is to make use of a multi-model ensemble (MME) based on some or all of the eleven models in the database. While the differences in the forecasting strategies (mainly the differing initialization strategies) between the S2S models make the formation of an MME difficult, the use of an MME can prove beneficial for subseasonal forecasting of extreme events (Wanders and Wood, 2016). The added value of using subseasonal MMEs has so far only been assessed in a small number of studies, mainly with a focus on precipitation (Vigaud et al., 2017; Specq et al., 2020). These studies show that the benefit of the MMEs was mainly to improve reliability but at the cost of sharpness of the predictions. However, Pegion et al. (2019) use an MME of the SubX models and demonstrate that the MME outperforms any single model ensemble at forecasting the NAO and the MJO at lead times of four weeks in terms of correlation skill. This indicates that the improvement from an MME can go beyond a mere increase in reliability.

Generally, the prediction of extreme events merits further research. In Chapter 2 we evaluated forecasts of extreme events using a measure of deterministic forecast skill. Arguably, a probabilistic verification would be more informative as probabilistic subseasonal forecasts are more valuable to users (Palmer and Richardson, 2014). However, when evaluating forecasts for extreme events, great care is due since many verification scores degenerate to trivial values for rare events (e.g. Ferro and Stephenson, 2011; Ferro and Stephenson, 2012). Also, evaluating forecasts with standard methods based on a subset of cases that only contains situations when an extreme happened can lead to discrediting skillful forecasts (favouring "alarmist" forecasts instead, Lerch et al., 2017). When stratifying on extreme observations, the scoring rules can loose the desired property of being proper. For instance, when evaluating a forecast only in situations when an extreme occurred, a forecast that always predicts the occurrence of an extreme can obtain a very high score although it is clearly useless. Gneiting and Ranjan (2011) present weighted scoring rules that highlight forecast performance for a specified region of the climatological distribution (e.g. the tails/extremes) but remain proper. These weighted scoring rules could be particularly useful for the evaluation of forecasts of extremes. Furthermore, analyses of the development of the ensemble spread leading up to extreme events can give insights into their predictability characteristics. It should additionally be noted that successful probabilistic forecasts of extreme events likely require ensemble sizes larger than those available for most hindcast systems (the ECMWF system, which was most extensively used in this study, has 11 ensemble members in hindcast mode). Thus, the verification is best done for the operational forecasts, which commonly have larger ensemble size (ECMWF: 51 members). However, the forecast period is significantly shorter than the hindcast period, limiting sample size further, which is already small for extreme, rare events by definition. These issues will have to be carefully addressed in future studies of the prediction and predictability of extreme events.

Even though the S2S database proved useful for the analyses conducted in this thesis, targeted model experiments have the potential to further advance the understanding of predictability. For instance, in Chapters 2 and 4 we chose different approaches to separate the forecasts into subsets and found that the prediction skill can significantly differ between forecasts from different subsets. While some of our analyses provide indications of the possible sources of predictability responsible for these differences, targeted model experiments are needed to isolate the involved mechanisms. In Chapter 2, we were able to show that part of the subseasonal prediction skill for the 2010 Russian heat wave stems from an increased persistence of temperatures that the forecast models are able to capture. Modeling studies targeting specific processes responsible for the enhanced persistence are required to resolve the question of how the extended predictability arises.

Appendix A

Supporting Information for "Higher Subseasonal Predictability of Extreme Hot European Summer Temperatures as Compared to Average Summers"

Text S1: Verification measures used in the SI

The Odds Ratio Skill Score (ORSS) as shown in Figure A.3 is computed following Hogan and Mason (2012) based on the Odds Ratio (OR) which is defined as:

$$\text{OR} = \frac{H}{1-F} \times \left(\frac{F}{1-H}\right)^{-1} \tag{A.1}$$

where H is the hit rate defined as the number of hits divided by the total number of observed events and F is the false alarm rate defined as the number of false alarms divided by the number of observed non-event days.

The ORSS is then defined with reference to a random forecast (which has OR $= 1$):

$$\text{ORSS} = \frac{\text{OR} - 1}{\text{OR} + 1} \tag{A.2}$$

The standard error of the ORSS is given by:

$$s_{\text{ORSS}} = \frac{2 \times \text{OR}}{\sqrt{n_h} \times (\text{OR} - 1)^2} \tag{A.3}$$

TABLE A.1: Characteristics of the four subseasonal forecasting systems used in this study.

Center	Horizontal Resolution	Levels (Model Top)	Hindcast Frequency	Time Range	Ens. Size	Ocean coupl.	Sea Ice coupl.
BoM	2.5° × 2.5° (T47)	L17 (10 hPa)	every 5 days	d0–62	33*	Yes	No
CMA	1.1° × 1.1° (T106)	L40 (0.5 hPa)	daily	d0–60	4	Yes	Yes
ECMWF	16km (Tco639), 32km (Tco319)	L91 (0.01 hPa)	twice weekly	d0–10, d11–46	11	Yes	Yes
NCEP	0.3° × 0.3° (T384),	L64 (0.02 hPa)	daily	d0–44	4	Yes	Yes

[*] The BoM system consists of three ensembles with 11 members each. For better comparison with the other forecasting systems, only one of these 11 member ensembles was used in the study.

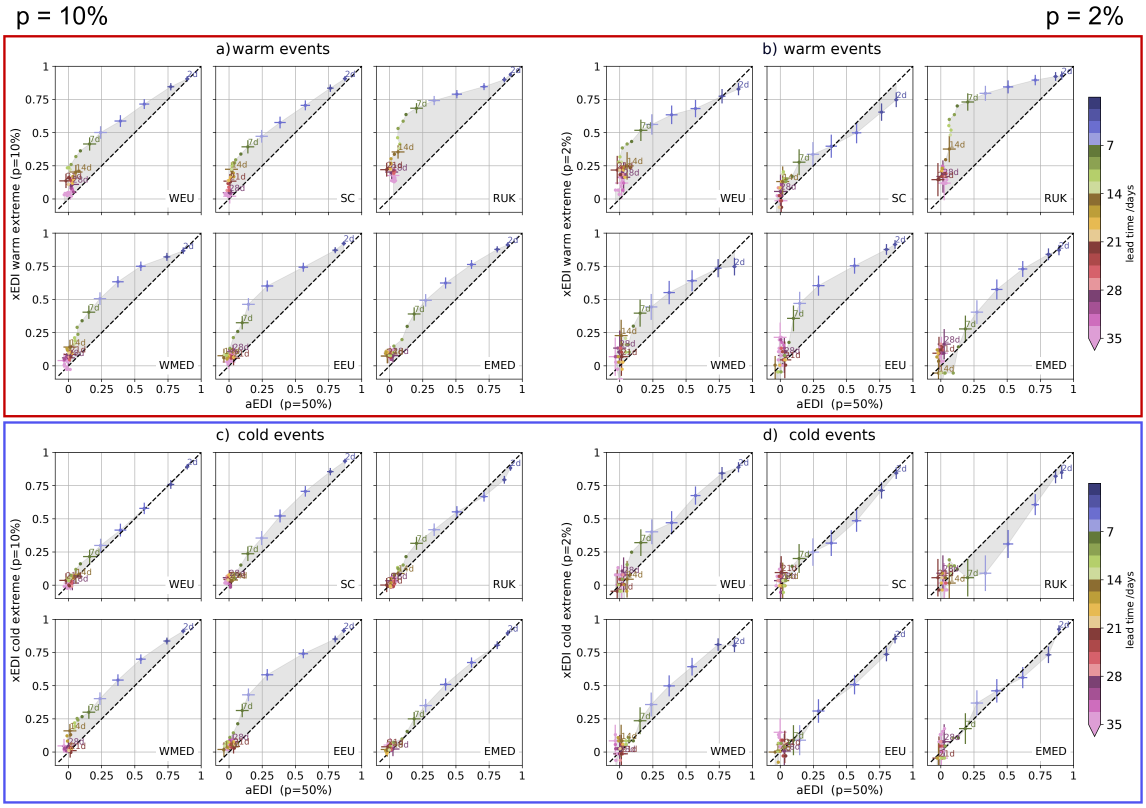

FIGURE A.1: Sensitivity of the xEDI to variations in the base rate. As in Figure 2.2 of the main manuscript but xEDI computed for events with a base rate p_x of 10% (a,c) and 2% (b,d). The comparison between xEDI and aEDI is shown for warm (a,b) and cold (c,d) events.

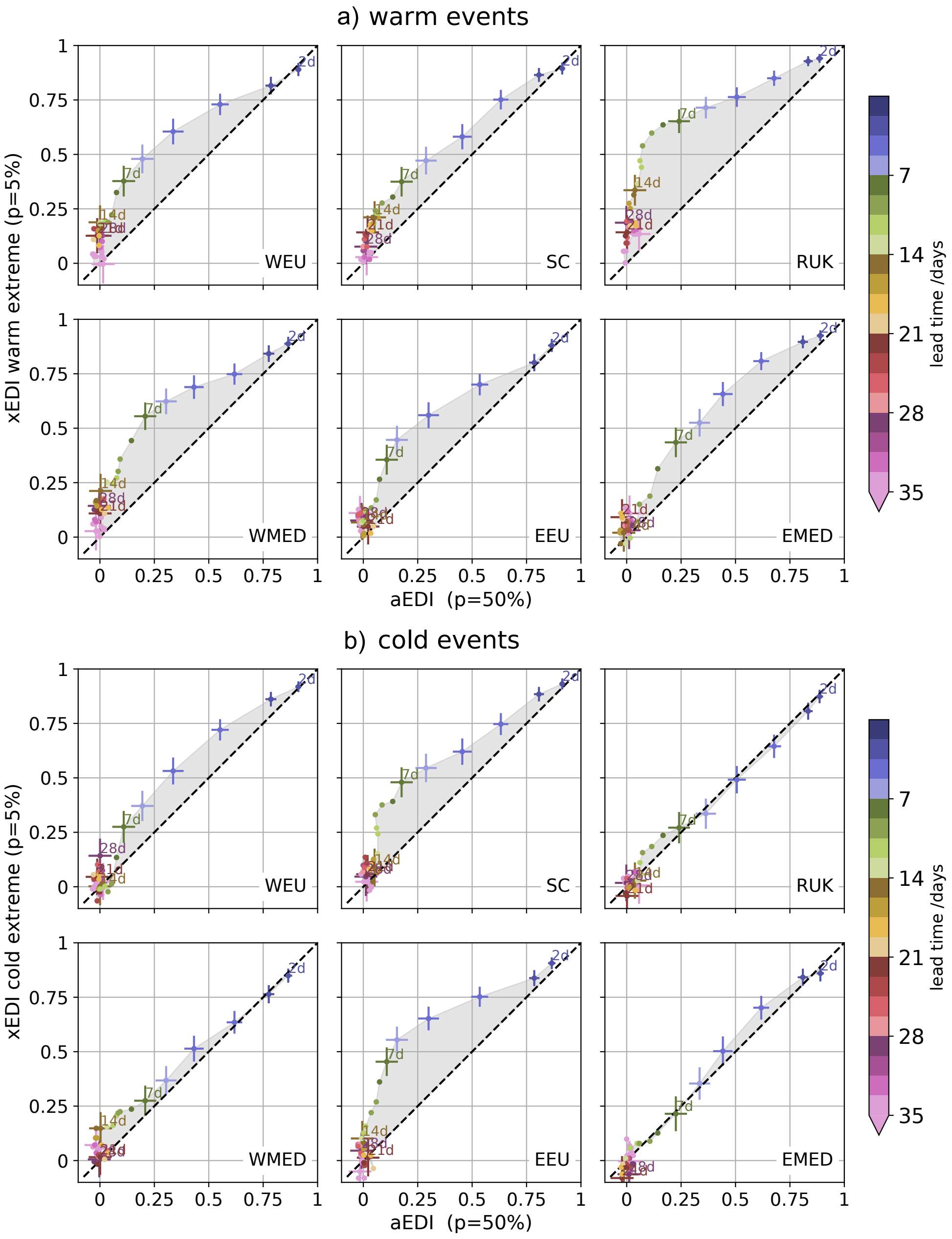

FIGURE A.2: aEDI vs xEDI for ECMWF as in Figure 2.2 of the main manuscript but here computed over the extended period 1998 – 2017.

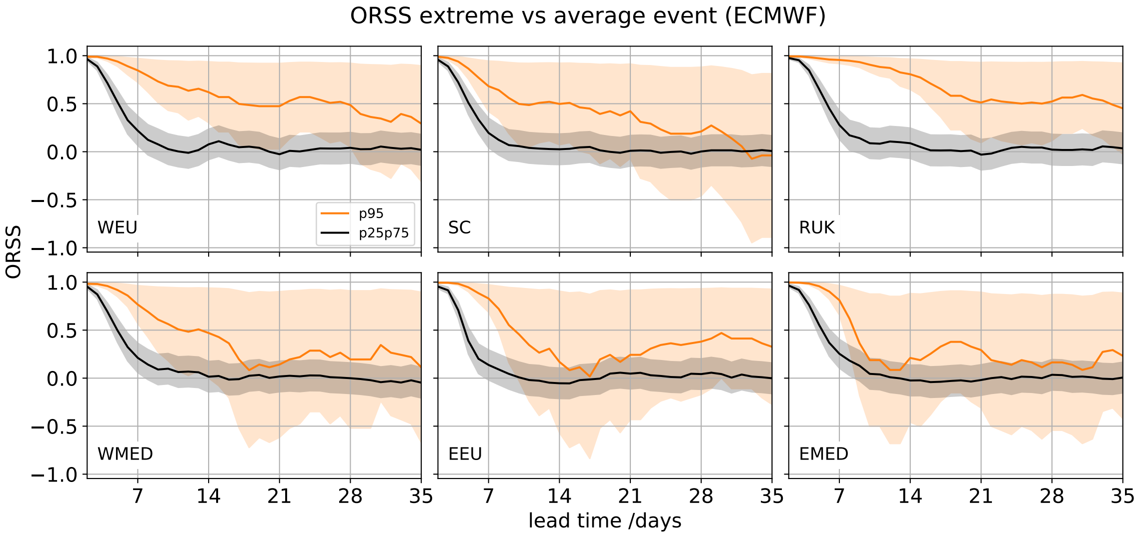

FIGURE A.3: Odds Ratio Skill Score (ORSS, Eq. A.2) as function of lead time for six European regions in the ECMWF system. In each panel, the ORSS for warm extreme events with a base-rate of $p = 5\%$ is shown in orange and ORSS for average events ($p = 50\%$) is shown in black. Shaded areas around the lines show the 95% confidence interval around the ORSS given by $1.96 \times s_{\mathrm{ORSS}}$ (see Eq. A.3).

Appendix B

Supporting Information for "Seasonal Variations in Subseasonal European Near-Surface Temperature Prediction Skill"

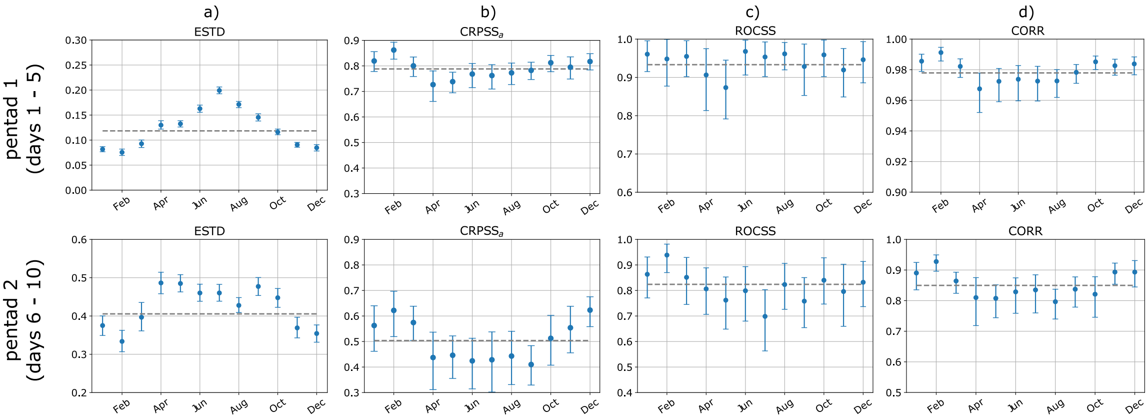

FIGURE B.1: As in Figure 4.1 but for pentads 1 (days 1 – 5, top) and 2 (days 6 – 10, bottom).

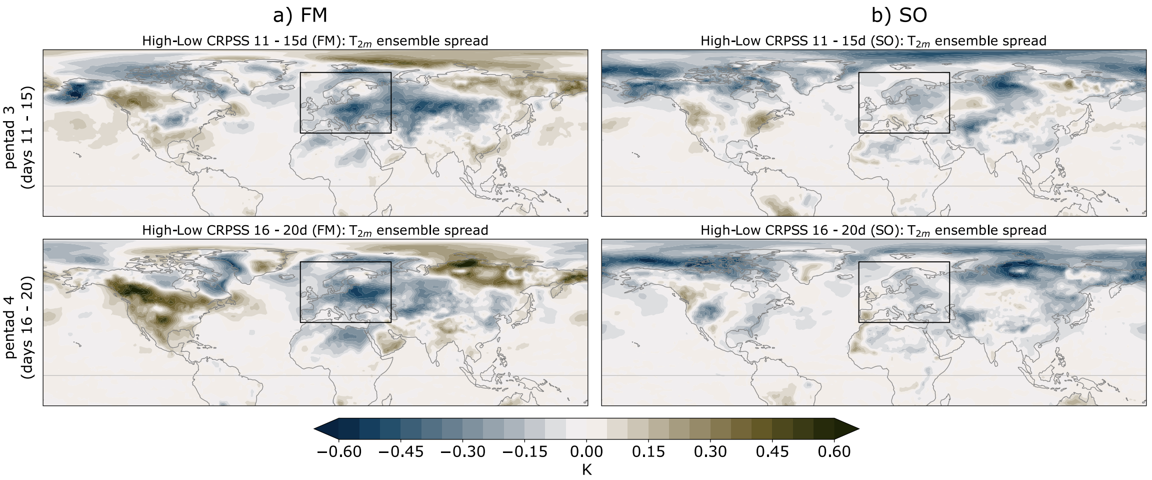
a) FM
b) SO
High-Low CRPSS 11 - 15d (FM): T_{2m} ensemble spread
High-Low CRPSS 11 - 15d (SO): T_{2m} ensemble spread
pentad 3 (days 11 - 15)
High-Low CRPSS 16 - 20d (FM): T_{2m} ensemble spread
High-Low CRPSS 16 - 20d (SO): T_{2m} ensemble spread
pentad 4 (days 16 - 20)
−0.60 −0.45 −0.30 −0.15 0.00 0.15 0.30 0.45 0.60
K

FIGURE B.2: Difference in T2m ensemble spread between best and most poorly predicted composites, i.e blue (brown) color indicates where the poor (best) predictions have higher spread than the best (poor) ones.

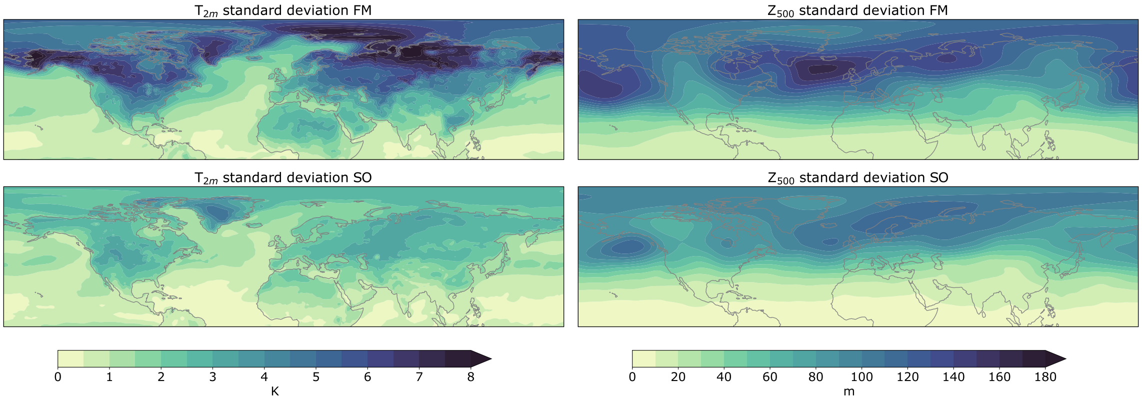
T_{2m} standard deviation FM
Z_{500} standard deviation FM
T_{2m} standard deviation SO
Z_{500} standard deviation SO
0 1 2 3 4 5 6 7 8
K
0 20 40 60 80 100 120 140 160 180
m

FIGURE B.3: Standard deviation of 5-day mean Z500 and T2m for FM and SO as indicated in the title of each panel.

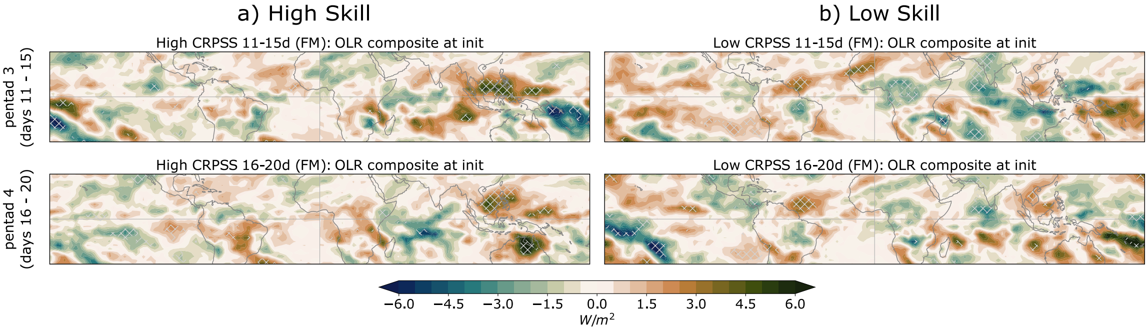
a) High Skill
b) Low Skill
High CRPSS 11-15d (FM): OLR composite at init
Low CRPSS 11-15d (FM): OLR composite at init
pentad 3 (days 11 - 15)
High CRPSS 16-20d (FM): OLR composite at init
Low CRPSS 16-20d (FM): OLR composite at init
pentad 4 (days 16 - 20)
−6.0 −4.5 −3.0 −1.5 0.0 1.5 3.0 4.5 6.0
W/m^2

FIGURE B.4: As in Figure 4.7 but for 20–100 days bandpass-filtered anomalies of outgoing longwave radiation (OLR) from the National Oceanic and Atmospheric Administration (NOAA, Liebmann and Smith, 1996).

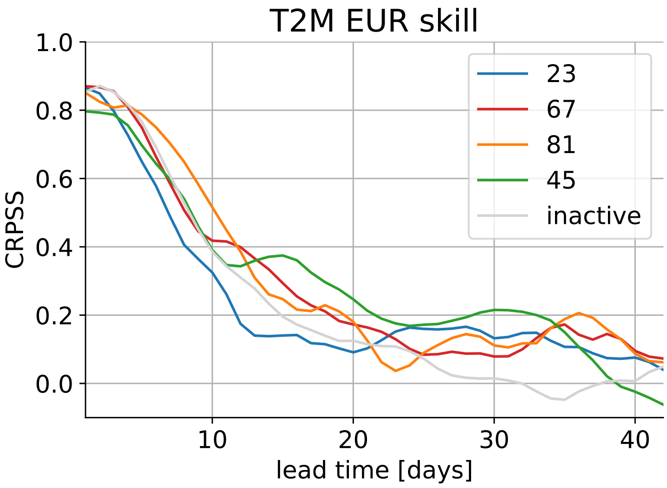

FIGURE B.5: As in Figure 4.8 but with the separation done based on the active MJO phase at forecast initialization (colored lines, see legend for the phases). The MJO phase is inferred from the Real-Time Multivariate MJO index (RMM, Wheeler and Hendon, 2004) computed from zonal wind at 200 and 850 hPa from ERA-Interim and OLR from NOAA. The MJO is considered active if the amplitude of the RMM index exceeds one. The grey line shows the skill of forecasts initialized under inactive MJO.

References

Abid, M. A., F. Kucharski, F. Molteni, I.-S. Kang, A. M. Tompkins, and M. Almazroui (2021). "Separating the Indian and Pacific Ocean Impacts on the Euro-Atlantic Response to ENSO and Its Transition from Early to Late Winter". In: *Journal of Climate* 34(4), 1531 –1548. DOI: 10.1175/JCLI-D-20-0075.1.

Alvarez, M. S., C. A. S. Coelho, M. Osman, M. Â. F. Firpo, and C. S. Vera (2020). "Assessment of ECMWF Subseasonal Temperature Predictions for an Anomalously Cold Week Followed by an Anomalously Warm Week in Central and Southeastern South America during July 2017". In: *Weather and Forecasting* 35(5), 1871 –1889. DOI: 10.1175/waf-d-19-0200.1.

Ardilouze, C., L. Batté, F. Bunzel, D. Decremer, M. Déqué, F. J. Doblas-Reyes, H. Douville, D. Fereday, V. Guemas, C. MacLachlan, W. Müller, and C. Prodhomme (2017a). "Multimodel assessment of the impact of soil moisture initialization on mid-latitude summer predictability". In: *Climate Dynamics* 49(11-12), 3959 –3974. DOI: 10.1007/s00382-017-3555-7.

Ardilouze, C., L. Batté, and M. Déqué (2017b). "Subseasonal-to-seasonal (S2S) forecasts with CNRM-CM: a case study on the July 2015 West-European heat wave". In: *Advances in Science and Research* 14, 115 –121. DOI: 10.5194/asr-14-115-2017.

Baehr, J., K. Fröhlich, M. Botzet, D. I. V. Domeisen, L. Kornblueh, D. Notz, R. Piontek, H. Pohlmann, S. Tietsche, and W. A. Müller (2015). "The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model". In: *Climate Dynamics* 44, 2723 —2735. DOI: 10.1007/s00382-014-2399-7.

Baldwin, M. P., L. J. Gray, T. J. Dunkerton, K. Hamilton, P. H. Haynes, W. J. Randel, J. R. Holton, M. J. Alexander, I. Hirota, T. Horinouchi, D. B. A. Jones, J. S. Kinnersley, C. Marquardt, K. Sato, and M. Takahashi (2001). "The quasi-biennial oscillation". In: *Reviews of Geophysics* 39(2), 179–229. DOI: 10.1029/1999RG000073.

Baldwin, M. P., B. Ayarzagüena, T. Birner, N. Butchart, A. H. Butler, A. J. Charlton-Perez, D. I. V. Domeisen, C. I. Garfinkel, H. Garny, E. P. Gerber, M. I. Hegglin, U. Langematz, and N. M. Pedatella (2021). "Sudden Stratospheric Warmings". In: *Reviews of Geophysics* 59(1), e2020RG000708. DOI: 10.1029/2020RG000708.

Baldwin, M. P. and T. J. Dunkerton (2001). "Stratospheric Harbingers of Anomalous Weather Regimes". In: *Science* 294(5542), 581 –584. DOI: 10.1126/science.1063315.

Baldwin, M. P. and J. R. Holton (1988). "Climatology of the Stratospheric Polar Vortex and Planetary Wave Breaking". In: *Journal of Atmospheric Sciences* 45(7), 1123 –1142. DOI: 10.1175/1520-0469(1988)045<1123:COTSPV>2.0.CO;2.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. Garcia-Herrera (2011). "The Hot Summer of 2010: Redrawing the Temperature Map of Europe". In: *Science* 332, 220 –224. DOI: 10.1080/10255842.2015.1069566.

Bauer, P., A. Thorpe, and G. Brunet (2015). "The quiet revolution of numerical weather prediction". In: *Nature* 525(7567), 47 –55. DOI: 10.1038/nature14956.

Becker, E. J. (2017). "Prediction of Short-Term Climate Extremes with a Multimodel Ensemble". In: *Climate Extremes*. American Geophysical Union (AGU). Chap. 21, 347–359. DOI: 10.1002/9781119068020.ch21.

Beerli, R. and C. M. Grams (2019). "Stratospheric modulation of the large-scale circulation in the Atlantic–European region and its implications for surface weather events". In: *Quarterly Journal of the Royal Meteorological Society* 145(725), 3732 –3750. DOI: 10.1002/qj.3653.

Beerli, R., H. Wernli, and C. M. Grams (2017). "Does the lower stratosphere provide predictability for month-ahead wind electricity generation in Europe?" In: *Quarterly Journal of the Royal Meteorological Society* 143, 3025 –3036. DOI: 10.1002/qj.3158.

Benedetti, A. and F. Vitart (2018). "Can the Direct Effect of Aerosols Improve Subseasonal Predictability?" In: *Monthly Weather Review* 146(10), 3481 –3498. DOI: 10.1175/MWR-D-17-0282.1.

Boer, G. J. (2009). "Climate trends in a seasonal forecasting system". In: *Atmosphere - Ocean* 47(2), 123 –138. DOI: 10.3137/AO1002.2009.

Bollasina, M. and G. Messori (2018). "On the link between the subseasonal evolution of the North Atlantic Oscillation and East Asian climate". In: *Climate Dynamics* 51, 3537 –3557. DOI: 10.1007/s00382-018-4095-5.

Bradley, A. A. and S. S. Schwartz (2011). "Summary Verification Measures and Their Interpretation for Ensemble Forecasts". In: *Monthly Weather Review* 139(9), 3075 –3089. DOI: 10.1175/2010MWR3305.1.

Branstator, G. (2002). "Circumglobal Teleconnections, the Jet Stream Waveguide, and the North Atlantic Oscillation". In: *Journal of Climate* 15, 1893 –1910. DOI: 10.1175/1520-0442(2002)015<1893:CTTJSW>2.0.CO;2.

Brönnimann, S. (2007). "Impact of El Niño–Southern Oscillation on European climate". In: *Reviews of Geophysics* 45(3). DOI: 10.1029/2006RG000199.

Brunner, L., N. Schaller, J. Anstey, J. Sillmann, and A. K. Steiner (2018). "Dependence of Present and Future European Temperature Extremes on the Location of Atmospheric Blocking". In: *Geophysical Research Letters* 45. DOI: 10.1029/2018GL077837.

Buehler, T., C. C. Raible, and T. F. Stocker (2011). "The relationship of winter season North Atlantic blocking frequencies to extreme cold or dry spells in the ERA-40". In: *Tellus A* 63(2), 212–222. DOI: 10.1111/j.1600-0870.2010.00492.x.

Büeler, D., R. Beerli, H. Wernli, and C. M. Grams (2020). "Stratospheric influence on ECMWF sub-seasonal forecast skill for energy-industry-relevant surface weather in European countries". In: *Quarterly Journal of the Royal Meteorological Society* 146, 3675 — 3694. DOI: 10.1002/qj.3866.

Buizza, R., M. Milleer, and T. N. Palmer (1999). "Stochastic representation of model uncertainties in the ECMWF ensemble prediction system". In: *Quarterly Journal of the Royal Meteorological Society* 125(560), 2887–2908. DOI: `10.1002/qj.49712556006`.

Buizza, R. and M. Leutbecher (2015). "The forecast skill horizon". In: *Quarterly Journal of the Royal Meteorological Society* 141(693), 3366–3382. DOI: `10.1002/qj.2619`.

Bunzel, F., W. A. Müller, M. Dobrynin, K. Fröhlich, S. Hagemann, H. Pohlmann, T. Stacke, and J. Baehr (2018). "Improved Seasonal Prediction of European Summer Temperatures With New Five-Layer Soil-Hydrology Scheme". In: *Geophysical Research Letters* 45, 346 –353. DOI: `10.1002/2017GL076204`.

Butler, A., A. Charlton-Perez, D. I. Domeisen, C. Garfinkel, E. P. Gerber, P. Hitchcock, A. Y. Karpechko, A. C. Maycock, M. Sigmond, I. Simpson, and S.-W. Son (2019). "Subseasonal Predictability and the Stratosphere". In: *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Ed. by A. W. Robertson and F. Vitart. 1st. Elsevier. Chap. 11, 223 –241.

Camargo, S. J., J. Camp, R. L. Elsberry, P. A. Gregory, P. J. Klotzbach, C. J. Schreck, A. H. Sobel, M. J. Ventrice, F. Vitart, Z. Wang, M. C. Wheeler, M. Yamaguchi, and R. Zhan (2019). "Tropical Cyclone Prediction on Subseasonal Time-Scales". In: *Tropical Cyclone Research and Review* 8(3), 150 –165. DOI: `10.1016/j.tcrr.2019.10.004`.

Cassou, C. (2008). "Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation". In: *Nature* 455(7212), 523 –527. DOI: `10.1038/nature07286`.

Cassou, C., L. Terray, and A. S. Phillips (2005). "Tropical Atlantic influence on European heat waves". In: *Journal of Climate* 18(15), 2805 –2811. DOI: `10.1175/JCLI3506.1`.

Charlton-Perez, A. J., L. Ferranti, and R. W. Lee (2018). "The influence of the stratospheric state on North Atlantic weather regimes". In: *Quarterly Journal of the Royal Meteorological Society* 144(713), 1140–1151. DOI: `10.1002/qj.3280`.

Charney, J. G. and P. G. Drazin (1961). "Propagation of planetary-scale disturbances from the lower into the upper atmosphere". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 66(1), 83 –109. DOI: `10.1029/JZ066i001p00083`.

Chevallier, M., F. Massonnet, H. Goessling, V. Guémas, and T. Jung (2019). "The Role of Sea Ice in Sub-seasonal Predictability". In: *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Ed. by A. W. Robertson and F. Vitart. 1st. Elsevier. Chap. 10, 201 –221.

Cohen, J. and D. Entekhabi (1999). "Eurasian snow cover variability and Northern Hemisphere climate predictability". In: *Geophysical Research Letters* 26(8), 1051. DOI: `10.1029/1999GL900200`.

Cohen, J., J. C. Furtado, J. Jones, M. Barlow, D. Whittleston, and D. Entekhabi (2014). "Linking Siberian snow cover to precursors of stratospheric variability". In: *Journal of Climate* 27(14), 5422 –5432. DOI: `10.1175/JCLI-D-13-00779.1`.

Compo, G. P., P. D. Sardeshmukh, and C. Penland (2001). "Changes of Subseasonal Variability Associated with El Niño". In: *Journal of Climate* 14(16), 3356 –3374. DOI: `10.1175/1520-0442(2001)014<3356:COSVAW>2.0.CO;2`.

Coumou, D. and S. Rahmstorf (2012). "A decade of weather extremes". In: *Nature Climate Change* 2(7), 491 –496. DOI: 10.1038/nclimate1452.

Dawson, A. (2016). "Windspharm: A High-Level Library for Global Wind Field Computations Using Spherical Harmonics". In: *Journal of Open Research Software* 4(1), e31. DOI: 10.5334/jors.129.

Dawson, A., T. N. Palmer, and S. Corti (2012). "Simulating regime structures in weather and climate prediction models". In: *Geophysical Research Letters* 39(21). DOI: 10.1029/2012GL053284.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. Mcnally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut, and F. Vitart (2011). "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system". In: *Quarterly Journal of the Royal Meteorological Society* 137, 553 –597. DOI: 10.1002/qj.828.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion (2017). "Predictability of week-3-4 average temperature and precipitation over the contiguous United States". In: *Journal of Climate* 30(10), 3499 –3512. DOI: 10.1175/JCLI-D-16-0567.1.

DelSole, T. and M. K. Tippett (2009a). "Average Predictability Time. Part I: Theory". In: *Journal of the Atmospheric Sciences* 66(5), 1172 –1187. DOI: 10.1175/2008JAS2868.1.

— (2009b). "Average Predictability Time. Part II: Seamless Diagnoses of Predictability on Multiple Time Scales". In: *Journal of the Atmospheric Sciences* 66(5), 1188 –1204. DOI: 10.1175/2008JAS2869.1.

Dettinger, M. (2011). "Climate change, atmospheric rivers, and floods in California – a multimodel analysis of storm frequency and magnitude changes". In: *Journal of the American Water Resources Association* 47(3), 514 –523. DOI: 10.1111/j.1752-1688.2011.00546.x.

Di Capua, G., M. Kretschmer, R. V. Donner, B. van den Hurk, R. Vellore, R. Krishnan, and D. Coumou (2020). "Tropical and mid-latitude teleconnections interacting with the Indian summer monsoon rainfall: a theory-guided causal effect network approach". In: *Earth System Dynamics* 11(1), 17 –34. DOI: 10.5194/esd-11-17-2020.

Diffenbaugh, N. S., D. Singh, J. S. Mankin, D. E. Horton, D. L. Swain, D. Touma, A. Charland, Y. Liu, M. Haugen, M. Tsiang, and B. Rajaratnam (2017). "Quantifying the influence of global warming on unprecedented extreme climate events". In: *PNAS* 114(19), 4881 –4886. DOI: 10.1073/pnas.1618082114.

Ding, Q. and B. Wang (2005). "Circumglobal teleconnection in the Northern Hemisphere summer". In: *Journal of Climate* 18, 3483 –3505. DOI: 10.1175/JCLI3473.1.

Ding, Q., B. Wang, J. M. Wallace, and G. Branstator (2011). "Tropical-extratropical teleconnections in boreal summer: Observed interannual variability". In: *Journal of Climate* 24(7), 1878 –1896. DOI: 10.1175/2011JCLI3621.1.

Dirmeyer, P. A., P. Gentine, M. B. Ek, and G. Balsamo (2019). "Land Surface Processes Relevant to Sub-seasonal to Seasonal (S2S) Prediction". In: *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Ed. by A. W. Robertson and F. Vitart. 1st. Elsevier. Chap. 8, 165 –181.

Dirmeyer, P. A., C. Peters-Lidard, and G. Balsamo (2015). "Land-Atmosphere Interactions and the Water Cycle". In: *Seamless Prediction of the Earth System: From Minutes to Months*. Ed. by G. Brunet, S. Jones, and P. M. Ruti. Vol. WMO-1156. World Meteorological Organization, 145 –155.

Doblas-Reyes, F. J., R. Hagedorn, T. N. Palmer, and J. J. Morcrette (2006). "Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts". In: *Geophysical Research Letters* 33(7), 1 –5. DOI: 10.1029/2005GL025061.

Dobrynin, M., D. I. V. Domeisen, W. A. Müller, L. Bell, S. Brune, F. Bunzel, A. Düsterhus, K. Fröhlich, H. Pohlmann, and J. Baehr (2018). "Improved Teleconnection-Based Dynamical Seasonal Predictions of Boreal Winter". In: *Geophysical Research Letters* 45(8), 3605–3614. DOI: 10.1002/2018GL077209.

Dole, R., M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, T. Zhang, X. W. Quan, T. Xu, and D. Murray (2011). "Was there a basis for anticipating the 2010 Russian heat wave?" In: *Geophysical Research Letters* 38(6), L06702. DOI: 10.1029/2010GL046582.

Domeisen, D. I. V., C. M. Grams, and L. Papritz (2020a). "The role of North Atlantic–European weather regimes in the surface impact of sudden stratospheric warming events". In: *Weather and Climate Dynamics* 1(2), 373 –388. DOI: 10.5194/wcd-1-373-2020.

Domeisen, D. I. V., A. H. Butler, A. J. Charlton-Perez, B. Ayarzagüena, M. P. Baldwin, E. Dunn-Sigouin, J. C. Furtado, C. I. Garfinkel, P. Hitchcock, A. Y. Karpechko, H. Kim, J. Knight, A. L. Lang, E.-P. Lim, A. Marshall, G. Roff, C. Schwartz, I. R. Simpson, S.-W. Son, and M. Taguchi (2020b). "The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling". In: *Journal of Geophysical Research: Atmospheres* 125(2), e2019JD030923. DOI: 10.1029/2019JD030923.

Domeisen, D. I., A. H. Butler, A. J. Charlton-Perez, B. Ayarzagüena, M. P. Baldwin, E. Dunn-Sigouin, J. C. Furtado, C. I. Garfinkel, P. Hitchcock, A. Y. Karpechko, H. Kim, J. Knight, A. L. Lang, E.-P. Lim, A. Marshall, G. Roff, C. Schwartz, I. R. Simpson, S.-W. Son, and M. Taguchi (2020c). "The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 1. Predictability of the Stratosphere". In: *Journal of Geophysical Research: Atmospheres* 125(2), e2019JD030920. DOI: 10.1029/2019JD030920.

Domeisen, D. I., C. I. Garfinkel, and A. H. Butler (2019). "The Teleconnection of El Niño Southern Oscillation to the Stratosphere". In: *Reviews of Geophysics* 57(1), 5–47. DOI: https://doi.org/10.1029/2018RG000596.

Domeisen, D. and A. Butler (2020). "Stratospheric drivers of extreme events at the Earth's surface". In: *Communications Earth & Environment* 1(59), 3732 –3750. DOI: 10.1038/s43247-020-00060-z.

Dorrington, J., I. Finney, T. Palmer, and A. Weisheimer (2020). "Beyond skill scores: exploring sub-seasonal forecast value through a case-study of French month-ahead energy prediction". In: *Quarterly Journal of the Royal Meteorological Society* 146(733), 3623–3637. DOI: 10.1002/qj.3863.

Douville, H. (2010). "Relative contribution of soil moisture and snow mass to seasonal climate predictability: a pilot study". In: *Climate Dynamics* 34, 797—818. DOI: 10.1007/s00382-008-0508-1.

Drijfhout, S., G. J. van Oldenborgh, and A. Cimatoribus (2012). "Is a Decline of AMOC Causing the Warming Hole above the North Atlantic in Observed and Modeled Warming Patterns?" In: *Journal of Climate* 25(24), 8373 –8379. DOI: 10.1175/JCLI-D-12-00490.1.

Duchez, A., E. Frajka-Williams, S. A. Josey, D. G. Evans, J. P. Grist, R. Marsh, G. D. McCarthy, B. Sinha, D. I. Berry, and J. Hirschi (2016). "Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 European heat wave". In: *Environmental Research Letters* 11(7), 074004. DOI: 10.1088/1748-9326/11/7/074004.

Ehrendorfer, M. (2006). "The Liouville equation and atmospheric predictability". In: *Predictability of Weather and Climate*. Ed. by T. Palmer and R. Hagedorn. 1st. Cambridge University Press. Chap. 4, 59 –98.

Ehrendorfer, M. (1994). "The Liouville Equation and Its Potential Usefulness for the Prediction of Forecast Skill. Part I: Theory". In: *Monthly Weather Review* 122(4), 703 –713. DOI: 10.1175/1520-0493(1994)122<0703:TLEAIP>2.0.CO;2.

Ferranti, L., T. N. Palmer, F. Molteni, and E. Klinker (1990). "Tropical-Extratropical Interaction Associated with the 30–60 Day Oscillation and Its Impact on Medium and Extended Range Prediction". In: *Journal of Atmospheric Sciences* 47(18), 2177 –2199. DOI: 10.1175/1520-0469(1990)047<2177:TEIAWT>2.0.CO;2.

Ferranti, L., S. Corti, and M. Janousek (2015). "Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector". In: *Quarterly Journal of the Royal Meteorological Society* 141(688), 916–924. DOI: 10.1002/qj.2411.

Ferranti, L., L. Magnusson, F. Vitart, and D. S. Richardson (2018). "How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe?" In: *Quarterly Journal of the Royal Meteorological Society* 144(715), 1788–1802. DOI: 10.1002/qj.3341.

Ferranti, L. and P. Viterbo (2006). "The European summer of 2003: Sensitivity to soil water initial conditions". In: *Journal of Climate* 19(15), 3659 –3680. DOI: 10.1175/JCLI3810.1.

Ferro, C. A. T. and D. B. Stephenson (2011). "Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events". In: *Weather and Forecasting* 26(5), 699 –713. DOI: 10.1175/waf-d-10-05030.1.

— (2012). "Deterministic forecasts of extreme events and warnings". In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Ed. by I. T. Jolliffe and D. B. Stephenson. 2nd. University of Exeter, UK: John Wiley and Sons, Ltd. Chap. 9.

Ferro, C. A., D. S. Richardson, and A. P. Weigel (2008). "On the effect of ensemble size on the discrete and continuous ranked probability scores". In: *Quarterly Journal of the Royal Meteorological Society* 15, 19 –24. DOI: 10.1002/met.45.

Fink, A. H., T. Bruecher, A. Krueger, G. C. Leckebusch, J. G. Pinto, and U. Ulbrich (2004). "The 2003 European summer heatwaves and drought-synoptic diagnosis and impacts". In: *Weather* 59(8), 209 –216. DOI: 10.1256/wea.73.04.

Fischer, E. M., S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär (2007a). "Soil Moisture-Atmosphere Interactions during the 2003 European Summer Heat Wave". In: *Journal of Climate* 20(20), 5081 –5099. DOI: 10.1175/JCLI4288.1.

Fischer, E. M., S. I. Seneviratne, D. Lüthi, and C. Schär (2007b). "Contribution of land-atmosphere coupling to recent European summer heat waves". In: *Geophysical Research Letters* 34, L06707. DOI: 10.1029/2006GL029068.

Folland, C. K., J. Knight, H. W. Linderholm, D. Fereday, S. Ineson, and J. W. Hurrel (2009). "The summer North Atlantic oscillation: Past, present, and future". In: *Journal of Climate* 22(5), 1082 –1103. DOI: 10.1175/2008JCLI2459.1.

Fouillet, A., G. Rey, F. Laurent, G. Pavillon, S. Bellec, C. Guihenneuc-Jouyaux, J. Clavel, E. Jougla, and D. Hémon (2006). "Excess mortality related to the August 2003 heat wave in France". In: *International Archives of Occupational and Environmental Health* 80, 16 –24. DOI: 10.1007/s00420-006-0089-4.

Fragkoulidis, G., V. Wirth, P. Bossmann, and A. H. Fink (2018). "Linking Northern Hemisphere temperature extremes to Rossby wave packets". In: *Quarterly Journal of the Royal Meteorological Society* 144, 553 –566. DOI: 10.1002/qj.3228.

Franzke, C., S. Lee, and S. B. Feldstein (2004). "Is the North Atlantic Oscillation a Breaking Wave?" In: *Journal of the Atmospheric Sciences* 61(2), 145 –160. DOI: 10.1175/1520-0469(2004)061<0145:ITNAOA>2.0.CO;2.

Garfinkel, C. I. and C. Schwartz (2017). "MJO-Related Tropical Convection Anomalies Lead to More Accurate Stratospheric Vortex Variability in Subseasonal Forecast Models". In: *Geophysical Research Letters* 44(19), 10,054–10,062. DOI: 10.1002/2017GL074470.

Garfinkel, C. I., J. J. Benedict, and E. D. Maloney (2014). "Impact of the MJO on the boreal winter extratropical circulation". In: *Geophysical Research Letters* 41(16), 6055–6062. DOI: 10.1002/2014GL061094.

Garfinkel, C. I., C. Schwartz, I. P. White, and J. Rao (2020). "Predictability of the early winter Arctic oscillation from autumn Eurasian snowcover in subseasonal forecast models". In: *Climate Dynamics* 55, 961 –974. DOI: 10.1007/s00382-020-05305-3.

Gastineau, G. and C. Frankignoul (2015). "Influence of the North Atlantic SST variability on the atmospheric circulation during the twentieth century". In: *Journal of Climate* 28(4), 1396 –1416. DOI: 10.1175/JCLI-D-14-00424.1.

Gneiting, T. and A. E. Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102(477), 359 –378. DOI: 10.1198/016214506000001437.

Gneiting, T. and R. Ranjan (2011). "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules". In: *Journal of Business & Economic Statistics* 29(3), 411–422. DOI: 10.1198/jbes.2010.08110.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli (2017). "Balancing Europe's wind-power output through spatial deployment informed by weather regimes". In: *Nature Climate Change* 7, 557–562. DOI: 10.1038/nclimate3338.

Grazzini, F. and F. Vitart (2015). "Atmospheric predictability and Rossby wave packets". In: *Quarterly Journal of the Royal Meteorological Society* 141, 2793 –2802. DOI: 10.1002/qj.2564.

Greatbatch, R. J., G. Gollan, T. Jung, and T. Kunz (2012). "Factors influencing Northern Hemisphere winter mean atmospheric circulation anomalies during the period 1960/61 to 2001/02". In: *Quarterly Journal of the Royal Meteorological Society* 138(669), 1970–1982. DOI: 10.1002/qj.1947.

Guimarães Nobre, G., J. E. Hunink, B. Baruth, J. C. Aerts, and P. J. Ward (2019). "Translating large-scale climate variability into crop production forecast in Europe". In: *Scientific Reports* 9, 1277. DOI: 10.1038/s41598-018-38091-4.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005). "The rationale behind the success of mulit-model ensembles in seasonal forecasting – I. Basic Concepts". In: *Tellus* 57A, 219 —233. DOI: 10.3402/tellusa.v57i3.14657.

Hamill, T. M. and J. Juras (2006). "Measuring forecast skill: Is it real skill or is it the varying climatology?" In: *Quarterly Journal of the Royal Meteorological Society* 132(621 C), 2905 – 2923. DOI: 10.1256/qj.06.25.

Hansen, F., R. J. Greatbatch, G. Gollan, T. Jung, and A. Weisheimer (2017). "Remote control of North Atlantic Oscillation predictability via the stratosphere". In: *Quarterly Journal of the Royal Meteorological Society* 143(703), 706–719. DOI: 10.1002/qj.2958.

Hansen, F., T. Kruschke, R. J. Greatbatch, and A. Weisheimer (2019). "Factors Influencing the Seasonal Predictability of Northern Hemisphere Severe Winter Storms". In: *Geophysical Research Letters* 46(1), 365–373. DOI: 10.1029/2018GL079415.

Hardiman, S. C., N. J. Dunstone, A. A. Scaife, D. M. Smith, S. Ineson, J. Lim, and D. Fereday (2019). "The Impact of Strong El Niño and La Niña Events on the North Atlantic". In: *Geophysical Research Letters* 46(5), 2874 –2883. DOI: 10.1029/2018GL081776.

Hauser, M., R. Orth, and S. I. Seneviratne (2015). "Role of soil moisture vs . recent climate change for heat waves in western Russia". In: *Geophysical Research Letters* 43, 2819 – 2826. DOI: 10.1002/2016GL068036.

Hersbach, H. (2000). "Decomposition of the continuous ranked probability score for ensemble prediction systems". In: *Weather and Forecasting* 15(5), 559 –570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu,

G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146(730), 1999–2049. DOI: `10.1002/qj.3803`.

Hitchcock, P. and I. R. Simpson (2014). "The Downward Influence of Stratospheric Sudden Warmings". In: *Journal of the Atmospheric Sciences* 1(10), 3856 –3876. DOI: `10.1175/JAS-D-14-0012.1`.

Hogan, R. J. and I. B. Mason (2012). "Deterministic forecasts of binary events". In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Ed. by I. T. Jolliffe and D. B. Stephenson. 2nd. University of Exeter, UK: John Wiley and Sons, Ltd. Chap. 3.

Holton, J. R. (2013). *An Introduction to Dynamic Meteorology*. Ed. by J. R. Holton and G. J. Hakim. 5th. Academic Press, Elsevier.

Hoskins, B. (2013). "The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science". In: *Quarterly Journal of the Royal Meteorological Society* 139(672), 573 –584. DOI: `10.1002/qj.1991`.

Hudson, D. and A. G. Marshall (2016). *Extending the Bureau's heatwave forecast to multi-week timescales*. Tech. rep. September, 42 pp. DOI: `10.22499/4.0016`.

Hurrell, J. W. and C. Deser (2009). "North Atlantic climate variability: The role of the North Atlantic Oscillation". In: *Journal of Marine Systems* 78(1), 28 –41. DOI: `10.1016/j.jmarsys.2008.11.026`.

Hurrell, J. W., Y. Kushnir, G. Ottersen, and M. Visbeck (2003). "An Overview of the North Atlantic Oscillation". In: *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*. American Geophysical Union (AGU), 1–35. ISBN: 9781118669037. DOI: `10.1029/134GM01`.

Imada, Y., H. Shiogama, C. Takahashi, M. Watanabe, M. Mori, Y. Kamae, and S. Maeda (2018). "Climate Change Increased the Likelihood of the 2016 Heat Extremes in Asia, In: Explaining Extremes of 2016 from a Climate Perspective". In: *Bulletin of the American Meteorological Society* 99, S97 –S101. DOI: `10.1175/bams-d-17-0109.1`.

IPCC (1990). *The IPCC Scientific Assessment*. Ed. by J. Houghton, G. Jenkins, and J. Ephraums. Cambridge University Press. URL: `https://www.ipcc.ch/report/ar1/wg1/`.

— (2013). "Notes: 1. global carbon dioxide budget, section 6.3.2". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 1535 pp.

Iza, M., N. Calvo, and E. Manzini (2016). "The Stratospheric Pathway of La Niña". In: *Journal of Climate* 29(24), 8899 –8914. DOI: `10.1175/JCLI-D-16-0230.1`.

Jeong, J. H., H. W. Linderholm, S. H. Woo, C. Folland, B. M. Kim, S. J. Kim, and D. Chen (2013). "Impacts of snow initialization on subseasonal forecasts of surface air temperature for the cold season". In: *Journal of Climate* 26(6), 1956 –1972. DOI: `10.1175/JCLI-D-12-00159.1`.

Jiménez-Esteve, B. and D. I. V. Domeisen (2018). "The Tropospheric Pathway of the ENSO-North Atlantic Teleconnection". In: *Journal of Climate* 31(11), 4563 –4584. DOI: `10.1175/JCLI-D-17-0716.1`.

— (2020). "Nonlinearity in the tropospheric pathway of ENSO to the North Atlantic". In: *Weather and Climate Dynamics* 1, 225 –245. DOI: `10.5194/wcd-1-225-2020`.

Jin, F. and B. J. Hoskins (1995). "The Direct Response to Tropical Heating in a Baroclinic Atmosphere". In: *Journal of Atmospheric Sciences* 52(3), 307 –319. DOI: `10.1175/1520-0469(1995)052<0307:TDRTTH>2.0.CO;2`.

Johnson, S. J., T. N. Stockdale, L. Ferranti, M. A. Balmaseda, F. Molteni, L. Magnusson, S. Tietsche, D. Decremer, A. Weisheimer, G. Balsamo, S. P. Keeley, K. Mogensen, H. Zuo, and B. M. Monge-Sanz (2019). "SEAS5: The new ECMWF seasonal forecast system". In: *Geoscientific Model Development* 12(3), 1087 –1117. DOI: `10.5194/gmd-12-1087-2019`.

Jolliffe, I. T. and D. B. Stephenson (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. University of Exeter, UK: John Wiley and Sons, Ltd.

Jung, T., M. A. Kasper, T. Semmler, and S. Serrar (2014). "Arctic influence on subseasonal midlatitude prediction". In: *Geophysical Research Letters* 41(10), 3676 –3680. DOI: `10.1002/2014GL059961`.

Kalnay, E., B. Hunt, E. Ott, and I. Szunyogh (2006). "Ensemble forecasting and data assimilation: two problems with the same solution?" In: *Predictability of Weather and Climate*. Ed. by T. Palmer and R. Hagedorn. 1st. Cambridge University Press. Chap. 7, 157 –180.

Karpechko, A. Y. (2015). "Improvements in statistical forecasts of monthly and two-monthly surface air temperatures using a stratospheric predictor". In: *Quarterly Journal of the Royal Meteorological Society* 141(691), 2444–2456. DOI: `10.1002/qj.2535`.

Kolstad, E. W., E. A. Barnes, and S. P. Sobolowski (2017). "Quantifying the role of land–atmosphere feedbacks in mediating near-surface temperature persistence". In: *Quarterly Journal of the Royal Meteorological Society* 143(704), 1620 –1631. DOI: `10.1002/qj.3033`.

Kolstad, E. W., C. O. Wulff, D. I. V. Domeisen, and T. Woollings (2020). "Tracing North Atlantic Oscillation Forecast Errors to Stratospheric Origins". In: *Journal of Climate* 33(21), 9145 –9157. DOI: `10.1175/JCLI-D-20-0270.1`.

Koster, R. D., S. P. P. Mahanama, T. J. Yamada, G. Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, G. Drewitt, C. T. Gordon, Z. Guo, J.-H. Jeong, W.-S. Lee, Z. Li, L. Luo, S. Malyshev, W. J. Merryfield, S. I. Seneviratne, T. Stanelle, B. J. J. M. van den Hurk, F. Vitart, and E. F. Wood (2011). "The Second Phase of the Global Land–Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill". In: *Journal of Hydrometeorology* 12(5), 805 –822. DOI: `10.1175/2011JHM1365.1`.

Koster, R. D., S. P. Mahanama, T. J. Yamada, G. Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, G. Drewitt, C. T. Gordon, Z. Guo, J. H. Jeong, D. M. Lawrence, W. S. Lee, Z. Li, L. Luo, S. Malyshev, W. J. Merryfield, S. I. Seneviratne, T. Stanelle, B. J. Van Den Hurk, F. Vitart, and E. F. Wood (2010). "Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment". In: *Geophysical Research Letters* 37, L02402. DOI: 10.1029/2009GL041677.

Kushnir, Y., W. A. Robinson, I. Bladé, N. M. J. Hall, S. Peng, and R. Sutton (2002). "Atmospheric GCM Response to Extratropical SST Anomalies: Synthesis and Evaluation". In: *Journal of Climate* 15(16), 2233 –2256. DOI: 10.1175/1520-0442(2002)015<2233:AGRTES>2.0.CO;2.

Labitzke, K. (1977). "Interannual Variability of the Winter Stratosphere in the Northern Hemisphere". In: *Monthly Weather Review* 105(6), 762 –770. DOI: 10.1175/1520-0493(1977)105<0762:IVOTWS>2.0.CO;2.

Lavaysse, C., G. Naumann, L. Alfieri, P. Salamon, and J. Vogt (2018). "Predictability of European heat and cold waves". In: *Climate Dynamics*. DOI: 10.1007/s00382-018-4273-5.

Lee, R. W., S. J. Woolnough, A. J. Charlton-Perez, and F. Vitart (2019). "ENSO Modulation of MJO Teleconnections to the North Atlantic and Europe". In: *Geophysical Research Letters* 46(22), 13535 –13545. DOI: 10.1029/2019GL084683.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting (2017). "Forecaster's Dilemma: Extreme Events and Forecast Evaluation". In: *Statistical Science* 32(1), 106–127. DOI: 10.1214/16-STS588.

Leutbecher, M. and T. Palmer (2008). "Ensemble forecasting". In: *Journal of Computational Physics* 227(7), 3515 –3539. DOI: 10.1016/j.jcp.2007.02.014.

Leutbecher, M. (2019). "Ensemble size: How suboptimal is less than infinity?" In: *Quarterly Journal of the Royal Meteorological Society* 145(S1), 107 –128. DOI: 10.1002/qj.3387.

Leutbecher, M., S.-J. Lock, P. Ollinaho, S. T. K. Lang, G. Balsamo, P. Bechtold, M. Bonavita, H. M. Christensen, M. Diamantakis, E. Dutra, S. English, M. Fisher, R. M. Forbes, J. Goddard, T. Haiden, R. J. Hogan, S. Juricke, H. Lawrence, D. MacLeod, L. Magnusson, S. Malardel, S. Massart, I. Sandu, P. K. Smolarkiewicz, A. Subramanian, F. Vitart, N. Wedi, and A. Weisheimer (2017). "Stochastic representations of model uncertainties at ECMWF: state of the art and future vision". In: *Quarterly Journal of the Royal Meteorological Society* 143(707), 2315–2339. DOI: 10.1002/qj.3094.

Liebmann, B. and C. A. Smith (1996). "Description of a Complete (Interpolated) Outgoing Longwave Radiation Dataset". In: *Bulletin of the American Meteorological Society* 77(6), 1275–1277.

Lim, Y., S.-W. Son, and D. Kim (2018). "MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models". In: *Journal of Climate* 31(10), 4075 –4094. DOI: 10.1175/JCLI-D-17-0545.1.

Limpasuvan, V., D. L. Hartmann, D. W. J. Thompson, K. Jeev, and Y. L. Yung (2005). "Stratosphere-troposphere evolution during polar vortex intensification". In: *Journal of Geophysical Research: Atmospheres* 110(D24). DOI: https://doi.org/10.1029/2005JD006302.

Lin, H. (2020). "Subseasonal Forecast Skill over the Northern Polar Region in Boreal Winter". In: *Journal of Climate* 33(5), 1935 –1951. DOI: 10.1175/JCLI-D-19-0408.1.

Lin, H., G. Brunet, and J. Derome (2009). "An Observed Connection between the North Atlantic Oscillation and the Madden–Julian Oscillation". In: *Journal of Climate* 22(2), 364 –380. DOI: 10.1175/2008JCLI2515.1.

Lin, H., G. Brunet, and J. S. Fontecilla (2010). "Impact of the Madden-Julian Oscillation on the intraseasonal forecast skill of the North Atlantic Oscillation". In: *Geophysical Research Letters* 37(19). DOI: 10.1029/2010GL044315.

Liniger, M. A., H. Mathis, C. Appenzeller, and F. J. Doblas-Reyes (2007). "Realistic greenhouse gas forcing and seasonal forecasts". In: *Geophysical Research Letters* 34(4), L04705. DOI: 10.1029/2006GL028335.

Liu, Z. and M. Alexander (2007). "Atmospheric bridge, oceanic tunnel, and global climatic teleconnections". In: *Reviews of Geophysics* 45(2). DOI: 10.1029/2005RG000172.

Livezey, R. E. (1999). "The evaluation of forecasts". In: *Analysis of Climate Variability*. Ed. by H. von Storch and A. Navarra. 2nd. Berlin: Springer-Verlag, 179 –198.

Lopez, A., E. Coughlan de Perez, J. Bazo, P. Suarez, B. van den Hurk, and M. van Aalst (2020). "Bridging forecast verification and humanitarian decisions: A valuation approach for setting up action-oriented early warnings". In: *Weather and Climate Extremes* 27, 100167. DOI: 10.1016/j.wace.2018.03.006.

Lorenz, E. N. (1969). "The predictability of a flow which possesses many scales of motion". In: *Tellus* 21(3), 289–307. DOI: 10.3402/tellusa.v21i3.10086.

MacLeod, D., C. O'Reilly, T. Palmer, and A. Weisheimer (2018). "Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics". In: *Atmospheric Science Letters* 19(5), e815. DOI: 10.1002/asl.815.

Madden, R. A. and P. R. Julian (1971). "Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific". In: *Journal of Atmospheric Sciences* 28(5), 702 –708. DOI: 0.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2.

Madonna, E., D. S. Battisti, C. Li, and R. H. White (2021). "Reconstructing winter climate anomalies in the Euro-Atlantic sector using circulation patterns". In: *Weather and Climate Dynamics Discussions* 2021, 1 –26. DOI: 10.5194/wcd-2021-6.

Madonna, E., C. Li, C. M. Grams, and T. Woollings (2017). "The link between eddy-driven jet variability and weather regimes in the North Atlantic-European sector". In: *Quarterly Journal of the Royal Meteorological Society* 143(708), 2960–2972. DOI: 10.1002/qj.3155.

Mann, M. E., S. Rahmstorf, K. Kornhuber, B. A. Steinman, S. K. Miller, and D. Coumou (2017). "Influence of Anthropogenic Climate Change on Planetary Wave Resonance and Extreme Weather Events". In: *Science Reports* 7, 45242. DOI: 10.1038/srep46822.

Manrique-Suñén, A., N. Gonzalez-Reviriego, V. Torralba, N. Cortesi, and F. J. Doblas-Reyes (2020). "Choices in the verification of S2S forecasts and their implications for climate services". In: *Monthly Weather Review* 148(10), 3995 –4008. DOI: 10.1175/mwr-d-20-0067.1.

Materia, S., Á. G. Muñoz, M. C. Álvarez-Castro, S. J. Mason, F. Vitart, and S. Gualdi (2020). "Multimodel subseasonal forecasts of spring cold spells: Potential value for the hazelnut agribusiness". In: *Weather and Forecasting* 35(1), 237 –254. DOI: 10.1175/WAF-D-19-0086.1.

Matsueda, M. and T. N. Palmer (2018). "Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts". In: *Quarterly Journal of the Royal Meteorological Society* 144(713), 1012 –1027. DOI: 10.1002/qj.3265.

McKinnon, K. A., A. Rhines, M. P. Tingley, and P. Huybers (2016). "Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures". In: *Nature Geoscience* 9(5), 389 –394. DOI: 10.1038/ngeo2687.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. Mitchell, R. J. Stouffer, and K. E. Taylor (2007). "The WCRP CMIP3 multimodel dataset: A new era in climatic change research". In: *Bulletin of the American Meteorological Society* 88(9), 1383 –1394. DOI: 10.1175/BAMS-88-9-1383.

Merryfield, W. J., J. Baehr, L. Batté, E. J. Becker, A. H. Butler, C. A. S. Coelho, G. Danabasoglu, P. A. Dirmeyer, F. J. Doblas-Reyes, D. I. V. Domeisen, L. Ferranti, T. Ilynia, A. Kumar, W. A. Müller, M. Rixen, A. W. Robertson, D. M. Smith, Y. Takaya, M. Tuma, F. Vitart, C. J. White, M. S. Alvarez, C. Ardilouze, H. Attard, C. Baggett, M. A. Balmaseda, A. F. Beraki, P. S. Bhattacharjee, R. Bilbao, F. M. de Andrade, M. J. DeFlorio, L. B. Díaz, M. A. Ehsan, G. Fragkoulidis, S. Grainger, B. W. Green, M. C. Hell, J. M. Infanti, K. Isensee, T. Kataoka, B. P. Kirtman, N. P. Klingaman, J.-Y. Lee, K. Mayer, R. McKay, J. V. Mecking, D. E. Miller, N. Neddermann, C. H. J. Ng, A. Ossó, K. Pankatz, S. Peatman, K. Pegion, J. Perlwitz, G. C. Recalde-Coronel, A. Reintges, C. Renkl, B. Solaraju-Murali, A. Spring, C. Stan, Y. Q. Sun, C. R. Tozer, N. Vigaud, S. Woolnough, and S. Yeager (2020). "Current and Emerging Developments in Subseasonal to Decadal Prediction". In: *Bulletin of the American Meteorological Society* 101(6), E869 –E896. DOI: 10.1175/BAMS-D-19-0037.1.

Michelangeli, P.-A., R. Vautard, and B. Legras (1995). "Weather Regimes: Recurrence and Quasi Stationarity". In: *Journal of Atmospheric Sciences* 2(8), 1237 –1256. DOI: 10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2.

Miralles, D. G., A. J. Teuling, C. C. Van Heerwaarden, and J. V. G. De Arellano (2014). "Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation". In: *Nat. Geosci.* 7(5), 345 –349. DOI: 10.1038/ngeo2141.

Miyakoda, K., T. Gordon, R. Caverly, W. Stern, J. Sirutis, and W. Bourke (1983). "Simulation of a Blocking Event in January 1977". In: *Monthly Weather Review* 111(4), 846 –869. DOI: 10.1175/1520-0493(1983)111<0846:SOABEI>2.0.CO;2.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger (2005). "A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes". In: *Journal of Climate* 18(10), 1513 –1523. DOI: 10.1175/JCLI3361.1.

Mundhenk, B., E. Barnes, E. Maloney, and C. F. Baggett (2018). "Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian oscillation and quasi-biennial oscillation". In: *npj Climate and Atmospheric Science* 1, 20177. DOI: 10.1038/s41612-017-0008-2.

Murphy, A. H. and E. S. Epstein (1989). "Skill Scores and Correlation Coefficients in Model Verification". In: *Monthly Weather Review* 117(3), 572 –582. DOI: 10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Narapusetty, B., T. Delsole, and M. K. Tippett (2009). "Optimal estimation of the climatological mean". In: *Journal of Climate* 22(18), 4845 –4859. DOI: 10.1175/2009JCLI2944.1.

Neddermann, N. C., W. A. Müller, M. Dobrynin, A. Düsterhus, and J. Baehr (2019). "Seasonal predictability of European summer climate re-assessed". In: *Climate Dynamics*. DOI: 10.1007/s00382-019-04678-4.

Newman, M., M. A. Alexander, T. R. Ault, K. M. Cobb, C. Deser, E. D. Lorenzo, N. J. Mantua, A. J. Miller, S. Minobe, H. Nakamura, N. Schneider, D. J. Vimont, A. S. Phillips, J. D. Scott, and C. A. Smith (2016). "The Pacific Decadal Oscillation, Revisited". In: *Journal of Climate* 29(12), 4399 –4427. DOI: 10.1175/JCLI-D-15-0508.1.

Nie, Y., H.-L. Ren, and Y. Zhang (2019). "The Role of Extratropical Air–Sea Interaction in the Autumn Subseasonal Variability of the North Atlantic Oscillation". In: *Journal of Climate* 32(22), 7697 –7712. DOI: 10.1175/JCLI-D-19-0060.1.

Osman, M. and M. S. Alvarez (2018). "Subseasonal prediction of the heat wave of December 2013 in Southern South America by the POAMA and BCC-CPS models". In: *Climate Dynamics* 50, 67 –81. DOI: 10.1007/s00382-017-3582-4.

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong (2017). "Observational evidence of European summer weather patterns predictable from spring". In: *PNAS* 115, 59 –63. DOI: 10.1073/pnas.1713146114.

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong (2020). "Development, Amplification, and Decay of Atlantic/European Summer Weather Patterns Linked to Spring North Atlantic Sea Surface Temperatures". In: *Journal of Climate* 33(14), 5939 –5951. DOI: 10.1175/JCLI-D-19-0613.1.

O'Reilly, C. H., T. Woollings, L. Zanna, and A. Weisheimer (2018). "The Impact of Tropical Precipitation on Summertime Euro-Atlantic Circulation via a Circumglobal Wave Train". In: *Journal of Climate* 31(16), 6481 –6504. DOI: 10.1175/JCLI-D-17-0451.1.

Pal, J. S. and E. A. B. Eltahir (2003). "A feedback mechanism between soil-moisture distribution and storm tracks". In: *Quarterly Journal of the Royal Meteorological Society* 129(592), 2279 –2297. DOI: 10.1256/qj.01.201.

Palmer, T. N. (1993). "Extended-Range Atmospheric Prediction and the Lorenz Model". In: *Bulletin of the American Meteorological Society* 74(1), 49 –65. DOI: 10.1175/1520-0477(1993)074<0049:ERAPAT>2.0.CO;2.

— (2006). "Predictability of weather and climate: from theory to practice". In: *Predictability of Weather and Climate*. Ed. by T. Palmer and R. Hagedorn. 1st. Cambridge University Press. Chap. 1, 1 –29.

Palmer, T. N. (2019). “Stochastic weather and climate models”. In: *Nature Reviews Physics* 1(7), 463 —471. DOI: 10.1038/s42254-019-0062-2.

Palmer, T. N. and D. L. T. Anderson (1994). “The prospects for seasonal forecasting – A review paper”. In: *Quarterly Journal of the Royal Meteorological Society* 120(518), 755 – 793. DOI: 10.1002/qj.49712051802.

Palmer, T. N. and D. Richardson (2014). “Decisions, decisions...!” In: *ECMWF Newsletter No. 141*. ECMWF, 12 –14. URL: https://www.ecmwf.int/sites/default/files/elibrary/2014/14584-newsletter-no141-autumn-2014.pdf.

Pegion, K., B. P. Kirtman, E. Becker, D. C. Collins, E. LaJoie, R. Burgman, R. Bell, T. DelSole, D. Min, Y. Zhu, W. Li, E. Sinsky, H. Guan, J. Gottschalck, E. J. Metzger, N. P. Barton, D. Achuthavarier, J. Marshak, R. D. Koster, H. Lin, N. Gagnon, M. Bell, M. K. Tippett, A. W. Robertson, S. Sun, S. G. Benjamin, B. W. Green, R. Bleck, and H. Kim (2019). “The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment”. In: *Bulletin of the American Meteorological Society* 100(10), 2043 –2060. DOI: 10.1175/BAMS-D-18-0270.1.

Pelletier, J. D. (1998). “The power spectral density of atmospheric temperature from time scales of 10^{-2} to 10^{6} yr”. In: *Earth and Planetary Science Letters* 158(3), 157 –164. DOI: 10.1016/S0012-821X(98)00051-X.

Pelly, J. L. and B. J. Hoskins (2003). “How well does the ECMWF Ensemble Prediction System predict blocking?” In: *Quarterly Journal of the Royal Meteorological Society* 129(590), 1683–1702. DOI: 10.1256/qj.01.173.

Perez, E. Coughlan de, B. van den Hurk, M. K. van Aalst, B. Jongman, T. Klose, and P. Suarez (2015). “Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts”. In: *Natural Hazards and Earth System Sciences* 15(4), 895 –904. DOI: 10.5194/nhess-15-895-2015.

Pfahl, S. and H. Wernli (2012). “Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Northern Hemisphere on (sub-)daily time scales”. In: *Geophysical Research Letters* 39, L12807. DOI: 10.1029/2012GL052261.

Plumb, R. A. (1989). “On the Seasonal Cycle of Stratospheric Planetary Waves”. In: *Pure and Applied Geophysics* 130(2/3), 233 –242. DOI: 10.1007/BF00874457.

Polvani, L. M. and P. J. Kushner (2002). “Tropospheric response to stratospheric perturbations in a relatively simple general circulation model”. In: *Geophysical Research Letters* 29(7), 18–1–18–4. DOI: 10.1029/2001GL014284.

Potts, J. M. (2012). “Basic concepts”. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Ed. by I. T. Jolliffe and D. B. Stephenson. 2nd. University of Exeter, UK: John Wiley and Sons, Ltd. Chap. 2.

Prodhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra (2016). “Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe”. In: *Climate Dynamics* 47(3), 919 –935. DOI: 10.1007/s00382-015-2879-4.

Quandt, L.-A., J. H. Keller, O. Martius, and S. C. Jones (2016). “Forecast Variability of the Blocking System over Russia in Summer 2010 and Its Impact on Surface Conditions”. In: *Weather and Forecasting* 32, 61 –82. DOI: 10.1175/waf-d-16-0065.1.

Quesada, B., R. Vautard, P. Yiou, M. Hirschi, and S. I. Seneviratne (2012). "Asymmetric European summer heat predictability from wet and dry southern winters and springs". In: *Nature Climate Change* 2(10), 736 –741. DOI: 10.1038/nclimate1536.

Ralph, F. M., P. J. Neiman, G. A. Wick, S. I. Gutman, M. D. Dettinger, D. R. Cayan, and A. B. White (2006). "Flooding on California's Russian River: Role of atmospheric rivers". In: *Geophysical Research Letters* 33(13), 3 –7. DOI: 10.1029/2006GL026689.

Richardson, D. S. (2001). "Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size". In: *Quarterly Journal of the Royal Meteorological Society* 127(577), 2473 –2489. DOI: 10.1002/qj.49712757715.

— (2012). "Economic value and skill". In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Ed. by I. T. Jolliffe and D. B. Stephenson. 2nd. University of Exeter, UK: John Wiley and Sons, Ltd. Chap. 9.

Robertson, A. W. and F. Vitart, eds. (2019). *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. 1st. Elsevier.

Robertson, A. W., F. Vitart, and S. J. Camargo (2020). "Subseasonal to Seasonal Prediction of Weather to Climate with Application to Tropical Cyclones". In: *Journal of Geophysical Research: Atmospheres* 125(6). DOI: 10.1029/2018JD029375.

Röthlisberger, M., S. Pfahl, and O. Martius (2016). "Regional-scale jet waviness modulates the occurrence of midlatitude weather extremes". In: *Geophysical Research Letters* 43, 10,989 –10,997. DOI: 10.1002/2016GL070944.

Scaife, A. A., J. R. Knight, G. K. Vallis, and C. K. Folland (2005). "A stratospheric influence on the winter NAO and North Atlantic surface climate". In: *Geophysical Research Letters* 32(18). DOI: 10.1029/2005GL023226.

Schaller, N., J. Sillmann, J. Anstey, E. M. Fischer, C. M. Grams, and S. Russo (2018). "Influence of blocking on Northern European and Western Russian heatwaves in large climate model ensembles". In: *Environmental Research Letters* 13, 054015. DOI: 10.1088/1748-9326/aaba55.

Schär, C. and G. Jendritzky (2004). "Hot news from summer 2003". In: *Nature* 432(7017), 559 –560. DOI: 10.1038/432559a.

Schär, C., P. L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. A. Liniger, and C. Appenzeller (2004). "The role of increasing temperature variability in European summer heatwaves". In: *Nature* 427(6972), 332 –336. DOI: 10.1038/nature02300.

Scher, S. and G. Messori (2019). "How Global Warming Changes the Difficulty of Synoptic Weather Forecasting". In: *Geophysical Research Letters* 46. DOI: 10.1029/2018GL081856.

Schiermeier, Q. (2018). "Climate as culprit". In: *Nature* 560, 20 –22. DOI: 10.1038/d41586-018-05849-9.

Screen, J. (2014). "Arctic amplification decreases temperature variance in northern mid- to high-latitudes". In: *Nature Climate Change* 4, 577—582. DOI: 10.1038/nclimate2268.

Screen, J. A. and I. Simmonds (2010). "The central role of diminishing sea ice in recent Arctic temperature amplification". In: *Nature* 464(7293), 1334 –1337. DOI: 10.1038/nature09051.

Seneviratne, S. I., D. Lüthi, M. Litschi, and C. Schär (2006). "Land-atmosphere coupling and climate change in Europe". In: *Letters to Nature* 443, 205 –209. DOI: 10.1038/nature05095.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling (2010). "Investigating soil moisture-climate interactions in a changing climate: A review". In: *Earth-Science Reviews* 99, 125 –161. DOI: 10.1016/j.earscirev.2010.02.004.

Shukla, J., J. Anderson, D. Baumhefner, C. Brankovic, Y. Chang, E. Kalnay, L. Marx, T. Palmer, D. Paolino, J. Ploshay, S. Schubert, D. Straus, M. Suarez, and J. Tribbia (2000). "Dynamical Seasonal Prediction". In: *Bulletin of the American Meteorological Society* 81(11), 2593 –2606. DOI: 0.1175/1520-0477(2000)081<2593:DSP>2.3.CO;2.

Sippel, S., F. E. L. Otto, M. Flach, and G. J. Van Oldenborgh (2016). "The Role of Anthropogenic Warming in 2015 Central Euopean Heat Waves, In: Explaining Extremes of 2015 from a Climate Perspective". In: *Bulletin of the American Meteorological Society* 97, S51 –S55. DOI: 10.1175/BAMS-D-16-0149.

Slingo, J. and T. Palmer (2011). "Uncertainty in weather and climate prediction". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 369(1956), 4751 –4767. DOI: 10.1098/rsta.2011.0161.

Sobolowski, S., G. Gong, and M. Ting (2010). "Modeled Climate State and Dynamic Responses to Anomalous North American Snow Cover". In: *Journal of Climate* 23(3), 785 –799. DOI: 10.1175/2009JCLI3219.1.

Sousa, P. M., R. M. Trigo, D. Barriopedro, P. M. Soares, and J. A. Santos (2018). "European temperature responses to blocking and ridge regional patterns". In: *Climate Dynamics* 50(1), 457 –477. DOI: 10.1007/s00382-017-3620-2.

Specq, D., L. Batté, M. Déqué, and C. Ardilouze (2020). "Multimodel Forecasting of Precipitation at Subseasonal Timescales Over the Southwest Tropical Pacific". In: *Earth and Space Science* 7(9), e2019EA001003. DOI: https://doi.org/10.1029/2019EA001003.

Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher (2017). "Review of Tropical-Extratropical Teleconnections on Intraseasonal Time Scales". In: *Reviews of Geophysics*, 902 –937. DOI: 10.1002/2016RG000538.

Sterk, A. E., M. P. Holland, P. Rabassa, H. W. Broer, and R. Vitolo (2012). "Predictability of extreme values in geophysical models". In: *Nonlinear Processes in Geophysics* 19(5), 529 –539. DOI: 10.5194/npg-19-529-2012.

Stott, P. A., D. A. Stone, and M. R. Allen (2004). "Human Contribution to the European Heatwave of 2003". In: *Letters to Nature* 432, 610 –614. DOI: 10.1029/2001JB001029.

Stott, P. A., N. Christidis, F. E. Otto, Y. Sun, J. P. Vanderlinden, G. J. van Oldenborgh, R. Vautard, H. von Storch, P. Walton, P. Yiou, and F. W. Zwiers (2015). "Attribution of extreme weather and climate-related events". In: *Wiley Interdisciplinary Reviews of Climate Change* 7, 23 –41. DOI: 10.1002/wcc.380.

Teuling, A. J., M. Hirschi, A. Ohmura, M. Wild, M. Reichstein, P. Ciais, N. Buchmann, C. Ammann, L. Montagnani, A. D. Richardson, G. Wohlfahrt, and S. I. Seneviratne (2009).

"A regional perspective on trends in continental evaporation". In: *Geophysical Research Letters* 36, L02404. DOI: 10.1029/2008GL036584.

Thomas, J. A., A. A. Berg, and W. J. Merryfield (2016). "Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring". In: *Climate Dynamics* 47(1-2), 49 –65. DOI: 10.1007/s00382-015-2821-9.

Tibaldi, S. and F. Molteni (1990). "On the operational predictability of blocking". In: *Tellus* 42A, 343 –365. DOI: 10.1034/j.1600-0870.1990.t01-2-00003.x.

Toth, Z. and R. Buizza (2019). "Weather Forecasting: What Sets the Forecast Skill Horizon?" In: *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Ed. by A. Robertson and F. Vitart. 1st. Elsevier. Chap. Chapter 2, 17 –45.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N. C. Lau, and C. Ropelewski (1998). "Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures". In: *Journal of Geophysical Research-Oceans* 103(C7), 14291 –14324. DOI: 10.1029/97jc01444.

Trigo, R. M., R. García-Herrera, J. Díaz, I. F. Trigo, and M. A. Valente (2005). "How exceptional was the early August 2003 heatwave in France?" In: *Geophysical Research Letters* 32, L10701. DOI: 10.1029/2005GL022410.

Tripathi, O. P., A. Charlton-Perez, M. Sigmond, and F. Vitart (2015). "Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions". In: *Environmental Research Letters* 10(10), 104007. DOI: 10.1088/1748-9326/10/10/104007.

van Straaten, C., K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits (2020). "The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures". In: *Quarterly Journal of the Royal Meteorological Society* 146, 2654—2670. DOI: 10.1002/qj.3810.

Ventrice, M. J., M. C. Wheeler, H. H. Hendon, C. J. Schreck, C. D. Thorncroft, and G. N. Kiladis (2013). "A Modified Multivariate Madden–Julian Oscillation Index Using Velocity Potential". In: *Monthly Weather Review* 141(12), 4197 –4210. DOI: 10.1175/MWR-D-12-00327.1.

Vigaud, N., A. W. Robertson, and M. K. Tippett (2017). "Multimodel Ensembling of Subseasonal Precipitation Forecasts over North America". In: *Monthly Weather Review* 145(10), 3913 –3928. DOI: 10.1175/MWR-D-17-0092.1.

Vitart, F. (2014). "Evolution of ECMWF sub-seasonal forecast skill scores". In: *Quarterly Journal of the Royal Meteorological Society* 140, 1889 –1899. DOI: 10.1002/qj.2256.

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A. W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. J. Won, H. Xiao, R. Zaripov, and L. Zhang (2017). "The subseasonal to seasonal (S2S) prediction project database". In: *Bulletin of the American Meteorological Society* 98(1), 163 –173. DOI: 10.1175/BAMS-D-16-0017.1.

Vitart, F., A. W. Robertson, and S2S Steering Group (2015). “Sub-seasonal to seasonal prediction: Linking weather and climate”. In: *Seamless Prediction of the Earth System: From Minutes to Months*. Ed. by G. Brunet, S. Jones, and P. M. Ruti. Vol. WMO-1156. World Meteorological Organization, 385 –401.

Vitart, F. and M. Balmaseda (2018). “Impact of sea surface temperature biases on extended-range forecasts”. In: *ECMWF Technical Memorandum* 830, 19. URL: `https://www.ecmwf.int/node/18659`.

Vitart, F., G. Balsamo, J.-R. Bidlot, S. Lang, I. Tsonevsky, D. Richardson, and M. Balmaseda (2019). “ERA5 reanalysis used to initialise re-forecasts”. In: *ECMWF Newsletter* (No. 161), 26 –31. DOI: `10.21957/g71fv083lm`.

Vitart, F. (2004). “Monthly Forecasting at ECMWF”. In: *Monthly Weather Review* 132(12), 2761 –2779. DOI: `10.1175/MWR2826.1`.

Vitart, F. and F. Molteni (2010). “Simulation of the Madden– Julian Oscillation and its teleconnections in the ECMWF forecast system”. In: *Quarterly Journal of the Royal Meteorological Society* 136(649), 842–855. DOI: `10.1002/qj.623`.

Wallace, J. M. and D. S. Gutzler (1981). “Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter”. In: *Monthly Weather Review* 109(4), 784 –812. DOI: `10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2`.

Wanders, N. and E. F. Wood (2016). “Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations”. In: *Environmental Research Letters* 11(9), 094007. DOI: `10.1088/1748-9326/11/9/094007`.

Weigel, A. (2012). “Ensemble forecasts”. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Ed. by I. Joliffe and D. Stephenson. 2nd. John Wiley & Sons, Ltd. Chap. Chapter 8, 141 –166.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008a). “Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?” In: *Quarterly Journal of the Royal Meteorological Society* 134, 241 –260. DOI: `10.1002/qj.210`.

Weigel, A. P., D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller (2008b). “Probabilistic verification of monthly temperature forecasts”. In: *Monthly Weather Review* 136(12), 5162 –5182. DOI: `10.1175/2008MWR2551.1`.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2007). “The discrete Brier and ranked probability skill scores”. In: *Monthly Weather Review* 135(1), 118 –124. DOI: `10.1175/MWR3280.1`.

Weisheimer, A. and T. N. Palmer (2014). “On the reliability of seasonal climate forecasts”. In: *Journal of the Royal Society Interface* 11, 20131162. DOI: `10.1098/rsif.2013.1162`.

Weisheimer, A., F. J. Doblas-Reyes, T. Jung, and T. N. Palmer (2011). “On the predictability of the extreme summer 2003 over Europe”. In: *Geophysical Research Letters* 38, L05704. DOI: `10.1029/2010GL046455`.

Wheeler, M. C. and H. H. Hendon (2004). “An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction”. In: *Monthly Weather Review* 132(8), 1917 –1932. DOI: `10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2`.

White, C. J., H. Carlsen, A. W. Robertson, R. J. Klein, J. K. Lazo, A. Kumar, F. Vitart, E. Coughlan de Perez, A. J. Ray, V. Murray, S. Bharwani, D. MacLeod, R. James, L. Fleming, A. P. Morse, B. Eggen, R. Graham, E. Kjellström, E. Becker, K. V. Pegion, N. J. Holbrook, D. McEvoy, M. Depledge, S. Perkins-Kirkpatrick, T. J. Brown, R. Street, L. Jones, T. A. Remenyi, I. Hodgson-Johnston, C. Buontempo, R. Lamb, H. Meinke, B. Arheimer, and S. E. Zebiak (2017). "Potential applications of subseasonal-to-seasonal (S2S) predictions". In: *Meteorological Applications* 24(3), 315 –325. DOI: `10.1002/met.1654`.

Wilks, D. S. (2019a). "Forecast Verification". In: *Statistical Methods in the Atmospheric Sciences*. 4th. Elsevier. Chap. 9, 369 –483.

— (2019b). "Frequentist Statistical Inference". In: *Statistical Methods in the Atmospheric Sciences*. 4th. Elsevier. Chap. 5, 143 –207.

Wilks, D. (2011). "Forecast Verification". In: *Statistical Methods in the Atmospheric Sciences*. Ed. by D. S. Wilks. 3rd ed. Vol. 100. International Geophysics. Academic Press. Chap. 8, 301–394. DOI: `10.1016/B978-0-12-385022-5.00008-7`.

Woollings, T. and M. Blackburn (2012). "The North Atlantic Jet Stream under Climate Change and Its Relation to the NAO and EA Patterns". In: *Journal of Climate* 25(3), 886 –902. DOI: `10.1175/JCLI-D-11-00087.1`.

Woollings, T., A. Hannachi, and B. Hoskins (2010). "Variability of the North Atlantic eddy-driven jet stream". In: *Quarterly Journal of the Royal Meteorological Sociecty* 136, 856 — 868. DOI: `10.1002/qj.625`.

Wulff, C. O., R. J. Greatbatch, D. I. V. Domeisen, G. Gollan, and F. Hansen (2017). "Tropical Forcing of the Summer East Atlantic Pattern". In: *Geophysical Research Letters* 44, 11,166 –11,173. DOI: `10.1002/2017GL075493`.

Wulff, C. O. and D. I. V. Domeisen (2019). "Higher Subseasonal Predictability of Extreme Hot European Summer Temperatures as Compared to Average Summers". In: *Geophysical Research Letters* 46(20), 11520 –11529. DOI: `10.1029/2019GL084314`.

Wulff, C., F. Vitart, and D. Domeisen (2021). "Trends inflate subseasonal temperature prediction skill". Under review in *Quarterly Journal of the Royal Meteorological Society*.

Xiang, B., S.-J. Lin, M. Zhao, N. C. Johnson, X. Yang, and X. Jiang (2019). "Subseasonal Week 3–5 Surface Air Temperature Prediction During Boreal Wintertime in a GFDL Model". In: *Geophysical Research Letters* 46(1), 416 –425. DOI: `10.1029/2018GL081314`.

Xiang, B., Y. Q. Sun, J.-H. Chen, N. C. Johnson, and X. Jiang (2020). "Subseasonal Prediction of Land Cold Extremes in Boreal Wintertime". In: *Journal of Geophysical Research: Atmospheres* 125(13), e2020JD032670. DOI: `https://doi.org/10.1029/2020JD032670`.

Xie, Y.-B., S.-J. Chen, I.-L. Zhang, and Y.-L. Hung (1963). "A preliminarily statistic and synoptic study about the basic currents over southeastern Asia and the initiation of typhoon (in Chinese)". In: *Acta Meteorologica Sinica* 33, 206 –217.

Yadav, P. and D. M. Straus (2017). "Circulation Response to Fast and Slow MJO Episodes". In: *Monthly Weather Review* 145(5), 1577 –1596. DOI: `10.1175/MWR-D-16-0352.1`.

Zhang, C. (2005). "Madden-Julian Oscillation". In: *Reviews of Geophysics* 43(2). DOI: `10.1029/2004RG000158`.

Acknowledgements

The completion of this thesis would not have been possible without the luck of receiving support from many people whose assistance I would like to acknowledge here.

First of all, I want to thank Daniela Domeisen for giving me the chance to do a PhD in the Atmospheric Predictability group and for being a great supervisor and mentor to me during these last four years. Daniela, I really appreciate the freedom you gave me in doing my research and that at the same time, you were always available to provide feedback and a sense of direction when I (thought I) had lost it. Thanks for showing confidence in my work.

I am very grateful to Antje Weisheimer and Christof Appenzeller for agreeing to be co-examiners for my thesis.

I also want to thank the whole Atmospheric Predictability group for being such a fun, diverse mix of personalities. I wish it would have been possible to spend more time with you all during this last year. Special thanks go to Bernat and Jake, whom I had the honor of sharing an office with for most of my PhD. Thanks to you guys it never got boring at work and I am grateful for the many interesting and intense discussions we had.

I am also grateful to many people that work(ed) on O and P floor of the IAC, especially Rob, Fabiola, Annika, Julie, Pavle, Gesa, Colin, Steffen, Jörg, Kristian, and Fabian. You made the time here and specifically the beginning of my PhD so much fun and I really appreciate that you took me in like family when I started here in 2017.

I would like to acknowledge the administrative and technical staff at the IAC, specifically Urs, Mathias, Eva, Bianca, Peter and Hans-Heini for doing an impressive job at making sure everything is running smoothly at the institute and for always being patient and ready to help.

I could not have gotten to this point without some activities outside of my PhD work that kept me sane during the most intense periods. Thanks to Davide, Etienne, Francesca, Hannes and Pierre-Shaq for many good moments on and off the basketball court. A big thank you goes to Philip for sharing his apartment, his great taste in music and many fun moments with me. I am also grateful to Eike and Laurène for being the kind people that they are.

I also want to thank my friends outside of Zurich that supported me remotely and on many visits and kept in touch even when I was (thinking I was) too busy to reach out. Thanks to Jan and his family, Jakob, Katherine, Elias and Bo, I am very lucky to have you.

Special thanks go to Iselin. Thank you for cheering me up, being my (home) office mate for the last year, editing parts of this thesis, having confidence in me and supporting me every day with your optimistic spirit.

Last but not least I would like to thank my family – Nele, Julietta und Eckhard, danke, dass ihr mir helft, wo immer ihr könnt und mich bedingungslos in jeder meiner Entscheidungen unterstützt, auch wenn ihr euch vielleicht an der ein oder anderen Stelle eine andere gewünscht hättet.