Machine Learning to improve and understand sub-seasonal forecasts of European temperature

van Straaten, Joachim Wilhelmus

2023

DOI (link to publisher)
10.5463/thesis.261

document version
Publisher's PDF, also known as Version of record

VRIJE UNIVERSITEIT

Machine Learning to improve and understand sub-seasonal forecasts of European temperature

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op maandag 20 november 2023 om 9.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Joachim Wilhelmus van Straaten

geboren te Nijmegen

promotoren:	prof.dr. B.J.J.M. van den Hurk prof.dr. D. Coumou
copromotoren:	dr. M.J. Schmeits dr. K. Whan
promotiecommissie:	prof.dr. P.J. Ward prof.dr. M. van Aalst dr. R.H. White dr. F. Vitart dr. M. Scheuerer

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE). The research was carried out at the Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, in close collaboration with the Royal Netherlands Meteorological Institute (KNMI), de Bilt, and supported by the Netherlands Organization for Scientific Research (NWO) in the form of project grant ALWOP.395.

Any model is at best a useful fiction

George Box
Statistical Control (1997) p.6

Initial states have later states
that depend on my technology
and later states have target states
until I forecast climatology

Poem based on 'Big whirls' by Lewis Fry Richardson
Weather Prediction by Numerical Process (1922) p.66

Cover map: Wikimedia Commons
Cover design and poem: Chiem van Straaten
Printed by: Ridderprint
DOI: http://doi.org/10.5463/thesis.261

Contents

Summary

Europe faces impactful droughts, heatwaves, and other extreme weather events every year. Ideally, these events would be predicted in advance so that society can take action and prevent the worst of the impacts. We desire long and skillful forecast windows, but prediction skill also decreases with the lead time with which an event is forecast. Sub-seasonal forecasts are issued two to six weeks before an event. They cannot predict the exact state of daily weather because rapidly growing disturbances and chaotic transitions dominate the evolution of weather, affecting the mid-latitudes in particular.

Occasionally however, there is a tendency towards a certain weather type. While invisible at individual locations or days, such tendencies are discernible at larger, aggregated scales. How many locations or days that scale involves varies, but you can think of a month with an above-normal number of hot days. In such a period, evolution of weather is partly governed by slow atmospheric variations, or by interactions between ocean, atmosphere and land. These processes shift the expected occurrence of a weather type (in this case toward the 'hot' weather type), which can be expressed with a probability forecast.

A conventional method to represent weather processes and make forecasts is with physics-based numerical models. Unfortunately, but inherently, their representations are imperfect, leading them to accumulate errors that can overwhelm the subtle tendencies that are the basis of sub-seasonal forecasts. Better forecasts could be possible when flawed representations of essential governing processes get corrected.

It is challenging for humans to determine which governing processes are essential. Interactions between them are many and shifts toward a certain weather type occur only occasionally. Machine Learning (ML) tools can sometimes more capably detect patterns to achieve a certain task. This thesis explores their use in improving sub-seasonal forecasts, particularly summertime forecasts of European temperature. I explore the

task of direct statistical forecasting, and of correcting numerical forecasts (i.e. statistical post-processing).

In Chapter 2 I begin by evaluating existing numerical forecasts, made by the European Centre for Medium-Range Weather Forecasts (ECMWF). With a clustering algorithm I aggregate individual forecasts, and evaluate whether their average temperature is predictable with sub-seasonal lead times. It confirms the intuition that predictable tendencies are most discernible at large spatial and temporal scales. However, the scales should match the process, as sub-seasonal forecast skill due to localized processes gets lowered by excessive spatial aggregation.

Importantly, skill is higher when shortcomings of the ECMWF model get corrected. Even a simple post-processing that only corrects the most pervasive errors (skipping the weather-dependent systematic errors) already extends the period that can be skillfully forecast by a few days.

To eventually correct more shortcomings, I want to detect and understand the essential governing processes. In Chapter 3 I therefore train an ML method to directly predict the probability that monthly Western European temperature will be above normal, which it can do two weeks ahead of the event. It achieves the task by detecting governing patterns in oceanic, land-surface and atmospheric states, and by representing their interactions. Just like with numerical forecasts, the input consists of information at initialization time, i.e. ERA5 data. Which exact patterns the ML forecast uses, is revealed with eXplainable AI (XAI) tools. Predominantly, the learned representation of essential processes is based on oceanic and land-surface patterns. This agrees with existing physical knowledge and with numerical model experiments, which increases trust in the ML method.

However, the ML method also uses patterns that deviate from the ones we know. These are either coincidental statistical connections, or represent essential processes that are prone to numerical errors, and have therefore remained undetected. In Chapter 4 I show that the latter is the case in sub-seasonal ECMWF forecasts of monthly Western European temperatures. I introduce an Artificial Neural Network architecture to correct systematic errors occurring in specific conditions only. The ML-detected initial patterns from Chapter 3 form part of the inputs, and again I assess their role with XAI. This reveals that pronounced correctable errors relate to conditions in tropical west Pacific sea surface temperatures (SSTs). The Pacific SST pattern can generate a delayed tendency in Euro-Atlantic weather that changes the probability of above normal temperatures, which is wrongly estimated by the ECMWF model.

Skill is gained by the correction, which confirms the known importance (and difficulty) of representing tropical to extra-tropical teleconnections in sub-seasonal forecasts.

In Chapter 5 I further diagnose whether the statistically implied connection is real. In that case a physical pathway should exist between the tropical west Pacific and Europe. First I show that the west-Pacific pattern is comprised of large-scale contrasts in SST. According to existing knowledge on Pacific SST variability, this captures a combination of long-term west Pacific warming (a response to anthropogenic climate change) and El Niño Southern Oscillation-like variability. These two can combine to tropical conditions that generate atmospheric waves. These waves can travel and partly determine the distribution of hot and cold airmasses over Europe, three to six weeks later. I show that the occasional connection has become more prominent since 1980, and induces sub-seasonal predictability since. It also partly explains the above-average increase in summertime European heat extremes in the last decades.

Overall this thesis traverses multiple applications of ML in sub-seasonal forecasting. The first application is detection: of predictable tendencies, and of preceding patterns in initial states. Second is direct forecasting: by learning a statistical representation of interacting processes. Third is the correction of numerical ECMWF forecasts. As I investigate the ML models with XAI, (tele)connections are gradually moved from 'statistically implied' to 'physically understood', and add to existing knowledge along the way. As forecast skill increases as well, I can conclude that ML helps to improve sub-seasonal forecasts in an interpretable manner.

Samenvatting

Europa kampt jaarlijks met impactvolle droogtes, hittegolven en andere extreme weersverschijnselen. Idealiter worden zulke gebeurtenissen enkele weken vooruit voorspeld zodat de samenleving vroeg in actie kan komen. Echter, hoe vroeger een voorspelling wordt afgeven hoe groter ook de onzekerheid rondom de voorspelde gebeurtenis. Sub-seizoenale voorspellingen worden drie tot zes weken voor een gebeurtenis afgegeven. Op zo'n termijn is het onmogelijk om een dagelijks situatie exact te voorspellen, aangezien snel groeiende, chaotische verstoringen de evolutie van het weer domineren, vooral in de gematigde breedtegraden.

Soms treden er echter tendensen richting een bepaald weertype op. Bezien vanuit individuele locaties of dagen zijn zulke tendensen vaak onzichtbaar, maar op grotere ruimte- en tijd-schalen komen ze aan het licht. Hoeveel locaties of dagen zo'n groot-schalige verzameling behelst is variabel, maar denk bijvoorbeeld aan een maand met een bovengemiddeld aantal hete dagen. In zo'n periode wordt de evolutie van het weer deels bepaald door trage atmosferische oscillaties, of door interactie tussen oceaan, atmosfeer en land. Zulke processen verschuiven het verwachte optreden van een weertype, wat uitgedrukt kan worden met een kansverwachting (in dit geval verschuift het zwaartepunt naar het 'hete' weertype).

Een gebruikelijke methode om weersverwachtingen te maken is met natuurkundige, numerieke modellen. Helaas, maar ook onvermijdelijk, zijn de proces-representaties in deze modellen imperfect. Zulke tekortkomingen leiden ertoe dat fouten zodanig kunnen accumuleren dat subtiele tendensen overweldigd worden. Voor betere sub-seizoenale voorspellingen zouden proces-representaties verbeterd moeten worden.

Mensen kunnen echter moeilijk overzien welke processen essentieel zijn voor sub-seizoenale voorspellingen. Processen interacteren namelijk veelvuldig, en de resulterende verschuivingen in weertype treden slechts sporadisch op. *Machine Learning* (ML) methoden kunnen vaak beter patronen detecteren die gerelateerd zijn aan een bepaalde uitkomst. In

deze thesis gebruik ik ML voor het verbeteren van sub-seizoenale voorspellingen, in het bijzonder voorspellingen van zomerse temperatuur in Europa. ML wordt ingezet om ofwel directe statistische verwachtingen te maken, ofwel numerieke verwachtingen te corrigeren (ook wel 'statistische nabewerking' genoemd).

In Hoofdstuk 2 start ik met een evaluatie van bestaande numerieke verwachtingen gemaakt door het European Centre for Medium-Range Weather Forecasts (ECMWF). Met een cluster-algoritme verzamel ik individuele verwachtingen en evalueer ik of hun gemiddelde accuraat verwacht wordt op de sub-seizoenale termijn. Dit bevestigt de intuïtie dat grootschalige gemiddelden het best de voorspelbare tendensen vatten. De schaal moet echter wel afgestemd zijn op het proces. Sub-seizoenale voorspelbaarheid dankzij lokale processen neemt namelijk af wanneer de ruimtelijke aggregatie overmatig is.

Een belangrijke bevinding is dat de accuraatheid toeneemt wanneer tekortkomingen in het ECMWF model gecorrigeerd worden. Een simpele statistische nabewerking die alleen de meest hardnekkige fouten corrigeert (dus niet de weers-afhankelijke fouten) verlengt de accurate verwachtingstermijn met een paar dagen.

Om uiteindelijk meer tekortkomingen te kunnen corrigeren wijd ik Hoofdstuk 3 aan het detecteren en begrijpen van essentiële processen. Allereerst train ik een ML methode om direct de kans te voorspellen dat maandelijkse west-Europeese temperatuur de klimatologische mediaan overschrijdt, hetgeen de methode succesvol doet twee weken voor de gebeurtenis. Gelijk numerieke modellen heeft de ML methode slechts beschikking tot initiële informatie (i.e. ERA5). In deze data detecteert de methode meerdere oceanische, landelijke en atmosferische patronen, en representeert vervolgens hun interacties. Met *eXplainable AI* (XAI) methoden onthul ik dat de geleerde representatie voornamelijk stoelt op oceanische en landelijke patronen die overeenkomen met bestaande kennis en numerieke model-experimenten. Dit vergroot het vertrouwen in de ML methode.

De methode gebruikt echter ook onbekende patronen. Deze moeten ofwel statistische toevalligheden zijn, of betreffen essentiële processen die voorheen ongedetecteerd bleven omdat ze vatbaar zijn voor numerieke fouten. In Hoofdstuk 4 laat ik zien dat het laatste het geval is voor subseizoenale ECMWF verwachtingen van maandelijkse west-Europeese temperatuur. Ik introduceer een kunstmatig neuraal netwerk dat systematische, weers-afhankelijke fouten kan corrigeren. De gedetecteerde patronen uit Hoofdstuk 3 vormen een deel van de invoer, en opnieuw evalueer ik hun

relevantie met XAI. De resultaten laten zien dat belangrijke, corrigeerbare ECWMF fouten samenhangen met west-Pacifische omstandigheden ter hoogte van de evenaar. De zee-oppervlakte-temperaturen aldaar lijken met enige vertraging een tendens te genereren in het Euro-Atlantische weer. Het resultaat is een verschuiving in overschrijdings-kansen die incorrect door het ECMWF model wordt gesimuleerd. Dat de verwachting vervolgens verbetert door een correctie, suggereert het belang (en de moeilijkheid) van een goede representatie van teleconnecties tussen de tropen en gematigde breedtegraden.

In Hoofdstuk 5 onderzoek ik of de gesuggereerde statistische connectie ook een fysische weerslag heeft. In dat geval zou er een oorzakelijk pad moeten bestaan tussen de westelijke Stille Oceaan en Europa. Eerst laat ik zien dat het west-Pacifische patroon bestaat uit grootschalige opposities in zee-oppervlakte temperaturen. De huidige wetenschap vat dit op als een combinatie van west-Pacifische opwarming (een reactie op antropogene klimaatverandering) en *El Niño Southern Oscillation*-achtige variabiliteit. Tezamen scheppen deze twee factoren tropische omstandigheden die atmosferische golven genereren. Deze golven kunnen zich verplaatsen en beïnvloeden gedeeltelijk de verdeling van warme en koude lucht over Europa drie tot zes weken later. Deze sporadische connectie is opgekomen sinds 1980 en zorgt voor sub-seizoenale voorspelbaarheid sindsdien. Gedeeltelijk verklaart de connectie ook de bovengemiddelde toename in de laatste decennia van extreme Europeese zomerhitte.

Als geheel verkent deze thesis meerdere toepassingen van ML in subseizoenale verwachtingen. De eerste toepassing is detectie: van voorspelbare tendensen en van voorafgaande patronen in de initiële informatie. De tweede is directe voorspelling: door een statistische representatie van interacterende processen te leren. De derde is correctie van numerieke ECMWF verwachtingen. Aangezien de toepassingen tevens gecombineerd worden met XAI, veranderen de gevonden (tele)connecties gaandeweg van 'statistisch gesuggereerd' naar 'fysisch begrepen', en vormen ze een aanvulling op bestaande kennis. Omdat de accuraatheid van de voorspellingen daarbij toeneemt, kan ik concluderen dat de toepassing van ML subseizoenale verwachtigen inzichtelijk verbetert.

Chapter 1

Introduction

1.1 Forecasting the weather

A weather forecast predicts how atmospheric properties like temperature, pressure, and moisture will vary in space and time. Suppose that it is two o'clock on a hot afternoon, and that a forecast gets made. For the next few minutes, the forecast could skillfully predict how a cloud blocks the sunshine on your balcony and gives you a chance to cool off. For the upcoming night, it could predict how the hot airmass will remain in the country. For the upcoming week, it could predict how all airmasses in the northern hemisphere midlatitudes will move more slowly than usual. For the upcoming year, it could predict the gradual change of seasons. For the upcoming decades, it could project how anthropogenic green-house-gas emissions affect the global climate.

Even for this last projection, which is on a wide-spread and long-lasting scale, a forecast still has to account for small-scale processes like the effect of a cloud on a sunny afternoon. This is because the climate system exists of strong interactions between processes operating at different spatial and temporal scales. The result is an intertwined continuum of variability that makes the climate system challenging to model (Frankignoul and Hasselmann, 1977; Hoskins, 2013; World Meteorological Organization, 2015; Lovejoy, 2015).

This challenge is particularly pronounced in the sub-seasonal part of the variability continuum, which forms the topic of this thesis. Sub-seasonal forecasts are roughly defined as forecasts of processes that are predictable beyond 2 weeks into the future. Historically, this subset of all processes fell in between 'weather'- and 'climate'-research and was there-

fore under-studied (Vitart and Robertson, 2019). But current interest is high, and fueled by the immense importance of early warnings (Merryfield et al., 2020; White et al., 2021). We know that taking action a few days in advance of a heatwave can reduce its impact (Casanueva et al., 2019). But even better preventative actions are possible with skillful forecasts more than two weeks in advance. Think of safeguarding crop harvests and preventing energy shortages (Coughlan de Perez et al., 2015; Grams et al., 2017; Guimaraes Nobre et al., 2019).

There are fundamental and practical reasons why only a subset of processes are predictable with lead times of more than two weeks. Fundamentally, the physical processes that make up the climate system, such as the principles of fluid dynamics on a rotating sphere, are deterministic. This means that exact solutions through time could be found when the initial state of the dynamical system is fully known. However, the principles are also nonlinear, which makes the solutions unstable with respect to small deviations in the initial state (Lorenz, 1963). As these deviations from perfect initial knowledge are inevitable (for instance due to irreducible measurement errors) any solution of the complex dynamical system will eventually diverge from the true solution. This so-called 'chaotic' property of the climate system affects certain processes more than others and places a fundamental limit on predictability.

Practically, predictability is further constrained by imperfections in our forecast models. The state-of-the-art in weather forecasting is formed by numerical weather prediction (NWP) models (Bauer et al., 2015). These are physics-based models that contain mathematical representations of the principles discussed above. However, their representations are approximate. They use spatial and temporal discretization, and statistically approximatate the bulk effects of unresolved small-scale processes (known as 'subgrid parametrizations') (Kalnay, 2003).

Unfortunately, errors from these imperfect representations and from imperfect knowledge of initial conditions accumulate. It is a property of the system that errors propagate from the small scales to the large scales (Lorenz, 1969; Vannitsem, 2017; Toth and Buizza, 2019). Eventually, forecast models thus reach a moment where the variable of interest, say average two-meter air temperature (t2m) in Germany, is completely de-railed by accumulating errors (see examples of 'forecast busts' in Magnusson, 2017). Though the growth rate of errors depends on the initial conditions and stability of the flow (Rodwell et al., 2018), we generally see that errors affect forecasts of individual clouds within minutes, forecasts of organized convection within hours, and forecasts of extra-tropical cy-

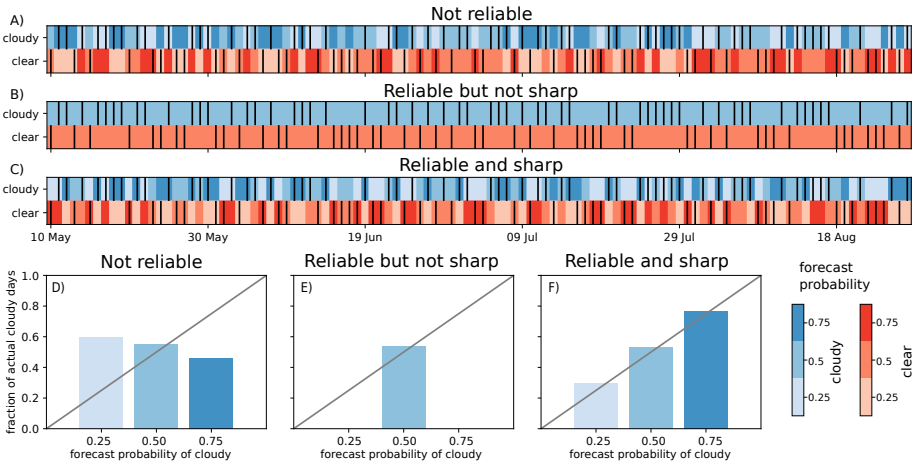


Figure 1.1: Synthetic example of three types of forecasting systems, each forecasting whether a day will be cloudy or clear (110 days are plotted). The daily synthetic observations are random: clear in fifty percent of cases, and cloudy in the other fifty percent (black vertical lines in panels A-C). The forecasting systems can only forecast probabilities of 0.25, 0.5 and 0.75 (light to dark colors), and for each day the probabilities of the cloudy and clear category (blue and red) must sum up to 1. Panels D-F are reliability diagrams. For each forecast probability of cloudiness (0.25,0.5,0.75) we count the number of observed cloudy days, and divide that by the total number of times that the forecast was issued. When this fraction of actual cloudy days is close to the 1:1 line, then the system is reliable and correctly expresses how likely it is to be correct and thus how likely it is to be wrong (e.g. only 25 percent of days for which a 75 percent probability of cloudiness was issued, will end up being clear). Sharpness concerns the ability of the forecast system to concentrate probability on one of both categories and thereby deviate from the generally expected fifty percent.

clones within days. Only the largest, hemispheric scales of motion remain predictable up to 2-3 weeks (Shukla, 1981; Privé and Errico, 2015; Buizza and Leutbecher, 2015; Vannitsem, 2017; Ying and Zhang, 2017; Zhang et al., 2019a; Toth and Buizza, 2019).

Since fundamental and practical unpredictability are so pervasive, forecasters have started to embrace the imperfection in their models (Palmer, 2017). Instead of just predicting an ultimately incorrect future state, they design their models to be 'reliable' or 'calibrated'. This means that the

model issues an uncertainty distribution that expresses the expected forecast error. In effect, the model issues a probability distribution of possible future weather states, which users can use to take action (Richardson, 2000; Palmer, 2017), as for reliable models, the true solution of the climate system will appear to be a random draw from the distribution (Gneiting et al., 2007; Wilks, 2011; Hopson, 2014; Buizza et al., 2018). This is illustrated by the synthetic forecasting problem in Figure 1.1, where only two of the three forecast systems issue reliable uncertainty distributions (panels B-C and E-F).

Each day the three systems provide a probability of it being cloudy and a probability of it being clear, which together sum up to one (Fig. 1.1). When a system issues a 75 percent probability of cloudiness, it implicitly says that there is a 25 percent chance that it is wrong about it being cloudy. Of the issued 75-percent forecasts, we thus expect in 75 percent of the cases a cloudy day to occur, and in 25 percent a clear day to occur. The first system does not correctly express its own errors, as verification of its 75-percent forecasts yields a set of days in which only 50 percent is cloudy (darkest blue Fig. 1.1D). The system is wrong about its wrongness, and thus unreliable.

Not wrong about its wrongness, but nonetheless wrong a lot of times, is the second system. It is reliable because it always issues probabilities of 50 percent (Fig. 1.1B), and is wrong in 50 percent of those cases. The difference with the third, also reliable system (Fig. 1.1C), is that system two has no concentration of probability or 'sharpness' in its uncertainty distributions. Given the reliability of the second and third system, sharpness can be seen as an expression of knowledge (Gneiting et al., 2007; Spiegelhalter, 2019). The second system has either no knowledge of the large-scale dynamical weather situation each day, or has no way to link it to the likelihood of cloudiness: it just issues the climatological probability each day. The third system on the other hand, can be thought of as one that does have knowledge of the initial conditions, and has a representation of how cloudiness is affected. Such knowledge and representations are always imperfect, but improved approximations generally result in more reliable, sharper uncertainty distributions and higher forecast skill (Toth et al., 2003; Gneiting et al., 2007).

An expression of uncertainty can be achieved in multiple ways. Statistical models often directly express the uncertainty associated with their approximations. Think of a linear regression: $Y = \alpha X + \epsilon$, with coefficient α that is the result of the least-squares method, and distribution ϵ to express unexplained variance (see Scher and Messori, 2021, for an example of

statistical models that do not do this). In NWP this is more involved, and requires Ensemble Predictions Systems (EPS) (Leutbecher and Palmer, 2008; Slingo and Palmer, 2011). Multiple NWP simulations (called 'ensemble members') are started from perturbed initial states. Jointly these perturbations express the uncertainty in our imperfect estimate of the initial state. Each of the simulations is also run with slightly altered (parameters of the) subgrid parametrizations (and possibly reduced precision) to express the uncertainty in the subgrid processes (Palmer, 2015; Leutbecher et al., 2017). After this, the members are treated as a sample of the forecast uncertainty distribution, each representing a possible solution of the dynamical system.

1.2 Subseasonal predictability

The goal for modelers and researchers is clear: Improve the representation of the climate system, such that the forecast models will forecast reliable uncertainty distributions that are as sharp as possible (Gneiting et al., 2007). Given comparable trustworthiness, it should not matter whether those models are statistical, numerical or a combination of the two (hybrid). For sub-seasonal forecasts the goal involves detecting and understanding those physical processes, that, with proper modeling and observation of their initial state (Merryfield et al., 2020), remain predictable beyond the first two weeks. Such processes are either internal atmospheric oscillations with long timescales (Baldwin et al., 2001; Ghil and Robertson, 2002; Stan and Krishnamurthy, 2019), or result from atmospheric interaction with the ocean and land surface (Shukla, 1998; Hoskins, 2013).

Conceptually, sub-seasonal processes can be understood as 'steering' the otherwise divergent atmospheric solutions towards a subset of possible states (Palmer, 1993). We illustrate this with an example of heavily cloudy and less cloudy days in Figure 1.2. Displayed are daily mean observations at a weather station in the Netherlands (De Bilt). We see that rapid (chaotic) alterations between heavily and less cloudy conditions are prevalent, and that isolated moments in time can be unpredictable. For instance: could a forecast made in June have foreseen the two heavily cloudy days on the 29th and 30th of July 2018? Probably not. As an aggregate however, atmospheric states can display tendencies, which in July 2018 is a tendency towards less cloudy skies (bright red, Fig. 1.2B). In fact the less cloudy skies related to persistent high pressure systems,

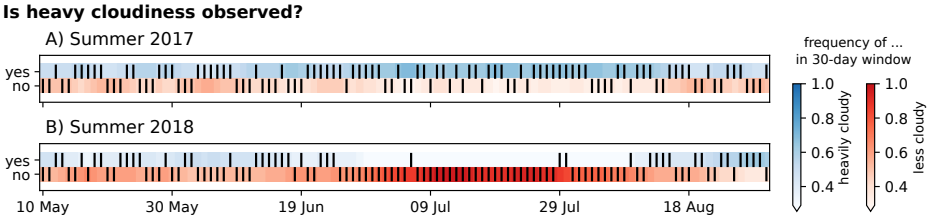


Figure 1.2: Illustration of sub-seasonal steering in daily atmospheric states. Plotted are daily measurements of cloudiness in De Bilt for two distinct summers (2017 and 2018). A day is classified as 'heavily cloudy' when cloudiness >= 7/8 [okta], and as 'less cloudy' otherwise (black vertical lines). Colored is the observed frequency of registering a heavily cloudy or less cloudy day, within a 30-day window surrounding each day.

which 'blocked' the usual westerly flow over Europe and resulted in a strong heatwave (Kueh and Lin, 2020; Kautz et al., 2021). A good model of such processes could have issued relatively sharp uncertainty distributions with most probability concentrated on the 'less cloudy' class. Note also that the aggregate tendency towards the 'less cloudy' class is absent in 2017 (Fig. 1.2A), and that sub-seasonal tendencies are thus state-dependent and only present when conditions are right (Palmer, 1993; Wilks and Vannitsem, 2018).

Another example of a state-dependent tendency is the Madden Julian Oscillation (MJO). This oscillation comprises an envelope of organized convection moving eastward along the equator, which can be active or inactive (Zhang, 2013). It potentially drives North-Atlantic atmospheric variability by generating a 5- to 15-day tendency towards a certain state (Cassou, 2008; Lin et al., 2009). Such a 'teleconnection' depends on the initial MJO activity and the initial state of the extra-tropical jet stream (Lin and Brunet, 2018).

From both examples we learn that sub-seasonal processes are intermittently active, and offer only conditional opportunities to sharpen the forecasts. These opportunities are referred to as 'windows of predictability' or 'forecasts of opportunity' (National Academies of Sciences, Engineering, and Medicine, 2016; Albers and Newman, 2019; Mariotti et al., 2020). They are also tied to specific scales of aggregate atmospheric variability, like the monthly tendency in July 2018 for Western Europe, and the 5- to 15-day North Atlantic tendency by MJO. Detecting a window of predictability and understanding the responsible sub-seasonal processes thus requires a characterization of 'low-frequency' variability. For the

mid-latitudes that generally happens in one of two ways.

The first way, coming from dynamical systems theory, is a characterization of large-scale 'regimes'. These are synoptic configurations that the atmosphere reverts to for extended periods of time (Vautard, 1990; Michelangeli et al., 1995; Grotjahn et al., 2016; Hannachi et al., 2017). A prominent configuration in the Euro-Atlantic region is one with high pressure in the south-west and low pressure in the north-west, termed the (Summer) North Atlantic Oscillation ((S)NAO), and relates to the strength and position of the jet stream (Wallace and Gutzler, 1981; Woollings, 2010; Folland et al., 2009; Bladé et al., 2012). Also important are the already mentioned 'blocked' regimes (Pfahl, 2014; Schaller et al., 2018; Sousa et al., 2018; Kautz et al., 2021). In blocked regimes the jet stream is diverted by a high pressure system, creating a region in which extreme surface temperatures can occur, like in July 2018 (Schneidereit et al., 2012; Brunner et al., 2017; Schaller et al., 2018).

The second way describes low-frequency variability in terms of atmospheric Rossby Waves (Platzman, 1968; Hoskins and Ambrizzi, 1993). Such waves consist of meridional undulations in the jet stream that encompass a succession of cyclonic (low pressure) and anti-cyclonic (high pressure) systems (Wirth et al., 2018; White et al., 2022). In between the 'standing' stationary waves, and fast 'transient' waves, are Rossby Wave Packets (RWPs). These are of sub-seasonal interest because they involve atmospheric tendencies over extended periods of time, either through recurrence (a repeated large amplitude wave, with the same phase over the same geographical location) or through quasi-stationarity (a wave that lingers in place) (Schubert et al., 2011; Kornhuber et al., 2017; Röthlisberger et al., 2019; Wolf et al., 2018). In the end, regime-based and wave-based characterizations are highly related, and can be seen as two sides of the same coin (Benedict et al., 2004; Michel and Rivière, 2011; Sousa et al., 2021).

It is beneficial that although the characterizations are simply descriptive, they can shed light on underlying processes. The first set of processes to drive RWPs and regimes, are tropical (Cassou et al., 2005; Cassou, 2008). Tropical deep convection and associated diabatic heating generate upper level divergence and vorticity anomalies that force Rossby Waves (Hoskins and Karoly, 1981; Sardeshmukh and Hoskins, 1988; Trenberth et al., 1998). Propagation into the westerly midlatitude flow commonly takes two weeks (Branstator, 2014; Stan et al., 2017). Often the deep convection is conditional on certain Sea Surface Temperature (SST) patterns, meaning that it is the coupling to ocean dynamics that provides

the potential sub-seasonal predictability (Liu and Alexander, 2007; Lee et al., 2019). Evidence for this has been found in diabatic heating in the tropical Pacific (Ding et al., 2011; O'Reilly et al., 2018), and in the Indian Summer Monsoon (Beverley et al., 2019; Di Capua et al., 2020a). Both lead to wave responses above Europe.

Second are atmosphere-ocean couplings in the extra-tropics (Black and Sutton, 2007; Della-Marta et al., 2007; Feudale and Shukla, 2011). With its large thermal capacity, the oceanic mixed layer can provide a memory-component to the atmospheric circulation. In the North Atlantic for instance, circulation in spring can create contrasts between relatively cool and warm SST regions, which in June and August can steer jet stream position and wave activity over the North Atlantic (Ossó et al., 2020; Wolf et al., 2020). Specifically, SST can reinforce existing RWPs and precede large amplitude wave events through diabatic heating (McKinnon et al., 2016; Vijverberg and Coumou, 2022; Kim and Lee, 2022). Influences have also been found for Arctic SSTs, which are closely coupled to sea ice concentrations (Kim et al., 2014; Kolstad and Årthun, 2018; Koenigk et al., 2016; Zhang et al., 2020).

Coupling to the land surface can also influence extra-tropical atmospheric variability. Snow cover can change the radiative energy balance at the surface, influencing quasi-stationary RWP amplitude remotely (Orsolini et al., 2013; Cohen et al., 2014; Hall et al., 2017; Henderson et al., 2018; Zhang et al., 2020). Also soil moisture, as an integrator of past hydro-meteorological conditions, can influence atmospheric circulation (Koster et al., 2010; van den Hurk et al., 2012; Prodhomme et al., 2016). Coupling particularly occurs at moments and locations where the partitioning between latent and sensible heating becomes limited by soil moisture content (Seneviratne et al., 2010). Effects can be local, like the exacerbation of heat extremes (Fischer et al., 2007; Miralles et al., 2019; Perkins, 2015), can be regionally displaced (Quesada et al., 2012; Schumacher et al., 2019), and can even reach beyond individual continents by amplifying RWPs (Teng and Branstator, 2019).

Lastly, low-frequency variability can originate from processes internal to the atmosphere. One example is the quasi-biennual oscillation of stratospheric winds over the equator (Baldwin et al., 2001). Another is the reflection and trapping of Rossby Waves by strong midlatitude jet streams (Hoskins and Karoly, 1981; Wirth et al., 2018; White et al., 2022). Under certain circumstances this allows transient waves to resonate circumglobally and assume a quasi-stationary character (Petoukhov et al., 2013; Kornhuber et al., 2017).

1.3 Role of Machine Learning methods

Research on the processes described above culminates in improved NWP representations of important couplings (Bauer et al., 2015; Merryfield et al., 2020). Improved representations of for instance soil moisture (Bunzel et al., 2018), and of ocean-atmosphere coupling during MJO (Henderson et al., 2017; Kim et al., 2018), now enable NWP models to issue more skillful forecasts in certain windows of predictability.

For other windows of predictability however, it remains hard to correctly model all driving processes. Approximations remain inherent in NWP and lead to hard-to-resolve biases, for example in Rossby Wave amplitude and the frequency of blocked flows (Matsueda, 2009; Gray et al., 2014; He et al., 2019; Quinting and Vitart, 2019). Conditional on some sub-seasonal steering, like the observed tendency towards blocked flow in July 2018 (Fig. 1.2B), such biases make forecast distributions less sharp and thus less informative than they could have been, and might even make them unreliable.

Instead of directly tinkering with the many underlying physical representations, a pragmatic approach has been to correct systematic NWP errors after forecasts have been made, via so-called 'statistical post-processing' (Glahn and Lowry, 1972; Vannitsem et al., 2018; Haupt et al., 2021; Vannitsem et al., 2021). In that approach a dataset of past forecasts and observations is used to statistically fit the forecasts to observations. This commonly involves fitting simple parametric models to correct the bias and uncertainty distributions of EPS systems, such that reliability is achieved (Gneiting et al., 2005; Wilks and Hamill, 2007; Gebetsberger et al., 2018). Parameters of an assumed distribution are modeled as linear functions of EPS mean, spread, and related quantities. A systematic bias in t2m is for instance corrected by forecasting a Gaussian uncertainty distribution, with a linear regression that makes its mean depend on the EPS mean and a constant value. The post-processed uncertainty distribution is thus a hybrid coalescence of dynamical computations in the NWP model and a statistical, bulk-wise correction afterwards. This approach is actively used to increase the reliability of sub-seasonal EPS systems (Vigaud et al., 2017; Ferrone et al., 2017; Monhart et al., 2018; Van Schaeybroeck and Vannitsem, 2018; Heinrich et al., 2021; Graham et al., 2022).

However, a bulk-wise correction determined on the majority of forecasts can be unsuited for certain weather conditions. Often, corrections are fitted per location and per season, but for sub-seasonal forecasts that

does not seem enough. Different windows of predictability are governed by different sub-seasonal processes, which will have different biases in the NWP model. Hence the NWP biases can be highly conditional. For instance, certain errors in the background flow are only relevant in situations when tropical diabatic heating generates Rossby Waves which fail to propagate in an erroneous background flow (Stan et al., 2017; O'Reilly et al., 2018). In general, but even more so for sub-seasonal forecasts, post-processing is better when it adapts to prevalent conditions (Allen et al., 2019; Strazzo et al., 2019; Specq and Batté, 2020; Vannitsem et al., 2021; Hewson and Pillosu, 2021). Such flexibility is attained by incorporating additional predictor variables that, either by themselves, or through interactions between them, represent the distinct conditions.

The capacity required for conditional post-processing is provided by Machine Learning (ML) methods (for an overview see Schulz and Lerch, 2022). With its conditional corrections, ML-based statistical post-procssing can produce reliable distribution that are sharper and more skillful than those from linear methods (Rasp and Lerch, 2018; van Straaten et al., 2018; Veldkamp et al., 2021; Hewson and Pillosu, 2021; Dai and Hemri, 2021; Grönquist et al., 2021; Kirkwood et al., 2021; Fan et al., 2021; Hewson and Pillosu, 2021; Schulz and Lerch, 2022; Li et al., 2022). Despite pioneering work by Scheuerer et al. (2020), ML-based post-processing of sub-seasonal forecasts is still rare.

Another advantage of applying ML is the possibility that in the multi-scale, multivariate coupled system, new windows of predictability can be found. ML methods could detect important sub-seasonal processes that are poorly represented in current NWP models (Cohen et al., 2018). For decades, statistical detection has already been practiced with multivariate methods like maximum covariance analysis (see overviews in Tippett et al., 2008; Wilks, 2014). These can for instance relate tropical heating or Arctic sea ice concentration to midlatitude atmospheric variability (Ding et al., 2011; Garcia-Serrano and Frankignoul, 2014; Di Capua et al., 2020b). Though the detected modes of covariability represent a predictable steering towards a subset of atmospheric states (and can thus be used for forecasting), they remain patterns that characterize all data at once. Those methods cannot detect highly conditional variability relevant to specific windows of predictability.

Besides detecting sub-seasonal variabilities, ML methods can also be used to disentangle the many dynamic dependencies between responsible processes (Saha et al., 2016; Vannitsem and Liang, 2022; Runge et al., 2015). Deep neural networks can in principle perform both learning tasks

at once, as their hierarchical stacks of universal function approximators enable them to construct high-level process representations, and model interactions between those (Reichstein et al., 2019; Schultz et al., 2021). Indeed, deep neural networks are increasingly succesful in predicting some aspects of the weather a few days in advance (Weyn et al., 2020; Rasp and Thuerey, 2021; Keisler, 2022; Lam et al., 2022). But sub-seasonal predictions remain a challenge (He et al., 2021; Mayer and Barnes, 2021; Weyn et al., 2021; Jacques-Dumas et al., 2022; Miloshevich et al., 2023; Lopez-Gomez et al., 2023). Part of the difficulty lies in detecting the scale at which aggregate atmospheric variability is affected, when expert-crafted characterizations like weather regimes and RWPs are not pre-supplied. On top of that, each occurring window of predictability is always slightly altered by small-scale unpredictable processes, which results in noisy data that ML methods find hard to navigate, given the amount available for training (Reichstein et al., 2019; He et al., 2021). This explains why simple statistical models are still widely used in sub-seasonal forecasting. With expert-crafted variables (and careful selection (Hwang et al., 2019)), simple models can for instance combine sea-ice- and SST-patterns to make skillful forecasts of jet stream position (Hall et al., 2017), and blocking frequency (Miller and Wang, 2019).

Just like expert knowledge can aid ML methods in their discovery, so can ML-based discovery increase expert knowledge (Karpatne et al., 2017). Especially simple statistical models often directly expose their hypothesized physical pathway (think of a regression equation representing the influence of sea ice on jet stream position (Hall et al., 2017)). If such a pathway implies a previously unknown sub-seasonal process, then it can be studied further and can potentially inform new dynamical and statistical models. Recent developments in data sciences now enable us to enhance our knowledge via methods like eXplainable AI (XAI) (Mueller et al., 2019; Arrieta et al., 2020; Camps-Valls et al., 2020; Molnar et al., 2020; Ras et al., 2022). The latter are statistical tools that can also be applied to complex ML models to present their predictions in a humanly understandable way.

Skillful ML models statistically represent a myriad of conditional interactions that enable them to issue sharp uncertainty distributions in certain windows of predictability. Exposing this statistical encoding with XAI, and interpreting it physically, can generate new insights (Wilby et al., 2003; McGovern et al., 2019; Toms et al., 2020; Clare et al., 2022; Silva et al., 2022; Beobide-Arsuaga et al., 2023). An example is the discovery that diabatic heating in a little-studied tropical region can precede

predictable sub-seasonal changes above the North Atlantic (Mayer and Barnes, 2021).

1.4 Research questions and content of this thesis

In this thesis I explore various Machine Learning methods to address the challenges in sub-seasonal forecasting. These challenges range from detecting windows of predictability, disentangling the sub-seasonal processes that drive them, and representing those in models to issue the sharpest (but still reliable) uncertainty distributions possible. I explore ML methods in three different roles:

- as a detection tool to increase our physical knowledge of the system
- as a stand-alone statistical forecast model
- as a post-processing method for NWP forecasts

I focus on European two-meter temperature (t2m) forecasts as t2m extremes in the region are rising at a rate faster than the global average (Christidis et al., 2015; van Oldenborgh et al., 2022), while early warning capacity is still low (Casanueva et al., 2019). On top of that, sub-seasonal predictability in Europe is generally seen as poor and little understood (compared to other regions on the globe) (Monhart et al., 2018; Lavaysse et al., 2019; Mastrangelo and Malguzzi, 2019; Wulff, 2021; Pyrina and Domeisen, 2022). Chapter by chapter, I address the following research questions:

RQ1 At what spatio-temporal scales can we detect predictable sub-seasonal t2m variability in Europe?

RQ2 Can a Machine Learning model forecast sub-seasonal variabilities in summer, disentangle the driving processes, and have its learned representation interpreted by humans?

RQ3 Are the driving processes properly simulated in an NWP model, and if not, can these conditional errors be corrected through statistical post-processing?

RQ4 Can we physically understand the ML-detected processes that drive a window of predictability in European summer t2m?

In Chapter 2 (RQ1) I look at the problem of detecting the spatio-temporal scale at which windows of predictability exist. With unsupervised clustering I group locations whose daily weather states (as recorded in t2m) are governed by the same overarching variability. As their aggregate behavior is potentially predictable, I extract the aggregate signal by averaging t2m in space and time. Actual predictability of each space-time combination is then quantified by establishing the maximum leadtime with which the aggregate can be skillfully forecast. The used NWP forecasts are from the state-of-the-art EPS of the European Center for Medium-range Weather Forecasts (ECMWF). The t2m forecasts are post-processed with only a simple linear model to increase skill. We therefore consider any sub-seasonal predictability at the detected space-time scales to be a lower bound that can be improved upon in the rest of the thesis.

Subsequently, in Chapter 3 (RQ2), I conduct a study of relevant sub-seasonal processes. As imperfectly represented processes cannot be diagnosed from ECMWF forecasts alone, I create a stand-alone ML-based forecasting model, using ERA5 data (Hersbach et al., 2020). This model learns an alternative representation of sub-seasonal processes, that, upon inspection with XAI, helps us diagnose which processes are relevant. First I ease the learning task, by applying the clustering technique from Chapter 2 such that I train the ML model on a spatial scale at which t2m has some predictability. The algorithm then has to characterize the sub-seasonal processes in the multi-variate and multi-scale data, and model the interactions between them. Deep learning struggles to perform both tasks at once, so I adopt a two-step procedure. First, I detect and transform oceanic, terrestrial and atmospheric sub-seasonal variabilities into predictors, with a data-driven dimensionality reduction based on clustering. Second, I fit a random forest (an algorithm based on decision trees (Breiman et al., 1984; Breiman, 2001)), to represent the interactions between the predictors and issue the actual forecasts. The algorithm is able to make skillful summertime forecasts for monthly t2m in West and Central Europe. XAI sheds light on on the sources of the predictable signal, pointing to some well known processes but also uncovering new ones.

In Chapter 4 (RQ3) I build a post-processing method based on neural networks, that corrects conditional errors of the ECMWF model, using the sub-seasonal predictors of Chapter 3. I show that this post-processing leads to sharper and more skillful sub-seasonal t2m forecasts. With XAI I attribute the conditional corrections and reductions of uncertainty to the responsible predictors. In this way the ML method could show that

pronounced correctable errors over Europe relate to initial conditions in the tropical West Pacific. This hints at the misrepresentation of a teleconnection in the ECMWF model.

In Chapter 5 (RQ4) I study whether the teleconnection, discovered by ML in the previous chapter, can be physically understood. The conditional teleconnection supposes a pathway between SST patterns in the tropical West Pacific and t2m in West and Central Europe three to six weeks later. Specifically, the teleconnection involves the arrival and lingering of a Rossby Wave Packet over the Euro-Atlantic sector. For each aspect of the supposed pathway, namely from SSTs, to tropical deep convection, to atmospheric wave, to European surface impact, I show that it can be embedded in existing literature on Pacific variability and summertime teleconnections. Through detailed analysis of the pathway in 70 years of ERA5 data, I further show that the connection has emerged as window of predictability in the last four decades. Overall I demonstrate that a teleconnection, discovered by ML, can confirm and expand existing physical knowledge.

In Chapter 6, the synthesis, I return to each of the research questions and discuss my findings. Specifically, I reflect on the way machine learning can be used to generate new knowledge. What are the implications for future forecasting systems?

Chapter 2

Detecting sub-seasonal predictability

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

Abstract

The succession of European surface weather patterns has limited predictability because disturbances quickly transfer to the large-scale flow. Some aggregated statistics, however, such as the average temperature exceeding a threshold, can have extended predictability when adequate spatial scales, temporal scales and thresholds are chosen. This study benchmarks how the forecast skill horizon of probabilistic 2-m temperature forecasts from the sub-seasonal forecast system of ECMWF evolves with varying scales and thresholds. We apply temporal aggregation by rolling-window averaging and spatial aggregation by hierarchical clustering. We verify 20 years of re-forecasts against the E-OBS dataset and find that European predictability extends at maximum into the fourth week. Simple aggregation and standard statistical post-processing extend the forecast skill horizon with two and three skillful days on average, respectively. The intuitive notion that higher levels of aggregation capture large-scale and low-frequency variability and can therefore tap into extended predictability holds in many cases. However, we show that the

effect can saturate and that there exist regional optimums beyond which extra aggregation reduces the forecast skill horizon. We expect such windows of predictability to result from specific physical mechanisms that only modulate and extend predictability locally. To optimize sub-seasonal forecasts for Europe, aggregation should thus be limited in certain cases.

2.1 Introduction

Extending skillful weather predictions beyond two weeks and into the sub-seasonal range is of great importance for humanitarian concerns such as safeguarding crop harvests and preventing energy shortages (Coughlan de Perez et al., 2015; Guimaraes Nobre et al., 2019; Grams et al., 2017). These efforts are propelled by the intuition that extreme, large-scale events can potentially be predicted in advance (Vitart and Robertson, 2018). However, producing skillful forecasts of such large-scale events remains notoriously difficult.

The atmosphere is a dynamical system that varies on many spatio-temporal scales. Its succession of instantaneous states is deterministic but chaotic. Small disturbances can transfer to larger scales, growing in such a way that they overwhelm signals that were originally present. This means that deterministic atmospheric forecasts draw on predictability arising from initial conditions but will at some point become inaccurate (Lorenz, 1969). The forecast error will then relate to the total variance in the predicted phenomenon. The saturation of forecast error occurs most quickly for the finest details, whereas at larger scales of motion variations are observed that have the potential for predictability at longer lead times (Toth and Buizza, 2019; Ying and Zhang, 2017; Privé and Errico, 2015; Hoskins, 2013).

These potentially predictable variations can be internal to the atmosphere, or they can form in interaction with other components of the Earth system. Internally, the mid-latitude tropospheric variability is often dominated by a few large-scale patterns that recur and evolve into each other (Vautard, 1990; Hannachi et al., 2017), which are associated with predominant weather types on the ground (Grotjahn et al., 2016). Variability in Europe is also steered into specific regions of phase space by slow-moving components such as Atlantic sea-surface temperatures (Czaja and Frankignoul, 2002), snow cover (Orsolini et al., 2013; Henderson et al., 2018), soil moisture (Prodhomme et al., 2016), the stratosphere (Baldwin and Dunkerton, 2001; Tripathi et al., 2015) and tropical vari-

ability like the Madden-Julian Oscillation (MJO) (Cassou, 2008; Lin and Brunet, 2018; Vitart, 2017; Yadav and Straus, 2017). These components often interact, so in the sub-seasonal forecast range they represent not only the slowly evolving boundary conditions but also the part of the internal variability that provides predictability by changing the statistics of the higher frequency weather. Thus, naturally, the seamless transition from short to extended range forecasts requires aggregations that capture the variability of the large-scale patterns in our meteorological variable of interest.

In practice, sub-seasonal forecasting aims to extend the time window of the predictand with increasing lead time (Wheeler et al., 2017; Ford et al., 2018; Bürger, 2019; Nicolis, 2016). It has been demonstrated that more aggregation indeed leads to a general predictability in upper air fields at longer lead times (Roads, 1986; Jung and Leutbecher, 2008; Buizza and Leutbecher, 2015) and in surface variables like precipitation (Wheeler et al., 2017). Studies have also tailored the aggregation to a single conditional source of predictability: rainfall events in Europe that are clustered in time due to the large-scale dynamics (Economou et al., 2015; Pasquier et al., 2019; Yang and Villarini, 2019) or extreme temperatures occurring simultaneously within a spatial region due to large-scale flow or sea-surface temperatures (Stefanon et al., 2012; McKinnon et al., 2016; Vijverberg et al., 2020). The forecast skill of such derived predictands can be high, but it is conditional on the occurrence of the source mechanism, and might also lose validity for less or more extreme events (Wulff and Domeisen, 2019). To improve skill under all physical circumstances, statistical post-processing is often used to correct systematic biases and under- or over-dispersion. This aligns the model error growth with the real uncertainty growth (Wilks and Vannitsem, 2018). In this way studies demonstrated a predictability of weekly aggregations into week 3 and 4 for the mid-latitudes (Ferrone et al., 2017; Vigaud et al., 2017; Monhart et al., 2018).

In conjunction with increased aggregation leading to increased predictability, based on physical understanding one would also expect an optimum to exist. When too many different situations are aggregated, the conditional predictability in either one of them is lost, for instance by spatially aggregating hotspots of soil-atmosphere coupling with non-hotspots (Ardilouze et al., 2017a) or by temporally aggregating beyond the time window in which flow configuration modulates precipitation significantly (Barton et al., 2016). Predictability is then only regained by aggregating even further, for example to multi-month values to capture

the modulation of Europe's seasonal state by ENSO or soil moisture (Lee et al., 2019; Bunzel et al., 2018).

This study benchmarks how the sub-seasonal predictability of the surface temperatures in Europe varies over the continent and changes with the amount of temporal and spatial aggregation applied. We hypothesize that the maximal extension of the 'forecast skill horizon' (Buizza and Leutbecher, 2015) occurs under certain optimum aggregation levels and by statistical post-processing of the raw ensemble forecasts. Section 2.2 introduces the forecast ensemble, the scores to determine the forecast horizon and the post-processing method. Section 2.3 shows the resulting influences of post-processing and aggregation for events with varying exceedence thresholds. Section 2.4 provides a discussion and Section 2.5 summarizes and concludes.

2.2 Data and methods

2.2.1 Datasets

The forecast ensemble is the European Centre for Medium-Range Weather Forecasts' (ECMWF's) extended range forecasting system cycle 45r1, which extends their medium range ensemble twice a week to +46 days (Buizza et al., 2018). A degradation of the resolution takes place at +16 days. We downloaded forecasts of daily mean 2-m temperatures on a regular grid of 0.38x0.38°, as it equates the degraded spectral resolution in large parts of the European domain and minimizes the need for MIR-interpolation on the ECMWF MARS archive. For forecasting in the extended range, a lead-time-dependent bias in the model climatology can be expected (Johnson et al., 2019). All 11 members in the re-forecast period from June 1998 till May 2019 were therefore used to calculate forecast climatological means specific to the day of the year (± 5 days) and the lead time, that were subtracted from the forecast values. This results in forecast anomalies with respect to the model re-forecast climatology, and that are free from potential drifts in that climatology.

Observed temperature anomalies were derived from the 19.0 ensemble version of E-OBS (Cornes et al., 2018). Its ensemble mean forms the best guess of observed daily mean 2-m temperatures on a 0.25x0.25° grid. From the more than 60 years in the dataset we retained those 20 years that overlap with the re-forecasts. At each location we subtracted the day-of-the-year-specific observed climatological mean (± 5 days), calculated from January 1998 till December 2018.

The daily gridded anomalies from E-OBS are then paired with the eleven forecast anomalies in the nearest neighbor forecast ensemble grid cell, representing an area that is only slightly different. The datasets span from June 1998 to December 2018 and are built separately for the winter and summer season, December-January-February (DJF) and June-July-August (JJA), respectively. We allow days of forecasts that were initialized before the start of the seasonal window to be included (Coelho et al., 2018).

2.2.2 Aggregation

The paired daily anomalies at all E-OBS grid cells in Europe were then averaged to multiple spatial and temporal levels, and all combinations of those levels. The temporal levels consist of rolling 1- to 11-day window averages. Each of these windows is applied to all lead times equally, and assigns the lead time of a given forecast to the window center, which is a compromise between the more accurate first days and the more uncertain last days in the window (Weigel et al., 2008; Buizza and Leutbecher, 2015). Thus, for the window of size 7 days, the first possible midpoint lead time is 4 days, which is assigned the average of anomalies from forecast days 1-7 (see Fig. 2.1).

The spatial levels are determined by hierarchical clustering (Hastie et al., 2009). This method begins with as many clusters as there are grid cells, and a dissimilarity defined between each of these, say, time series A and B:

$$d_{A,B} = 1 - \max_{\tau=-20,\ldots,20} \rho(A_{t-\tau}, B_t). \tag{2.1}$$

This maximum in a set of correlations ρ with lags τ ranging from -20 to +20 days allows cells to be similar when experiencing the same (but temporally displaced) dominant weather features (Pfleiderer and Coumou, 2018). Each level of spatial averaging is then determined by grouping all sets of grid cells below a certain dissimilarity level (e.g. the level of 0.025 requires a minimum similarity, namely lagged correlation exceeding 0.975, between each of the cells) into single clusters, until the whole of Europe is one single cluster. We opt for an average linking rule (see Hastie et al., 2009). Our progression through the dissimilarity levels from 0.025 to 1 avoids the common problem of assuming a fixed number of clusters at the beginning of a study (e.g. Yiou et al., 2008). We perform the cluster extraction for winter and summer separately, using the observed daily temperatures from January 1989 to December 2018. The

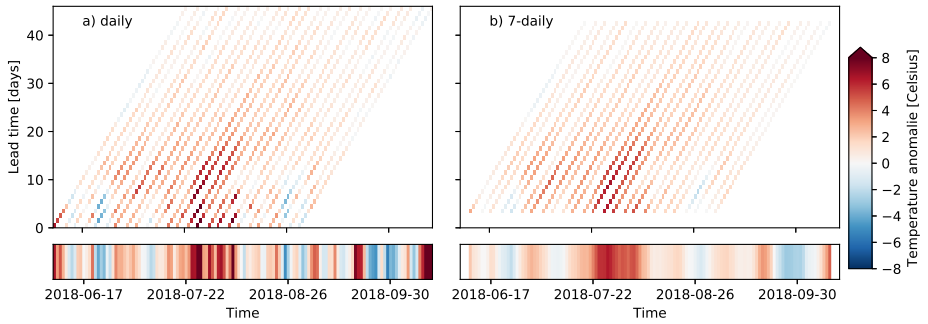


Figure 2.1: The temporal aggregation of ECMWF extended range forecasts and corresponding E-OBS observations, exemplified with daily mean 2-m temperature anomalies during the 2018 European heatwave. Top panels: ensemble mean of forecasts initialized each Monday and Thursday, aggregated to daily (a) and 7-day rolling means (b). Note that the window size remains 7 days for all lead times indicated on the vertical axis. Bars at bottom: corresponding E-OBS observations. Data taken from the grid cell closest to 52 degree latitude and 7 degree longitude.

use of daily time series separates our spatial aggregation from the temporal aggregation and allows a separate investigation of their effects. The supposed independence was briefly tested, and similarly shaped spatial clusters appeared at other time aggregations.

2.2.3 Scoring and forecast skill horizon

Each set of anomalies, averaged to a spatial and temporal level, is evaluated by comparing the distribution of forecast anomalies to the observed one. First we extract the forecast probability that a temperature anomaly will exceed a certain quantile in the 20-year model re-forecast climatology of averaged anomalies. The Brier Score (BS) then averages the squared difference between this probabilistic prediction p_i and the binary observation o_i (whether the observed anomaly exceeded the equivalent quantile in the observed 20-year climatology of averaged anomalies) over the n forecast-observation pairs per lead time and per spatial cluster in each set:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2. \tag{2.2}$$

Using two separate but equivalent thresholds, this BS extends the mean de-biasing, performed to create anomalies, with an implicit calibration of the raw ensemble forecasts to match the observed climatological spread. Additionally the BS of a reference forecast based on only the observed climatology is computed. It has a fixed p_i, namely 1 minus the quantile probability itself.

The full 11-member distribution is scored with the Continuous Ranked Probability Score (CRPS). The implicit calibration mentioned above has no effect on the CRPS as that score can be regarded as the BS integrated over all possible thresholds y, and accounts for reliability and sharpness (e.g. Wilks, 2011):

$$CRPS(F, y) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (F_{for,i}(y) - F_{obs,i}(y))^2 \, dy. \tag{2.3}$$

F_{for} is the forecast cumulative distribution function (cdf) and F_{obs} is the observed single-step cdf. As the forecast distribution is a discrete ensemble of 11 members, it receives worse CRPS scores than a version of the same underlying distribution with more members. A fair reference score is thus formed by sampling the same number of members ($M = 11$) from the empirical climatological distribution F of the observed anomalies, at intervals determined by a Weibull estimator (Wilks, 2011):

$$F^{-1}(\frac{m}{M+1}) \quad \text{for } m \text{ in } 1, ..., M. \tag{2.4}$$

A persistence reference forecast might be harder to beat, but since the construction of its probability distribution is non-trivial (Smith et al., 2015), we use the climatological reference to transform both scores to a skill score: $BSS = 1 - BS/BS_{clim}$ and $CRPSS = 1 - CRPS/CRPS_{clim}$. For each cluster and lead time we determine a confidence interval around these skill scores by scoring random samples (with replacement) from the set of n forecast-observation pairs. Because of auto-correlation, which will differ between clusters and which will increase with larger rolling-window sizes at higher temporal aggregation levels, this bootstrapping is done with different block lengths for each. The block lengths are based on a measure of the characteristic timescale T_0 (Fig. 2.2, Feng et al. (2011)):

$$T_0 = 1 + 2 * \sum_{\tau=1}^{D} (1 - \tau/D) * \rho_\tau, \tag{2.5}$$

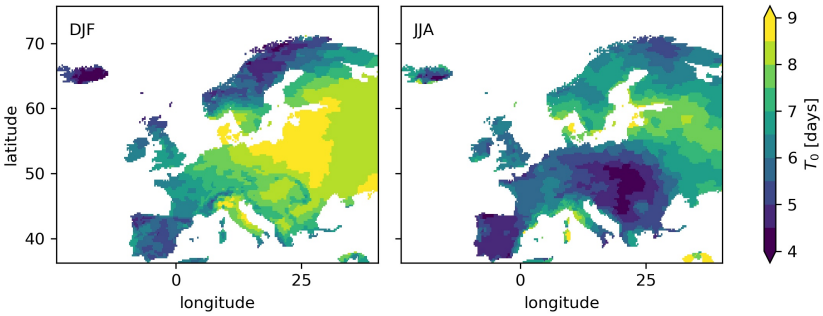


Figure 2.2: Characteristic timescale in the daily observed temperature anomalies at the 0.025 spatial aggregation level. Left: winter, 1158 clusters. Right: summer, 977 clusters.

where D is a cutoff lag, which similar to the hierarchical clustering we set to 20 days. ρ_τ is the auto-correlation between the lagged and unlagged time series of a cluster. The block bootstrapping is repeated only 200 times because of computational limitations. With these skill confidence intervals per cluster and per lead time we then deduce the local forecast skill horizons, defined as the lead time at which the lower bound of the interval (the 2.5th percentile) first crosses the zero skill line (Buizza and Leutbecher, 2015). This means: the lead time at which the forecast ceases to be statistically better than the climatological reference forecast, according to a one-tailed test at a 0.025 significance level.

2.2.4 Statistical post-processing

Besides scoring the aggregated, but otherwise unprocessed, forecast anomalies and scoring the climatological reference, we also score a version of the ensemble that is post-processed with a Non-homogeneous Gaussian Regression (NGR), which is a standard post-processing method for temperatures (Wilks, 2018). Its Gaussian distribution is assumed to have a location parameter μ_i and scale parameter σ_i that vary, respectively, with ensemble mean m_i and ensemble standard deviation s_i.

$$\mu_i = \alpha_1 + \alpha_2 \cdot m_i \tag{2.6}$$

$$\ln(\sigma_i) = \beta_1 + \beta_2 \cdot \ln(s_i) \tag{2.7}$$

The model is fitted using a three-fold cross-validation by minimizing the CRPS (Gneiting et al., 2005; Gebetsberger et al., 2018) on two thirds

of the 20-year dataset and validating on the other third. The model is fitted separately for each season, aggregation level, cluster and lead time. For scoring the post-processed distribution with CRPS we extracted 11 members with the estimator in (2.4) (a robustness test with 100 members gave similar results).

Obviously, NGR is a simple method that uses simple predictors and assumes normality even when it is inappropriate. Some studies have demonstrated the usefulness of more advanced predictors and post-processing methods in the sub-seasonal to seasonal range (Rodney et al., 2013; Hwang et al., 2019; Yoo et al., 2018; Strazzo et al., 2019; Kämäräinen et al., 2019). Such extensions are often specific to single sources of predictability or to a fixed time aggregation. In this study we compare the general predictability at varying aggregations, and aim to do this in a way that is simple but corrects for systematic errors.

2.3 Results

2.3.1 The effect of post-processing

In Figure 2.3 the lower bound of bootstrapped BSS is plotted for the exceedance of four climatological quantiles: for cold anomalies (0.15,0.33) and for warm anomalies (0.66,0.85). The lowest skill is seen for the full-color lines, which are the more extreme quantiles that are harder to predict than the more moderate terciles. At short lead times and the daily aggregation level (left panels in Figure 2.3), post-processing adds skill to the raw ensemble forecasts, even as the implicit calibration made the raw forecasts 'climatologically reliable' (Van Schaeybroeck and Vannitsem, 2015). What happens is that in these first five days the spread of the under-dispersed raw forecasts is increased by NGR, adding 'ensemble reliability' to the 'climatological reliability' leading to increased overall reliability (confirmed by a CRPS decomposition (Hersbach, 2000), not shown). After 5 days the added value becomes smaller as the raw forecast has better dispersion properties. At the 9-day aggregation level in winter, between lead times 5 to 13 days, the BSS values of the NGR and raw forecasts are even comparable (Figure 2.3b). Afterwards NGR forces ensemble spread to be similar to observed climatological spread when uncertainty is greatest at large lead times. In this unskillful range the zero BSS line is contained between the 2.5th and 97.5th percentile (upper bound not shown). The upper bounds of the bootstrapped BSS distribution of the raw ensemble are close to those of NGR (not shown) while

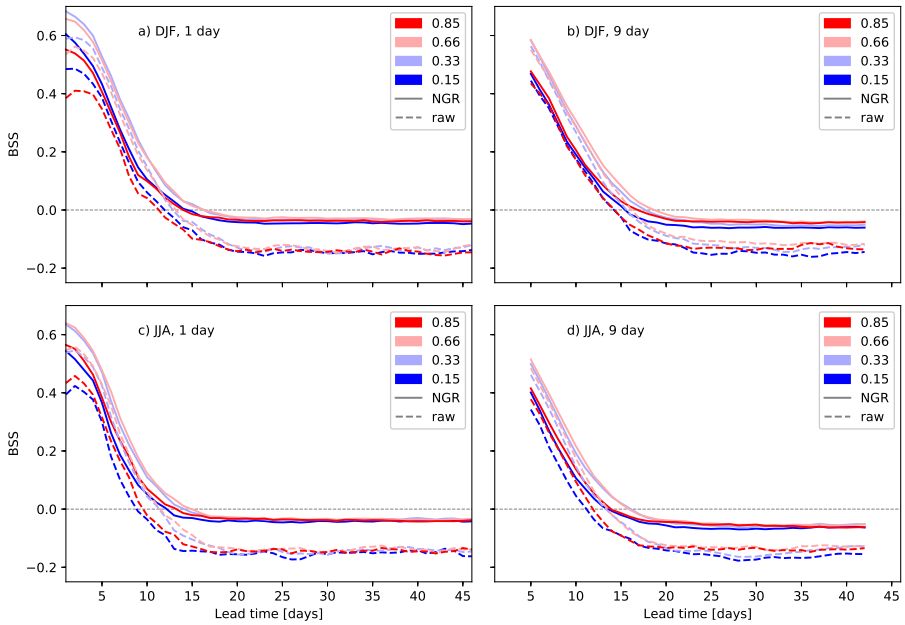


Figure 2.3: Area weighted average of the 2.5th percentile of bootstrapped BSS over all clusters at the 0.025 spatial aggregation level. Plotted for four different climatological quantiles, in winter (top), summer (bottom), and daily (left) and 9 day (right). Red lines are for the warm mean temperature anomalies, blue lines are for the cool anomalies. Post-processed in solid and the raw forecasts in dashed lines.

its lower bounds are lower due to its negatively skewed BSS distribution. The forecast horizon is defined by these lower bounds, meaning that NGR extends the lead time at which skill crosses zero by about 3 days. In the following we therefore only present results from post-processed forecasts.

2.3.2 The effect of aggregation

In Figure 2.4 we see how aggregation affects predictability in winter, as measured by the forecast skill horizon in the CRPSS. For each row, increasing time aggregation tends to increase predictability. The longest forecast skill horizon is obtained for the 11-day rolling average, except for Iceland. In Iceland the temperature range within a season is pretty narrow and is shifted by multi-annual variability. The climatological dis-

tribution obtained by pooling the years 1998 to 2018 is thus wider than the range of possibilities at each point in time, leading in comparison to overly skillful post-processed anomaly forecasts (see also the discussion of Figure 2.8). Other regions that for instance have 19-day predictability when a 9-day aggregation is applied to all lead times (light green, indicating a day 15 to 23 mean) and 20-day predictability when an 11-day aggregation is applied to all lead times (yellow, indicating a day 15 to 25 mean) imply that the forecast skill horizon can be extended by using the 11-day window. The longest forecast horizons are obtained for hardly any spatial aggregation (top row, 1158 clusters) or full aggregation to the European scale (bottom row, 1 cluster). At an intermediate level of spatial aggregation (from 483 to 20 clusters) some local regions have reducing and later increasing forecast skill horizons, indicating that space aggregation can work both as a benefit and as a disadvantage.

Similar results for summer are shown in Figure 2.5. Skillful forecast do not extend as long as for winter, showing the lower general predictability of summer temperature anomalies. Time aggregation has the largest influence at the lowest spatial aggregation. At larger spatial aggregations, the forecast skill horizon of regions sometimes hardly changes, meaning that the averaging works equivalent to a smoother.

Whether aggregation changes the forecast skill horizon just due to a smoothing of skill of underlying regions/days or due to the extraction of a signal with a truly different predictability is illustrated in Figures 2.6 and 2.7. In Figure 2.6 the lower bound of bootstrapped CRPSS in each cluster is plotted against lead time. At the 0.025 aggregation level the clusters form a spatial distribution, at the European level only one value per lead time bin. Both in winter and in summer at the daily time aggregation (Fig. 2.6a,c) the European CRPSS is clearly higher than the average of the underlying clusters. Particularly at lead times shorter than 11 days the European aggregate has more predictability than the ensemble of regions. Near the forecast horizon (where most clusters cross the zero line) it tends to the inter-quartile range, which implies that spatial aggregation acts as a smoother. Some individual regions have more skill and a more extended forecast horizon when the degree of spatial aggregation is limited. The dots above the zero line at very long lead times are locations with variable scores, not with interminable forecast horizons, their lower bound will have equally jumped below zero at earlier lead times. At longer time aggregations (Fig. 2.6b,d) the mentioned effects of space aggregation are less pronounced but still present.

In Figure 2.7 the spatial CRPSS distributions belonging to two time-

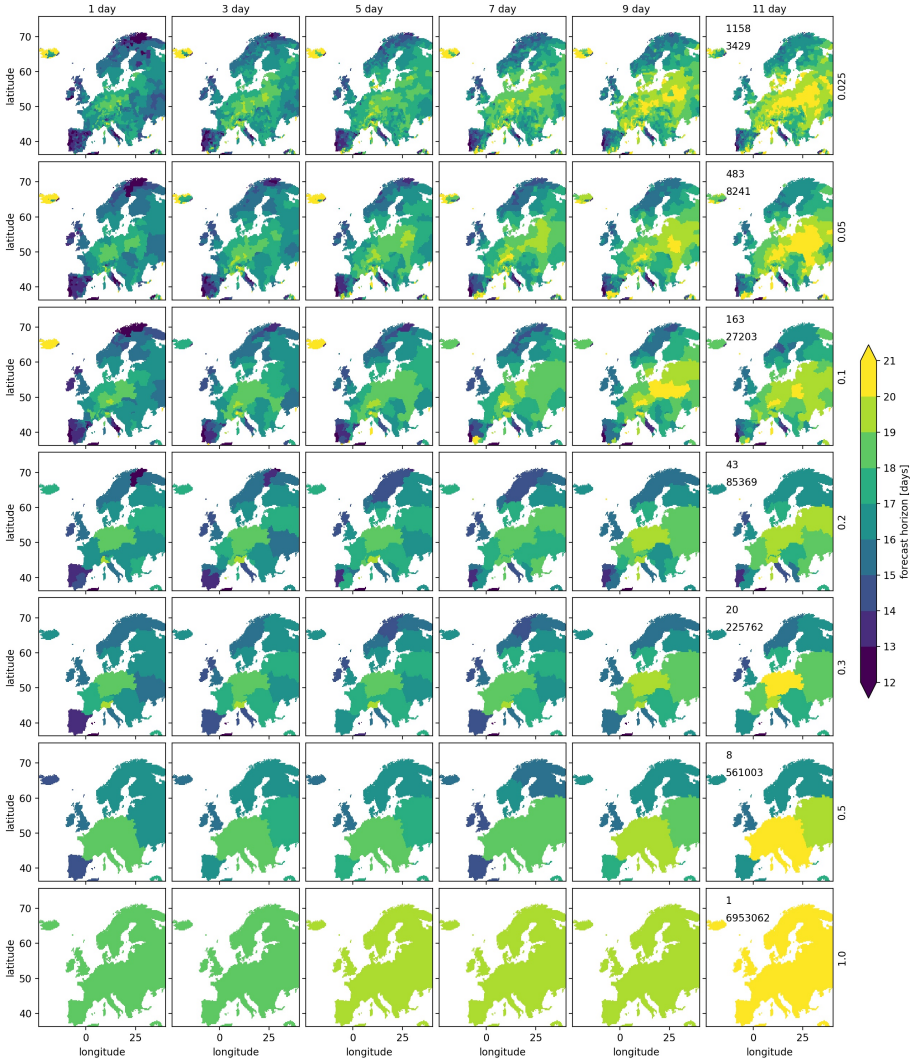


Figure 2.4: Forecast skill horizon [days] for post-processed winter temperature anomalies for different levels of spatial aggregation from small-scale to Europe-wide averages (rows), and temporal aggregation from daily to 11-day rolling averages (columns). In the right column the dissimilarity threshold, the number of clusters and their median size [km^2] are annotated.

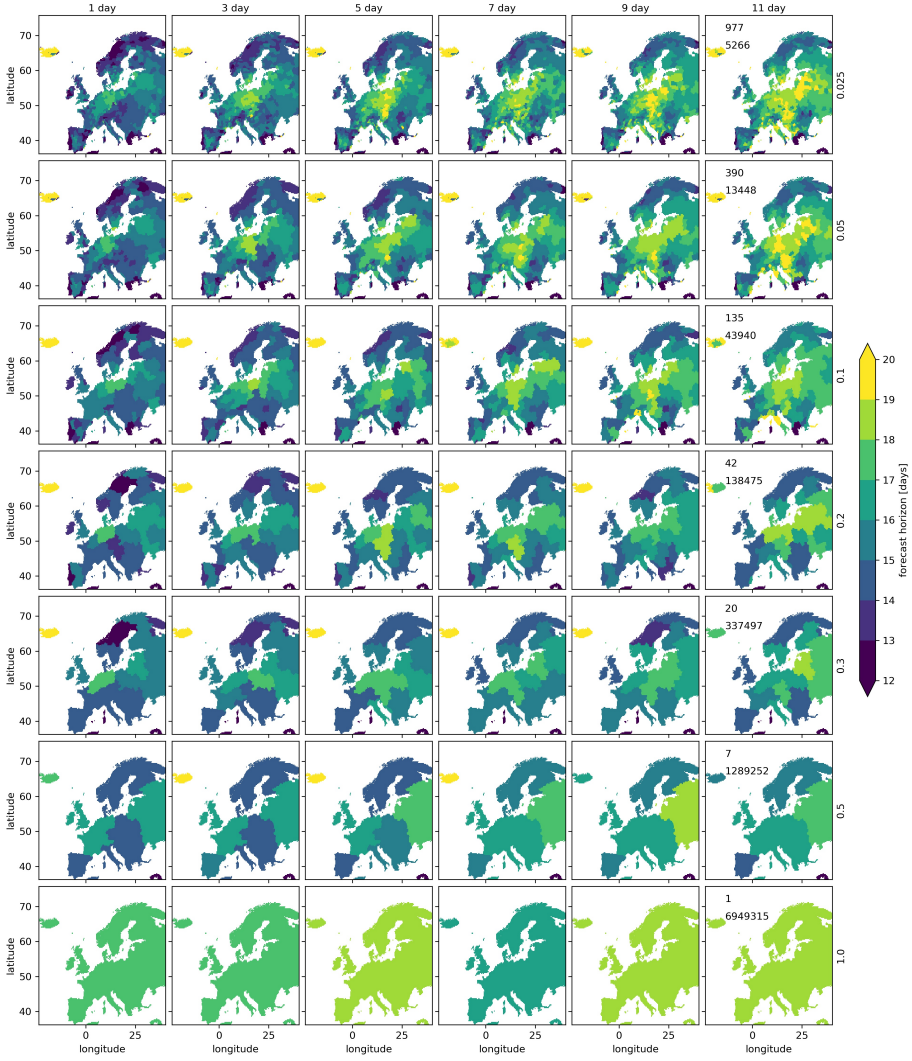


Figure 2.5: As Figure 2.4 but for summer. Note that a different spatial clustering and a different color scale has been applied than for the winter season.

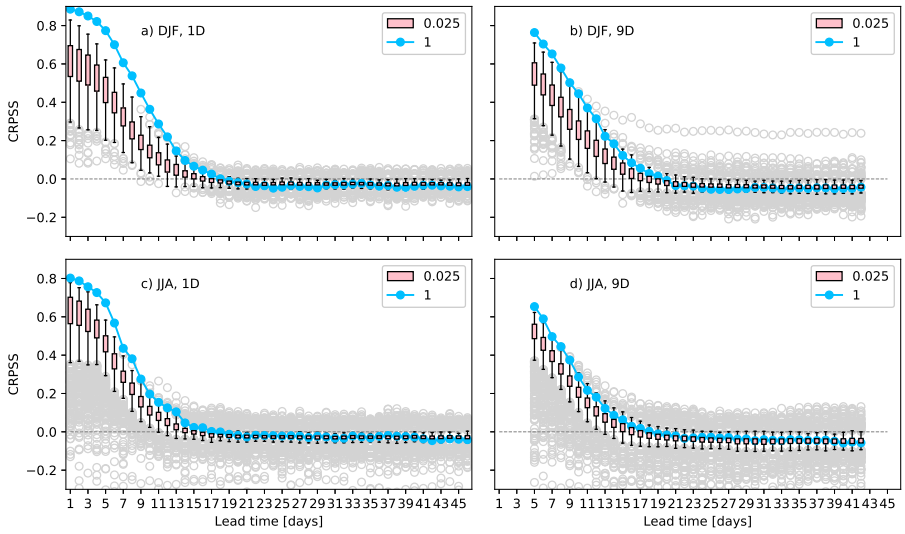


Figure 2.6: The influence of spatial aggregation on the 2.5th percentile of bootstrapped CRPSS at a 1-day (left) and 9-day (right) time aggregation. The spatial distribution at the 0.025 aggregation level (1158 clusters in winter (top) and 977 clusters in summer (bottom)) is shown with a pink box for the inter-quartile range, whiskers extending to 1.5 times that range and with outliers in grey. The European averaged aggregation (1 cluster) is shown in lightblue.

aggregations are compared (outliers are not shown for clarity). The inter-quartile ranges of the daily and the 9-day scores only become distinguishable at lead times exceeding 6 days, indicating that beyond that lead time a predictable multi-day variability was well initialized and was captured by the simple 9-day average. It extends the median forecast horizon by about two days. Some of the differentiation between the time aggregation levels can also be related to the convex shape of the curve between lead times 9 and 15 days. There the temporal window is a favorable mixture of initial days that are much more predictable than its center day (to which its lead time is assigned) and final days that are only slightly less predictable. However, higher CRPSS values also appear along the straight section between lead times 5 and 9 days and Buizza and Leutbecher (2015) also demonstrated that the skill of time-averaged variables is higher than the skill of time-averaged scores. Therefore, we are confident that the increased forecast horizon can be attributed to the temporal aggregation

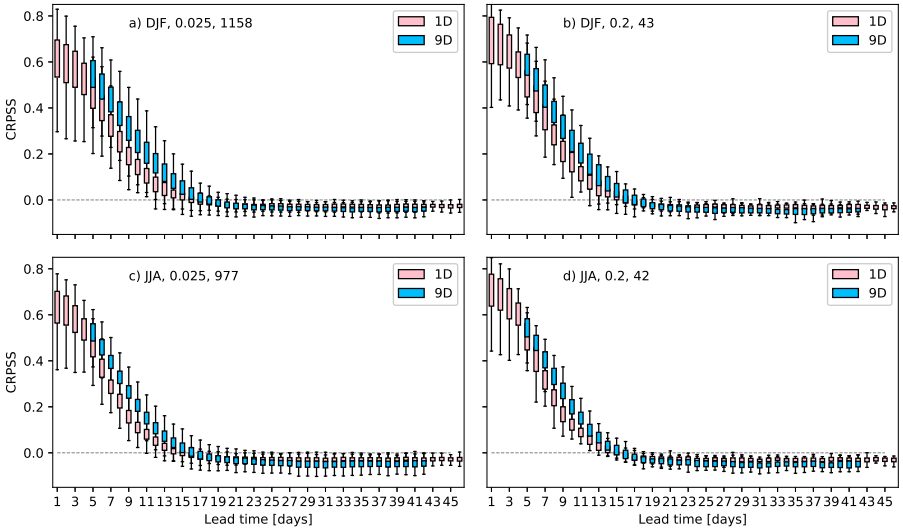


Figure 2.7: The influence of time aggregation on the 2.5th percentile of bootstrapped CRPSS. The spatial distribution of daily time series is shown in pink and that of 9-day rolling averages in lightblue box-and-whisker plots. Boxes represent the inter-quartile range, whiskers extend to 1.5 times that range and outliers are not shown. Season, spatial aggregation level and number of clusters are annotated.

applied.

The extensions and regional optimums that we find have to be related to sources of predictability. It can be expected that such sources are related to particular types of events, and that their conditional predictability emerges at a certain intensity level. Becker et al. (2013) found for instance increased signal to noise ratios for extreme events, despite an also increasing error in predicting them. In Figure 2.8 we show the BSS forecast skill horizon for predicting the exceedance of varying climatological quantiles. Note that the forecast skill horizon for Iceland is now strongly reduced compared to Figures 2.4 and 2.5. This difference is not surprising because the CRPS is an integration of the BS over all possible thresholds per point in time, while the BSS in Figure 2.8 is created by summing over time first. The inflated CRPSS for Iceland followed from a reference that was too wide for the varying set of possibilities at each point in time. In case of the BSS the varying exceedance probability is over-estimated in some years and under-estimated in others, but aggre-

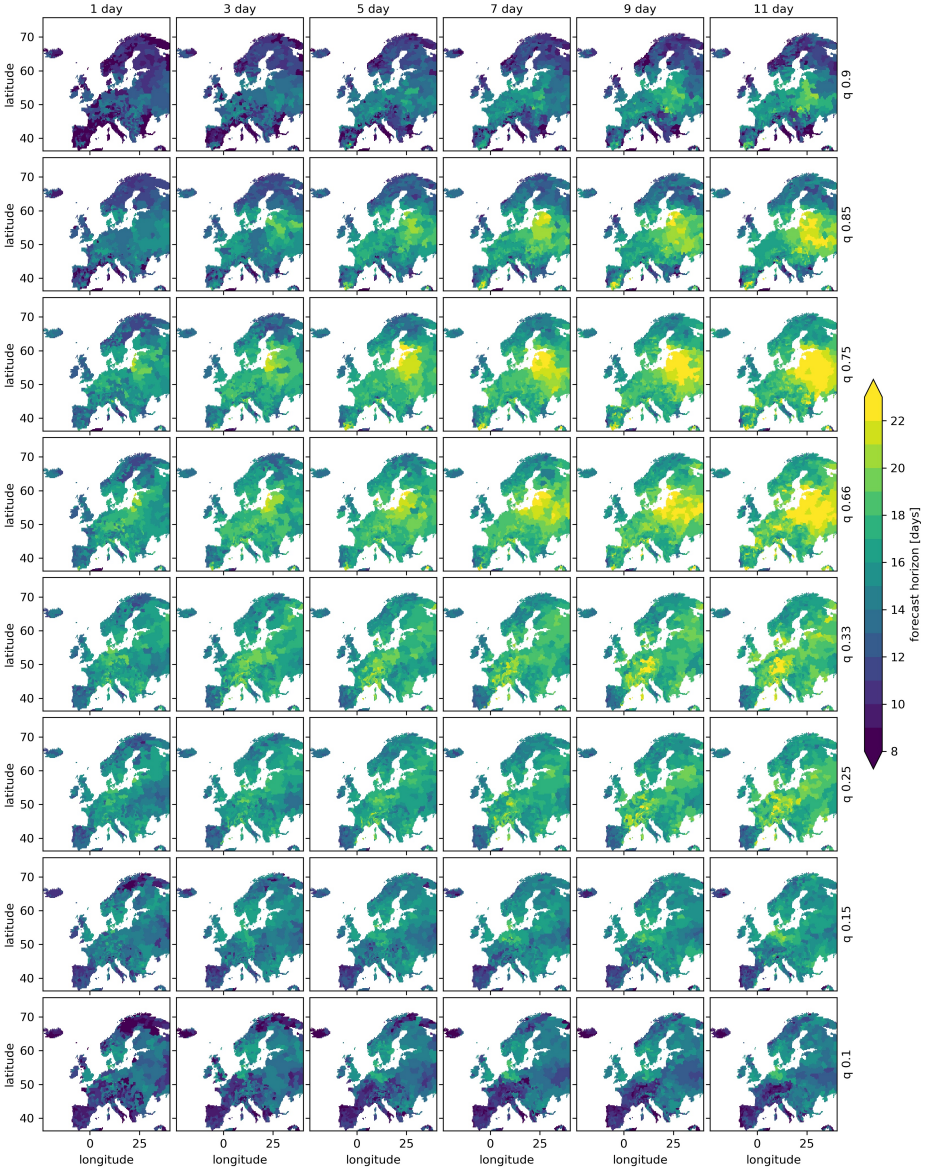


Figure 2.8: Forecast skill horizon [days] of post-processed forecasts in winter for different climatological quantiles (rows), for varying levels of time aggregation (columns), and at the 0.025 spatial aggregation level (1158 clusters). Bottom: cold tail, top: warm tail.

gated over all forecast occasions the reference is by definition exactly right and skill is not inflated. The source of these multi-annual changes is to be sought in the sea-surface temperatures (Frajka-Williams et al., 2017). These are able to dominate because E-OBS includes only coastal stations in Iceland (Cornes et al., 2018).

Particularly for the largest time aggregation (Fig. 2.8, right column) the forecast horizon for the different quantiles shows an asymmetry. The upper tercile is more predictable than the lower and this predictability is primarily located in the east of the domain. In the raw forecasts this spatial structure is also present, but less pronounced (Fig. 2.9), indicating that the asymmetry is partially caused by NGR. Closer investigation revealed that the climatological distribution in the regions with long forecast skill horizons is negatively skewed. Initially NGR corrects the underdispersion of the raw forecast and does well, but when uncertainty increases with long lead times and dispersion approaches climatology, the thicker lower tail is badly represented by the post-processed Gaussian shape and we see the performance for cold quantiles drop relative to the warm quantiles.

For summer (Fig. 2.10) time aggregation does not clearly reveal an asymmetry in the forecast skill horizons. At quantiles 0.1 and 0.9 some regions show consistently short forecast horizons, despite time aggregation. Generally the moderate events in the bulk are better predictable than events in the tails. This ordering contrasts the study of Wulff and Domeisen (2019), who found that European warm extremes in summer, exceeding the 90th percentile and at a 5 day temporal aggregation, are more predictable than moderate events between the 25th and 75th percentile. They found this for the warm tail only, so they hypothesized that the emergent source of conditional predictability related to land-atmosphere feedbacks and large-scale circulation. Here we find no indication for an emergent source.

2.4 Discussion

The forecast skill horizons presented above are in agreement with other estimates of European unconditional predictability in bias-corrected forecasts (Ferrone et al., 2017; Monhart et al., 2018). We find the forecast skill horizon for the full forecast distribution to be highest in winter where midpoint lead times extend to slightly above 21 days, meaning that the windows of predictability can be extended up to weeks 3 and 4. In this

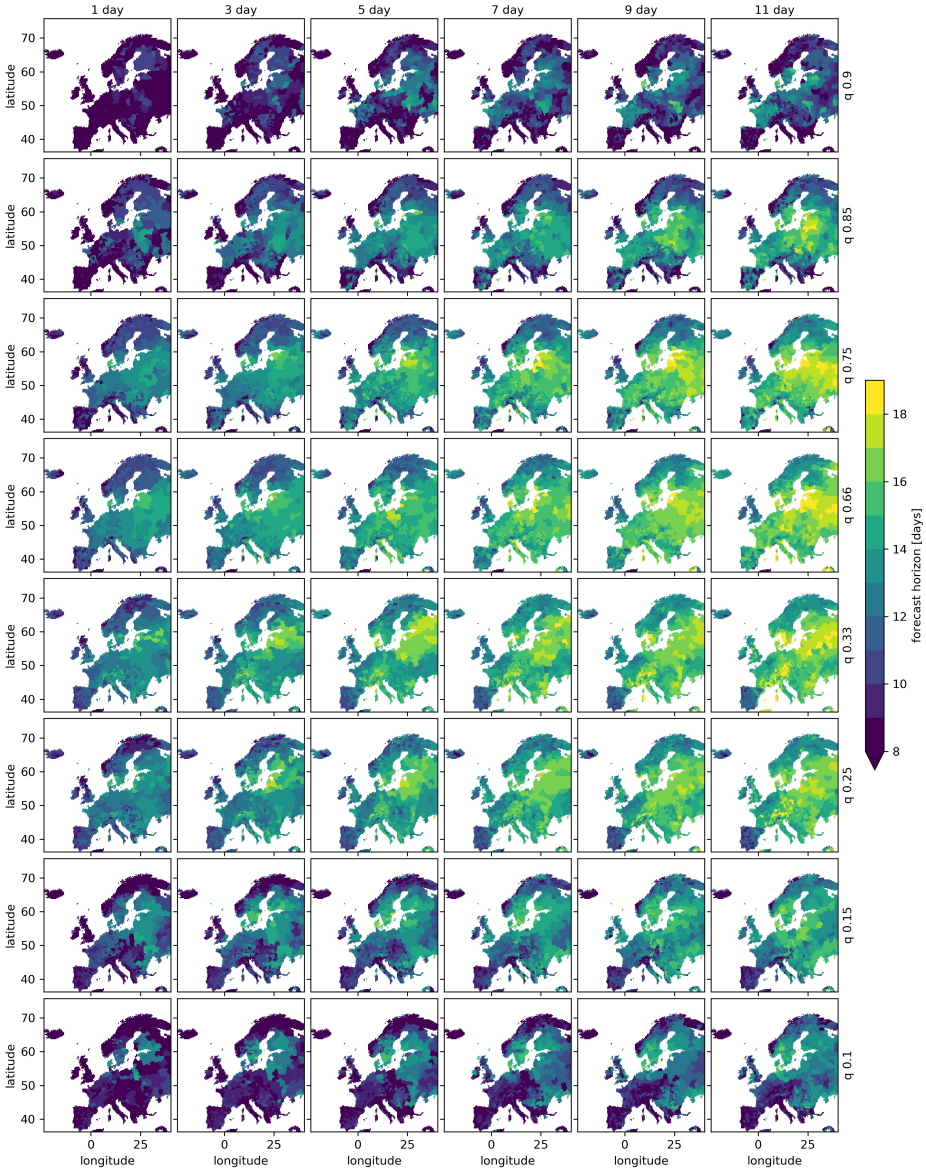


Figure 2.9: As Figure 2.8 but for raw DJF forecasts. Note the different color scale.

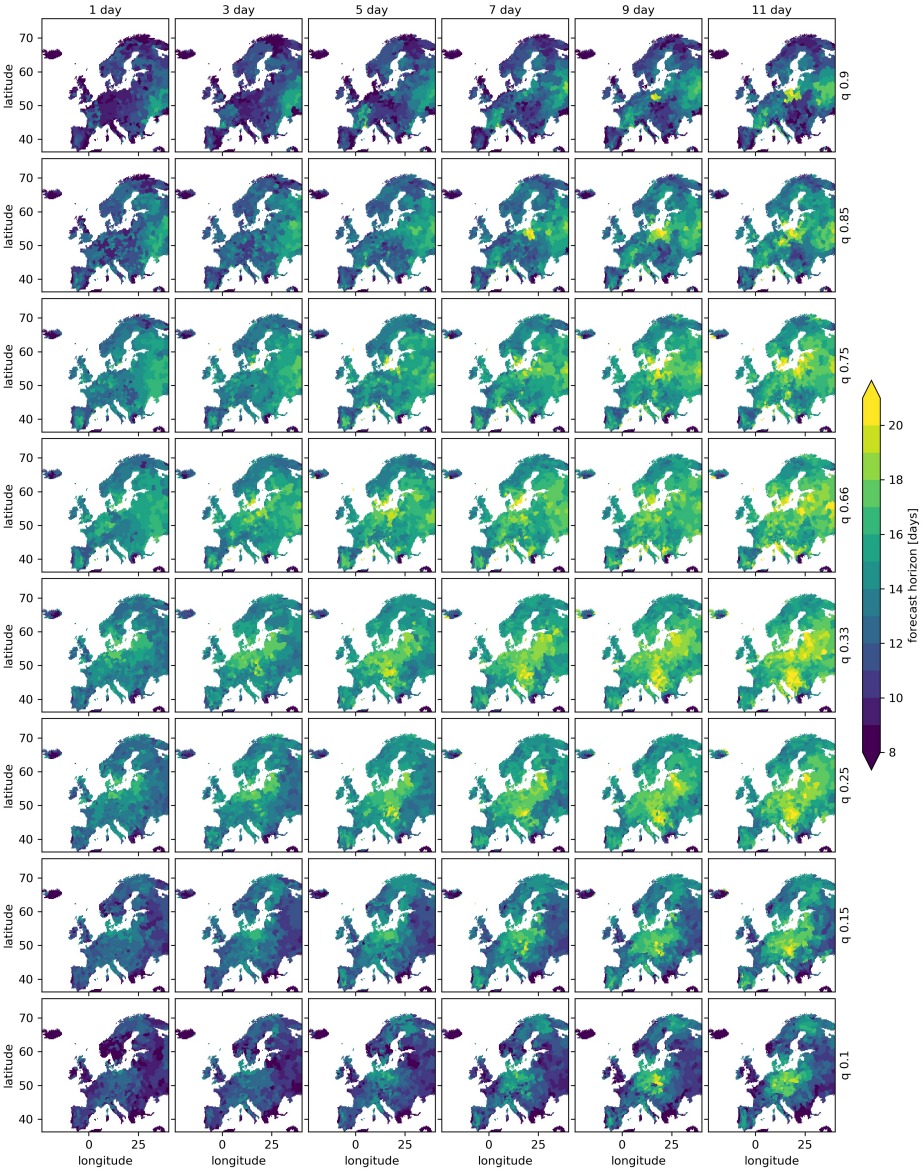


Figure 2.10: As Figure 2.8 but for summer. Note the different color scale.

study we have varied the level of aggregation to test its impact on the predictability horizon. Our finding shows that no distinct aggregation captures the one and only predictable sub-seasonal signal in Europe. Results suggest that the predictable mode of variability varies over the domain and that aggregation can increase predictability (but not always).

For areas where sub-seasonal predictability exists, time aggregation increases skill, predominantly beyond a given lead time (Fig. 2.7). This confirms that it is optimal to apply aggregation only when uncertainty has increased with lead time and when the predictable low-frequency signal remains (Ford et al., 2018; Bürger, 2019). In other places however, especially for more extreme quantiles, time aggregation hardly affected the forecast skill horizon. It just smoothed the skill (or the absence thereof) over time and no predictable signal appeared that is captured by simple averaging.

In contrast, space aggregation changed the signal considerably at short lead times (Fig. 2.6). For the first 11 days the European average is easier to predict than the ensemble of regions. At this largest spatial aggregation, winter results (Fig. 2.4) showed that it is best to also aggregate in time. It confirms that the dominant features are best captured by changing both the space and time filter as both scales are related (NAO varies slowly and influences temperatures over the entire European continent) (World Meteorological Organization, 2015). However, this study also found the opposite evidence, namely that for certain regions the forecast horizon is not maximized by increasing spatial aggregation. We think that predictability in either of the underlying regions is then lost by mixing the physical mechanisms that modulate locally.

We hypothesized that specific sources of predictability could be identified from anomalies exceeding specific climatological quantiles. One school of thought is that extreme events are related to predictable large-scale drivers and therefore better predictable themselves (Sillmann et al., 2017). The other is that extremes are actually harder to predict because they need a rare synchronization of processes at all relevant scales. We did not investigate extremes beyond the 10th and 90th percentile, but our results support both schools of thought. The BSS curves (Fig. 2.3) showed that tail events are harder to predict and we also found no indication of the increased predictability of summer warm extremes ((Wulff and Domeisen, 2019), Fig. 2.10). On the other hand, the winter BSS (Fig. 2.8) displayed a regional signal in the upper quantiles (visible only at larger time aggregations) which we might relate to an emergent predictable phenomenon.

A candidate mechanism associated to above normal temperatures in winter for a ±10-day time aggregation could be the early disappearance of the snow pack, as this increases the absorption of shortwave radiation and will take time before it can rebuild itself again. Certainly, the regionally extended forecast skill horizons are at least also partly due to the persistence of weather. The region around the Baltic Sea and Denmark is persistent in winter (Fig. 2.2) and strongly imprinted by the first principal component of the large-scale Euro-Atlantic atmospheric variability (Ferranti et al., 2018). This results in a relatively skillful region in our analysis (as in Monhart et al. (2018)).

Clusters that displayed weak predictability can be interpreted as devoid of sources of real extended predictability, but might also just indicate biases in the model. With NGR we attempted to remove biases and under- and over-dispersion for each lead time, season and cluster. This led to noticeable increases of skill, but also to a bias for winter cold anomalies at long lead times in regions with a skewed climatological distribution (Fig. 2.8). A correction approach that can better handle such distributions and for example the multi-annual variability change in Iceland, might be realized with other simple (Vigaud et al., 2017; Ferrone et al., 2017) or more advanced (Hwang et al., 2019; Yoo et al., 2018; Strazzo et al., 2019; Kämäräinen et al., 2019) post-processing methods.

2.5 Conclusion

This study has demonstrated that the forecast skill horizon for average temperatures varies over the European domain and can be extended to week 3 and 4 without preconditioning. A standard non-homogeneous Gaussian regression post-processing step added three skillful forecast days on average. The influence of space and time aggregation was explored by a protocol that allowed a clean comparison of different aggregation levels. We found that simple averaging captures predictable large-scale patterns in the high frequency weather and that this aggregation becomes especially effective beyond a few days lead time, adding two skillful days on average. For some regions however, time aggregation simply smoothed skill over time, showing that not everywhere a signal gets extracted by aggregation. Also space aggregation, when applied at an intermediate level, was found to lead to smoothing, therefore discarding the local extended forecast horizons present in some regions. To optimize sub-seasonal predictability in Europe, aggregation should in certain cases thus be limited,

especially when it is important to trace back the signals to the associated sources of predictability. This tracing is further eased when not only particular spatio-temporal scales are known but also the type and intensity of the events. We have demonstrated that quantiles can be used for such a stratification, but that not always a source of extended predictability emerges for the more extreme ones. A recommended extension of this study is to explore other statistics than the average (e.g. DelSole and Tippett, 2009). The predictable modes of variability might be better detected with meteorological index variables such as the clustering of warm days or rainfall events, than with temperature or rainfall averages.

Chapter 3

Disentangling driving processes with ML

Published as:
van Straaten, C., Whan, K., Coumou, D., van den Hurk, B. and Schmeits, M. (2022) Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central Europe *Monthly Weather Review* 150, 1115-1134

Abstract

Reliable sub-seasonal forecasts of high summer temperatures would be very valuable for society. Although state-of-the-art numerical weather prediction (NWP) models have become much better in representing the relevant sources of predictability like land- and sea-surface states, the subseasonal potential is not fully realized. Complexities arise because drivers depend on the state of other drivers and on interactions over multiple time-scales. This study applies statistical modeling to ERA5 reanalysis data, and explores how nine potential drivers, interacting on eight time-scales, contribute to the sub-seasonal predictability of high summer temperatures in western and central Europe. Features and target temperatures are extracted with two variations of hierarchical clustering, and are fitted with a machine learning (ML) model based on Random Forests. Explainable AI methods show that the ML model agrees with physical understanding. Verification of the forecasts reveals that a large part of predictability comes from climate change, but that reliable and valuable sub-seasonal forecasts are possible in certain windows, like forecasting

monthly warm anomalies with a lead time of 15 days. Contributions of each driver confirm that there is a transfer of predictability from the land- and sea-surface state to the atmosphere. The involved time-scales depend on lead time and the forecast target. The explainable AI methods also reveal surprising driving features in sea-surface temperature and 850 hPa temperature, and rank the contribution of snow-cover above that of sea-ice. Overall, this study demonstrates that complex statistical models, when made explainable, can complement research with NWP models, by diagnosing drivers that need further understanding and a correct numerical representation, for better future forecasts.

3.1 Introduction

In the recent two decades Europe faced a multitude of impactful heat extremes (e.g. Barriopedro et al., 2011; Russo et al., 2014), that exceeded the modeled expectation (van Oldenborgh et al., 2013). These were disasters that ended in loss of life and a severe disruption of activities. If forecasts would exist that reliably predict such events more than two weeks in advance, then new forms of anticipatory risk management can be realized (White et al., 2017). Hints of such predictability have already been found at lead times ranging from sub-seasonal (2-6 weeks) (Wulff and Domeisen, 2019) to seasonal (Weisheimer et al., 2011; Prodhomme et al., 2016, 2021), but the current time window to act upon operational forecasting systems remains limited to shorter lead times (Coughlan de Perez et al., 2018; Casanueva et al., 2019).

The complication with sub-seasonal lead-times is that climate variables have time-varying contributions to predictands. Involvement of a variable is conditional on the state of other variables (Mariotti et al., 2020). Only occasionally do windows of predictability appear in the larger background of chaotic variation. These are conditions in which sub-seasonal forecasts have a greater opportunity to succeed (Albers and Newman, 2019; Mariotti et al., 2020; Mayer and Barnes, 2021). Sahelian heatwaves are for instance more successfully forecast during active modes of tropical variability (Guigma et al., 2021). But windows of predictability are hardly regular. Contributions of key drivers of heat extremes do vary from event to event (Wehrli et al., 2019). Also, interactions that lead up to an event, take place over a range of timescales and locations (Sillmann et al., 2017). Such complexities have hampered our current understanding of the set of potential heatwave drivers, and of interactions

between them (Perkins, 2015). To therefore discover and leverage new opportunities for European sub-seasonal forecasts, we need to learn what methods of sufficient complexity learn. When we supply the climate variables that we know are important, machine learning (ML) tools can help us to extract better driving features from them (Cohen et al., 2018).

Clearly involved is atmospheric variability, more specifically the synoptic high-pressure 'blocking' systems that are associated to high surface temperatures (Brunner et al., 2017; Schaller et al., 2018). In the mid-latitudes these systems are part of quasi-stationary Rossby Wave Packets (RWPs) (Schubert et al., 2011; Röthlisberger et al., 2019). In transient state such groups of waves travel eastwards along the waveguide of the jet stream, instigating low- and high pressure systems downstream (Wirth et al., 2018). But in quasi-stationary state the pattern persists over a geographical location, leading to potentially large temperature anomalies (Wolf et al., 2018). Air parcels enter the stagnant high pressure system, descend and adiabatically heat the lower atmosphere (Zschenderlein et al., 2019).

Variability in the land surface is also involved, and conditionally enhances or negates high temperatures. When the land surface is dry, the excess energy from atmospheric temperature and clear-sky conditions will primarily go into diabatic sensible heating, but when the land surface is wet the excess will be used for evapo-transpiration too (Seneviratne et al., 2010; Miralles et al., 2019). The potential feedback from a dry land surface in summer can follow after winter and spring precipitation deficits (Quesada et al., 2012). Antecedent drought can also be driven by temporally enhanced transpiration when vegetation greens anomalously in spring (Fischer et al., 2007; Ma et al., 2016). Apart from inter-seasonal interactions, the land surface is also involved in shorter term interactions that can amplify developing events (Schumacher et al., 2019). Fischer et al. (2007), Haarsma et al. (2009) and Zampieri et al. (2009) find that drier soils can increase upper level high pressure, and could partially be responsible for stagnating the RWP pattern.

The third type of variability involved at multiple timescales comes from the global ocean. Two-way interaction with anomalous Sea Surface Temperatures (SSTs) can drive meridional jet stream position over the North Atlantic in June, July and August (Duchez et al., 2016; Ossó et al., 2020). More specifically within the season, SST patterns can predictably precede hot events by sourcing involved RWPs (McKinnon et al., 2016), which might be the way the tropical Atlantic influenced the 2003 heatwave (Cassou et al., 2005). SST patterns can also reinforce and thereby stag-

nate existing RWPs (Black and Sutton, 2007; Della-Marta et al., 2007; Feudale and Shukla, 2011). At Arctic latitudes the ocean state is equally important, with sea-ice cover influencing summertime jet stream position and quasi-stationary RWP amplitude in the Euro-Atlantic region (Hall et al., 2017; Wolf et al., 2020). The same effect, originating from Arctic land mass, is found for snow-cover.

Clearly, the features leading up to hot events are not limited to a single timescale or region. Part of this is due to interaction between climate variables, but the involvement of multiple timescales is also just a fact of atmospheric dynamics itself (Schneidereit et al., 2012). Numerical model experiments can be used to disentangle different contributions (e.g. Koster et al., 2010; Stéfanon et al., 2012; van den Hurk et al., 2012; Wehrli et al., 2019; Osborne et al., 2020). Through one-by-one manipulation, a variable's role in feedbacks or as source of predictability can be diagnosed. In observations such causal 'what-if' manipulation is not possible (Runge et al., 2019). Empirical analysis in observations is more likely to measure 'association'. The empirical approach however, will include mechanisms that are imperfectly represented in numerical models. Successful statistical diagnosis has happened through composite driver statistics conditioned on high temperature events (e.g. Stefanon et al., 2012; Brunner et al., 2017), composite temperature statistics conditioned on Euro-Atlantic circulation (e.g. Cassou et al., 2005; Jézéquel et al., 2018), plain predictive association like regression in a small set of variables (e.g. Quesada et al., 2012; Hall et al., 2017; Suarez-Gutierrez et al., 2020; Kueh and Lin, 2020), analysis of dominant modes in the full multivariate space (e.g. Della-Marta et al., 2007; O'Reilly et al., 2018), or using a potential driver as covariate in a modeled distribution of temperature (e.g. Whan et al., 2015).

What all empirical approaches have in common is a limited scope in terms of variables and timescales. Subsets of variables are selected a-priori and interaction between them is often ruled out. This leads to a setup that is easy for humans to understand, but that could falsely attribute underlying, undiscovered features to ones that are only partially involved. Such partial information, not representing the source of predictability itself, might vary from forecast occasion to forecast occasion. When operationalized, we would not know whether the forecasts can be trusted, and whether its learned patterns are actionable, should they arise in real-time.

This study presents a data-driven method to extract information from nine important climate variables related to high European temperatures

Table 3.1: Nine climate variables related to high summer 2-meter air temperatures in Europe. Resampled to daily values from the hourly ERA5 reanalysis

variable	abbreviation	ERA-5	daily resampling	unit
2m temperature	t2m	single level	24hr mean	K
300hPa geopotential	z300	pressure level	12UTC	$m^2 s^{-2}$
850hPa temperature	t850	pressure level	12UTC	K
total cloud cover	tcc	single level	24hr mean	-
sea surface temperature	sst	single level	24hr mean	K
sea-ice concentration	siconc	single level	24hr mean	-
snow cover	snowc	Land	24UTC	-
transpiration	transp	Land	24hr accumulation	m
shallow volumetric soil water	swvl13	Land	24hr mean, depth-weighted average of upper three layers (0 - 100cm)	$m^3 m^{-3}$
deep volumetric soil water	swvl4	Land	24hr mean, bottom layer (100 - 289 cm)	$m^3 m^{-3}$

in June-July-August (JJA), limiting a-priori choices. Subsequently, an ML method based on Random Forests reconstructs the interaction between variables on a range of timescales, and forecasts the exceedance of regional temperature above a given threshold. The forecasts are verified in terms of skill and potential value to users. Foremost we are interested in the climate variable features that the method extracts and leverages as sources of predictability. Usually such complex 'black box' ML methods are hard to understand, making them un-trustworthy in their own way. Here we demonstrate how explainability tools, which have become more and more mature in recent times (McGovern et al., 2019; Molnar et al., 2020), can be used to query the ML method for prior knowledge. For instance for the theory that relevant information is carried across variables and timescales, from oceanic RWP sources and antecedent land surface conditions long before the event, to the atmospheric state close to the event. Lastly, we show how the method could behave in real-time. To that end we study its predictions of the European heatwave in 2015 (Duchez et al., 2016; Ardilouze et al., 2017b). Section 3.2 introduces the data processing steps and ML methodology. Section 3.3 presents the verification results, found sources of predictability and 2015 case study. Section 3.4 discusses and concludes.

3.2 Data and Methods

3.2.1 Reanalysis data

Nine climate variables and the target 2-meter air temperature (t2m) are obtained from the ERA5 reanalysis. This modeled reality, based on as-

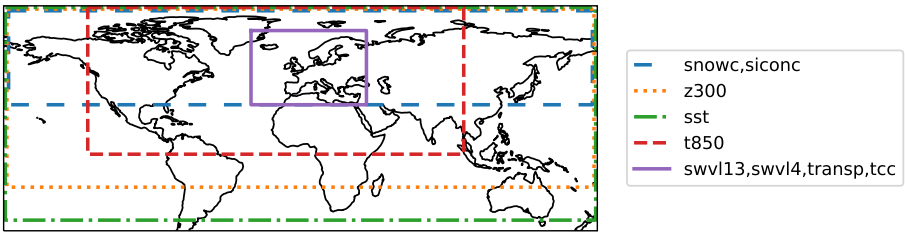


Figure 3.1: Spatial domains used to extract climate variables from ERA5. See Table 3.1 for the meaning of variable abbreviations.

similated observations (Hersbach et al., 2020), has the benefit of being multi-variate, spatially dense and temporally homogeneous. Atmospheric variability is represented by 300hPa geopotential (z300), 850hPa temperature (t850) and total cloud cover fraction (tcc). Respectively these capture upper level wave patterns, lower level heating and the clear-sky conditions of high pressure systems. Oceanic states are represented by sea ice concentration (siconc) and sea surface temperature (sst), also from ERA5. Land surface variability is represented by snow-cover fraction (snowc), transpiration (transp) and deep- and shallow volumetric soil-moisture (swvl4 and swvl13), all from ERA5-Land (Muñoz-Sabater et al., 2021).

Since ERA5-Land starts in 1981, two years later than ERA5, we extract the data from 1981 till 2019, at high spatial resolution (0.25*0.25 and 0.1*0.1 degrees for ERA5 and ERA5-Land, respectively). Domains are variable-specific (Fig. 3.1). When teleconnections, like influence from the tropical Atlantic, were expected, the domain was kept large enough for the ML-method to use potential features from these regions. For variables without expected global teleconnections we used smaller domains, to ease the total computational load. Hourly values at all gridcells were resampled to a daily resolution (Table 3.1). For each variable we removed the seasonal cycle by subtracting for each calendar date the average value of similar calendar dates (+- 5 days), found in the period 1981-2019.

At this point in the data processing we grouped the resulting gridded fields of daily anomalies into 5 distinct datasets or ‘folds’. The 39 summer seasons were repeatedly split into a 31 or 32 season subset for training and a 7 or 8 season subset for verification. The verification sets consist of consecutive and complete summers. Temperature evolution in one part of summer can depend on the evolution in other parts, and information leakage between training and verification sets needs to be avoided.

3.2.2 Target variables

After resampling from ERA5, the gridded t2m anomalies are at daily temporal resolution. A target at such resolution would appear unpredictable at sub-seasonal lead times. Sub-seasonal signals can only be extracted from the total variability by aggregating multiple days or even weeks (Hoskins, 2013). However, the optimal level of temporal aggregation as well as the level of spatial aggregation, is hard to establish a-priori (van Straaten et al., 2020). For summer temperatures, spatial domains smaller than the continent are preferred (Jézéquel et al., 2018), because hot spell types and their relation to sub-seasonal drivers vary across Europe (Sousa et al., 2018; Stefanon et al., 2012). Besides, sub-continental domains are also preferable from a forecast user perspective.

We obtain our sub-European target region by means of agglomerative hierarchical clustering (also used in Gómez-Orellana et al., 2023). This algorithm groups grid-cells that are similar, starting strict, with single-cell groups only, but gradually allowing more and more dissimilarity, merging those groups that comply. First we let binary timeseries indicate whether local t2m anomalies exceed the grid-cell's 95th climatological percentile. A high synchronicity between two such series, i.e. two grid cells sharing many daily exceedances, indicates that they are governed by the same regional hot spells. We therefore measure the number of non-shared daily exceedances relative to the number of shared ones, with the Jaccard dissimilarity (as in McKinnon et al., 2016). We compute it between all grid-cell pairs, and cluster them hierarchically. At the level where on average, in each cluster, 10 percent of the exceedances are shared, a suitable central-west European cluster emerged. This region is selected as target for this study (Fig. 3.2B). A similar geographic region appeared when climatological percentiles like the 66th were used to define dissimilarity.

From the spatially averaged t2m anomalies over this region, binary prediction targets can be defined. These targets have a value of 1 when the weekly (7 day), bi-weekly (15-day) or monthly (31 day) average temperature exceeds the 50th, 66th, or 90th climatological percentile (Fig. 3.2B), otherwise they are 0. The thresholds are computed for each temporal averaging window separately. In addition, we want the ML method to detect at which time scales the climate variables relate to the weekly, bi-weekly and monthly temperature target. The eight possible relation time scales are set to 1, 3, 5, 7, 11, 15, 21 and 31 days. Each time scale is defined as a rolling time window applied to the spatially averaged t2m and climate variables simultaneously, when climate variable features

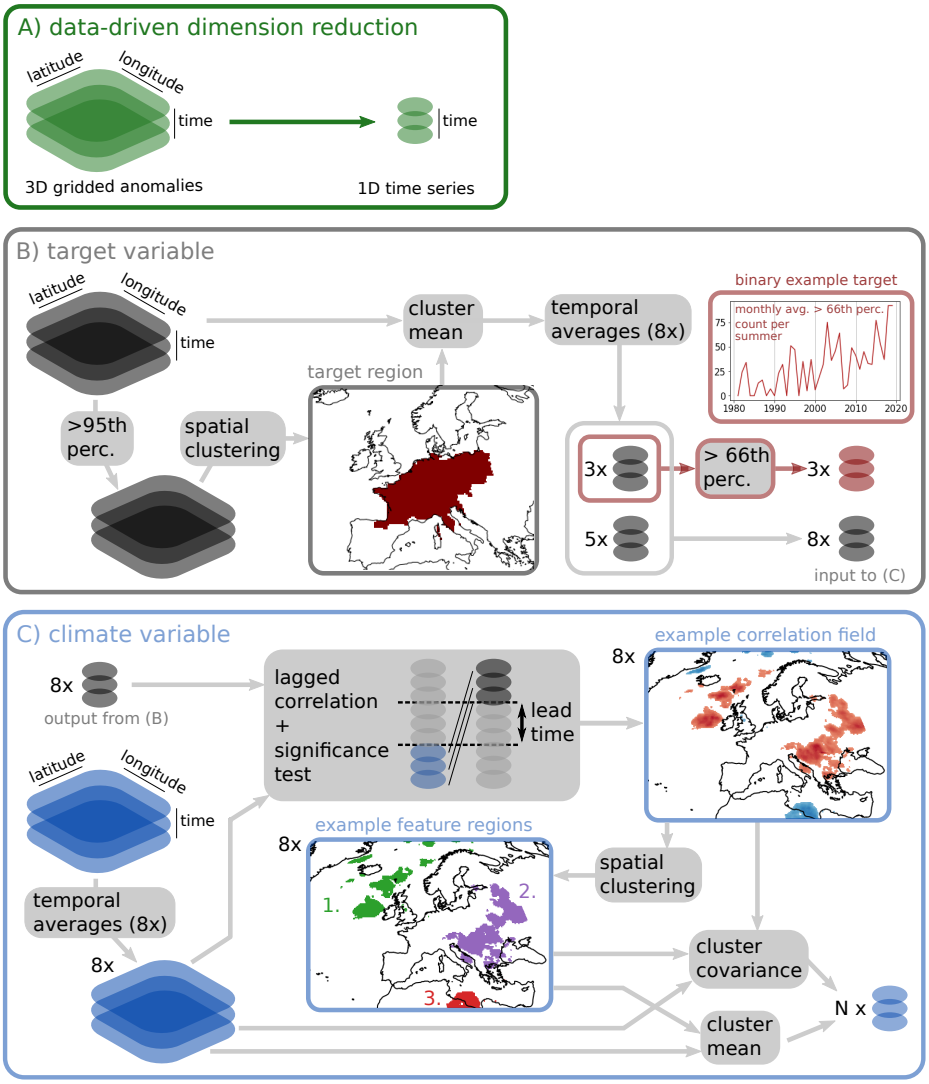


Figure 3.2: Method for data-driven feature extraction from gridded climate variables. A) Example of reducing the dimensionality of 3D gridded anomalies to 1D feature time series. B) Extraction of three target time series: average regional temperature, also averaged in time, exceeding a given threshold, resulting in the red binary series (see also section 3.2.2). Eight intermediate continuous time series, the result of averaging with windows of varying length, are used in (C). C) The extraction of potentially predictive features from a climate variable (blue, total cloud cover in this example), through correlation with temperature across a lead time gap, at eight different timescales. Multiple discovered feature regions, with two dimension reduction steps each (covariance and mean) lead to N time series (see also section 3.2.3).

belonging to the time scale get extracted (Fig. 3.2C) (section 3.2.3).

3.2.3 Data-driven features

The climate variable features related to high temperatures will likely depend on the lead time at which they are sought. We allow for this possibility without assuming it a-priori, by always supplying all nine variables on all eight time scales to the feature extraction in each lead time. Eight timescales are available because the rolling windows provide an average value each day, regardless of the applied window size. Crucially, this allows the complex ML method (described later and also fitted per lead time) to let features from multiple time scales interact. Chosen lead times mirror the eight aggregation time scales (1, 3, 5, 7, 11, 15, 21 and 31 days) and are defined as the number of days between the end of input information and the start of (predicted) output information. The rolling windows are thus applied in backward mode to the climate variables (timestamp at the end of each period) and in forward mode to the predicted two-meter temperatures (timestamp at the start of each period). Also short lead times (up to 7 days) are examined to confirm whether our method shows the expected high skill under fairly deterministic conditions.

The extraction applied per lead time is a dimensionality reduction of the gridded variable data (Fig. 3.2A), that differs from conventional tools for dimension reduction. Principal Component Analysis for instance, reduces variability to a few dominant components (e.g. Kämäräinen et al., 2019), but without guarantee that the variables projecting highly into those components contain information relevant to a particular forecasting problem (Lakshmanan et al., 2015). Theoretically, the same issue occurs when upper level (e.g. 300 hPa) geopotential is reduced to a set of regimes first, from which temperature is predicted at second instance. Although such methods have established conceptual modes of variability, like the Summer North Atlantic Oscillation (Folland et al., 2009; Bladé et al., 2012), that proved insightful for heatwave research (e.g. Cassou et al., 2005; Kueh and Lin, 2020), this study deliberately uses a dimension-reduction that ensures direct relevance to the target variable (e.g. Kretschmer et al., 2017). This choice penalizes simplicity and interpretability in terms of well-established modes, but serves the goal of discovering (potentially new) sources of predictability in nine climate variables.

First, we compute for each lead time and each climate variable grid-cell (c) a partial rank correlation $r_{partial,c}$ between the time series of the

climate variable $x_{t,c}$ and the spatial mean temperature $\bar{y}_t$ in the target region, where the time dimension for each entity is aggregated to one of the eight time scales (Fig. 3.2C, largest grey box). For example, at a lead time of 15 days we correlate 31-day cloud cover with 31-day temperature, and we correlate 1-day cloud cover with 1-day temperature. The rank correlation is called partial because we remove common drivers that can inflate the correlation (Runge et al., 2014). Seasonality was already removed in the creation of anomalies, so at each time t we use linear regression to further remove the long term climatic trend and the influence of auto-correlation from values at $t - \tau$:

$$\hat{\bar{y}}_t = \beta_1 \bar{y}_{t-\tau} + \beta_2 t \tag{3.1}$$

$$\hat{x}_{t,c} = \beta_{3,c} x_{t-\tau,c} + \beta_{4,c} t \tag{3.2}$$

$$r_{partial,c} = r(\bar{y}_t - \hat{\bar{y}}_t, x_{t,c} - \hat{x}_{t,c}), \tag{3.3}$$

where τ is the rolling window size and where $\beta_1, \beta_2, \beta_3, \beta_4$ are regression coefficients. $\beta_{3,c}$ and $\beta_{4,c}$ are estimated for each grid cell separately, in order to apply the detrending procedure for every grid cell separately.

Each correlation value is tested for significant two-sided difference from zero, with a confidence level α that becomes increasingly strict with window size: $\alpha = 5 \cdot 10^{-4-0.3(\tau-1)}$. This relation was found experimentally, and provides good correction for the increase in dependence by rolling window aggregation. The joint result of all correlation values is a correlation field (example in Fig. 3.2C) to which we apply an extra False Discovery Rate correction (Benjamini and Hochberg, 1995). This corrects for the inclusion of grid cells that are significant by chance and results in a single field per lead time, climate variable and time scale. And a single field per cross-validation fold, because we repeat the computation for each training set, as it is known that correlation patterns depend on the years in which they are computed (DelSole and Shukla, 2009; Garcia-Serrano and Frankignoul, 2014).

The number of features extracted from each field is data-driven and therefore variable. The cloud cover example (Fig. 3.2C) illustrates that multiple contiguous groups of significantly correlated cells can be present in a single field. The patterns of negatively and positively related anomalies need not arise at the same time. More likely, especially at larger distances between cell-groups, is that each group arises as an independent regional feature. We use the 'Hierachical density-based spatial clustering of applications with noise' (HDBSCAN) algorithm (McInnes et al., 2017) to identify these features. HDBSCAN is preferred over standard

Table 3.2: Parameters in feature extraction: clustering gridded correlation patterns on a sphere. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) considers three parameters. Minimum final cluster size, minimum number of samples for a sub-clusters to not be considered noise, and the spherical distance ϵ below which clusters are not split up.

variable	min. cluster size [cells]	min. samples [cells]	ϵ [radians]
t850	3000	1000	0.17
z300	6000	2000	0.17
tcc	400	200	0.09
sst	1000	300	0.20
siconc	400	200	0.10
swvl13	450	200	0.08
swvl4	450	200	0.08
transp	600	200	0.08
snowc	3000	1000	0.11

clustering methods like k-means to better handle the large grid cell density differences pertaining to the regular latitude-longitude grid (Saunders et al., 2021). Variants of HDBSCAN have also been applied by e.g. Zhang et al. (2019b) and Tilloy et al. (2022). Here we use haversine to measure the geographical distance between all grid cell pairs.

In the example HDBSCAN leads to three distinct clusters (Fig. 3.2C). We call these clusters 'feature regions' to distinguish them from the 'target region' produced by the hierarchical clustering procedure for the target. A close look at region 1 reveals how it can comprise both negatively and positively related anomalies (respectively blue and red in the correlation field). Though not necessarily the case in this cloud cover example, such dipoles are common driving features, for instance in Atlantic SST (Ossó et al., 2020), but also in the sequence of low and high pressure systems that constitute an RWP (Schubert et al., 2011; Wolf et al., 2018). In these cases we allow a larger geographical distance to exist within the cluster, such that all negative and positive constituents become a single regional feature. While the complex model could in principle learn the dependency between such constituents, we prefer to treat well-known dipole- or wave-features as one, which gives room for the model to learn unknown links. To this end, we experimentally established a specific set of HDBSCAN parameters for each climate variable (Table 3.2). These parameters are: minimum size of the feature region, minimum number of samples for a sub-region to not be considered noise, and the distance ϵ below which a

region is not split up any further. The different parameter values reflect differences in a variable's characteristic length scale, e.g. z300 being a much smoother atmospheric field than tcc, and the higher grid-cell density of ERA5-Land as compared to ERA5.

After clustering, we extract two time series per feature region. First is the mean anomaly at each timestep. Second is the covariance between the time-varying anomalies $x_{t,c}$ and the static correlation values r_c from the correlation field:

$$\text{covariance}_t = \frac{1}{n}\sum_{c=1}^{n}(x_{t,c} - \frac{1}{n}\sum_{c=1}^{n}x_{t,c})(r_c - \frac{1}{n}\sum_{c=1}^{n}r_c), \tag{3.4}$$

where n is the number of grid cells in the cluster. This covariance expresses the spatial coherence of anomalies with the corresponding correlation field. If anomalies are an exact copy (inverse) of the correlation pattern, that timestep would be assigned a high positive (negative) value. Covariance thus ensures that dipole features, whose positive and negative anomalies would cancel in a cluster mean, are extracted too. As the number of regions is data-driven, the feature extraction results in a total of N feature time series per lead time and training fold (Fig. 3.2C).

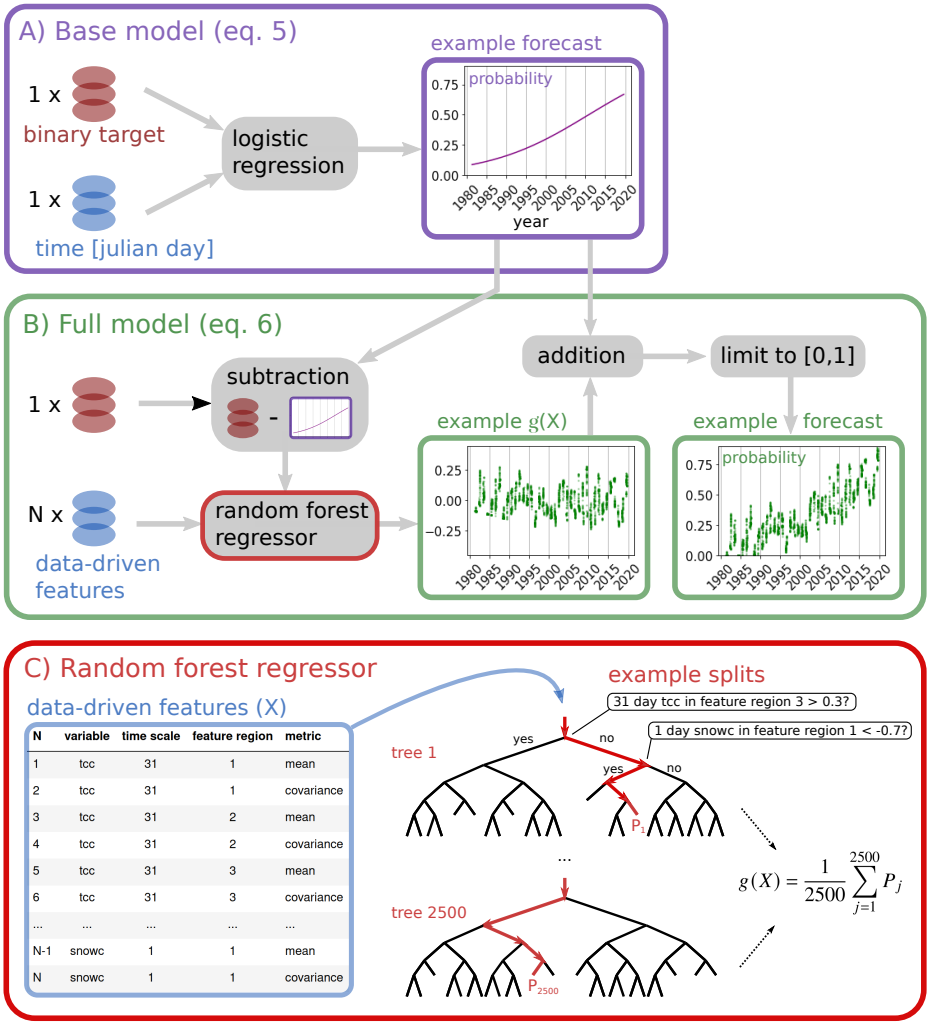
3.2.4 Machine learning model

Per lead time and cross-validation fold one ML model is fitted to predict either the weekly, bi-weekly or monthly binary target. Because of the way we defined the target, this prediction is simplified by climate change. The frequency with which temperature exceeds a fixed threshold increases with time due to thermodynamic effects (Vogel et al., 2020). Forecasts can make reliable use of this (Suarez-Gutierrez et al., 2020). Climate change is thus the most simple prediction a model can make, as it does not need to understand or leverage any other driver of high temperatures. Therefore we define our 'base' model as the Logistic Regression of climate-change driven probability p_{base} against time t (julian day):

$$p_{base} = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 t)}}, \tag{3.5}$$

where γ_0 and γ_1 are regression coefficients (see also Fig. 3.3A).

Sub-seasonal windows of predictability, being sourced from the features and interactions that we wish to research, would exist on top of this climate change signal (Hamill and Juras, 2006). Our 'full' machine



N	variable	time scale	feature region	metric
1	tcc	31	1	mean
2	tcc	31	1	covariance
3	tcc	31	2	mean
4	tcc	31	2	covariance
5	tcc	31	3	mean
6	tcc	31	3	covariance
...	...	...	...	...
N-1	snowc	1	1	mean
N	snowc	1	1	covariance

Figure 3.3: Method for fitting the sources of predictability with an ML model that forecasts a temperature exceedance target (red-brown) from climate variable features (blue, see also Fig. 3.2). (A) Base model predicting the increased probability of events due to climate change (equation (3.5)). (B) Full model predicting deviations from the base, by leveraging the large set of N data-driven features available at a single lead time (equation (3.6)). (C) Random Forest component of the full model, fitted per lead time and cross-validation fold, consisting of 2500 decision trees with a maximum depth of 5, each using a random selection of 35 features in its splits. The predicted value for $g(X)$ is the average over the trees. The table on the left illustrates some of the N available features, for instance the three cloud cover features illustrated in Fig. 3.2C. Random forest diagram after Mecikalski et al. (2021).

learning model is thus defined as the complex function $g(X)$ with which a Random Forest regressor (Breiman, 2001) lowers or heightens the base probability, by leveraging the set of N data-driven features X:

$$p_{full} = p_{base} + g(X) \quad \text{with } p \in [0, 1] \tag{3.6}$$

A combination of a Random Forest on top of a base model has been applied earlier (Kirkwood et al., 2021). However, as seen in Fig. 3.3B, the target variable of this Random Forest regressor is not fully continuous. Instead it is the residual between (trended) binary observations o and the base probability that increases with time: $o - p_{base}$. o is either 0 or 1, so the regression target is bounded between [-1,1]. This approach is not elegant, because after addition any final negative probabilities need to be transformed to 0, but it works in practice. Other reasons for this approach, and possible alternatives, are discussed in section 3.4.

The set of about $N \approx 300$ input features at each lead time, is large compared to the number of independent observations. Although daily resolution was retained by the rolling window averaging, the actual amount of non-overlapping observations ranges from about 3000 for forecasts of the 1-day average target, to about 100 for forecasts of the 31-day average target. Many statistical models are prone to over-fitting with such low observation-to-feature ratios (100/300 at worst). Random Forests however, consist of an ensemble of decision trees (Fig. 3.3C) and are suited for this task (Wei et al., 2015). Each tree uses a random subset of data and features to split its target values into collections (also 'nodes') with maximum homogeneity. Splitting rules are combinations of a feature and a threshold (see example rules in Fig. 3.3C). The trees then keep partitioning the data into smaller and smaller nodes until a stopping criterion is reached. As an ensemble, the trees converge to stable average estimates of the target, and have proven useful for meteorological applications (e.g. Taillardat et al., 2016; Whan and Schmeits, 2018; van Straaten et al., 2018; Bakker et al., 2019; Hill et al., 2020; Mecikalski et al., 2021).

Individual trees are tuned by setting their maximum tree depth, the minimum number of samples required to split a node, and the number of features considered per split. The ensemble size is determined by setting the number of trees. These hyperparamaters are usually tuned during the training process and then kept fixed when the final model is trained. For each and every lead time, we constrained the tree-depth to a maximum of 5, required a minimum number of 30 samples, set a random pick of 35 out of 300 features, and used 2500 trees (Fig. 3.3C). The first two settings are

known to limit model capacity, resulting in shallower trees, which avoids over-fitting in data sets where over-fitting is possible (Segal, 2004).

The first three hyperparameters were found by iteratively testing different combinations for performance in terms of the Brier score (section 3.2.5), measured on the verification folds. This introduces the risk of selecting hyperparameters that may over-fit the same verification folds, also used for final verification (section 3.2.5). However, this risk was reduced by accepting only combinations that display a low generalization error (similar performance in both the training and the verification set). A test of our approach with an unseen data set (a backward 1950-1979 ERA5 extension) demonstrated the robustness of found hyperparameters (not shown).

Our fourth hyperparameter, i.e. 2500 for the number of trees, was chosen for practical reasons. The random pick of 35 (out of 300) features implies a 35/300 chance to be available for one of the top-most splits in a decision tree. The first splits dominate the predictions, which means that any feature has little chance to be important in a single tree (Wei et al., 2015). To distribute feature selection probability equally, we choose a large but computationally feasible number of 2500 trees.

3.2.5 Verification

The existence of sub-seasonal predictability can be evaluated by comparing full- to base-model performance. Both models produce probability forecasts for which we compute the Brier Score (BS) over all forecast-observation pairs:

$$BS_{[full,base]} = \sum_{i=1}^{K} (o_i - p_{i,[full,base]})^2. \tag{3.7}$$

Where K is the total amount of pairs present in the 5 verification folds. The BS is then converted to a Brier Skill Score: $BSS = 1 - BS_{full}/BS_{base}$. Besides BS_{base} we also compare against the BS of the climatological frequency $\bar{o}$ observed over all data, which is a reference probability forecast that is commonly used. Uncertainty in the BSS can be large due to dependence between samples, and due to the relatively small number of positive cases when the most extreme temperature threshold (90th percentile) is used to define the target variable. We therefore repeatedly re-compute BSS on K forecast-observation pairs that are drawn at random with replacement, i.e. bootstrapping. The dependence between samples due to

rolling window aggregation would cause uncertainty to be underestimated when we draw day-by-day. Therefore we draw in consecutive blocks, with sizes ranging from 1 to 60 days.

A single verification metric is often not enough to understand performance (Gneiting et al., 2007). We complement the BSS with reliability diagrams and an evaluation of forecast value. Reliability diagrams are graphic tools to assess a forecast's reliability and resolution (Wilks, 2011). The Potential Economic Value (PEV) is the value that a forecast has to a hypothetical decision maker, compared to having no forecast available (Richardson, 2000). The user's decision problem is characterized by a cost-loss ratio c, being the cost of taking action over the potential loss if no action is taken. PEV becomes:

$$PEV = \frac{\min(c, \bar{o}) - Fc(1 - \bar{o}) + H\bar{o}(1 - c) - \bar{o}}{\min(c, \bar{o}) - c\bar{o}}, \tag{3.8}$$

where hit rate H and false alarm rate F are obtained from a contingency table after binarizing the forecast with probability thresholds 0.1, 0.3, 0.5, 0.7 and 0.9, and where $\bar{o}$ is the observed frequency of the event. We evaluate PEV for a range of cost-loss ratios between 0 and 1.

3.2.6 Explainability

Enhanced full model performance as compared to the base, can only occur when the full model has learned to leverage some of its input features as sources of predictability, either direct, or as an interaction on multiple time scales. We investigate what the model has learned in two ways, one being the permutation importance of each feature for the overall correctness of full-model predictions (Fig. 3.4A), the other being the contribution of each feature to a low or high forecast probability, as quantified by TreeSHAP, an application of SHapley Additive exPlanations specifically designed for tree-based methods (Lundberg et al., 2020) (Fig. 3.4B).

Permutation Importance quantifies the decrease in performance over all predictions when a feature is wrongly assigned (i.e. permuted) (Breiman, 2001; Lakshmanan et al., 2015). This means it results in one 'global importance' per feature, 'global' meaning 'over all samples'. We express the decrease in performance in terms of BS. These BS values depend on lead-time, so to equalize situations far-before and close-to the event, we rank importance within each model from 0 to 1 (from least important, lowest increase in BS, to most important, highest increase in BS). We permute in a repeated 'multi-pass' manner, which, as opposed to 'single-pass', can

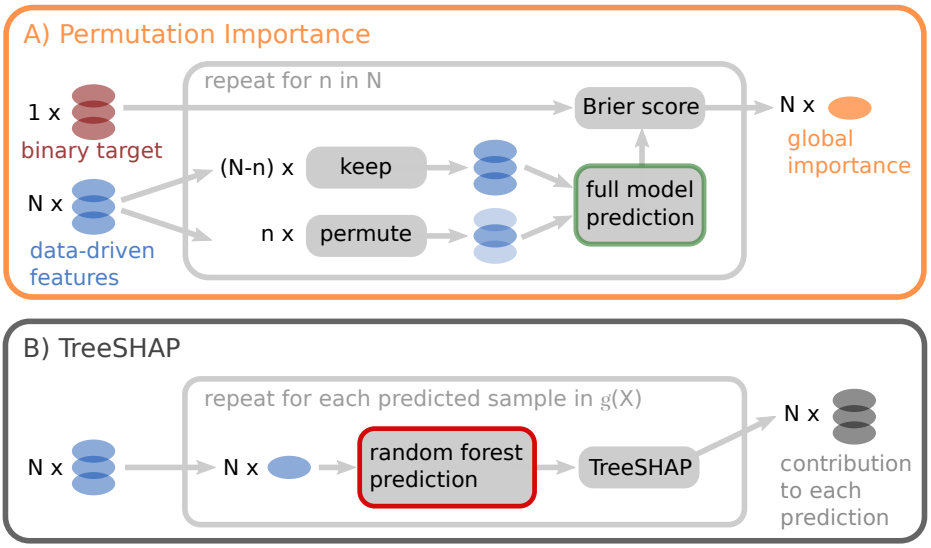


Figure 3.4: Explainability tools to interpret fitted source of predictability. (A) Explanation of the full model with Permutation Importance. Sources of predictability are identified iteratively. It are the first n features that after random re-ordering of their time series result in the worst forecast scores. (B) Explanation of the full model's Random Forest Regressor with TreeSHAP (see also section 3.2.6).

better discriminate between correlated features (Lakshmanan et al., 2015; McGovern et al., 2019). It involved iteratively searching and permuting the next-most important feature, given a set of already permuted features, until a total number of $n = 30$ features were found (Fig. 3.4A). We also found little difference in the results of 'multi-pass' and 'single-pass', when evaluated on training data.

The application of TreeSHAP to our Random Forest produces a different, 'local' measure of importance (Lundberg et al., 2020). In each sample, the Random Forest receives N feature values, and produces a single prediction $g(X)$ (see eq. (3.6) and Fig. 3.3C). TreeSHAP is a method originating from game theory that can attribute a game's single outcome to the contributions from each player. In this case it computes the set of positive and negative contributions from all features, that together add up to the predicted probability (more details in Lundberg et al., 2020). When repeated for all samples, we obtain a time series of contributions for each feature. We use these series in our case-study of summer 2015, but also extract a measure of global importance by averaging the absolute

values of all contributions in each series.

3.3 Results

3.3.1 Verification and predictability

We consider the performance of the full machine-learning model, fitted to forecast different types of target events at multiple lead times. Events are mean temperatures in a given averaging window (columns of Fig. 3.5), exceeding a given temperature threshold (rows of Fig. 3.5). Since the threshold is fixed in time, we expect a big part of the total predictability of events to come from climate change. We compare the full model BSS computed with two different reference forecasts (Fig. 3.5).

The first reference forecast is the commonly used climatological event frequency over all samples in the verification sets. It assumes that the probability of occurrence is fixed, and not low in the beginning and high towards the end, as in reality. Relative to this reference, the full model shows positive BSS at many lead times (grey shading in Fig. 3.5). At extended lead times (15 till 31 days) the monthly target shows larger BSS values than other targets, presumably because noise has been suppressed by a larger averaging window, increasing the usability of the climate change signal in forecasts (e.g. Fischer et al., 2013). We confirm the contribution of climate change to predictability with BSS relative to the base reference. This reference does model the gradual change in probability. Skill at extended lead times is hardly different from zero (green shading in Fig. 3.5), meaning that when we define sub-seasonal predictability as 'the ability to forecast deviations from the climate change signal at extended lead times', it is low.

The low amount of apparent sub-seasonal predictability can be expected. First, it is characteristic of the target region (Prodhomme et al., 2021). Second, the occasional part of predictability that exists in forecasts of opportunity, might be masked by the fact that we use all samples to compute BSS (Mariotti et al., 2020). Still, BSS does indicate a skillful lead time window in forecasting median-exceedance in the monthly target. The full model BSS is higher at a lead time of 15 days than at other, also shorter, lead times (Fig. 3.5A). Usually we expect forecasts to be more skillful for shorter lead times, as events extend less far into the future. The situation gets for instance increasingly certain from a 5- to 1-day lead time (Fig. 3.5C). But this is not the case for the monthly target (Fig. 3.5A).

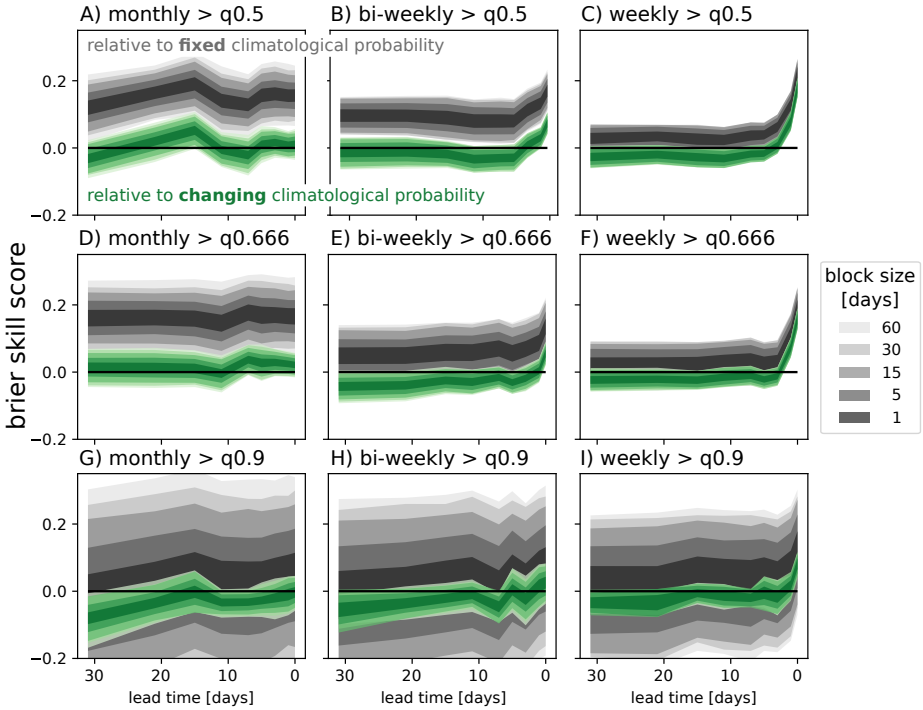


Figure 3.5: BSS-based identification of skillful lead time windows. Full model machine learning forecasts are made for different target events, being mean temperature in a given averaging window (columns), exceeding a given climatological quantile threshold (rows). Performance is measured as a function of lead time and relative to two different reference forecasts. Grey: relative to the event's fixed probability of occurrence over all samples. Green: relative to the gradual change in probability due to climate change, as forecast with the base model. Values above the zero line indicate positive skill. Uncertainty in the metric is illustrated with the 5th till 95th percentile uncertainty bounds, obtained by bootstrapping available samples with different block sizes (see legend, 5000 repeats).

Differences in skill relate to the discriminatory power of the forecasts. The skillful window at a lead time of 15 days suggests that at 15 days before the event a physical link from input features to the target can be leveraged and distinguishes high probabilities from low probabilities of monthly exceedance. In verification terms, the model would temporarily show a better 'resolution' at this lead time (for a mathematical definition of resolution see Wilks, 2011).

In Fig. 3.6 we plot reliability diagrams for the skillful monthly exceedance of the median, 66th and 90th percentile, forecast with a 15-day lead time. The higher quantiles are of interest to explore the predictability of upper tail-events with metrics that provide a more complete picture than BSS (as in Dorrington et al., 2020). The reliability diagrams show that the upper-tail forecasts of the full model outperform the base model (Fig. 3.6D,E and 3.6G,H). A reliability diagram compares the forecast probability with the observed frequency: for a forecast probability p ($0 \leq p \geq 100$ percent), the event should be observed in p percent of the cases. Forecasts that are reliable in that sense, lie on the diagonal 1:1 line. Panel D shows that the base model probabilities range from about 0.1 to 0.6 and are close to the 1:1 line. This implies that the forecasts are reliable, but not that the forecasts are perfect (which would only be realized when binary probabilities of either 0 or 1 were issued). The full model's range is wider than the base model, with for example probabilities of 0 or 0.8 being more frequently issued (Fig. 3.6D,E). Since the full model remains close to the 1:1 line and has widened the probability range, we can conclude that it has reliably increased the resolution of forecasts, on top of the climate change signal.

That the full model adds value to the base model, is visible in the vertical difference between their PEV curves (Fig. 3.6F). Base model upper tercile forecasts are valuable for decision makers with cost-loss ratios ranging from 0.1 to 0.6. The full model widens this range, and especially adds value for users with cost-loss ratios <0.2 (typical for many real-world users). Also for predictions of extremer events, namely exceedance of the 90th percentile, value is added (Fig. 3.6I). We see that the full model has learned to issue probability forecasts up to 0.6, compared to the maximum of 0.3 in the base model (Fig. 3.6H). This extension is not perfectly reliable (Fig. 3.6G shows deviations from the perfect reliability curve), but is still adding value to users with cost-loss ratios of 0.2 to 0.5 (Fig. 3.6I). The useful increase of the resolution shown in Fig. 3.6D,E also extends to monthly forecasts at different lead times (not shown). This performance improvement can not be derived from BSS values alone.

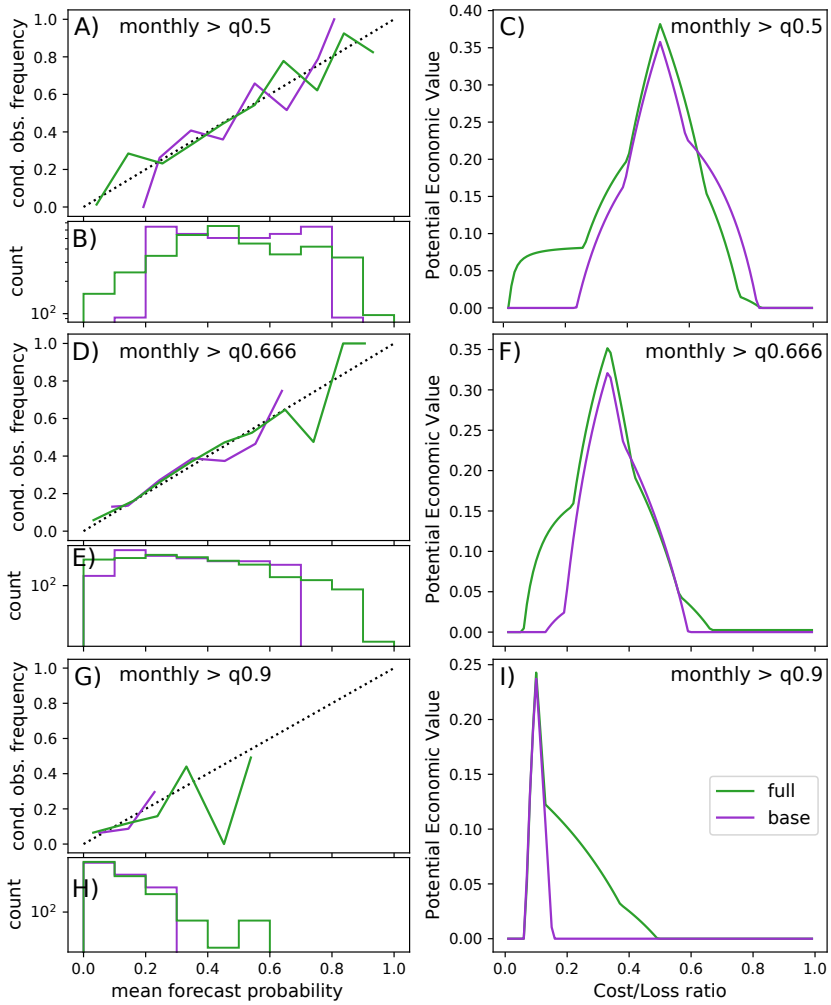


Figure 3.6: Verification of probabilistic machine learning forecasts for an identified skillful lead time window. Forecasts are for monthly temperature exceeding the 50th, 66th or 90th percentile (top, middle and bottom, respectively), made with a 15-day lead time. The full model (green) is compared to the base model which forecasts only the gradual change in probability due to climate change (purple). (A,D,G) Reliability diagrams of the conditional observed frequency per bin versus binned forecast probabilities. Perfect reliability is depicted by the dotted 1:1 line. (B,E,H) Histograms of forecast probabilities. Given good reliability in the panel above, a wider histogram shows better resolution. (C,F,I) The maximum Potential Economic Value over probability thresholds 0.1, 0.3, 0.5, 0.7 and 0.9, for users acting on these forecasts in the context of different cost-loss ratios (x-axis). Vertical difference between green and purple shows value added by the full model.

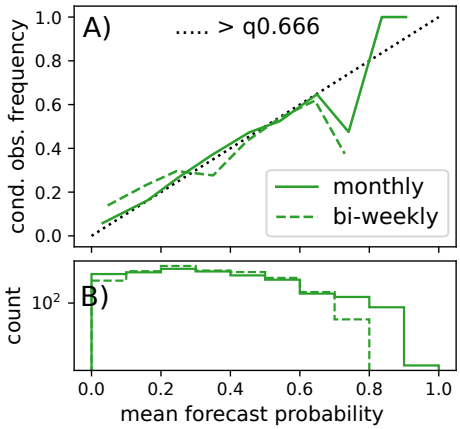


Figure 3.7: Time scale dependence for an identified skillful lead time window. As Fig. 3.6D,E, but with full model forecasts for temperature exceedance in two different averaging windows: monthly and bi-weekly (solid and dashed respectively), made with a 15-day lead time.

The ability to leverage features for forecasting is expected to not only depend on lead time but also on the properties of the target. The full model BSS at a 15-day lead time is higher for the monthly target (Fig. 3.5A) than for the bi-weekly target (Fig. 3.5B). The former event extends 46 days into the future (15-day lead time and 31 days of event), while the latter event extends 30 days into the future. Again we could expect the shortest extension into the future to be the most certain. However, in Fig. 3.7 we see that forecasts of the monthly (solid green) get closer to perfect reliability, and with a wider range of probabilities, than forecasts of the bi-weekly (dashed green), using the exact same set of features extracted on eight time scales. It needs saying that the monthly target, due to the larger averaging window, enables the full model to use a more apparent climate change signal in its base (not shown). But still the results suggest that driving features exist and are leverageable in especially this skillful lead time window. Predictability in a monthly target thus does not need to stem from successful prediction of its first two weeks.

3.3.2 Sources of predictability

The convincing resolution enhancement in the reliability diagrams at a 15-day lead time, and the moderate enhancement at other lead times, lead to a logical question: which of the features has the ML model learned to

leverage as sources of predictability? We consider this question over a range of lead times.

We expect the learned relations to depend on lead time in two ways. Physically we expect a transfer of predictability across variables, from oceanic RWP sources and antecedent land surface conditions long before the event, to the atmospheric state close to the event. Second, we expect a transfer of predictability across time scales: Multiple time scales are involved in lead up to the event, and close to the event we expect the short time scales, representing the state closest in time to the event, to become dominant in the interaction. If per the second expectation, all forecasts would be dominated by a time scale equal to the lead time, then all panels in Fig. 3.8 would look like the top-left example. Unused features are in blue and leveraged features are in orange and black (permutation importance and TreeSHAP, respectively).

We first discuss the forecasts of monthly temperature, exceeding the 66th percentile threshold (left columns, Fig. 3.8). Judging by the amount of black and orange dots, and the patterns that both metrics agree on, SST is the largest source of predictability, followed by snow-cover, 850 hPa temperature and sea-ice concentration. Of the antecedent land surface conditions, deep soil moisture is a more important source than shallow soil moisture and transpiration. Interestingly, we also see that the black dots are distributed horizontally, along the 21- and 31-day feature time scale, instead of diagonally like in the example. It means that the model keeps preferring the longer time scales despite getting closer to the event with decreasing lead time. The likely reason is that the monthly target still extends far into the future. The necessary long term information is not sufficiently present in the short term states. With t850 being an exception, this statement applies to atmospheric features at all timescales. Especially TreeSHAP shows that z300 and tcc do not contribute to forecasts of the 31-day target (almost no black dots in the left atmospheric column of Fig.3.8).

This changes when we look at predictions of the shorter weekly target (right columns, Fig. 3.8). SST, snow-cover and deep soil moisture still are dominant sources of predictability at lead times longer than 5 days. The model however does not use the 21- and 31-day feature time scales exclusively. The black and largest orange dots now lie in the window range of 10 to 31 days for SST, 5 to 21 days for snow-cover and deep soil moisture. At lead times shorter than 5 days, when the event gets closer, features from z300, followed by t850 and transpiration, become the dominant sources of predictability: they increase or decrease the chance of

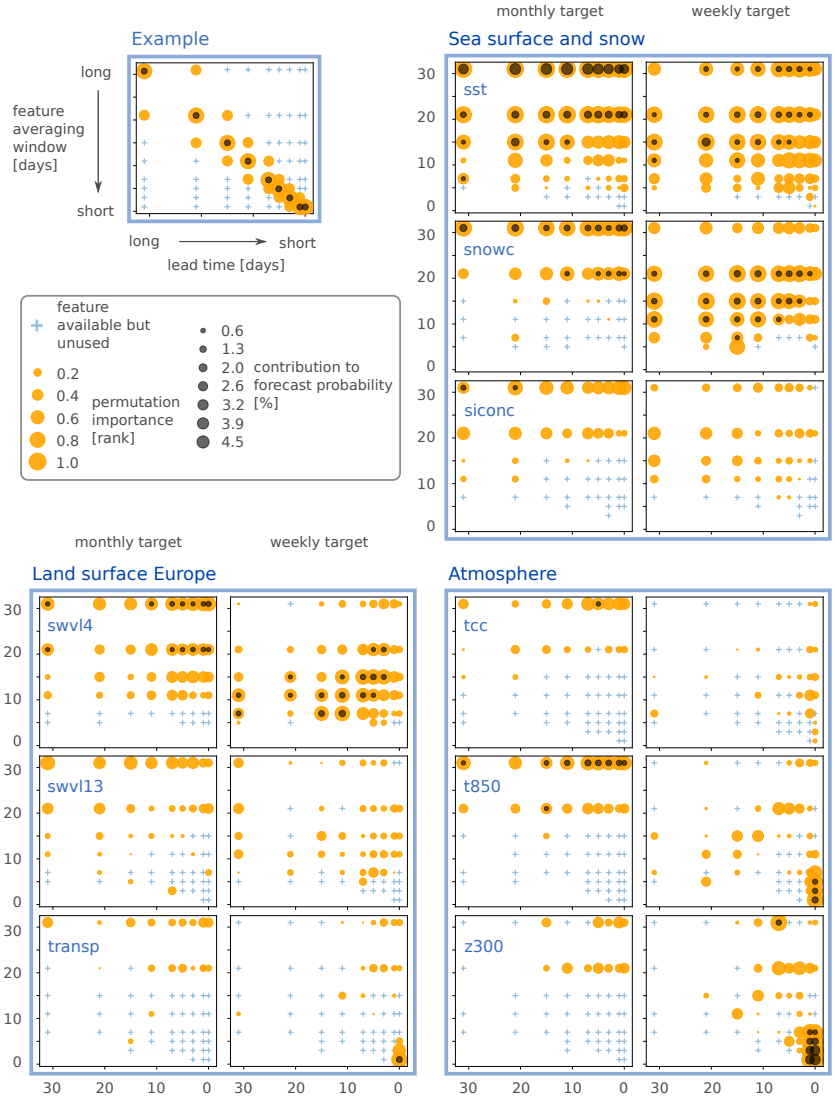


Figure 3.8: Sources of predictability learned by the full model. Features from nine climate variables are available on eight time scales, depending on lead time. Their importance is revealed by two metrics: average absolute TreeSHAP (black) and multi-pass permutation importance (orange). Unused features are in blue. The emphasized time scales (y-axis) and climate variables (rows) vary with the lead time at which the forecast model operates (x-axis) and the type of forecast target. The two targets are: 66th percentile exceedance in monthly temperatures (left column) and weekly temperatures (right column). TreeSHAP and permutation importance values are the maximum over all features per time scale and climate variable (comprising multiple time series of mean and covariance from multiple feature regions). TreeSHAP contributions below 0.5 percent are not plotted.

weekly high temperature events. In other words, the ML-models confirm, by being able to pick and learn freely, that information transfers from longer term oceanic and land surface conditions to the atmospheric state.

Before further physical interpretation, we consider the difference between the two importance measures. Usually, the highest ranking variables according to permutation importance, i.e. the largest orange dots, are also important according to TreeSHAP and accordingly accompanied by a black dot (Fig. 3.8). The two methods often do not agree on lower ranking variables. It appears from Fig. 3.8 that average absolute TreeSHAP is a stricter measure of global importance, whereas permutation importance admits a more wide-spread set of important features and time scales. Part of the reason lies in the difference between rank and TreeSHAP contribution. A feature that is conditionally the n-most important variable, might not contribute noticeably to forecast probability in each and every sample. Qualitative differences in emphasized features can also follow from permutation importance itself. As permutation breaks dependencies between features, it can move the model to situations it was not trained for, leaving one to interpret extrapolation behavior instead of normal predictive links (Hooker and Mentch, 2019). Consequentially, we have most confidence in the patterns that both measures agree on.

The visualized importances in Fig. 3.8, like the surprising usability of 31-day average atmospheric t850, are linked to specific regions. Each dot is namely the maximum global importance of the multiple possible feature regions with two time series each, i.e. one for the mean and one for the covariance, that were all at the full model's disposal.

So in Fig. 3.9 we map where the 31-day average input features of SST, t850, snowc, and siconc are important for predicting monthly temperature exceedance of the 66th percentile threshold at a 15-day lead. Features from SST are present in a large portion of its domain, as many grid-cells correlate significantly to temperature, even after auto-correlation and linear trends have been accounted for (Fig. 3.9A). Only a selection of these cells are robust and present in at least 4 of the 5 sets of training data (orange to yellow). Robust groups of cells for instance appear east of the Maritime Continent, suggesting that Pacific SST variability in that region can be important. However, such a potential relation does not imply that a feature will be a useful source of predictability. So for SST and the other variables (in rows), we plot each feature's permutation and TreeSHAP importance, for the robust cells only, averaged over at least 4 of the 5 cross-validation sets (Fig. 3.9, middle and right column). Consequentially, the plotted shapes do not perfectly resemble the feature

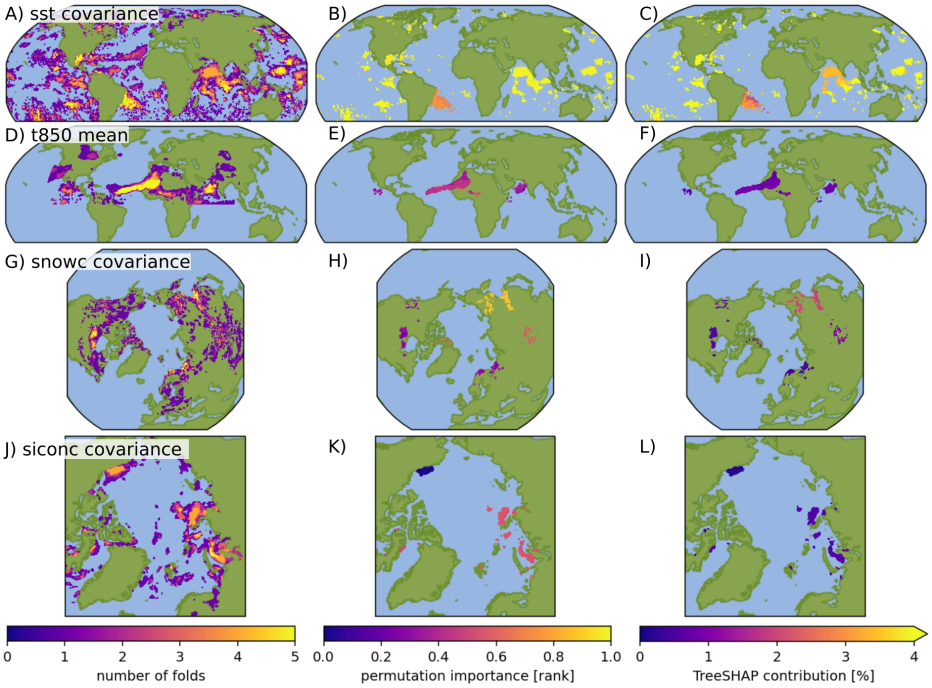


Figure 3.9: Geographic distribution of sources of predictability, learned by the full model. Visible are the feature regions in SST, T850, snow cover and sea-ice concentration (top to bottom) that the model leverages to predict monthly average temperature exceedance with a lead time of 15 days (threshold is q0.66). The climate variables' grid-cells, belonging to distinct regional features, are colored with each feature's average permutation importance and with the average absolute TreeSHAP contribution to the forecast probability (middle and right column respectively). The importance values are presented only for gridcells that are found significant in at least 4 of the 5 cross-validation folds (left column). The type of time-series used by the model is annotated (mean or covariance). Time scale of the features is also monthly.

regions of a single subset.

Features from SST (Fig. 3.9B) are stronger sources of predictability than features of other variables (Fig. 3.9E,H,K). But not all SST cells will be of equal relevance. Especially smaller patches are often part of a feature region like the whole Indian Ocean, and therefore share the importance of a predictive signal coming from large patches that dominate a feature's mean or covariance. Focusing for that reason on large contiguous patches, we notice that besides the mentioned Pacific features, the Indian Ocean provides an extensive source of predictability, and that a lesser source of predictability lies off the east coast of South America (Fig. 3.9B,C).

One of T850's predictive features is co-located with an SST feature over the Indian Ocean (Fig. 3.9E,F). Their coincidence off India's west coast hints at monsoon dynamics, which previously have been found to affect Euro-Atlantic summer circulation (Beverley et al., 2019). The most surprising T850 feature is the region extending from the tropical Atlantic into the Western Sahara. In T850, the feature mean is more predictive than feature covariance, which in combination with climate change, gives the impression that the feature is leveraged to explain a thermodynamic trend. However, our full model learns deviations from the climate trend (i.e. deviations from the base model), so this is likely not the case. Without a causal framework it remains speculative how this source of predictability, present at the relatively low level of 850 hPa, affects western and central European temperature. Given its location a link to upper-level disturbances in the tropical Atlantic and Sahel region is expected (Cassou et al., 2005; Nakanishi et al., 2021). Crucially, it would not have been discovered, had we not applied a data-driven dimension reduction and let our ML-model process many features with little a-priori selection.

The location of snow cover features differs from fold to fold. This is visible in the minor extent to which they overlap (Fig. 3.9G). Still, these are important sources of predictability (Fig. 3.9H,I). Importance of Eurasian snow cover anomalies trumps that of the North-American ones (Hall et al., 2017). And within Eurasia the regions located farther east are more important than those located in the west. The most eastern snow cover feature is also more important than any of the features in sea-ice concentration (Fig. 3.9H,K), which is the reverse of the order suggested by Zhang et al. (2020). Nonetheless, within sea ice concentration, the ML-model has diagnosed the relative importance of the Kara Sea, whose decreased ice concentration is known to relate to a more southern and stronger polar front jet (Hall et al., 2017), inhibiting stagnant high pressure systems over the target region.

3.3.3 Summer of 2015

Physical interpretation of sources of predictability remains limited when, like above, emphasized features are seen, but the sign of their predictive relations are not. One needs to know whether a specific emerging anomaly inhibits or increases the likelihood of an event, by how much, and what the state of the other features is, since links can be conditional. This information is usually hard to extract from complex ML models. With so-called 'local' explanations of individual forecasts we can uncover such details (see also Lundberg et al., 2020; Gibson et al., 2021). When a feature's TreeSHAP value shows increased forecast contribution from one point in time to the next, we can trace that in real-time to a specific set of shifting anomalies, that for instance has started to resemble the feature's underlying correlation pattern. We demonstrate such a breakdown of contributions with forecasts of the hot summer in 2015.

The summer of 2015 was characterized by exceptional temperatures in western and central Europe, that were clustered in two intense periods. Duchez et al. (2016) found that a cold North Atlantic SST anomaly from the months before might have displaced the sub-tropical jet to a stationary southern position, favoring buildup of heat over the continent. They find that the displacement commenced in the last days of June and persisted into September. High temperatures followed on 1-6 July. Despite the claim that the temperatures were driven by the cold SST anomaly, this first period was hard to predict by an operational ensemble (Ardilouze et al., 2017b). The high temperatures were followed by a rainy intermission, before a second temperature peak started on 16 July, for which Wehrli et al. (2019) found that SST was of low importance. Their analysis also showed that although soil moisture (on a monthly scale) was decreased during the event, no significant feedback to atmospheric heat was found.

For both peak periods, we examine the features contributing most to the probability forecasts of the ML-model. A horizontal bar (Fig. 3.10A) shows contributions to the probability that monthly temperature exceeded the 66th percentile, from June 6 till July 6, a period that encompasses the first heatwave period at its end. Below is the forecast that encompasses the second period at its beginning (Fig. 3.10F). The forecasts were made with a 15-day lead-time, which was shown to have predictive value in Section 3.3.1.

For both periods the full model raises forecast probability above the 59 percent that we expect from the climate change base model (Fig 3.10A,F).

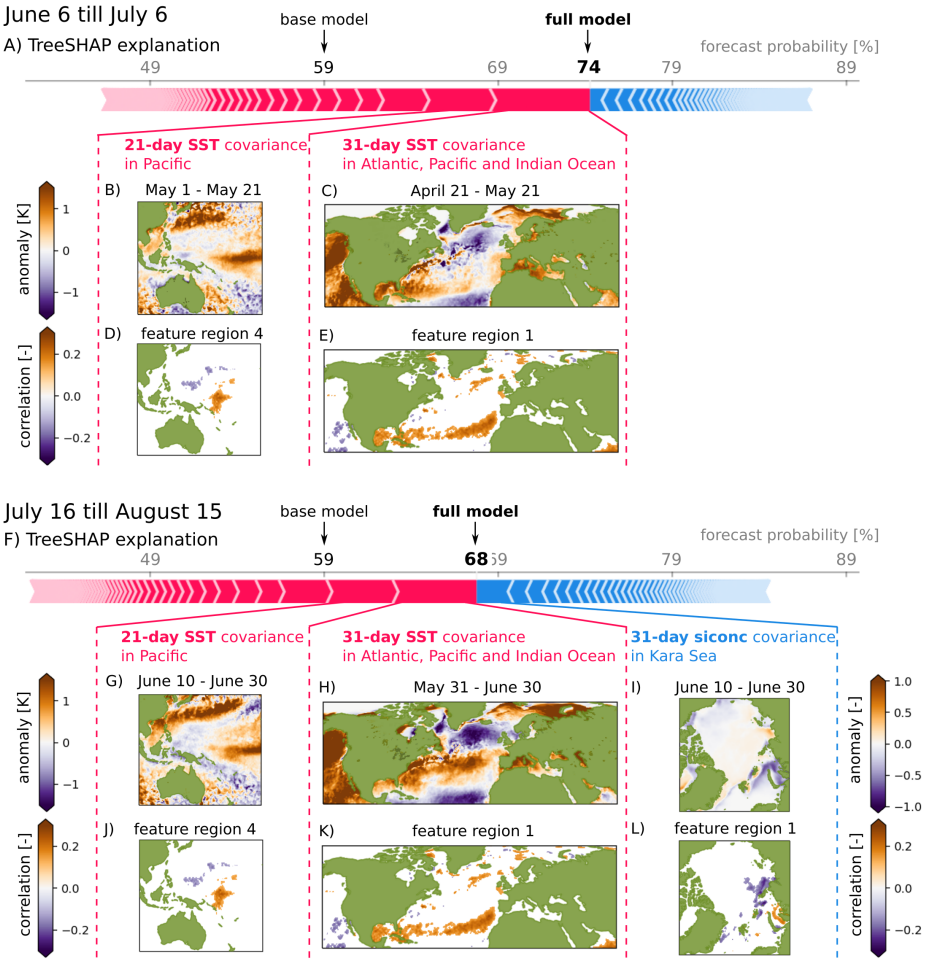


Figure 3.10: Summer 2015. Explaining the forecast probability that monthly temperature exceeds the 0.66 quantile, from June 6 till July 6, and from July 16 till August 15, made with a lead time of 15 days. (A,F) TreeSHAP explanation of the full model forecast, i.e. the decomposition of the forecast probability (bold) into additive contributions from all driving features. Features increasing the probability are in pink, features decreasing the probability are in blue. The type (i.e. covariance) and time scale of the largest contributors are annotated. (B,C) Anomalies leading to the two largest contributions. (D,E) Correlation patterns underlying the contributing features. (G-I) as B,C but including the largest negative contributor. (J-L) as D,E but including the largest negative contributor.

The increase occurs because the joint state of all driving features produces larger positive contributions (pink), than negative inhibitory contributions (blue, Fig. 3.10A,F). In both Fig. 3.10A and F the largest positive contributions come from the state of 21-day average SST, in feature region 4, and from 31-day average SST, in feature region 1. This suggests the influence of a long term SST anomaly (Duchez et al., 2016). However, region 1, where the 31-day covariance signal originates, encompasses more than just the Atlantic. When we look at the respective SST anomaly (Fig. 3.10C) we see the cold North Atlantic temperatures south of Iceland. A close look at the underlying correlation pattern (Fig. 3.10E) shows that at this location, no significantly negative, even slightly positive SST anomalies are associated to higher target temperatures, for a driving effect at this lead time. The positive contribution of covariance (resemblance to an underlying correlation pattern), has thus to be sought elsewhere, for instance in the positive anomalies extending from the Gulf of Mexico eastward. Not shown for the 31-day state, but shown for the 21-day state, is the feature east of the Maritime Continent (Fig. 3.10B). The correlation pattern here (Fig. 3.10D), in SST's feature region 4, is partly resembled by the SST at this point in time. Seeing the model put forward such an emergent feature from a relatively small region, human forecasters can study its trustworthiness. The influence of the feature does remain the 2nd largest factor also for the second part of the summer (Fig. 3.10G,J)

What does change from the first to the second high temperature period, according to our model, is the strength of inhibiting drivers, that together decrease forecast probability from 0.74 to 0.68 (Fig. 3.10A to F). The state of 31-day average sea ice concentration is dominant in this. We see that during the first predicted period (June 10 till June 30), a negative sea ice concentration anomaly was present in the Kara Sea (Fig. 3.10I) which, as discussed in the last section, can project negatively on the target temperatures (Fig. 3.10L). Accordingly the full model lowered probability slightly, though still kept it elevated with respect to the base model. Exceedence of the threshold did happen in both high temperature periods of the summer of 2015.

3.4 Discussion and conclusion

It is understood that there is a large number of climate variables and time scales involved in lead up to high summer temperatures. In this

study we have extracted data-driven features from nine variables, varying on eight time scales. Interactions in such a set can source sub-seasonal predictability, but often at a complexity level that is beyond direct human understanding. We have explored whether an ML-model can integrate and discover the sources of predictability for us. To our knowledge this has been the first attempt using such a large set of features.

Relative to the amount of features, the ERA5 reanalysis provides a low amount of samples. Whereas the use of ERA5 helps to account for processes that are inadequately captured in numerical simulations, a full integration of important sub-seasonal interactions is a great challenge for machine learning models (He et al., 2021). Combined with the limited samples to train on, a machine learning model does not attain the maximum skill possible. In certain lead time ranges, operational numerical models in combination with statistical post-processing, will probably do better (e.g. Ferrone et al., 2017; van Straaten et al., 2020). But as this study has demonstrated, even purely statistical forecasts for west and central Europe, can be reliable and valuable in certain windows, like forecasting exceedance of monthly summer temperature with a lead-time of 15 days (Fig. 3.6). We have further shown that this long-term predictability is not due to successful prediction of its short-term (two-week) constituents (Fig. 3.7).

Limited data makes certain types of statistical association hard to estimate. In feature extraction we correlated climate variables with temperatures (Fig. 3.2C). This appears sub-optimal, as the features are later supposed to predict only the upper tail of the temperature distribution, namely binary exceedance of a relatively high threshold. Rank correlation remains usable because non-linear associations specific to the upper tail are accounted for, as long as they are monotonic. Better suited quantities like extremal-dependence metrics (Coles et al., 1999) and Kendall's Tau weighted towards the tail of the distribution, were also tried, but their estimates were found to be too unstable for this amount of data. An interesting alternative to our correlation-based dimension reduction is to directly apply a predictive ML model to the raw input data (e.g. He et al., 2021). In our setting with large domains, and nine climate variables for eight time scales, that will be challenging, as even after dimension reduction a low observation-to-feature ratio was obtained.

A suitable ML model to identify sources of predictability in a set with a low observation-to-feature ratio is a Random Forest (Wei et al., 2015). The resulting picture of importance might however be more diffuse than the importance in reality. We had to mitigate the large number of

features by drawing 35 at random. This can make single dominant sources of predictability (e.g. from 31-day SST) un-available at prime splits in the decision trees, a role that is then taken by correlated features (e.g. from 21-day SST). This is also the reason that TreeSHAP contributions have small values. After computation of contributions, even the most important source of predictability alters the forecast probabilities by a mere 5 percent on average (legend of Fig. 3.8). Clear importance patterns have nevertheless emerged. The importance of long term variability in SST is in line with physical understanding. Surprising is the dominance of deep soil moisture over shallow soil moisture, at all targets and time scales, and that of snow cover over sea ice.

An inherent challenge for sub-seasonal forecasts is the non-stationarity of driving features. For sea-ice it is known that a link to European temperature can exist at certain moments in time (Kolstad and Årthun, 2018). When such a potential for predictability is not systematic over all possible samples, it can happen that empirical models are trained on a subset with its presence and will make false forecasts on a set without. TreeSHAP can show in real-time whether such features keep contributing to the forecasts (Fig. 3.10). A later computation of permutation importance can then reveal whether the important TreeSHAP contributions were valid or not. This needs a series of observed outcomes first, but features that have already started to degrade the scores will show no importance after further permutation. We expect that differences between TreeSHAP and permutation importance (Fig. 3.8) might inform one about potential non-stationarity.

We have treated sub-seasonal predictability as 'deviations from a changing climatology that are predictable at extended lead-times'. As in other studies (Dole et al., 2014; Prodhomme et al., 2021), this separation from the trend is influential (Fig. 3.5), because climate change provides a large part of total predictability, and is even valuable to certain users (Fig. 3.6). Our first attempt at separation involved detrending average temperature before creating the binary (threshold-exceedance) target. This distributes events uniformly over time, and thus supposes that moderate temperatures of the past and current more extreme temperatures comprise a homogeneous class with similar dynamics. Depending on the definition of extremity, this is likely not the case (Vogel et al., 2020). It led to bad performance in the last decade. We therefore modeled the probabilistic deviations as additions to a base model (Eq. 3.6; Fig. 3.3B). More elegant tools than subtraction and addition exist, like logarithmic transformations (e.g. Scheuerer et al., 2020) and Bayesian methods. But

unfortunately those were not (yet) compatible with the Random Forests we needed to handle the low observation-to-feature ratio.

There is good reason to keep applying complex ML-methods to subseasonal prediction. First, we have shown that such a method can reliably increase forecast resolution by leveraging features from reanalysis data. Second, we demonstrated that an explainable method gives conceptual grip on the complexity. It confirmed many physical expectations, like the weighing of time scales and the transfer of information across variables with lead time. It also discovered surprising features like the long term predictability originating from 850 hPa temperature. Overall, the associative learning of an ML-method will complement research with NWP models (e.g. Quinting and Vitart, 2019). It shows which links need further understanding, and which variables need correct representation in NWP models for better future forecasts.

Chapter 4

Correcting numerical forecast errors with ML

Published as:
van Straaten, C., Whan, K., Coumou, D., van den Hurk, B. and Schmeits, M. (2023) Correcting sub-seasonal forecast errors with an explainable ANN to understand misrepresented sources of predictability of European summer temperatures *Artificial Intelligence for the Earth Systems*

Abstract

Sub-seasonal forecasts are challenging for numerical weather prediction (NWP) and machine learning models alike. Forecasting two-meter temperature (t2m) with a lead-time of two or more weeks requires a forward model to integrate multiple complex interactions, like oceanic and land surface conditions leading to predictable weather patterns. NWP models represent these interactions imperfectly, meaning that in certain conditions, errors accumulate and model predictability deviates from real predictability, often for poorly understood reasons. To advance that understanding, this paper corrects conditional errors in NWP forecasts with an artificial neural network (ANN). The ANN post-processes ECMWF extended-range summer temperature forecasts by learning to correct the ECMWF-predicted probability that monthly t2m in western and central Europe exceeds the climatological median. Predictors are objectively selected from ECMWF forecasts themselves, and from states at initialization, i.e. the ERA5 reanalysis. The latter allows the ANN to account for sources of predictability that are biased in the NWP model itself. We

attribute ANN-corrections with two explainable AI tools. This reveals that certain erroneous forecasts relate to tropical west Pacific sea surface temperatures at initialization. We conjecture that the atmospheric teleconnection following this source of predictability is imperfectly represented by the ECMWF model. Correcting the associated conditional errors with the ANN improves forecast skill.

4.1 Introduction

Sub-seasonal to seasonal (S2S) forecasts are made with a lead time of more than two weeks. S2S forecasts of variables like atmospheric temperature and precipitation are crucial in the anticipation of heatwaves and droughts (White et al., 2021). Often however, the detailed temporal evolution of temperature and precipitation at lead times beyond two weeks is not predictable (e.g. Buizza and Leutbecher, 2015; Zhang et al., 2019a), as atmospheric motion is sensitive to small variations in initial conditions (Lorenz, 1963). In these cases ensemble members of numerical weather prediction (NWP) models show widely diverging possible states (Leutbecher and Palmer, 2008).

Only in certain conditions can low-frequency internal variability of the atmosphere, and interaction with persistent Earth system components create so-called 'windows of predictability' (Mariotti et al., 2020). In such windows, a source of sub-seasonal predictability constrains the range of states that the atmosphere can visit (Palmer, 1993; Toth and Buizza, 2019). Figure 4.1A illustrates how atmospheric and oceanic sources of predictability can steer the probability of an event at a valid time more than two weeks into the future. Mandatory is that the sources of predictability are adequately represented in the NWP model (Fig. 4.1A).

Certain sources of sub-seasonal predictability are particularly important to represent. For instance oceanic influences, as Sea Surface Temperature (SST) patterns are known to interact with, and steer the position of the North Atlantic jet stream in summer (Ossó et al., 2020; Osborne et al., 2020; Carvalho-Oliveira et al., 2022). Another example is land surface drought. Its feedback on the atmosphere during high pressure 'blocking' systems (Kautz et al., 2021) makes high two-meter air temperatures (t2m) more likely (Quesada et al., 2012).

The existence of such time-dependent sources of predictability can in theory lead to better S2S forecasts of heatwaves and droughts (Hoskins, 2013). In practice this predictability is not achieved because NWP models

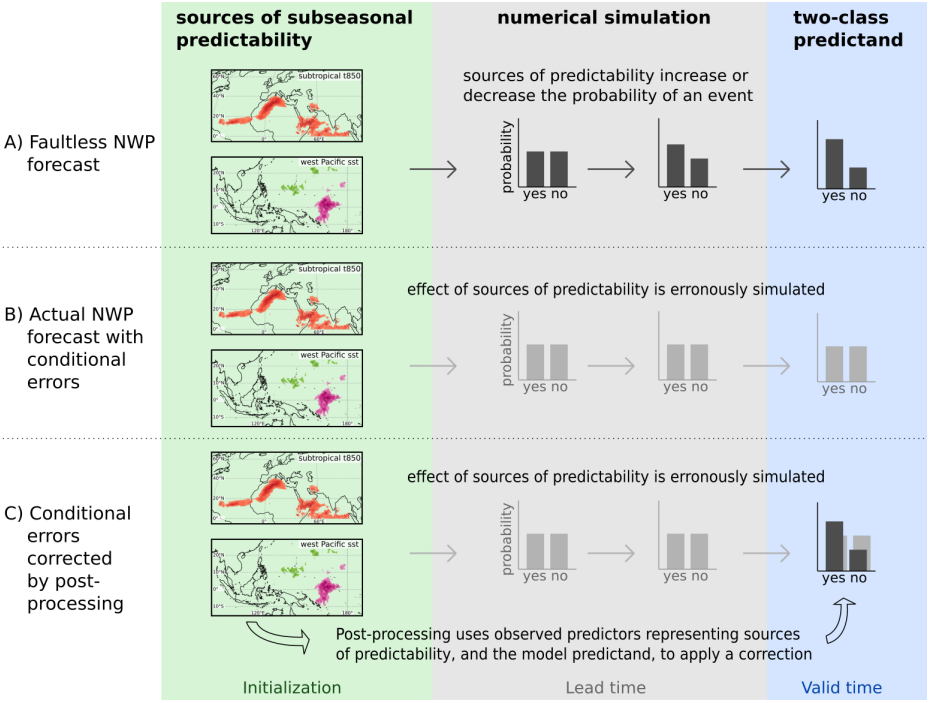


Figure 4.1: Modeling time-dependent sources of predictability of an event-based, two-class (yes/no) predictand. (A) An ideal NWP model perfectly represents the sources of predictability and correctly simulates their effect from initialization until valid time. (B) Due to imperfect representations, actual NWP models make conditional errors. This generates errors in simulated states and results in erroneous predictions of the predictand (light grey event probabilities at valid time). (C) As a solution, statistical post-processing corrects the predictand after the model has run, utilizing the predictand itself and observed predictors that represent sources of predictability in their initial state. Note: the displayed sources of subseasonal predictability are examples of variables found to be important for forecasts of summer two-meter temperature in West and Central Europe (discussed later in this study).

are imperfect. A first reason is the mentioned sensitivity to initial conditions in the atmosphere and in other Earth system components. These are hard to initialize correctly (Merryfield et al., 2020), meaning that errors will grow and transfer to S2S time scales (Lorenz, 1969). A second reason is that physical interactions need to be approximated numerically. The representations of sources of predictability are therefore imperfect, and lead to conditional errors in the forecasts. Figure 4.1B illustrates how in an actual NWP model, errors from misrepresentations can affect simulated states and the predictand. Any sub-seasonal steering of the event-probability at valid time, that was present in Fig. 4.1A, becomes biased or absent.

As an example, we know that the prediction of t2m in Europe in the ECMWF model is affected by too low model soil moisture in spring and summer, which has consequences for the heating of the atmosphere (Dutra et al., 2021). We also know that the decay of Rossby Wave Packets (RWPs) above Europe is underestimated, resulting in too infrequent blocking (He et al., 2019; Quinting and Vitart, 2019). The errors of NWP models can thus relate to highly specific conditions, like the arrival of an RWP or a soil moisture deficit in summer. This means that such conditional sources of predictability can potentially also predict the errors that their misrepresentation will result in. In this study we hypothesize that conditional errors can, in principle, be predictable, correctable, and better understood, when we relate them to the involved sources of predictability, as it was their imperfect representation that resulted in the errors.

The idea of correcting errors is not new. Predictable NWP errors have been corrected with statistical methods since the 1970's (Glahn and Lowry, 1972). In so-called 'statistical post-processing' the estimate of a predictand of interest (in our case European t2m, valid two weeks after the forecast is made) gets corrected after the NWP model has run (see Haupt et al., 2021; Vannitsem et al., 2021, for current reviews of statistical post-processing). As S2S forecasts have attracted attention only recently (Vitart and Robertson, 2018), most S2S post-processing has been simple, correcting only unconditional errors, i.e. applying the same (e.g. additive) bias correction to each and every forecast (Ferrone et al., 2017; Vigaud et al., 2017; Monhart et al., 2018; van Straaten et al., 2020; Graham et al., 2022).

To correct conditional errors, statistical post-processing needs to be 'weather-dependent'. This is generally achieved with predictors representing different weather conditions, and with methods from Machine Learning (ML) that are capable of modeling non-linear relations (e.g.

Allen et al., 2019; Schulz and Lerch, 2022; Hewson and Pillosu, 2021). Post-processing of S2S forecasts with ML often employs predictors representing large-scale patterns, such as upper-level geopotential height (in case of Scheuerer et al., 2020; Fan et al., 2021), the El Niño Southern Oscillation (ENSO) (in case of Strazzo et al., 2019; Specq and Batté, 2020), or large-scale atmospheric circulation regimes (in case of Manzanas et al., 2018; Lavaysse et al., 2018; Richardson et al., 2020; Mastrantonas et al., 2022). Such predictors represent sources of predictability for a predictand defined at the surface.

The employed predictors are often model predictors, derived from NWP forecasts. For instance, simulated large-scale circulation is used to correct simulated t2m, both more than two weeks into the future. Since the predictor also comes from the NWP forecast, such a correction becomes problematic if conditional errors affect both predictor and predictand. In these cases a better predictor might be found in the initial state of the atmosphere, ocean and/or land. Figure 4.1C illustrates how such observed predictors could be used to correct the predictand at valid time.

In this study we present a statistical post-processing method that uses both observed predictors at initialization, and model predictors at valid time. Relating predictors at initialization to errors at valid time is a novelty compared to many recent post-processing studies. Most studies feed their ML-based correction method with forecast states only (Rasp and Lerch, 2018; Strazzo et al., 2019; Scheuerer et al., 2020; Fan et al., 2021; Veldkamp et al., 2021). An exception is Grönquist et al. (2021) who used states at initialization and +24 hours to correct a +48 hr forecast. Such an approach has not been applied to correcting conditional forecast errors at S2S lead times. In this paper we develop an artificial neural network (ANN) that relates predictors from initialization time to errors at valid time, to correct and understand conditional errors in sub-seasonal ECMWF forecasts of 2m temperature in Europe during summer. Our ANN (Section 4.3.1) is based on the architecture proposed by Scheuerer et al. (2020). We adapt the architecture to directly relate learned conditional corrections to responsible predictors, using eXplainable AI (XAI). These are tools used for interpreting complex ML models (Mueller et al., 2019; McGovern et al., 2019; Arrieta et al., 2020; Molnar et al., 2020; Toms et al., 2020; Clare et al., 2022), and have been succesfully applied in the S2S range (Mayer and Barnes, 2021; Gibson et al., 2021; van Straaten et al., 2022).

Relating each type of conditional error to predictors with XAI, be-

comes especially potent in combination with observed predictors. When predictors from the initial state are found to be most predictive of accumulated NWP model errors at valid time, we can deduce that the sources of predictability that they stand for, are crucial, but imperfectly represented in the NWP model. In addition to improving skill, the post-processing method presented here can help us understand conditional NWP errors and advance the numerical model representation of poorly understood sources of sub-seasonal predictability (Vitart and Robertson, 2018; Vitart et al., 2019; Merryfield et al., 2020).

4.2 Data

We post-process NWP forecasts from ECMWF's Integrated Forecasting System. A single model version, namely 45r1, is used to avoid changes in systematic model error that occur when different versions are utilized (Bauer et al., 2015). Cycle 45r1 ran operationally from the 5th of June 2018 to the 10th of June 2019, and produced a 51-member, 46-day ensemble forecast every Monday and Thursday. Each starting date ECMWF also produced 20 years of re-forecasts, by starting additional runs from the same date, but in the 20 previous years. These additional years can be used for investigating model errors and fitting a statistical post-processing method. As in the operational setting, re-forecasts are initialized from an analysis, in this case a reanalysis. The control member starts from the analysis, and ten other members start from slightly perturbed initial states, according to the estimated initial condition uncertainty (Leutbecher and Palmer, 2008). At any valid time the 11-member ensemble forms a sample from the distribution of possible future states. We merge one year of forecasts (after extracting the control member and 10 randomly sampled members) with 21 years of re-forecasts, to obtain a dataset spanning 22 years, from 1998 till 2019. We focus on summer only (June-July-August (JJA)) because we want to improve forecasts of anomalously warm conditions during that season.

The model predictand, i.e. the ECMWF forecast that will be post-processed, is derived from gridded daily two-meter temperature (t2m). These forecasts are retrieved in a domain encompassing Europe and the North Atlantic, from 20 to 80 degrees north and from -90 to 30 degrees east, at a spatial resolution of 0.32x0.32°.

As data for model predictors we retrieve forecasts of four variables that represent different sources of predictability, in the same domain as

above, at a resolution of 1.5x1.5°. We retrieve geopotential at 300 hPa (z300) as representation of the high-pressure 'blocking' systems and quasi-stationary Rossby Waves that relate to surface temperature extremes (Schaller et al., 2018; Wolf et al., 2018; Kautz et al., 2021). We extract sea surface temperature (SST) because of its potential to influence the North Atlantic jet stream in summer (e.g. Ossó et al., 2020). Lastly we extract shallow soil moisture in the top-three model-layers (0-100 cm) (swvl13) and deep soil moisture from layer four (100-289 cm) (swvl4). Both can influence t2m through the surface energy and water balance (Seneviratne et al., 2010; Miralles et al., 2019).

The data for the observed predictors are retrieved from the ERA5 and ERA5-Land reanalyses (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). These reanalyses are based on assimilated observations and closely correspond to the states from which reforecasts are initialized. Similar to the model predictors we retrieve daily gridded values of z300, SST, swvl13 and swvl4. The domain in which they are retrieved is larger (for a reason discussed in Section 4.3.4) and happens at a spatial resolution of 0.25x0.25° and 0.1x0.1°, for ERA5 and ERA5-Land respectively. For additional observed predictors we retrieve 850 hPa temperature (t850) and total cloud cover (tcc), respectively related to the low-level heating and clear-sky conditions in summertime blocking systems. We also extract sea ice concentration (siconc) and snow cover (snowc), as they can be relevant for the summertime jet stream position (Hall et al., 2017; Zhang et al., 2020). Lastly, we retrieve transpiration from vegetation (transp) as an extra indicator of the land surface water balance. This selection of nine variables is the same as in an earlier study on S2S predictability of European summer temperatures (van Straaten et al., 2022).

From ERA5 we also extract gridded daily t2m, which will later form the 'observation' that is used as truth in the training, validation and verification of the post-processing method.

For all gridded daily values described above, we subtract the local seasonal cycle. The climatological value per grid-point and per day-in-the-year is computed by averaging values from the same day-in-the-year (+- 5 days) recorded in the 22 years of the dataset. For the ECMWF model the average value is computed by pooling all members, but is stratified per lead-time to account for gradual drift in the model climatology, which is a known phenomenon in (sub)seasonal forecasts (Johnson et al., 2019). The result is a transformation to gridded daily anomalies relative to the mean seasonal cycle. As Manrique-Suñén et al. (2020) showed, this prevents seasonality to inflate any signals, and will lead to a fairer assessment of

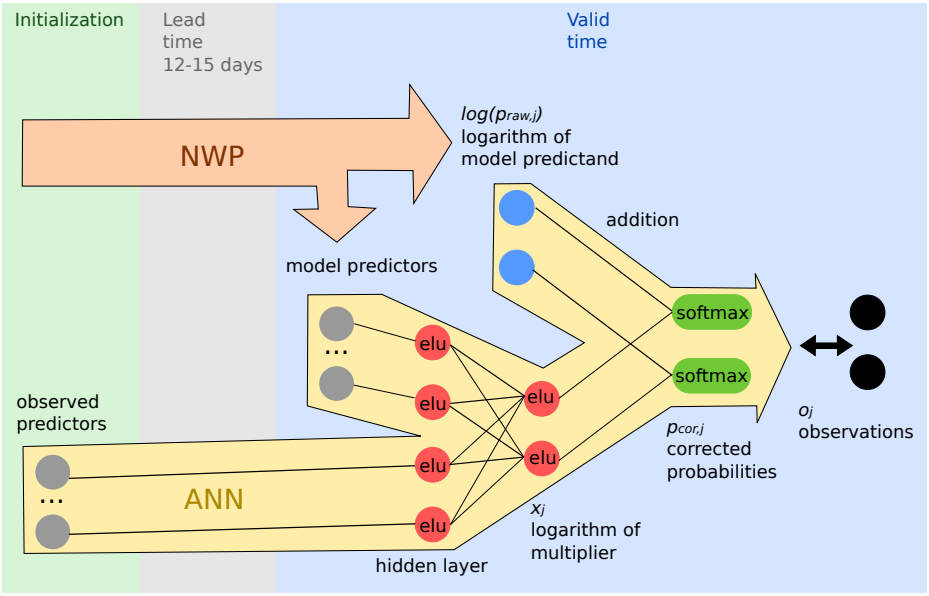


Figure 4.2: Artificial neural network (ANN) for post-processing a two-class probability distribution. The fully connected network is forced to use the distribution forecast by the NWP model (blue) and learns to apply a multiplicative correction (red). The input layer is formed by a mixture of model predictors derived from forecasts at valid time and observed predictors representing the state at initialization (grey). The input layer is connected with single links for illustration purposes only, in reality all predictors are fully connected to the four-node hidden layer. Activation functions are either Exponential Linear Unit (elu) or softmax and are annotated inside the nodes.

sub-seasonal forecast skill.

4.3 Methodology

4.3.1 ANN-based post-processing architecture

Our adaptation of the ANN architecture from Scheuerer et al. (2020), involves model predictors from valid time, and observed predictors from initialization (grey nodes in Fig. 4.2). It post-processes the model predictand at valid time (blue nodes in Fig. 4.2), which expresses the raw forecast probability that monthly average t2m does not ($p_{raw,0}$), or does ($p_{raw,1}$) exceed a threshold, in a period starting 12-15 days in the future

(further details on the predictand in section 4.3.2). These probabilities (blue nodes) are supplied to the neural network right before the final output layer (green nodes in Fig. 4.2). The correction of the model forecasts is learned in the fully connected part of the network (red nodes in Fig. 4.2). We now describe this formally.

Let x_j be one of the nodes in an output layer that predicts whether the observation will fall in the non-exceedance ($j = 0$) or exceedance class ($j = 1$). Softmax activation transforms the two nodes to two probabilities that sum to one:

$$p_j = \frac{\exp(x_j)}{\sum_{k=0}^{1} \exp(x_k)}, \quad j = 0, 1. \tag{4.1}$$

As Scheuerer et al. (2020) demonstrates, and as is done in our ANN, we can take the logarithm of the model predictand (hereafter $p_{raw,j}$) and add it to x_j before softmax activation to obtain the following:

$$p_{cor,j} = \frac{\exp(x_j + \log(p_{raw,j}))}{\sum_{k=0}^{1} \exp(x_k + \log(p_{raw,k}))} = \frac{\exp(x_j) p_{raw,j}}{\sum_{k=0}^{1} \exp(x_k) p_{raw,k}}, \quad j = 0, 1. \tag{4.2}$$

Comparing (4.2) to (4.1), we see that the pre-activation output x_j learned by the fully connected network (red in Fig. 4.2) can be interpreted as the logarithm of a weather dependent multiplier of ECMWF's prior probability $p_{raw,j}$. After activation one obtains the post-processed distribution $p_{cor,j}$ (green in Fig. 4.2). The fact that multiplier x_j is a complex function of the predictors, can be used to understand the physical circumstances leading up to each conditional correction, and thus the weather-dependence of forecast errors. In section 4.3.8 we explain how the multiplier's value gets attributed with XAI.

On a side-note, we concede that two output nodes can seem redundant for a two-class prediction, as the same might be achieved with one output node and sigmoid activation. But we prefer this general architecture because multi-class distributions can be obtained by adding more nodes.

4.3.2 Predictand

The choice of the predictand in S2S forecasting is a compromise between the desire for detail and the limited predictability in such details. Local daily values of t2m in Europe are hardly predictable at lead times beyond two weeks (van Straaten et al., 2020). This means that a certain amount of spatial and temporal aggregation is always needed (Shukla, 1981; Roads,

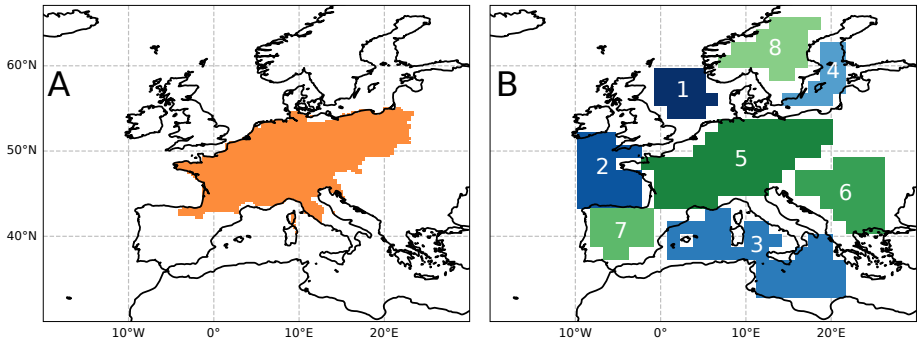


Figure 4.3: A) Region used for spatial averaging of gridded t2m anomalies to obtain predictand. B) Regions used to obtain model predictors at valid time. One statistic is computed to summarize the spatio-temporal state in each region: either the ensemble mean SST anomaly in oceanic regions in blue, or the ensemble mean swvl13 and swvl4 anomaly in terrestrial regions in green. Cluster ids are annotated.

1986; Wheeler et al., 2017). Aggregated values represent the mean effect that sources of predictability have on daily t2m values that by themselves would be unpredictable. In this study we limit ourselves to monthly (31-day) average t2m, averaged over a West and Central European region (Fig. 4.3A), exceeding a certain quantile. We opt for relatively long, monthly averages, because an earlier analysis showed that for this region t2m averaged over weeks 3, 4, 5 and 6 after initialization was better predictable than a shorter-term average over week 3 and 4 (van Straaten et al., 2022). Apparently, the predictable process is better detectable on the longer time scale. We focus on a lead time range of 12 to 15 days before the start of the 31-day period, as combining multiple lead times gives us more samples to work with than a single lead time, and as 15 days is the maximum lead time in the 46-day ECMWF forecasts.

To derive the two-class probability estimate $p_{raw,j}$ from the ECMWF ensemble members, and the corresponding binary observation o_j from the ERA5 reanalysis, we apply three similar methodological steps to both datasets. First, the spatial average of gridded t2m anomalies in the region is taken. Second, 31-day temporal averaging is applied as a rolling window. Third, we estimate climatological quantile levels $q \in 0.5, 0.66, 0.75, 0.9$, of which one serves as exceedance threshold. Quantile estimation is done separately for forecasts and reanalysis, and occurs per day-in-the-year (+-15 days) to account for seasonality. The +-15 day

window is larger than the +-5 day window used for computing anomalies (section 4.2) because estimating high quantiles is more susceptible to noise then estimating a mean. For the ECMWF model the quantile estimation is also stratified per lead time, thereby implicitly correcting for drift in its climatological spread.

After threshold estimation the 11-member ensemble ($M = 11$) is transformed to $p_{raw,1}$ by counting the number of members m above the threshold in a Tukey plotting position estimator (Wilks, 2011):

$$p_{raw,1} = \frac{m + 1 - a}{M + 2 - 2a} \tag{4.3}$$

where $a = 1/3$. Accordingly, $p_{raw,0} = 1 - p_{raw,1}$. We prefer the Tukey estimator over dividing m by M (also known as 'democratic voting', $a = 1$), as the former leads to slightly higher forecast skill (Ferrone et al., 2017). It also prevents probabilities that are zero, which is required for transforming $p_{raw,j}$ to logarithms (equation (4.2)).

4.3.3 Model predictors

Model predictors are one of two types available to the ANN. The model predictors are derived from simulated SST, swvl13, swvl4 and z300 (section 4.2), at valid time: a period that starts 12-15 days into the future. For each of these variables, we capture distinct sources of predictability with tailored summary statistics, such that one predictor/statistic summarizes one spatio-temporal state that can relate to t2m in our target region (and to its forecast errors).

At valid time, predicted SST, swvl13 and swvl4 anomalies are simultaneous in time to the predictand. This means that they do not influence the predictand in a temporally lagged and spatially remote fashion, but that the dominant influence will be more local instead. As relevant spatio-temporal states we thus extract 21-day and 31-day temporal and spatial averages in regions close by or inside the predictand region (Fig. 4.3B). As summary statistic we only extract the ensemble mean prediction, as the additional use of ensemble spread does not always improve post-processing skill (Rasp and Lerch, 2018; Schulz and Lerch, 2022). With four regions for SST, four regions for swvl13 and swvl4 (Fig. 4.3B), and two timescales, this results in 24 predictors.

Predictors from Z300 are obtained differently. The circulation at this upper tropospheric level is namely known to exhibit recurring synoptic configurations over extended periods of time, that have specific imprints

on the surface weather (Hannachi et al., 2017; Casanueva et al., 2014; Grotjahn et al., 2016). Among these 'regimes' are the ridges or blocks responsible for summer heat (Cassou et al., 2005; Pfahl, 2014; Sousa et al., 2018; Kueh and Lin, 2020; Kautz et al., 2021). Consequently, succesful S2S prediction of surface weather is sometimes possible when forecast tropospheric circulation gets classified into appropriate regimes (Lavaysse et al., 2018; Richardson et al., 2020; Mastrantonas et al., 2022). Here we obtain predictors summarizing the appropriate spatio-temporal states, by classifying predicted z300 anomalies into four distinct regimes (details in Appendix A.1). By recording the predicted frequency of each regime in a 21-day and 31-day period, this results in 8 additional predictors.

4.3.4 Observed predictors

The second type of predictors that we supply to the ANN are observed predictors at initial time. These represent the initial state of atmosphere, ocean and land, unaffected by NWP model errors.

The domains from which these predictors are obtained are larger than above because initialization occurs 12 to 15 days before our predictand period. In those two intervening weeks, mid-latitude circulation can be influenced by short-lived heating from as far as the tropics (Branstator, 2014). The lagged influence of a sub-seasonal source of predictability is not only local. Signals originating in tropical Pacific SST have long been known to affect the distribution of air masses elsewhere (Bjerknes, 1969). The tropics are therefore included in the SST domain, and since such teleconnections not only originate from SST but also from other variables, such as snow cover and sea ice (Hall et al., 2017; Zhang et al., 2020), nine reanalysis variables are taken into account (section 4.2). These variables and the associated domains are taken from van Straaten et al. (2022) (their Fig. 1).

To capture the spatio-temporal states of the different sources of predictability in those domains, we first identify distinct regions in which gridded anomalies from the nine variables relate to our predictand of interest (Bello et al., 2015; Kretschmer et al., 2017). These regions resulted from the analysis of van Straaten et al. (2022), in which we computed the lagged correlation between the gridcells of each variable and the average t2m in the predictand region, starting 12-15 days later. Temporally, the correlation was performed on average states ranging from 1-day averages to 31-day averages of the daily gridded anomalies, in order to extend the range of possibly useful time scales beyond the monthly timescale of

the predictand. Spatially, groups of neighbouring, significantly correlated grid cells were then clustered into feature regions. Assuming that each of those represented a distinct source of predictability, two statistics were extracted per region to summarize its spatio-temporal state: the spatial mean anomaly, and a spatial covariance measuring the resemblance between the reanalysis state and the correlation pattern that was clustered (more details in van Straaten et al., 2022). With two statistics, a variable number of feature regions, nine variables and eight scales of temporal averaging, the procedure results in 226 predictors.

As final observed predictors, we also include the state of the Madden Julian Oscillation (MJO), 12-15 days before our target period. This tropical oscillation is a dominant mode of sub-seasonal variability (Zhang, 2013), and is known to conditionally affect atmospheric circulation over the Euro-Atlantic region (Cassou, 2008; Lin et al., 2009). In NWP models the amplitude of the connection is often found to be too weak (Vitart, 2017). We use the daily two-component RMM index as supplied by the Australian Bureau of Meteorology (Wheeler and Hendon, 2004). This adds two further predictors and results in a grand total of 260 (model and observed) predictors, which are subject to a predictor selection described in section 4.3.7.

4.3.5 Benchmark models

With its mixture of predictors the ANN corrects raw ECMWF probabilities p_{raw}. The skill of corrected probabilities p_{cor} is evaluated against p_{raw} itself and benchmarks that are based on ERA5 t2m data only.

Our first benchmark to compare against, uses the threshold quantile level q. Exceedance of the 0.75 quantile happens (on average) in 25 percent of the data. Assuming stationarity, we create a constant probability forecast valid for each sample:

$$p_{constant,1} = 1 - q, \tag{4.4}$$

However, when climate is changing, a comparison against a stationary benchmark can artificially inflate skill (Hamill and Juras, 2006; Manrique-Suñén et al., 2020; Wulff et al., 2022). Anthropogenic global warming makes exceedance of a stationary threshold less likely in the early part of the climatological period, and more likely at the end of it. A more competitive non-stationary benchmark takes that into account, which is why a second benchmark models the climate-change driven probability

with Logistic Regression. p_{trend} is a function of time t (Julian day), and regression coefficients β and γ:

$$p_{trend,1} = \frac{1}{1 + e^{-(\beta+\gamma t)}}. \tag{4.5}$$

Such a trend model signifies the climatologically expectected probability of exceedance, not conditioned on any weather information. This means that we can introduce that conditioning by adjusting the expected probability up and down, based on a set of predictors, similar to adjustments made in our post-processing of model-predicted probabilities. In the third benchmark model we therefore supply $\log(p_{trend})$ instead of $\log(p_{raw})$ to the ANN as prior probability estimate (blue nodes in Fig. 4.2), and leave everything else the same. The resulting benchmark forecasts are called $p_{trend+ann}$.

4.3.6 Performance metrics

We verify the probabilistic exceedance forecasts $p_{[raw,cor,constant,trend,trend+ann],1}$ against binary observations o_1. With the Brier Score (BS) we compute the mean squared error over the n forecast-observation pairs: $BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$ (Brier, 1950). Additionally, we use reliability diagrams to visually assess the forecast reliability and resolution (Wilks, 2011).

We also use the Area Under the ROC-Curve (AUC). For this score the forecasts are discretized for a range of probability thresholds, each providing a contingency table with True- and False Positives, and True- and False Negatives (respectively TP, FP, TN, FN). This leads to an ROC-diagram with false alarm rates ($FP/(FP+TN)$) on the x-axis and hit rates ($TP/(TP+FN)$) on the y-axis. For each set of probability forecasts AUC quantifies whether increases in p discriminate observed occurrences from non-occurrences. AUC is however in-sensitive to the magnitude of the increase and therefore only forms a measure of the forecast's potential usability (Kharin and Zwiers, 2003). We also compute the Hanssen-Kuipers or Pierce Score (PS), which is the hit rate minus the false alarm rate. It expresses the maximal potential economic value a reliable forecast can have to users making yes-or-no decisions (Richardson, 2000).

All numerical scores $S \in \{BS, AUC, PS\}$ are transformed to skill scores ($BSS, AUCSS, PSS$) by normalizing the difference between S and $S_{constant}$, the score computed for the constant benchmark model (equation

4.4):

$$SS = \frac{S_{constant} - S}{S_{constant} - S_{perfect}}, \tag{4.6}$$

where $S_{perfect}$ is {0,1,1} respectively.

4.3.7 Data partitioning and ANN tuning

We measure the performance of our ANN-based post-processing method on independent test data. For tuning, the performance on a validation set is used. We split the set of 22 summers into four seasons for testing, and 18 seasons for cross-validation (Fig. 4.4). This means that the 18 seasons are further subdivided into three folds with 6 seasons each. Repeatedly, a model is trained on two of them, and makes predictions for the third. After three repeats, the concatenated validation predictions are assessed for performance and used to select predictors and choose hyperparameter values.

The data splits are subject to two criteria. First, our observed predictors from initialization are based on clustered correlation patterns. These correlations were computed on data from 1981 till 2013 (more details in van Straaten et al., 2022). For independent testing the test set should thus be post-2013. Second, all sets need to contain a balanced mix of cold, normal and warm seasons, such that each is representative of the data distribution. Given that p_{raw} displays a warming trend (Fig. 4.4)), an equal balance cannot be achieved with chronological splits. A training set from 1998 to 2004 would be dominated by lower-tercile or 'cold' seasons (minus symbols in Fig. 4.4). So instead we split the data such that each set contains seasons from all tercile classes (result visible in Fig. 4.4).

Note that although we use terciles to partition the data, the predictand is still a two-class variable, with a separate ANN being trained for each quantile exceedance threshold $q \in 0.5, 0.66, 0.75, 0.9$. For each of these ANNs, we select a set of l predictors based on combined validation performance. First, each of the 260 predictors was scaled to lie between 0 and 1 (validation and test sets remain unseen). Then a greedy stepwise forward algorithm was applied: Suppose $l-1$ predictors are already in use. Add an unused candidate, train the model on two cross-validation folds, make predictions for the remaining fold, and repeat the training and prediction to also cover the other folds. Compute the Ranked Probability Score (RPS) $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}(p_{i,j} - o_{i,j})^2$ over the n concatenated validation samples and the two classes j. Repeat the steps above three

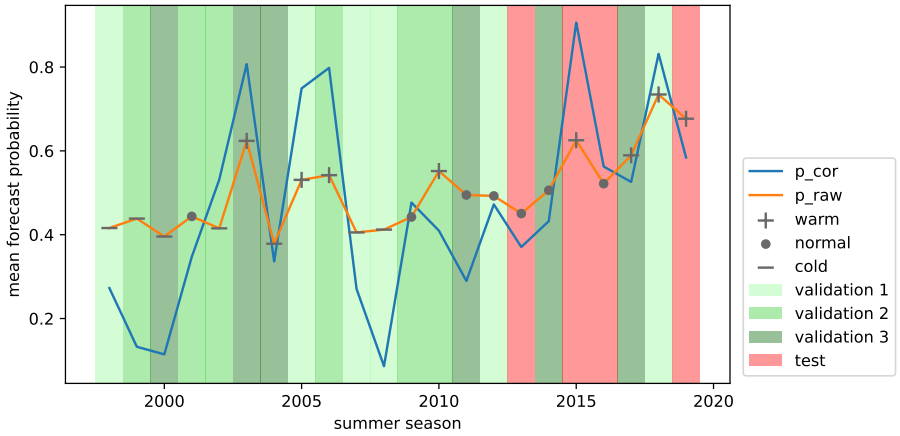


Figure 4.4: Summer average probability that monthly temperature exceeds the 0.5 quantile, forecast with a lead time of 12 to 15 days before the start of the period. The dataset of 22 summer seasons is divided into sub-sets that balance the amount of warm, normal and cold summers (tercile classes of p_{raw}, symbols). Test set summers are colored red and the three-fold cross-validated sets green, of which two folds are used for training and the remaining for optimizing the model. Raw ECMWF forecasts are indicated by the orange line, probabilities from an optimized post-processing model are indicated by the blue line (p_{cor}, trained on all cross-validation folds).

times to account for randomness due to weight initialization. Record the scores, switch the candidate predictor for a new one, and proceed from the start until all predictors are tested. Choose the one leading to the lowest mean RPS as the l-th predictor. Continue to find the $l+1$ most imporant predictor, but only if validation RPS decreases by more than 6 percent. This subjectively chosen stopping criterion resulted in 2 to 4 selected predictors per ANN.

Each ANN is trained with exponential linear unit (elu) activation functions, categorical cross-entropy loss, and the ADAM-algorithm for gradient descent (Kingma and Ba, 2014). Tuning was required for (i) the number of hidden layers, (ii) the number of nodes per hidden layer, (iii) the learning rate, (iv) the batch size and (v) the amount of epochs for which we tolerate an increasing validation loss before stopping training early. Settings for these parameters were explored using a large random search, by means of the SHERPA python package (Hertel et al., 2020).

The number of trials was set to 200. Combinations leading to the lowest RPS over the three cross-validation folds were chosen (repeated 8 times to mitigate randomness due to weight initialization). For all quantile thresholds, a simple model was the result, namely with 1 hidden layer, 4 hidden layer nodes, a learning rate of 0.0014, a batch size of 32 and a patience of 7 epochs before stopping training early (see the architecture in Fig. 4.2). With these hyperparameters and after predictor selection, each ANN is trained one final time on the combined cross-validation folds, with 33 percent of the data left out for the early stopping algorithm.

4.3.8 Explainable AI

Besides improving forecasts by post-processing p_{raw}, we also want to understand the conditional NWP errors that the ANN corrects. Predictor contributions can be used to learn about the physical circumstances of each error.

After the ANNs have been trained we apply two XAI techniques to attribute the learned corrections to the 2 to 4 predictors involved. Recall that the learned correction factor $\exp(x_1)$ is a weather-dependent multiplier of $p_{raw,1}$ (equation (4.2)). Predictor contributions to this factor will thus vary from forecast to forecast, depending on the state of the sources of predictability that the observed- and model predictors represent. Positive predictor contributions signify conditions in which $p_{raw,1}$ is usually an underestimation, meaning the ANN attempts to increase the probability of exceedance. Negative contributions signify conditions in which $p_{raw,1}$ is an overestimation.

The first method with which we quantify contributions is the model-agnostic KernelSHAP, where SHAP stands for Shapley Additive Values (Lundberg and Lee, 2017). These values originate from game-theory, and are solutions to the problem of dividing a game's single payout over multiple contributing players. In statistical modeling the equivalent is the division of a single predicted value over the contributions from each predictor. As these contributions are defined relative to a 'normal' background, they can be negative or positive. Together with the background expectation they will add up to the predicted value, in this case the multiplication factor. Determining a background from all samples is computationally demanding so we reduce the full dataset to 30 representative centroids with k-means and compute the background from those. SHAP values are then computed for each train and test sample.

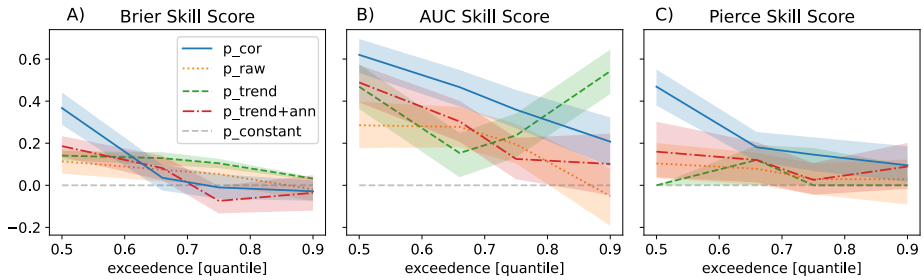


Figure 4.5: Test set performance according to three probabilistic skill scores, (a) BSS, (b) AUCSS and (c) PSS. Predictand is monthly average European temperature exceeding a quantile threshold (x-axis). One ANN is fitted per exceedance threshold and to the combined lead times of 12-15 days before the start of the monthly period. p_{cor} is the result of post-processing p_{raw} with the ANN, $p_{trend+ann}$ is the result of adjusting p_{trend}, also with the ANN. Comparison is made to p_{raw} and p_{trend} themselves. $p_{constant}$ is the zero skill reference. Shading denotes the 5th to 95th percentile uncertainty bounds, as obtained by bootstrapping forecast-observation pairs before computation of the scores (1000 repeats).

The second XAI method is specific to differentiable models like neural networks. We quantify the gradient of the fully connected output (x_1) with respect to the predictor values. Such a gradient is measured in the vicinity of the predicted value and represents a sensitivity. However, sensitivity does not always entail relevance. For instance, daytime t2m can be sensitive to cloud cover, with clearer skies leading to more solar irradiance and surface heating. But on a persistently cloudy day, the real cause of high t2m can be a process like warm air advection, to which it is also sensitive. Actual relevance is better approximated by multiplying the gradient with the predictor value, rather than using the gradient itself. This so-called 'input times gradient' has shown good reconstruction of a known ground truth in a climate-like prediction problem (Mamalakis et al., 2022b).

4.4 Results

4.4.1 The benefit of post-processing

The ANN-based post-processed forecast (p_{cor}) is supposed to improve upon the NWP forecast (p_{raw}) by using information from both initializa-

tion and valid time. Skill scores for the two, and other benchmarks, are presented in Fig. 4.5. Extremity of the predictand varies along the x-axis, showing the quantile used as exceedance threshold in monthly temperature (section 4.3.2). One ANN is fitted per quantile, to the combined lead times of 12 to 15 days before the start of the predictand period. A separate predictor selection is performed for each ANN.

In Figure 4.5 skill scores generally decrease with increasing exceedance threshold. p_{cor} (solid blue) outperforms all benchmarks in all three scores for median exceedance, and only in AUCSS and PSS for the 0.66 and 0.75 quantile. For the 0.9 quantile the AUCSS of the trend-benchmark drastically increases (dashed green, Fig. 4.5B). That behaviour might be caused by AUC's trapezoidal approximation, leading to distorted results for extremes (Ben Bouallègue and Richardson, 2022). Comparing the scores, BSS can seem overly conservative. However, the Brier score is a proper score (Gneiting and Raftery, 2007) and it is common that skill for 90th percentile events does not extend beyond two weeks after initialization (Lavaysse et al., 2019). Absence of skill can come from ANN's inability to learn the right conditional corrections in a limited set of samples, or from an intrinsic lack of predictability in these extremes.

For median exceedances it is clear that p_{cor} provides the most skillful forecasts (solid blue, Fig. 4.5), also when comparing to $p_{trend+ann}$ (dash dot red). The latter ANN uses p_{trend} as prior probability estimate instead of p_{raw} (p_{raw} is in that case supplied as regular predictor, but never selected). The difference between the two does not stem from their trend-awareness, as p_{raw} is trended as well (Fig. 4.4) (see also Shao et al., 2022). Also, both are equally weak when un-adjusted (dotted orange and dashed green, Fig. 4.5A). The skill gained by p_{cor} over $p_{trend+ann}$ therefore shows that p_{raw} forecasts from ECMWF contain a predictive value that emerges when its shortcomings are corrected. In the remainder of the paper we focus on the ANN-based p_{cor} forecasts for median exceedance.

A more complete look on the performance of p_{cor} is given by reliability diagrams (Fig. 4.6). On the training set the reliability achieved by post-processing is near perfect (solid blue, Fig. 4.6A). This is obvious because the ANN trains to mimic the observations, until it is stopped early by increasing validation loss. The early stopping is visible in the fact that p_{cor} is not concentrated at 0 and 1 (which would be a perfect mimicry), but that the issued probabilities of exceedance cover a range in between (solid blue, Fig. 4.6B). The ANN is able to issue sharper forecasts (i.e. more probabilities close to 0 and 1) than p_{raw} (dotted orange, Fig. 4.6B). These improvements transfer to the unseen test data, as p_{cor} lies closer

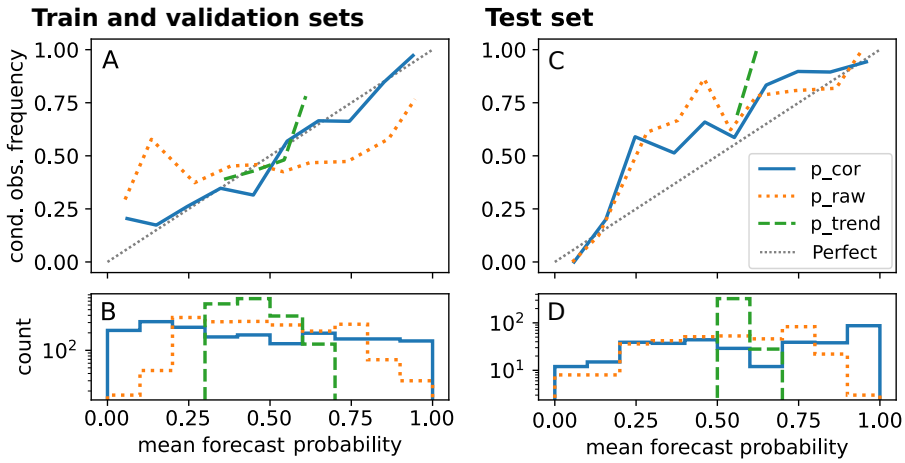


Figure 4.6: Reliability diagrams of ANN-based post-processing of monthly t2m exceeding q0.5 (blue), measured against raw NWP forecasts (orange) and the trended benchmark (green). Lead time is 12 to 15 days before the start of the monthly period. Left: performance on the combined cross-validation sets (green summers in Fig. 4.4). Right: test set. A,C) Reliability diagrams of the conditional observed frequency per bin versus binned forecast probabilities. The dotted 1:1 line shows perfect reliability. B,D) Histogram of binned forecast probabilities, where concentration close to 0 and 1 hints at high forecast sharpness.

to the 1:1 line and results in more probabilities close to 1 than p_{raw} (Fig. 4.6C,D).

4.4.2 Selected predictors

The ANN-based forecast presented in Figure 4.6 is made with a set of 3 predictors, besides the model predictand (p_{raw}) that is is being corrected. Figure 4.7 depicts the top-20 predictors that the forward predictor selection found. Only the top 3 decreased RPS on the combined validation folds by more than 6 percent, and were therefore included in the ANN (marked blue dots in bottom row of Fig. 4.7). The figure shows whether the predictors are observed or model-predicted (top row), whether they relate to atmospheric, oceanic or land surface variability (second row), and at what time scale they are defined (third row).

The highest ranked predictor is 21-day SST from ERA5. It represents the state of SST in the western equatorial Pacific at initialization (see the

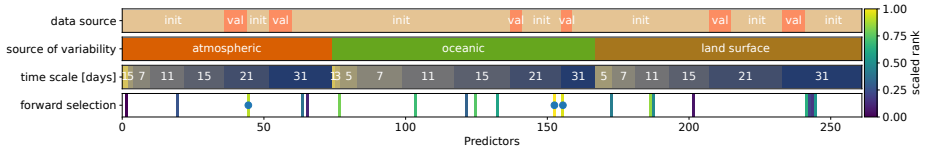


Figure 4.7: Characteristics of predictors (x-axis) selected for the ANN-based post-processing of monthly temperature exceeding q0.5 (lead time of 12-15 days before the start of the monthly period). Bottom row: rank of the top-20 forward selected predictors. Marked by blue dots are the top three predictors that decrease RPS by more than 6 percent. Top row: data source (whether observed at initialization, 'init', or model-predicted at valid time, 'val'). Middle rows: source of variability, and the timescale of the predictors (i.e. the average in days before the lead time gap for observed predictors, and the average in days after the gap for model predictors).

bottom source of predictability mapped in Fig. 4.1). In the construction of observed predictors, both the mean anomaly and spatial covariance were extracted, but it is the spatial-covariance statistic that is selected here (we will discuss what it represents in section 4.4.3). The primacy of this predictor is confirmed by two other ways of predictor selection in which it also appeared first (not shown). The second predictor is model-predicted 31-day average SST in the North Sea (region 1, Fig. 4.3B), and therefore comes from valid time. The third predictor is ERA5 850 hPa temperature at initialization, specifically the mean 21-day anomaly in a subtropical region stretching from the Atlantic, over the Sahel to the Indian Ocean (see the upper source of predictability mapped in Fig. 4.1).

Two of the predictors responsible for the corrections, represent an observed state at initialization time. The relation of these sources of predictability to NWP errors suggests that their effects are potentially misrepresented by the NWP model, but can be corrected. Only one predictor, namely North Sea SST, is a model-predicted state at valid time. Remarkably, model-predicted candidates like soil moisture and circulation regime, are not selected. Their influence on t2m is either already properly resolved in the ECMWF model, meaning they do not relate to systematic t2m errors compared to reanalysis, or they themselves are biased or lack predictability at these lead times. Either reason would render them useless for post-processing (this topic is further discussed in Appendix A.1).

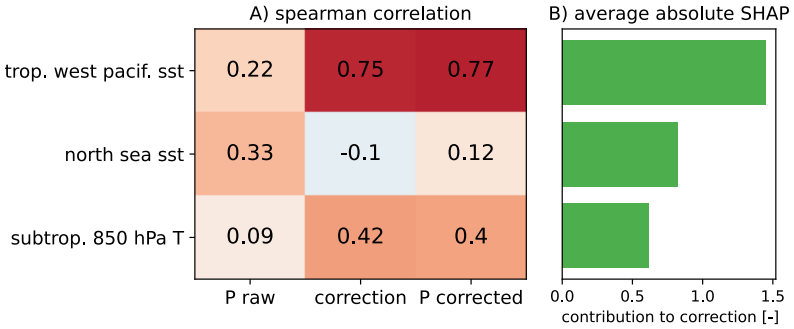


Figure 4.8: Summary of the role of the three selected predictors in post-processing monthly temperature > q0.5 with a lead time of 12-15 days before the start of the monthly period. A) spearman rank correlation with p_{raw} from ECMWF, with the correction by the ANN ($p_{cor} - p_{raw}$), and with post-processed probability (p_{cor}). B) Importance for the learned multiplicative correction as measured with the absolute SHAP value over all samples.

4.4.3 Understanding forecast errors

Next we aim at understanding conditional NWP errors, by attributing the ANN's corrections to the three predictors. Fig. 4.8A displays the three predictors and their Spearman rank correlation with p_{raw}, the correction ($p_{cor} - p_{raw}$), and p_{cor}. The tropical west Pacific SST predictor shows a weak positive correlation with p_{raw} and a strong positive correlation with the correction and p_{cor} (Fig. 4.8A, first row). This amplification by the ANN indicates that the connection in the ECMWF model between this initial SST pattern and our monthly predictand, is weaker than it should be.

Of the three predictors North Sea SST is most positively correlated with p_{raw} (Fig. 4.8A, first column). This is understandable given the geographical proximity of the North Sea to the predictand region and the fact that simulated cold or warm spells are likely to extend over both (Fig. 4.3). Direct correlation between this predictor and the correction is close to zero, meaning that the role of North Sea SST in the ANN is likely non-linear and not captured by a monotonic correlation metric. Average absolute SHAP values (interpretable as the average relevance over all samples) show that it is an important predictor (Fig. 4.8B). It is more important than subtropical t850, even though the latter shows higher correlation with the correction (Fig. 4.8A, second column). We will see

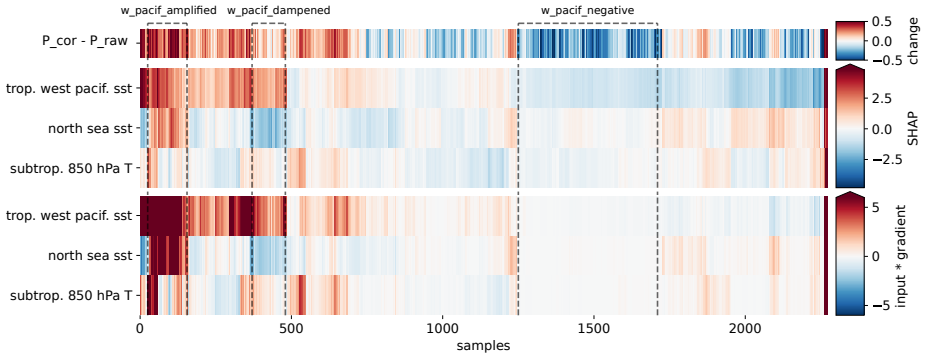


Figure 4.9: Contributions of the three predictors to the ANN-based correction of forecasts of monthly temperature above the median, made with a lead time of 12-15 days before the start of the period. Top row displays the change in probability applied by the ANN ($p_{cor} - p_{raw}$). Three rows below shows contributions as quantified by SHAP (summing up to the multiplicative correction factor). Bottom three rows are contributions as quantified by input times gradient (inputs are standardized instead of min-max scaled between 0 and 1). Samples (x-axis) are sorted by the leaf order that results from a hierarchical clustering of SHAP values, such that situations requiring similar corrections for the same reason, are close.

later that this is because predicted North Sea SST plays a modulating, conditional role.

Figure 4.9 shows the contributions from each predictor per corrected forecast. The applied correction is in the top row, with SHAP contributions and input times gradient contributions below. As we are interested in the physical circumstances of different errors, we re-order the samples along the x-axis. The order is the so-called 'leaf order' of a hierarchical clustering algorithm (applied to the SHAP values with a Euclidian distance metric). This order refers to the lowest level in the hierarchy, where each sample is still its own cluster and the algorithm has shuffled similar samples close to each other. We perform this grouping because NWP errors, and therefore the learned corrections, are highly conditional. This means that the physical circumstances can only be understood when looking at similar corrections that are applied for the same reason. To achieve this, samples should not be grouped on predictor values only, as that would weigh west Pacific SST and t850 equally, even though their importance in the correction differs (Fig. 4.8B). Neither can we group samples by their correction factor only, as that would mix samples with

similar corrections that are applied for very different reasons. Hierarchical clustering of SHAP values reconciles these two approaches, as predictor values are replaced by predictor contributions to the output, and are in units that sum up to the applied correction. Lundberg et al. (2020) showed that this makes the clustering supervised, and weighs variables according to their impact.

The hierarchical re-ordering enables us to distinguish different groups by eye. We distinguish and annotate three groups (Fig. 4.9). In 'West Pacific amplified' circumstances are such that the tropical west Pacific and North Sea SST predictors contribute positively to the correction (both red), with a large increase from p_{raw} to p_{cor} as a consequence. In 'West Pacific dampened' the west Pacific contributes positively but North Sea SST contributes negatively (blue). The eventual applied correction is close to zero. In 'West Pacific negative' the west Pacific contributes negatively while other predictors remain neutral, leading to distinct negative corrections (more visible in SHAP than in input times gradient).

Note that the sign of the t850 contribution varies within the delineated groups, which means that the hierarchical algorithm considers t850 of lesser importance for finding similar physical circumstances. Quantitatively it is also a less important predictor than West Pacific and North Sea SST (Fig. 4.8B). In the remainder of the paper we therefore focus on the two SST predictors.

The three identified correction groups point to different physical circumstances that we investigate by creating a composite of the samples in each group, for the variables z300, SST, swvl13 and t2m. Figure 4.10 displays these variables (in rows) at different moments in time (columns), namely the analysis at initialization time (first column), the analysis at valid time (second column), and the ensemble mean forecast at valid time (third column).

In 'West Pacific amplified' the development of high pressure (and high t2m) over west and central Europe is under-estimated by the ECMWF forecasts. The ERA5 data show that z300 from initialization onward (Fig. 4.10a) develops into a zonal quasi-stationary wave-pattern, with a core of high pressure over western Europe, flanked by two low pressure regions (Fig. 4.10b). The raw forecasts misplace the high pressure core over the north Atlantic (Fig. 4.10c). In ERA5 the heat situated over the Iberian peninsula expands into west and central Europe (Fig. 4.10j to k). Comparatively the forecast of t2m in the target region is too cold (Fig. 4.10l). Since the ECMWF model retains a correct soil moisture pattern (Fig. 4.10g,h,i), it seems that the missed or misplaced atmospheric wave

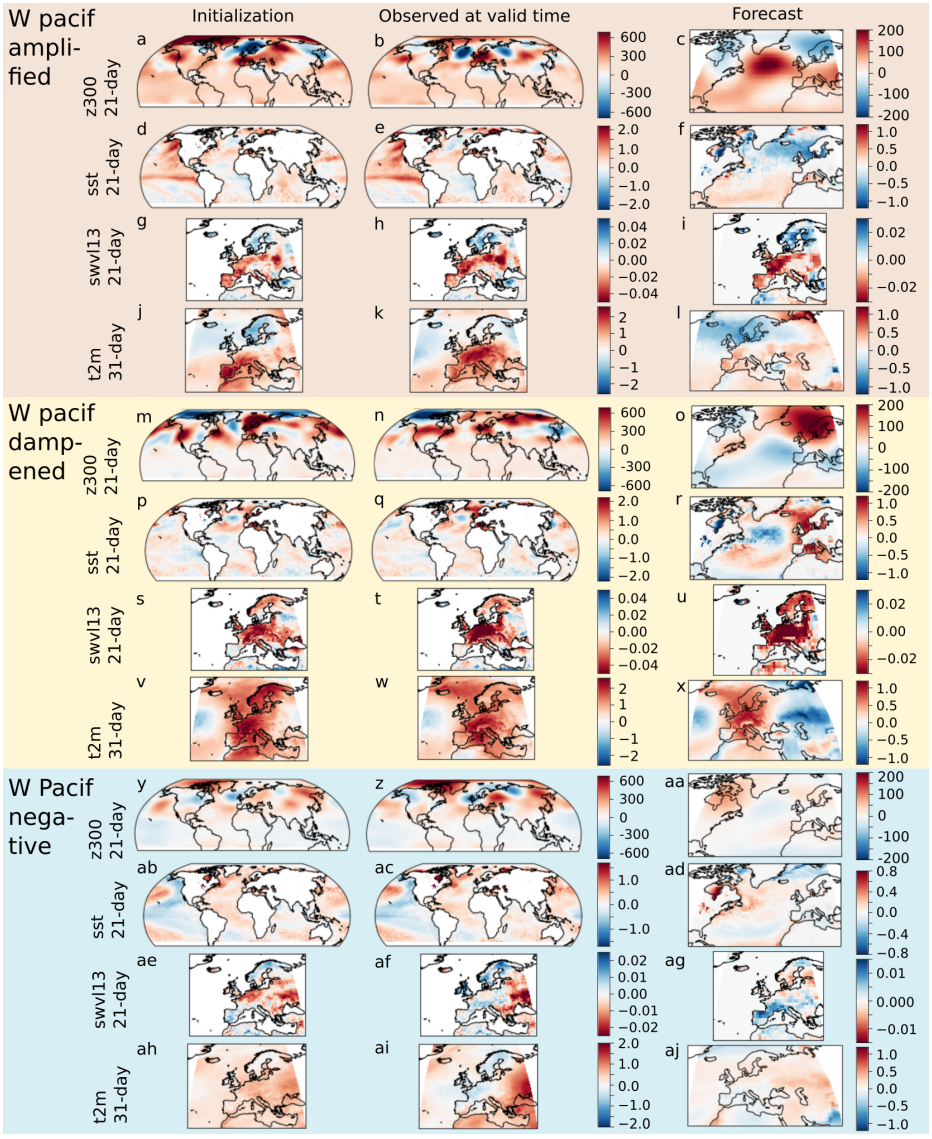


Figure 4.10: Composite anomalies for samples belonging to the three groups delineated in Fig. 4.9. Rows display different variables, columns display different moments in time. Left: the 21- or 31-day reanalyzed state before initialization. Middle: the 21- or 31-day reanalyzed state during our predictand period (starting two weeks later, at valid time). Right: 21- or 31-day ECMWF ensemble mean forecast over the Euro-Atlantic domain (also at valid time). Note that the color bar scale for forecasts is reduced because we take the ensemble mean. Units are $[m^2 s^{-2}]$, [K], $[m^3 m^{-3}]$, [K] for z300, SST, swvl13 and t2m respectively. Time scale for t2m is 31 days, to overlap fully with the t2m-based predictand. Time scale for z300, SST and swvl13 is 21 days, to overlap fully with our most important SST predictor from initialization.

is the prime reason for the ANN to apply a strong upward correction in these situations (top row, Fig. 4.9). As noted, the west Pacific SST predictor contributes positively to this correction. As a whole, tropical Pacific SST resembles an El Niño state (Fig. 4.10d,e).

An informative contrast appears with 'West Pacific negative', where Pacific SST shows an opposite, La Niña like pattern (Fig. 4.10ab,ac). Again an initial z300 pattern develops into a wave-like pattern with strong anomalies (Fig. 4.10y to z). Low pressure is now situated over the British Isles and high pressure over western Russia (Fig. 4.10z). For the predictand region this means that soils wetten (Fig. 4.10ae to af), and that heat accumulates east of it, over western Russia (Fig. 4.10ai). The ECMWF model does not capture this pattern, as it places the wettening of soils around the Mediterranean (Fig. 4.10ag). Moreover, the low pressure system is too weak and located too far to the west over the Atlantic (Fig. 4.10aa). The slightly cool t2m anomaly in the predictand region and the warm anomaly over Russia (Fig. 4.10ai) are missing in the forecasts (Fig. 4.10aj), which is why the ANN learns to decrease the p_{raw} exceedance probability in this group of samples (top row, Fig. 4.9).

In the physical circumstances of 'West Pacific amplified' and 'West Pacific negative' the ECMWF model thus appears unable to simulate a developing atmospheric wave, associated with west/central European heat in the former, and Russian heat in the latter. The ANN learns that the necessary, opposing corrections are predicted by west Pacific SST. Figure 4.11 summarizes the NWP forecast errors for the three groups ('West Pacific dampened' will be discussed later) and the state of the west Pacific predictor. As a covariance predictor it measures correspondence between anomalies (red-blue shading) and correlation pattern (green-purple contours), in the gridcells that correlate significantly to European t2m (see also section 4.3.4). Positively correlated cells lie at the western edge of the Nino-4 region (0°N, 160°E) (Trenberth and Stepaniak, 2001). Negatively correlated cells lie at 10 and 20°N (green contours in Fig. 4.11A). In 'West Pacific amplified' the positive-negative correlation dipole, is appearing as a hot-cold dipole in the anomalies. This leads to positive covariance (0.027, Fig. 4.11A). In 'West Pacific negative', the hot-cold dipole in the anomalies is inverted, leading to negative covariance (-0.033, Fig. 4.11C). The (inversion of the) pattern predicts the under- (over-) estimation in p_{raw}, as compared to the observed frequency of exceedance (Fig. 4.11E,G).

Based on the composite SST pattern, we infer that the west Pacific hot-cold dipole comprises an ENSO-like source of predictability, with a

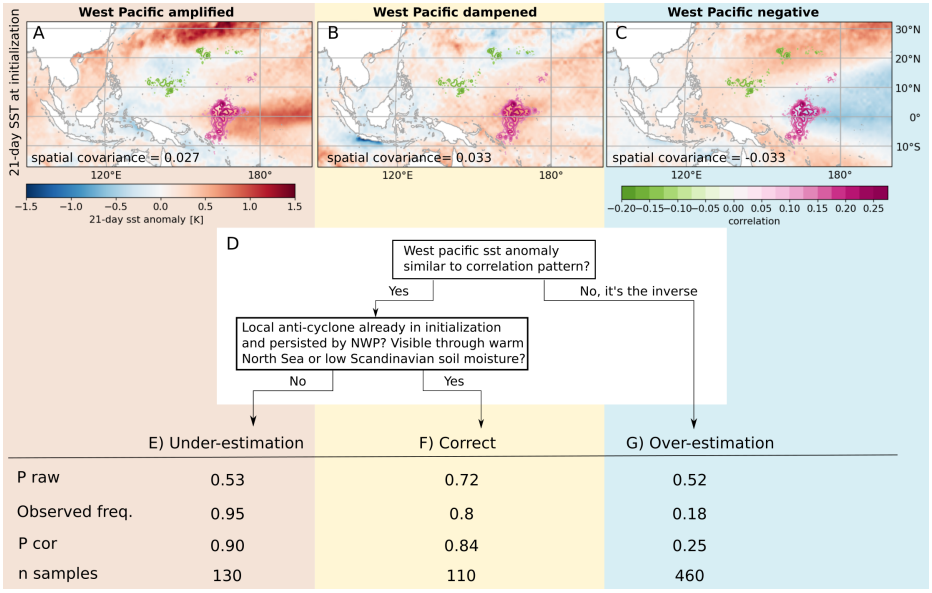


Figure 4.11: (A-C) Composite of initial 21-day SST anomalies, in the three groups of Fig. 4.9. Contours display correlation values for grid cells that correlate significantly to t2m in the predictand region. The west Pacific predictor measures correspondence between anomalies and correlation pattern. Its composite value is annotated (unit [K]). (D) Human summary of learned weather-dependent corrections. (E-G) Table with associated forecast errors and their correction: mean p_{raw}, observed frequency of exceedance, mean p_{cor} and number of samples in each group.

contrast between SST anomalies in the central tropical Pacific and those in the Maritime Continent. It is a source of predictability for the development of an atmospheric wave in the reanalysis, and errors at valid time in the ECMWF model. Some aspect of the predictable physical pathway is thus misrepresented. Indeed, tropical convection patterns, closely related to ENSO, are known to influence Euro-Atlantic summer circulation (Ding et al., 2011; O'Reilly et al., 2018; Ma and Franzke, 2021). Extratropical circulation gets affected by patterns of tropical convection leading to diabatic heating and upper level divergence, which forces Rossby waves (Bjerknes, 1969; Sardeshmukh and Hoskins, 1988; Ting, 1994; Trenberth et al., 1998). Aspects of such teleconnection responses, like the propagation of Rossby Waves, are known to be misrepresented in NWP models (Beverley et al., 2019) (see O'Reilly et al., 2018; Strazzo et al., 2019, for misrepresentation of ENSO-like teleconnections).

Our results however also reveal that the NWP error is conditional. Corrections in the 'West Pacific dampened' group suggest that the error can be absent. In these samples the west Pacific SST dipole is in a positive state (covariance of 0.033, Fig. 4.11B), but predicted p_{raw} is already close to the observed frequency of exceedance (Fig. 4.11F). Looking at Figure 4.10, we see that a strong Scandinavian blocking is present at initialization (Fig. 4.10m), causing high t2m anomalies in the predictand region (Fig. 4.10v). This situation persists over the region until valid time (Fig. 4.10n,w) and is accompanied by further depletion of soil moisture and warming of North Sea SST (Fig. 4.10t,q), all of which is correctly captured by the ECMWF forecasts (Fig. 4.10r,u,x). Only the predicted low pressure over the Iberian peninsula is not observed (Fig. 4.10n,o). Especially the model-predicted warm North Sea and depleted Scandinavian soil moisture, distinguish this group of correct forecasts from the cold North Sea and wet Scandinavia in 'West Pacific amplified' (Fig. 4.10f,i). This suggests that sub-seasonal prediction of median t2m exceedance is more successfull when local anti-cyclonic circulation is already present in the initial conditions (including derivatives like a warm North Sea) (also noted in Vitart et al., 2019). As there is no need to apply a correction, the ANN modulates contributions from the west Pacific predictor (which would lead to an upward correction), with a second predictor, namely predicted North Sea SSTs, such that the overall correction is minor. Figure A.1 in the appendix further illustrates that statistically, the ANN dampens the West Pacific influence when model-predicted North Sea SST is high. The capacity for conditional correction is further demonstrated by the comparison in Appendix Table A.1, in which the ANN performs better than a logistic regression which is incapable of modeling conditional interactions. Overall we can summarize ECMWF's conditional errors and the learned corrections with the decision tree in Figure 4.11D.

4.5 Discussion

Our ANN-based post-processing technique corrected forecasts with a mixture of observed predictors from initialization (ERA5) and model predictors from valid time (ECMWF forecasts). Subsequent application of XAI highlighted a connection between west Pacific SST anomalies at initialization, and west/central European t2m more than two weeks later. The atmospheric wave associated with the west Pacific dipole, is absent in the ECMWF model (Fig. 4.10), meaning that its forecast (p_{raw}) lacks the

observed shift to high and low probabilities of median exceedance (Fig. 4.11E,G). Post-processing applies this shift through conditional correction, effectively strengthening the relation between the west Pacific SST pattern and exceedance probability (first row, Fig. 4.8A). The robustness of this correction is supported by increased skill on unseen test data, and by the fact that the ANN used only three predictors besides p_{raw}. This makes it unlikely that the ANN exploited spurious interactions between predictors for their coincidental alignment with variability in the predictand.

The discovered relation between the initial west Pacific state and forecast errors, highlights the importance of correctly representing tropical to extra-tropical teleconnections in the NWP model. Previous studies have also found that tropical Pacific sources of predictability are important for European summer weather (Ding et al., 2011; O'Reilly et al., 2018; Ma and Franzke, 2021). The particular source of predictability that our predictor represents, appears related to ENSO, as correction occurs during El Niño (Fig. 4.10d-e) and La Niña (Fig. 4.10ab-ac) conditions. Also when SHAP values are ordered by time (Appendix, Fig. A.2), we see that the 'West Pacific negative' condition predominantly occurs in the 1999, 2008, end 2010, and begin 2011 summer seasons, which are years with a persisting La Niña (Jong et al., 2020). This implies that ENSO's inter-annual variability can be important for sub-seasonal forecasts, and forms a cross-timescale connection (Liu and Alexander, 2007; Hoskins, 2013). ENSO is known to display predictable month-to-month evolution (Chapman et al., 2021), and to be capable of modulating shorter-term variability like summer monsoons (Di Capua et al., 2020b). Surprisingly, the relation of MJO to forecast errors seems weak. We find that the RMM MJO index gets disregarded in the predictor selection (and does not improve scores when added manually).

The current analysis related west Pacific SSTs at initialization directly to errors at valid time. This enabled the ANN-based post-processing to use sources of predictability whose states in the NWP model are potentially biased. A consequence was however that our interpretation of NWP errors lacked intermediate predictors, making it difficult to diagnose which exact part of the pathway is imperfectly represented. Shortcomings in the representation of teleconnections can namely reside in many model components. Shortcomings in tropical convection often relate to multiple physical parametrizations (Kim et al., 2018). Diabatic modification of Rossby waves over the mid-latitudes is often poorly represented (Gray et al., 2014). Future studies should investigate the implied pathway in

more detail, particularly the information shared between ENSO and the west Pacific predictor.

Nonetheless, the ANN-architecture combined with XAI proved highly insightful for understanding conditional NWP errors and their physical circumstances. Three aspects should be minded for future applications: First, balanced train-validation-test splits are required. Without training on samples from the full p_{raw} range (and the global warming trend it contained), robust corrections of p_{raw} could not be learned. One avenue to prevent out-of-range values can be to train on simulations of the future climate. This will require 'perfect model' assumptions to generate pseudo-observations. Second, the method cannot be immediately applied in operational settings. To summarize the initial state of relevant sources of predictability, we based our observed predictors on multi-day anomalies from reanalyses, which are not available in real-time. Operational analyses usually only stretch for a few hours, meaning that multiple analyses would need to be concatenated. Third, in operational practice there is also a risk that unprecedented events lead to out-of-range input values. Thorough knowledge about the ANN's failure conditions would be required.

4.6 Conclusion

This study demonstrated that ANN-based post-processing improves probabilistic forecasts of monthly summer temperature exceeding the median, with a lead time of 12 to 15 days before the start of the period. Raw ECMWF probability forecasts did not always outperform the climatological and climate-change trend benchmark, but the ANN-corrected predictions outperformed both. The ANN bases its corrections on a shallow neural network architecture and three predictors, besides the raw ECMWF temperature forecast. One of these predictors represents a source of subseasonal predictability with influence on the Euro-Atlantic circulation. Using the state of this predictor at initialization, the ANN is able to correct conditional errors that are made when forecasts lack an atmospheric response.

Detailed explanations were obtained with XAI, specifically 'input times gradient' and SHAP, which quantifies the contributions to each ANN-based correction. Hierarchical clustering subsequently groups the forecasts into groups with similar corrections for the same physical reasons. These analyses revealed that SSTs in the tropical west Pacific were pre-

dictive of ensuing errors in the NWP forecast. The pattern appears to represent an ENSO-like tropical variability, for which other studies have also shown that NWP models do not perfectly represent the teleconnections. It also appeared that the NWP error is conditional, because raw ECMWF forecasts did not need to be corrected when local anti-cyclonic conditions were already present in the initial conditions.

The ANN architecture developed in this paper corrects two-class NWP probability forecasts of monthly temperature exceedance for 2m temperature over Europe during summer. The ANN learns conditional corrections that can be explained with XAI, demonstrating its ability to identify and understand conditional errors in NWP models.

Chapter 5

Understanding driving processes

To be published as:
van Straaten, C., Coumou, D., Whan, K., van den Hurk, B. and Schmeits, M. (2023) Strengthening gradients in the tropical west Pacific connect to European summer temperatures on sub-seasonal timescales *Weather and Climate Dynamics Discussions* in revision

Abstract

Recent work has shown that (sub-)seasonal variability in tropical Pacific convection, closely linked to ENSO, relates to summertime circulation over the Euro-Atlantic. The teleconnection is non-stationary, probably due to long-term changes in both the tropical Pacific and extra-tropical Atlantic. It also appears imperfectly captured by numerical models. A dipole in west Pacific sea surface temperatures (SSTs) was found to be the best predictor of errors in numerical sub-seasonal forecasts of European temperature. In this diagnostic study we use reanalysis data to further investigate the teleconnection pathway and the processes behind its non-stationarity. We show that SST gradients associated with the dipole represent a combination of ENSO variability and west Pacific warming, and have become stronger since 1980. Associated patterns of suppressed and enhanced tropical heating are followed by quasi-stationary waves that linger for multiple weeks. Situations with La Niña-like gradients are followed by high pressure centers over eastern Europe and Russia, three to six weeks later. Inverted situations are followed by high pressure over

western Europe, three to six weeks later. The latter situation is conditional on a strong meridional tripole in north Atlantic SST and a co-located jet stream. Overall, the sub-seasonal pathway diagnosed in this study connects to patterns detected at seasonal scales, and confirms earlier findings that the summertime connectivity between the Pacific and Europe has shifted in recent decades.

5.1 Introduction

Sub-seasonal forecasts are made with a lead time of 2-6 weeks. For weather at any mid-latitude location, part of the predictability at that lead time originates in the tropics (Vitart and Robertson, 2018). Tropical deep convection and associated diabatic heating generate upper level divergence and vorticity anomalies that force Rossby Waves (Hoskins and Karoly, 1981; Sardeshmukh and Hoskins, 1988; Trenberth et al., 1998). These waves can propagate into the westerly mid-latitude flow and steer associated weather patterns, commonly taking two weeks to establish (Liu and Alexander, 2007; Branstator, 2014; Stan et al., 2017). Impacts on mid-latitude surface weather are especially pronounced when waves are quasi-stationary (Schubert et al., 2011; Wolf et al., 2018; Röthlisberger et al., 2019). An example of a teleconnection from tropical heating is the Madden Julian Oscillation, which in the Euro-Atlantic sector, has a phase-dependent influence on the North Atlantic Oscillation (Cassou, 2008; Henderson et al., 2017; Vitart, 2017).

Since its origins, teleconnection research has mostly focused on winter, as this is a season with higher baroclinicity and more potent Rossby wave propagation (Bjerknes, 1969; Trenberth et al., 1998; Branstator and Teng, 2017). Teleconnections are also thought to have an influence in summer (Cassou et al., 2005). Heating patterns over the western tropical Pacific are of primary importance for summertime quasi-stationary Rossby waves (QSRWs) (Ting, 1994; Behera et al., 2013; Ma and Franzke, 2021). The dipole of enhanced convective activity over the Maritime continent, in conjunction with reduced activity over the west and central Pacific is related to circulation over the Euro-Atlantic sector (O'Reilly et al., 2018; Fuentes-Franco and Koenigk, 2020). One feature known for producing such contrasts in sea surface temperature (SST) and atmospheric heating, is the El Niño Southern Oscillation (ENSO). When ENSO's atmospheric component (the Walker circulation) strengthens, convection over the Maritime continent increases and that over the tropical central

Pacific decreases (Bjerknes, 1969). Such a (developing) La Niña episode was found to have supported the prominent blocking that was part of the Russian heatwave of 2010 (Schneidereit et al., 2012).

Such an ENSO-induced contrast in convection over the tropical central Pacific and Maritime continent, is oriented along the equator. Diagnosis of heating differentials also reveals that meridional contrasts have a role in steering mid-latitude flow (Ding et al., 2011). Such contrasts reflect activity of the Indian Summer Monsoon, and the western north Pacific monsoon. Both meridional and equatorial SST contrasts have increased rapidly since 1990 by what is called the 'west Pacific warming mode' (Funk and Hoell, 2015). It consists of concentrated warming over the Maritime continent and the western north Pacific (WNP). This is found to be a response to anthropogenic emissions (Funk and Hoell, 2015) and has strengthened the Walker circulation (Funk et al., 2018; Lee et al., 2022).

Coinciding with west Pacific warming, observed connectivity between the Pacific and Euro-Atlantic circulation appears to have strengthened (O'Reilly et al., 2019b; Sun et al., 2022). This strengthening is not a consequence of internal atmospheric variability but is a response to the SST trends (O'Reilly et al., 2019b), potentially influencing current and future Euro-Atlantic circulation. Unfortunately, numerical climate models seem unable to reproduce the observed changes in Euro-Atlantic circulation (Boe et al., 2020), meaning they underestimate the rapid increase of European summer temperature extremes (van Oldenborgh et al., 2022). Also numerical weather prediction (NWP) models have shortcomings in simulating Rossby Wave teleconnections (O'Reilly et al., 2018; Quinting and Vitart, 2019). We previously found that west Pacific SSTs could relate to QSRWs that conditionally led to monthly European summer temperature anomalies. This teleconnection appeared imperfectly represented in the numerical model of the European Center for Medium-range Weather Forecasts (ECMWF) (van Straaten et al., 2023).

The shortcomings of weather- and climate-models can reside in many processes associated with Rossby Wave teleconnections. Whether QSRWs influence a remote location can be modulated by processes related to the forcing of waves, and processes related to wave propagation and amplification (White et al., 2022). In the Pacific region, or further along the propagation trajectory, other sources of heating and vorticity can strengthen the QSRW when they are in-phase, and negate it when they are out-of-phase, dependent on the optimal forcing pattern of the QSRW (Schubert et al., 2011; Kim and Lee, 2022). Extra-tropical Pacific SST anomalies

provide such feedback when preceding dynamics have left in place an anomaly pattern that is in-phase with a summertime QSRW (Vijverberg and Coumou, 2022). In the same way the state of the north Atlantic can permit or hinder QSRW propagation towards Europe (Fuentes-Franco et al., 2022). Also soil moisture depletion, for example over the US, can amplify wave patterns and make them circumglobal (Teng and Branstator, 2019).

A second potential modulator is the way atmospheric jets function as waveguides (Hoskins and Ambrizzi, 1993; White et al., 2022). Waveguides are sharp gradients in the background flow, along which Rossby Waves propagate (Wirth et al., 2018; Manola et al., 2013). The role of jets is noticeable as strong waves often emanate at their exit regions, of which the Euro-Atlantic sector is one (Stan et al., 2017). The characteristics of the background flow, in the form of jet position, width and strength can determine whether a wave response will be of limited longitudinal extent, or circumglobal (Branstator and Teng, 2017).

In this study we want to characterize the west Pacific to Europe teleconnection and evaluate if changes in the Pacific, and associated strengthening of the connection can indeed be an explanation of observed Euro-Atlantic circulation changes. To that end we build an index for SST dipoles over the west Pacific, and an index for European surface temperature more than two weeks later. Such a lagged, sub-seasonal timeframe is different from the concurrent, seasonal diagnostics used in earlier studies (Ding et al., 2011; Behera et al., 2013; O'Reilly et al., 2018). We examine the pathway between the two end-points, and investigate whether the pathway is modulated by (a combination of) the processes described above. Such modulation might explain the conditional subseasonal occurrence. Overall, we hope that our characterization can help in targeting weather- and climate-model evaluations, such that long term projections and sub-seasonal forecasts of European summer extremes can be improved.

5.2 Data

In this study we use the ERA5 reanalysis (Hersbach et al., 2020). We extract daily values of sea surface temperature (SST) and two-meter temperature (t2m), as they will form the respective start- and end-point of the teleconnection. Top-of the atmosphere outgoing longwave radiation (OLR) was extracted as indicator of tropical deep convection, zonal wind

at 300 hPa (u300) as an indicator of jet stream strength and position, and geopotential height at 300 hPa (z300) as an indicator of the QSRW itself. Values were extracted from 1950 until 2021, in a domain that spans from 20°S to 90°N, and 180°W to 180°E, at a spatial resolution of 0.25x0.25°.

Daily values were de-seasonalized by approximating the seasonal cycle with a polynomial that is a function of the day-in-the year (as in Mayer and Barnes, 2021, 2022). We found that a 7-degree polynomial was most suited to accommodate the changing shape of seasonal cycles with latitude. At polynomials below 5 degrees the residuals remained visibly dependent on the season. For each grid cell the polynomial was fitted to the complete set of years from 1950 till 2021. Daily anomalies were then averaged to four-week (31-day) values for t2m and three-week (21-day) values for the other variables. In previous studies we namely found that three-week averages of SST and other variables are at least as related to four-week European t2m as four-week averages (van Straaten et al., 2022, 2023). Both aggregations were executed as a rolling-window averaging, such that one value was recorded each day

The western north Pacific (WNP) region, whose warming as part of the west Pacific warming mode was found to have strong influence on the Walker Circulation, is defined from 10°N–30°N 130°E–170°W (Funk et al., 2018). In this region we record each day the spatial mean, three-week average SST anomaly.

To investigate the role of ENSO we use pre-computed monthly relative ENSO indices (van Oldenborgh et al., 2021), based on the ersstv5 dataset of ocean observations (Huang et al., 2017), in regions Niño 3 and Niño 4 (Trenberth and Stepaniak, 2001). Relative ENSO indices are defined relative to the global average SST between 20°N and 20°S, which makes them less distorted by global warming. We interpolate them linearly to obtain a monthly value each day.

For the Pacific Decadal Oscillation (PDO) we use standardized values of the first principle component of monthly Pacific SST anomalies north of 20°N (Mantua et al., 1997). The pre-computed index provided by NOAA is based on ersstv5, and has the global mean sst anomaly subtracted to make it less distorted by global warming. We interpolate the PDO index linearly to obtain a monthly value each day.

The sub-seasonal predictand is derived from gridded t2m through spatial and temporal aggregation. This is required to capture the overarching variability that affects daily anomalies at multiple moments and locations. When prediction is attempted for individual samples, the variability would be harder to detect and harder to predict with sub-seasonal lead

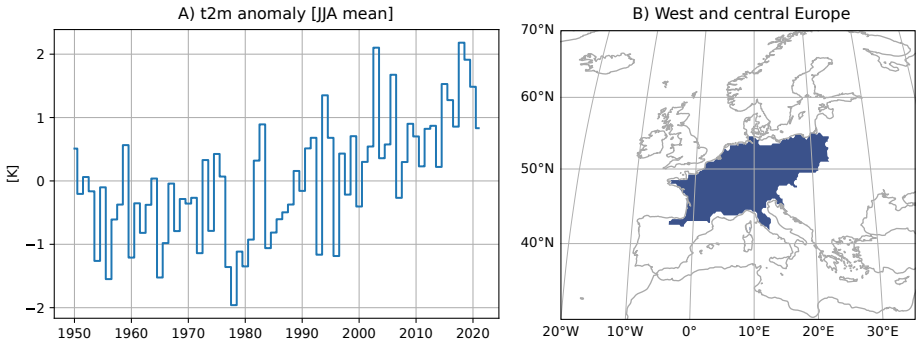


Figure 5.1: Two-meter temperature (t2m) anomaly in west and central Europe. A) Seasonal mean (JJA) ERA5 t2m anomaly in the region, from 1950 to 2021. B) west and central European region.

times (Buizza and Leutbecher, 2015; Wheeler et al., 2017; van Straaten et al., 2020). In a previous study we used hierarchical clustering to find a west and central European region (Fig. 5.1B), in which the average t2m anomaly is predictable when also aggregated to the four-week or monthly time scale (van Straaten et al., 2022). Using rolling averages as described above, we thus create a four-week average response variable which we will refer to as 't2m in week 3,4,5 and 6'. A diagnostic plot of June-July-August (JJA) averages of this variable shows that summer temperatures have been warming. In fact, western Europe has been warming faster than the global average (Christidis et al., 2015), especially since the 1990's.

5.3 West Pacific dipole index

Figure 5.2 displays Pacific SST grid-cells whose variability precedes European t2m by more than two weeks. We correlate three-week-average SST anomalies ('SST in week -2 to 0') to lagged European four-week-averaged t2m in week 3 to 6, while correcting for factors inflating the correlation, like global warming and auto-correlation (Fig. 5.2A). Specifically, we compute the partial correlation between residual SST and residual t2m. Those residuals remain after a linear regression predicts observed SST and t2m anomalies using time and the value of the previous time step (details can be found in van Straaten et al., 2022). Within the full dataset of 1950 to 2021, we test robustness of the partial correlation with a five-fold crossvalidation, leaving out consecutive blocks of 14 years. Grid cells with correlations significantly different from zero in five out of five subsets are

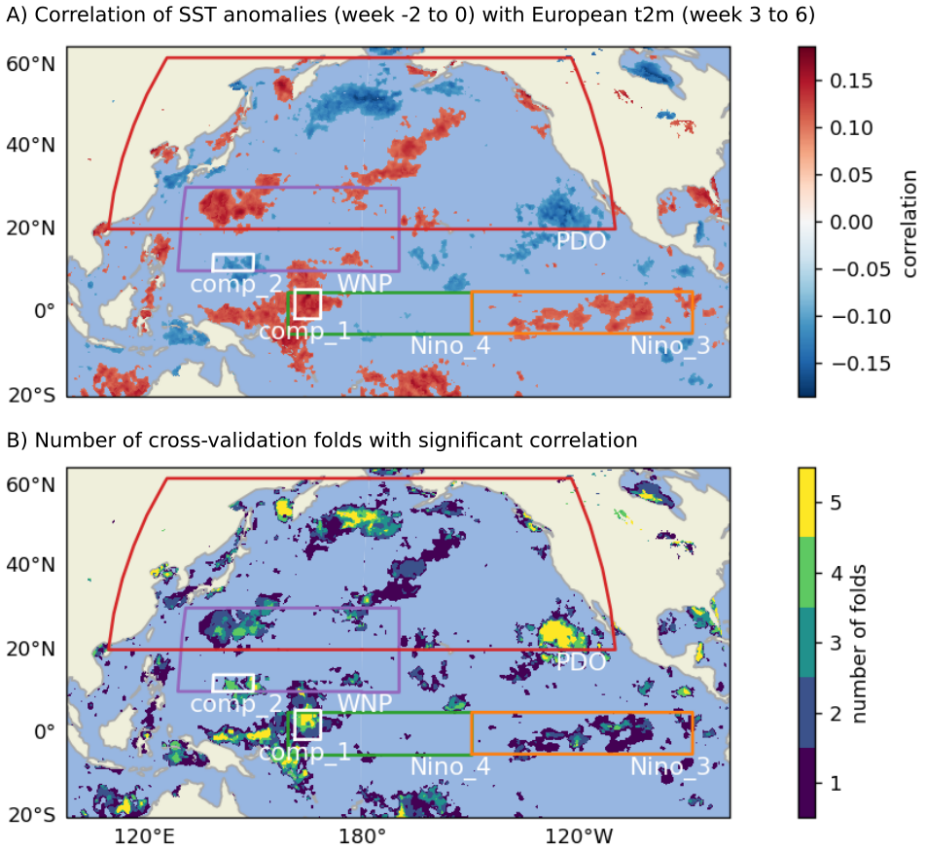


Figure 5.2: Connection between ERA5 Pacific SST anomalies and European t2m anomalies at sub-seasonal timescales in summer (1950-2021). A) Spearman rank correlation between week -2 to 0 SST anomalies and week 3 to 6 western European t2m, corrected for inflation by linear trends, seasonality and auto-correlation. Reported is the mean correlation over 5 crossvalidation folds, constructed by leaving out consecutive blocks of 14 years. B) Robustness of the correlation as measured by the number of folds with significant correlation. Annotated are regions commonly used to capture Pacific variability. The two components of the west Pacific Dipole (WPD) index are highlighted by white squares.

highlighted in yellow (Fig. 5.2B) ($\alpha = 5 \cdot 10^{-12}$, corrected for the false discovery rate (Benjamini and Hochberg, 1995)).

The highlighted SST regions relate to European t2m, and do not correspond perfectly to patterns that explain the highest amount of Pacific variability, like PDO and ENSO. Aspects are however captured. A large cluster of significant cells resides in the region used to define the PDO (Mantua et al., 1997) (Fig. 5.2B). Also at the western edge of the Niño 4 area we see a cluster of cells (called 'component 1'). This cluster is flanked by 'component 2' which lies at 10°N in the WNP area (Fig. 5.2). The respective positive and negative correlations of component 1 and 2 (Fig. 5.2A), hint at situations with anomalously warm SSTs in component 1 and anomalously cold SSTs in component 2, and in which week 3 to 6 European t2m would later be above normal (and vice-versa). With component 2 located in the WNP, and component 1 at the western edge of the central Pacific, their combination appears to capture the combined equatorial and meridional heating contrast known to force a teleconnection towards Europe (Ding et al., 2011; Behera et al., 2013; O'Reilly et al., 2018; Ma and Franzke, 2021). In van Straaten et al. (2023) the contrast was captured with a spatial co-variance predictor and was found to relate to a failure of the ECMWF model to represent the teleconnection. Here we simply capture the contrast by defining a west Pacific Dipole (WPD) index, namely the three-week anomaly in component 1 minus the three-week anomaly in component 2. As SSTs in the tropical west Pacific are, in absolute terms, generally warmer than those in the tropical central Pacific, this definition dictates that: positive WPD values represent a weaker equatorial SST gradient, and negative WPD values represent a stronger equatorial gradient. The reason that positive WPD is defined as 'positive', is its association to above-normal western European t2m (Fig. 5.3C).

We expect the WPD index to relate to other Pacific modes. PDO for instance, comprises a combination of remote ENSO-induced variability in combination with local atmosphere-ocean interactions (Newman et al., 2016). Especially due to spatial overlap in regions, we see that WPD relates to Niño 4 and WNP, which is illustrated by cross-correlations of 0.61 and -0.29 respectively, in the period from 1979 until 2021 (Fig. 5.3B). We do not include the pre-1979 period because it lacks the west Pacific warming mode and the stronger Pacific-Atlantic connections (Funk and Hoell, 2015; O'Reilly et al., 2019b), which would muddy the illustration of current inter-relations. The real difference with for instance Niño 4, is that the WPD-index is designed to specifically target the teleconnec-

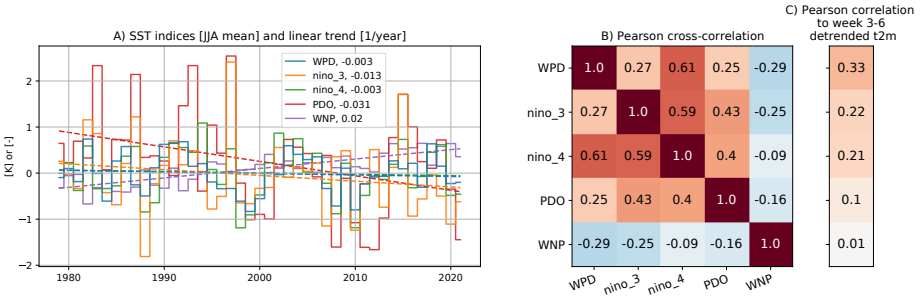


Figure 5.3: Correspondence between the west Pacific Dipole index (WPD: SST in 'component_1' minus SST in 'component_2', Fig. 5.2), and other Pacific indices. A) Seasonal mean time series and linear trend from 1979 till 2021. B) Cross-correlation matrix between the SST indices (JJA only). C) Lagged correlation between SST indices and European t2m in week 3 to 6, corrected for a linear trend in t2m. The SST indices capture SSTs from week -2 to 0 for WPD and WNP, and SSTs from week -3 to 0 for PDO and Niño (see also section 5.2).

tion towards Europe, which is confirmed by the highest correlation of all indices to week 3-6 t2m (Fig. 5.3C).

Correlation between Pacific modes is further illustrated by a year-to-year representation of the time series (Fig. 5.3A). Over the 1979-2021 period the standardized PDO is moving towards a negative phase (-0.031 std/yr). Also Niño 3 displays a slight negative trend (-0.013 K/yr), which is different from Niño 4 and WPD, and opposed to the positive trend in WNP (0.02 K/yr). These opposing trends reflect documented warming in the western Pacific, while the central to eastern tropical Pacific has not warmed (Wills et al., 2022; Seager et al., 2022; Sun et al., 2022; Lee et al., 2022). The result is an increase of equatorial and meridional gradients, and Pacific SST states that are more 'La Niña-like', with a stronger Walker circulation (Funk and Hoell, 2015; Lee et al., 2022), and potentially important influences on tropical-extratropical teleconnections (Schubert et al., 2014; O'Reilly et al., 2019b; Sun et al., 2022).

5.4 Emergence of a teleconnection

We classify the WPD index into three tercile-based categories. The positive phase is when SSTs during week -2 to 0 are anomalously warm in component 1 and anomalously cold in component 2. In the negative

phase this is inverted, and a neutral phase is in-between. Because the WPD index does not have a significant trend (Fig. 5.3A, -0.003 K/yr), we determine tercile thresholds over the entire 1950-2021 dataset. For week 3-to-6 t2m on the other hand, tercile thresholds are computed per rolling window of 21 summers. Otherwise the large trend (Fig. 5.1A) would fill the 'positive' class mostly with samples from recent years. This would distort the lagged t2m response that forms part of our measure of teleconnection strength: namely the number of occasions with (i) *positive WPD* AND *positive t2m response*, plus the number of occasions with (ii) *negative WPD* AND *negative t2m response*.

Temporal evolution of the WPD classes reveals that since 1950 (the beginning of the dataset) the negative WPD phase has strongly increased in frequency (blue line in Fig. 5.4A). Until 2000 this happens mostly at the expense of the neutral phase, and after that also at the expense of the positive phase. The dwindling occurrence of the neutral phase can explain why the tercile distribution shows large changes (Fig. 5.4A), whereas the seasonal mean of WPD had no significant trend (Fig. 5.3A). The increased occurrence of negative WPD phases does agree with WNP warming and the strengthening of equatorial gradients over the Pacific (Section 5.3, Fig. 5.3A).

Concurrent with this development we see that lagged t2m phases in week 3 to 6 increasingly follow the WPD phases in week -2 to 0. The sum of positive and negative responses to respectively, positive and negative WPD, rises beyond values found by chance (grey area, Fig. 5.4B). This roughly occurs for 21-year windows centered in 1990 and beyond, meaning it starts in 1980. The emergence of a significant teleconnection between the Pacific and Euro-Atlantic circulation is reported by other studies, both for the past decades (Wu and Lin, 2012; Lim et al., 2019; O'Reilly et al., 2019b; Sun et al., 2022), and for the near-future (Mayer and Barnes, 2022).

In the remainder of this paper we focus on the period 1980-2021 in which a significant teleconnection has emerged. The joint distribution of WPD and t2m phases is given in Figure 5.4C. Presented are the number of samples in each category (3- and 4-week averages for WPD and t2m respectively, with values recorded each day). Note that the marginal distribution of t2m is uniform, because tercile thresholds are re-estimated in each rolling 21-year window, and that the marginal of WPD is not uniform and shows the relative dominance of negative WPD phases in this period.

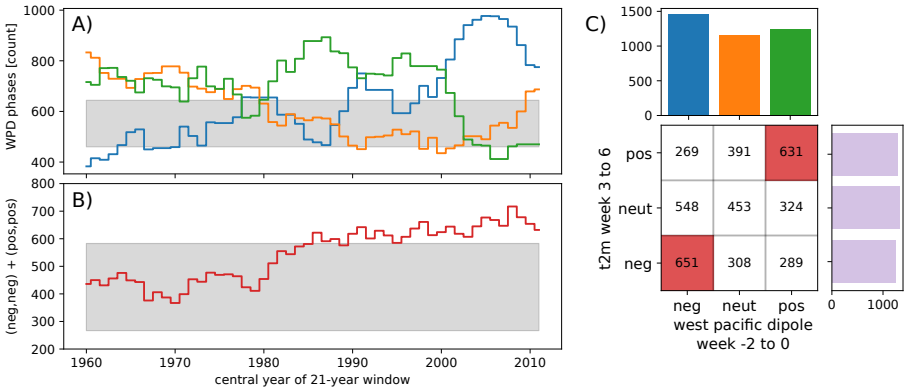


Figure 5.4: Teleconnection between week -2 to 0 west Pacific Dipole index and week 3-6 European t2m, as measured by correspondence in tercile classes (negative, neutral, positive). A) Prevalence of the three WPD classes in a moving window of 21 summer seasons (blue: negative WPD; orange: neutral WPD; green: positive WPD), with tercile thresholds determined over the entire dataset, from 1950-2021. B) Total occurrence of the negative and positive phase of the teleconnection, i.e. a negative WPD preceding a negative t2m response plus a positive WPD preceding a positive t2m response, as counted for each 21-season window, whereby the t2m tercile thresholds are re-computed in each window. Grey areas in (A) and (B) denote the 0.025 and 0.975 quantiles of the uncertainty distribution when counting is performed on 21 summer seasons that are randomly sampled from the 1950-2021 dataset (500 repeats, with replacement). C) Distribution of tercile classes for WPD (x-axis) and European t2m (y-axis) as counted for each 21-season window from 1980 till 2021 (central years 1990-2011) when the teleconnection (both in negative and positive phase) was present in a statistically significant way (as shown in panel B).

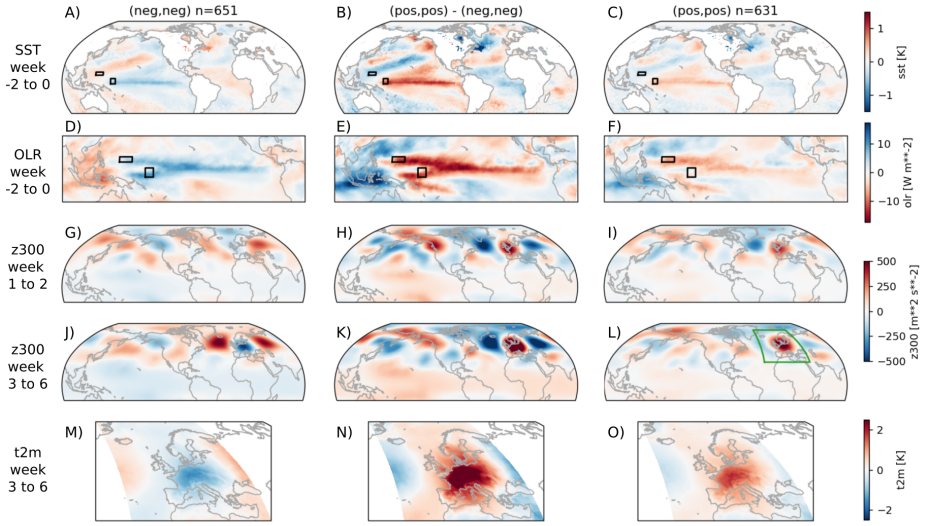


Figure 5.5: Composite plots illustrating the teleconnection between week -2 to 0 west Pacific SST and week 3 to 6 west-central European t2m, based on composite anomalies from the period 1980-2021 (climate normal also estimated from 1980-2021). Left column: Samples in which negative WPD phases precede the negative t2m class (n=651). Right column: Samples in which positive WPD phases precede the positive t2m class (n=631). Middle column: Difference between panels in the right and left column. From top to bottom: SST pattern and OLR in week -2 to 0, the subsequent z300 response in week 1 and 2, and the eventual impact on z300 and surface temperature in week 3 to 6. Black insets in panels A-F show the two components of the West Pacific Dipole index. Green inset in panel L serves as reference for the spatial extent of surface t2m in the bottom row.

5.5 Forcing, quasi-stationary wave, and surface imprint

We now investigate the spatial and temporal patterns that occur during the period that the teleconnection is significant. Different components of the pathway (i.e. SST, OLR, z300, t2m) are plotted as rows in Figure 5.5. The negative phase, positive phase and their difference are plotted in the left, right and middle column, respectively.

It is clear that the WPD index captures a large geographical pattern with pronounced SST and OLR anomalies across the Pacific ocean, de-

spite being defined by small boxes (Fig. 5.5A-F). Visually, the SST states in week -2 to 0 represent a combination of three patterns: (i) ENSO, which is visible in the La-Niña-like equatorial contrast between anomalously cool SSTs in Niño 4 and anomalously hot SSTs around the Maritime continent in Fig. 5.5A, and in the El-Niño-like contrasts in Fig. 5.5C, (ii) the west Pacific warming mode, which resembles a tilted red 'V' and connects the above-normal SSTs around the Maritime continent to the extra-tropics, in north-eastward and south-eastward direction (our Fig. 5.5A and Fig. 1A of Funk and Hoell (2015)), and (iii) the PDO-like pattern that the north-eastward extension of warm SSTs also resembles, which is known to provide summertime predictability for the eastern US (our Fig. 5.5A and Fig. 1F of Vijverberg and Coumou (2022)).

In OLR we see that the WPD phases correspond to clear contrasting signals in anomalous deep convection, but that these contrasting signals need not be aligned with the dipole in anomalous SST and the WPD boxes themselves (Fig. 5.5D-F). The heating during weeks -2 to 0 reflects equatorial contrasts like the enhanced heating over Niño 3 and 4, in combination with reduced heating over the Maritime Continent (Fig. 5.5F). But OLR also reflects meridional contrasts like increased activity of the western north Pacific monsoon (heating centered around 15°N, east of the Philippines, in region component 2, Fig. 5.5F). Both types of heating contrasts were found to be important by Ding et al. (2011). Noteworthy is that the prevalent heating anomaly in the Maritime Continent (either suppressed or enhanced) extends into the Indian ocean (Fig. 5.5E), confirming that west Pacific forcing is not always separable from forcing in the Indian Ocean (as in Behera et al., 2013; Ma and Franzke, 2021).

The forcing at weeks -2 to 0 translates into an initial atmospheric response in weeks 1-2 (Fig. 5.5G-I). Centres of action (locations with the largest z300 anomalies) are found from the north Pacific to Eurasia, showing the considerable longitudinal extent of the QSRW (Fig. 5.5H). Over the north American continent the wave pattern resembles the pattern found by Vijverberg and Coumou (2022) (our Fig. 5.5H, their Fig. 1b), which here extends towards Europe. The extension involves a prominent center of action south of Greenland, in the north Atlantic, which was also found by O'Reilly et al. (2018). Furthermore, this cyclonic anomaly, in combination with the west/central European high pressure (Fig. 5.5I), bears similarity to the Summer East Atlantic (SEA) pattern (Fig. 1a of Wulff et al. (2017)). The tropical forcing of SEA was thought to come primarily from the Eastern Pacific and Carribean (Wulff et al., 2017), but our results indicate a possibility of west Pacific forcing as well.

After development in weeks 1 to 2, the lifetime of the QSRW extends into week 3 to 6 (Fig 5.5J-L). This results in negative t2m anomalies in west and central Europe when WPD was negative and positive t2m anomalies when WPD was positive (Fig 5.5M,O). In the negative phase a cool western Europe (relative to the climate normal from 1980-2021) is flanked by high pressure and high t2m in Eastern Europe and Russia (Fig. 5.5J,M). This cold-warm t2m dipole has been reported by earlier studies as well (Behera et al., 2013). In fact, given that the negative phase of WPD is occurring more frequently since 1990 (Fig. 5.4A), we expect that in recent summers this t2m dipole pattern has become more frequent, which is exactly what Lee et al. (2017) have reported.

5.6 Modulation

As above, we can look at corresponding WPD and t2m phases and study the teleconnection pathway when it is fully present. But we are also interested in cases where positive or negative WPD phases produce Pacific heating contrasts, but where the full atmospheric response is lacking and the QSRW fails to reach Europe because of a modulating process.

ENSO's influence on the large scale background circulation might be such an indirect modulator (Ding et al., 2011; Schneidereit et al., 2012; O'Reilly et al., 2019b). An example is the cooling of the atmosphere in the entire tropical belt by a strong La Niña episode. This changes the zonal mean equator-to-pole gradient and therefore the waveguide (Ding et al., 2011). Modulation by ENSO can also happen in a zonally asymmetric way by strengthening the western north Pacific Monsoon, or Indian Summer Monsoon (Ding et al., 2011; Di Capua et al., 2020b). The evolution of ENSO, i.e. whether it is strengthening, decaying, or persisting, is important as well (e.g. Jong et al., 2020). Strong expression of a pattern in the preceding winter can namely leave an imprint on extra-tropical SSTs, which leads to favourable or unfavourable diabatic interaction with QSRWs in subsequent seasons (Vijverberg and Coumou, 2022).

We investigate whether positive and negative phases of the summertime teleconnection occur under distinct ENSO evolutions, particularly focusing on ENSO evolution in the central to eastern Pacific (Niño 3) which is less directly related to our WPD index (Fig 5.3B). Plotted in Figure 5.6 is the monthly mean evolution of the relative Niño 3 index in each combination of phases (negative WPD phases in blue, positive WPD phases in red, with opacity denoting whether or not we see a Eu-

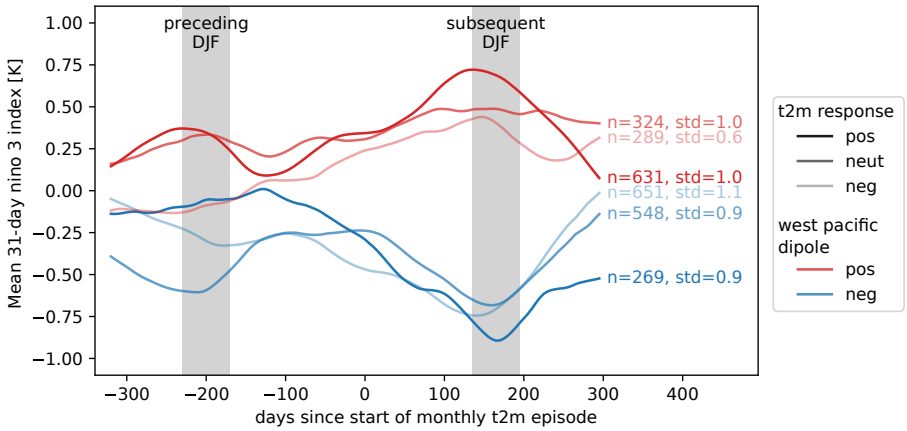


Figure 5.6: Evolution of the monthly mean relative Niño 3 index, computed per group of samples (determined in JJA). Plotted are groups with positive (red) and negative (blue) WPD phases (based on the SST anomaly in week -2 to 0, sometime in JJA). The phases tend to occur when ENSO evolves from neutral to El Niño or La Niña. Plotted with opacity is the European t2m response in week 3 to 6 (a moment in JJA that we set to zero on the x-axis). This response seems independent of the ENSO state. Samples come from the period 1980-2021. Grey boxes indicate moments that are guaranteed to fall in winter (DJF), as this depends on whether x=0 occurs early or late in summer. Number of samples and standard deviation of DJF values are annotated per group.

ropean t2m response). As expected, positive and negative WPD phases are El-Niño-like and La-Niña-like, respectively (red lines are generally on top of blue lines, Fig. 5.6). This differentiation between red and blue actually increases when moving from preceding DJF, via JJA (zero on the x-axis), to subsequent DJF. This suggests that summers with non-neutral WPD phases relate to the strengthening of ENSO: The negative summertime WPD phase occurs during years when ENSO migrates from an average neutral state to a pronounced La Niña state. Vice versa, the positive WPD phase occurs when ENSO migrates from an average neutral state to an El Niño state. This does not imply that ENSO also fully determines the resulting t2m response: A negative t2m response is likely after a negative WPD phase (and vice versa) given the significance of the teleconnection (Fig. 5.4C). But within each WPD group (comparing blue lines of different opacity, and red lines of different opacity), there are no

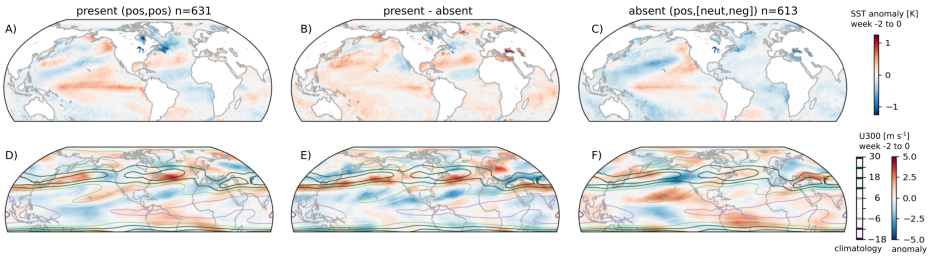


Figure 5.7: Modulation of the teleconnection when the west Pacific dipole in week -2 to 0 is in its positive phase. Left column: samples with positive t2m response in week 3 to 6, meaning a positive phase teleconnection (n = 631). Right column: samples resulting in a neutral or negative t2m response, meaning an absence of the teleconnection (n = 613). Middle column: difference between panels in the left and right column. Top row: composite anomalies of SST in week -2 to 0. Bottom row: composite U300 anomalies in week -2 to 0. Contour-overlay shows the summertime climatological U300 value. Composites are extracted from the period 1980-2021 (climate normal also defined from 1980-2021).

distinct ENSO states related to the 'pos,pos' (darkest red) and 'neg,neg' (lightest blue) teleconnection occurrences (Fig. 5.6).

If not by ENSO, the QSRWs from west Pacific heating anomalies can also be modulated by other factors. For negative teleconnection phases little was found, but for positive WPD phases we plot the spatial composite patterns of SST and u300 (Fig. 5.7). The composite plots show the state in week -2 to 0, so before the QSRW occurs, with on the left samples leading to a lagged, positive European t2m response, and on the right all those that do not. The difference plot in the middle column shows that large SST differences are found in the Atlantic (Fig. 5.7B). A strong meridional cold-warm-cold tripole in the north Atlantic is associated with positive t2m responses (Fig. 5.7A). Diagnosed by earlier studies, this tripole pattern relates to the strength of the oceanic gyres in the north Atlantic, which are partly driven by wind-stress on the ocean surface (Häkkinen et al., 2011). In the configuration of Fig. 5.7A, cyclonic circulation and cold SSTs prevail south of Greenland, while heat is transferred from ocean to atmosphere (Häkkinen et al., 2011). The tripole pattern can occur already in late winter and early spring and is known to precede the summertime SEA pattern (Fig. 5.5I) (Gastineau and Frankignoul, 2015; Ossó et al., 2020; Wolf et al., 2020; Beobide-Arsuaga et al., 2023). Spring SST anomalies are particularly strengthened by a two-way

coupling between ocean and jet stream. This happens as soon as the north Atlantic jet migrates northward with the change of seasons (Wolf et al., 2020; Ossó et al., 2020). Indeed we see that sharp meridional SST gradients are co-located with jet stream position over the Atlantic (Fig. 5.7A,D). Strong u300 anomalies are present in week -2 to 0, and show a stronger and narrower Atlantic jet as compared to the broader climatological mean jet (green contours, Fig. 5.7D). In the case without SST tripole, and without a t2m response in week 3-6, the jet is less strong (Fig. 5.7F). Strong and narrow jets form better waveguides (Manola et al., 2013; White et al., 2022), and in this case precede the QSRW and the t2m response. We therefore deduce that the interplay of SST tripole and Atlantic jet modulates the teleconnection by longitudinally guiding QSRWs from the west Pacific towards Europe.

5.7 Discussion and conclusion

In this study we have defined a west Pacific Dipole (WPD) index that captures the strength of a dipole SST pattern in the western Pacific. These contrasts in SST occur in both the equatorial and meridional direction, and were shown to relate to changing patterns of deep convection and large-scale SST changes across the Pacific. Specifically, the WPD targets those heating patterns that excite Quasi-Stationary Rossby Waves (QSRW) that potentially affect west and central Europe more than two weeks later. Such teleconnections are known mechanisms for sub-seasonal predictability. Cross-correlations make clear that the emphasis of the WPD is different than, but still relates to, well-known Pacific modes like ENSO and PDO (Fig. 5.3B,C). Particularly, the positive and negative WPD phases in summer coincide with the strengthening of ENSO (Fig. 5.6). Also, the index is not overly sensitive to the exact placement of its boxes on the correlation map (Fig. 5.2). An index based on three components, with one northward box inside the PDO region, led to similar results.

A prominent result of this study is that negative WPD phases have become more frequent over recent decades. In this phase, convection over the Maritime Continent and western north Pacific is enhanced, and that over the central Equatorial Pacific suppressed. Its increased occurrence reflects a long-term shift towards 'La Niña-like' states with stronger equatorial SST gradients, as a consequence of the west Pacific warming relative to the central tropical Pacific. This west Pacific warming mode is thought

to be a response to anthropogenic forcing, and to strengthen the Walker circulation (Funk and Hoell, 2015).

Coinciding with the long term SST changes we find that the WPD phase during weeks -2 to 0 becomes an important predictor for the European t2m response in week 3 to 6 (Fig. 5.4B) (also van Straaten et al., 2023). We diagnose that this summertime Pacific to Euro-Atlantic teleconnection emerged after 1980, which agrees with other studies (O'Reilly et al., 2019b; Sun et al., 2022). We should however be cautious about concluding that the teleconnection has been absent before 1980, as it can also relate to data quality, which improved with the advent of satellite observations (Hersbach et al., 2020). Nonetheless, we show that for the period of 1980 till 2021 connectivity is significant, and that the teleconnection pathway can be well understood. Following the heating anomalies, a QSRW develops in week 1 to 2 and consists of known centers of action. These centers span the north American continent (known from Vijverberg and Coumou (2022)), and then extend eastward to encompass a large center south of Greenland, and one over west and central Europe (resembling the known Summer East Alantic (SEA) pattern (Wulff et al., 2017)). An additional center of action is located over eastern Europe and Russia, and has a sign that opposes the sign over west and central Europe (known from Behera et al., 2013; Lee et al., 2017).

Given this significant emergence in recent decades, the teleconnection explains some recent changes in the summer circulation over Europe. The increased frequency of the negative WPD phase, would according to its corresponding QSRW, induce a warming in eastern Europe and Russia. Indeed, high pressure has become more prevalent there (Lee et al., 2017; Kim and Lee, 2022), and the region has seen a very strong increase in heatwaves (Rousi et al., 2022), with average summer t2m increasing more than in our Western European target region (Teng et al., 2022).

To a smaller extent also Western European t2m has been increasing (Boe et al., 2020; van Oldenborgh et al., 2022). During that time negative WPD has roughly doubled in frequency (Fig. 5.4A), and has related to the cold Western European t2m tercile, relative to the warming trend. This means that if the teleconnection is influencing Western Europe through circulation changes, then its effect among all other factors, woud be a dampening of the warming.

With its reduction in frequency, positive WPD has become less of a factor for Euro-Atlantic circulation (Fig. 5.4A). However, we have seen that the QSRW following positive WPD is potentially modulated by the situation in the Atlantic (Fig. 5.7). A combination of SST tripole and

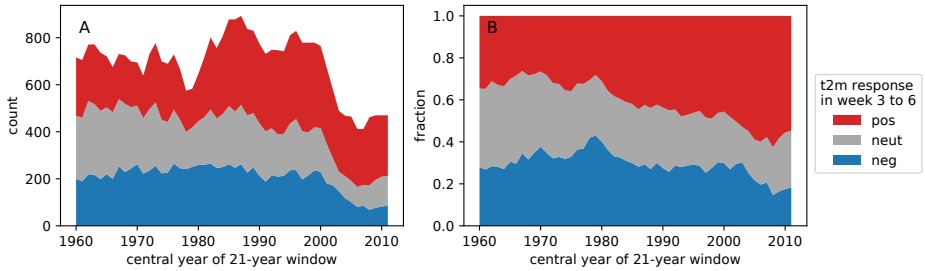


Figure 5.8: A) Count of positive WPD phases and its resulting t2m responses in a moving window of 21 summer seasons. Positive WPD phases are recorded when the west Pacific Dipole index in week -2 to 0 is in its upper tercile class (thresholds are determined over entire dataset, from 1950-2021). The total count is subdivided into positive WPD phases resulting in negative, neutral or positive t2m anomalies in week 3 to 6 (respectively: blue, grey and red). Tercile thresholds for t2m are recomputed in each window (as in Fig. 5.4). B) As panel A, but presented as a fraction of the total count.

a strong and narrow jet in the north Atlantic allow the QSRW to reach western Europe. Prominent in this modulation are the relatively cold SSTs south of Greenland (agreeing with Fuentes-Franco et al., 2022). Such relatively cold SSTs have become prevalent since the 1980's (Chemke et al., 2020). This implies that although the total count of positive WPD phases decreases, those that do occur are more likely to generate a QSRW that reaches western Europe and result in warm t2m states there (red, Fig. 5.8B), as compared to positive WPD phases resulting in neutral or negative t2m anomalies (grey and blue, Fig. 5.8B).

We suggest that the interplay between long-term Pacific changes and long-term north Atlantic changes be researched further. Relevant in this interplay, are the cold north Atlantic SSTs, which can become more prevalent if Atlantic meridional overturning circulation slows down. Climate model experiments suggest that such slowdown might follow from further anthropogenic forcing and can generate a summertime high pressure response not far from the UK (Haarsma et al., 2015; Rousi et al., 2021). Relevant for the interplay are also long term changes that potentially affect jet stream strength, like Arctic amplification and aerosols (Coumou et al., 2018; Dong et al., 2022).

Such research should probably not be conducted with numerical climate models only. Our analysis namely emphasized features that climate

models have difficulty capturing. One is the strengthening of equatorial gradients in Pacific SST, which we detected as the increase of negative WPD phases. We know that climate models are unable to reproduce the observed strengthening, and simulate a weakening instead (Funk and Hoell, 2015; Wills et al., 2022; Seager et al., 2022; Lee et al., 2022). Worrying is that even with prescribed SSTs, climate models fail to reproduce the associated dynamical response: an increased prevalence of high pressure over western and eastern Europe (Boe et al., 2020). It thus appears that climate models do not represent the detected QSRWs well. A similar failure to represent the teleconnection pathway happens in the ECMWF model, despite being initialized with observed SSTs (van Straaten et al., 2023).

Resolving these issues is challenging because the shortcomings of climate models force us to use observations or reanalysis products. These are of limited length, with few independent samples as a result. It is also challenging because a full theoretical understanding of QSRWs on different space and time scales does not exist (White et al., 2022). We hope that the targeted WPD index provides future diagnostic studies with a starting point. An improved representation of the Pacific-to-Europe connection in weather and climate models has a lot to offer. One is a better projection of European summer extremes (van Oldenborgh et al., 2022). The other is a conditional opportunity to forecast European summer circulation more than two weeks in advance (Mariotti et al., 2020; van Straaten et al., 2023).

Chapter 6

Synthesis

6.1 Introduction

In Chapters 2 to 5 of this thesis I explored various roles that Machine Learning can play in improving sub-seasonal forecasts. These roles involved detecting sub-seasonal variabilities, issuing stand-alone forecasts, post-processing NWP forecasts, and understanding sub-seasonal sources of predictability. Below I first discuss the findings for each of the four research questions. Then I discuss the broader context in which my methods exist and raise some considerations for future forecasting systems.

6.2 Research findings

6.2.1 At what spatio-temporal scales can we detect predictable sub-seasonal t2m variability in Europe?

In Chapter 2 I evaluated how well the state-of-the-art EPS from ECMWF predicts different combinations of space-time scales. The spatial dimension of these scales was data-driven, as I tasked a hierarchical clustering algorithm to group locations whose daily weather states are governed by the same overarching variability. For each combination I then quantified the maximum lead time with which the average t2m could be skillfully forecast. This so-called 'forecast skill horizon' was determined both with, and without simple statistical post-processing using standard non-homogeneous gaussian regression.

My first finding was that this standard post-processing method extended the forecast skill horizon by about three days. Second, forecast

horizons were longer in winter than in summer (also found in Goutham et al., 2022). Thirdly and importantly, forecast horizons lengthened when spatial and temporal aggregation were applied, up to the point that they became sub-seasonal for some regions in Europe (i.e. forecast skill extended into the third and fourth week after initialization). The answer to the research question thus appears to be a general intuition: namely that sub-seasonal t2m variability is best discernible and predictable at large spatial and temporal scales. However, whereas the effect of temporal aggregation was an unambiguous lengthening of the forecast horizon, the effect of spatial aggregation was two-sided (also noted in Young et al., 2020). Heavy spatial aggregation in both winter and summer tended to decrease the forecast horizon in some regions. Only in winter could sub-seasonal predictability emerge again with further aggregation to continental-scale t2m (Fig. 2.4). I posed that the reason for the initial loss of local predictability could be the existence of sub-seasonal processes that only affect t2m states in their vicinity.

This supposition of a process based on a measure of skill, can be treacherous. In essence the discovery of skill only implies that aggregate atmospheric variability at a certain scale is predictable with sub-seasonal lead times. It does not imply that the process responsible for the variability is also 'sub-seasonal', meaning that it occupies a well-defined place within the variability continuum at exactly the detected scale. For instance, a predictable tendency captured by an 11-day aggregate in Chapter 2, could be a full 'slave' of a seasonal phenomenon that enhances the probability of a certain weather state during the entire summer ('slave' terminology is borrowed from Hoskins, 2013). I will now discuss to what degree the detected predictable variabilities seem to be well-defined within the variability continuum.

First, detection can be muddied by the 'fast' end of the continuum. Daily and local predictability can namely give rise to sub-seasonal predictability, despite the fact that predictability strongly decreases with lead time, which is a fact found from surface variables (Vigaud et al., 2017; Ferrone et al., 2017; Lavaysse et al., 2019; Pyrina and Domeisen, 2022), to upper atmospheric geopotential (Buizza and Leutbecher, 2015; Mastrangelo and Malguzzi, 2019), to regimes (Büeler et al., 2021; Cortesi et al., 2021). Indeed, the decline of daily predictability is also present in my results, but it is also true that beyond day 10, daily t2m values are still slightly predictable (Fig. 2.3a). This means that predictability of a larger-scale aggregate, say the average t2m from day 11 to 21, could result from predictability in days 11, 12 and 13, combined with their weight on

the average. And similarly, the predictability of a spatial aggregate could result from excellent predictability at a few locations.

Such spatial smearing is indeed what happened with the disappearance of local predictability discussed earlier. However, sub-seasonal predictability also emerged again at the continental scale in winter, without any of the immediate underlying regions displaying that same predictability (Fig. 2.4). Similarly but for time, results in Chapter 3 showed that monthly average West and Central European t2m (week 3 to 6) was more predictable than its initial two-week constituent (week 3 to 4) (Fig. 3.7). These two gains in predictability make clear that when variability is most skillfully forecast at the largest of a set of scales, it is unlikely that smaller-scale processes are responsible. Hence, my results do attest to some separation at the 'fast' end.

Second, definitions can also be muddied by the 'slow' end of the continuum. Attesting to this are the sub-seasonal forecast horizons found for daily t2m values in Iceland, which I suggested could relate to multi-annual oceanic variability (Fig. 2.5). Also results of Chapter 3 demonstrated that a chunk of sub-seasonal forecast skill can come from a long-term global warming trend (see also Wulff et al., 2022). The skill that remained when comparing to a global warming benchmark was however still significant, at least for predictions of West and Central European t2m in week 3 to 6 (Fig. 3.5). The process responsible for that predictability is therefore not (just) global warming. This attests to some separation at the 'slow' end.

So far the results thus support the endeavor laid out in the remaining research questions. That is namely to go from the predictable monthly variability in West and Central European t2m to the driving processes, and interpret those as (sub-)seasonal.

6.2.2 Can a Machine Learning model forecast sub-seasonal variabilities in summer, disentangle the driving processes, and have its learned representation interpreted by humans?

Results in Chapter 3 demonstrate that the answer to this question is in principle 'yes'. The search for driving processes is however challenging because the continuum to investigate is larger than in the previous question. It not only encompasses the predictable sub-seasonal t2m variability, but includes oceanic, land-surface and other atmospheric variables as well.

Practically, Chapter 3 employed a continuum of nine variables, eight time scales, and large geographical domains. This is beyond what previous

data-driven studies of European sub-seasonal predictability attempted. Meeting the challenge, while using only 40 summers worth of ERA5 data, limited the set of suitable ML methods. My results show that a two-step approach can be an option. First, a correlation- and clustering-based dimensionality reduction detects potentially relevant oceanic, terrestrial and atmospheric predictors. This means that information contained in for instance the joint variability of neighbouring oceanic locations, is supplied as a single predictor. Without this step, the ML model would have to represent all joint variabilities within an atmospheric, oceanic or terrestrial variable from the bottom up. Now it only needs to learn the conditional interactions between predictors to make actual forecasts. Still, even with the dimension reduction, the number of remaining predictors (300) was large relative to the number of independent samples. I showed that Random Forests are suitable ML models in this setting (see also Weirich Benet et al., 2023), because with the right hyper-parameters they did not overfit. Hence they could skillfully and valuably forecast whether average t2m in week 3 to 6 in West and Central Europe would exceed the climatological median or upper tercile.

Such successful statistical forecasts should be interpreted as a fortunate combination of three facts. First, a sub-seasonal window of predictability exists (this is not equally true for all scales of variability, see previous question). Second, the signal of the driving and interacting processes is sufficiently present in the available data. Third, the complexity of the interaction does not exceed the capacity of the ML model to represent it.

Regarding the latter point, it is clear that the answer to the research question thus heavily depends on whether a suitable ML model and hyperparameters were chosen. These models namely exist on a scale of different capacities: from high-capacity deep learning to low-capacity linear regression.

At the deep-learning end of the scale, all temporal and spatial information of each variable is supplied as separate inputs, leaving the algorithm to learn and represent all relevant variability from the bottom up (Schultz et al., 2021). The associated benefit is that it can implicitly represent almost any conditional interaction (a topic we will revisit in the outlook). The downside is that deep learning cannot work without large quantities of independent data (He et al., 2021; Miloshevich et al., 2023).

At the other end of the scale, linear-regression has a low capacity to represent interactions and requires an elaborate filtering of the input data. This means that a linear model can only leverage a complex combination

of processes when the interaction is explicitly represented and supplied in the form of a single predictor. Such representations are usually not learned from data, but are constructed by experts. They therefore preclude the discovery of unknown processes. That is unfortunate given that discovery is one of the main reasons for applying ML in this thesis. Obviously, the employed two-step model was chosen to lie inbetween these two endpoints.

Even with more data, the model with the highest capacity might not be the most suitable one. With increasing capacity, models can be prone to representing spurious relations. Either because these relations occur by chance in a specific portion of data (Caldwell et al., 2014), or because they are confounded by a third factor. Think of the correlation (and not causation) between the number of cell-phone towers and newborns within a geographical area. These do not represent causal pathways, though they are often interpreted that way (Lapuschkin et al., 2019). This means that the learned representations of ideal ML models include as few spurious relations as possible (Trenary and DelSole, 2022) and also account for the observational uncertainties that they have incorporated (Buizza et al., 2022; Geer, 2021). An ML model can only help humans in disentangling the driving processes when it makes its forecasts for the correct (physical) reasons.

In Chapter 3 I investigated the correctness of the learned representation with TreeSHAP and permutation importance. These two eXplainable AI (XAI) methods attribute the forecasts to the statistically responsible predictors. Many of the statistical relations aligned with existing physical knowledge: namely that patterns in SST, deep soil moisture, snow cover and sea ice are of prime importance. These variables evolve slowly and can, across time and space, generate an atmospheric tendency toward a certain weather type (i.e. can generate a window of predictability). Because of the delay inherent in these interactions, physical intuition also dictates that the relevant information for next-day weather should mostly reside in atmospheric states. The ML model reproduces this behavior as well. These two findings give confidence that also the more surprising statistical connections in the ML model represent actual physical processes. These might have been previously undiscovered because numerical models poorly represented them.

6.2.3 Are the driving processes properly simulated in an NWP model, and if not, can these conditional errors be corrected through statistical post-processing?

From previous research it is already known that NWP models have shortcomings that influence sub-seasonal predictions. Examples are shortcomings in stratosphere-troposphere coupling in winter (Cohen et al., 2018) and soil moisture processes in spring (Dutra et al., 2021). Generally we can thus expect that not all driving processes are properly simulated. In Chapter 4 I evaluated whether this also concerns ECMWF forecasts of a particular window of predictability: the probability that monthly average t2m in west and central Europe exceeds a certain quantile, in summer. This predictand was found to be predictable by the stand-alone ML forecasts of the previous chapter.

The adopted strategy was to first correct conditional errors of the ECMWF model with ML-based post-processing (i.e. tackle the last part of the question), then to understand under what physical circumstances the corrections are applied, and lastly to infer from those circumstances what the likely misrepresented driving process is (tackle the first part of the question). Logically, the conditional corrections required a more elaborate post-processing method than the simple un-conditional post-processing used in Chapter 2. The solution was an artificial neural-network (ANN) architecture based on Scheuerer et al. (2020).

The first innovation over Scheuerer et al. (2020) was the inclusion of predictors from the state observed at initialization (i.e. ERA5), besides predictors derived from ECMWF forecasts themselves. Most post-processing applications only use the latter, even though it can be suspected that errors accumulate from initialization onward, which leads to biased NWP states at valid time (in this case the state in week 3-6). Indeed, my results show that predictors representing the initial state of sub-seasonal processes (before errors can accumulate) are useful for correcting t2m forecasts at valid time.

The second innovation was to force the ANN to learn a correction factor, which it must apply to the EPS probability forecast to come up with a post-processed forecast. First, this makes the post-processing trend-aware, which is beneficial in a changing climate (Heinrich et al., 2021; Shao et al., 2022). Second, this allowed me to directly investigate the processes related to each correction: not by attributing probability forecasts to the statistical predictors (like in Chapter 3), but by attributing the

correction factor directly (comparable to Silva et al., 2022). KernelSHAP and 'input times gradient' were used as appropriate XAI tools.

Overall, Chapter 4 finds that the conditional corrections improve forecast skill for t2m exceeding the climatological median. The ANN specifically learned that certain correctable errors related to a dipole in tropical west Pacific SSTs at initialization. Isolating those circumstances, I show that the SST pattern relates to the observed development of atmospheric Rossby waves that generate a tendency in Western European t2m towards a high or low probability of median exceedence. The ECMWF model lacks the developing wave and the associated window of predictability in t2m. The answer to the research question thus appears to be that this sub-seasonal teleconnection is imperfectly represented.

Statistically, the shortcoming is corrected by the ANN, but physically, the story is not yet complete. From the important dipole predictor we understand that the implied process is an interaction between Pacific variability and the atmosphere. That finding agrees with earlier research that NWP models have difficulty representing teleconnections from ENSO-like Pacific variability (O'Reilly et al., 2018). But since such teleconnections are more potent and mostly studied in the winter season, the discovery is still somewhat surprising (Branstator and Teng, 2017). So to ultimately refute that the connection is a spurious statistical artifact, a coherent physical understanding of the implied process is needed.

6.2.4 Can we physically understand the ML-detected processes that drive a window of predictability in European summer t2m?

Results in Chapter 3 already showed that the stand-alone ML model uses patterns that are generally consistent with physically understood processes. In Chapter 4 the ANN detects a surprising conditional connection between a dipole in tropical West Pacific SST and European t2m. In Chapter 5 I show that it can be physically understood.

First I showed that the SST dipole is representative of large-scale, partly ENSO-like, Pacific SST patterns, with pronounced equatorial and meridional gradients in deep convection. It represents La-Niña-like phases with enhanced convection over the Maritime continent and western subtropical Pacific, and reduced convection over the central tropical Pacific, and vice-versa for the El-Niño-like phase. In the two weeks following these phases a quasi-stationary atmospheric wave materializes, which lingers in place into week 3 to 6, and is associated with t2m anomalies in Western-

Europe, and anomalies of opposite sign in Eastern Europe and Russia. Both the patterns of convection, and the locations of the wave's low and high pressure systems were in part diagnosed in earlier studies (Ting, 1994; Ding et al., 2011; Behera et al., 2013; Wulff et al., 2017; O'Reilly et al., 2018; Ma and Franzke, 2021; Vijverberg and Coumou, 2022).

Such correspondence embeds the ML-discovered process in existing physical understanding. Still, the longitudinal extent of the atmospheric wave, from the Pacific, via north America and the north Atlantic to Europe, is surprising for the summer season (Branstator and Teng, 2017). Results showed that the arrival of the wave in Europe can be conditional on the presence of a strong and narrow jet stream over the extra-tropical Atlantic, which acts as a waveguide. As these conditions grow out of an interaction with the Atlantic sea surface, the result thus showcases how much a window of predictability can be conditional on interacting variability in the Pacific and Atlantic.

Further evidence of inter-relation was found in the fact that the window of predictability has emerged since the 1980's (also detected by Wu and Lin, 2012; Lim et al., 2019; O'Reilly et al., 2019a; Sun et al., 2022). This coincides with a strong warming of West Pacific SSTs in response to anthropogenic climate change (Funk and Hoell, 2015), which has aggravated the large-scale gradients, and has potentially enabled the sub-seasonal teleconnection. This idea of a sub-seasonal teleconnection transferring long-term changes in the Pacific to European summer weather, is conceptually quite new. Overall, the investigation of the physical pathway in Chapter 5 thus both confirms existing physical understanding and highlights new aspects.

6.3 Short term outlook

The findings above demonstrate a core conclusion of this thesis, namely that a window of predictability can be the consequence of interacting time scales, from long-term anthropogenic climate change, to inter-annual ENSO-like variability, to multi-week quasi-stationary waves. Forecast models need to accurately represent all of this, but do so imperfectly. I demonstrated that representation can be improved through conditional statistical post-processing. But in fact, post-processing is just one of the ways of combining numerical models with Machine Learning: a procedure generally referred to as 'hybrid modeling' (Slater et al., 2023). I discuss two other existing hybrid designs.

By themselves, current numerical EPS systems are often sub-optimally used. The problem is that they are not reliable and do not consistently indicate occasions of higher forecast skill with lower spread (Albers and Newman, 2019). Especially at sub-seasonal lead times, decision makers are therefore left guessing whether forecasts with low spread indicate a rare window of predictability, meaning they should pay extra attention, or not. In contrast, stand-alone statistical forecasts based on initial conditions, such as those made in Chapter 3, can often better link driving processes to occasions of higher certainty (Albers and Newman, 2019; Robertson et al., 2020; Mayer and Barnes, 2021). A hybrid solution is therefore to let the statistical model determine *when* users should attend to the numerical forecasts. Albers and Newman (2019) use a linear inverse model to determine those occasions, but I suggest that non-linear models like the one of Chapter 3 could prove useful for sub-seasonal setups. After all, the processes that drive a window of predictability are conditional and interact.

A second hybrid solution is to let the statistical model determine *which* forecasts the users should attend to. It is known that not all ensemble members of an EPS system are equally affected by errors, given that the occurrence of conditional errors depends on the trajectory that is simulated. This means that a statistical model can forecast the most likely physical state, such that the least erroneous ensemble members can be selected to provide a more skillful sub-ensemble. The forecast state by which to make the selection is often that of a dominant phenomenon like the North Atlantic Oscillation (Dobrynin et al., 2018; Neddermann et al., 2019). Instead of only selecting or weighing members from a single EPS, the procedure can even be scaled to multiple forecast systems, resulting in so-called opportunistic mixing models. Such a model, mixing post-processed EPS forecasts and climatological forecasts, outperformed all other models in the S2S challenge hosted in 2021 (Vitart et al., 2022).

Given the importance of comparing the performance of models to further their development, this thesis reiterates the need do so carefully and with good benchmarks (Coelho et al., 2019; Manrique-Suñén et al., 2020; Mouatadid et al., 2021). Chapter 3 showed that sub-seasonal predictability is only properly assessed when separated from trend-related predictability, by comparing against a climate-change benchmark. In Chapter 2 and 4 I also implicitly accounted for simple biases in EPS mean and spread, such that the EPS formed a better and 'climatologically reliable' benchmark (Van Schaeybroeck and Vannitsem, 2015). This was done by determining the t2m exceedence thresholds separately for the

EPS and observations. I suggest that others take up these two practices as well.

6.4 Longer term outlook

The presented hybrid combination of ML and numerical forecasts in Chapter 4 not only improved skill, but was also useful to learn about a numerically misrepresented tropical to extra-tropical teleconnection. Hence, hybrid modeling is not just a way to achieve high accuracy, it is a way to investigate existing data more deeply than with only the numerical models (Slater et al., 2023). This goal of knowledge creation should be central to future ML applications, which is why I want to raise some considerations about their design. Especially the capacity of the ML model, which ranges from explicit to implicit (section 6.2.2), can make a learned representation more or less informative to humans, when investigated with XAI.

First, the representation should not be highly implicit. Such representations are found in NWP models, as they evolve one multi-variate state of prognostic variables (such as moisture, pressure, temperature, velocity) to the next, at very short time steps. Aggregate phenomena such as propagating Rossby waves emerge when the simulation is 'rolled out' over many steps. Roll-out deep learning models also exist and perform well for lead times of at least a few days (Weyn et al., 2021; Keisler, 2022; Lam et al., 2022). Both the roll-out NWP and ML models fit well within the seamless forecasting paradigm (World Meteorological Organization, 2015). But as demonstrated in this thesis, shortcomings in important interactions can limit the NWP model's performance at sub-seasonal lead times. Also the deep-learning models accumulate errors over many roll out steps, leading to limited sub-seasonal skill (Weyn et al., 2021). And even if a deep learning model would achieve good sub-seasonal performance, the numerous roll out steps would make it difficult to trace back a correctly predicted sub-seasonal window of predictability to the processes that drove it from initialization onward. Current XAI methods are not up to this task, meaning that more effort has go into their development and benchmarking, should we want to embark on this path (Ras et al., 2022; Mamalakis et al., 2022b,a).

A more promising path regarding knowledge creation, could lie in more explicit ML models. These should not be overly explicit, because then the problem with expert-crafted representations and precluded dis-

covery arises again (section 6.2.2). An ML model of medium capacity is suited because it cannot let sub-seasonal phenomena emerge as aggregate behaviour only, but has to represent them more directly. The medium capacity model presented in Chapter 3 involved a dimension reduction first, which then allowed a Random Forest to freely learn inter-variable and inter-timescale interactions. As shown in Chapter 3 for the 2015 heatwave, the drivers of the ML forecast can then be presented in terms of anomalies that resemble patterns about which theories exist, such as the influence of ice in the Kara sea (Hall et al., 2017) (Fig. 3.10). The XAI explanation is then close to a meteorologist's mental image of the physical system, which was found to be a requirement for trustworthiness (Harrison, 2022).

Certainly, there will be future ways to create even more informative medium-capacity ML models, for instance through alternative definition of the predictand. For the bulk of this thesis the predictand was defined as whether monthly average t2m in west and central Europe exceeded a threshold. Such static predictands with fixed spatio-temporal scales are currently very common (e.g. Weirich Benet et al., 2023; Gómez-Orellana et al., 2023). In fact, Chapter 2 was devoted to detecting the right static scale: the level of aggregation at which sub-seasonal variability would be captured.

Consider however that a sequence of daily t2m anomalies is never exactly the same, and that the composition of a monthly anomaly can vary (this varying 'sub-structure' is found for Western European summers in particular, Röthlisberger et al., 2020). This is because a quasi-stationary wave's exact location, amplitude, phase speed, interaction with the surface, and spatial imprint, will differ from one window of predictability to the next (e.g. Barriopedro et al., 2011; Böhnisch et al., 2023). All of that differentiation gets lost when considering a static predictand and its so-called 'climatic predictability' (Toth and Buizza, 2019). The ML model will only learn those aspects of the driving processes that are shared, limiting what humans can learn from the representation.

An alternative would be to consider predictability in its 'traceable' sense (Toth and Buizza, 2019). In this case predictands represent objects themselves, which for sub-seasonal forecasts could be a collection of air-masses embedded in a quasi-stationary wave, interacting with ocean and land surface. Potentially, this makes the learned representation closer to the actual driving processes, and therefore more revealing to humans. However, as objects capture more spatio-temporal particularities, the number of observed duplicates can quickly decrease. This is detrimen-

tal for ML, as it needs many identically distributed samples to train on.

In a profound sense, this requirement of identical distribution is already being violated in most ML applications. Anthropogenic global warming has namely been altering the data distributions for numerous decades. In Chapter 3 and 4 this non-stationarity was tackled by providing the trend as a first-order estimate to the ML model. But still, a model trained on historical data is not guaranteed to remain valid in the future, as it can receive out-of-sample 'unseen' input values, even when the responsible processes themselves do not fundamentally change (Gessner et al., 2021). An obvious solution could be to augment the training set with numerical simulations of the future climate (it is already common practice to use simulations of current climate to obtain large data sets, Irrgang et al., 2021; Miloshevich et al., 2023; Pan et al., 2022; Trenary and DelSole, 2023). Being physics-based, the numerical models have greater extrapolative strength than ML models, and can sample the unobserved climates that arise from internal variability and anthropogenic forcing (Balaji, 2021; Frame et al., 2022). However, the difference in strength is not always clear (Razavi, 2021; Frame et al., 2022), as also NWP models contain many empirical elements subject to non-stationary: think of subgrid parametrizations and the observation operator in data-assimilation. Obviously, more research is needed into learning climate-invariant models (Beucler et al., 2022).

To conclude, this thesis has shown two ways in which hybrid modeling blurs the distinction between physics-based and ML models. One way is the direct improvement of forecasts by statistical post-processing with ML models. The second, indirect way is the use of ML in combination with XAI to gain knowledge on important sub-seasonal processes and errors in their numerical representation. This eventually enables structural NWP improvements to be made. All signs point towards a further blurring of the distinction between physics-based and ML models (Düben et al., 2021; Karpatne et al., 2017; Camps-Valls et al., 2020; Irrgang et al., 2021). Purists may see this as a problem. But as any model is a fiction, the focus should be on making the most useful ones.

Appendix A

Supplementary material for Chapter 4

A.1 Regime classification in z300

We distinguish four regimes in the Euro-Atlantic region. Four is commonly seen as the minimum amount needed in this region (Michelangeli et al., 1995; Zampieri et al., 2017; Falkena et al., 2020). We use the same domain as Cassou et al. (2005), namely from 20 to 80 degrees north and from -90 to 30 degrees east, and derive regimes from daily ERA5 z300 anomalies spanning from May to August. First, the ERA5 fields are regridded to the resolution of 1.5x1.5°, at which we extracted the ECMWF forecasts. Then we linearly detrend all gridcells at once to account for thermodynamic expansion of the air column due to global warming, disregarding grid-cell specific trends like the North Atlantic warming hole (Chemke et al., 2020). We follow the approach of Ferranti et al. (2015) and Michelangeli et al. (1995) in reducing the phase space to ten leading Empirical Orthogonal Functions (EOFs), and computing 4 k-means clusters.

After computation we assign the daily reanalysis fields to the closest centroid, measured in terms of Euclidean distance in EOF-space. If a daily distance to all k-means centroids is found to be larger than the median distance of all samples to that centroid, the pattern is labeled 'unclassified'. This unclassified regime populates about 12 percent of reanalysis states. Forecast z300 anomalies are detrended similarly and assigned to the ERA5 centroids. Each member at each lead time gets classified as one of the four, or the unclassified regime. At 1-day lead

time about 10 percent of forecasts are unclassified, against 18 percent at +40 days. The NWP model thus shows a drift towards unclassified flow patterns.

As daily regime forecasts can be subject to timing errors (e.g. blocking develops a day too late), the relation to monthly t2m (section 4.3.2) is stronger if we quantify a period's tendency towards a certain regime. To that end we count the relative frequency of each regime over the 11 members and a succession of lead times (similar to Lavaysse et al., 2018; Cortesi et al., 2021). We extract frequencies for 21-day and 31-day periods.

In the end these predictors were not selected as part of the top predictors (Fig. 4.7), even though NWP-predicted regimes were found useful in other S2S studies (Richardson et al., 2020; Mastrantonas et al., 2022). It might relate to our simplistic regime classification (for improvements see Grams et al., 2017; Falkena et al., 2020; Dorrington and Strommen, 2020). Alternatively, we know that summer circulation is continuous in phase space and less 'regime-like' than in winter (Rousi et al., 2021). This can make our division into four classes arbitrary, harder to predict, and of little use for S2S post-processing. The limited use of ECMWF regime states is however consistent with our suggestion that misrepresented sources of predictability affect a multitude of model-predicted states at valid time, among which is atmospheric circulation (Fig. 4.1B). Results indeed indicate the missed development of an atmospheric wave (Fig. 4.10c,aa). With NWP models lacking such connections, it is not surprising that skill in regime predictions often does not extend beyond the third forecast week (Cortesi et al., 2021; Büeler et al., 2021).

A.2 Additional XAI plots

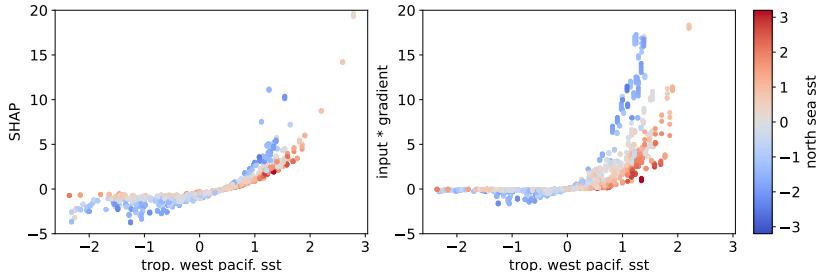


Figure A.1: Effect of the initial west Pacific SST pattern on the ANN-based correction, in terms of SHAP value (left) and input*gradient (right). X-axis: standardized value of the West Pacific predictor. Y-axis: contribution to the multiplicative correction factor. Color scale: standardized value of predicted North Sea SST. Curves: High West Pacific values (positive correspondence between initial anomalies and correlation pattern) result in positive corrections, negative values in negative corrections, but impact is modulated by North Sea SST (West Pacific contributions are brought closer to zero when predicted North Sea SST is high).

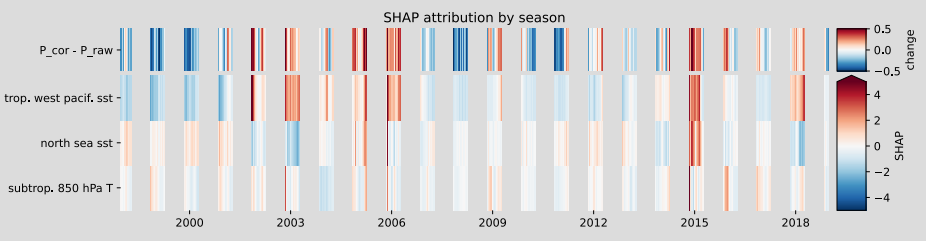


Figure A.2: As Fig. 4.9, but with samples ordered by time instead of the leaf order from hiearchical clustering, and with SHAP values only.

A.3 Additional Logistic Regression benchmarks

Table A.1: Brier Score comparison on the test set (lower is better) of ANN-based post-processed forecasts (p_{cor}) to two simple logistic regression models using the same forward selected predictors as the ANN, with either p_{raw} or $\log(p_{raw})$ as extra predictor. ANN characteristics, like the amount of hidden layers, and the amount of predictors are varied over rows. Benchmarks p_{trend} and p_{raw} have a BS of 0.214 and 0.224 respectively

	p_{cor}	Logistic + p_{raw}	Logistic + $\log(p_{raw})$
3 predictors, 1 hidden layer	0.168	0.188	0.190
3 predictors, 0 hidden layer	0.171	0.188	0.190
1 predictor, 1 hidden layer	0.188	0.195	0.196
1 predictor, 0 hidden layer	0.188	0.195	0.196

Bibliography

Albers, J. R. and Newman, M. (2019). A priori identification of skillful extratropical subseasonal forecasts. *Geophysical Research Letters*, 46(21):12527–12536.

Allen, S., Ferro, C. A., and Kwasniok, F. (2019). Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(725):3535–3552.

Ardilouze, C., Batté, L., Bunzel, F., Decremer, D., Déqué, M., Doblas-Reyes, F. J., Douville, H., Fereday, D., Guemas, V., MacLachlan, C., et al. (2017a). Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate dynamics*, 49(11-12):3959–3974.

Ardilouze, C., Batté, L., and Déqué, M. (2017b). Subseasonal-to-seasonal (s2s) forecasts with cnrm-cm: A case study on the july 2015 west-european heat wave. *Advances in Science and Research*, 14:115–121.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Bakker, K., Whan, K., Knap, W., and Schmeits, M. (2019). Comparison of statistical post-processing methods for probabilistic nwp forecasts of solar radiation. *Solar Energy*, 191:138–150.

Balaji, V. (2021). Climbing down charney's ladder: machine learning and the post-dennard era of computational climate science. *Philosophical Transactions of the Royal Society A*, 379(2194):20200085.

Baldwin, M., Gray, L., Dunkerton, T., Hamilton, K., Haynes, P., Randel, W. J., Holton, J. R., Alexander, M., Hirota, I., Horinouchi, T., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, 39(2):179–229.

Baldwin, M. P. and Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, 294(5542):581–584.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., and García-Herrera, R. (2011). The hot summer of 2010: redrawing the temperature record map of europe. *Science*, 332(6026):220–224.

Barton, Y., Giannakaki, P., Von Waldow, H., Chevalier, C., Pfahl, S., and Martius, O. (2016). Clustering of regional-scale extreme precipitation events in southern switzerland. *Monthly Weather Review*, 144(1):347–369.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47.

Becker, E. J., Van Den Dool, H., and Peña, M. (2013). Short-term climate extremes:

Prediction skill and predictability. *Journal of Climate*, 26(2):512–531.

Behera, S., Ratnam, J. V., Masumoto, Y., and Yamagata, T. (2013). Origin of extreme summers in europe: the indo-pacific connection. *Climate dynamics*, 41(3):663–676.

Bello, G. A., Angus, M., Pedemane, N., Harlalka, J. K., Semazzi, F. H., Kumar, V., and Samatova, N. F. (2015). Response-guided community detection: Application to climate index discovery. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 736–751. Springer.

Ben Bouallègue, Z. and Richardson, D. S. (2022). On the roc area of ensemble forecasts for rare events. *Weather and Forecasting*.

Benedict, J. J., Lee, S., and Feldstein, S. B. (2004). Synoptic view of the north atlantic oscillation. *Journal of the atmospheric sciences*, 61(2):121–144.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Beobide-Arsuaga, G., Düsterhus, A., Müller, W. A., Barnes, E. A., and Baehr, J. (2023). Spring regional sea surface temperatures as a precursor of european summer heatwaves. *Geophysical Research Letters*, 50(2):e2022GL100727. e2022GL100727 2022GL100727.

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., and Gentine, P. (2022). Climate-invariant machine learning. *Proceedings of the National Academy of Sciences*.

Beverley, J. D., Woolnough, S. J., Baker, L. H., Johnson, S. J., and Weisheimer, A. (2019). The northern hemisphere circumglobal teleconnection in a seasonal forecast model and its relationship to european summer forecast skill. *Climate Dynamics*, 52:3759–3771.

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial pacific. *Monthly weather review*, 97(3):163–172.

Black, E. and Sutton, R. (2007). The influence of oceanic conditions on the hot european summer of 2003. *Climate dynamics*, 28(1):53–66.

Bladé, I., Liebmann, B., Fortuny, D., and van Oldenborgh, G. J. (2012). Observed and simulated impacts of the summer nao in europe: implications for projected drying in the mediterranean region. *Climate dynamics*, 39(3-4):709–727.

Boe, J., Terray, L., Moine, M.-P., Valcke, S., Bellucci, A., Drijfhout, S., Haarsma, R., Lohmann, K., Putrasahan, D. A., Roberts, C., et al. (2020). Past long-term summer warming over western europe in new generation climate models: role of large-scale atmospheric circulation. *Environmental Research Letters*, 15(8):084038.

Böhnisch, A. F., Felsche, E., and Ludwig, R. (2023). European heatwave tracks: Using causal discovery to detect recurring pathways in a single-regional climate model large ensemble. *Environmental Research Letters*, 18(1):014038.

Branstator, G. (2014). Long-lived response of the midlatitude circulation and storm tracks to pulses of tropical heating. *Journal of Climate*, 27(23):8809–8826.

Branstator, G. and Teng, H. (2017). Tropospheric waveguide teleconnections and their seasonality. *Journal of the Atmospheric Sciences*, 74(5):1513–1532.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly*

weather review, 78(1):1–3.

Brunner, L., Hegerl, G. C., and Steiner, A. K. (2017). Connecting atmospheric blocking to european temperature extremes in spring. *Journal of Climate*, 30(2):585–594.

Buizza, C., Casas, C. Q., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur, T., Ruiz, L. B., et al. (2022). Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, 58:101525.

Buizza, R., Balsamo, G., and Haide, T. (2018). Ifs upgrade brings more seamless coupled forecasts. *ECMWF newsletter*, 156:18–22.

Buizza, R. and Leutbecher, M. (2015). The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, 141(693):3366–3382.

Bunzel, F., Müller, W. A., Dobrynin, M., Fröhlich, K., Hagemann, S., Pohlmann, H., Stacke, T., and Baehr, J. (2018). Improved seasonal prediction of european summer temperatures with new five-layer soil-hydrology scheme. *Geophysical Research Letters*, 45(1):346–353.

Bürger, G. (2019). A seamless filter for daily to seasonal forecasts, with applications to iran and brazil. *Quarterly Journal of the Royal Meteorological Society*.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J. F., and Grams, C. M. (2021). Year-round sub-seasonal forecast skill for atlantic-european weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 147(741):4283–4309.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, 41(5):1803–1808.

Camps-Valls, G., Svendsen, D. H., Cortés-Andrés, J., Moreno-Martínez, Á., Pérez-Suay, A., Adsuara, J. E., Martín, I., Piles, M., Muñoz-Marí, J., and Martino, L. (2020). Living in the physics and machine learning interplay for earth observation. *arXiv*.

Carvalho-Oliveira, J., Borchert, L. F., Zorita, E., and Baehr, J. (2022). Self-organizing maps identify windows of opportunity for seasonal european summer predictions. *Frontiers in Climate*, 4.

Casanueva, A., Burgstall, A., Kotlarski, S., Messeri, A., Morabito, M., Flouris, A. D., Nybo, L., Spirig, C., and Schwierz, C. (2019). Overview of existing heat-health warning systems in europe. *International journal of environmental research and public health*, 16(15):2657.

Casanueva, A., Rodríguez-Puebla, C., Frías, M., and González-Reviriego, N. (2014). Variability of extreme precipitation over europe and its relationships with teleconnection patterns. *Hydrol. Earth Syst. Sci*, 18:709–725.

Cassou, C. (2008). Intraseasonal interaction between the madden–julian oscillation and the north atlantic oscillation. *Nature*, 455(7212):523.

Cassou, C., Terray, L., and Phillips, A. S. (2005). Tropical atlantic influence on european heat waves. *Journal of climate*, 18(15):2805–2811.

Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.-P., Lerch, S., and Hayatbini, N. (2021). Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*.

Chemke, R., Zanna, L., and Polvani, L. M. (2020). Identifying a human signal in the north atlantic warming hole. *Nature communications*, 11(1):1–7.

Christidis, N., Jones, G. S., and Stott, P. A. (2015). Dramatically increasing chance of

extremely hot summers since the 2003 european heatwave. *Nature Climate Change*, 5(1):46–50.

Clare, M. C. A., Sonnewald, M., Lguensat, R., Deshayes, J., and Balaji, V. (2022). Explainable artificial intelligence for bayesian neural networks: Towards trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, 14(11):e2022MS003162.

Coelho, C. A., Brown, B., Wilson, L., Mittermaier, M., and Casati, B. (2019). Chapter 16 - forecast verification for s2s timescales. In Robertson, A. W. and Vitart, F., editors, *Sub-Seasonal to Seasonal Prediction*, pages 337–361. Elsevier.

Coelho, C. A., Firpo, M. A., and de Andrade, F. M. (2018). A verification framework for south american sub-seasonal precipitation predictions. *Meteorologische Zeitschrift*, 27(6):503–520.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., and Tziperman, E. (2018). S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdisciplinary Reviews: Climate Change*, 10(2):e00567.

Cohen, J., Furtado, J. C., Jones, J., Barlow, M., Whittleston, D., and Entekhabi, D. (2014). Linking siberian snow cover to precursors of stratospheric variability. *Journal of Climate*, 27(14):5422–5432.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., and Jones, P. D. (2018). An ensemble version of the e-obs temperature and precipitation datasets. *Journal of Geophysical Research: Atmospheres*.

Cortesi, N., Torralba, V., Lledó, L., Manrique-Suñén, A., Gonzalez-Reviriego, N., Soret, A., and Doblas-Reyes, F. J. (2021). Yearly evolution of euro-atlantic weather regimes and of their sub-seasonal predictability. *Climate Dynamics*, 56:3933–3964.

Coughlan de Perez, E., Van Aalst, M., Bischiniotis, K., Mason, S., Nissan, H., Pappenberger, F., Stephens, E., Zsoter, E., and Van Den Hurk, B. (2018). Global predictability of temperature extremes. *Environmental Research Letters*, 13(5):054017.

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

Coumou, D., Di Capua, G., Vavrus, S., Wang, L., and Wang, S. (2018). The influence of arctic amplification on mid-latitude summer circulation. *Nature Communications*, 9(2959):1–12.

Czaja, A. and Frankignoul, C. (2002). Observed impact of atlantic sst anomalies on the north atlantic oscillation. *Journal of Climate*, 15(6):606–623.

Dai, Y. and Hemri, S. (2021). Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149(12):3923–3937.

Della-Marta, P. M., Luterbacher, J., von Weissenfluh, H., Xoplaki, E., Brunet, M., and Wanner, H. (2007). Summer heat waves over western europe 1880–2003, their relationship to large-scale forcings and predictability. *Climate Dynamics*, 29(2-3):251–275.

DelSole, T. and Shukla, J. (2009). Artificial skill due to predictor screening. *Journal of Climate*, 22(2):331–345.

DelSole, T. and Tippett, M. K. (2009). Average predictability time. part ii: Seamless diagnoses of predictability on multiple time scales. *Journal of the atmospheric sciences*, 66(5):1188–1204.

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

Di Capua, G., Runge, J., Donner, R. V., van den Hurk, B., Turner, A. G., Vellore, R., Krishnan, R., and Coumou, D. (2020b). Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales. *Weather and Climate Dynamics*, 1(2):519–539.

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

Dobrynin, M., Domeisen, D. I., Müller, W. A., Bell, L., Brune, S., Bunzel, F., Düsterhus, A., Fröhlich, K., Pohlmann, H., and Baehr, J. (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophysical Research Letters*, 45(8):3605–3614.

Dole, R., Hoerling, M., Kumar, A., Eischeid, J., Perlwitz, J., Quan, X.-W., Kiladis, G., Webb, R., Murray, D., Chen, M., et al. (2014). The making of an extreme event: putting the pieces together. *Bulletin of the American Meteorological Society*, 95(3):427–440.

Dong, B., Sutton, R. T., Shaffrey, L., and Harvey, B. (2022). Recent decadal weakening of the summer eurasian westerly jet attributable to anthropogenic aerosol emissions. *Nature Communications*, 13(1):1148.

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

Dorrington, J. and Strommen, K. (2020). Jet speed variability obscures euro-atlantic regime structure. *Geophysical Research Letters*, 47(15):e2020GL087907.

Düben, P., Modigliani, U., Geer, A., Siemen, S., Pappenberger, F., Bauer, P., Brown, A., Palkovic, M., Raoult, B., Wedi, N., and Baousis, V. (2021). Machine learning at ecmwf: A roadmap for the next 10 years. *ECMWF Technical Memoranda*, 878.

Duchez, A., Frajka-Williams, E., Josey, S. A., Evans, D. G., Grist, J. P., Marsh, R., McCarthy, G. D., Sinha, B., Berry, D. I., and Hirschi, J. J. (2016). Drivers of exceptionally cold north atlantic ocean temperatures and their link to the 2015 european heat wave. *Environmental Research Letters*, 11(7):074004.

Dutra, E., Johannsen, F., and Magnusson, L. (2021). Late spring and summer subseasonal forecasts in the northern hemisphere midlatitudes: biases and skill in the ecmwf model. *Monthly Weather Review*.

Economou, T., Stephenson, D. B., Pinto, J., Shaffrey, L., and Zappa, G. (2015). Serial clustering of extratropical cyclones in a multi-model ensemble of historical and future simulations. *Quarterly Journal of the Royal Meteorological Society*, 141(693):3076–3087.

Falkena, S. K., de Wiljes, J., Weisheimer, A., and Shepherd, T. G. (2020). Revisiting the identification of wintertime atmospheric circulation regimes in the euro-atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 146(731):2801–2814.

Fan, Y., Krasnopolsky, V., van den Dool, H., Wu, C.-Y., and Gottschalck, J. (2021). Using artificial neural networks to improve cfs week 3-4 precipitation and 2-meter air temperature forecasts. *Weather and Forecasting*.

Feng, X., DelSole, T., and Houser, P. (2011). Bootstrap estimated seasonal potential predictability of global temperature and precipitation. *Geophysical Research Letters*, 38(7).

Ferranti, L., Corti, S., and Janousek, M. (2015). Flow-dependent verification of the ecmwf ensemble over the euro-atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141(688):916–924.

Ferranti, L., Magnusson, L., Vitart, F., and Richardson, D. S. (2018). How far in advance can we predict changes in large-scale flow leading to severe cold conditions over europe? *Quarterly Journal of the Royal Meteorological Society*, 144:178–1802.

Ferrone, A., Mastrangelo, D., and Malguzzi, P. (2017). Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Advances in Science and Research*, 14:123–129.

Feudale, L. and Shukla, J. (2011). Influence of sea surface temperature on the european heat wave of 2003 summer. part i: an observational study. *Climate dynamics*, 36(9-10):1691–1703.

Fischer, E. M., Beyerle, U., and Knutti, R. (2013). Robust spatially aggregated projections of climate extremes. *Nature Climate Change*, 3:1033–1038.

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., Lüthi, D., and Schär, C. (2007). Soil moisture–atmosphere interactions during the 2003 european summer heat wave. *Journal of Climate*, 20(20):5081–5099.

Folland, C. K., Knight, J., Linderholm, H. W., Fereday, D., Ineson, S., and Hurrell, J. W. (2009). The summer north atlantic oscillation: past, present, and future. *Journal of Climate*, 22(5):1082–1103.

Ford, T. W., Dirmeyer, P. A., and Benson, D. O. (2018). Evaluation of heat wave forecasts seamlessly across subseasonal timescales. *npj Climate and Atmospheric Science*, 1(20).

Frajka-Williams, E., Beaulieu, C., and Duchez, A. (2017). Emerging negative atlantic multidecadal oscillation index in spite of warm subtropics. *Scientific reports*, 7(1):11224.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S. (2022). Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392.

Frankignoul, C. and Hasselmann, K. (1977). Stochastic climate models, part ii application to sea-surface temperature anomalies and thermocline variability. *Tellus*, 29(4):289–305.

Fuentes-Franco, R. and Koenigk, T. (2020). Identifying remote sources of interannual variability for summer precipitation over nordic european countries tied to global teleconnection wave patterns. *Tellus A: Dynamic Meteorology and Oceanography*, 72(1):1–15.

Fuentes-Franco, R., Koenigk, T., Docquier, D., Graef, F., and Wyser, K. (2022). Exploring the influence of the north pacific rossby wave sources on the variability of summer atmospheric circulation and precipitation over the northern hemisphere. *Climate Dynamics*, pages 1–15.

Funk, C., Harrison, L., Shukla, S., Pomposi, C., Galu, G., Korecha, D., Husak, G.,

Magadzire, T., Davenport, F., Hillbruner, C., et al. (2018). Examining the role of unusually warm indo-pacific sea-surface temperatures in recent african droughts. *Quarterly Journal of the Royal Meteorological Society*, 144:360–383.

Funk, C. C. and Hoell, A. (2015). The leading mode of observed and cmip5 enso-residual sea surface temperatures and associated changes in indo-pacific climate. *Journal of Climate*, 28(11):4309–4329.

Garcia-Serrano, J. and Frankignoul, C. (2014). Retraction note: High predictability of the winter euro–atlantic climate from cryospheric variability. *Nature Geoscience*, 7(6):E2–E2.

Gastineau, G. and Frankignoul, C. (2015). Influence of the north atlantic sst variability on the atmospheric circulation during the twentieth century. *Journal of Climate*, 28(4):1396–1416.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338.

Geer, A. J. (2021). Learning earth system models from observations: machine learning or data assimilation? *Philosophical Transactions of the Royal Society A*, 379(2194):20200089.

Gessner, C., Fischer, E. M., Beyerle, U., and Knutti, R. (2021). Very rare heat extremes: quantifying and understanding using ensemble reinitialization. *Journal of Climate*, 34(16):6619–6634.

Ghil, M. and Robertson, A. W. (2002). "waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proceedings of the National Academy of Sciences*, 99(suppl 1):2493–2500.

Gibson, P. B., Chapman, W. E., Altinok, A., Delle Monache, L., DeFlorio, M. J., and Waliser, D. E. (2021). Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, 2(159):1–13.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, 11(8):1203–1211.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118.

Goutham, N., Plougonven, R., Omrani, H., Parey, S., Tankov, P., Tantet, A., Hitchcock, P., and Drobinski, P. (2022). How skillful are the european sub-seasonal forecasts of wind speed and surface temperature? *Monthly Weather Review*.

Graham, R. M., Browell, J., Bertram, D., and White, C. J. (2022). The application of sub-seasonal to seasonal (s2s) predictions for hydropower forecasting. *Meteorological Applications*, 29(1):e2047.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H. (2017). Balancing europe's wind-power output through spatial deployment informed by weather

regimes. *Nature climate change*, 7(8):557.

Gray, S. L., Dunning, C., Methven, J., Masato, G., and Chagnon, J. M. (2014). Systematic model forecast error in rossby wave structure. *Geophysical Research Letters*, 41(8):2979–2987.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., and Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, 379(2194):20200092.

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., Gershunov, A., Gutowski, W. J., Gyakum, J. R., Katz, R. W., et al. (2016). North american extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 46(3-4):1151–1184.

Guigma, K. H., MacLeod, D., Todd, M., and Wang, Y. (2021). Prediction skill of sahelian heatwaves out to subseasonal lead times and importance of atmospheric tropical modes of variability. *Climate Dynamics*, 57:537—-556.

Guimaraes Nobre, G., Hunink, J. E., Baruth, B., Aerts, J. C. J. H., and Ward, P. (2019). Translating large-scale climate variability into crop production forecast in europe. *Scientific Reports*, 9(1):2045–2322.

Gómez-Orellana, A. M., Guijo-Rubio, D., Pérez-Aracil, J., Gutiérrez, P. A., Salcedo-Sanz, S., and Hervás-Martínez, C. (2023). One month in advance prediction of air temperature from reanalysis data with explainable artificial intelligence techniques. *Atmospheric Research*, 284:106608.

Haarsma, R. J., Selten, F., Hurk, B. V., Hazeleger, W., and Wang, X. (2009). Drier mediterranean soils due to greenhouse warming bring easterly winds over summertime central europe. *Geophysical research letters*, 36(4).

Haarsma, R. J., Selten, F. M., and Drijfhout, S. S. (2015). Decelerating atlantic meridional overturning circulation main cause of future west european summer atmospheric circulation changes. *Environmental Research Letters*, 10(9):094007.

Hall, R. J., Jones, J. M., Hanna, E., Scaife, A. A., and Erdélyi, R. (2017). Drivers and potential predictability of summer time north atlantic polar front jet variability. *Climate Dynamics*, 48(11-12):3869–3887.

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

Hannachi, A., Straus, D. M., Franzke, C. L., Corti, S., and Woollings, T. (2017). Low-frequency nonlinearity and regime behavior in the northern hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55(1):199–234.

Harrison, D. (2022). *Machine Learning Co-Production in Operational Meteorology*. phdthesis, University of Oklahoma.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer series in statistics.

Haupt, S. E., Chapman, W., Adams, S. V., Kirkwood, C., Hosking, J. S., Robinson, N. H., Lerch, S., and Subramanian, A. C. (2021). Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the oxford 2019 workshop. *Philosophical Transactions of the Royal Society A*, 379(2194):20200091.

He, B., Liu, P., Zhu, Y., and Hu, W. (2019). Prediction and predictability of northern

hemisphere persistent maxima of 500-hpa geopotential height eddies in the gefs. *Climate Dynamics*, 52:3773–3789.

He, S., Li, X., DelSole, T., Ravikumar, P., and Banerjee, A. (2021). Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1).

Heinrich, C., Hellton, K. H., Lenkoski, A., and Thorarinsdottir, T. L. (2021). Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *Journal of the American Statistical Association*, 116(535):1048–1059.

Henderson, G. R., Peings, Y., Furtado, J. C., and Kushner, P. J. (2018). Snow–atmosphere coupling in the northern hemisphere. *Nature Climate Change*, 8(11):954–963.

Henderson, S. A., Maloney, E. D., and Son, S.-W. (2017). Madden–julian oscillation pacific teleconnections: The impact of the basic state and mjo representation in general circulation models. *Journal of Climate*, 30(12):4567–4587.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hertel, L., Collado, J., Sadowski, P., Ott, J., and Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, 12:100591.

Hewson, T. D. and Pillosu, F. M. (2021). A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, 2(1):1–10.

Hill, A. J., Herman, G. R., and Schumacher, R. S. (2020). Forecasting severe weather with random forests. *Monthly Weather Review*, 148(5):2135–2161.

Hooker, G. and Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv*.

Hopson, T. (2014). Assessing the ensemble spread–error relationship. *Monthly Weather Review*, 142(3):1125–1142.

Hoskins, B. (2013). The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672):573–584.

Hoskins, B. J. and Ambrizzi, T. (1993). Rossby wave propagation on a realistic longitudinally varying flow. *Journal of Atmospheric Sciences*, 50(12):1661–1671.

Hoskins, B. J. and Karoly, D. J. (1981). The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of the atmospheric sciences*, 38(6):1179–1196.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M. (2017). Extended reconstructed sea surface temperature, version 5 (ersstv5): upgrades, validations, and intercomparisons. *Journal of Climate*, 30(20):8179–8205.

Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., and Mackey, L. (2019). Improving subseasonal forecasting in the western us with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2325–2335, Anchorage, AK, USA. Association for Computing

Machinery, Association for Computing Machinery.

Häkkinen, S., Rhines, P. B., and Worthen, D. L. (2011). Atmospheric blocking and atlantic multidecadal ocean variability. *Science*, 334(6056):655–659.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J. (2021). Towards neural earth system modelling by integrating artificial intelligence in earth system science. *Nature Machine Intelligence*, 3(8):667–674.

Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., and Bouchet, F. (2022). Deep learning-based extreme heatwave forecast. *Frontiers in Climate*, 4.

Jézéquel, A., Yiou, P., and Radanovics, S. (2018). Role of circulation in european heatwaves using flow analogues. *Climate Dynamics*, 50:1145–1159.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., et al. (2019). Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117.

Jong, B.-T., Ting, M., Seager, R., and Anderson, W. B. (2020). Enso teleconnections and impacts on us summertime temperature during a multiyear la niña life cycle. *Journal of Climate*, 33(14):6009–6024.

Jung, T. and Leutbecher, M. (2008). Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 134(633):973–984.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Kämäräinen, M., Uotila, P., Karpechko, A. Y., Hyvärinen, O., Lehtonen, I., and Räisänen, J. (2019). Statistical learning methods as a basis for skillful seasonal temperature forecasts in europe. *Journal of Climate*, 32(17):5363–5379.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331.

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., and Woollings, T. (2021). Atmospheric blocking and weather extremes over the euro-atlantic sector–a review. *Weather and Climate Dynamics Discussions*, pages 1–43.

Keisler, R. (2022). Forecasting global weather with graph neural networks.

Kharin, V. V. and Zwiers, F. W. (2003). On the roc score of probability forecasts. *Journal of Climate*, 16(24):4145–4150.

Kim, B.-M., Son, S.-W., Min, S.-K., Jeong, J.-H., Kim, S.-J., Zhang, X., Shim, T., and Yoon, J.-H. (2014). Weakening of the stratospheric polar vortex by arctic sea-ice loss. *Nature communications*, 5(4646):1–8.

Kim, D. W. and Lee, S. (2022). The role of latent heating anomalies in exciting the summertime eurasian circulation trend pattern and high surface temperature. *Journal of Climate*, 35(2):801–814.

Kim, H., Vitart, F., and Waliser, D. E. (2018). Prediction of the madden–julian oscillation: A review. *Journal of Climate*, 31(23):9425–9443.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*.

Kirkwood, C., Economou, T., Odbert, H., and Pugeault, N. (2021). A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. *Philosophical Transactions of the Royal Society A*, 379(2194).

Koenigk, T., Caian, M., Nikulin, G., and Schimanke, S. (2016). Regional arctic sea ice variations as predictor for winter climate conditions. *Climate Dynamics*, 46(1-2):317–337.

Kolstad, E. W. and Årthun, M. (2018). Seasonal prediction from arctic sea surface temperatures: Opportunities and pitfalls. *Journal of Climate*.

Kornhuber, K., Petoukhov, V., Petri, S., Rahmstorf, S., and Coumou, D. (2017). Evidence for wave resonance as a key mechanism for generating high-amplitude quasi-stationary waves in boreal summer. *Climate Dynamics*, 49(5-6):1961–1979.

Koster, R. D., Mahanama, S., Yamada, T., Balsamo, G., Berg, A., Boisserie, M., Dirmeyer, P., Doblas-Reyes, F., Drewitt, G., Gordon, C., et al. (2010). Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophysical Research Letters*, 37(2):L02402.

Kretschmer, M., Runge, J., and Coumou, D. (2017). Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophysical Research Letters*, 44(16):8592–8600.

Kueh, M.-T. and Lin, C.-Y. (2020). The 2018 summer heatwaves over northwestern europe and its extended-range prediction. *Scientific Reports*, 10(1):1–18.

Lakshmanan, V., Karstens, C., Krause, J., Elmore, K., Ryzhkov, A., and Berkseth, S. (2015). Which polarimetric variables are important for weather/no-weather discrimination? *Journal of Atmospheric and Oceanic Technology*, 32(6):1209–1223.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Stott, J., Vinyals, O., Mohamed, S., and Battaglia, P. (2022). Graphcast: Learning skillful medium-range global weather forecasting. *arXiv*.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Lavaysse, C., Naumann, G., Alfieri, L., Salamon, P., and Vogt, J. (2019). Predictability of the european heat and cold waves. *Climate Dynamics*, 52(3):2481–2495.

Lavaysse, C., Vogt, J., Toreti, A., Carrera, M. L., and Pappenberger, F. (2018). On the use of weather regimes to forecast meteorological drought over europe. *Natural Hazards and Earth System Sciences*, 18(12):3297–3309.

Lee, M.-H., Lee, S., Song, H.-J., and Ho, C.-H. (2017). The recent increase in the occurrence of a boreal summer teleconnection and its relationship with temperature extremes. *Journal of Climate*, 30(18):7493–7504.

Lee, R. W., Woolnough, S. J., Charlton-Perez, A. J., and Vitart, F. (2019). Enso modulation of mjo teleconnections to the north atlantic and europe. *Geophysical Research Letters*.

Lee, S., L'Heureux, M., Wittenberg, A. T., Seager, R., O'Gorman, P. A., and Johnson, N. C. (2022). On the future zonal contrasts of equatorial pacific climate: Perspectives from observations, simulations, and theories. *npj Climate and Atmospheric Science*, 5(1):82.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H. M., Diamantakis, M., Dutra, E., et al. (2017). Stochastic representations of model uncertainties at ecmwf: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143(707):2315–2339.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, 227(7):3515–3539.

Li, W., Pan, B., Xia, J., and Duan, Q. (2022). Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, 605:127301.

Lim, Y.-K., Cullather, R. I., Nowicki, S. M., and Kim, K.-M. (2019). Inter-relationship between subtropical pacific sea surface temperature, arctic sea ice concentration, and north atlantic oscillation in recent summers. *Scientific reports*, 9(1):1–11.

Lin, H. and Brunet, G. (2018). Extratropical response to the mjo: Nonlinearity and sensitivity to the initial state. *Journal of the Atmospheric Sciences*, 75(1):219–234.

Lin, H., Brunet, G., and Derome, J. (2009). An observed connection between the north atlantic oscillation and the madden–julian oscillation. *Journal of Climate*, 22(2):364–380.

Liu, Z. and Alexander, M. (2007). Atmospheric bridge, oceanic tunnel, and global climatic teleconnections. *Reviews of Geophysics*, 45(2).

Lopez-Gomez, I., McGovern, A., Agrawal, S., and Hickey, J. (2023). Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, 2(1):e220035.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307.

Lovejoy, S. (2015). A voyage through scales, a missing quadrillion and why the climate is not what you expect. *Climate Dynamics*, 44(11):3187–3210.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ma, Q. and Franzke, C. L. (2021). The role of transient eddies and diabatic heating in the maintenance of european heat waves: a nonlinear quasi-stationary wave perspective. *Climate Dynamics*, 56(9):2983–3002.

Ma, S., Pitman, A. J., Lorenz, R., Kala, J., and Srbinovsky, J. (2016). Earlier green-up and spring warming amplification over europe. *Geophysical Research Letters*, 43(5):2011–2018.

Magnusson, L. (2017). Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, 143(706):2129–2142.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I. (2022a). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, 1(4):e220012.

Mamalakis, A., Ebert-Uphoff, I., and Barnes, E. A. (2022b). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, 1:e8.

Manola, I., Selten, F., de Vries, H., and Hazeleger, W. (2013). "waveguidability" of

idealized jets. *Journal of Geophysical Research: Atmospheres*, 118(18):10–432.

Manrique-Suñén, A., Gonzalez-Reviriego, N., Torralba, V., Cortesi, N., and Doblas-Reyes, F. J. (2020). Choices in the verification of s2s forecasts and their implications for climate services. *Monthly Weather Review*, 148(10):3995 – 4008.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). A pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the american Meteorological Society*, 78(6):1069–1080.

Manzanas, R., Lucero, A., Weisheimer, A., and Gutiérrez, J. (2018). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dynamics*, 50(3-4):1161–1176.

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., Dirmeyer, P. A., Ferranti, L., Johnson, N. C., Jones, J., et al. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, 101(5):E608–E625.

Mastrangelo, D. and Malguzzi, P. (2019). Verification of two years of cnr-isac subseasonal forecasts. *Weather and Forecasting*, 34(2):331–344.

Mastrantonas, N., Magnusson, L., Pappenberger, F., and Matschullat, J. (2022). What do large-scale patterns teach us about extreme precipitation over the mediterranean at medium- and extended-range forecasts? *Quarterly Journal of the Royal Meteorological Society*, n/a(n/a):1–16.

Matsueda, M. (2009). Blocking predictability in operational medium-range ensemble forecasts. *Sola*, 5:113–116.

Mayer, K. and Barnes, E. A. (2021). Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, 48(10):e2020GL092092.

Mayer, K. J. and Barnes, E. A. (2022). Quantifying the effect of climate change on midlatitude subseasonal prediction skill provided by the tropics. *Geophysical Research Letters*, 49:e2022GL098663.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

McKinnon, K. A., Rhines, A., Tingley, M., and Huybers, P. (2016). Long-lead predictions of eastern united states hot days from pacific sea surface temperatures. *Nature Geoscience*, 9(5):389.

Mecikalski, J. R., Sandmæl, T. N., Murillo, E. M., Homeyer, C. R., Bedka, K. M., Apke, J. M., and Jewett, C. P. (2021). A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar–goes satellite–lightning observations. *Monthly Weather Review*, 149(6):1725–1746.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., et al. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, 101(6):E869–E896.

Michel, C. and Rivière, G. (2011). The link between rossby wave breakings and weather regime transitions. *Journal of the Atmospheric Sciences*, 68(8):1730–1748.

Michelangeli, P.-A., Vautard, R., and Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of the atmospheric sciences*, 52(8):1237–1256.

Miller, D. E. and Wang, Z. (2019). Skillful seasonal prediction of eurasian winter blocking and extreme temperature frequency. *Geophysical Research Letters*, 46(20):11530–11538.

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., and Bouchet, F. (2023). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Phys. Rev. Fluids*, 8:040501.

Miralles, D. G., Gentine, P., Seneviratne, S. I., and Teuling, A. J. (2019). Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Annals of the New York Academy of Sciences*, 1436(1):19.

Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. (2018). Skill of sub-seasonal forecasts in europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres*.

Mouatadid, S., Orenstein, P., Flaspohler, G., Oprescu, M., Cohen, J., Wang, F., Knight, S., Geogdzhayeva, M., Levang, S., Fraenkel, E., and Mackey, L. (2021). Learned benchmarks for subseasonal forecasting. *ArXiv*.

Mueller, S. T., Hoffman, R. R., Clancey, W. J., Emrey, A., and Klein, G. (2019). Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv*.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., S. Boussetta, M., Choulga, Harrigan, S., Hersbach, H., Martens, B., Miralles, D., Piles, M., Fernández, N. R., and Zsoter, E. (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383.

Nakanishi, T., Tachibana, Y., and Ando, Y. (2021). Possible semi-circumglobal teleconnection across eurasia driven by deep convection over the sahel. *Climate Dynamics*, 57:2287–2299.

National Academies of Sciences, Engineering, and Medicine (2016). *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*. The National Academies Press, Washington, DC.

Neddermann, N.-C., Müller, W. A., Dobrynin, M., Düsterhus, A., and Baehr, J. (2019). Seasonal predictability of european summer climate re-assessed. *Climate Dynamics*, 53(5):3039–3056.

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., Mantua, N. J., Miller, A. J., Minobe, S., Nakamura, H., et al. (2016). The pacific decadal oscillation, revisited. *Journal of Climate*, 29(12):4399–4427.

Nicolis, C. (2016). Error dynamics in extended-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 142(696):1222–1231.

O'Reilly, C. H., Weisheimer, A., Woollings, T., Gray, L. J., and MacLeod, D. (2019a). The importance of stratospheric initial conditions for winter north atlantic oscillation predictability and implications for the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*, 145(718):131–146.

O'Reilly, C. H., Woollings, T., Zanna, L., and Weisheimer, A. (2019b). An interdecadal

shift of the extratropical teleconnection from the tropical pacific during boreal summer. *Geophysical Research Letters*, 46(22):13379–13388.

Orsolini, Y., Senan, R., Balsamo, G., Doblas-Reyes, F., Vitart, F., Weisheimer, A., Carrasco, A., and Benestad, R. (2013). Impact of snow initialization on sub-seasonal forecasts. *Climate dynamics*, 41(7-8):1969–1982.

Osborne, J. M., Collins, M., Screen, J. A., Thomson, S. I., and Dunstone, N. (2020). The North Atlantic as a Driver of Summer Atmospheric Circulation. *Journal of Climate*, 33(17):7335–7351.

Ossó, A., Sutton, R., Shaffrey, L., and Dong, B. (2020). Development, amplification and decay of atlantic/european summer weather patterns linked to spring north atlantic sea surface temperatures. *Journal of Climate*, 33(14):5939–5951.

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

Palmer, T. (2015). Modelling: Build imprecise supercomputers. *Nature*, 526(7571):32–33.

Palmer, T. (2017). The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, 9(2):730–734.

Palmer, T. N. (1993). Extended-range atmospheric prediction and the lorenz model. *Bulletin of the American Meteorological Society*, 74(1):49–66.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., and Lee, J. (2022). Improving seasonal forecast using probabilistic deep learning. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002766. e2021MS002766 2021MS002766.

Pasquier, J., Pfahl, S., and Grams, C. M. (2019). Modulation of atmospheric river occurrence and associated precipitation extremes in the north atlantic region by european weather regimes. *Geophysical Research Letters*, 46(2):1014–1023.

Perkins, S. E. (2015). A review on the scientific understanding of heatwaves—their measurement, driving mechanisms, and changes at the global scale. *Atmospheric Research*, 164:242–267.

Petoukhov, V., Rahmstorf, S., Petri, S., and Schellnhuber, H. J. (2013). Quasiresonant amplification of planetary waves and recent northern hemisphere weather extremes. *Proceedings of the National Academy of Sciences*, 110(14):5336–5341.

Pfahl, S. (2014). Characterising the relationship between weather extremes in europe and synoptic circulation features. *Natural Hazards and Earth System Sciences*, 14(6):1461–1475.

Pfleiderer, P. and Coumou, D. (2018). Quantification of temperature persistence over the northern hemisphere land-area. *Climate dynamics*, 51(1-2):627–637.

Platzman, G. W. (1968). The rossby wave. *Quarterly Journal of the Royal Meteorological Society*, 94(401):225–248.

Privé, N. C. and Errico, R. M. (2015). Spectral analysis of forecast error investigated with an observing system simulation experiment. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1):25977.

Prodhomme, C., Doblas-Reyes, F., Bellprat, O., and Dutra, E. (2016). Impact of land-surface initialization on sub-seasonal to seasonal forecasts over europe. *Climate dynamics*, 47:919–935.

Prodhomme, C., Materia, S., Ardilouze, C., White, R. H., Batt'e, L., Guemas, V., Fragkoulidis, G., and García-Serrano, J. (2021). Seasonal prediction of european summer heatwaves. *Climate dynamics*, 58:2149–2166.

Pyrina, M. and Domeisen, D. I. (2022). Sub-seasonal predictability of onset, duration, and intensity of european heat extremes. *Quarterly Journal of the Royal Meteorological Society*.

Quesada, B., Vautard, R., Yiou, P., Hirschi, M., and Seneviratne, S. I. (2012). Asymmetric european summer heat predictability from wet and dry southern winters and springs. *Nature Climate Change*, 2(10):736.

Quinting, J. and Vitart, F. (2019). Representation of synoptic-scale rossby wave packets and blocking in the s2s prediction project database. *Geophysical Research Letters*, 46(2):1070–1078.

Ras, G., Xie, N., van Gerven, M., and Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900.

Rasp, S. and Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405.

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling & Software*, 144:105159.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204.

Richardson, D., Fowler, H. J., Kilsby, C. G., Neal, R., and Dankers, R. (2020). Improving sub-seasonal forecast skill of meteorological drought: a weather pattern approach. *Natural Hazards and Earth System Sciences*, 20:107–124.

Richardson, D. S. (2000). Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563):649–667.

Roads, J. O. (1986). Forecasts of time averages with a numerical weather prediction model. *Journal of the atmospheric sciences*, 43(9):871–893.

Robertson, A. W., Vigaud, N., Yuan, J., and Tippett, M. K. (2020). Toward Identifying Subseasonal Forecasts of Opportunity Using North American Weather Regimes. *Monthly Weather Review*, 148(5):1861–1875.

Rodney, M., Lin, H., and Derome, J. (2013). Subseasonal prediction of wintertime north american surface air temperature during strong mjo events. *Monthly Weather Review*, 141(8):2897–2909.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., and Wernli, H. (2018). Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, 99:1015–1026.

Röthlisberger, M., Frossard, L., Bosart, L. F., Keyser, D., and Martius, O. (2019). Recurrent synoptic-scale rossby wave patterns and their effect on the persistence of cold and hot spells. *Journal of Climate*, 32(11):3207–3226.

Röthlisberger, M., Sprenger, M., Flaounas, E., Beyerle, U., and Wernli, H. (2020). The substructure of extremely hot summers in the northern hemisphere. *Weather and*

Climate Dynamics, 1(1):45–62.

Rousi, E., Kornhuber, K., Beobide-Arsuaga, G., Luo, F., and Coumou, D. (2022). Accelerated western european heatwave trends linked to more-persistent double jets over eurasia. *Nature communications*, 13(1):1–11.

Rousi, E., Selten, F., Rahmstorf, S., and Coumou, D. (2021). Changes in north atlantic atmospheric circulation in a warmer climate favor winter flooding and summer drought over europe. *Journal of Climate*, 34(6):2277–2295.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11).

Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., Hartman, D., Marwan, N., Paluš, M., and Kurths, J. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, 6:8502.

Runge, J., Petoukhov, V., and Kurths, J. (2014). Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27(2):720–739.

Russo, S., Dosio, A., Graversen, R. G., Sillmann, J., Carrao, H., Dunbar, M. B., Singleton, A., Montagna, P., Barbola, P., and Vogt, J. V. (2014). Magnitude of extreme heat waves in present climate and their projection in a warming world. *Journal of Geophysical Research: Atmospheres*, 119(22):12–500.

Saha, M., Mitra, P., and Nanjundiah, R. S. (2016). Autoencoder-based identification of predictors of indian monsoon. *Meteorology and Atmospheric Physics*, 128(5):613–628.

Sardeshmukh, P. D. and Hoskins, B. J. (1988). The generation of global rotational flow by steady idealized tropical divergence. *Journal of the Atmospheric Sciences*, 45(7):1228–1251.

Saunders, K., Stephenson, A., and Karoly, D. (2021). A regionalisation approach for rainfall based on extremal dependence. *Extremes*, 24(2):215–240.

Schaller, N., Sillmann, J., Anstey, J., Fischer, E., Grams, C., and Russo, S. (2018). Influence of blocking on northern european and western russian heatwaves in large climate model ensembles. *Environmental Research Letters*, 13(5):054015.

Scher, S. and Messori, G. (2021). Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, 13(2).

Scheuerer, M., Switanek, M. B., Worsnop, R. P., and Hamill, T. M. (2020). Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over california. *Monthly Weather Review*, 148(8):3489–3506.

Schneidereit, A., Schubert, S., Vargin, P., Lunkeit, F., Zhu, X., Peters, D. H., and Fraedrich, K. (2012). Large-scale flow and the long-lasting blocking high over russia: Summer 2010. *Monthly Weather Review*, 140(9):2967–2981.

Schubert, S., Wang, H., and Suarez, M. (2011). Warm season subseasonal variability and climate extremes in the northern hemisphere: The role of stationary rossby waves. *Journal of Climate*, 24(18):4773–4792.

Schubert, S. D., Wang, H., Koster, R. D., Suarez, M. J., and Groisman, P. Y. (2014). Northern eurasian heat waves and droughts. *Journal of Climate*, 27(9):3169–3207.

Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097.

Schulz, B. and Lerch, S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257.

Schumacher, D. L., Keune, J., Van Heerwaarden, C. C., de Arellano, J. V.-G., Teuling, A. J., and Miralles, D. G. (2019). Amplification of mega-heatwaves through heat torrents fuelled by upwind drought. *Nature Geoscience*, 12(9):712–717.

Seager, R., Henderson, N., and Cane, M. (2022). Persistent discrepancies between observed and modeled trends in the tropical pacific ocean. *Journal of Climate*, 35(14):4571 – 4584.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. Technical report, University of California.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161.

Shao, Y., Wang, Q. J., Schepen, A., and Ryu, D. (2022). Introducing long-term trends into sub-seasonal temperature forecasts through trend-aware post-processing. *International Journal of Climatology*, pages 1–17.

Shukla, J. (1981). Dynamical predictability of monthly means. *Journal of the Atmospheric Sciences*, 38(12):2547–2572.

Shukla, J. (1998). Predictability in the midst of chaos: A scientific basis for climate forecasting. *science*, 282(5389):728–731.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., Seneviratne, S. I., Vautard, R., Zhang, X., and Zwiers, F. W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and climate extremes*, 18:65–74.

Silva, S. J., Keller, C. A., and Hardin, J. (2022). Using an explainable machine learning approach to characterize earth system model errors: Application of shap analysis to modeling lightning flash occurrence. *Journal of Advances in Modeling Earth Systems*, 14(4):e2021MS002881. e2021MS002881 2021MS002881.

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M. (2023). Hybrid forecasting: blending climate predictions with ai models. *Hydrology and Earth System Sciences*, 27(9):1865–1889.

Slingo, J. and Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4751–4767.

Smith, L. A., Du, H., Suckling, E. B., and Niehörster, F. (2015). Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1085–1100.

Sousa, P. M., Barriopedro, D., García-Herrera, R., Woollings, T., and Trigo, R. M. (2021). A new combined detection algorithm for blocking and subtropical ridges. *Journal of Climate*, 34(18):7735–7758.

Sousa, P. M., Trigo, R. M., Barriopedro, D., Soares, P. M., and Santos, J. A. (2018). European temperature responses to blocking and ridge regional patterns. *Climate dynamics*, 50:457–477.

Specq, D. and Batté, L. (2020). Improving subseasonal precipitation forecasts through a statistical–dynamical approach: application to the southwest tropical pacific. *Cli-*

mate Dynamics, 55(7):1913–1927.

Spiegelhalter, D. (2019). Probability - the language of uncertainty and variability. In *The art of statistics: Learning from data*, chapter 8. Penguin UK.

Stan, C. and Krishnamurthy, V. (2019). Intra-seasonal and seasonal variability of the northern hemisphere extra-tropics. *Climate Dynamics*, pages 1–19.

Stan, C., Straus, D. M., Frederiksen, J. S., Lin, H., Maloney, E. D., and Schumacher, C. (2017). Review of tropical-extratropical teleconnections on intraseasonal time scales. *Reviews of Geophysics*, 55(4):902–937.

Stéfanon, M., Drobinski, P., d'Andrea, F., and de Noblet-Ducoudré, N. (2012). Effects of interactive vegetation phenology on the 2003 summer heat waves. *Journal of Geophysical Research: Atmospheres*, 117(D24).

Stefanon, M., D'Andrea, F., and Drobinski, P. (2012). Heatwave classification over europe and the mediterranean region. *Environmental Research Letters*, 7(1):014023.

Strazzo, S., Collins, D. C., Schepen, A., Wang, Q., Becker, E., and Jia, L. (2019). Application of a hybrid statistical–dynamical system to seasonal prediction of north american temperature and precipitation. *Monthly Weather Review*, 147(2):607–625.

Suarez-Gutierrez, L., Müller, W. A., Li, C., and Marotzke, J. (2020). Dynamical and thermodynamical drivers of variability in european summer heat extremes. *Climate Dynamics*, 54:4351—-4366.

Sun, X., Ding, Q., Wang, S.-Y. S., Topál, D., Li, Q., Castro, C., Teng, H., Luo, R., and Ding, Y. (2022). Enhanced jet stream waviness induced by suppressed tropical pacific convection during boreal summer. *Nature Communications*, 13(1):1–10.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393.

Teng, H. and Branstator, G. (2019). Amplification of waveguide teleconnections in the boreal summer. *Current Climate Change Reports*, 5(4):421–432.

Teng, H., Leung, R., Branstator, G., Lu, J., and Ding, Q. (2022). Warming pattern over the northern hemisphere midlatitudes in boreal summer 1979–2020. *Journal of Climate*, 35(11):3479–3494.

Tilloy, A., Malamud, B., and Joly-Laugel, A. (2022). A methodology for the spatiotemporal identification of compound hazards: Wind and precipitation extremes in great britain (1979–2019). *Earth System Dynamics*, pages 1–45.

Ting, M. (1994). Maintenance of northern summer stationary waves in a gcm. *Journal of the atmospheric sciences*, 51(22):3286–3308.

Tippett, M. K., DelSole, T., Mason, S. J., and Barnston, A. G. (2008). Regression-based methods for finding coupled patterns. *Journal of Climate*, 21(17):4384–4398.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002.

Toth, Z. and Buizza, R. (2019). Weather forecasting: What sets the forecast skill horizon? In Robertson, A. W. and Vitart, F., editors, *Sub-Seasonal to Seasonal Prediction*, pages 17–45. Elsevier.

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y. (2003). Probability and ensemble forecasts. In Jolliffe, I. T. and Stephenson, D. B., editors, *Forecast verification: A practitioner's guide in atmospheric science*, pages 137–163. Wiley.

Trenary, L. and DelSole, T. (2022). Advancing interpretability of machine-learning

prediction models. *Environmental Data Science*, 1:e14.

Trenary, L. and DelSole, T. (2023). Skillful statistical prediction of subseasonal temperature by training on dynamical model data. *Environmental Data Science*, 2:e7.

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

Trenberth, K. E. and Stepaniak, D. P. (2001). Indices of el niño evolution. *Journal of climate*, 14(8):1697–1701.

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

van den Hurk, B., Doblas-Reyes, F., Balsamo, G., Koster, R. D., Seneviratne, S. I., and Camargo, H. (2012). Soil moisture effects on seasonal temperature and precipitation forecast scores in europe. *Climate Dynamics*, 38:349–362.

van Oldenborgh, G. J., Hendon, H., Stockdale, T., L'Heureux, M., De Perez, E. C., Singh, R., and Van Aalst, M. (2021). Defining el niño indices in a warming climate. *Environmental research letters*, 16(4):044003.

van Oldenborgh, G. J., Reyes, F. D., Drijfhout, S., and Hawkins, E. (2013). Reliability of regional climate model trends. *Environmental Research Letters*, 8(1):014055.

van Oldenborgh, G. J., Wehner, M. F., Vautard, R., Otto, F. E. L., Seneviratne, S. I., Stott, P. A., Hegerl, G. C., Philip, S. Y., and Kew, S. F. (2022). Attributing and projecting heatwaves is hard: We can do better. *Earth's Future*, 10:e2021EF002271.

Van Schaeybroeck, B. and Vannitsem, S. (2015). Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688):807–818.

Van Schaeybroeck, B. and Vannitsem, S. (2018). Chapter 10 - postprocessing of long-range forecasts. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 267–290. Elsevier.

van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., and Schmeits, M. (2020). The influence of aggregation and statistical post-processing on the subseasonal predictability of european temperatures. *Quarterly Journal of the Royal Meteorological Society*, 146(731):2654–2670.

van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., and Schmeits, M. (2022). Using explainable machine learning forecasts to discover sub-seasonal drivers of high summer temperatures in western and central europe. *Monthly Weather Review*, 150(5):1115–1134.

van Straaten, C., Whan, K., Coumou, D., van den Hurk, B., and Schmeits, M. (2023). Correcting sub-seasonal forecast errors with an explainable ann to understand misrepresented sources of predictability of european summer temperatures. *Artificial Intelligence for the Earth Systems*.

van Straaten, C., Whan, K., and Schmeits, M. (2018). Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *Journal of Hydrometeorology*, 19(11):1815–1833.

Vannitsem, S. (2017). Predictability of large-scale atmospheric motions: Lyapunov exponents and error dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(3):032101.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699.

Vannitsem, S. and Liang, X. S. (2022). Dynamical dependencies at monthly and interannual time scales in the climate system: Study of the north pacific and atlantic regions. *Tellus A: Dynamic Meteorology and Oceanography*, 74(1):141–158.

Vannitsem, S., Wilks, D. S., and Messner, J. (2018). *Statistical postprocessing of ensemble forecasts*. Elsevier.

Vautard, R. (1990). Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Monthly weather review*, 118(10):2056–2081.

Veldkamp, S., Whan, K., Dirksen, S., and Schmeits, M. (2021). Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149(4):1141–1152.

Vigaud, N., Robertson, A. W., and Tippett, M. K. (2017). Multimodel ensembling of subseasonal precipitation forecasts over north america. *Monthly Weather Review*, 145(10):3913–3928.

Vijverberg, S. and Coumou, D. (2022). The role of the pacific decadal oscillation and ocean-atmosphere interactions in driving us temperature predictability. *npj Climate and Atmospheric Science*, 5(1):1–11.

Vijverberg, S., Schmeits, M., van der Wiel, K., and Coumou, D. (2020). Subseasonal statistical forecasts of eastern us hot temperature events. *Monthly Weather Review*, 148(12):4799–4822.

Vitart, F. (2017). Madden—julian oscillation prediction and teleconnections in the s2s database. *Quarterly Journal of the Royal Meteorological Society*, 143(706):2210–2220.

Vitart, F., Alonso-Balmaseda, M., Ferranti, L., Benedetti, A., Balan-Sarojini, B., Tietsche, S., Yao, J., Janousek, M., Balsamo, G., Leutbecher, M., Bechtold, P., Polichtchouk, I., Richardson, D., Stockdale, T., and Roberts, C. D. (2019). *Extended-range prediction*, volume 854. European Centre for Medium-Range Weather Forecasts.

Vitart, F., Robertson, A., Spring, A., Pinault, F., Roškar, R., Cao, W., Bech, S., Bienkowski, A., Caltabiano, N., Coning, E. D., Denis, B., Dirkson, A., Dramsch, J., Dueben, P., Gierschendorf, J., Kim, H. S., Nowak, K., Landry, D., Lledó, L., Palma, L., Rasp, S., and Zhou, S. (2022). Outcomes of the wmo prize challenge to improve sub-seasonal to seasonal predictions using artificial intelligence. *Bulletin of the American Meteorological Society*.

Vitart, F. and Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate and Atmospheric Science*, 1(1):3.

Vitart, F. and Robertson, A. W. (2019). Chapter 1 - introduction: Why sub-seasonal to seasonal prediction (s2s)? In Robertson, A. W. and Vitart, F., editors, *Sub-Seasonal to Seasonal Prediction*, pages 3–15. Elsevier.

Vogel, M. M., Zscheischler, J., Fischer, E. M., and Seneviratne, S. (2020). Development of future heatwaves for different hazard thresholds. *Journal of Geophysical Research: Atmospheres*, 125(9):e2019JD032070.

Wallace, J. M. and Gutzler, D. S. (1981). Teleconnections in the geopotential height

field during the northern hemisphere winter. *Monthly weather review*, 109(4):784–812.

Wehrli, K., Guillod, B. P., Hauser, M., Leclair, M., and Seneviratne, S. I. (2019). Identifying key driving processes of major recent heatwaves. *Journal of Geophysical Research: Atmospheres*, 124:11746–11765.

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 142:399–432.

Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitart, F., and Appenzeller, C. (2008). Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, 136(12):5162–5182.

Weirich Benet, E., Pyrina, M., Jiménez-Esteve, B., Fraenkel, E., Cohen, J., and Domeisen, D. I. (2023). Sub-seasonal prediction of central european summer heatwaves with linear and random forest machine learning models. *Artificial Intelligence for the Earth Systems*, pages 1 – 52.

Weisheimer, A., Doblas-Reyes, F. J., Jung, T., and Palmer, T. (2011). On the predictability of the extreme summer 2003 over europe. *Geophysical Research Letters*, 38(5):L05704.

Weyn, J. A., Durran, D. R., and Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, 13(7):e2021MS002502.

Whan, K. and Schmeits, M. (2018). Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Monthly Weather Review*, 146(11):3651–3673.

Whan, K., Zscheischler, J., Orth, R., Shongwe, M., Rahimi, M., Asare, E. O., and Seneviratne, S. I. (2015). Impact of soil moisture on extreme maximum temperatures in europe. *Weather and Climate Extremes*, 9:57–67.

Wheeler, M. C. and Hendon, H. H. (2004). An all-season real-time multivariate mjo index: Development of an index for monitoring and prediction. *Monthly weather review*, 132(8):1917–1932.

Wheeler, M. C., Zhu, H., Sobel, A. H., Hudson, D., and Vitart, F. (2017). Seamless precipitation prediction skill comparison between two global models. *Quarterly Journal of the Royal Meteorological Society*, 143(702):374–383.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., et al. (2017). Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological applications*, 24(3):315–325.

White, C. J., Domeisen, D. I., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw, D. J., et al. (2021). Advances in the application and utility of subseasonal-to-seasonal predictions. *Bulletin of the American Meteorological Society*, pages 1–57.

White, R. H., Kornhuber, K., Martius, O., and Wirth, V. (2022). From atmospheric waves to heatwaves: A waveguide perspective for understanding and predicting concurrent, persistent, and extreme extratropical weather. *Bulletin of the American Meteorological Society*, 103(3):E923–E935.

Wilby, R., Abrahart, R., and Dawson, C. (2003). Detection of conceptual model rainfall—runoff processes inside an artificial neural network. *Hydrological Sciences Journal*, 48(2):163–181.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

Wilks, D. S. (2014). Comparison of probabilistic statistical forecast and trend adjustment methods for north american seasonal temperatures. *Journal of Applied Meteorology and Climatology*, 53(4):935–949.

Wilks, D. S. (2018). Univariate ensemble postprocessing. In *Statistical postprocessing of ensemble forecasts*, pages 49–89. Elsevier.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, 135(6):2379–2390.

Wilks, D. S. and Vannitsem, S. (2018). Uncertain forecasts from deterministic dynamics. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical Postprocessing of Ensemble Forecasts*, pages 1–13. Elsevier, 1 edition.

Wills, R. C., Dong, Y., Proistosecu, C., Armour, K. C., and Battisti, D. S. (2022). Systematic climate model biases in the large-scale patterns of recent sea-surface temperature and sea-level pressure change. *Geophysical Research Letters*, page e2022GL100011.

Wirth, V., Riemer, M., Chang, E. K., and Martius, O. (2018). Rossby wave packets on the midlatitude waveguide—a review. *Monthly Weather Review*, 146(2018):1965–2001.

Wolf, G., Brayshaw, D. J., Klingaman, N. P., and Czaja, A. (2018). Quasi-stationary waves and their impact on european weather and extreme events. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2431–2448.

Wolf, G., Czaja, A., Brayshaw, D., and Klingaman, N. (2020). Connection between sea surface anomalies and atmospheric quasi-stationary waves. *Journal of Climate*, 33(1):201–212.

Woollings, T. (2010). Dynamical influences on european climate: an uncertain future. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1924):3733–3756.

World Meteorological Organization (2015). *Seamless prediction of the Earth system: from minutes to months*. World Meteorological Organization.

Wu, Z. and Lin, H. (2012). Interdecadal variability of the enso–north atlantic oscillation connection in boreal summer. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1668–1675.

Wulff, C. O. and Domeisen, D. I. (2019). Higher subseasonal predictability of extreme hot european summer temperatures as compared to average summers. *Geophysical Research Letters*, 46:11520–11529.

Wulff, C. O., Greatbatch, R. J., Domeisen, D. I., Gollan, G., and Hansen, F. (2017). Tropical forcing of the summer east atlantic pattern. *Geophysical Research Letters*, 44(21):11–166.

Wulff, C. O., Vitart, F., and Domeisen, D. I. (2022). Influence of trends on subseasonal temperature prediction skill. *Quarterly Journal of the Royal Meteorological Society*.

Wulff, C. O. W. (2021). *Subseasonal Prediction and Predictability of European Temperatures*. phdthesis, ETH Zurich.

Yadav, P. and Straus, D. M. (2017). Circulation response to fast and slow mjo episodes.

Monthly Weather Review, 145(5):1577–1596.

Yang, Z. and Villarini, G. (2019). Examining the capability of reanalyses in capturing the temporal clustering of heavy precipitation across europe. *Climate Dynamics*, pages 1–13.

Ying, Y. and Zhang, F. (2017). Practical and intrinsic predictability of multiscale weather and convectively coupled equatorial waves during the active phase of an mjo. *Journal of the Atmospheric Sciences*, 74(11):3771–3785.

Yiou, P., Goubanova, K., Li, Z., and Nogaj, M. (2008). Weather regime dependence of extreme value statistics for summer temperature and precipitation. *Nonlinear Processes in Geophysics*, 15(3):365–378.

Yoo, C., Johnson, N. C., Chang, C.-H., Feldstein, S. B., and Kim, Y.-H. (2018). Subseasonal prediction of wintertime east asian temperature based on atmospheric teleconnections. *Journal of Climate*, 31:9351–9366.

Young, M., Heinrich, V., Black, E., and Asfaw, D. (2020). Optimal spatial scales for seasonal forecasts over africa. *Environmental Research Letters*, 15(9):094023.

Zampieri, M., D'andrea, F., Vautard, R., Ciais, P., de Noblet-Ducoudré, N., and Yiou, P. (2009). Hot european summers and the role of soil moisture in the propagation of mediterranean drought. *Journal of Climate*, 22(18):4747–4758.

Zampieri, M., Toreti, A., Schindler, A., Scoccimarro, E., and Gualdi, S. (2017). Atlantic multi-decadal oscillation influence on weather regimes over europe and the mediterranean in spring and summer. *Global and Planetary Change*, 151:92–100.

Zhang, C. (2013). Madden–julian oscillation: Bridging weather and climate. *Bulletin of the American Meteorological Society*, 94(12):1849–1870.

Zhang, F., Qiang Sun, Y., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., and Emanuel, K. (2019a). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*.

Zhang, R., Sun, C., Zhu, J., Zhang, R., and Li, W. (2020). Increased european heat waves in recent decades in response to shrinking arctic sea ice and eurasian snow cover. *NPJ Climate and Atmospheric Science*, 3(1):1–9.

Zhang, Y., Huang, W., and Zhong, D. (2019b). Major moisture pathways and their importance to rainy season precipitation over the sanjiangyuan region of the tibetan plateau. *Journal of Climate*, 32(20):6837–6857.

Zschenderlein, P., Fink, A. H., Pfahl, S., and Wernli, H. (2019). Processes determining heat waves across different european climates. *Quarterly Journal of the Royal Meteorological Society*, 145(724):2973–2989.

Acknowledgments

Welcome to the acknowledgments. Truth is, I found this section really hard to write. This is not because I've forgotten who's been important to me, but actually because one of those people is dearly missed. The story of my PhD, is, among other things, the story of losing my sister Milou. I will briefly tell this story, and then proceed with the acknowledgments as you know them.

It was roughly six years ago that, as part of my master's degree, I was doing an internship under Maurice and Kiri at KNMI. The internship was going well, but other than that, times were pretty tough. One day was particularly fortunate though. I remember that Maurice called me into his office. That week I had been quite distracted, so I expected Maurice to say some angry words about my lack of progress. Instead he expressed his fondness of my work and proposed to jointly write a PhD proposal. I was flattered of course, and said yes. It set in motion a chain of events that resulted in this thesis.

Less fortunate was that towards the end of the internship another chain of events ended with my sister's suicide. I received the news when I was on my way to visit her.

Skip forward a few weeks and I had a master's degree but little to look forward to. The result was a hard and empty year, at the end of which NWO funded the proposal. I could start if I wanted, but I was not sure. The only thing I knew was that 'I had to get back on track' somehow, and that my supervisors would be good people.

In practice the following years became both easy and hard. Hard was the isolation involved in doing a one-person project, particularly during corona. Also hard was the abstraction, away from personal needs, and into the distant position from which science is conducted. Of course I liked weather and climate science and believed in its relevance, but this could never fully bridge the gap. Luckily, three therapists and one reverend have helped me navigate this reality. These are Kees, Peter, Denise and

Riekje.

Friends, colleagues and supervisors have also made the experience easier. First of all, this is due to the very large number of engaging, dedicated people walking around at IVM and KNMI. From all these colleagues I want to highlight current and past members of IVM's Climate Extremes group, KNMI's statistical post-processing group, IVM's running group, people from KNMI's 'weather chamber' and jong-KNMI. Let's also not forget that I've enjoyed the support of no less than four supervisors. Thank you Dim, Maurice, Kiri and Bart. You gave me feedback that I sometimes found hard to swallow, but it definitely made the work better, and you came with a lot of interest, empathy and good company. More broadly, it has also been rewarding to meet the international community at conferences and workshops. Here I want to highlight events at ECMWF, EGU, the Trieste summer school and the Lorentz-workshop that we organized.

Then of course, life is also happening thanks to a great number of other important people. Among these are former happyChaos members, former Kairos members, ekte-138-ers, Frits-Coers roomies, assorted Nimma- and Groningen-crews, 'the Huricoons', my cousins, my family-in-law, and my parents. I have a lot of love for you all.

Appropriately, the last acknowledgment will go to Jet. Parallel to the PhD there has been this other story... and it's a good one.

SENSE

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

D I P L O M A

for specialised PhD training

The Netherlands research school for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Joachim Wilhelmus van Straaten

born on 9 June 1995 in Nijmegen, The Netherlands

has successfully fulfilled all requirements of the
educational PhD programme of SENSE.

Amsterdam, 20 November 2023

Chair of the SENSE board

Prof. dr. Martin Wassen

The SENSE Director

Prof. Philipp Pattberg

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Joachim Wilhelmus van Straaten** has successfully fulfilled all requirements of the educational PhD programme of SENSE with a work load of 44.6 EC, including the following activities:

SENSE PhD Courses

- Environmental research in context (2018)
- Research in context activity: 'Weather Stories: applying for EGU outreach grant, writing two stories, live storytelling for NVTG and `Boom open Mic' (2019-2022)

Other PhD and Advanced MSc Courses

- Dynamical Meteorology, Utrecht University (2018)
- Predictability and ensemble forecasting, ECMWF (2019)
- Scientific Integrity, VU Amsterdam (2020)
- Python in High-Performance-Computing., PRACE / FutureLearn (2020)
- Artificial Intelligence for Earth System Science , NCAR (2020)
- Trustworthy Artificial Intelligence for Environmental Science, AI2ES/NCAR (2021)
- Artificial Intelligence for Detection and Attribution of Climate Extremes, ICTP (2022)

Management and Didactic Skills Training

- Writing a multidisciplinary workshop proposal for Machine learning in research (2019)
- Giving a KNMI workshop at the UU-career day (2020)
- Participated in a networking day organized by "Beta in bestuur en beleid" (2022)
- Co-writing of proposal and organising a workshop on explainable AI for Subseasonal forecasting. Lorentz Center Leiden and eScience Center (2021-2022)
- Teaching in the MSc course 'Methods of Global Environmental Change' (2020-2021)

Selection of Oral Presentations

- *The influence of aggregation and statistical post-processing on the sub-seasonal predictability of European temperatures*. EGU 2020 4-8 May 2020, Online
- *Using explainable machine learning forecasts to discover drivers of high summer temperatures in western and central Europe.* ECMWF workshop on machine learning for numerical weather prediction, 14-15 June 2021, Online
- *Improving sub-seasonal forecasts by correcting missing teleconnections using ANN-based post-processing*. EGU, 23-27 May 2022, Vienna, Austria
- *Understanding Climate extremes through AI*. Innovation Center for Artificial Intelligence, 16 November 2022, Utrecht, The Netherlands

SENSE coordinator PhD education

Dr. ir. Peter Vermeulen