



Hidden amongst chaos

Dynamics and predictability of
weather on subseasonal-to-seasonal
timescales

Sem Vijverberg

VRIJE UNIVERSITEIT

Hidden amongst Chaos

Dynamics and predictability of weather on subseasonal-to-seasonal timescales

Hidden amongst Chaos: Dynamics and predictability of weather on subseasonal-to-seasonal timescales

PhD thesis, Vrije Universiteit Amsterdam, The Netherlands

ISBN: 978-94-6483-158-0

Artwork: Sander van der Valk

Printed by: Ridderprint

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE). The research for the thesis was carried out at the Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, made possible with financial support from the Dutch Research Council (NWO), project number 016.Vidi.171.011.

VRIJE UNIVERSITEIT

Hidden amongst Chaos

Dynamics and predictability of weather on subseasonal-to-seasonal timescales

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 vrijdag 26 mei 2023 om 11.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Sebastiaan Pieter Vijverberg

geboren te Naaldwijk

promotor:

prof.dr. D. Counou

copromotor:

dr. M.J. Schmeits

promotiecommissie:

prof.dr. D.I.V. Domeisen

dr. M.G. Donat

prof.dr.ir. B.J.J.M. van den Hurk

prof.dr. J. Runge

dr. A. Weisheimer

Contents

Summary	v
Summary	vii
1 Introduction	1
1.1 Predicting the weather	1
1.2 Research challenges	6
1.2.1 Dynamical changes due to heterogeneous warming	6
1.2.2 Lack of forecast skill on subseasonal-to-seasonal timescales	6
1.2.3 Understanding the main source of predictability	8
1.2.4 Challenges that hamper the uptake of S2S forecasting	8
1.3 Research questions	9
2 Projections and Hazards of Future Extreme Heat - Dynamical Mechanisms	13
2.1 Large-scale controls	14
2.2 Observed and projected dynamical changes	16
3 Sub-seasonal statistical forecasts of eastern United States hot temperature events	19
3.1 Introduction	20
3.2 Method	22
3.2.1 Data	22
3.2.2 Defining the target variable and the Pacific Extreme Pattern	22
Hot day events	22
PEP	23
3.2.3 Composite-based Precursor Pattern Algorithm	23
3.2.4 Forecasting method	24
3.2.5 Forecast Validation	26
3.3 Results	27
3.3.1 Spatial clustering and hot days in ERA-5 and EC-Earth	27
3.3.2 Comparison between CPPA, PEP and climate indices	28
3.3.3 Using multiple validation metrics	29
3.3.4 Temporal aggregation to improve signal-to-noise ratio	30
3.3.5 Using a window probability and spatial aggregation to improve event forecasts	32
3.3.6 Sub-seasonal forecasts of moderate heatwaves using both SST and soil moisture	34
3.4 Discussion	35

3.4.1	Using multiple validation metrics	35
3.4.2	Improving statistical forecasts for events	36
3.4.3	Physical interpretation of the CPPA pattern	37
3.5	Conclusions	38
3.6	Acknowledgements	39
4	The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability	41
4.1	Introduction	42
4.2	Method	44
4.2.1	Data	44
4.2.2	Clustering North American temperature events	44
4.2.3	Link between temperature, circulation, and sea surface temperature	45
4.2.4	Causal Effect Network using PCMCI	46
4.2.5	Partial correlation maps	47
4.2.6	Forecasting	47
4.3	Results	48
4.3.1	Explaining the long-lead causal link	52
4.3.2	Temperature predictability	52
4.3.3	Window of opportunity by winter-to-spring PDO state	53
4.4	Discussion	56
4.4.1	Data Availability Statement	58
5	Skillful US Soy-yield forecasts at pre-sowing lead-times	61
5.1	Introduction	62
5.2	Method	64
5.2.1	Data	64
5.2.2	Clustering of soybean production regions and derivation of spatial mean soy-yield	65
5.2.3	Cross-validation and pre-processing	66
5.2.4	Response-guided dimensionality reduction (RGDR)	67
5.2.5	Causal Inference as precursor selection method	68
5.2.6	Baseline dimensionality reduction approach	69
5.2.7	Statistical models and hyperparameter tuning	70
5.2.8	Skill metrics	71
5.3	Results	71
5.3.1	Conditionally Dependent (C.D.) precursor regions	71
5.3.2	Leave-three-out hindcast skill	72
5.3.3	One-step-ahead forecast skill	76
5.3.4	Synthesis of results	76
5.3.5	State-level forecast skill	76
5.4	Discussion	79
5.4.1	Physical interpretation	81
5.5	Conclusion	81
5.5.1	Data Availability Statement	82
6	Next steps for S2S forecasting: Scientific and Societal valorization	85

6.1	Introduction	86
6.2	Opportunities, pitfalls, and challenges of S2S weather forecasting using statistical models	86
6.2.1	Opportunities of machine learning	86
	Reducing the noise	86
	Improving the signal	87
	Impact-based forecasting	88
6.2.2	Pitfalls of machine learning	89
	Modelling & Verification Pitfalls (MVP)	89
	Model understanding	89
	Reproducibility and transparency	91
6.2.3	Limitations and challenges	91
6.3	Way forward	92
6.3.1	Scientific valorization: High-level community software	92
6.3.2	Societal valorization: Spin-off company	94
	Energy Transition & Climate Extremes	95
	Products	97
6.4	Conclusion	98
7	Synthesis & Outlook	99
7.1	Main findings	100
7.2	Directions for future research	103
7.2.1	Systematic evaluation of ocean-atmosphere interaction	103
7.2.2	Causality and machine learning methods for future impact and risk projections	104
7.2.3	Seasonal-to-Decadal predictions	105
7.3	Concluding remarks	106
	Appendices	107
3.A	Spatial clustering of heat extremes	109
3.B	Double cross-validation	109
3.C	CPPA vs. linear point correlation map approach	111
3.D	Soil moisture timeseries	113
3.E	Climate indices	113
3.F	Supporting information forecasts	115
4.A	clustering simultaneous warm temperature periods	121
4.B	Defining the Rossy wave timeseries	121
4.C	Comparison to atmospheric modes of variability	121
4.D	SST-RW coupling in winter and spring	124
4.E	Seasonal dependence of temperature predictability	126
4.F	Window of Predictability emerging from PDO state	128
5.A	Pre-processing of crop yield data and cross-validation	131
5.B	Response-guided dimensionality reduction and causal precursor selection	131
5.C	Forecast verification	135
	List of Publications	143
	References	167

Samenvatting

Het weer heeft niet alleen van invloed op ons dagelijks leven, maar het vormt samenlevingen op talloze manieren: van de infrastructuur tot culturele gewoontes zoals siësta's. De snelheid waarmee het klimaatsysteem verandert als gevolg van antropogene emissies legt de kwetsbaarheid van samenlevingen voor weersextremen pijnlijk bloot. Nu het weer onze samenleving op ongekende wijze beïnvloedt, wordt het belang van vroegtijdige waarschuwingen steeds duidelijker.

Ons vermogen om weersvoorspellingen op middellange termijn te maken (2 tot 10 dagen vooruit) is de afgelopen decennia gestaag toegenomen. Wetenschappers zijn ook in staat geweest om decennia van tevoren betrouwbare projecties te maken voor wat zich momenteel op wereldschaal afspeelt. Maar ons vermogen om weken tot maanden vooruit te voorspellen blijft achter, vandaar dat deze subseizoens-tot-seizoens (S2S) tijdschalen 'de voorspelbaarheidswoestijn' wordt genoemd.

In dit proefschrift heb ik datagedreven methoden gebruikt om zowel de S2S weersvoorspellingen als het fysisch begrip ervan te verbeteren. Het belangrijkste aandachtsgebied in dit proefschrift is Noord-Amerika, waarvoor al relatief veel kennis beschikbaar was aan het begin van dit promotieonderzoek, wat dit gebied een goede test bed maakte voor datagedreven methoden. Ik onderzocht (1) de voorspelbaarheid van extreme temperaturen in het oosten van de Verenigde Staten (VS), (2) de oceaan-atmosfeer interactie - wat een belangrijk proces is voor de voorspelbaarheid in het oosten van de VS, (3) de mogelijkheid om oogstmislukkingen in het oosten van de VS te voorspellen, en tot slot (4) identificeer ik valkuilen, kansen en een visie voor het onderzoeken van dynamische processen en voorspelbaarheid met behulp van datagedreven methoden.

(1) Eerder werk toonde aan dat hittegolven in het oosten van de Verenigde Staten (VS) vaak werden voorafgegaan door een specifiek patroon in het oppervlakte zeewater temperatuur (Sea Surface Temperature, SST) in de noordelijke Stille Oceaan. In ons onderzoek introduceerden we een nieuw algoritme dat de extractie van het SST-patroon automatiseerde, waardoor de voorspelling voor hittegolven verbeterde. Helaas, blijft het voorspellen van weersextremen op erg lastig. Daarom onderzochten we ook de verbetering van de voorspelling wanneer we (1) de extremitet van de hittegolf verlagen en (2) de ruimtelijke aggregatie verhogen. Wij concludeerde dat compromissen moeten worden gesloten op het gebied van extremitet en ruimtelijke aggregatie om voorspellingen voor extremen voldoende betrouwbaar te maken op deze tijdschalen. Een andere optie is het voorspellen van de kans dat de hittegolf zich voordoet binnen een breder tijdsvenster van bijvoorbeeld 10 dagen. Dit verbetert de betrouwbaarheid van de voorspelling en geeft nog steeds informatie over het risico van de toekomstige extreme gebeurtenis, hoewel de exacte timing binnen het toekomstige tijdvenster onbekend is. Door deze flexibiliteit toe te laten, kunnen we de kans op gematigde hittegolven in het oosten van de VS tot 60 dagen vooruit

op bekwame wijze voorspellen.

(2) Puur statistische modellen kunnen soms verkeerde verbanden leren, wat het vertrouwen in dit soort modellen belemmert. Natuurkundig inzicht is een essentieel component dat nodig is om de betrouwbaarheid van de modellen te kunnen beoordelen. Causale ontdekkingsmethoden kunnen een handje helpen door natuurkundige relaties te leren uit data, waardoor ze in hetzelfde domein opereren als 'machine learning'. Causale ontdekkingsalgoritmen kunnen statische modellen betrouwbaarder maken door de modellen te voeden met relaties die met hogere zekerheid echte fysieke verbanden weergeven.

Een hoefijzervormige oppervlakte zeewater temperatuur patroon in de noordelijke Stille Oceaan (oftewel 'horseshoe SST' patroon) speelt een belangrijke rol voor de zomertemperatuur in de VS. Dit effect wordt gemodereerd via een stationaire atmosferische (Rossby-)golf, maar de precieze fysica die leidt tot voorspelbaarheid op zo'n lange termijn (d.w.z. ver van tevoren) is nog onduidelijk. Met behulp van een causaal ontdekkingsalgoritme, onderzocht ik hoe de atmosferische Rossby-golf interacteert met de Stille Oceaan. Hieruit blijkt dat de voorspelbaarheid van de zomertemperatuur ontstaat vanuit laagfrequente variabiliteit die (deels) in de winter wordt geïnitieerd en zich gedurende het voorjaar en de vroege zomer ontwikkelt. Wij laten zien dat de laagfrequente variabiliteit in de Stille Oceaan vooral wordt aangedreven door de atmosfeer-naar-oceaan forcering en de positieve feedback interactie in de winter en het voorjaar. In de zomer is er echter een forcering van de oceaan naar de atmosfeer. De aanwezigheid van een sterk SST patroon, zorgt voor frequentere en persistentere atmosferische golven, die gepaard gaan met een hogedrukgebied boven het oosten van de VS, met bijbehorende hogere temperaturen en minder neerslag. Volgend uit deze hypothese, hebben wij bevestigd dat de voorspelbaarheid aanzienlijk toeneemt in jaren met een sterk horseshoe SST patroon.

(3) Het belang van het 'horseshoe SST patroon' van de winter tot in het voorjaar suggereert dat warm en droog weer in het midden en oosten van de VS op erg lange termijn voorspelbaar zou kunnen zijn. Zo ver vooruit kunnen voorspellen zou nieuwe mogelijkheden kunnen bieden voor de landbouwsector, aangezien belangrijke beslissingen m.b.t. plantenbeheer of financiële beslissingen moeten worden genomen vóór het plantseizoen. Ik heb de dimensionaliteit van oceaan data gereduceerd en vervolgens een selectiestap gebaseerd op causale inferentie gedaan. Dit had als doel om het statistische model te kunnen trainen met betrouwbaardere input data. Het statische model was in staat om slechte soja oogst jaren al op 1 februari te voorspellen, wat 3 maanden voor het inzaaien is. Dit stelt landbouwers in staat een geïnformeerde beslissing te nemen over bijvoorbeeld de zaaidichtheid, het wel of niet zaaien van droogtegevoelige landbouwgrond of de aankoop van meer droogte-bestendige zaden.

(4) De resultaten in dit proefschrift suggereren dat we de voorspelbaarheid en potentiële waarde van lange-termijn weersvoorspellingen wellicht hebben onderschat. Hoewel deze resultaten veelbelovend zijn, zijn er belangrijke valkuilen die we moeten vermijden. Er is bijvoorbeeld nog onduidelijkheid over de 'best practices' van datagedreven voorspellingen. Ik denk dat speciale open-source software en een actieve gemeenschap hierbij kunnen helpen. Bovendien hebben we - voordat samenlevingen kunnen profiteren van deze innovaties - een organisatie/bedrijf nodig dat dit soort voorspellingen maakt, operationaliseert, de software onderhoudt en de IT-infrastructuur ondersteunt. Beide aspecten worden besproken in hoofdstuk 6.

Summary

Weather is not only affecting our daily lives, but it shapes societies in a myriad of ways: from its infrastructure to cultural habits like siestas. The speed at which the climate system is changing due to anthropogenic emissions is painfully exposing vulnerabilities of societies to weather extremes. As weather is impacting our society in unprecedented ways, the importance of early warnings is becoming more evident.

Our ability for medium-range weather forecasting (~ 2 to ~ 10 days in advance) has steadily increased over the last decades. Likewise, scientists were able to make reliable projections for what is currently unfolding on a global scale decades in advance. However, our ability to forecast weeks up to months ahead is lagging, which is why these subseasonal-to-seasonal (S2S) timescales are referred to as 'the predictability desert'.

In this thesis, I have explored the use of data-driven methods to improve both S2S forecast skill and physical understanding. The main focus area in this thesis is North-America, for which substantial knowledge was available at the start of this Phd research, thereby providing a good testbed for data-driven methods. I investigate (1) the predictability of eastern United States (US) temperature extremes, (2) the ocean-atmosphere interaction that drives eastern US predictability, (3) the ability to predict harvest failure in the eastern US, and finally, (4) I identify pitfalls, opportunities, and a vision on a way forward for exploring S2S dynamics and predictability using data-driven methods.

(1) Previous work showed that eastern United States (US) heatwaves were often preceded by a specific north-Pacific sea surface temperature (SST) pattern. In our research, I introduced a new algorithm that automated the extraction of a robustly preceding SST pattern, which improved forecast skill for the target heatwave events. Predicting extremes is notoriously difficult. I therefore explore the increase in forecast skill when lowering the extremity of the heatwave definition and increasing the spatial aggregation. We concluded that, at longer lead-times, compromises have to be made in terms of extremity and spatial aggregation in order to make forecasts for extremes sufficiently reliable. Another option is to predict the probability of the event occurrence within a wider time window of e.g., 10 days. This improves forecast skill and it still informs on the risk of a future extreme event, although the exact timing within the future time window is unknown. Using this so-called 'window probability', we could skillfully predict moderate hot events up to 60-days ahead.

(2) Purely statistical machine learning models can sometimes learn implausible relationships, which hampers the trust in the model. Hence, physical understanding is a vital component that is needed to build more trustworthy forecast models. Causal discovery methods can learn physical relationships purely from data, thereby operating in the same paradigm as machine learning. Causal discovery algorithms can guide machine learning models to focus on relationships that are more likely to represent the real physical linkages. For example, we know that the effect of the north-Pacific (horseshoe-shaped) SST pattern

on eastern US summer temperature is moderated via a stationary atmospheric (Rossby) wave, yet the exact physics that leads to predictability at such long lead-times (i.e., far in advance) is still unclear. I introduce a framework based on a causal discovery algorithm to study how the atmospheric Rossby wave interacts with its underlying ocean. I find that the summer temperature predictability originates from low-frequency variability in the north Pacific that is (partly) initiated in winter and develops throughout spring and early summer. We show that the low-frequency Pacific variability is mainly driven by the atmosphere-to-ocean forcing and two-way feedbacks in winter and spring. However, in summer, there is an upward forcing from the ocean to the atmosphere. The presence of a strong horseshoe SST pattern that emerges from spring is driving more frequent and persistent atmospheric waves, which are associated with a high-pressure system, higher temperatures, and reduced rainfall in the eastern US. Since the horseshoe SST pattern acts as a boundary forcing to the atmosphere in summer, I show that the predictability is substantially enhanced during years with a pronounced horseshoe pattern.

(3) The importance of the winter-to-spring horseshoe SST pattern suggests that mid-to-eastern US hot and dry weather could be predictable at long lead-times. Predictability at these lead-times could open-up new opportunities for the agricultural sector, as important plant management or financial decisions need to be made prior to the planting season. I implemented a response-guided dimensionality reduction method in combination with a causal inference-based selection step to extract reliable input features from observational SST and soil moisture gridded datasets. The forecast model was able to predict poor soybean harvest years already on the 1st of February, which is \sim months prior to sowing. This allows farmers to make a more informed decision on e.g., sowing density, avoiding to plant in drought-prone lands, or buying drought-resistant seeds.

(4) The results in this thesis suggest that we may have been underestimating the predictability and potential value of S2S forecasting. Although these results look promising, there are important pitfalls we need to avoid. For example, there is still ambiguity on best practices for data-driven forecasting. I believe dedicated open-source software and an active data-driven S2S forecasting community could help with this issue. Furthermore, before societies can benefit from these innovations, we need an organization/company that builds forecasts, operationalizes these forecasts, maintains the software, and supports the IT infrastructure. Both these aspects are discussed in Chapter 6.

Introduction

1.1 Predicting the weather

Climate change is straining societies' resilience against weather extremes. What was a hot day in the past, can now become an extreme heatwave that can harm society and ecological systems in unprecedented ways. The record shattering 2021 western North American heatwave is a prime example. Its occurrence would not have been possible without global warming (Philip et al., 2021). It promoted large wildfires with 1 million hectares being burned as of mid-July (Guardian, 2021a; Wikipedia, 2021), heat-related deaths are estimated at 1400, economic impact is estimated at 8 billion euro, and the heatwave has killed roughly 1 billion marine animals (Guardian, 2021b). In line with the warming trend due to anthropogenic emissions, the summer of 2022 exposed Europe to extreme heat and drought conditions. The hot and dry conditions within Europe surpassed infrastructure-boundaries for numerous sectors within Europe. For example, extremely warm river temperatures forced power plants to operate at reduced capacity. Extremely low water levels reduced the efficiency of inland shipping and hydropower plants. In some regions of the Global South which are already exposed to harsh climate conditions, climate change is even triggering migration (Kaczan and Orgill-Meyer, 2020). The strain climate change is exerting on societal and ecological systems comes on top of the increasing demand for resources due to population growth and increasing welfare.

As climate change is increasingly exposing our vulnerability to weather, the value of weather predictions is unmistakable. They save lives by enabling anticipatory action (WMO, 2021) and prevent huge economic as well as ecological damage (Golding, 2022). The majority of the UN members that signed the Paris Agreement identify the development of early warning systems as "top priority" (Golding, 2022). Advances in physical understanding, the higher number of observations, and more computational power has led to a substantial increase in forecast skill of numerical weather predictions. The term 'skill' refers to the

ability of a forecast to reliably predict what is observed. A prediction is only 'skillful' when it can perform significantly better than a proper benchmark forecast. A proper benchmark forecast depends on the context, generally it should be intuitive and simple (e.g., assuming persistence or climatology). The horizon up to which forecasts are skillful have steadily increased at about one additional day each decade, meaning that our 6-day forecast in 2020 is as good as the 5-day forecast in 2010 (Bauer et al., 2015).

Medium-range weather forecasts focuses on predicting weather ~ 2 to ~ 14 days in advance. Such forecasts mainly rely on the memory of the atmospheric state. The further we are trying to predict into the future, the poorer the forecast quality will be due to chaotic atmospheric interactions. This behavior exists due to the growth of random small-scale disturbances, effecting larger and larger - finally - synoptic-scale weather systems. This phenomenon is colloquially known as 'the butterfly effect' (Lorenz, 1969). Due to the butterfly effect, the current state is largely independent from the state ~ 14 days ago.

However, for some early-action interventions, the time horizon of 14 days is too short to act upon: for example, whether a farmer prior to the planting season decides to buy the more drought-resistant - but more expensive - seeds (Crane et al., 2010); or whether a non-governmental organization (NGO) decides to allocate resources *prior* to the occurrence of seasonal drought conditions, thereby attempting to avoid famine instead of responding to it (Coughlan de Perez, 2018); or whether gas reserves need to be increased due to an upcoming cold winter. Going beyond the forecast horizon of ~ 14 days, we enter the domain of subseasonal-to-seasonal (S2S) weather predictions.

introduction into subseasonal-to-seasonal (S2S) dynamics

Subseasonal-to-Seasonal dynamics and predictability

The physics of weather predictability on subseasonal-to-seasonal (S2S) timescales is fundamentally different from weather predictions up to two weeks. Weather predictions up to two weeks can be skillful by solving the 'initial value problem': using the equations of motion and thermodynamics to determine how the atmosphere is likely to evolve. However, beyond two weeks, the predictable part from the atmospheric dynamics alone starts to become very small. Predictability of weather on subseasonal-to-seasonal timescales, i.e., two weeks up to months ahead, stems from slowly varying components of the climate system that interact with the atmosphere. With 'slow', we refer to the persistence of the dynamics. Slow components have a much higher persistence compared to the more chaotic dynamics of atmospheric circulation. A slow component can be seen as a slowly varying boundary condition acting on and interacting with the atmosphere (WMO, 2020). The memory (i.e., persistence) of these slow components and the interaction with the atmosphere allows us to predict (the statistics of) weather weeks up to months ahead. Such components are the ocean (Vijverberg and Coumou, 2022; Kushnir et al., 2002; Rodwell and Folland, 2002; Li et al., 2022), monsoon systems (Di Capua et al., 2019a; Di Capua et al., 2019b; Ding and Wang, 2005), the stratosphere (Cai et al., 2016; Thompson et al., 2002) and the Earth's surface (Seneviratne et al., 2010; Dirmeyer, 2003; Teng et al., 2019; Chevallier et al., 2019; Walsh, 2005).

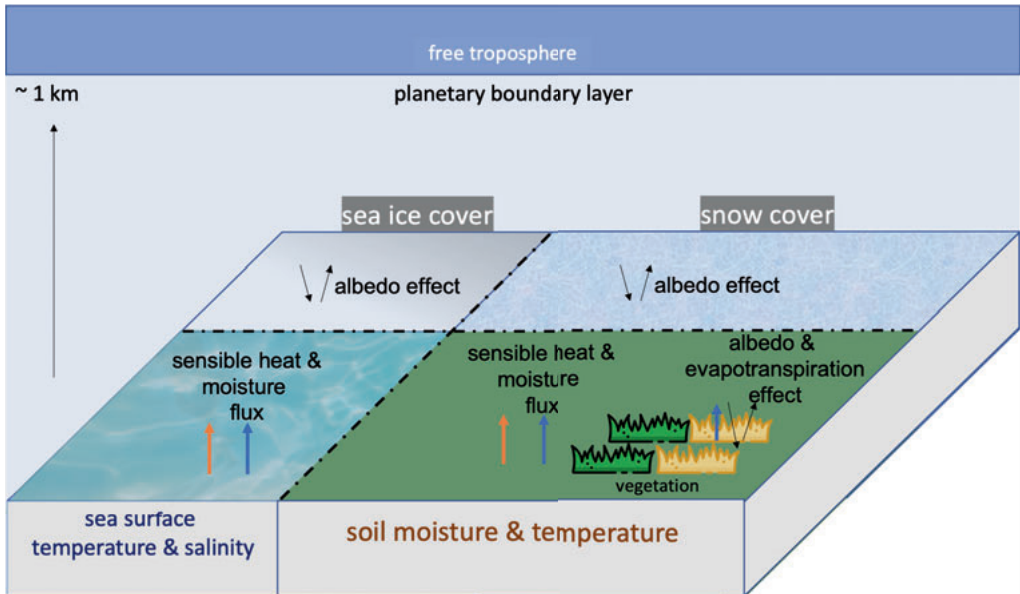


Figure 1.1: Sources of predictability that have a direct impact on the atmospheric column aloft (planetary boundary layer) and the associated mechanisms/fluxes.

Local influence of slow components

The earth's surface has a strong influence on the well-mixed lowest layer of the atmospheric column, known as the planetary boundary layer. The memory of the earth's surface can be a source of predictability by having a strong effect on the weather conditions in the planetary boundary layer. The most well known sources of predictability are [soil moisture, soil temperature, sea ice cover, snow cover, sea surface temperature and sea surface salinity]. These are shown in Figure 1.1 together with their associated influence on the thermal heat/moisture budget of the planetary boundary layer.

Remote influence of slow components

Slow components can also have a non-local effect by modulating the distribution of heat (i.e., energy) of the free troposphere (the atmospheric column above the planetary boundary layer and below the stratosphere). Gradients in heat are 'the fuel' that drive the atmospheric circulation. Redistributions of heat can thus change the atmospheric circulation (geostrophic and thermal wind balance) and trigger atmospheric waves. Particularly atmospheric Rossby waves play a crucial role in the communication of synoptic circulation disturbances across large distances by creating a train of high- and low-pressure systems (Hoskins and Karoly, 1981; Hoskins and Ambrizzi, 1993).

The El Niño Southern Oscillation (ENSO) is the most well-known slow component that induces non-local effects. ENSO variability arises due to interactions between the tropical Pacific Ocean dynamics and the atmosphere aloft. During the positive ENSO state (el Niño), the sea surface temperatures (SSTs) over the tropical east Pacific are substantially warmer compared to the negative ENSO state (la Niña). In the tropics, warm SSTs

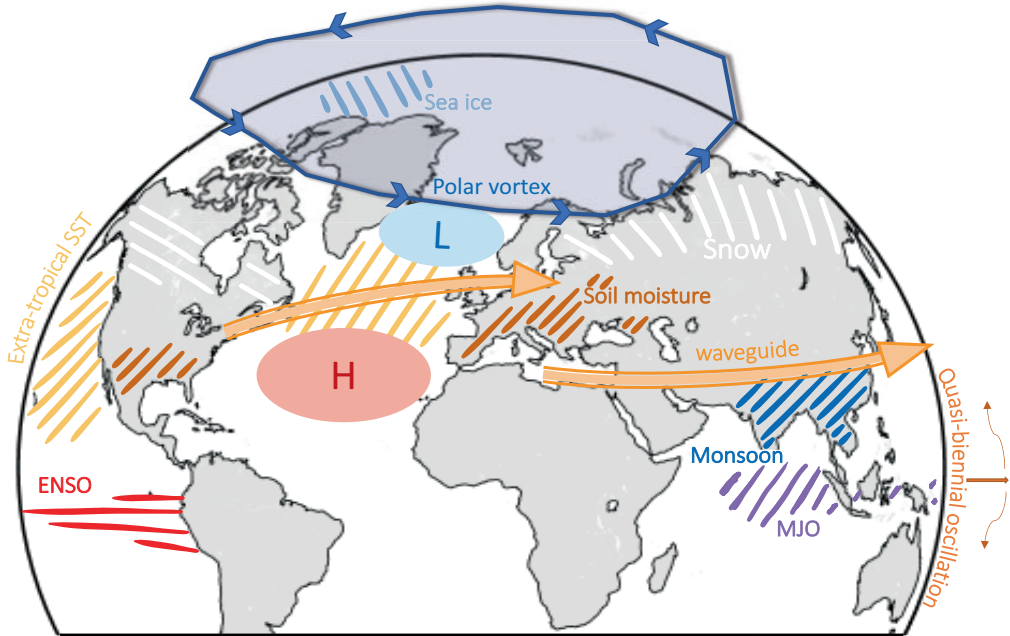


Figure 1.2: Schematic (non-exhaustive) overview of important sources of predictability and mechanisms that have an indirect/remote impact on the atmosphere and thereby influence remote surface weather conditions, often called teleconnections.

drive deep convection resulting in enhanced latent heat release at high altitudes. This perturbation of heat triggers a Rossby wave response, affecting circulation over many regions across the globe (Qin and Robinson, 1993). The Rossby waves response to a perturbation is relatively fast, i.e., clearly seen within a day and is fully developed after ~ 2 weeks (Schubert et al., 2011). Importantly, the persistence of the source is what makes these Rossby waves more prevalent not only in the next days, but also weeks up to months ahead. Such circulation changes can lead to a regional increase in the prevalence of high- and low-pressure systems. In turn, such systems can affect the frequency of heatwaves, droughts, floods or cold spells. Hence, such an atmospheric perturbation is the starting point of how slow components affect large-scale circulation, and thereby the probability of extremes (Schubert et al., 2011).

Complexity of interactions on different timescales

Figure 1.2 illustrates various important low-frequency processes of the climate system that can have a remote (i.e., non-local) impact on surface weather conditions. As the climate system is a fully connected system, these components affect each other in a myriad of ways. Consequently, stating that process X causes Y is always a simplification of the true underlying dynamics, but we can use this simplification for communicating our understanding of the dominant physical links.

To illustrate the complexity, I start with describing a low-frequency process, the North Atlantic Oscillation (NAO), which exists in an atmosphere-only model (not coupled to 'slow components' like the ocean or the land surface) (Franzke, 2002). The NAO, schematically

depicted by the red (H) and blue (L) oval in Figure 1.2, is characterized by bimodal behavior: the jet stream¹ (orange arrows in Figure 1.2) tends to persist longer at either more southern *or* more northern latitudes over the Atlantic domain. This low-frequency NAO variability exists due to the feedback between small scale eddies (i.e., storms), the zonal mean flow (i.e., the jet stream) and planetary-scale Rossby waves (Robert et al., 2017; Barnes and Hartmann, 2011; Gerber and Vallis, 2007; Rivière et al., 2016). In an atmosphere-only setting, the e-folding time (a metric for persistence) of the NAO is ~ 7 days (Franzke, 2002). In reality, numerous slow components influence - and thereby further enhance - the persistence of NAO variability: for example, (1) via ocean-atmosphere interactions (Barnes and Hartmann, 2010), (2) via ENSO affecting the meandering of the northern hemisphere jet stream by triggering Rossby waves (Wolf et al., 2020), (3) via Rossby waves triggered by the Madden-Julian Oscillation (MJO), or (4) via winter stratospheric (polar vortex) variability (Scaife et al., 2016). There are multiple modes of atmospheric variability similar to the NAO (which are not shown in Figure 1.2), such as the Pacific North American (PNA) pattern, Northern Annular Mode (NAM) (Garfinkel et al., 2020), and others. These modes are all - with varying degrees of strength - affected by (interaction with) slow components. Due to the persistence of these slow components, we can sometimes predict the probability of anomalous surface weather conditions weeks up to months in advance. As the strength of these physical links varies for different regions and seasons, so does our ability to predict the weather vary for different regions and seasons.

===== end of introduction into subseasonal-to-seasonal (S2S) dynamics =====

The value of S2S forecasting has so far remained limited, despite its clear potential (White et al., 2021). The main bottleneck reported by various (potential) end-users is the lack of forecast skill (Coughlan de Perez, 2018; Vigo et al., 2019; Kusunose and Mahmood, 2016). Traditionally, numerical weather models are coupled to slow components such as the ocean and land, and are then used to enable predictions on S2S timescales. The skill of these ensemble predictions can be improved with bias-correction methods, yet the skill of bias-corrected numerical models on S2S timescales is limited (Johnson et al., 2019; Kirtman et al., 2014; van Straaten et al., 2020). Luckily, knowledge on the dynamics on S2S timescales and predicting weather weeks up to seasons ahead is growing (Vitart and Robertson, 2018b). Particularly, the use of advanced statistical techniques such as machine learning and causal discovery show promising results. These techniques can learn relationships from observations, which can be used for predictability, as well as improved physical understanding. Increasingly skillful predictions on S2S timescales could pave the way to solve important societal problems. Anticipating weather-related risk supports better humanitarian action plans, agricultural management, (renewable) energy resource management, industry/insurance risk portfolios and public health management. In section 1.2, I specify some major challenges related to S2S dynamics and predictability in more detail.

¹The jet stream is a band of fast narrow eastward winds in the upper troposphere, which is mainly zonally oriented.

1.2 Research challenges

For this thesis, I have identified 4 research challenges that involve our physical understanding of climate dynamics and predictability of weather on subseasonal-to-seasonal timescales. A good understanding of climate dynamics on S2S timescales is important for regional climate projections, as well as for S2S predictability. The first chapter reviews the literature on dynamical changes and implications the regional climate projections. The remainder of the thesis, however, focuses on S2S predictability and the underlying physical mechanisms.

1.2.1 Dynamical changes due to heterogeneous warming

Understanding S2S dynamics, i.e., how these slow components interact with the atmosphere, is not only important for predictability, but also for future climate change projections (Raymond et al., 2019; Shepherd, 2014). On a regional scale, the decadal varying slow components increase the ensemble spread and thereby add to the uncertainty of regional climate projections (Zaplotnik et al., 2022; Watanabe et al., 2020; O’Reilly et al., 2021). In some regions across the globe, climate extremes are intensifying beyond what would be expected from the thermodynamic warming effect induced by greenhouse gases (Horton et al., 2016; Mann et al., 2017; Petoukhov et al., 2013; Rousi et al., 2022). This could be explained by a change in circulation patterns that is favoring the occurrence of a certain extreme events. Even relatively small changes in the atmospheric circulation can have important consequences for the frequency and intensity of extreme events (Coumou et al., 2015; Lehmann and Coumou, 2015; Luo et al., 2022). The process of global warming is - on a global scale - well understood and faithfully simulated by the climate models (IPCC, 2021). On a more regional scale, however, the future climate projections are much more uncertain. A major component of this uncertainty is associated with the role of atmospheric circulation (Shepherd, 2014). Dynamical changes could arise due to heterogeneous warming of the climate system. On a global scale, heterogenous warming occurs due to the faster rate of warming in the lower troposphere of the Arctic region (Coumou et al., 2018) and the faster warming in the upper-part of the tropical troposphere (Yuval and Kaspi, 2020). On a more regional scale, changes in the land-sea thermal contrasts, static stability² and moisture content can drive dynamical changes (Baker et al., 2019; Shaw and Voigt, 2015; Donges et al., 2016). To improve regional climate projections, we must better understand the dynamical changes that could be induced by anthropogenic global warming (Shaw and Voigt, 2015; Coumou et al., 2017; Lau and Kim, 2015; Baker et al., 2019).

1.2.2 Lack of forecast skill on subseasonal-to-seasonal timescales

State-of-the-art dynamical forecast systems have little or no forecast skill on S2S timescales. The question is whether this lack of skill is an actual characteristic of the climate system, or whether dynamical models are missing some important processes (Scaife and Smith,

²Static stability refers to ability of atmospheric flow to become turbulent due to the effects of buoyancy that can lead to vertical displacement, it is proportional to the ease or difficulty at which air parcels can become convective.

2018). On S2S timescales, seasonal dynamical models are incorporating the interaction with the relevant slow components. To simulate the end-of-chain effect on our surface weather, the models need to correctly simulate ocean, land and cryosphere dynamics including the associated boundary fluxes, (eddy) momentum transport, and convective systems. So far, this remains a very challenging task, especially for extreme events. For example, the correct magnitude of Russian 2010 heatwave was predicted 'only' 8 days before the onset (Vitart and Robertson, 2018a) and the 2021 western North American heatwave was correctly predicted 5 days prior to its peak temperature (Lin et al., 2022). On weekly mean timescales, the lead-time increases and dynamical models can succeed in predicting weather anomalies (such as upper terciles) 2-3 weeks in advance (Vitart and Robertson, 2018a; Bloomfield et al., 2021; Lin et al., 2022). On (multi) monthly mean timescales, predictability up to a few months ahead is possible, but only some mid-latitude regions (Kirtman et al., 2014; Johnson et al., 2019). Many regions, such as the eastern US and Europe, show no or very limited skill on these timescales (Kirtman et al., 2014; Johnson et al., 2019). Data-driven studies suggest that the predictable signal might be underestimated in dynamical models (Di Capua et al., 2021; Vijverberg et al., 2020; Merryfield et al., 2020; National Academies of Sciences, 2016; Scaife and Smith, 2018). Next to ongoing improvements of dynamical seasonal forecast models (Merryfield et al., 2020), ML approaches are thus widely considered to have a big potential to add value in this domain.

A prominent study reported a remarkable skillful forecast for eastern United States (US) hot days at 50 days lead-time (McKinnon et al., 2016). The high skill could be explained by their reliance on observational data and their discovery of a new specific sea surface temperature (SST) pattern in the mid-latitude Pacific. The study suggests the high potential of data-driven approaches to learn from observations, thereby circumventing the low predictable signal that hampers the skill of dynamical models. However, there are several technical pitfalls associated with data-driven forecasting. For example, a clear train-test splitting is important when training a statistical model, and a proper forecast verification requires multiple metrics as different metrics measure different aspects of the forecast quality.

One of the challenges with predicting weather extremes is that the timescale of a typical extreme event is generally (much) shorter (lasting hours up to a couple of days) than the timescales of the sources for S2S predictability. If extremes are happening within relatively short time windows in which the slow (forcing) components are approximately constant, this would mean there is no signal or trigger that can be used to predict exactly when the extreme event will happen. How to translate the signal of slow (forcing) components into a useful prediction for (extreme) weather events remains a challenge. Besides the need to improve our ability to model the slow components (and the interaction with the atmosphere) to boost S2S forecast skill, we also need to learn how to handle and communicate the inherent uncertainty that is part of our climate system.

On a more fundamental level we may never know when we are capable of perfectly extracting the signal from the climate system. The definition of predictability is "an a priori estimate of our ability to make skillful forecasts and quantifies the inherent property of nature; it does not depend on the quality of the forecast system" (WMO, 2020). Thus, predictability is the upper limit of predictive skill that a perfect forecast system can achieve. As we will arguably never know if the forecast system is perfect, the true predictability of

the weather remains an unanswered question.

1.2.3 Understanding the main source of predictability

The ocean is arguably the most important source of low-frequency variability in the climate system and plays a fundamental role in many processes shown in Figure 1.2. Modelling studies have examined the atmospheric response to mid-latitude SST anomalies, and hypothesized about the potential predictability since the late 1960s (Namias, 1959; Namias and Cayan, 1981; Pitcher et al., 1988; Kushnir and Lau, 1992). In 2002, results suggested that the atmospheric response was rather weak and mostly restricted to the local marine boundary layer aloft (Kushnir et al., 2002). However, they did not have access to the high-resolution models of today. Later, it was shown that small-scale eddies play an important role in transferring momentum and vorticity to the upper troposphere (Deser et al., 2007; Ferreira and Frankignoul, 2005). Modern research shows that a large-scale barotropic Rossby wave-like response is possible (Zhou, 2019). However, the ocean-atmosphere interaction appears to be dependent on many factors, such as the persistence of the SST anomaly, location of the SST anomaly and the background atmospheric state. This highlights a challenge with modelling experiments in which often differences in the magnitude of the atmospheric response are found (Zhou, 2019). Hence, besides improved forecast skill, data-driven methods can also help to better understand the physics by learning relationships from observations.

The physical mechanism that leads to the predictability of eastern US temperature (section 1.2.2) is not fully understood. The proposed mechanism was that the mid-latitude sea surface temperature (SST) is responding to strong atmospheric Rossby wave forcing and subsequently maintained/amplified that atmospheric Rossby wave pattern via positive feedbacks. An alternative hypothesis, presented in chapter 3, is that the Rossby wave is forced by low-frequency SST variability. A framework to learn the coupling strength between atmospheric Rossby waves and SST anomalies from observations - and thereby circumvent the potential errors made by dynamical model experiments - has not been developed yet.

1.2.4 Challenges that hamper the uptake of S2S forecasting

As introduced in section 1.2.2, the lack of forecasting skill on S2S timescales has been an outstanding challenge for the last two decades. However, fully capitalizing on the potential for S2S forecasting for society requires a scope beyond predictive skill alone. Numerous studies investigated the end-user value of S2S forecasts, mostly focusing on applications for the agricultural sector (Crane et al., 2010; Kusunose and Mahmood, 2016; Lemos et al., 2002; Templeton et al., 2018), energy sector (Vigo et al., 2019; Troccoli, 2018; Goodess et al., 2019; Bett et al., 2018) and disaster risk management (Coughlan de Perez, 2018; Guimarães Nobre, 2019; Weingärtner and Wilkinson, 2019). These studies highlight two additional core challenges, i.e., trust and stakeholder interaction.

Trust

Trust in the forecast is essential for the uptake of S2S forecasting products. The poor past performance is likely still shaping how end-users perceive the reliability of modern seasonal forecast systems (Kusunose and Mahmood, 2016), although the skill might have

improved substantially in recent years (WMO, 2020). While the advent of machine learning might lead to higher forecast skill, there are threats and weaknesses associated with a statistical modelling approach. For example, dynamical models may derive their credibility by being based on well-known physical laws, while ML models may have learned incorrect relationships that are not physically plausible (Li et al., 2020; McGovern et al., 2019). Another issue is that certain technical pitfalls that lead to an overestimation of the true forecast skill (see e.g. (Schauberger et al., 2020; García-Serrano and Frankignoul, 2014; Vijverberg et al., 2020)). Making false claims, likely by mistake, pose a major threat to the trust that end-users (will) have in statistical techniques.

Stakeholder dialogue

To justify early-action, stakeholders depend on clear and effective communication of a given forecast. An important aspect of clear communication is the translation of the predicted weather statistics to impact (Dorrington et al., 2020). If a farmer is confronted with the prediction of "an 80% probability of above normal August temperatures", it is non-trivial what this implies in terms of e.g., harvest loss. Another challenge lies in the fact that stakeholders need to be able to deal with the inherent uncertainty and interpretation of probabilistic forecasts (Kusunose and Mahmood, 2016; Donkor et al., 2019). How to best communicate the varying degrees of confidence to simplify decision making remains an open question.

Co-development can be an essential component to support effective early-action plans, especially if the stakeholder has specific expectations or needs (White et al., 2021). The need of stakeholders can be rather straight forward, such as the need for reliable surface temperature forecasts, or be very specific. In the latter case, tailored forecasts such as a hydropower plant manager needing the accumulated precipitation forecasts for its specific catchment over the next months are needed. However, the expectations of stakeholders can be more challenging to manage, as they might not be realistic. A stakeholder might have developed an early action plan in response to a rare high impact event which is not predictable on S2S timescales. This could lead to a number of false positive predictions and high costs associated with unnecessary early actions. To enable effective intervention, it could be wise to lower the extremity of the event to increase the reliability of the forecast. In response, the stakeholder will need to revise the early-action plan accordingly, especially if early-action is relatively costly and complex (involving multiple potential actions that could be undertaken). Hence, a two-way dialogue and clear communication with the stakeholder is important.

1.3 Research questions

The goal of this thesis is to improve our ability to predict (the impact of) weather on subseasonal-to-seasonal timescales by gaining a better understanding of the physics and leveraging the power of data-driven techniques. I pursue this goal via a case-study towards predicting eastern US summer heatwaves and associated crop failures. The developed methods, however, are generic and can be applied to other regions, seasons, and impacts. This goal is addressed through the following research questions:

- ❖ How can we improve the sub-seasonal forecast skill for eastern United States (US) heatwaves?








	Climate change	S2S forecast skill	Physical understanding	Valorization
Chapter 2: Circulation changes				
Chapter 3: Predicting extremes				
Chapter 4: US temp. predictability				
Chapter 5: Soy yield predictions				
Chapter 6: Vision S2S forecasting				

Figure 1.3: Overview of research challenges and valorization activities covered per chapter in this thesis.

- ❖ What is the role of ocean-atmosphere feedbacks in generating predictability for eastern US summer temperature?
- ❖ Can we predict soy harvest failure in the eastern US with sufficient lead-time to enable farmers to take anticipatory action?

Besides these research questions, my overarching aim is to bring those innovations to society via effective AI-based forecasts. To do so requires overcoming several challenges and in chapter 6 I present ideas on how this could be done:

- ❖ How can we promote new AI innovations for subseasonal-to-seasonal forecasting and how can we bring those to society in an effective way?

These research questions are addressed in chapters 3 to 6 of this thesis. In addition, chapter 2 reviews the observed and projected dynamical changes and their potential future risk for heatwaves, whereas the main body of the thesis will focus on S2S dynamics and predictability. The individual chapters cover the following topics, with keywords in bold text referring to the overview in Figure 1.3.

- In Chapter 2, I discuss how and why **climate change** is leading to heterogeneous warming patterns across the Earth. Heterogeneous warming patterns can lead to **circulation changes**. We discuss the physical mechanisms and confidence in observed and projected circulation changes.

- In Chapter 3, I investigate how to tackle the challenge of **forecasting** eastern US **extreme temperature** events on **S2S timescales** using statistical modelling techniques. I develop a new algorithm to automatically extract a sea surface temperature pattern that is used to predict eastern US heatwaves.
- In Chapter 4, I describe the role of ocean-atmosphere coupling in forcing a Rossby wave, thereby explaining why the eastern US summer temperature is predictable by north-Pacific sea surface temperature, while the western US temperature is not. Additionally, I show how this improved **physical understanding** can be used to increase **S2S forecast skill**.
- In Chapter 5, I apply a causal inference-based feature selection method on an observational sea surface temperature and soil moisture dataset, enabling reliable and **skillful forecasts of poor soy harvest, already 8 months in advance**. Predicting the impact, i.e. probability on reduced end-of-year crop yield, helps with communicating future risk to agricultural stakeholders. The prediction is skillful even prior to sowing, enabling agricultural stakeholders to make better-informed decisions on (costly) anticipatory actions that are no longer possible after sowing. Such impact forecasts are valuable for the process of **valorization**, as it helps with 'transferring knowledge to actors with an industrial and/or societal perspective'.
- In Chapter 6, I focus on scientific and societal **valorization**. In that chapter, I identify (1) opportunities of data-driven S2S forecasting, (2) bottlenecks and pitfalls, and subsequently, (3) I present my vision for the future of data-driven S2S forecasting.

2

Projections and Hazards of Future Extreme Heat - Dynamical Mechanisms

This subchapter is first authored by S.P. Vijverberg and part of the following book:

C. Raymond, D. Coumou, T. Foreman, A. King, K. Kornhuber, C. Lesk, C. Mora, S. Perkins-Kirkpatrick, S. Russo, and S. Vijverberg (2019). “Projections and Hazards of Future Extreme Heat”. In: *The Oxford Handbook of Planning for Climate Change Hazards*. Ed. by W. T. Pfeffer, J. B. Smith, and K. L. Ebi. December. Oxford University Press, pp. 1–43. DOI: [10.1093/oxfordhb/9780190455811.013.59](https://doi.org/10.1093/oxfordhb/9780190455811.013.59)

2.1 Large-scale controls

Evidence is mounting that, partly due to dynamical changes, midlatitude heat waves are intensifying more than what would be expected from only the thermodynamic warming effect induced by greenhouse gases (Horton et al., 2016; Mann et al., 2017; Petoukhov et al., 2013). Recent heat waves in Russia in 2010 and Europe in 2015 and 2017 are exemplary. They were intensified by anomalously persistent dynamics and concomitant land-atmosphere feedbacks (Miralles et al., 2014; Lhotka et al., 2017). In the Text Box we elaborate on how specific midlatitude circulation states favor heat waves. The following text describes why dynamical changes are expected; what changes have been observed; what changes are projected for the future; and how these changes affect heat wave genesis and persistence. We focus on the midlatitudes since extreme heat there is much more driven by large-scale dynamics than is the case in the tropics.

Midlatitude heat waves are substantially more likely to occur during persistent high-pressure systems (anticyclonic circulations) (Alvarez-Castro et al., 2018; Jézéquel et al., 2017; Pfahl, 2014; Horton et al., 2016). Trends in the frequency of persistent anticyclones have sometimes been quantified using methods designed to capture 'blocking circulations' (e.g. Barnes et al., 2014), which refer to slow-moving or stationary portions of the jet stream that result in a persistent anticyclone over a region (Altenhoff et al., 2008), thus favoring heat-wave genesis. However, this approach is more statistically than physically based (Scaife et al., 2010; Nakamura and Huang, 2018). More importantly, blocking circulation is only one type of persistent circulation that favors warm anomalies, particularly in the higher latitudes (Sousa et al., 2018). Therefore, this review focuses on dynamical mechanisms. It is well established that persistent anticyclones originate from a complex synergy between quasi-stationary Rossby waves, jet streams (Duchez et al., 2016; Kennedy et al., 2016) and storm tracks (Lehmann and Coumou, 2015; Woollings, 2010). Jet streams and Rossby waves in particular are strongly affected by large-scale temperature gradients (Molnos et al., 2017) (see Text Box).

Large-scale temperature gradients are changing due to the heterogeneous warming of the atmosphere (Screen et al., 2013; Petrie et al., 2015; Wang et al., 2015; Barnes and Screen, 2015; Oudar et al., 2017), see Fig. 2.3, which provides the first-order reason to expect mid-latitude circulation changes (Horton et al., 2016). Another arises from the projected strengthening of deep convection in the tropics (Lau and Kim, 2015).

Mid-latitude dynamics that favors heat waves

Due to differences in incoming solar radiation, there exists an equator-to-pole temperature gradient. Because the sub-tropical, tropical and polar air masses are quite well separated due to large-scale circulation, this temperature gradient is strongest at the interfaces between these air masses. This high temperature gradient results in strong winds at high altitudes, leading to the formation of the sub-tropical and polar jet stream (Fig. 2.1). Especially at the polar front, the temperature contrast between sub-tropical and polar air is high. Thus the latitudinal position of the polar jet, varying between 30° and 75° , has a large impact on surface weather conditions (e.g. temperature/precipitation), especially when the **position of the jet stream is persistent** (Hoskins and Woollings, 2015; Mahlstein et al., 2012). As schematically shown in fig. 2b, the polar jet stream displays wavy patterns, induced by large scale Rossby waves (due to their large scale also called 'planetary waves'). These can be generated by atmospheric instability (free Rossby waves), or induced by mountains, sea surface temperature patterns, land-ocean temperature differences, and diabatic heating. Since these forcings are generally stationary in space, they lead to the formation of 'quasi-stationary Rossby waves'. These waves can promote **high amplitude poleward excursions of the polar jet stream**, which favors hot extremes (Screen and Simmonds, 2014; Teng et al., 2013). The aforementioned high temperature gradient at the polar front also provides energy for smaller scale Rossby waves, which can lead to the formation of smaller scale rotating circulation (i.e. eddies) (O'Gorman, 2010). Particularly the atmosphere over the Atlantic and Pacific oceans shows strong eddy activity, i.e. storms, therefore these regions are called the storm tracks^a. In the summer season, **weakened storm tracks** carry less moist and cool air from oceans to land, thereby favoring heat build up over land (Lehmann et al., 2014). The persistent high-amplitude jet stream and weakened storm tracks are closely related to heatwave-inducing '**blocking circulation**', yet their exact interaction is debated and proper blocking definitions are lacking (Nakamura and Huang, 2018).

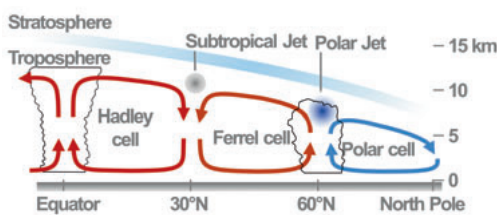


Figure 2.1: Simplified large scale circulation, depicting the sharp temperature contrast between sub-tropical (orange arrows) and polar air (blue arrows), and the deep convection near the equator that results in the descending air over the sub-tropics at $\approx 30^\circ\text{N}$, which suppresses cloud formation (clear sky conditions).

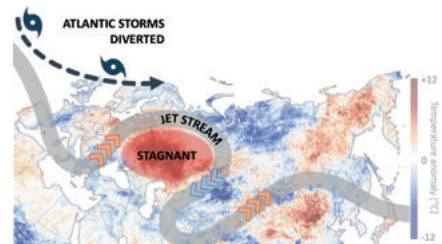


Figure 2.2: The 2010 Russian heat wave and co-occurring Pakistan flood was characterized by a persistent (high amplitude) wave pattern of the jet stream, together with diverted storm tracks (Lau and Kim, 2012). Such a constellation is often referred as 'blocking circulation', since it blocks the west-to-east flow. Image adapted from NASA Earth Observatory.

^aThe atmosphere over the Atlantic and Pacific ocean generates much stronger eddies due to energy release from warm sea surface temperatures on the west side of these basins related to ocean currents (Hoskins and Valdes, 1990).

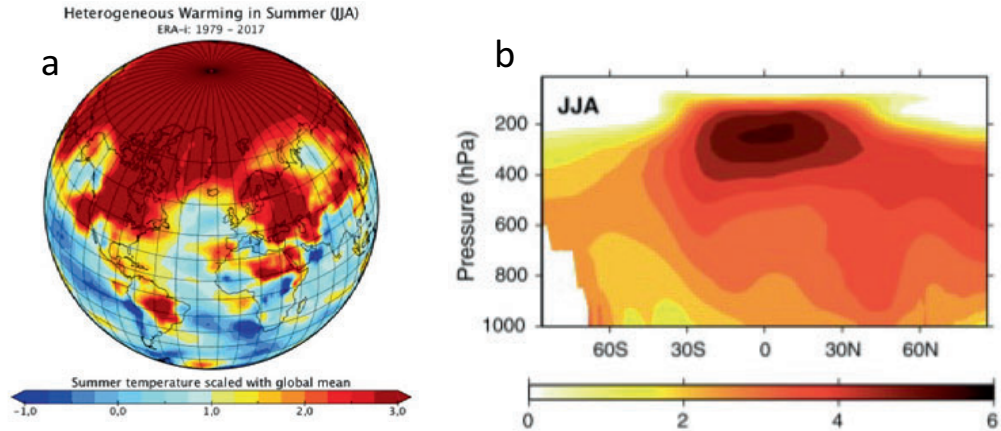


Figure 2.3: (a) Warming rates of regions warming faster (> 1) or slower (< 1) than the global mean temperature at 900hPa in summer, i.e. in figure (a) dark red areas have warmed 3 times as fast compared to global average. (b) From O’Gorman (2010), mean summer (JJA) temperature difference in $^{\circ}\text{C}$ between future (2080-2100) minus present climate (2000-2020) from CMIP3 climate model simulations. The zonal mean at different altitudes (pressure levels) and latitudes shows the Upper Tropospheric Warming signature.

2.2 Observed and projected dynamical changes

There are competing and interacting processes that influence temperature gradients, which complicates the final dynamical outcome (Shaw and Voigt, 2015; Shaw et al., 2016; Peings et al., 2017). Accurately simulating dynamical processes on a large scale (Haarsma et al., 2015; Lau and Kim, 2015), and more so on a regional scale (Sigmond et al., 2007; Lhotka et al., 2017; Plavcová and Kyselý, 2016), is difficult for global climate models. Furthermore, detection of dynamical changes is statistically difficult due to the large internal variability of the climate system. Despite these aspects, robust circulation changes in summer are already detectable in our current climate and are generally expected to become more pronounced in the future.

For example, over the recent (1979-2013) period storm tracks have significantly weakened 8 to 15% in summer (Coumou et al., 2015), meaning that on average, less cool and moist air is transported from ocean to land, thus favoring the buildup of hot and dry conditions (Lehmann and Coumou, 2015) (see text-box). The weakening is attributed to the recent reduction in the equator-to-pole temperature gradient and is also seen in the weakening of the zonal (west-to-east) mean wind, which serves as a proxy for the jet stream strength (Coumou et al., 2015). How this weakening will affect quasi-stationary Rossby waves and persistent blocking is still fairly uncertain.

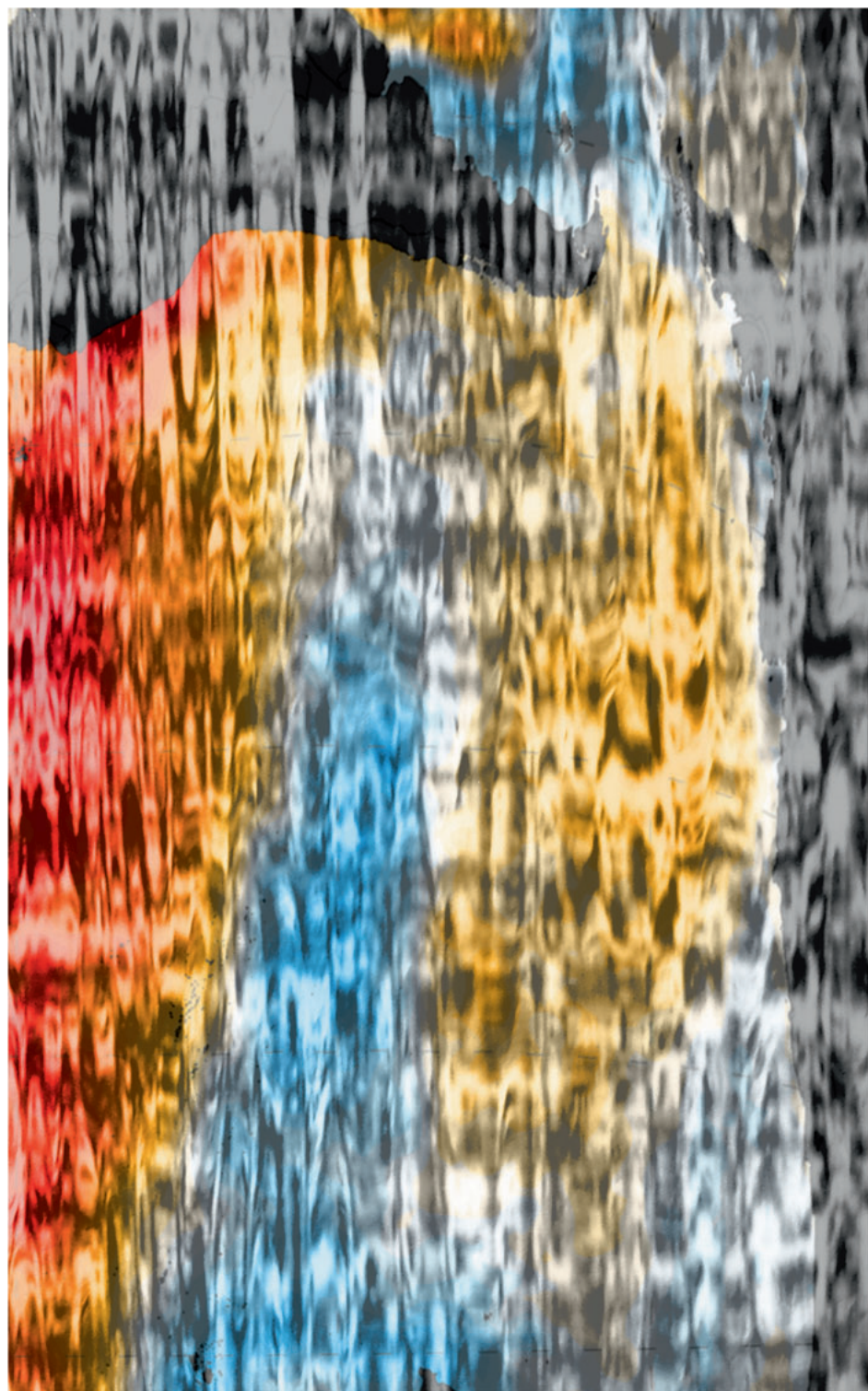
Future (end of the twenty-first century) projections of summer storm tracks also show a continued weakening over the Atlantic and Pacific Ocean (Lehmann et al., 2014; Chang et al., 2012; Simpson et al., 2014; Zappa et al., 2013). The projected northern hemispheric poleward shift in the summertime polar jet is fairly robust over the Atlantic and eastern

U.S. (Simpson et al., 2014; Lorenz and DeWeaver, 2007; Brewer and Mass, 2016), which will cause climate regimes to shift accordingly (see text-box).

The jet stream position can be nudged into forming persistent anticyclones by quasi-stationary Rossby waves (Screen and Simmonds, 2014; Teng et al., 2013) (see Text Box). The frequency and amplitude of some Rossby waves have increased in recent decades (Coumou et al., 2014; Coumou et al., 2017; Lee et al., 2017), although this trend is not robust.

Importantly, the mean jet stream can also interact with forced Rossby waves, thereby creating coherent spatial wave patterns around the entire hemisphere, inducing alternating patterns of persistent high- and low-pressure anomalies called circumglobal wavetrains (Hoskins and Ambrizzi, 1993; Branstator, 2002; Branstator and Teng, 2017). Quasi-resonant amplification [QRA] can be interpreted as a dynamical mechanism that promotes 'extreme' circumglobal wavetrains. During QRA, a stationary free Rossby wave resonates in concert with a forced circumglobal wavetrain (see Text Box), thereby favoring the occurrence of persistent and high-amplitude excursions of the jet stream during certain background atmospheric states (Petoukhov et al., 2013; Coumou et al., 2017; Kornhuber et al., 2017a; Kornhuber et al., 2017b). Ongoing Arctic Amplification, Hadley Cell expansion, and changes in land-sea temperature contrast appear to favor these background-state conditions, potentially explaining the increasing QRA occurrences in recent decades (Coumou et al., 2014; Coumou et al., 2017) and the projected additional increase in the future (Mann et al., 2017).

Atmospheric dynamics are changing mainly due to heterogeneous warming of the climate and a more vigorous tropical convection. Dynamical changes can regionally either mitigate or exacerbate heat wave genesis substantially. Some evidence suggests that dynamical changes are favoring more persistent heat waves in the midlatitudes (Coumou et al., 2018; Mann et al., 2017; Horton et al., 2015; Lhotka et al., 2018; Pfliederer and Coumou, 2018), but uncertainties are large about this and not all studies have come to similar conclusions (Barnes and Screen, 2015; Horton et al., 2016; Screen and Simmonds, 2013; Cattiaux et al., 2016).



Sub-seasonal statistical forecasts of eastern United States hot temperature events

Extreme summer temperatures can cause severe societal impacts. Early warnings can aid societal preparedness, but reliable forecasts for extreme temperatures at subseasonal-to-seasonal (S2S) timescales are still missing. Earlier work showed that specific sea surface temperature (SST) patterns over the northern Pacific are precursors of high temperature events in the eastern United States, which might provide skillful forecasts at long-leads (~ 50 days). However, the verification was based on a single skill metric and a probabilistic forecast was missing. Here, we introduce a novel algorithm that objectively extracts robust precursors from SST linked to a binary target variable. When applied to reanalysis (ERA-5) and climate model data (EC-Earth), we identify robust precursors with the clearest links over the North-Pacific. Different precursors are tested as input for a statistical model to forecast high temperature events. Using multiple skill metrics for verification, we show that *daily* high temperature events have no predictive skill at long leads. By systematically testing the influence of temporal and spatial aggregation, we find that noise in the target timeseries is an important bottleneck for predicting extreme events on S2S timescales. We show that skill can be increased by a combination of (1) aggregating spatially and/or temporally, (2) lowering the threshold of the target events to increase the base-rate, or (3) add additional variables containing predictive information (soil-moisture). Exploiting these skill-enhancing factors, we obtain forecast skill for moderate heatwaves (i.e. 2 or more hot days closely clustered together in time) up to 50 days lead-time.

This Chapter is published as:

S. Vijverberg, M. Schmeits, K. van der Wiel, and D. Coumou (2020). “Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events”. In: *Monthly Weather Review* 148.12, pp. 4799–4822. DOI: 10.1175/MWR-D-19-0409.1

3.1 Introduction

Subseasonal to seasonal (S2S) predictions offer society valuable information on weather-related risk, allowing decision-makers to initiate early warning action plans for extreme events (WMO, 2017) and to optimize resource management (Vitart and Robertson, 2018a; Vitart et al., 2017). Predictability on these timescales stems from variables or more regularly varying oscillations which are evolving at lower temporal frequencies compared to the regular, more chaotic, weather (Doblas-Reyes et al., 2013; Mitchell et al., 2016; Krishnamurthy, 2019). This predictability can be exploited by (1) initializing a dynamical model with these slowly evolving variables such as soil moisture, sea-ice, snow cover and sea surface temperature (Jaiser et al., 2012; Seo et al., 2019; Vitart and Robertson, 2018a) or (2) select low-frequency variables directly as input for purely statistical forecasting models, using past climate data to train them (Kretschmer et al., 2017a; Cohen et al., 2018; Totz et al., 2017; Guimarães Nobre et al., 2019; Alfaro et al., 2006) (3) or a combination of both (Dobrynin et al., 2018).

S2S predictability can be improved by post-processing the output of dynamical models, which is conventionally done by compensating for systematic biases (Finnis et al., 2012; Doblas-Reyes et al., 2013). Alternatively, statistical models can be directly trained to make S2S predictions and offer computational efficiency, flexibility, and the precursor timeseries can be further analyzed to provide process information (Runge et al., 2019). In this paper, we use the word *precursor* to refer to an anomalous pattern or geographical region, while the *precursor timeseries* refers to the timeseries that results from a dimensionality reduction of this pattern or region. On S2S timescales, their forecast skill can be comparable to that of dynamical models (Hall et al., 2017). A better understanding of important precursors can also help with the (bias) correction of dynamical models, either by using the precursors directly or by using the statistical model to sub-sample only the reliable forecasting pathways of the dynamical model output (Dobrynin et al., 2018; Strazzo et al., 2019).

The ocean is the most important source of long-term memory that interacts with the atmosphere (Frankignoul, 1985; Kushnir et al., 2002; Kaspi and Schneider, 2011; Putrasahan et al., 2013; Thomson and Vallis, 2018). The atmospheric response to SST anomalies (SSTA) in the tropics is more direct and local, i.e. via thermally driven deep convection and associated latent heat release (Kushnir et al., 2002). In the mid-latitudes, the lower specific humidity content and smaller Rossby radius of deformation hinders the formation of strong deep convection as seen in the tropics, resulting in a weaker direct and local atmospheric response to SSTA (Kushnir et al., 2002; Hewitt et al., 2017). The mid-latitude atmospheric response to a SST anomaly is mainly driven by the adjustment to thermal wind balance and an indirect response due to eddy feedbacks (Nie et al., 2016). The latter makes the atmospheric response depend on the background zonal mean climate, and thus also on the season and location of the SSTA (Kushnir et al., 2002; Nie et al., 2016; Putrasahan et al., 2013).

These nuances for the atmospheric response, suggest that statistical forecasting tools should not solely rely on known modes of variability in the climate system, often referred to as climate indices [e.g. Hessler et al. (2004), Steptoe et al. (2018), and Guimarães Nobre et al. (2019)]. Those indices represent spatially large-scale, and temporally low-frequent processes. A priori, one does not know if they contain the information that is relevant for

the target of interest. This reasoning is supported by statistical studies and dynamical model results (McKinnon et al., 2016; Deng et al., 2018).

Following this rationale, McKinnon et al. (2016) showed that ‘hot day’ events in the eastern U.S. are preceded by a specific SST pattern over the Pacific. This SST pattern (called the Pacific Extreme Pattern, i.e. PEP) was found by analyzing composite SST anomalies that co-occurred at certain lags with heat events. The PEP pattern is characterized by a zonally oriented tripole cold-warm-cold pattern in the Pacific at approximately 35°N, related to the forcing and/or amplification of a Rossby wave train (Wirth et al., 2018). The PEP was shown to outperform the conventional climate indices, such as the Pacific Decadal Oscillation (PDO) or the El Niño Southern Oscillation (ENSO). Using the PEP pattern, the study claimed remarkable long-lead predictability up to 50 days lead time for extreme events defined at the daily timescale. In their main text, they assessed the skill of the PEP timeseries using only a single validation metric (area under the Relative Operating Characteristic). However, it is generally recommended to use multiple skill metrics that measure different aspects of the forecast quality to assess predictability (Wilks, 2011). Further, the study used a rectangular box over the tripole SST region to define the spatial extent of their precursor whereas a more objective SST pattern detection tool will verify if a link with the response variable is indeed robust and therefore might provide better physical understanding and more predictive skill (Bello et al., 2015; Kretschmer et al., 2017a).

A response-guided statistical forecast tool that searches for precursors that explain the full variability of a continuous response (i.e. target) variable has been developed already (Kretschmer et al., 2017a), but consequently, it cannot handle a binary target timeseries. Here, we introduce a novel response-guided algorithm that objectively extracts precursors that are directly related to our binary target variable, i.e. the eastern U.S. hot day event timeseries (section 23.2.3). We train this algorithm on reanalysis ERA-5 and 160 years of data from the coupled ocean-atmosphere model EC-Earth.

In this manuscript, we present (1) a comparison between our response-guided algorithm with the PEP pattern of McKinnon et al. (2016) and their relation to the relevant climate indices (section 33.3.2); (2) the verification of hot day forecasts, thereby stressing the importance of using multiple skill metrics (section 33.3.3). Section 33.3.4 shows forecast skill can be boosted by using temporal aggregation and lower-threshold events. To enable forecasts of events defined on a daily resolution, we no longer aggregate our target timeseries in time, but we use a window probability approach and we increase the signal-to-noise ratio by increasing the domain for spatial aggregation (section 33.3.5). Finally, while our focus lies on retrieving predictability from the ocean, in section 33.3.6 we also include additional information from soil moisture, since it is known to be a potentially important precursor of heatwaves (Seneviratne et al., 2010; Miralles et al., 2014; Ardilouze et al., 2017).

3.2 Method

3.2.1 Data

Our analysis relies on data from the ERA-5 reanalysis, 1979 – 2018 (Copernicus Climate Change Service (C3S), 2017) and from the EC-Earth v2.3 earth system model (coupling between ocean, atmosphere, land surface and sea ice) (Hazeleger et al., 2012) with 160 yrs of simulated present-day climate (van der Wiel et al., 2019). We calculate the daily maximum 2 meter air temperature (mx2t) in ERA-5 ($0.25^\circ \times 0.25^\circ$) by calculating the daily maximum of the 'maximum 2m temperature since previous post-processing', with a step size of 1 hour. For SST in ERA-5 we use daily means on a $1^\circ \times 1^\circ$ grid. We additionally use information from ERA-5 soil moisture ($1^\circ \times 1^\circ$) for the forecasts, i.e. the volumetric soil water levels of the 2nd [7 - 28 cm] and 3rd [28 - 72 cm] layer of the land surface model. To remove the seasonal cycle and the global warming trend (of which the strength might vary throughout the year), all variables are linearly detrended for each day-of-year. Because a single day-of-year across 40 years is insufficient to reliably estimate the climatological mean value and trend, we apply a 25-day rolling mean (using a Gaussian window with a standard deviation of 12.5) to the raw ERA-5 data. From the raw data, we then subtract climatological mean and trend based on the smoothed data.

For EC-Earth, we use daily mean T2m and SST data ($1.125^\circ \times 1.125^\circ$). The coupled ocean-atmosphere climate model experiment consisted of 2000 years of simulated present-day weather, from this we sampled 160 years for our study. The selected years are not chronological, which is a desired property for making good splits between training and test data, because no inter-annual information is passed from the previous to the subsequent years. For more information on the model simulation set-up, see van der Wiel et al. (2019). For EC-Earth, the seasonal cycle and a potential long-term trend is directly removed for each day-of-year (no prior smoothing), since 160 years should be enough to reliably estimate the trend and climatological mean value.

3.2.2 Defining the target variable and the Pacific Extreme Pattern

We define our target variable following McKinnon et al. (2016) and determine it for ERA-5 and EC-Earth based on the detrended temperature data. The study period consists of the climatological 60 hottest days of year, ranging from 24th of June to the 22nd of August (McKinnon et al., 2016). The target variable is retrieved by, firstly, performing an objective identification of spatial clusters within the U.S., where gridcells are clustered together if they tend to experience extreme events simultaneously. This clustering approach is expected to increase the signal-to-noise ratio and thereby helps to identify precursors, for more information on the clustering see Appendix A.

Hot day events

McKinnon et al. (2016) calculated the spatial 95th percentile of daily maximum temperature anomalies within the eastern U.S. cluster. Hence, for each day, the spatial 95th percentile of all observations was calculated, which in practice means that each day contained the temperature value of only a single observation. This introduces some unwanted noise into

the target timeseries since small scale processes can affect the maximum temperature at a *single* observation and *single* moment in time. To improve the signal-to-noise ratio and at the same time stay close to the original definition, we calculate the spatial mean over the 10% warmest grid cells. This way we still end up with a very similar timeseries as compared to the T95 timeseries used by McKinnon et al. (2016) (figure 3.3). We refer to this timeseries as $T90_m$ in the remainder of this article, with the lower case m referring to the spatial mean that is calculated. The hot day timeseries (HD) is defined as,

$$HD(t) = HD : 1 \text{ if } T90_m > (\overline{T90_m} + \sigma_{T90_m}) \text{ else } E : 0, \quad (3.1)$$

with $\overline{T90_m}$ being the temporal mean and σ_{T90_m} being the standard deviation of $T90_m$. This results in a base-rate of approximately 16%.

PEP

The PEP pattern is retrieved by taking the area weighted SSTA composite mean of hot day events at lag τ . The spatial region is defined by the rectangular box as depicted by green stippled lines in figure 3.4, the coordinates are [145°E, 130°W, 20°S, 50°N]. The PEP timeseries at lag τ is defined as the spatial covariance between the PEP pattern and the SSTA field at each timestep ($SSTA(t)$),

$$PEP_\tau(t) = \frac{1}{N} \sum_i^N w_i \left[\left(PEP_\tau(t, i) - \overline{PEP_\tau(t)} \right) \cdot \left(SSTA(t, i) - \overline{SSTA(t)} \right) \right] \quad (3.2)$$

where i denotes a gridcell of in total N gridcells within the rectangular box, w_i denotes the weight proportional to the gridcell area, the overbar denotes the spatial mean.

3.2.3 Composite-based Precursor Pattern Algorithm

This is a response-guided algorithm in the sense that it searches for a signal that directly relates to a response (i.e. target) variable of interest, in this case, the hot day timeseries. It is inspired by the approach presented in McKinnon et al. (2016), who created a composite mean of hot day events, i.e. calculating the mean SSTA that co-occurred with hot day events at a certain lag. The null hypothesis would be that the SSTAs are unrelated to heat wave events, meaning that one would be randomly sampling anomalies with respect to climatology, which should approximate zero. However, a distinct pattern of significantly deviating SSTA was found. This algorithm automatically infers the precursor regions based on robust anomaly patterns in the composite mean. The algorithm is described in step 1 and 2 (figure 3.1) and the parameters are listed in Table 3.1.

Detecting robust SST precursors

Robust anomalous gridcells should (i.) be insensitive to the exclusion of a (number of) year(s) and (ii.) SST anomalies should persist through time for at least a few days. Criterion (i.) is tested by creating sub-sampled composite mean (SCM) maps and setting gridcells exceeding a percentile threshold (param

= SCM_perc_thres ; Table 3.1) to 1 and the rest to 0. We iteratively remove a number of training years based on a percentage. If we are, for example, removing 7.5% of the N_{yrs} (i.e. 36 for ERA-5) training years, we delete 7.5% of 36 = 2.7, which we round to 3 years. This is done N_{yrs} times, each time removing a different subset while ensuring that the deleted years are uniformly sampled, thereby avoiding that a certain year is recurrent in many of the SCMs, while others are not. This procedure of removing a percentage is done multiple (N_{perc}) times, once for each percentage in the list $perc_yrs_out$ (i.e. for ERA-5, the list of percentages are [5, 7.5, 10, 12.5, 15], thus $N_{perc}=5$). Criterion (ii.) is tested by re-doing the previous step, but then the composite dates are shifted by n_{days} in time. These date shifts with respect to the composite dates are listed in param = $days_before$. For ERA-5, date shifts are [0, 7, 14], thus $N_{shifts}=3$). In total, the sub-sampling leads to $N_{yrs} * N_{perc} * N_{shifts} = N_{tot}$ SCMs. For ERA-5 data, N_{tot} is equal to 540.

Next we calculate (and normalize) the frequency for each gridcell to obey criterion i. and ii., and we reject those which are not extracted at least 80% (param = FSP_thres) of the N_{tot} SCM maps. We found that using other reasonable parameter settings lead to qualitatively the same results. To form individual precursor regions, we use Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) (Schubert et al., 2017), which assigns separate labels to groups of adjacent robust gridcells of the same sign (see figure B2). To achieve this, we use the Haversine formula as the distance metric, which calculates the great-circle distance between two points on a sphere.

To summarize, CPPA searches for a robust SSTA pattern associated with the events of interest and using *DBSCAN* assigns separate labels to adjacent robust gridcells, thereby grouping them into precursor regions. Each of these precursor regions are reduced to a 1-dimensional timeseries by calculating the area-weighted mean. Similar to equation (3.2), we also calculate a spatial covariance timeseries of all the precursor regions together, referred to as CPPA spatial pattern timeseries, or short CPPAsp. Hence, CPPA outputs both the spatial pattern timeseries *and* a single timeseries for each precursor region. For more detailed information about the output of the algorithm and a comparison to using the linear Pearson correlation metric, see Appendix C.

3.2.4 Forecasting method

We implement a logistic regression (Varoquaux et al., 2015), which tunes the regularization coefficient using cross-validation. Conventional logistic regression optimizes the coefficients to minimize the loss function, which tends to lead to overfitting. The regularization improves generalizability to unseen validation data by minimizing the loss function + $1/C$ times the sum of the squared coefficients (L_2 regularization), with $1/C$ being the regularization coefficient. $1/C$ is tuned by a second stratified 5-fold cross-validation, i.e. the model is trained on a subset of the data and subsequently, the generalizability of the model is tested on validation data. The regularization coefficient that renders the best average score on the 5 validation sets is chosen. Using the best value for C , the model is re-fitted on all 36 years. We choose this statistical model because it does not have as

many degrees of freedom compared to complex machine learning models and therefore we are less prone to a limitation by datapoints.

Note that we are first separating the train-test split via stratified cross-validation and subsequently split each training set into train-validation sets via another stratified cross-validation. This allows us to efficiently use as much data as possible for training, while the test data is always strictly separated. For more information on this double cross-validation framework and a schematic overview, see Appendix B. All precursor timeseries are standardized, where the mean and standard deviation are based on training data.

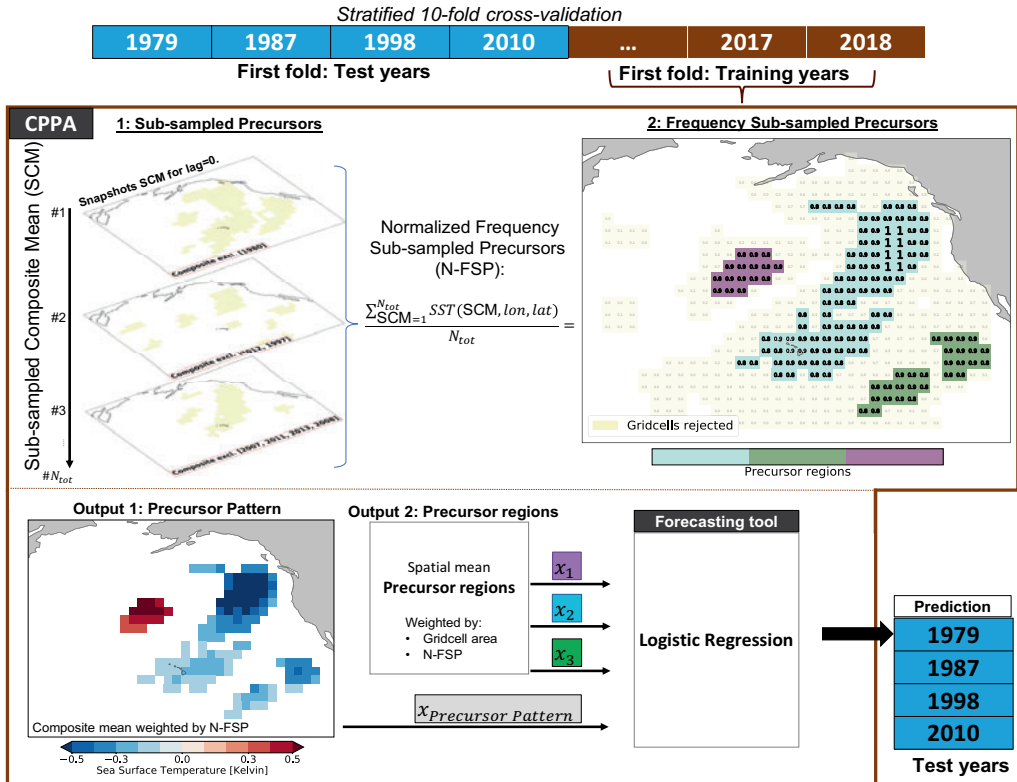


Figure 3.1: Schematic illustration of the Composite-based Precursor Algorithm (CPPA). In the upper-left we illustrate step 1, which detects gridcells with robust composite anomalies for a given lead time. We define robustness by selecting those gridcells that consistently exceed a percentile threshold, irrespective of the sub-sample used (i.e. by leaving out some years or by shifting the composite lead-times by a couple of days). In step 2, we reject all non-robust gridcells and the remaining (robust) gridcells are grouped into precursor regions (shown in different colors). The output of the algorithm consists of the spatial co-variance of the full precursor pattern ($x_{Precursor Pattern}$), and the spatial mean of all individual precursor regions (x_1, x_2, x_3).

Table 3.1: Parameters of the Composite-based Precursor Pattern Algorithm.

	Parameter names	Settings for ERA-5 data	Settings for EC-Earth data
1	<i>SCM_percentile_thres</i>	0.95	0.95
2	<i>perc_yrs_out</i>	[5, 7.5, 10, 12.5, 15]	[10, 20, 30, 40]
3	<i>FSP_thres</i>	0.80	0.80
4	<i>days_before</i>	0, 7, 14 [days]	0, 7, 14 [days]
4	<i>min_area_in_degrees2</i>	5° ²	5° ²
5	<i>distance_eps</i>	500 [km]	500 [km]

Table 3.2: Contingency table

Contingency table		Event observed	
		Yes	No
Forecast	Yes	true positive (tp) / Hit (H)	false positive (fp) / False Alarm (FA)
	No	false negative (fn) / Misses (M)	true negative (tn) / Correct Negatives (CN)

3.2.5 Forecast Validation

According to Wilks (2011), "forecast skill refers to the relative accuracy of a set of forecasts, with respect to some set of standard reference forecasts". A good quality forecast should meet a number of requirements, which cannot be summarized by a single scalar quantity (Wilks, 2011). The World Meteorological Organization set up standard guidelines (WMO, 2006) for verification of long-range forecasts, encouraging the use of Relative Operating Characteristic (ROC), reliability curves, and a Mean Squared Skill Score (i.e. Brier Skill Score). We argue that one can only claim predictive skill if it performs well on all metrics. In addition, the forecast should perform better than an appropriate reference forecast, which for sub-seasonal predictions is the climatological probability. We use Area Under the Curve Relative Operating Characteristic (AUC-ROC) and Area Under Curve Precision-Recall (AUC-PR), Brier Skill Score and reliability plots.

The AUC-ROC was also used by McKinnon et al. (2016). The ROC represents a balance between true positive rate (TPR) and false positive rate (FPR) for different thresholds of the binary forecasting time series (Table 3.3 and 3.2). The ROC area can be interpreted as "the probability that the forecast probability assigned to the event is higher than to the non-event" (Mason and Graham, 2002). See also Kharin and Zwiers (2003), Fawcett (2006), and Wilks (2011) for more information.

The AUC-ROC does not take into account the precision, reliability and resolution of the forecast. Although the precision-recall curve still does not take into account the reliability and resolution, it is more suitable for imbalanced classes (Saito and Rehmsmeier, 2015) and has a focus on evaluating the positive predictions, i.e. the forecasted events. It quantifies the balance between precision and the Recall (or TPR) for different thresholds. If we forecast events using a low threshold, it is easy to get a very high precision, but difficult

Table 3.3: Summary of verification metrics used in this article, see Table 3.2 for the contingency table.

	Calculation	Description
BSS	$(BLS_f - BLS_c) / BLS_c$	Mean Squared Error for binary classification (forecast vs. climatology).
Precision	$tp / (tp + fp)$	Correct positive predictions vs. all positive predictions.
Accuracy	$(tp + tn) / (tp + tn + fn + fp)$	Ratio of total correct predictions.
TPR (Recall)	$tp / (tp + fn)$	Correct positive predictions vs. total number of events.
FPR or (1 - specificity)	$fp / (fp + tn)$ or $1 - tn / (tn + fp)$	Incorrect positive predictions vs. incorrect positive predictions + correct negative predictions
AUC-ROC	Aura under curve TPR vs. FPR points	Forecast probability assigned to event higher than to non-event.
AUC-PR	Aura under curve Precision vs. TPR points	Does not consider true negatives (Misses), focus on positive predictions.

to get a high TPR (the denominator will be high due to many False Negatives).

The Brier Skill Score (BSS) is a commonly used metric for quantifying the quality of a probabilistic forecast. It takes into account both the reliability and resolution (Wilks, 2011). The reliability quantifies to what extent forecast y_i deviate from the conditional average observation (mean of distribution of observations (\bar{o}_i), conditioned on the forecast (y_i), i.e. $\bar{o}_i = p(o_i|y_i)$). Resolution quantifies the difference between the conditional average observation (\bar{o}_i) and the climatological probability ($\bar{o}_i - \bar{o}$), i.e. forecasts with high resolution can, on average, more confidently distinguish events from non-events. The BSS is calculated using the Brier score (BS, equation (3.3)) for a given probability timeseries.

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2, \quad (0 \leq BS \leq 1) \quad (3.3)$$

With p_i being the forecast probability at timestep i , and o_i , the observed event (1 or 0). The climatological Brier score is calculated for each train-test split by assuming a constant climatological probability (BS_c) based on the concomitant training dataset, i.e. the same as used to fit the statistical model. Using BS_c and BS_f we calculate the Brier skill score (eq. 3.4); if the BSS is significantly above 0, the forecast system is better than climatology.

$$BSS = \frac{BS_c - BS_f}{BS_c} = 1 - \frac{BS_f}{BS_c}, \quad (BSS \leq 1) \quad (3.4)$$

The reliability diagram (e.g. the last row of figure 3.6) is used to visualize how reliable and resolute the forecast is. On the x-axis we plot the forecast probability ranging from 0 to 1 (with 1 being 100% probability). For the reliability curve we use 10 equally sized bins (stepsize=0.1) and plot the forecast probability on the x-axis and observed frequency on the y-axis. A perfectly reliable probabilistic forecast would always match the observed frequency, i.e. show a diagonal line. A histogram is plotted below the curve to show the forecast distribution, which informs about the sharpness of the model. The sharpness refers to the ability of the forecast model to substantially deviate from the climatological probability. The dark grey area shows where the forecast is better than climatology ($BSS > 0$), and the light grey area shows where the forecast is only doing better than a random forecast.

Confidence intervals are created by bootstrapping ($n=2000$, unless stated otherwise), where we bootstrap blocks to account for autocorrelation, thereby avoiding over-sampling dependent datapoints. The block-size is objectively defined by the lag at which the autocorrelation becomes significantly different from zero.

3.3 Results

3.3.1 Spatial clustering and hot days in ERA-5 and EC-Earth

We performed a parameter sweep to test for robustness of the eastern U.S. cluster in the ERA-5 and EC-Earth datasets, as further detailed in Appendix 3.A. Overall, we conclude that the eastern U.S. cluster is robust, i.e. it is generally categorized as a separate cluster, with only small differences in the exact boundaries and size (depending on minor

perturbation of the clustering parameters). As described in Appendix 3.A, we choose the eastern U.S. cluster that is most similar to McKinnon et al. (2016), see figure 5.2. We calculate the spatial $T90_m$ and the associated hot days as explained in section 3.2.2. Figure 3.3 shows that the ERA-5 $T90_m$ timeseries and associated frequency of hot days (year^{-1}) matches closely with the original T95 and hot day timeseries found by McKinnon et al. (2016).

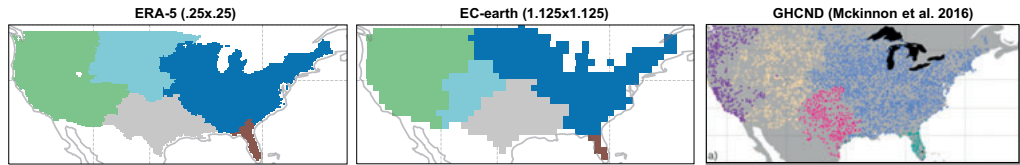


Figure 3.2: Result of clustering each location based on a binary timeseries, containing information on the timing of exceedances of a large anomaly (1 or 0). The datasets differ in spatial resolution (ERA-5 : 0.25° , EC-Earth : 1.125°) and time period (ERA-5 : 1979-2018, EC-Earth : 160x1 yr (present-day climate). The original clusters as presented in McKinnon et al. (2016) by using GHCND station data : 1980 - 2015 .

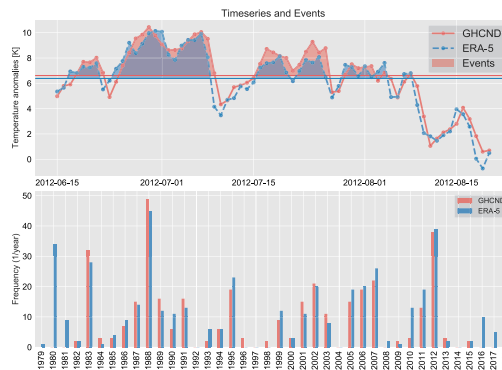


Figure 3.3: Upper plot: year 2012 of original $T95$ timeseries and hot day events based on observational GHCND station from McKinnon et al. (2016) and the $T90_m$ timeseries based on ERA-5 reanalysis data. Events are defined as exceeding one standard deviation (σ) of $T90_m$. Bottom plot: frequency of hot days per year. For comparison in the upper plot the mean and σ of both timeseries is calculated over same period (1982 – 2015) as is available from the original $T95$ timeseries, for the bottom plot and elsewhere we use the whole ERA-5 timeseries (1979 – 2018).

3.3.2 Comparison between CPPA, PEP and climate indices

Figure 3.4 shows the hot day events composite mean of SST gridcells (mean over 10 training datasets) for both ERA-5 and EC-Earth, where the stippled green rectangle depicts the PEP pattern and the black contour lines show the robust anomalous grid cells detected by CPPA. As can be seen from figure 3.3, there is a lot of inter-annual variability in the amount of hot days, with 4 years together accounting for 33% of the events and

with 9 years having less than 1% of the events in the ERA-5 reanalysis. The output of CPPA, however, is robust across the 10 training datasets, as detailed in Appendix B and figure 3.B.1. For ERA-5, the labels that are (randomly) assigned to each region by the DBSCAN clustering algorithm are shown in figure 3.C.1.

We observe that, in the tropical Pacific, a La Nina-like pattern is picked up in EC-Earth and not in ERA-5 and also the tropical Atlantic precursor regions are different. Both ERA-5 and EC-Earth do share the cold-eastern and warm mid-Pacific features. These are also the main features of the PDO pattern and are part of the PEP pattern as presented by McKinnon et al. (2016). Yet the cold western-Pacific of the PEP pattern is considered non-robust according to CPPA.

We also analyzed how the $T90_m$, PEP, Nino3.4, PDO, and CPPA spatial pattern (CPPAsp) timeseries are linked to each other via a cross-correlation matrix (figure 3.5). See Appendix E for background information on the calculation of the PDO and ENSO indices. We observe that the PEP timeseries show a higher correlation coefficient with $T90_m$ compared to the CPPAsp timeseries and the climate indices (PDO and Nino3.4), particularly during the summer days. In the following section, we will compare the forecast skill between PEP, the climate indices and the CPPA timeseries (CPPAsp and CPPA precursor regions timeseries). The difference between EC-Earth and ERA-5, the link between PEP, CPPAsp and the climate indices and the potential physical mechanism are further discussed in section 5.4 and 3.4.3.

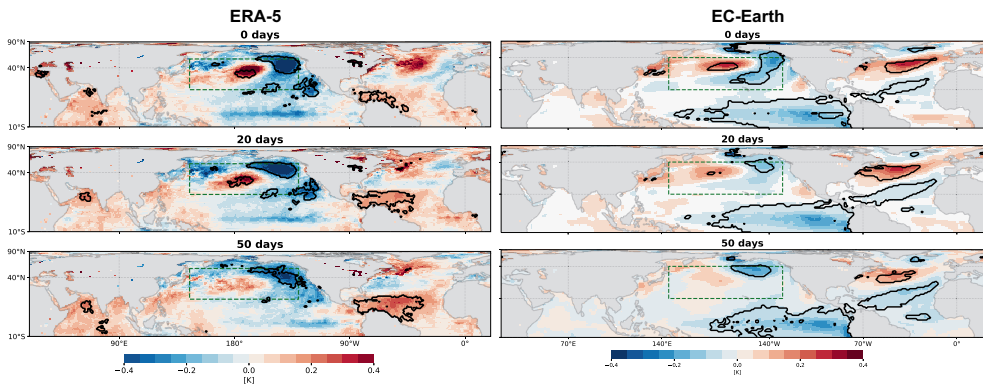


Figure 3.4: Composite mean of hot day events (mean over 10 training datasets presented) for both (a) ERA-5 and (b) EC-Earth. The lag with respect to hot day events is presented in the subtitles. The stipled green rectangle depicts the PEP pattern. The contour lines show the robust anomalous gridcells which are extracted at least 5 out of 10 training datasets.

3.3.3 Using multiple validation metrics

Figure 3.6 shows the verification of hot day event forecasts, comparing the use of the PEP timeseries versus the CPPA output to fit the statistical model. As explained in section 23.2.5, we objectively determine the block window size for bootstrapping by calculating up to which lag the autocorrelation is significantly different from 0 (see figure 3.7), for

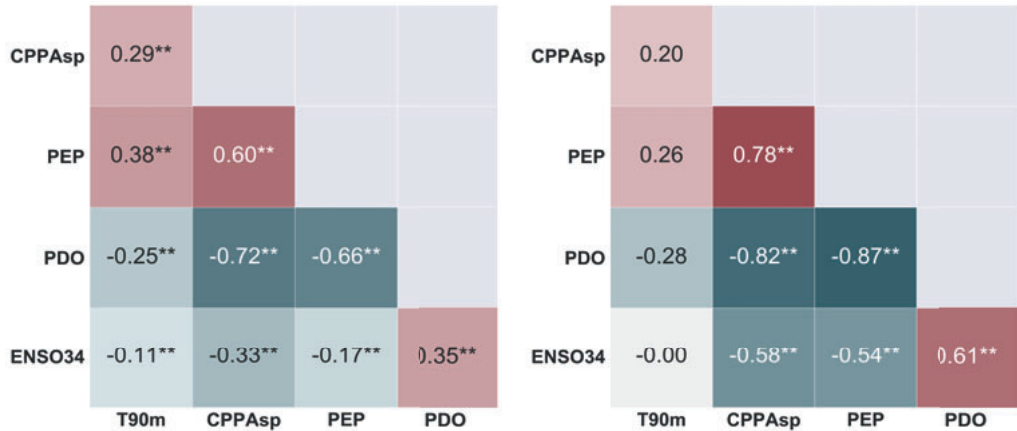


Figure 3.5: (a) Out of sample cross-correlation matrix for daily data during the study period (24th of June to the 22th of August) and for (b) annual mean values. The ENSO timeseries refers to the Nino3.4 timeseries, defined by the area-weighted SSTA mean between 5°S - 5°N and 170°W - 120°W. Based on ERA-5 data. The ** indicates significance at p value < 0.01.

the ERA-5 daily $T_{90,m}$ timeseries this is 32 days (figure 3.7a) and for the EC-Earth daily $T_{90,m}$ timeseries this is 71 days (figure 3.7c).

We observe that forecasts based on either PEP or CPPA perform better than random chance, rendering approximately the same skill for ERA-5. For EC-Earth data, we observe that CPPA is a better precursor compared to PEP. We also see lower skill for EC-Earth, even though the climate model data has 4 times as many datapoints. Both datasets, however, do not render a significantly better forecast compared to the climatological probability, as is evident from the near-zero BSS values and the reliability diagrams (figure 3.6). This non-existent predictability for hot days is not surprising given the fact that we are trying to predict the exact day at which the hot day event should occur. Even for the EC-Earth data, where we have many datapoints available, the statistical model cannot resolutely discriminate between events and non-events. Since we know the EC-Earth model has its limitations in representing the real climate, especially extremes, we will now only focus on the ERA-5 dataset.

3.3.4 Temporal aggregation to improve signal-to-noise ratio

To improve the signal-to-noise ratio, we aggregate over time with the trade-off of a reduction in temporal precision and the number of datapoints. We aggregate the daily data into bins of 15 days and calculate the mean of all bins. The window size of 15 days is commonly used in the literature when studying Rossby wave dynamics (Kornhuber et al., 2017a; Röthlisberger et al., 2018). Since we are now working with time-windows, the lead time is defined from the day that the forecast would be issued, using only information prior and including that day, to the center date of a forecasted time-window, see Appendix F for more information. We will compare the forecasts with the conventional approach, i.e. using the relevant climate modes of variability from SST (PDO+ENSO)

Figure 3.8 shows the verification when we first calculated 15-day means of $T_{90,m}$ and then

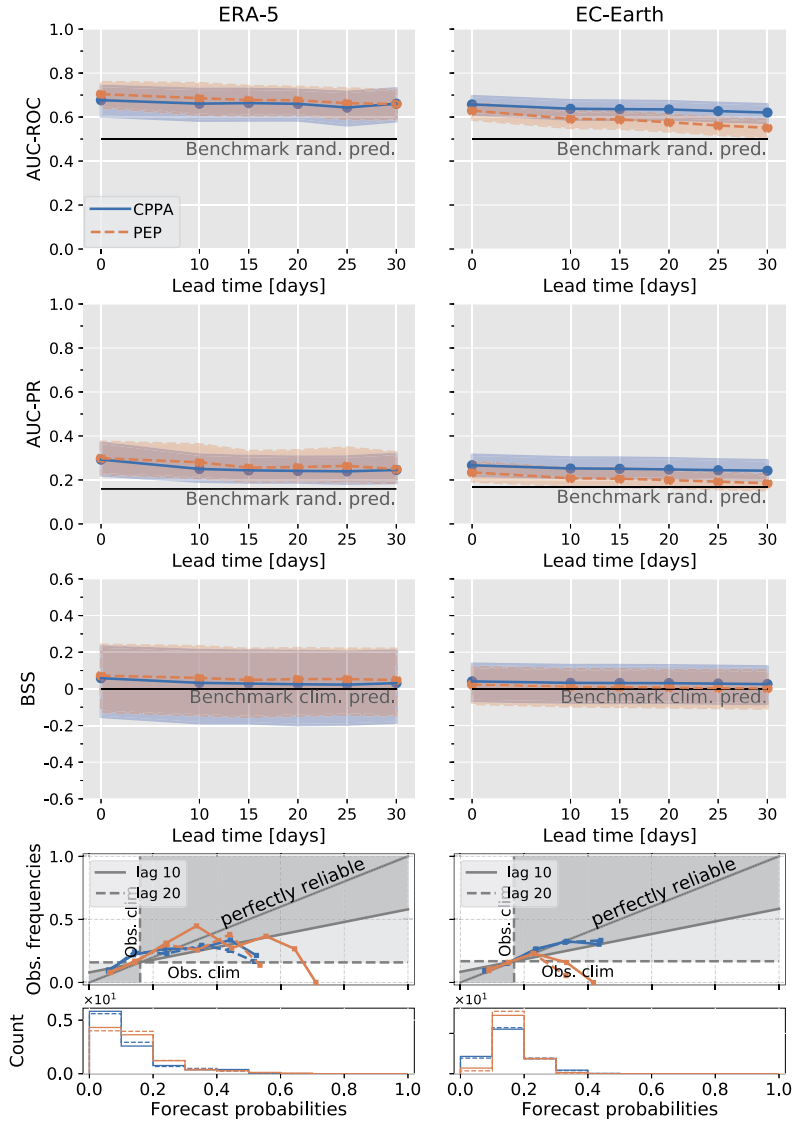


Figure 3.6: Forecast validation for hot days, using only information from SSTA. We compare using the PEP pattern with the CPPA precursors for forecasting and show the importance of using multiple skill metrics.

used the event definition that was also used to define hot days (see equation (3.1)), thus having a base-rate of approximately 16%. The block window size is 5 time-steps, i.e. 75 days. For these so-called 'hot 15-day mean events', we observe a decline in skill, not an improvement. The histogram shows that almost all values are close to the climatological probability, especially for the PDO+ENSO forecast. Ostensibly, we still have insufficient information to fit a reliable model and/or the reduction in datapoints seems to dominate the benefit of a better signal-to-noise ratio.

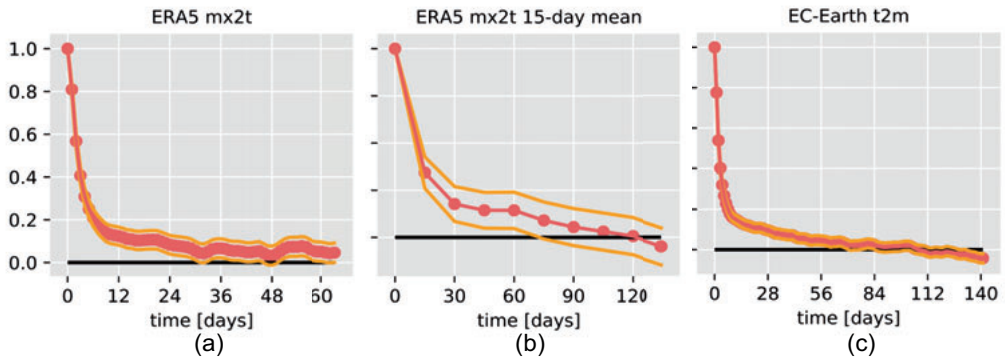


Figure 3.7: The autocorrelation of $T90_m$ and the 15-day mean $T90_m$ in ERA-5 and the $T90_m$ in EC-Earth. The autocorrelation is used to determine block window size for bootstrapping.

Thus, next, we lower the extremity of the events (which increases the base-rate) and define the target based on 15-day upper tercile and 15-day above median events. Figure 3.9 shows that the statistical models that are fitted using the CPPA precursors outperform the ones that use PEP, or PDO+ENSO. For the upper tercile events (right two columns of figure 3.9), skill is better compared to the hot 15-day mean events (figure 3.8), but is slightly lower compared to the above median events, which shows skill up to at least 30 days lead-time (left two columns of figure 3.9). To summarize, we improved forecast skill by (1) finding better precursors using CPPA, (2) using temporal aggregation in combination with increasing the base-rate (i.e. lowering the threshold for events).

3.3.5 Using a window probability and spatial aggregation to improve event forecasts

To increase predictability of extreme events, we relax the temporal precision by using a 'window probability', meaning that we predict the occurrence of a relatively short heatwave event within a longer time-window. Hence, the exact date of occurrence within this time-window is flexible. When using a 15-day time-window, a predicted heat wave event may thus occur 7 days earlier or later. We define a heatwave when 2 or more hot days occur with at most one non-hot day in between. With this approach, we still smoothen out noise in the *precursor* timeseries (by using 15-day means), while still predicting relatively short-lived events consisting of daily temperature extremes.

Still, the target variable is not smoothened in time. To increase the signal-to-noise of the target variable we apply spatial aggregation. We do this in a similar manner as was done for $T90_m$ (section 3.2.3.2.2), defined as the *spatial* mean over the 10% warmest gridcells within the eastern U.S. cluster. Here, we define two additional target timeseries with increased spatial aggregation by calculating the mean over the 35% warmest ($T65_m$) and 50% ($T50_m$) warmest gridcells. Subsequently, the hot days are defined for each timeseries using the equivalent of equation (3.1).

The Brier skill score for the $T90_m$ heatwave forecast (figure 3.10, left column) is lower compared to that of upper tercile 15-day mean $T90_m$ events (figure 3.9, right columns),

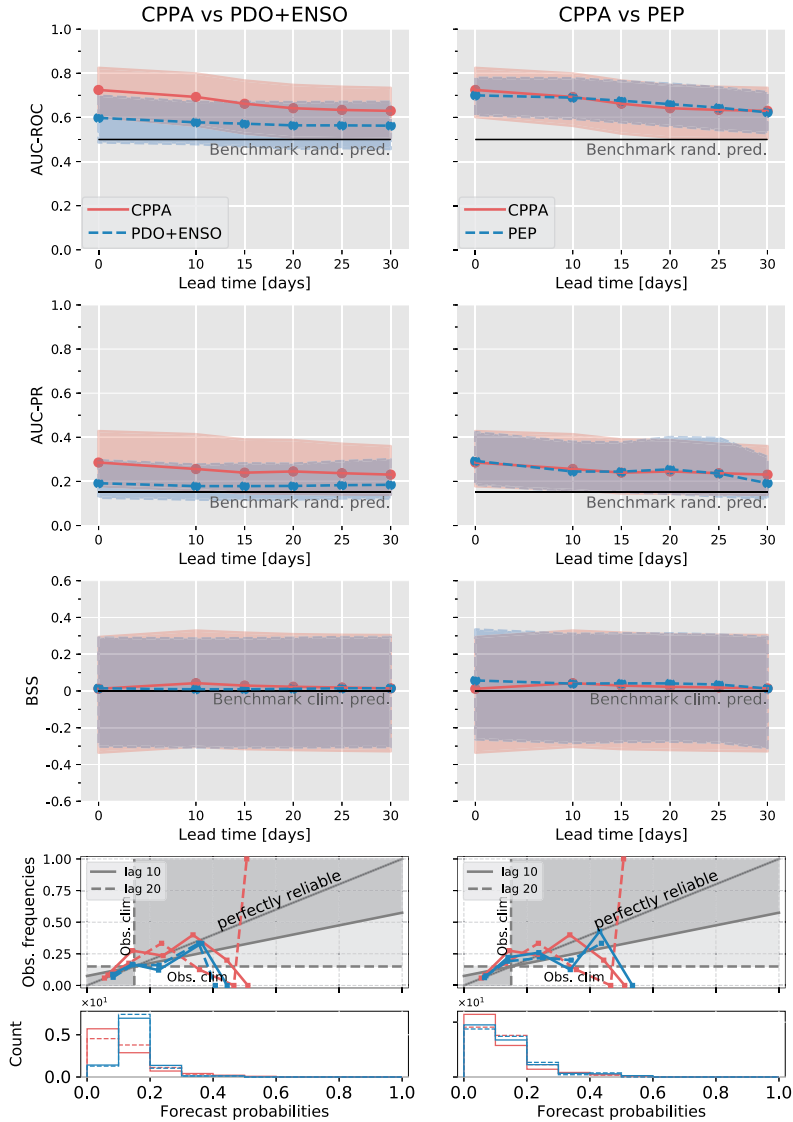


Figure 3.8: Forecast validation for 'hot 15-day mean events'. Here we show the comparison between using the PDO+ENSO versus the CPPA precursors (left column) and the PEP pattern versus the CPPA precursors (right column).

even though the base-rate of the $T90_m$ heatwave window probability is higher (41%). By aggregating over space ($T65_m$ and $T50_m$, i.e. second and third column of figure 3.10) one reduces the noise in the target timeseries and thereby enhances forecast skill.

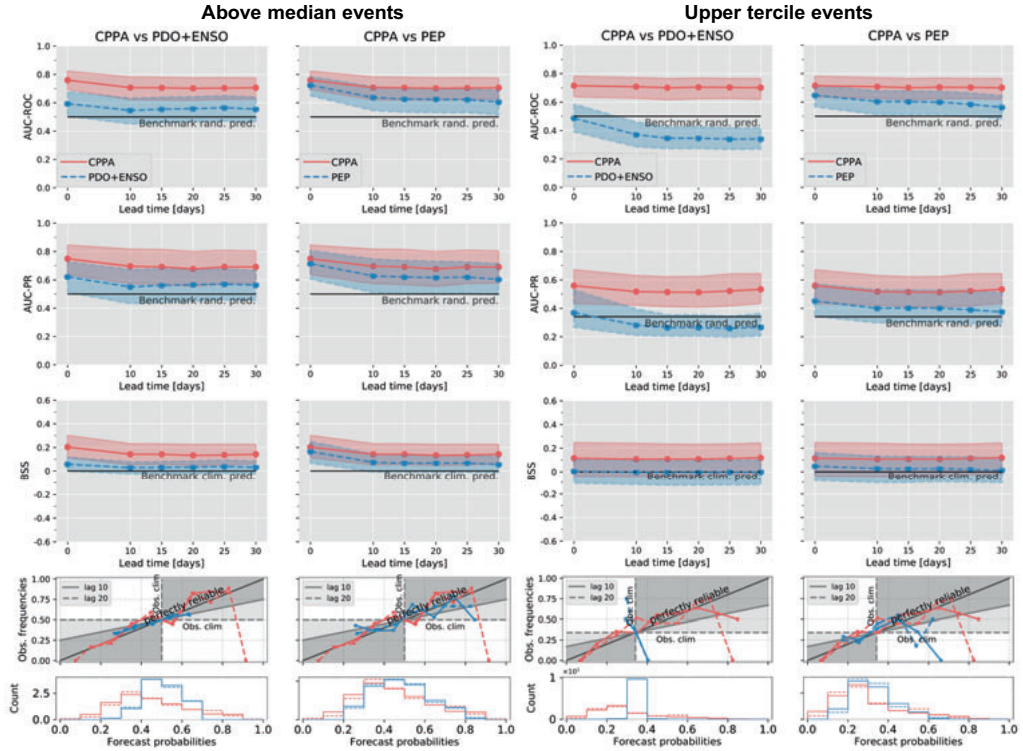


Figure 3.9: Similar to figure 3.8, but on the 2 left columns for above median events and on the 2 right columns for upper tercile events of the 15-day mean T_{90_m} timeseries. Based on ERA-5 data.

3.3.6 Sub-seasonal forecasts of moderate heatwaves using both SST and soil moisture

Previous results focused on quantifying predictability from only SST. Now we aim at enhancing forecast skill by including additional information from soil moisture. We proceed with T_{65_m} as it has significant skill up to at least 30 days (figure 3.10, central column), while still being relevant for temperature extremes. During the summer days, the daily T_{65_m} timeseries¹ has a temporal mean value of 2.3°C and standard deviation of 2.1°C . Using the equivalent of equation (3.1), the events have an average anomaly of 5.6°C ranging between 4.4 and 9.9°C . 466 Days belong to these events (base-rate of 15.5 %), and after grouping these days into multi-day events (as defined in section 33.3.5) there are 103 events left. Because the threshold is now less extreme, we will call these events moderate heatwaves.

Figure 3.11 shows the verification results for forecasts when using precursors both from CPPA and soil moisture (orange dashed line) and when using only CPPA (blue solid

¹Note, T_{65m} refers to a timeseries from calculating the spatial mean of the 35% warmest eastern U.S. grid cells on each day; it has a temporal mean and standard deviation (just like T_{90_m} is a timeseries, as shown in Figure 3.3).

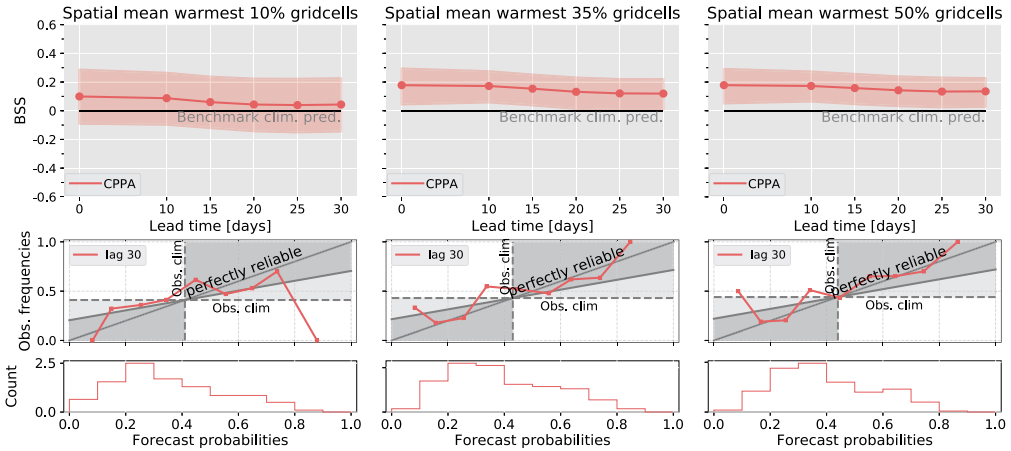


Figure 3.10: Forecasting heatwaves (defined in the text) within a 15-day window. Columns: 3 different spatial aggregation sizes are used to define our continuous temperature timeseries (T_{90m} , T_{65m} , T_{50m}), after which the associated moderate heatwave events are calculated. Based on ERA-5 data.

line). We observe that soil moisture contributes to a small increase in skill up to 30 days lead-time, but for longer lead-times all information can be retrieved from the SST precursors. Tables 3.C.1 and 3.D.1 show all precursors that were used for this prediction. Figure 3.F.2 shows that the 10 models with a lead-time of 50 days that were learned based on different training datasets are robust, i.e. the 10 models generally learned the same regression coefficients. Figure 3.F.3 shows that also the forecast quality is robust when using different train-validation combinations, see also Appendix B.

For this forecast, we achieve predictive skill 50 days in advance at the 2.5 to 97.5th confidence interval ($n=5000$). This forecast for moderate heatwaves is more capable of discriminating between event and non-event occurrences (higher resolution) compared to the original hot day definition, as is evident from the reliability diagram and Brier Skill Score.

3.4 Discussion

3.4.1 Using multiple validation metrics

A proper forecast validation for eastern U.S. hot days (i.e. forecasting individual days), shows that the forecast does not perform significantly better than the climatological probability. The probabilistic forecasted values for hot days were not able to confidently discriminate between events and non-events, i.e. low resolution, $p(o_j|y_i)$, where y_i is the forecast probability, and o_j are the observed values (Wilks, 2011). This can be seen from the reliability diagrams in figure 3.6. Contrarily to McKinnon et al. (2016), we conclude that there is no predictive skill for individual hot days.

The AUC-ROC metric measures discrimination (see section 23.2.5), also the forecasted values are only sorted and their actual value is neglected. Thus, resolution will not

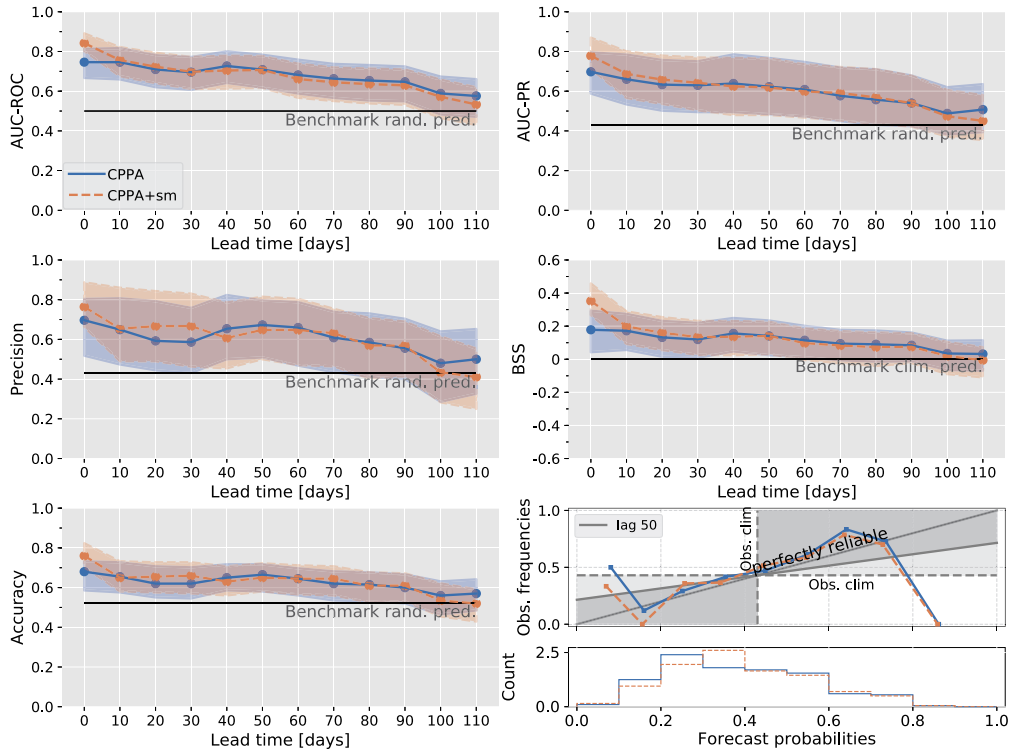


Figure 3.11: Verification results for forecasting T_{65_m} heatwaves (defined in the text) within a 15-day window. Solid blue line shows the results for the forecast when using the CPPA timeseries, the stippled orange line when including precursor timeseries from both CPPA and soil moisture (see Table 3.C.1 and 3.D.1 list of all precursors that were used). Bootstrap sample size is 5000. Based on ERA-5 data.

be measured, and consequently the forecasted probability might be always close to the climatological probability (p_c), which makes them of low practical value. If one wants to assess predictive skill, the AUC-ROC is an improper validation metric if used by itself as it only measures *potential* skill (see chapter 8 in Wilks (2011)).

3.4.2 Improving statistical forecasts for events

The problem when predicting extreme events on S2S timescales lies between a boundary condition problem and an initial value problem (Vitart et al., 2019), i.e. the boundary conditions that we use to constrain a target distribution (in this case temperature) changes over time. From this perspective, we believe there are 3 limiting factors for these statistical forecasts: (1) missing information of low-frequency drivers, (2) the chaotic nature of the atmosphere, i.e. knowing the full constrain of the boundary condition(s) is still not strong enough to reliably predict extreme events (Krishnamurthy, 2019; Vitart et al., 2019), (3) the statistical model is sub-optimal due to insufficient datapoints and/or the complex non-linear interactions cannot be described by the model.

When using only PDO and ENSO for forecasting (see figure 3.9), we clearly miss information compared to using the CPPA timeseries. While we have many datapoints when using the EC-Earth dataset for the forecast of individual hot day events (figure 3.6) or 'hot 15-day mean events' (figure 3.F.4), we are still unable to produce reliable and resolute forecasts. However, improved precursors (using CPPA) and temporal aggregation (relating to point 1 of the limiting factors) can only contribute to a pronounced improvement in skill when we predict events that are not too extreme (relating to point 2) (compare figure 3.8 and 3.9). We do not think the linear statistical model (point 3) was inadequate, we have tried tuning a tree-based Gradient Boosting Regressor (GBR) by doing an extensive parameter grid search (results not shown). However, the best performance of the GBR was only as good as the regularized logistic regression.

To enable forecasts of more extreme events, we use a window probability definition for the target variable (i.e. the probability of a relatively short-lived heatwave occurring within a longer prediction window) and, in addition, apply stronger spatial smoothing. This combination effectively reduces the noise in the target timeseries and increases the base-rate, while still predicting societally relevant high-temperature events. Thus, we conclude that S2S predictions of high-temperature events are possible, but also fundamentally limited by the chaotic nature of the atmosphere constraining the signal-to-noise ratio and the availability of data which hampers the detection of the signal. Nevertheless, with the techniques presented here, a stakeholder can be helped to decide on the preferred balance between spatial-aggregation, temporal-aggregation and extremity of the to-be-forecasted events. Thus, given the stakeholder needs, optimal aggregation and threshold levels can be found to attempt to render skill at the desired lead-times.

3.4.3 Physical interpretation of the CPPA pattern

In a response-guided approach the features are learned objectively, which can improve forecast skill compared to using e.g. climate indices, as is shown in this paper. Another important advantage is that the features remain physically interpretable, hence, they can be evaluated with physical understanding. Both ERA-5 and EC-Earth render a SSTA pattern that strongly resembles the main features of the PDO pattern in its negative phase (see Appendix figure D5). This is in line with the physical mechanism that low level heating can effect the position of the jet stream (Thomson and Vallis, 2018; Teng et al., 2019).

In the Atlantic, the relationship between hot days and SSTA differs between EC-earth and ERA-5. We suspect EC-Earth to suffer from biases, since model perturbation experiments have shown a reduction in precipitation due to a warm Gulf of Mexico state (Wang et al., 2010), which overlaps with the warm Caribbean region our analysis finds in the ERA-5 data (Figure 3.4, left column). The lower amount of precipitation is linked to an increase in temperature due to a stronger soil moisture-temperature feedback (Wang et al., 2010). Their analysis describes the complexity of the physical links, indicating that it is difficult to simulate the teleconnection between U.S. temperature and Atlantic SSTA. EC-Earth has to simulate the entire chain of interactions accurately to get the correct temperature impact, e.g. the circulation, cloud and precipitation response, land surface fluxes and the soil-moisture temperature feedback.

In general, we also observe that the pattern anomalies are stronger for the ERA-5 dataset,

which could be due to the sampling size of 40 years. However, we suspect it is more likely that EC-Earth is underestimating the link between SSTA and hot days, which is also supported by the lower forecast skill of EC-Earth presented in section 33.3.3. The ostensible underestimation of the atmospheric response to SSTA could be the result of unresolved smaller scale processes due to insufficient spatial resolution in climate models (Hodson et al., 2010; Van Der Linden et al., 2019; Thomson and Vallis, 2018).

McKinnon et al. (2016) proposed that the PEP pattern arises due to atmosphere-to-ocean heat fluxes in spring/summer, which are indeed directed towards strengthening of the pattern (see figure S12 in McKinnon et al. (2016)). This suggests a mechanism acting on a sub-seasonal timescale, separate from the PDO. However, using annual mean values, the cross-correlation matrix based upon ERA-5 data in figure 3.5 shows high correlation coefficients between the PEP and PDO, suggesting that PEP does not arise in a 60-day window, but is in fact, strongly related to the presence of the negative PDO phase.

We propose that the presence of the right background SSTA pattern favors the occurrence and persistence of a wavy jet stream resulting in a high pressure system over the eastern U.S., and ocean-to-atmosphere heat fluxes are likely amplifying the final response (a wavy jet stream). The correlation of the SSTA pattern with temperature is likely strongest in summer (figure 3.5) because the impact of a high pressure system on temperature is exacerbated due to the higher solar irradiation and potentially stronger soil moisture-temperature feedbacks (when the evaporation becomes strongly limited by the available soil moisture, the impact on temperature becomes most apparent, which generally happens at the end of the summer) (Seneviratne et al., 2010).

3.5 Conclusions

In this work, we focussed on (1) a comparison between the data-driven & response-guided CPPA approach, the PEP pattern and using climate indices, (2) the importance of using multiple skill metrics and (3) how one can make reliable statistical S2S forecast for high-temperature events. Firstly, we presented an algorithm that objectively extracts robust SST anomalies (SSTA) from a target event timeseries. We conclude that CPPA can successfully detect robust SSTA regions. We note that, using continuous timeseries instead of a binary one for the target variable, correlation maps appear more robust and render similar results (see discussion in Appendix C). Boschat et al. (2016) also concluded that correlation maps are more robust compared to a composite approach, although, they did not perform a sub-sampling as done by the CPPA to check for robustness.

The use of the AUC-ROC score as a single metric to assess skill should be avoided because it measures only potential skill. Based on multiple skill metrics, we showed that long-lead predictability does not exist for *individual* hot days (section 33.3.3). To generate reliable S2S forecasts, one needs to improve the signal-to-noise ratio, either by temporal aggregation, spatial aggregation or statistical filtering techniques (e.g. wavelet transformations (Deo et al. 2017)). Here, we have shown that a low signal-to-noise ratio in the target timeseries is indeed a bottleneck when trying to forecast extreme events defined on a daily resolution. Using a window probability, we were able to forecast moderate heatwaves with an average anomaly of 5.6°C above July-Aug climatology.

Forecast skill improved when using the CPPA precursor regions as compared to using

modes of variability (PDO, Nino3.4). A key advantage of this response-guided approach compared to some other feature extraction techniques, is that precursors remain physically interpretable. With this approach, one can benefit from a data-driven tool to optimize skill and also use physical understanding to e.g. identify plausible physical relationships, select variables, estimate the associated timescale of dynamics, understand limitations of predictability from physics (Mariotti et al., 2020). Hence, we recommend a response-guided approach to learn your input features for statistical forecasting models, as was also done by Kretschmer et al. (2017b), for which Python code is being developed and shared on Github². The Github release contains the code and ERA-5 timeseries to reproduce the forecasts in this manuscript.

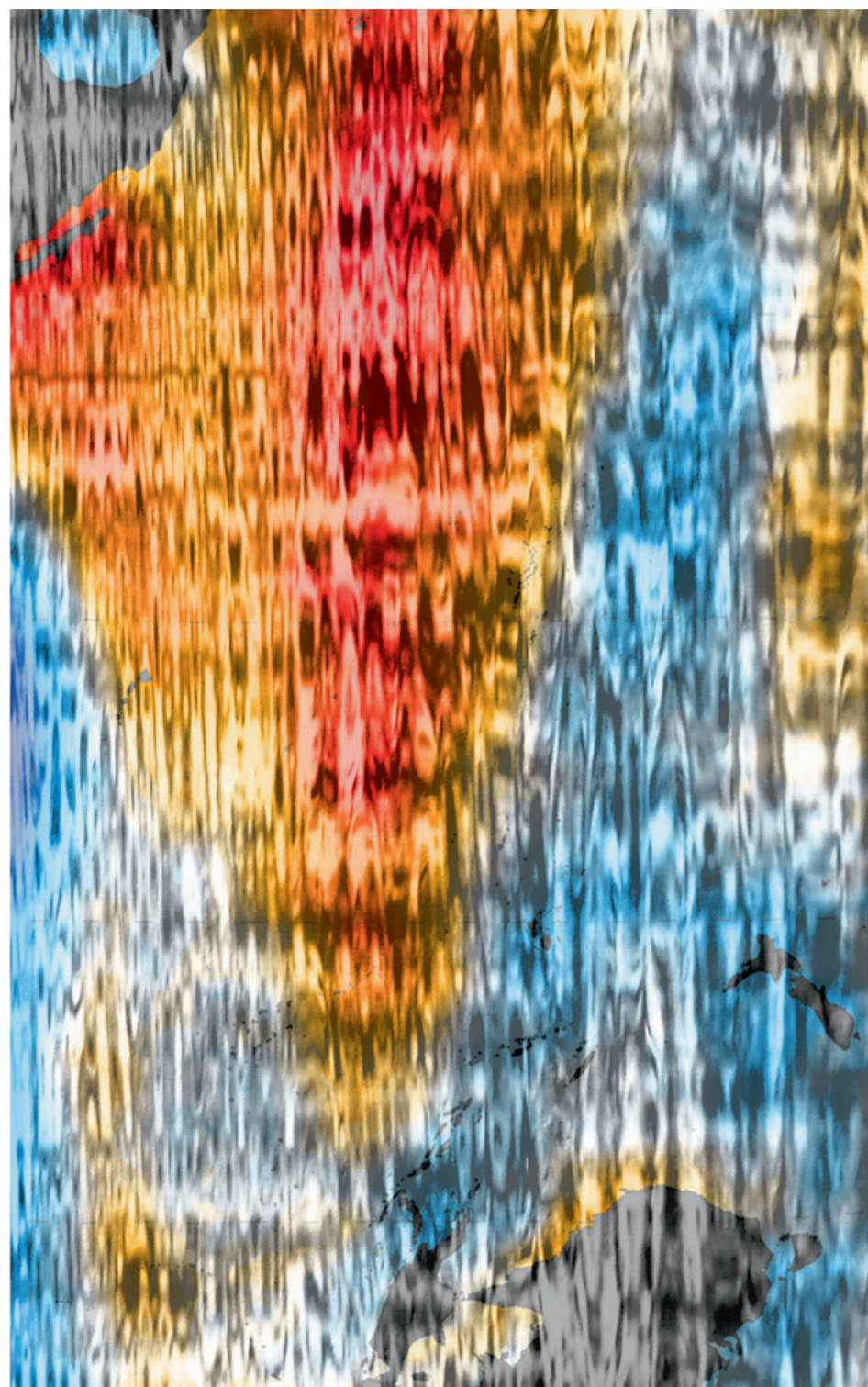
Future work could look into implementing statistical methods to obtain a better signal-to-noise ratio. Using an automated response-guided approach as presented here in combination with dynamical model output, i.e. producing hybrid forecasts, could be the next step to make operational, improved S2S forecasts.

From a physical perspective, the link between SSTA and temperature is complex and appears to be affected by (1) the soil moisture-temperature feedback, (2) ocean-to-atmosphere interaction leading to a feedback between Rossby waves and the SSTA, (3) the direct circulation response to the SSTA pattern excluding the effect of ocean-to-atmosphere feedbacks and potentially (4) a dependence of the atmospheric response on the wind field (Thomson and Vallis, 2018). The physical interaction and relative importances of these processes will be subject of future work.

3.6 Acknowledgements

Sem Vijverberg would like to thank Anais Couason for the help with phrasing the results, Bram Kraaijeveld for his earlier work on the forecasting part and the two anonymous reviewers who gave good feedback and suggestions.

²The code that was used for this work is published in a separate release (Vijverberg, 2020). For the most recent version: <https://github.com/semvijverberg/RGCPD>.



4

The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability

Heatwaves can have devastating impact on society and reliable early warnings at several weeks lead time are needed. Previous studies showed that north-Pacific sea surface temperatures (SST) can provide long-lead predictability for eastern US temperature, mediated by an atmospheric Rossby wave. The exact mechanisms, however, are not well understood. Here we analyze two different Rossby waves associated with temperature variability in western and eastern US, respectively. Causal discovery analyses reveal that both waves are characterized by positive ocean-atmosphere feedbacks at daily timescales. Only for the eastern US, a long-lead causal link from SSTs to the Rossby wave exists, which generates summer temperature predictability. We show that this SST forcing mechanism originates from the evolution of the winter-to-spring Pacific Decadal Oscillation (PDO). During pronounced winter-to- spring PDO phases (either positive or negative) eastern US summer temperature forecast skill more than doubles, providing a temporary window of enhanced long-lead predictability.

This Chapter is published as:

S. Vijverberg and D. Coumou (2022). “The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability”. In: *npj Climate and Atmospheric Science* 5.1, p. 18. DOI: [10.1038/s41612-022-00237-7](https://doi.org/10.1038/s41612-022-00237-7)

4.1 Introduction

Quasi-stationary or recurrent Rossby waves in boreal summer play an important role in the development of high impact heat waves (Wolf et al., 2018; Röthlisberger et al., 2019). Such Rossby waves create persistent clear-sky high pressure systems, which, in combination with soil desiccation and land-atmosphere feedbacks, can lead to extreme heatwaves such as seen in Russia 2010 (Lau and Kim, 2012; Petoukhov et al., 2013) and the United States 2012 (Wang et al., 2014). These Rossby waves (RW) can arise due to internal atmospheric variability, with a preferred phase (Kornhuber et al., 2020) that largely depends on orography (Hoskins and Karoly, 1981), land-ocean boundaries (Kornhuber et al., 2017b) and atmospheric waveguidability (Kornhuber et al., 2017b; Hoskins and Ambrizzi, 1993; Branstator and Teng, 2017). Vorticity anomalies induced by e.g. tropical convection or mid-latitude sea surface temperature (SST) anomalies can also force quasi-stationary Rossby waves (Hoskins and Karoly, 1981; Ding et al., 2011; Di Capua et al., 2020; Ferreira and Frankignoul, 2005).

Understanding the role of mid-latitude ocean-atmosphere interactions in generating and maintaining Rossby waves is needed to improve subseasonal-to-seasonal (S2S) predictions (Switanek et al., 2020; McKinnon et al., 2016; Vijverberg et al., 2020) and climate change projections (Simpson et al., 2014; Baker et al., 2019; Raymond et al., 2019). Currently, these interactions are not well understood. We know that, especially on intra-seasonal timescales, mid-latitude SST anomalies are predominantly forced by atmospheric variability (Frankignoul and Hasselmann, 1977; Kushnir et al., 2002), yet the ocean can also influence the atmosphere (Peng and Robinson, 2001; Frankignoul and Sennéchaël, 2007). The initial atmospheric response to diabatic heating at the ocean surface is baroclinic, with a low-level trough and high-level ridge slightly downstream of an initial warm SST anomaly (Hoskins and Karoly, 1981). Subsequently, the baroclinic response is modified to a local or slightly downwind-shifted warm ridge (barotropic) response via a transient eddy feedback (Ferreira and Frankignoul, 2005; Liu and Wu, 2004). In the upper atmosphere, the warm ridge is associated with a negative vorticity anomaly. The atmosphere responds to this negative vorticity anomaly by moving air equatorward, mainly at the downstream edge of the warm ridge. This adjustment can lead to a downstream Rossby wave response consisting of alternating highs and lows (Zhou et al., 2017).

The atmospheric response to SST anomalies is thus complicated due to the transient eddy feedback, which strongly depends on the strength of the background flow, and therefore also on season and location of the anomaly (Peng and Robinson, 2001). A stronger atmospheric response is expected when the SST anomaly is close to the storm tracks and when the storm tracks are strong (e.g. in winter) (Zhou, 2019). This sensitivity of the atmospheric response to the storm track's characteristics is also linked to the waveguidability of the jet stream (Zhou, 2019). Vorticity disturbances in the storm track near the core of the jet will be refracted to the core (Hoskins and Ambrizzi, 1993; Branstator, 2002), thereby generating a more zonally elongated Rossby wave response. A higher waveguidability is found for a strong and/or more narrow jet stream, leading to a stronger atmospheric wave-response (Manola et al., 2013). The jet and storm track are tightly coupled (Lorenz and Hartmann, 2003), and it is thus likely that both strongly affect the atmospheric response to an SST induced vorticity anomaly.

The atmospheric response also depends on the persistence of an imposed SST anomaly.

While the timescale of the baroclinic adjustment is only a few days, to reach the equilibrium barotropic adjustment takes approx. 1 to 2 months (Ferreira and Frankignoul, 2005; Deser et al., 2007). The SST persistence is governed by the oceanic Rossby wave response to atmospheric forcing (Newman et al., 2016), yet it is also affected by the thermal inertia of the ocean mixed layer and the turbulent heat fluxes (Deser et al., 2010). The mixed layer is shallower during summer, and therefore SST anomalies are less persistent (dissipating within a couple of months) compared to winter (>1 year) (Namias and Born, 1970; Deser et al., 2003). Vice versa, the shallower mixed layer also means that the persistence of summer SST is more sensitive to atmospheric forcing (Deser et al., 2003). All these factors illustrate the complexity of the coupled ocean-atmosphere Rossby wave interactions, with (1) a seasonally varying ocean-atmosphere coupling strength, (2) a seasonally varying persistence of SST, (3) the slow atmospheric baroclinic-to-barotropic adjustment to an SST anomaly, and (4) the dependence on the location of the SST anomaly (and background atmospheric state).

Here, we focus on United States (US) temperature variability and its relationship with atmospheric Rossby waves, and how these Rossby waves interact with north-Pacific SST anomalies. Previous work showed that extra-tropical Pacific SSTs, associated with a Rossby wave, provide long-lead predictability for eastern US hot temperatures (McKinnon et al., 2016). Vijverberg et al. (2020) showed that using only SST precursors to predict high temperature events renders predictive skill up to 50 days lead-time, while adding local soil moisture only slightly improves skill at shorter lead-times (up to 30 days lead-time). Thus, we focus here on interactions between SST and Rossby Waves (SST-RW), and how those interactions affect the long-lead SST signal for eastern US temperature. It was hypothesized that, in summer, amplifying two-way feedbacks between the Rossby wave and the underlying SST pattern can generate long-lead predictability (McKinnon et al., 2016). The SST pattern would initially arise as response to a strong atmospheric Rossby wave, and subsequently amplify via positive ocean-atmosphere feedbacks. An alternative hypothesis is that the long-lead SST signal predominantly originates from the Pacific Decadal Oscillation (PDO), since the robust SST precursor pattern as found by Vijverberg et al. (2020), projects strongly onto the PDO pattern. This suggests that the low-frequency PDO dynamics leads to a continuous and persistent boundary condition for the atmosphere (Vijverberg et al., 2020). Both processes might also simultaneously contribute to the long-lead signal between SST and eastern US temperature.

To test these hypotheses, we use a causal discovery technique to quantify the SST-RW coupling strength of the Rossby wave associated with eastern US temperature variability. As a comparison, we perform the same analyses for the western US and show that this region is modulated by a different Rossby wave pattern with different dynamical characteristics. Importantly, we show that only for the eastern RW, a long-lead SST signal exists, and thus long-lead predictability is possible. If the long-lead signal for the eastern RW originates from a positive SST-RW feedback, we expect to find a stronger SST-RW coupling on (sub)-synoptic timescales (1 to 10 days) compared to the western RW. On the other hand, if the ocean is forcing the atmosphere by acting as a boundary forcing, we expect to find a pronounced upward ocean-to-atmosphere link for the eastern RW.

To measure the coupling strength, lagged univariate correlation analyses are inadequate since the autocorrelation of both the Rossby wave and especially the SST variability will spuriously inflate the correlation coefficient (Runge, 2018). Therefore, we will use a

causal discovery algorithm which has been specifically developed to deal with strongly autocorrelated climate data (Runge et al., 2014).

4.2 Method

4.2.1 Data

Our analysis relies on 42 years of data (1979 - 2020) from the ERA-5 reanalysis (Copernicus Climate Change Service (C3S), 2017). Daily maximum 2-meter temperature (on a $0.25^\circ \times 0.25^\circ$ grid) is calculated by computing the daily maximum of the 'maximum 2m temperature since previous post-processing?', with a step-size of 1 hour. We use daily mean sea surface temperature (SST), geopotential height at 500 hPa (z500) and meridional wind at 300 hPa (v300), all on a 1×1 degree grid. Z500 Denotes the thickness of the atmospheric layer at 500 hPa, therefore, it clearly discriminates between high- and low-pressure systems, and it is directly affected by vortex stretching/compression that can arise due to diabatic anomalies (Holton, 2004). Meridional wind (v) at 300 hPa is less affected by lower tropospheric disturbances, therefore, v300 is often used to investigate large-scale Rossby wave patterns (Kornhuber et al., 2017a; Teng and Branstator, 2019). For the daily data, we determine the seasonal cycle by (1) applying a 25-day rolling mean, (2) calculating the multi-year mean of each day-of-year of the smoothed timeseries and (3) then construct the final seasonal cycle by fitting the first 6 annual harmonics to the calculated seasonal cycle based on the smoothed data. The 1-dimensional target timeseries for the analyses in section 4.3.1 to 4.3.3 are aggregated from pre-processed daily data to 2-month means. We also use raw monthly mean SST data as input when analyzing data on the 2-monthly timescale for computational efficiency. For the raw monthly mean SST data, we construct the seasonal cycle by calculating the multi-year mean of each month. We subtract the seasonal cycle from the daily/monthly data and remove the climate change signal by subtracting the long-term linear trend of each gridcell.

4.2.2 Clustering North American temperature events

We use Hierarchical agglomerative clustering to identify coherently behaving regions (McKinnon et al., 2016) (Figure 4.1). We apply the clustering on US gridcells and a part of Canada and Greenland (up to 70°N). Regions that tend to experience temperature above the 66th percentile simultaneously are clustered together. Because the dynamics behind temperature variability might be different at high elevation, we excluded all gridcells with an altitude above 1500 m (e.g., the Rocky Mountains). We performed the clustering for a range of temporal aggregations [5, 10, 15, 30 days] and number of clusters [4,5,6,7,8,9,10] to test for robustness. From the results presented in the Appendix 4.A, we choose the two robust clusters, to simplify notations we refer to this as the western and eastern US cluster. By testing the spatial decorrelation radius within each cluster Figure 4.A.2, we verified that the size of the clusters is appropriate. Of these two clusters an area-weighted spatial mean temperature is calculated, rendering two 1-dimensional timeseries. The timeseries convey the western and eastern US daily maximum temperature variability, these are referred to by T^W and T^E , respectively.

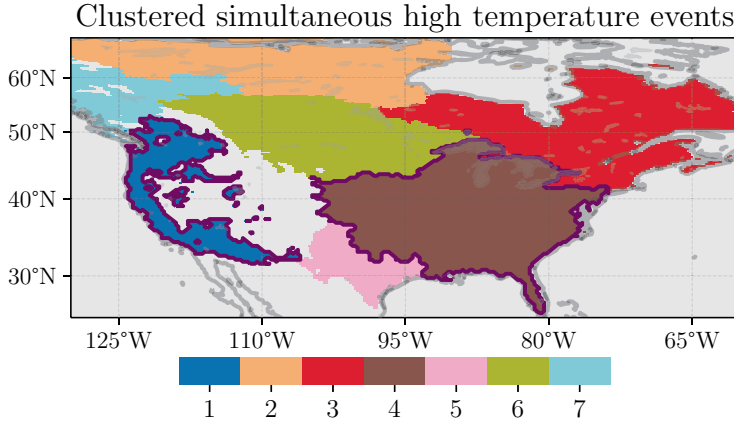


Figure 4.1: Gridcells which show more frequent simultaneous warm temperature periods occurrences are combined using Hierarchical agglomerative clustering. Warm temperature periods are defined as 15-day mean temperature exceeding the 66th percentile. The white gridcells indicate (the Rocky) mountains (altitude > 1500 metre). These were left out of the analysis because temperature variability at high altitudes might have a different relationship with Rossby wave variability compared to low altitude gridcells. The purple contour lines indicate the western and eastern US spatial clusters.

4.2.3 Link between temperature, circulation, and sea surface temperature

To quantify the temperature versus z500 relationship, we aggregated to 15-day means and calculate one-point correlation maps ($\alpha_{\text{FDR}}=0.05$) at lag 0 for both the west (T^W) and eastern US (T^E) temperature timeseries. We account for the False Discovery Rate using the Benjamini/Hochberg correction (Benjamini and Hochberg, 1995; Wilks, Dan, 2016). For Figure 4.2 and Figure 4.3a and g, we test for robustness of the correlation maps by re-calculating them on 70 subsets (36 years) sampled from the 42 years of data. See section 5.2.7 for more information. In the one-point correlation maps, gridcells are only presented as significant if they are found significant in 60/70 subsets of data. The RW pattern ($\text{RW}_{\text{pattern}}$) is defined by the significantly correlating gridcells within the green rectangle as shown in the z500 correlation maps (Figure 4.2a and b). We reduce it to a 1-dimensional timeseries by calculating the area-weighted spatial covariance, i.e.,

$$\text{RW}(t) = \frac{1}{N} \sum_i^N w_i \left\{ \left[\text{RW}_{\text{pattern}}(t, i) - \overline{\text{RW}_{\text{pattern}}(t)} \right] \cdot \left[z(t, i) - \overline{z(t)} \right] \right\} \quad (4.1)$$

using only the N significantly correlating grid cells. Where w_i denotes the area weight at grid cell i , $\text{RW}_{\text{pattern}}$ denotes a vector with the correlation values of the significantly correlating grid cells, the overbar denotes the spatial mean and $z(t)$ denotes the geopotential height field at time t . Temperature correlates strongest with local geopotential height. The higher correlation values result in a much stronger weight for the local high-pressure system compared to adjacent lows and highs of the RW pattern. To obtain a RW timeseries where the high and lows have equal weights, we set all significant positively (negatively)

correlating gridcells to 1 (-1). This is done only for the RW timeseries. We tested other options, but this method led to a timeseries that was best capable to reproduce the RW pattern that the timeseries is supposed to capture (section 4.B). We use this procedure to calculate both the west (RW_t^W) and eastern US (RW_t^E) RW timeseries, which we will use for the PCMCI and (partial) correlation analysis.

4.2.4 Causal Effect Network using PCMCI

To obtain the link between the RW timeseries and north-Pacific SST, we first calculate one-point correlation maps with SSTA versus both the west and eastern RW timeseries (Eq. 4.1). These correlation maps show the RW imprint on the SSTA. Substituting RW_{pattern} for SST_{pattern} and z for SST in Eq. 4.1, we obtain a 1-dimensional timeseries for the SST pattern. At this point, we have the SST pattern timeseries and the RW pattern timeseries, i.e. (SST_t^W and RW_t^W) and (SST_t^E and RW_t^E).

To quantify the SST-RW coupling strength, we use the PCMCI algorithm (Runge et al., 2019) in combination with conditional independence (CI) tests based on partial correlation (Runge et al., 2015). For each significantly correlating link, a partial correlation analysis is performed which is conditioning on all relevant information that might statistically inflate the correlation link strength.

The relevant information is found by step (1) of the PCMCI algorithm, which attempts to find the parents (\mathcal{P}) of both to-be-tested variables through an iterative process of CI tests. In this case, we only have two variables (x_t^i and x_t^j). For example, the first parent subset $\mathcal{P}^0(x_t^i)$ consists of all possible lagged correlations (the lag is indicated by τ) with $p_{\text{value}} < 0.05$. The maximum lag for lagged correlations is restricted by the parameter τ_{max} . Next, each timeseries in subset $\mathcal{P}^0(x_t^i)$ is only kept if it passes all partial correlations tests (e.g. $\text{parcorr}(x_{t-\tau}^j, x_t^i | z)$), where z is a single variable of the subset \mathcal{P}^0 that is not $x_{t-\tau}^j$. All variables that are conditionally dependent are stored in the second subset $\mathcal{P}^1(x_t^i)$. Next, all possible CI tests are performed again with the cardinality of z increased from 1 to 2. Using our settings, the cardinality may increase up to the total size (minus one) of the first parent subset $\mathcal{P}^0(x_t^i)$. Once the final sets of parents are estimated, i.e., $\widehat{\mathcal{P}}(x_t^i)$ and $\widehat{\mathcal{P}}(x_t^j)$, step (2) of PCMCI calculates the Momentary Conditional Information (MCI), which is using partial correlation and is defined as,

$$\text{parcorr} \left(x_{t-\tau}^j, x_t^i \mid \left\{ \widehat{\mathcal{P}}(x_t^i) \setminus x_{t-\tau}^j, \widehat{\mathcal{P}}(x_{t-\tau}^j) \right\} \right) \quad (4.2)$$

where $\widehat{\mathcal{P}}(x_t^i) \setminus x_{t-\tau}^j$ are the estimated parents of x_t^i , excluding the to-be-tested variable $x_{t-\tau}^j$. All links are tested in both directions, as well as instantaneous, i.e., from $\tau_{\text{min}} = 0$ up to $\tau_{\text{max}} = \tau_{\text{max}}$. If the MCI is significant ($\alpha_{\text{FDR}}=0.05$) when conditioning on all the parents and assuming the underlying assumptions are satisfied (Runge, 2018), the link is deemed causal. When we state there is a causal link, it should be interpreted as causal within the context of the experiment, i.e., not the result of a spurious link due to the past SST evolution or RW occurrences, with the past (i.e., maximum lag considered) being limited by τ_{max} .

Sensitivity analyses are performed by re-iterating the analysis workflow, i.e., from calculating the RW pattern and timeseries (section 4.2.3) up to the CEN, each time using a

unique set of 36 out of the 42 years of data. Since we apply PCMCI repeatedly on different subsets of data and PCMCI tests many different dependency tests within the algorithm, we correct for the False Discovery Rate (FDR) using the Benjamin/Hochberg correction. With these sensitivity analyses, we are propagating uncertainties due to leaving out data through the entire workflow. Similar types of robustness/stability tests are becoming more common in the machine learning community (Szegedy et al., 2013; Belgiu and Drăgu, 2016). The results of the sensitivity analyses are also used to quantitatively compare the western vs the eastern CENs in Table 4.1.

4.2.5 Partial correlation maps

We use the partial correlation conditional independence tests to construct latitude, longitude maps where we test the influence of a potential confounder of interest. We use these maps to regress out the influence of the RW at lag 2, the low-frequency ENSO and the low-frequency PDO timeseries when testing the link between SST_{t-1} and RW_t^E . The low-frequency variability is obtained by apply a 6-month rolling mean (indicated by \overline{ENSO}_t and \overline{PDO}_t). When selecting the dates at lag 2 we approx. select the winter-to-spring mean timeseries. We ensure that the rolling mean is based on data prior and including lag 2 to avoid information leakage. The ENSO is calculated using the area-weighted nino3.4 bounding box [5°S - 5°N, 170°E - 120°W]. The PDO pattern is found by calculating the first area-weighted Empirical Orthogonal Function of Pacific SSTA [115-250°E, 20-70°N]. For the (partial) correlation maps in Figure 4.5, we use the same cross-validation as introduced in section 5.2.7 to obtain different subsets of data. Hence, the partial correlation maps are calculated 10 times on subsets of 36 years.

4.2.6 Forecasting

To investigate the seasonal dependence, we use a response-guided approach (Vijverberg et al., 2020; Kretschmer et al., 2017a; Bello et al., 2015; Di Capua et al., 2019a). This approach encompasses methods that reduce dimensionality of the precursor field based on a relationship to a target, instead of using some statistic of the precursor field (e.g. maximizing the explained variance). First, we calculate one-point correlation maps based on training data (at lag 1). Secondly, adjacent regions of the same correlation sign are grouped together into precursor regions. This is done using the Density-based spatial clustering of applications with noise (DBSCAN) (Schubert et al., 2017). Thirdly, for each precursor region, an area-weighted and correlation-value weighted spatial mean is calculated.

The resulting 1-dimensional timeseries are standardized and then fitted on the training data using a Ridge regression. The regularization parameter λ is tuned using the default Generalized Cross-Validation (Varoquaux et al., 2015). The alphas range between .1 and 1.5, with 25 steps spaced evenly on a log scale with base=10. The standardizing and fitting are done on the same training data as is used to calculate the correlation maps.

We use the Pearson-r correlation coefficient and mean absolute error skill score (MAE-SS) for verifying the deterministic forecasts. The MAE gives equal weights to each observation/forecast pair, making the analysis we present in Figure 4.8 a fairer comparison between the two data subsets. It is defined as, $MAE = \frac{1}{n} \sum_{i=1}^n |y_{pred,t} - y_{true,t}|$, where n

are the number of observation/forecast pairs, $y_{pred,t}$ is the predicted value at timestep t and $y_{true,t}$ is the observation at timestep t .

We implement a stratified 10-fold cross-validation (training sets consist of 36 or 35 yrs, test sets 4 or 5 yrs). The stratification is achieved by creating the training sets, such that these contain similar statistics in terms of the magnitude of July-August temperature values. This ensures that the training/test sets are good approximations of the climatological US temperature dynamics. Since we cannot reliably estimate the skill score based on a single test set of 4 years. We implement a double cross-validation for tuning the regularization parameter within each training sample, as done in (Vijverberg et al., 2020). This means that we fit (and tune) a statistical model on each training set and use that to forecast the test set. The verification metrics are computed on the 10 concatenated test sets.

4.3 Results

4.2a-b shows that western (T^W) and eastern (T^E) US summer temperatures strongly correlate with two distinct Rossby wave patterns, here called the western (RW^W) and eastern RW (RW^E) pattern. These are phase-shifted by about half a wavelength with respect to each other. T^W and T^E are the area-weighted 15-day mean anomaly temperature timeseries of the western and eastern US spatial cluster, respectively. The western and eastern US temperature clusters are based on gridcells that tend to show simultaneous occurrences of warm temperature periods (Vijverberg et al., 2020; McKinnon et al., 2016), see section 5.2.2.

The western RW pattern is more zonally elongated and resembles a dominant Northern Hemispheric mode of variability. The eastern RW consists of an arcing pattern over the Pacific and North America. This wave-pattern is reminiscent of the winter Pacific North American (PNA) pattern in its negative phase (Liu et al., 2015) and the ENSO-forced atmospheric bridge response Lopez and Kirtman, 2019. Interestingly, it does not resemble the summer PNA pattern, and it does not appear to be related to a circumglobal mode of variability. See Appendix 4.C for a more detailed discussion and evidence. Hence, while the RW^W clearly relates to an atmospheric mode of variability, the RW^E does not appear to match any summer mode of variability.

Figure 4.2e-f show there is a strong instantaneous and lagged SST correlation with eastern US temperature (T^E). For the west (T^W), no long-lead signal SST signal is detected (Figure 4.2d), only an instantaneous one (Figure 4.2c). To investigate the role of ocean-atmosphere feedbacks, we quantify the coupling strength between SST and the western and eastern Rossby waves, respectively, using the Peter and Clark - Momentary Conditional Information (PCMCI) algorithm. To visualize the causal dependencies found by PCMCI, we plot Causal Effect Networks (CEN), which are directed network graphs.

By calculating the spatial covariance of the RW patterns within the green rectangles (shown in Figure 4.2a and Figure 4.2b), we quantify timeseries that capture the RW variability for both the western and eastern RW, referred to as RW_t^W and RW_t^E (Method section 4.2.3). Figure 4.3a and g show the SST correlation with the RW_t^W and RW_t^E timeseries, respectively. By calculating the spatial covariance within the green rectangle of these SST correlation patterns, we capture the SST variability (SST_t^W and SST_t^E) associated with the two Rossby waves (Method section 4.2.4). We use the PCMCI algorithm that

consists of two-steps: (1) an adaptation of the PC (Sprites et al., 2001) (Peter and Clark) algorithm and (2) the Momentary Conditional Information metric (Runge et al., 2012). If one is interested to test the causal relationship between the timeseries x_{t-1} and y_t , first, the PC-step estimates the lagged parents of both timeseries (x_{t-1} and y_t) by iteratively performing conditional independence tests. The MCI-step tests if x_{t-1} and y_t are conditionally independent, given the influence of the lagged parents of both x_{t-1} and y_t . To measure conditional independence, we use partial correlation analyses, e.g. $\text{parcorr}(x_{t-1}, y_t | Z)$, in which Z is a single or set of timeseries to condition on. For more information, see Method section 4.2.4. To measure the SST-RW feedback on (sub)-synoptic timescales we perform the PCMCI analysis on 1, 5, 10 and 15 day means. To measure the effect of lower-frequency variability, we aggregate to 30, 45, 60 day mean timescales. For conciseness, we present the CENs for 1, 5, 10, 30 and 60 day means, as for the other timescales (15 and 45 day mean) we only found instantaneous links (not informing us about the directionality of the forcing). Note that we focus on summer (June, July, August), yet when using 60-day means, we extend into May, June, July, and August to increase sample-size (2 datapoints per year instead of 1).

Since a CEN depicts causal links as a yes/no answer (dependent on the significance threshold used), we implement a sensitivity analysis to ensure we present robust links only. We do this by repeating the PCMCI analysis 70 times on slightly different subsets of data (Method section 4.2.4). For the CENs, a link is only shown if it is significant ($\alpha_{\text{FDR}}=0.05$) in 60 out of 70 perturbation experiments. The spread in the link strength that results from the sensitivity experiments also represents the uncertainty due to sampling. Given this spread, a double-sided t-test is used to measure if the western SST-RW link strengths ($n=70$ perturbations) are significantly different ($\alpha = 0.05$) from the eastern SST-RW coupling ($n=70$ perturbations), indicated with a * in Table 4.1.

The CENs for the west and east both show a positive two-way coupling between SST and the Rossby waves on daily timescales, but there are important differences (Figure 4.3). The causal influence of the atmosphere on the ocean is more pronounced for the western US Rossby wave as compared to the eastern one. On the daily and 5-day timescale, the downward causal link for the west ($\text{RW}^W \rightarrow \text{SST}^W$) is stronger by about a factor 2 compared to the east (Table 4.1). Similarly, the instantaneous RW – SST link on these timescales is also stronger for the western RW (again by about a factor 2). On longer timescales than 5-day means, no robust directed links for the western RW are found (only instantaneous links).

On timescales longer than 5-day means, the influence of the ocean to the atmosphere is consistently stronger for the eastern Rossby wave compared to the western one. Using 60-day aggregated data, the 60-day lagged causal link $\text{RW}^E \rightarrow \text{SST}^E$ is stronger by roughly a factor 10 (Table 4.1). This 60-day aggregated $\text{SST}_{t-1}^E \rightarrow \text{RW}_t^E$ link is very robust and found to be causal by PCMCI in all the 70 perturbation experiments.

Figure 4.3 also shows that the SST_{t-1}^E timeseries has a higher persistence compared to SST_{t-1}^W , as indicated by the higher auto-strength values (colour of SST nodes). The auto-strength value is similar to the conventional autocorrelation, but accounts for factors that might artificially inflate it. For Figure 4.3f and 4.3l, the auto-strength is calculated by the partial correlation of SST_t versus SST_{t-1} , conditioned on SST_{t-2} ($\text{parcorr}(\text{SST}_t, \text{SST}_{t-1} | \text{SST}_{t-2})$).

Note that for the eastern RW, for timescales longer than daily, there are no causal links

from the atmosphere to the ocean. More precisely, the $\text{parcorr}(\text{RW}_{t-1}^E, \text{SST}_t^E | \text{SST}_{t-1}^E)$ link is non-significant. From this, it can be deduced that the higher persistence of SST_t^E in the first place results from the past SST state (SST_{t-1}^E) with only a minor influence of antecedent atmospheric forcing (RW_{t-1}^E).

To get a better understanding on the interaction between RW^E and SST prior to the summer, we calculate the ocean-atmosphere coupling between SST and the eastern RW for winter and spring (see Appendix 4.D). To ensure that we focus on the same RW pattern, we project the summer eastern RW pattern (as defined Figure 4.2b) onto the winter and spring z500 field. Figure 4.B.1 verifies how the Rossby wave timeseries correlates with z500 variability. For JJA we retrieve a pattern very similar to what is shown in Figure 4.2b, confirming that our RW_t^E index is a good proxy for the eastern Rossby wave. In winter and spring, the RW_t^E index projects strongly on the same eastern Rossby wave and additionally correlates with the tropical belt and is again similar to the Pacific-North-American pattern and the ENSO-forced teleconnection called the atmospheric bridge, which starts above the ENSO region and arcs over the Pacific-North American domain (Alexander et al., 2002). In Figure 4.D.2 the correlation maps in winter (panel a) and spring (panel g) between SST and the RW_t^E index show a clear resemblance to main features of the PDO pattern in its negative phase. The CENs in panels b-f show that during winter, we find a strong downward forcing (atmosphere-to-ocean). For spring, in addition to a strong downward forcing, we also observe two-way coupling on the 5-day mean timescale (panels h-l). In the next section, we use partial correlation to further investigate the importance of different processes in winter and spring for the upward (ocean-to-atmosphere) forcing that we find in summer (Figure 4.3l).

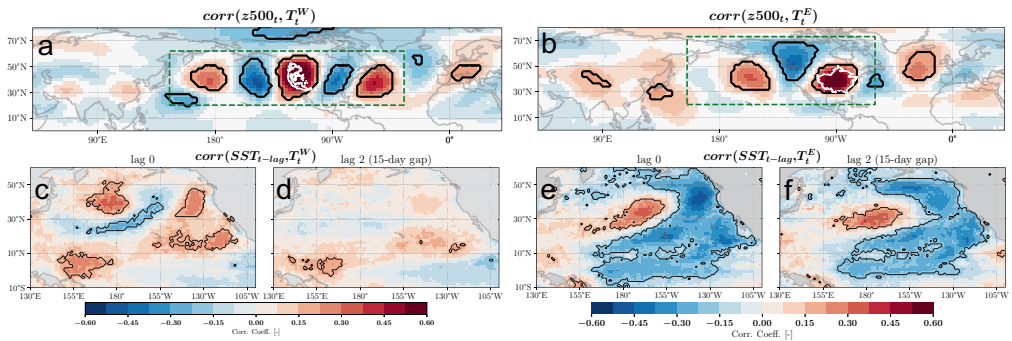


Figure 4.2: Correlation maps for different variables against western (T^W) and eastern (T^E) US temperature variability. Correlation map between geopotential height at 500 hPa (z500) and T^W (panel a) and T^E (panel b). Panel c, d, e and f are similar to the upper panels, but versus SSTA, showing instantaneous and lag 2 correlation values. Based on 15-day mean data. For the significance ($\alpha_{\text{FDR}}=0.05$), we correct for the False Discovery Rate using the Benjamini/Hochberg correction. Gridcells are highlighted by black contour lines if they are significant at least 60 out of 70 data subsets (Method section 4.2.4). The white contour line indicates the western US cluster (panel a) and eastern US cluster (panel b). The green rectangles in a and b indicates the region that is used to calculate the spatial covariance, see section 4.2.3.

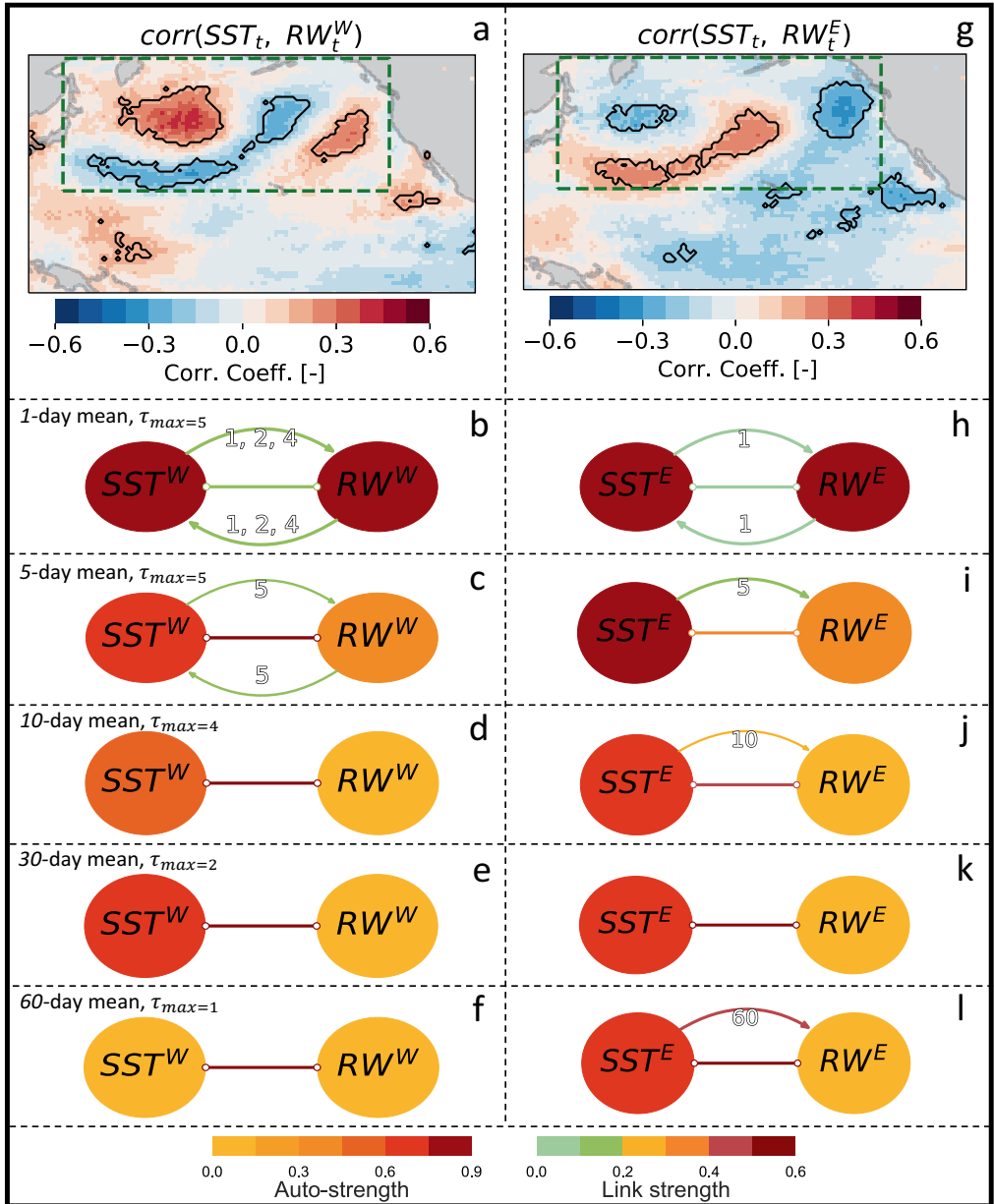


Figure 4.3: Quantifying SST-RW coupling at different timescales in summer (JJA). Panel a (b): lag 0 correlation maps of SST versus the western (eastern) RW timeseries. Panels a-f and h-l: CENs between the respective SST pattern timeseries and the RW pattern timeseries for different temporal aggregations [1, 5, 10, 15, and 60]. The Link strength (link colour) shows the MCI value (mean over significant links), which is the correlation strength, after removing the information of the parents of both variables. The Auto-strength (node colour) shows the autocorrelation after regressing out the influence of its parents. The link labels indicate at which lags [in days] there was a causal link. CEN link is only plotted if they are significant at least 60 out of 70 perturbation experiments, similar for the indication of significance in panel a and g by black contour lines.

Table 4.1: Comparing SST-RW coupling strength in summer by the mean ratio of the Momentary Conditional Information (MCI) values $\mu_{MCI_{west}}/\mu_{MCI_{east}}$.

Direction of link	$\frac{\mu_{MCI_{west}}}{\mu_{MCI_{east}}}$ 1-day mean	$\frac{\mu_{MCI_{west}}}{\mu_{MCI_{east}}}$ 5-day mean	$\frac{\mu_{MCI_{west}}}{\mu_{MCI_{east}}}$ 15-day mean	$\frac{\mu_{MCI_{west}}}{\mu_{MCI_{east}}}$ 60-day mean
Instantaneous link	2,1*	1,7*	1,4*	1,4*
Rossby wave to SST forcing	1,6*	2,8*	2,0*	1,3*
SST to Rossby wave forcing	1,6*	0,8*	0,8*	0,1*

The * indicates a significant ($\alpha_{FDR}=0.05$) difference given the uncertainty due to sampling (see "Results").

4.3.1 Explaining the long-lead causal link

Here we show that the long-lead upward ocean-forcing that drives the eastern US Rossby wave in summer (as identified above), is closely related to low-frequency PDO variability. From here on, we work with 2-month mean data (instead of 60-day means) to ease interpretation. On the 2-month mean timescale, eastern RW correlation pattern is clearly in phase with the PDO pattern, while the western RW is not. This can be seen in Figure 4.4 which shows the PDO (1st EOF loading) pattern together with the instantaneous and lag-1 (corresponding to a 2-months lag) correlations maps between SSTs and the western and eastern RW, respectively.

In Figure 4.5, we investigate how the intra-seasonal evolution of the eastern Rossby wave, ENSO and PDO (at lag 2) affect the SSTA signal at lag 1. We do so by creating lag-1 correlation maps that are conditioned on different actors at lag 2. Figure 4.5a shows the SST – RW^E correlation pattern at lag 1 (2 months). The correlation values and their significance are reduced when conditioning on RW^E at lag 2 (Figure 4.5b), implying that the Rossby wave activity at lag 2 plays some role in forcing the SST signal at lag 1. We calculate the low-frequency (winter-to-spring mean) ($\overline{ENSO_{t-2}}$) and ($\overline{PDO_{t-2}}$) timeseries as described in Method section 4.2.5. When conditioning on $\overline{ENSO_{t-2}}$, the SST signal does not weaken indicating that the mean ENSO state has little or no effect (Figure 4.5c). Figure 4.5d shows the SST_{t-1} influence on RW^E_t is most effectively weakened when conditioning on the winter-to-spring PDO variability ($\overline{PDO_{t-2}}$). Thus, most of the information originates from winter-to-spring PDO variability. These results show that the lagged SST signal - relevant for forcing the May-August eastern RW (RW^E_t) - is influenced in the first place by the winter-to-spring PDO state. An additional (but smaller) influence is provided by the prior atmospheric wave forcing. That they both show an influence is in line with the strong co-variability between the PDO and RW^E in spring (Figure 4.C.1).

4.3.2 Temperature predictability

A robust lagged Pacific SST signal can only be found for temperature in the eastern US cluster, but not for the west (Figure 4.6). This is in line with the lack of any long-lead causal links to the western RW (Figure 4.3) or any significant lagged SST correlation for western US temperature variability (Figure 4.2d). For the eastern US, the lagged SST signal is clearly strongest for July-August (Figure 4.6). The correlating regions (Figure 4.6, July-August mean) are clustered into the mid- and eastern Pacific region (Figure 4.7), these two regions are used as a mask to calculate 1-dimensional spatial mean timeseries (Method section 5.2.7). These timeseries will be used for predictions in section 4.3.3. This method is referred as the response-guided dimensionality reduction (DR), i.e., the dimensionality reduction is based on precursor regions that correlate with the target

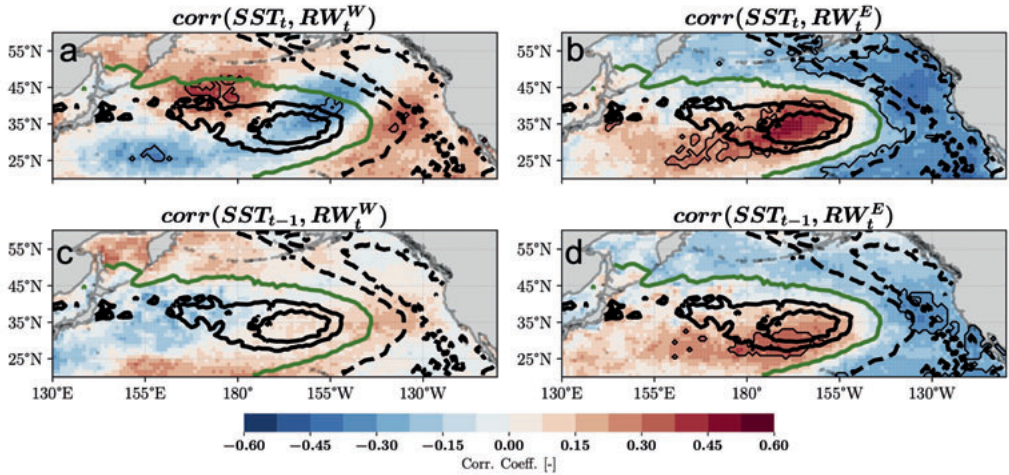


Figure 4.4: Instantaneous and lag 1 correlation maps of SSTA versus the May-August Rossby wave timeseries. Panel a and c show the western RW, panel b and d show the eastern RW. The Rossby wave patterns are still the as depicted in Figure 4.2, yet the data is now aggregated to 2-month means. For the significance ($\alpha_{\text{FDR}}=0.05$), we correct for the False Discovery Rate using the Benjamini/Hochberg correction. Gridcells are highlighted by the thin black contour lines if they are significant at least 5 out of 10 training subsets. The thick contour lines indicate the negative PDO pattern (1st EOF loading pattern) ranging from -0.7 (black dashed) to 0 (green solid) to 0.7 (black solid).

variable (Kretschmer et al., 2017a; Vijverberg et al., 2020)).

We make out-of-sample Ridge Regression forecasts for July-August mean T^E using two different precursor sets, i.e., (a) precursors found by the response-guided DR and (b) the PDO climate index timeseries (Figure 4.7). The test data is obtained by splitting the data into training and test years using a 10-fold stratified cross-validation (see Method section 5.2.7). Both DR methods are done on only training data and extrapolated to the test set to make predictions. We use the correlation coefficient and the Mean Absolute Error Skill Score (MAE-SS) for the verification. The latter is defined as $\text{MAE-SS} = 1 - \frac{\text{MAE}_{\text{forecast}}}{\text{MAE}_{\text{clim}}}$, where MAE_{clim} is the error when always predicting the climatological mean of the T^E anomalies (≈ 0).

The response-guided DR leads to better forecast skill compared to using the PDO as a predictor (Table 4.2 and Discussion for more details). We construct Ridge Regressions using lag 1, lag 1 and 2, and lag 1, 2 and 3 (Table 4.2). We observe that using lag 1 and 2 (May-June and March-April) and the Response-guided DR provides the best predictive skill for July-August temperatures in eastern US (out-of-sample correlation value of 0.62). Hence, although there is a clear link to PDO variability, using a more tailored method to extract the signal renders a clear boost in skill (see Discussion).

4.3.3 Window of opportunity by winter-to-spring PDO state

July-August mean eastern US temperature are substantially more predictable when the forecasted summer is preceded by a strong (instead of a weak) winter-to-spring PDO state

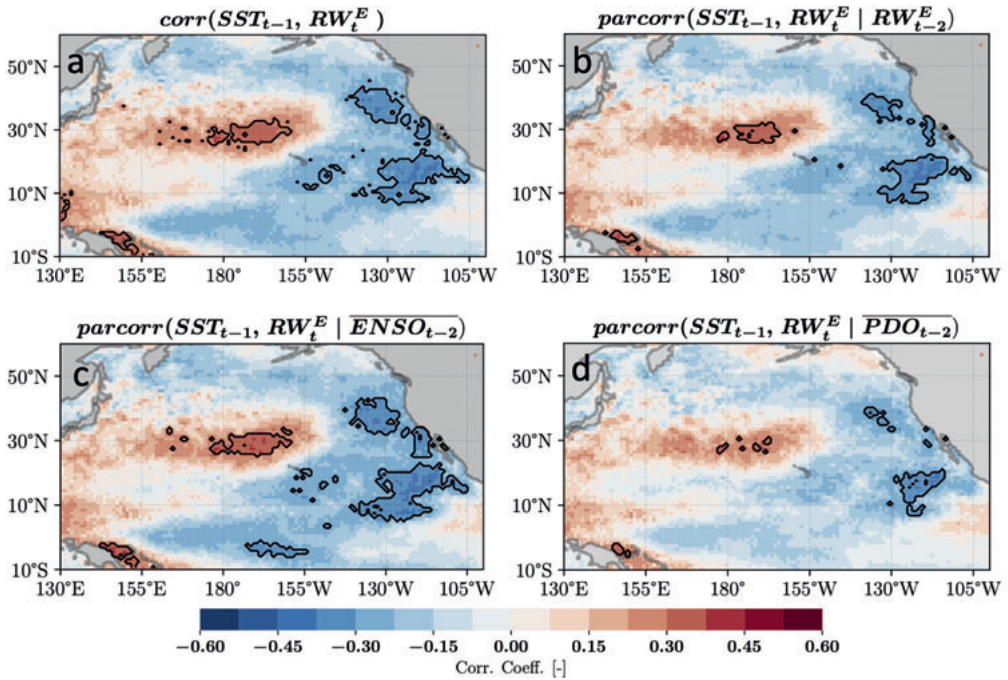


Figure 4.5: (partial) correlation map between SST_{t-1} and RW_t^E using 2-month mean data with RW_t^E defined from May to August. Panel a shows the correlation map between SST_{t-1} and RW_t^E , whereas panel b, c and d show the partial correlation maps that are removing the effect of (b) the RW_{t-2}^E timeseries, the 6-month rolling mean (c) ENSO and (d) PDO timeseries. The rolling mean is defined at (and prior to) lag 2. Gridcells are highlighted by contour lines if they are significant ($\alpha_{FDR}=0.05$) at least 5 out of 10 training subsets.

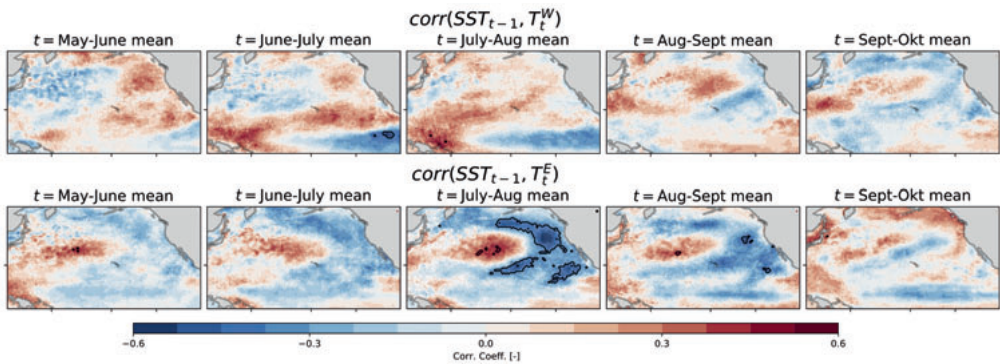


Figure 4.6: Correlation maps of SST at lag 1 versus western (T^W) and eastern (T^E) US temperature as function of target months, using 2-month mean data. Contour lines indicate significantly ($\alpha_{FDR} = 0.05$) correlating gridcells in 9/10 training subsets.

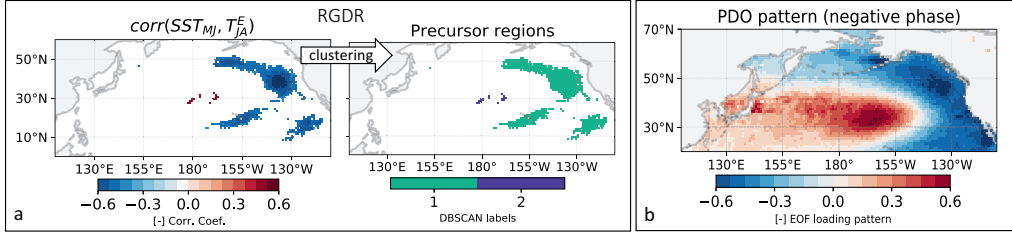


Figure 4.7: Two dimensionality reduction approaches to extract precursor timeseries from the north Pacific SST; (a) Response guided dimensionality reduction (RGDR) and (b) (climate index) PDO timeseries. For the RGDR method, the same lag 1 (May-June SST) correlation analysis is done as shown for the T^E July-August temperature in Figure 4.6. For the EOF analysis, all months (Jan-Dec) are used.

(Figure 4.8). We demonstrate this by comparing the skill during the 50% of years with strongest winter-to-spring PDO states (either positive or negative) with the skill of the 50% with weak PDO states. The winter-to-spring PDO state is defined by the DJFMAM mean PDO. During strong positive or negative PDO states, there is a 54% reduction in the MAE compared to years with weak PDO states (Figure 4.8, left column). When comparing the forecast skill to the climatological benchmark (MAE-SS), we observe that most skill is present in these strong PDO state years (Figure 4.8, right column). This result is robust when using different train-test splits. If we make a stricter selection of anomalous winter-to-spring PDO states (top 30%), the skill further increases with MAE-SS values ranging between 0,48 and 0,57 and correlation values ranging between 0,85 and 0,89 (Table 4.F.1). During weak PDO state years, the model hardly outperforms a climatological mean temperature forecast.

Similarly, using partial correlation to remove the $(\overline{\text{PDO}}_{t-2})$ from the lag 1 SST timeseries and the $(\overline{\text{PDO}}_{t-3})$ from the lag 2 SST timeseries causes the forecast skill to vanish (mean MAE-SS = $0.03_{-0.23}^{0.20}$, with the lower and upper subscript denoting the C.I. at $\alpha = 0.05$). Once more this indicates that the low-frequency antecedent PDO evolution is the background mechanism that is vital for predictability and that it can be used to identify a window of opportunity at the time of the forecast.

Table 4.2: Verification of July-August (JA) mean eastern US temperature predictions using Ridge Regression.

Dimensionality reduction method (lags used)	Corr. coeff.	MAE-SS
Response-guided (lag 1)	0,52	0,11
Response-guided (lag 1 and 2)	0,62	0,19
Response-guided (Lag 1, 2 and 3)	0,56	0,16
PDO (lag 1)	0,32	0,04
PDO (lag 1 and 2)	0,28	0,01
PDO (lag 1, 2 and 3)	0,21	0,01

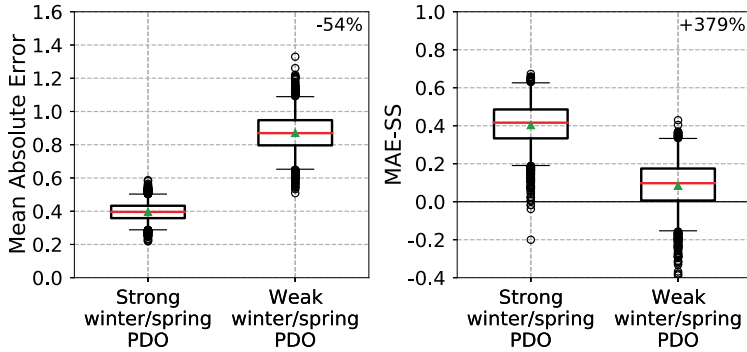


Figure 4.8: Boxplots of the bootstrapped ($n=2000$) Mean Absolute Error (MAE) and the MAE-skill score (MAE-SS) calculated for two different sub-sets. The strong PDO sub-set contains 21 (50%) years with the most anomalous DJFMAM mean PDO states. The weak PDO sub-set contains the other 21 years. The whiskers indicate the 95% confidence intervals, data outside the confidence interval are shown as outliers, red line shows the median, black line shows the quartiles, and the green triangle shows the mean.

4.4 Discussion

We show that two different Rossby waves are important drivers of temperature variability in western and eastern US, respectively (Figure 4.2a-b). While both Rossby waves correlate equally strong with surface temperatures over the US on synoptic timescales (15-day means), a long-lead signal between temperature and SST is only present for the eastern US (Figure 4.2c-f). As hypothesized in the introduction, the CEN analyses confirms that the associated summer eastern RW is forced by the low-frequency north-Pacific SST variability (Figure 4.3).

We show that low-frequency PDO variability is a crucial aspect for this long-lead signal, and thus for predictability (section 4.3.1 and 4.3.3). In our view, the mid- and eastern Pacific timeseries are the direct causal precursors, while the antecedent low-frequency PDO dynamics are vital to develop the persistent and high amplitude signal that is needed to force a persistent RW-like response in summer. This is in line with modelling experiments which show that a persistent [order of 2 months] SST forcing is needed for a barotropic (RW-like) response to develop and that a stronger boundary forcing (higher amplitude SSTA) results in a stronger atmospheric response (Ferreira and Frankignoul, 2005). To first order, the PDO pattern arises from extra-tropical atmospheric forcing and the corresponding oceanic Rossby wave response (Zhang and Delworth, 2015; Liu and Di Lorenzo, 2018). This downward forcing is strongest in winter (Newman et al., 2016), as is also observed in our winter CEN (Appendix 4.D). Multiple processes are important for strengthening the PDO variability, such as (1) the re-emergence mechanism (Deser et al., 2010), (2) the ENSO teleconnection named 'the atmospheric bridge' (Alexander et al., 2002; Newman et al., 2003; Lau and Nath, 2001) and (3) active ocean-atmosphere coupling (Zhang and Delworth, 2015; Lau and Nath, 2001) and the associated local positive feedbacks (Luo et al., 2020). However, the relative importance of these processes is uncertain. Our CEN analysis quantifying the SST-RW coupling for winter and spring indeed support that processes (2) and (3) strengthen the PDO pattern (Supplementary

Note 2). The CENs show that forcing in winter and spring is predominantly downward (from atmosphere to ocean). In spring, we also observe a more pronounced two-way feedback. Finally, the forcing is predominantly upward in summer. This is consistent with observational findings (their Fig. 3) (Kushnir et al., 2002) and previous work (Frankignoul and Sennéchaël, 2007; Liu et al., 2006).

Since persistence is a requirement to get a clear barotropic RW response, the spatial patterns of any low-frequency mode of SST variability will provide a physical constraint on the location and phase-position of quasi-stationary Rossby waves and therefore on the downstream surface impact of the Rossby waves. Results for the western RW supports the latter hypothesis, as its wave pattern is not in phase with the PDO pattern (Figure 4.4). We argue that this is the reason why the western RW is not forced by the north-Pacific SSTs at longer timescales (Figure 4.3), and therefore no long-lead SST signal is found for western US temperature (Figure 4.2d). In contrast, the eastern RW is in phase with the PDO pattern (Figure 4.4) resulting in a long-lead SST signal that forces the atmosphere (Figure 4.3). The persistent SST forcing originates from the co-evolution of winter-to-spring PDO dynamics and the associated ocean-atmosphere interactions. Hence, these are the key process behind predictability for the eastern US summer temperature.

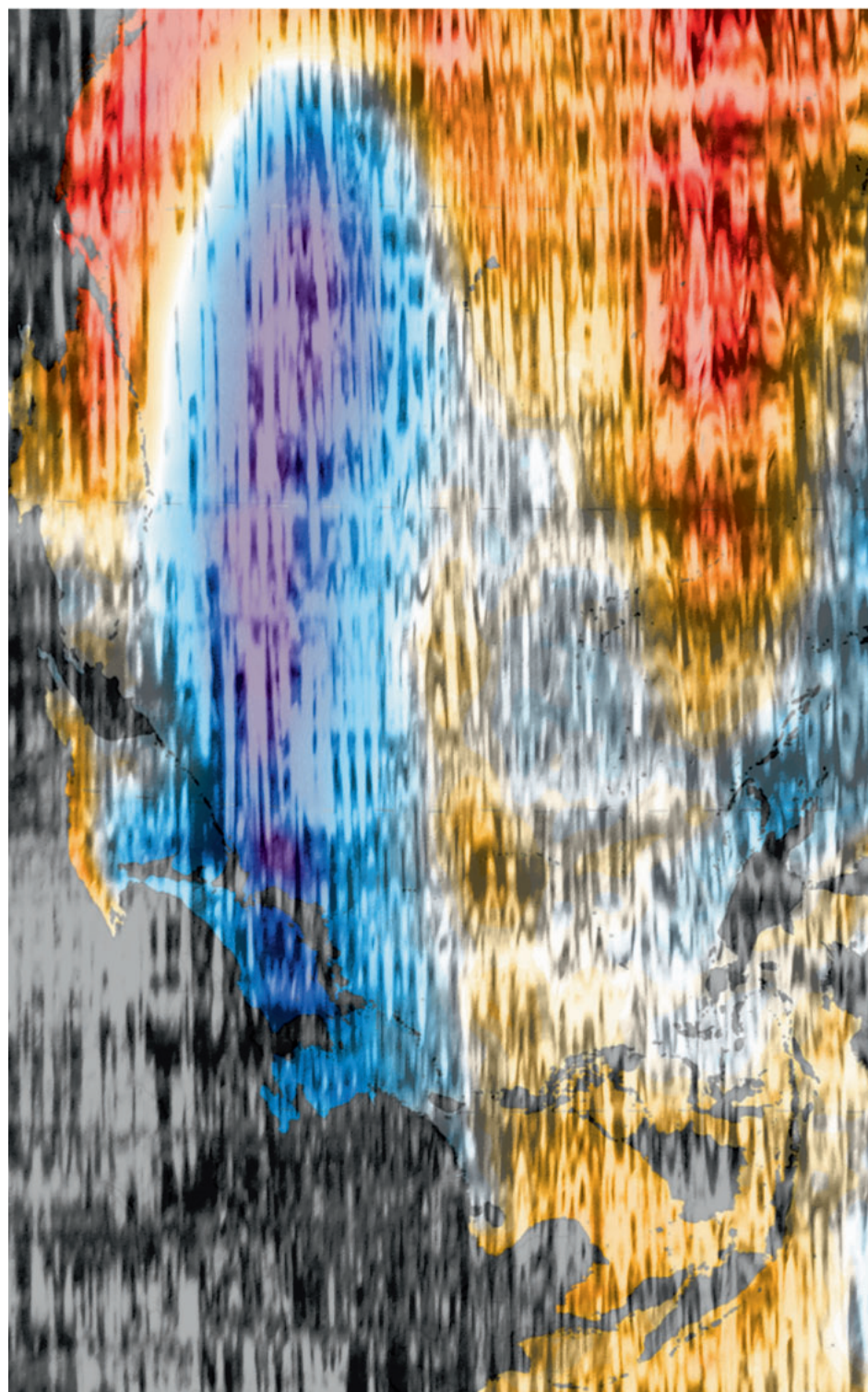
We show that using the mid- and eastern Pacific SST timeseries yields higher forecast skill compared to using the PDO index as a predictor (section 4.3.2), in line with previous work (Vijverberg et al., 2020). The PDO timeseries captures variability in a much larger domain over the Pacific and therefore includes disturbances that are irrelevant to the Rossby wave forcing mechanisms described in the introduction. In contrast, the mid- and eastern Pacific regions are the core PDO regions which ? based on theoretical and modelling experiments ? are expected to force an eastern RW-like response (Ferreira and Frankignoul, 2005; Kushnir et al., 2002). Moreover, while the PDO (simply explaining most variance of SSTA in the North Pacific) suggests that the mid- and eastern Pacific regions are part of the same variability, the correlation between the mid- and eastern Pacific timeseries is only -0.56. Using a separate mid- and eastern Pacific SST timeseries (extracted by the Response-Guided Dimensionality Reduction, RGDR), enables the Ridge regression to (1) learn a more detailed model and (2) use timeseries that are more directly related to the forcing of the RW. Nevertheless, the importance of the background PDO state is further illustrated by the considerable increase in forecast skill for the July-August mean temperature for years with a persistent high amplitude winter-to-spring PDO state (section 4.3.3).

Seasonal dependence of the lagged SST signal for eastern US temperature is evident from Figure 4.6. The exact reason for this specific window of predictability is not fully understood yet. It might be explained by (1) the atmosphere being less chaotic in summer which results in a higher signal-to-noise-ratio or, (2) the seasonal cycle of solar radiation resulting in a stronger impact of high-pressure systems on surface temperature during summer months or (3), potentially amplifying effects of soil-moisture deficits become important near the end of summer (Seneviratne et al., 2010). We also note that it is likely that the persistent summer eastern RW – forced by spring SSTs – leads to both higher temperatures and reduced rainfall, thereby simultaneously affecting summer soil-moisture content. Similarly, the winter-to-spring atmospheric variability that is associated with a strong winter-to-spring PDO state might already affect rainfall over eastern US in those months.

We unravelled the role of ocean-atmosphere feedbacks that are driving long-lead predictability of eastern US summer temperature based on careful analyses with causal discovery algorithms. As shown in section 4.3.3, understanding the sources of predictability paves the way for identifying windows of enhanced S2S predictability and our approach might be successful in finding other potential windows of predictability.

4.4.1 Data Availability Statement

All data used in this study is publicly available. ERA-5 SST and mx2t is available at <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview> and geopotential height at 500 hPa (z500) and meridional wind at 300 hPa (v300) data are available at <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview>. The PNA timeseries from the Climate Prediction Center of NCEP/National Oceanic and Atmospheric Administration (NOAA) was downloaded from the KNMI Climate Explorer, https://climexp.knmi.nl/getindices.cgi?WMO=NCEPData/cpc_pna_daily&STATION=PNA&TYPE=i&id=someone@somewhere&NPERYEAR=366on01-09-2020.



Skillful US Soy-yield forecasts at pre-sowing lead-times

Soy harvest failure events can severely impact farmers, insurance companies and raise global prices. Reliable seasonal forecasts of mis-harvests would allow stakeholders to prepare and take appropriate early action. However, especially for farmers, the reliability and lead-time of current prediction systems provide insufficient information to justify within-season adaptation measures. Recent innovations increased our ability to generate reliable statistical seasonal forecasts. Here, we combine these innovations to predict the 1-3 poor soy harvest years. We use a clustering algorithm to spatially aggregate crop producing regions within the eastern US that are particularly sensitive to hot-dry weather conditions. Next, we use observational climate variables (sea surface temperature (SST) and soil moisture) to extract precursor timeseries at multiple lags. This allows the machine learning model to learn the low-frequency evolution, which carries important information for predictability. A selection based on causal inference allows for physically interpretable precursors. We show that the robust selected predictors are associated with the evolution of the so-called 'horseshoe Pacific SST pattern', in line with previous research. We use the state of the horseshoe Pacific to identify years with enhanced predictability. We achieve high forecast skill of poor harvests events, even 3 months prior to sowing, using a strict one-step-ahead train-test splitting. Over the last 25 years when the horseshoe Pacific SST pattern was anomalous, 65% of the in February predicted poor harvests were correct. When operational, this forecast would enable farmers to make better-informed decisions on adaption measures.

This Chapter is published as:

S. Vijverberg, R. Hamed, and D. Coumou (2022b). "Skillful US Soy-yield Forecasts at Pre-sowing Lead-times". In: *American Meteorological Society: Artificial Intelligence for the Earth Systems* Early Online Release, pp. 1–44. DOI: <https://doi.org/10.1175/AIES-D-21-0009.1>

5.1 Introduction

Seasonal forecasts of United States (US) soy yield play a crucial role in the decision making of numerous stakeholders (Klemm and McPherson, 2017). Reliable yield forecasts can improve crop management of local farmers and inform (non-)governmental organizations on total supply and expected prices (Basso and Liu, 2019). Monetary value of soy is ranked first among all staple crops, and the US supplies one third of globally-traded soybean (Jin et al., 2017), making forecasts highly relevant for commodity traders (Torreggiani et al., 2018). The ongoing increase in soy demand is expected to continue in the future (Fehlenberg et al., 2017), while on the other hand, climate change is expected to threaten US production by increasing average and extreme temperatures (Dirmeyer et al., 2013; Winter et al., 2015).

Reliable and timely predictions can mitigate the impacts from climate extremes (WMO, 2020; Alley et al., 2019), and are expected to be the most cost-efficient way to increase resilience against the projected impacts of climate extremes (Mbow et al., 2019). The most frequently used adaptation measure by farmers is crop or cultivar selection and adjusting planting timing and/or management (Crane et al., 2010). With reliable and timely forecasts, farmers could (1) better manage irrigation schedules (Villani et al., 2021), (2) buy insurance against crop failure (Li et al., 2019), (3) lower the sowing density (Carter et al., 2018; Lobell et al., 2020), (4) decide to only plant in lower (i.e., wetter) altitude areas (Crane et al., 2010), or (5) decide to order more drought resistant crops or soy cultivars (normally done already in January/February) (Dong et al., 2019; Arya et al., 2021; Crane et al., 2010). Commodity traders can buy the soy seasons prior to the harvesting period in October at an expected future price, which shifts the risk of future price fluctuations from the farmer to the trader (Bhardwaj et al., 2015). Thus, also traders have a key interest in information on expected yields already several seasons before harvest time. Knowing the risk well in advance enables both farmers and traders to make more informed decisions. However, to the best of our knowledge, skillful predictions for harvest failures at lead times before planting in May, currently do not exist (Basso and Liu, 2019).

Current operational forecasts are based upon surveys, which rely on local observations by experts (Schnepf, 2017; National Agricultural Statistics Service, 2012). Although these survey-based forecasts can be skillful (Beguería and Maneta, 2020), a logical consequence is that they are done during the growing season, seriously limiting adaptation options. Moreover, such surveys do not take into account long-range weather forecasts or any other relevant climatic information (National Agricultural Statistics Service, 2012).

Recent studies have shown that US soybean production is vulnerable to the combined effect of hot and dry weather conditions in July and August (Ortiz-Bobea et al., 2019; Haqiqi et al., 2020; Goulart et al., 2021), particularly in the mid-to-southern producing regions (Hamed et al., 2021). Extreme conditions (defined by the 95th percentile of temperature and 5th percentile of soil moisture) reduce soy yields by about two standard deviations. This crop-sensitivity to hot-dry conditions is 4 times larger compared to hot conditions alone and 3 times larger compared to dry conditions alone (Hamed et al., 2021). The trend in hot-dry conditions seem to have increased over time and is expected to further increase with future climate change (Hamed et al., 2021; Zscheischler and Seneviratne, 2017).

To inform stakeholders on the hot-dry hazards, dynamical seasonal forecast models can be used in isolation, or they can be combined with a crop simulation model to predict

end-of-season yields. However, the dynamical forecast models do not have sufficient long-lead skill in the US that is necessary to give warning well in advance (i.e., 3-4 months ahead of sowing) (Kirtman et al., 2014; Jong et al., 2021). On these seasonal timescales, a predictable signal can emerge from low-frequency processes of the climate system that can affect surface weather via teleconnections (Krishnamurthy, 2019). However, this predictable signal is generally underestimated in dynamical models (Di Capua et al., 2021; Vijverberg et al., 2020; Merryfield et al., 2020; National Academies of Sciences, 2016; Scaife and Smith, 2018). The poor skill of dynamical seasonal forecasts (Ramírez-Rodriguez et al., 2016) combined with imperfect crop simulation models generates a cascade of uncertainty making the approach unsuited for pre-sowing harvest predictions (Brown et al., 2018; Iizumi et al., 2018).

Machine learning techniques have the potential to circumvent the problem of low signal-to-noise ratios of dynamical models by directly learning from observations. Of course, this is based on the premise that there is inherent predictability in the system. A first essential step is to minimize the unpredictable noise in the target timeseries, which generally involves spatial and temporal averaging (Krishnamurthy, 2019). Secondly, dimensionality reduction methods are often needed to extract the signal(s) from relevant precursor datasets. The traditional approach is to use the known climate indices, since those capture the first-order low-frequency processes in the climate system (e.g., the nino3.4 index to describe the El Niño Southern Oscillation, ENSO). However, these climate indices can easily miss important, more detailed, information relevant for a specific target variable of interest, as they can oversimplify the state of a complex dynamical system to a single number (Vijverberg and Coumou, 2022). For example, the north-Pacific sea surface temperature (SST) is known to affect surface weather in eastern US (Liu et al., 2006; McKinnon et al., 2016), and correlations are indeed found with the Pacific Decadal Oscillation (PDO) index (Kurtzman and Scanlon, 2007; Yu and Zwiers, 2007). However, the PDO pattern is designed to maximize the explained variability over the entire north-Pacific, while ostensibly only a sub-region of the Pacific is physically connected to US surface weather via the forcing of atmospheric Rossby waves (Vijverberg and Coumou, 2022).

Recent studies have shown that eastern US July-August temperature is well predictable by a horseshoe-like SST pattern in the north-Pacific when using machine learning techniques (McKinnon et al., 2016; Vijverberg et al., 2020; Vijverberg and Coumou, 2022). Eastern US heatwaves are predictable up to 50 days lead-time (Vijverberg et al., 2020), and the July-August mean temperature is highly predictable by the winter-to-spring horseshoe SST pattern (Vijverberg and Coumou, 2022). The horseshoe north-Pacific SST pattern resembles the Pacific Decadal Oscillation pattern (Newman et al., 2016) and therefore shows approximately similar decadal variability (Vijverberg and Coumou, 2022). Still, there is a lot of variability occurring at inter- and intra-annual timescales. On seasonal timescales, the horseshoe-like SST pattern can force an arcing Rossby wave over the north American continent in summer, which subsequently leads to more persistent and/or frequent high-pressure systems over the mid and eastern US. High pressure systems are associated with reduced rainfall and higher surface temperature, thereby increasing the risk of hot-dry weather. Besides the ocean-to-atmosphere forcing in summer, the winter-to-spring atmosphere-to-ocean forcing, and two-way ocean-atmosphere feedbacks, also play a role in strengthening the horseshoe-like SST pattern (Vijverberg and Coumou, 2022). This strengthening is important for predictability, since during both persistent

and anomalous mid-latitude SST states there is a stronger atmospheric response (Ferreira and Frankignoul, 2005). Hence, during persistent and anomalous horseshoe Pacific states, a window of predictability exists (Mariotti et al., 2020), since this higher signal results in a pronounced increase in forecast skill for eastern US temperatures (Vijverberg and Coumou, 2022).

Feeding a statistical model with information from multiple lags (instead of only the most recent lag) can often help to improve forecast skill (Vijverberg and Coumou, 2022; Switanek et al., 2020). For example, considering the past evolution informs on the life-cycle of the El Niño Southern Oscillation (ENSO) (if El Niño is in a growing, decaying, or persistent phase) and thus contains more information compared to only a single snapshot, which is important for forecast skill.

Here, we aim at forecasting poor harvest years of US soybean by combining three recent insights: (1) the mid-to-south producing regions are vulnerable to hot and dry conditions in summer (Hamed et al., 2021), and (2) seasonal July-August temperatures are well predictable at seasonal timescales in years with a pronounced horseshoe Pacific SST state (Vijverberg and Coumou, 2022) and (3) using features at multiple lags can further boost forecast skill (Switanek et al., 2020; Vijverberg and Coumou, 2022). We investigate the potential of a forecasting system that can inform stakeholders on the risk of a poor harvest. Hence, we aim at directly predicting impact, defined as poor soy yield years over an aggregated domain (see method section 5.2.2). Building upon state-of-the-art data-driven techniques, we introduce a new framework that applies a causal inference-based method to select specific precursors to reduce overfitting and improve interpretability and reliability (Fig. 5.1). Combining these new insights and techniques allows us to achieve high forecast skill for yield in the mid-to-southern producing region already in February, eight months prior to the harvest period, and 3 months before sowing.

5.2 Method

5.2.1 Data

For the precursor datasets, we use ERA-5 monthly mean sea surface temperature (10°S - 60°N) and the volumetric soil water layer [m^3/m^3] (0-7 cm depth, 135°W - 60°W , 10°S - 60°N) spanning from 1950 up to 2019, both on a 1.0° spatial resolution (Copernicus Climate Change Service (C3S), 2017). The seasonal cycle, calculated as the multi-year mean per month, is removed from the data and subsequently we remove the linear trend. Finally, the SST data is aggregated to 2-month means. Since we are using regularization for our statistical model training, we - after fitting a Gamma distribution - transform the 2-month aggregated soil water data to a standard normal distribution (McKee et al., 1993), known as the Standardized Soil moisture Index (SSI-2). Since SSI-2 is a proxy for the soil moisture levels, we simply refer to the precursor soil moisture (SM). The motivation for using SST is described in the introduction. Soil moisture was initially added to inform on the exposure to droughts occurring from April to July. However, results show that soil moisture at longer lead-times adds value by 'integrating exposure to (dominant) weather patterns' (see Results and Discussion).

For the US soybean yields, we start with county scale census data spanning from 1950 up

to 2019 from the US Department of Agriculture (USDA) National Agriculture Statistics Survey (NASS) Quick Stats database (http://www.nass.usda.gov/Quick_Stats, last access: 1 March 2021). The dataset is first regridded to 0.5° spatial resolution. A grid cell contained within a county is assigned the yield value of that county. Otherwise, if several counties are contained within a grid cell, the grid cell is assigned the average yield value of the contained counties. In a second step, we select for grid cells with common sowing dates (i.e., mid-April to mid-May) and a soybean production area share of at least 90% rainfed agriculture. The period from April to May represents sowing dates for the majority of soybean production across the US. Information on the soybean growing season dates and the production system used is obtained from the monthly irrigated and rainfed crop area database around the year 2000 (MIRCA2000), a global gridded dataset at a 0.5° resolution (Portmann et al., 2010).

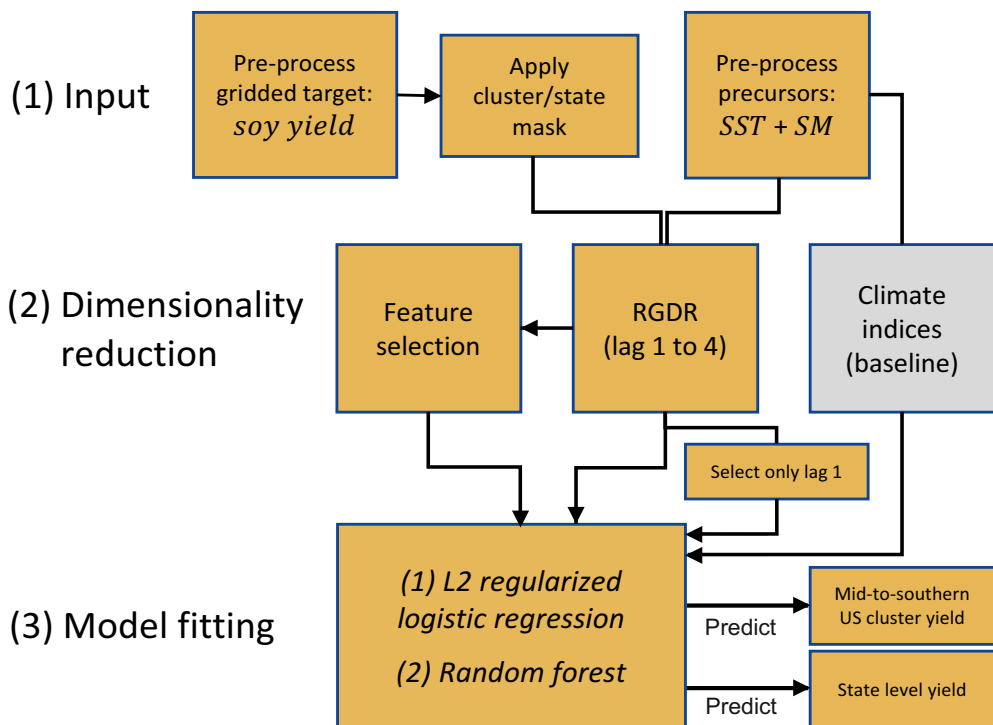


Figure 5.1: Overview of data-driven pipeline(s). The pre-processing steps and input data are described in sections 5.2.1 to 5.2.3. We test four different feature input sets (sections 5.2.4 to 5.2.6) to predict the mid-to-southern US cluster yield timeseries. The selected feature timeseries are also used to predict yield at the state level. Statistical models and verification metrics are described in section 5.2.7 and 5.2.8, respectively.

5.2.2 Clustering of soybean production regions and derivation of spatial mean soy-yield

Previous work showed that the mid- and southern regions are particularly sensitive to summer high maximum temperatures, while the northern regions are mainly sensitive

to early summer cold minimum temperatures (Hamed et al., 2021). This north-south separation in weather sensitivity is ostensibly leading to a detectable difference in yield variability. We use a clustering algorithm to separate the northern soy producing regions from the mid- and southern regions (Fig. 5.2). For our gridded yield data, some gridcells span from 1950 to 2019, but not all. However, the clustering algorithm requires complete timeseries (no Not-a-Number). Hence, a trade-off exists between using fewer-but-longer observational timeseries versus more-but-shorter timeseries. We select data from 1975 onwards since many regions cover the post-1975 period. We interpolate missing data using a second order spline, extrapolation is done using a linear trend line. If more than 7 years of the timeseries needs to be extrapolated, the observational timeseries is excluded from the clustering analysis. Irrespective of the clustering method used (k-means, hierarchical agglomerative clustering optimizing intra-cluster correlation or minimizing intra-cluster variability), we consistently found the same clusters when setting n -clusters to 2. Section 5.3.3 to 5.3.4 focuses on predicting poor yield events in the southern cluster 1, as this domain is coherently sensitive to weather (hot-dry) conditions (Hamed et al., 2021). We do not construct localized (county-level) forecasts, as predicting at such a small spatial scale will decrease the signal-to-noise (S/N) ratio. Such localized forecasts would require a more dedicated analyses and more in-depth knowledge to account for external factors that are specific for each county, although the S/N ratio issue will remain a challenge. Section 5.3.5 explores the skill on the US state level.

We can now use all data (since 1950) and calculate the spatial mean timeseries of the observations that fall within cluster 1 or one of the US states. The data is first detrended and standardized at the grid-cell level to focus exclusively on crop yield anomalies relative to the local expected value (Figure 5.A.1). In this way, the time-mean yield value at any grid-cell is zero irrespective of varying yield potentials across the spatial domain. To get a sufficiently reliable estimation of the mean and standard deviation, we select data with a minimal length of 30 years. To simplify interpretability, we focus on relative poor harvest events within the spatial domain, i.e., we do not apply weighting proportional to the area or the production potential per gridcell. Poor harvest events are defined as years in which the spatial mean timeseries falls below the 33rd percentile threshold, i.e., the 1-in-3 poor harvest years.

As mentioned in the introduction, low-frequency variability in the Pacific plays a role in the mechanism that leads to predictability in the eastern US. Therefore, the detrending is done linearly, since any non-linear detrending method will remove the small amount of slow variability in the crop timeseries that is in-phase with the Pacific variability. Consequently, non-linear detrending methods would to some extent disrupt the low-frequency co-variability that exists between Pacific SST and crop yield fluctuations, and therefore disrupt the training of a statistical model.

5.2.3 Cross-validation and pre-processing

The results presented in the main text are based upon the leave-three-out (LTO) (Iizumi et al., 2021) and the (operational-like) one-step-ahead (OSA) cross-validation (CV) techniques (Lehmann et al., 2020). LTO does not use the year prior and after the test year for training to reduce information leakage from adjacent years that could be present due to temporally correlated timeseries. This is repeated for each test year, meaning we have 69 training

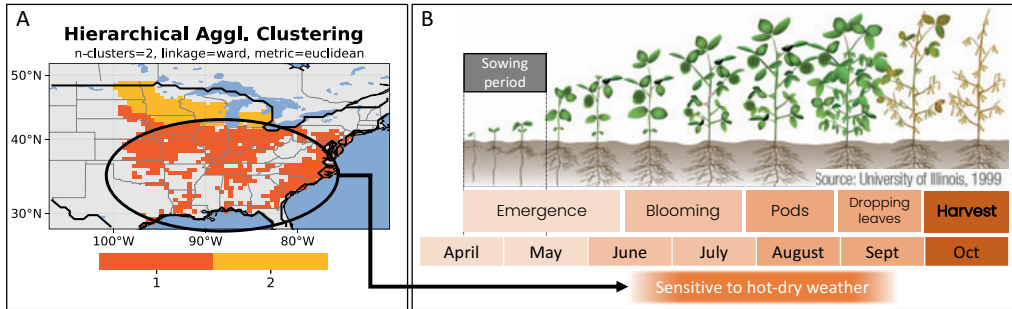


Figure 5.2: The clustering algorithm aims at minimizing the intra-cluster variability. The two detected regions are in-line with previous results documenting the geographical differences in sensitivity to weather anomalies between northern and southern states. (B) The mid-to-southern cluster is sensitive to hot-dry weather, while the northern cluster is not (Hamed et al., 2021). Result sections 5.3.3 to 5.3.4 only focus on predicting spatial averaged yield in cluster 1. Growing stages are adapted from Figure S3 of (Ortiz-Bobea et al., 2019).

folds, with each 66 (or 67 at the edges) years of data available for training (Figure 5.A.2). With one-step-ahead CV, we aim at emulating an operational-like setting. Thus, the crop yield pre-processing is done using only (and all available) training data from the past (detrending, standardizing and calculating the event thresholds), see Figure 5.A.4. For example, with the 'One-Step-Ahead-25' CV, we forecast the recent 25 years (1995-2019). When predicting the year 1995, we only use data from 1950-1994. When predicting the year 1996, we only use data from 1950-1995, and so forth. Thus, the size of the training dataset varies between 44 and 68 years. Only extracting the two spatial clusters of soy yield (section 5.2.2) and the pre-processing of SST and SM precursor datasets (detrending, removing seasonal cycle and calculating the SSI-2) is done in-sample. We expect the latter effect to be small given the large trend and high variability in crop yield compared to SST and SM. Hence, this approach should give a good estimation of the operational forecast skill that would have been achieved over the recent years.

5.2.4 Response-guided dimensionality reduction (RGDR)

In a RGDR method, the dimensionality reduction method takes into account the target variability (Bello et al., 2015). A successful and data-efficient approach is to compute pairwise correlations with a target timeseries (Kretschmer et al., 2017a) and cluster spatially adjacent gridcells into precursor regions using via a clustering approach (Lehmann et al., 2020; Vijverberg et al., 2020). For the SST data, one-dimensional precursor timeseries are calculated for each cluster (here called precursor regions) by calculating an area-weighted and correlation-weighted spatial average (Figure 5.3). For the soil moisture data, we calculate the spatial covariance timeseries of the correlation pattern, only considering significantly correlating gridcells (Figure 5.4).

The clustering of the correlating gridcells is done via the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996; Schubert et al., 2017). DBSCAN iteratively groups together points that are close to each other, while separate and distant points are seen as outliers. With respect to a single point (i.e., gridcell), the radius parameter determines the distance at which points are grouped together. The

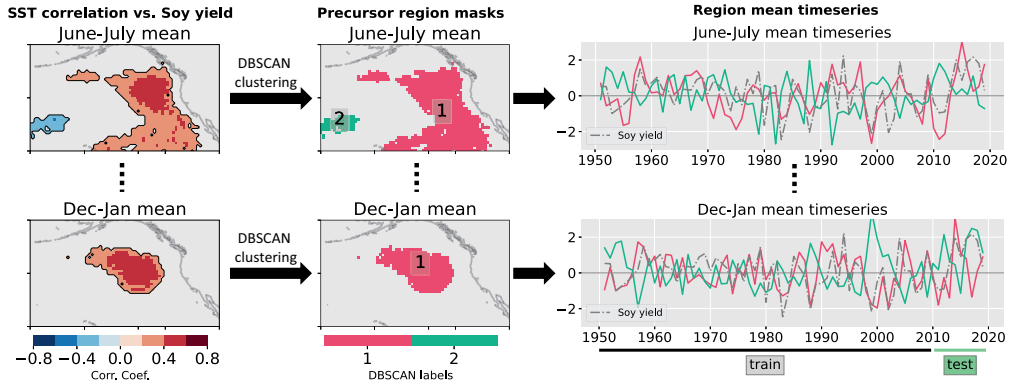


Figure 5.3: Schematic of the Response-Guided Dimensionality Reduction (RGDR) method applied to the pre-processed soy yield and SST data (subdomain shown), with the forecast issued the 1st of August. Correlation maps are computed at 4 lags (here; JJ, AM, FM, and DJ), only the first (JJ) and last (DJ) lag are shown. The significantly correlating gridcells ($\alpha_{FDR} = 0.05$) are clustered using DBSCAN. These clusters are used as masks to calculate area- and correlation value-weighted region mean timeseries. All timeseries are standardized based on the training data.

distance is measured as the great-circle distance in km. If there are less than 3 points in its vicinity, the significantly correlating gridcell is discarded as an outlier. Note, the clustering of positively and negatively correlating gridcell are treated separately. DBSCAN tends to create large clusters as the reachability of a cluster increases via the points at the edges, which can iteratively search for their nearby points. We avoid this by setting the radius parameter to a relatively low value, at 250 km. However, this can generate relatively nearby - but separate - clusters. For the model fitting and feature selection this is undesired since the strong spatial correlations can make these timeseries dependent, i.e., carrying the same signal. Hence, after clustering with a low radius parameter, we search for closely located clusters and group those who show a strong inter-correlation ($\sim > 0.4$). See Appendix 5.B for more information.

For the correlation maps, we set $\alpha_{FDR} = 0.05$ and account for multiple-hypothesis testing by applying the Benjamin-Hochberg correction (Benjamini and Hochberg, 1995; Wilks, 2006). The advantage of the RGDR is that the timeseries are tailored towards the target variable, while the detection power is high (not data-hungry since initial step is based on correlation). The RGDR method can flexibly search for precursor timeseries at multiple lags (Vijverberg and Coumou, 2022), thereby also considering the evolving spatial extent of the precursor regions (see e.g., Figure 5.3). We will search for precursor timeseries at four lags, with the lag 1 being the 2-month mean of the 2 months prior to the forecast release date. The last lag (lag 4) is 8 months prior to the forecast release date. The added value of multiple-lag input is also benchmarked versus using only the most recent lag to forecast soy yield (see Table 5.3).

5.2.5 Causal Inference as precursor selection method

Removing spurious precursors from the input features reduces the risk of overfitting (Kretschmer et al., 2017a). We rely on a causal inference approach to remove spurious

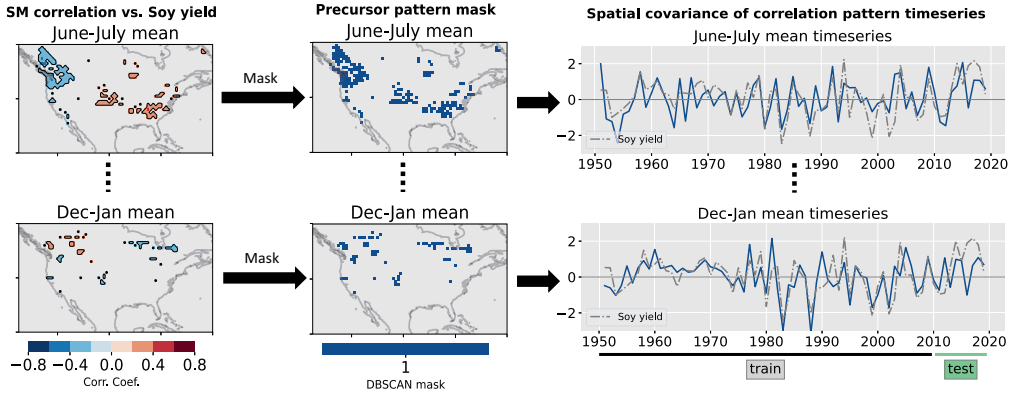


Figure 5.4: Schematic of the Response-Guided Dimensionality Reduction (RGDR) method applied to the pre-processed soy yield and SM data, with the forecast issued the 1st of August. Similar to Figure 5.3, but here the spatial covariance timeseries of the correlation pattern is computed to create a single precursor timeseries.

precursor timeseries and thereby remove redundant information. The selection step is purposefully made not very strict as missing physical drivers is more detrimental to the forecast quality than having a few spurious ones. Via cross-validation and hyperparameter tuning, we aim at assigning low weights to the small number of potential spurious features that pass the selection step (section 5.2.7). To improve interpretability, we use one simple rule: the timeseries of a precursor region at any given lag should always be dependent given the influence of every other precursor detected at any lag. This way we will not regress out the influence of autocorrelation as this information is needed to learn the evolution of a precursor region which can enhance predictability (Switanek et al., 2020; Vijverberg and Coumou, 2022). We use partial correlation for our conditional independence tests (Ebert-Uphoff and Deng, 2012). Although we do not rely on more sophisticated causal discovery (Runge et al., 2019), the causal inference step is expected to keep the strongest correlating precursor timeseries and filter out (most) timeseries that are correlating due to a common driver effect or an indirect link. Due to this simplification, we refer to the selected precursors not as causal but as conditionally dependent, i.e., significantly correlated with the target variable even when conditioned on each timeseries in vector Z (Figure 5.5). The result of this selection step are visualized in Figure 5.6 and 5.7.

5.2.6 Baseline dimensionality reduction approach

We compare our rather complex dimensionality reduction approach with a simple benchmark. We compute, for each training dataset, the first Empirical Orthogonal Function (EOF) over the Pacific decadal Oscillation (PDO) domain (110°E , 260°E , 20°N , 70°N) on our pre-processed SST dataset using the November to March months, closely resembling the PDO index (Trenberth and Fasullo, 2013). By projecting the EOF loading pattern on observations, we extrapolate to the test sets. We also compute the area-weighted spatial mean timeseries over the ENSO3.4 domain (190°E , 240°E , 5°S , 5°N) as a proxy for ENSO variability. We refer to this baseline approach as 'climate indices'.

Selection algorithm

```

 $X = (X_{\tau_1}^1, X_{\tau_2}^1, \dots, X_{\tau_4}^n)$ 
 $X$ : set of all RGDR features
for  $x_{\tau}^i$  in  $X$ :
     $Z = X \setminus \{x_{\tau_1}^i, \dots, x_{\tau_4}^i\}$ 
     $Z$ : set of all features, excl.  $x_{\tau}^i$  and lag shifted timeseries of  $x_{\tau}^i$ 
    list_pvals = []
    for z in  $Z$ :
         $p_{val} = \text{parcorr}(x_{\tau}^i, y|z)$ 
        list_pvals.append( $p_{val}$ )
    If  $\max(\text{list\_pvals}) > 0.05$ :
         $x_{\tau}^i$  was always conditionally dependent given z
         $x_{\tau}^i \in S$ 
 $S$ : set of Conditionally Dependent features

```

Figure 5.5: Pseudo-code of selection algorithm. Where `parcorr` indicates a partial correlation analysis, x_{τ}^i is the to-be-tested timeseries, y is the Soy yield timeseries, z is the to-be-tested confounding timeseries, and p_{val} is the p-value.

Table 5.1: Exhaustive search over hyperparameters is tested within a 10-fold CV aiming to minimize the Brier (error) score.

Statistical Model	Hyperparameters	
L2 regularized LR	C (inverse regularization parameter)	[1E-3, 1E-2, 5E-2, 1E-1, 0.5, 1, 1.2, 4, 7, 10, 20]
Random Forest Classifier (n-estimators = 300)	Max depth	[2, 5, 8, 15]
	Max features	[.4, .8]
	Max samples	[.4, .7]

5.2.7 Statistical models and hyperparameter tuning

We tested both a regularized logistic regression (LR) and Random Forest (RF) to make probabilistic forecasts. For the tuning of hyperparameters, we apply a double cross-validation approach (Vijverberg et al., 2020). In this approach, each training dataset of the 'outer' cross-validation is split into training and validation sets using a second 'inner' cross-validation. The methods for the 'outer' cross-validation are introduced in section 5.2.3, the 'inner' cross-validation is always a 10-fold CV. A schematic of the double cross-validation approach is shown in Figure 5.A.2. The parameters are tuned to minimize the brier score (BS). The statistical models, cross-validation and parameter-search software are used from the Scikit-learn Python package (Varoquaux et al., 2015). To ensure equal penalty of the regularization (parameter C in Table 5.1), all features are standardized (based on the training data) prior to model fitting. For the random forest, we tune the max depth, indicating the depth of each tree (number of splits). With max features, we limit the percentage of features used to create a single tree to 40 or 80%. With max samples, we limit the percentage of samples used to construct each tree to 40 or 70%. Lower percentages of the latter two parameters can improve generalizability. N-estimators refer to the number of trees build.

Table 5.2: Verification metrics presented in this manuscript. The first row shows the Brier Skill Score (BSS) with the climatological probability as the benchmark forecast (0.33). The Brier Score (BS) is calculated for both the benchmark (BS_{clim}) as well as the forecast (BS_f). f_i represents the forecast at timestep i , N is the number of forecast/observation pairs.

Brier Skill Score	$BS_f = \sum_{i=1}^N (f_i - o_i)^2$	$BSS = 1 - \frac{BS_f}{BS_{clim}}$
Accuracy	$\frac{\text{True Positive} + \text{True Negative}}{N}$	
Precision	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	

5.2.8 Skill metrics

Multiple metrics are needed for proper verification (Vijverberg et al., 2020; Wilks, 2011) and we use (1) the Brier Skill Score (BSS), where the benchmark is the observed climatological probability after the trend line has been subtracted (i.e., 33% probability of poor harvest), (2) the Accuracy and (3) the Precision metric (Table 2). The probability threshold used for the latter two metrics is equal to the climatological probability (33%). Given the probability of poor harvest, the Accuracy and Precision of a random guess is 54 and 33, respectively.

5.3 Results

The results of the feature selection algorithm, using the Leave-Three-Out (LTO) CV, are discussed in section 5.3.3. The corresponding figures for the One-Step-Ahead (OSA) CV can be found in Appendix 5.B. The forecast skill of LTO and OSA are presented in section 5.3.2 and 5.3.3, respectively. Section 5.3.4 shows an overview of the forecast skill when using different input features, different CV methods and the two statistical models. The forecast skill on a smaller spatial domain (state-level) is verified in section 5.3.5.

5.3.1 Conditionally Dependent (C.D.) precursor regions

The Pacific horseshoe-like region is the most robust SST precursor region detected at all lead-times (Figure 5.6), even up to 15 months prior to the start of the harvest period (1st of October). The spatial extent of the horseshoe region decreases as function of lead-time, but the magnitude remains strong. We also observe a robust Atlantic signal in summer and late spring (i.e., at short lead times), which is interpreted as the SST response to the westward extension of Rossby wave (high- and low-pressure system over the Atlantic) that is associated with the low-pressure system (linked to less hot-dry conditions) over the eastern US. However, the fact that the Atlantic regions passes all conditional independence tests (in all training samples) suggests that the western Atlantic SST variability is not solely the result of the Rossby wave that is forced by the horseshoe Pacific pattern (see Discussion).

The positive soil moisture correlation pattern in August shows that locally higher soil moisture levels correlate with higher end-of-year yields. For the north-west of North America (i.e., outside the harvest area), we observe regions with negative correlations (Figure 5.7) indicating wet soils in the north-west are linked to higher yields in our target area (cluster 1). The correlation maps prior to August no longer show a soil moisture signal within the harvest area, and they show a positive correlation over the north-west of North America domain (opposite sign compared to August). The circulation patterns associated with the soil moisture correlation patterns are shown in Figure 5.B.1. We find that the soil moisture pattern correlates with a circulation pattern that shows a consistent low-pressure system over the mid-Pacific and high-pressure system over the north/north-eastern Pacific. Via ocean-atmosphere interaction, such a circulation pattern is likely able to strengthen the Pacific horseshoe pattern (see Discussion). At very long lead-times, this indirect signal becomes weaker, and the soil moisture is generally filtered out by the conditional independence tests.

5.3.2 Leave-three-out hindcast skill

Using the Random Forest model, the out-of-sample hindcasted soy harvest failures between 1950 and 2019 can be predicted with good accuracy already in February, which is approximately 3 months prior to sowing (Figure 5.8). This long-lead predictability is possible during specific windows of predictability. We use the state of the horseshoe-like Pacific precursor pattern (Figure 5.6) to quantify the signal strength which is used to identify the windows of predictability (i.e., when the signal is strong). For a given year, the state of the horseshoe precursor pattern is calculated by taking the mean over all lags that passed the conditional independence tests. We use both the horseshoe Pacific and the soil moisture conditions for the August forecast, since we know that local end-of-summer SM can strongly impact crop growth (Hamed et al., 2021). Using soil moisture earlier in the season to identify the window of predictability did not improve results. The signal strength (S) timeseries are plotted below each forecast in Figure 5.8. The years with an anomalously strong state, either negative or positive, are indicated by red and blue dots, respectively (see legend Figure 5.8). During these states, we expect larger deviations from the climatological mean weather, and therefore better yield predictability.

The tables next to each forecast in Figure 5.8 show 3 skill metrics, i.e., brier skill score (BSS), the accuracy and the precision (see section 5.2.8). The columns refer to the subset of years that are included when calculating the verification skill: All (all years), top 50% (the 50% years with strongest signal), and top 30% (the 30% years with strongest signal). Hence, these subsets of years indicate the windows of predictability in time. During the top 30% years, i.e., when the signal is very strong, a high brier skill score (≤ 0.5) shows that the forecasts are both reliable and confident in terms of its assigned probability. Overall, we observe a decline in skill as function of lead-time. However, during the top 30% window of predictability the skill for poor harvest years ($< 33^{rd}$ percentile) is remarkably high across lead-times, with the August forecast achieving a BSS of 0.80 and precision of 88% and the February forecast achieves a BSS of 0.59 and precision of 62%. We obtain substantially less forecast skill when predicting good harvest years ($> 66^{th}$ percentile), result not shown (see Discussion).

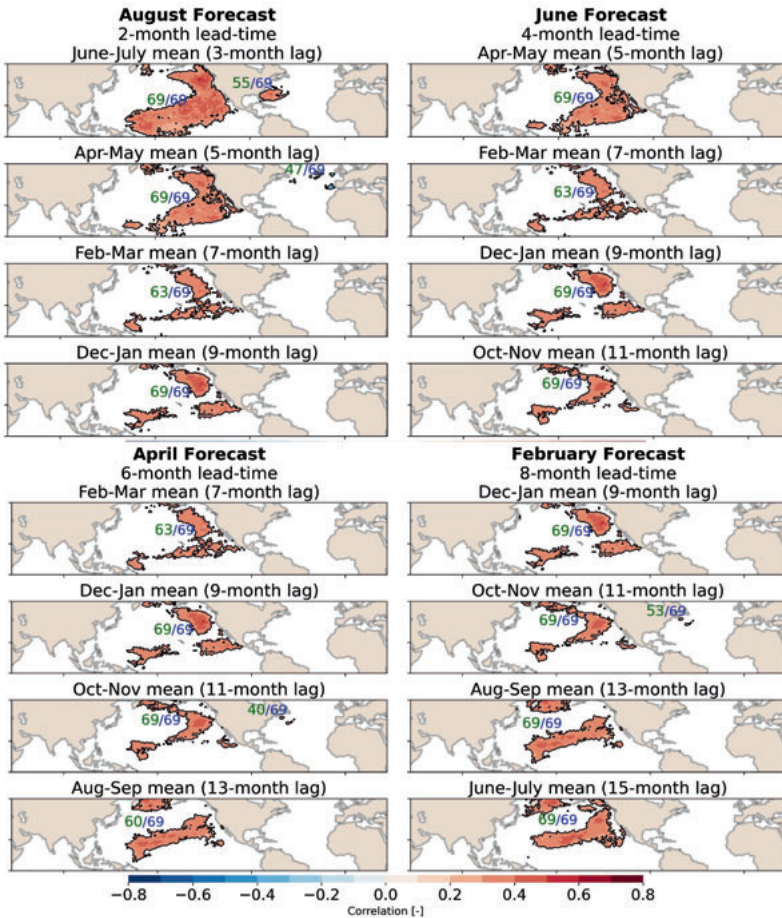


Figure 5.6: SST (2-month mean) correlation maps versus the crop yield variability in cluster 1 (see Figure 5.2) for each forecast month. A correlation value is only shown if a gridcell is significantly correlating in one of the 69 training datasets. The green integers denote the number of training samples the precursors timeseries has passed all the conditional independence tests. The blue integers denote the number of training samples the precursor timeseries is detected by the response-guided dimensionality reduction (RGDR). For clarity, we only show the regions which were conditionally dependent in at least 50% of the training samples. The precursor region labels assigned by DBSCAN are shown in Figure 5.B.2.

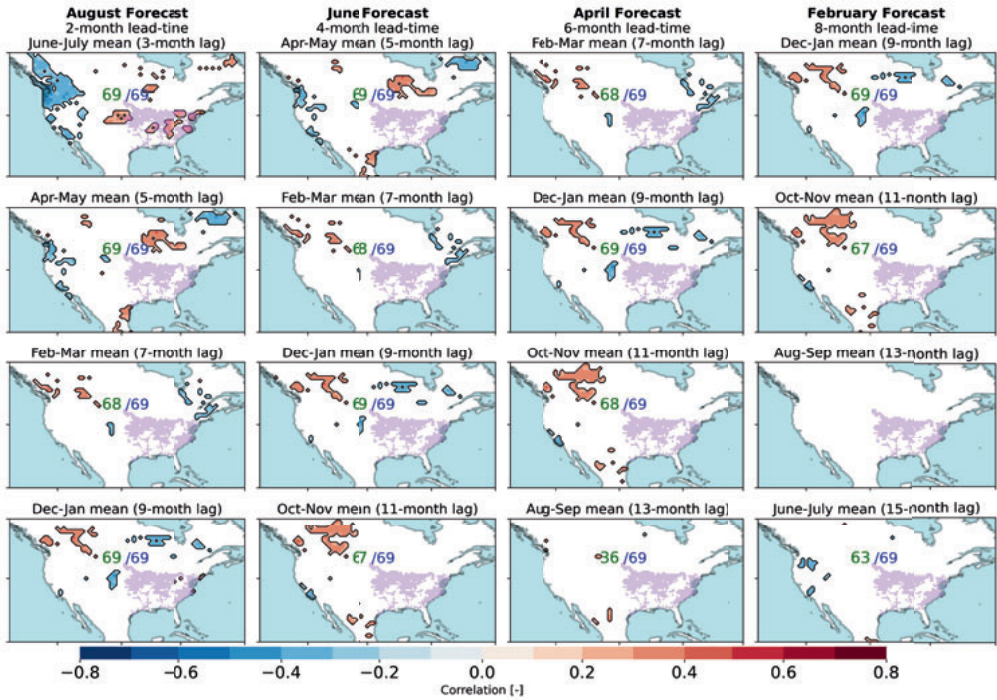


Figure 5.7: SM (SSI-2) correlation maps versus the crop yield variability in cluster 1 (see Figure 5.2) for each forecast month. A correlation value is only shown if a gridcell is significantly correlating in one of the 69 training datasets. The SM precursor timeseries is based upon the spatial covariance of the (significant) correlation values. The ratio shows the conditional dependent/detected precursor timeseries, similar to Figure 5.6. If the SM time series is not conditionally dependent in at least 50% of the training samples, the SM correlation pattern is completely masked. The spatial domain of cluster 1 is shown in opaque pink.

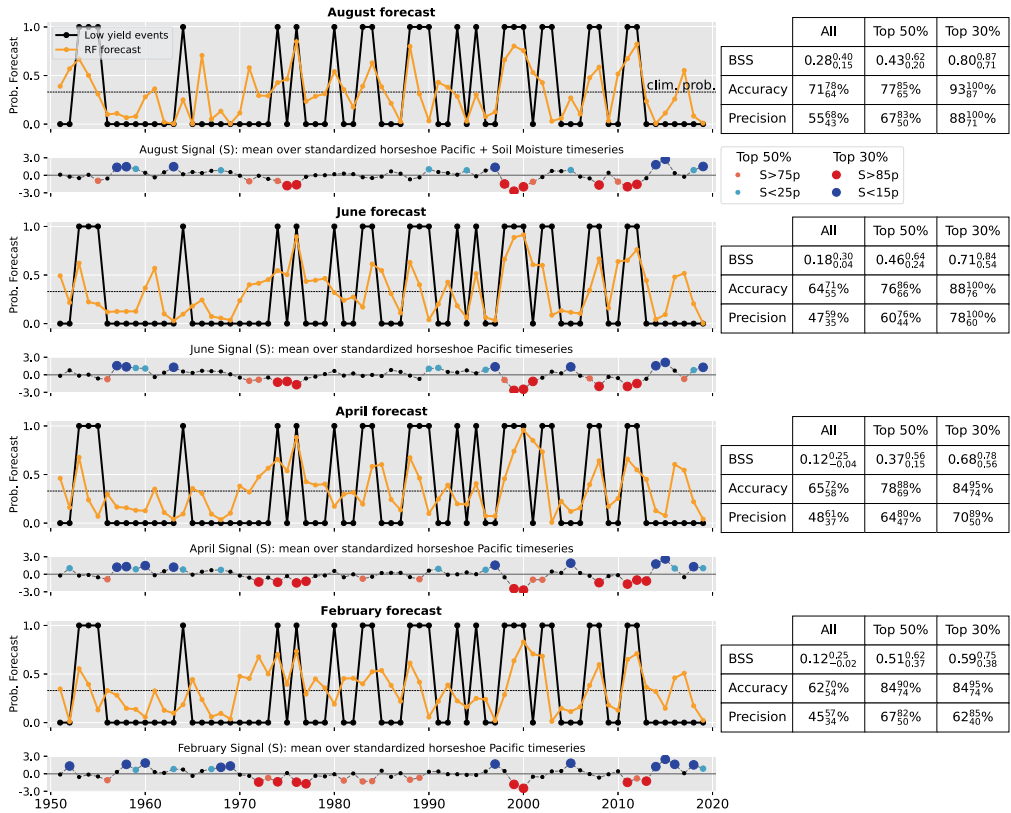


Figure 5.8: Probabilistic out-of-sample forecast of low yield events (below 33rd percentile) by a Random Forest model using a Leave-Three-Out cross validation over 69 years of data. Only the conditionally dependent timeseries from RGDR are used as input. Solid black line indicates the observed low yield events, yellow line indicates the out-of-sample probabilistic forecast, thin horizontal black line indicates the mean probability (0.33). The top 50% and 30% indicate windows of predictability in time, identified by a strong signal which is plotted below each forecast timeseries. The sub- and superscript indicate the 90% confidence intervals from 2000 bootstrapped samples.

5.3.3 One-step-ahead forecast skill

In an operational-like setting, our forecast system would have been able to predict low yield years with high skill over the last 25 years (Figure 5.9). Again, we observe a systematic boost in forecast skill during the windows of predictability. The top 50% forecast (consisting of 14 years) in February achieves a precision of 67%, i.e., 4 true positives out of 6 forecasted low-yield events (see Figure 5.9). 12 out of the total 15 predictions were correct, i.e., an accuracy of 80%. The precursors that are selected (Figure 5.B.4 and 5.B.5) are similar to precursors presented in section , although the horseshoe Pacific SST region is more often removed after conditioning on the soil moisture pattern timeseries for the August and June forecasts.

For the one-step-ahead (OSA) cross-validation, we are more limited in the amount of datapoints for training and verification. Calculating the skill metrics based on only 25 (or less) forecast/observation pairs can introduce a (sampling) bias by being (un)lucky in the sample set. To get a feeling for the sampling bias, we also calculated the skill metrics over the recent 30 and 20 years (Figure 5.C.1) and we slightly changed the event timeseries by changing the quantile threshold from 0.31 up to 0.35 with steps of 0.01 (Figure 5.C.2). Given these perturbations, the precision did not drop below 65% for the February forecast during the window of opportunity (top 50%). Overall, the skill is robust across months and quantiles, yet there are some (positive and negative) outliers (Figure 5.C.1 and 5.C.2). We also tested the influence of a varying event frequency within the verification periods and found this to be minor (Appendix 5.C).

5.3.4 Synthesis of results

Within our set of experiments shown in Table 5.3, using combined lead-time models (features at multiple lags) show the largest boost in forecast skill compared to using only the most recent lag (indicated by 'lag 1 RGDR precursors' in Table 5.3). Furthermore, we observe that, with enough data (66 years), the random forest performs best when using all precursor timeseries extracted by the RGDR method, i.e., no feature selection (vector \mathbf{X} in Figure 5.5). With limited training data (44 up to 68 years), the performance of the regularized LR and the RF is very similar. The RF tends to be a bit more careful in its assigned probabilities (lower sharpness) and we believe this causes it to perform a bit more stable. For example, we tested the impact of in-sample versus out-of-sample pre-processing of the target variable, and the RF was less sensitive to this change (Figure 5.C.2).

5.3.5 State-level forecast skill

When predicting the state-level yield (Figure 5.10), forecast skill is good (BSS of ≥ 0.3) inlands of our mid-to-southern target cluster (section 5.2.2), yet underperforms compared to predicting the average of the entire target cluster (Figure 5.9). The state-level forecasts are made using the one-step-ahead-25 cross validation and the random forest (similar skill was obtained using the regularized logistic regression). Again, we use only the selected - more trustworthy - features and refitted the model for each state.

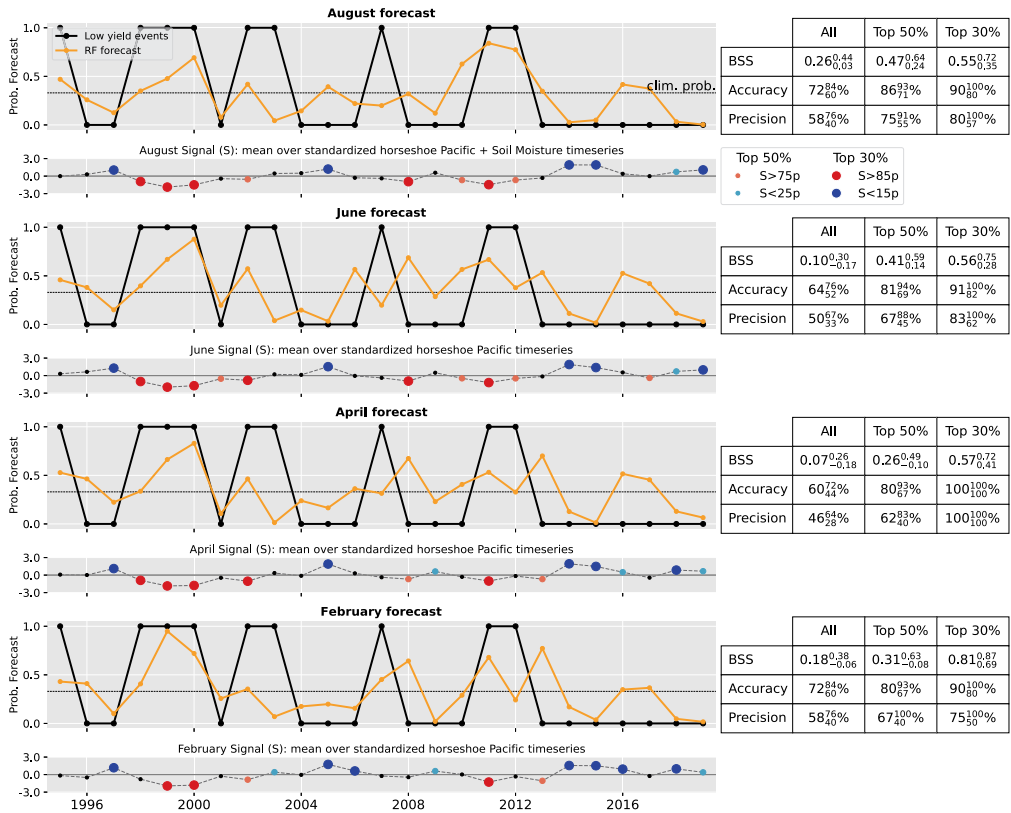


Figure 5.9: Same as Figure 5.8, but using the one-step-ahead cross-validation scheme over the last 25 years.

Table 5.3: Overview of Brier Skill Scores for two cross-validation schemes (LTO and OSA-25), two statistical models (LR and RF) and four different types in input training data. Our baseline model is called 'climate indices' and consists of the PDO and ENSO3.4 timeseries. For 'lag 1 RGDR precursors' we only use the timeseries extracted at the most recent lag w.r.t. the forecast month. 'RGDR precursors' refers to using all timeseries at all lags found by the RGDR method, i.e., vector \mathbf{X} in Figure 5.5. 'C.D. RGDR precursors' refers to all precursor timeseries that were found conditionally dependent, i.e., vector \mathbf{S} in Figure 5.5.

Statistical model	training input	August	June	April	February	Mean
Leave Three Out (LTO) cross-validation						
Reg. LR	Climate indices	0.03	-0.05	-0.10	-0.09	-0.05
Reg. LR	Lag 1 RGDR precursors	0.11	0.05	0.01	0.24	0.11
Reg. LR	RGDR precursors	0.24	0.17	0.15	0.12	0.17
Reg. LR	C.D. RGDR precursors	0.25	0.04	0.05	0.09	0.11
RF	Climate indices	0.05	0.02	0.02	0.05	0.04
RF	Lag 1 RGDR precursors	0.12	-0.04	0.06	0.23	0.09
RF	RGDR precursors	0.32	0.18	0.18	0.13	0.20
RF	C.D. RGDR precursors	0.28	0.18	0.12	0.12	0.18
One-Step-Ahead-25 (OSA-25) cross-validation						
Reg. LR	Climate indices	0.09	0.05	0.01	-0.00	0.04
Reg. LR	Lag 1 RGDR precursors	0.01	0.22	0.07	0.22	0.13
Reg. LR	RGDR precursors	0.19	0.24	0.20	0.18	0.20
Reg. LR	C.D. RGDR precursors	0.17	0.10	0.07	0.26	0.15
RF	Climate indices	0.13	0.07	0.02	0.10	0.08
RF	Lag 1 RGDR precursors	0.08	0.14	-0.01	0.19	0.10
RF	RGDR precursors	0.28	0.20	0.15	0.12	0.19
RF	C.D. RGDR precursors	0.26	0.10	0.07	0.18	0.15

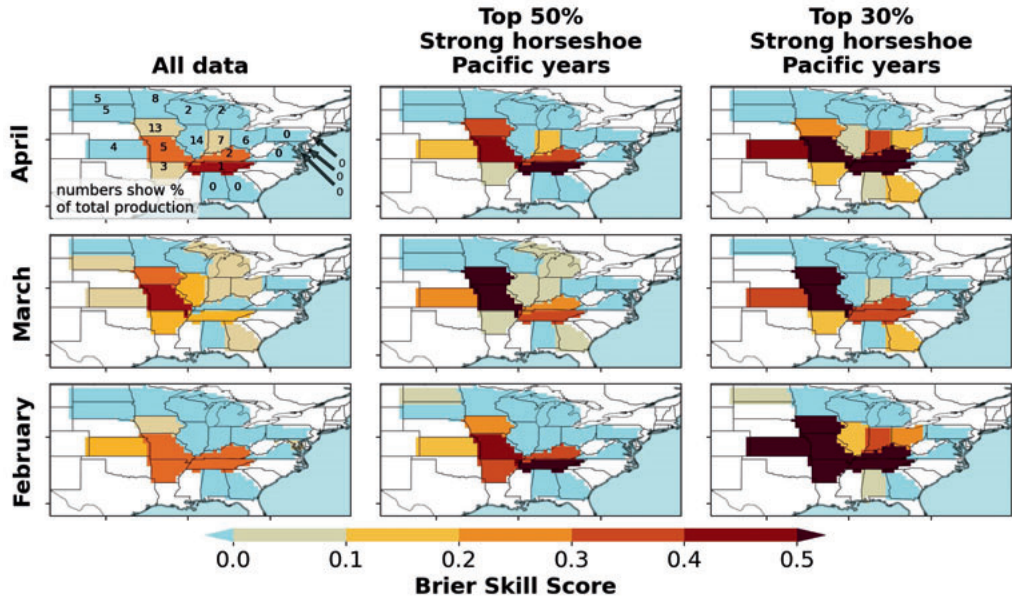


Figure 5.10: State-level forecast skill in terms of the Brier Skill Score. Data-driven pipeline is similar to Figure 10, yet here we forecast the poor (1-3) yield events of the out-of-sample pre-processed state-level aggregated yield. The 3 rows show the forecast release months, April, March, and February, always released on the first day of each month.

5.4 Discussion

Our forecast framework provides new opportunities for stakeholders to make better-informed decisions already early in the season. When a poor harvest (defined as a 1-in-3 year event) is predicted the 1st of February during a window of opportunity (top 50%), i.e. ~ 3 months before sowing, the risk of a poor harvest increases from 33% to $\sim 65\%$ (Figure 5.9). Using the window of predictability concept, we can communicate when we have high confidence in a forecast and when not. Based on a forecasted poor harvest in February, farmers still have sufficient time to change their planting timing, avoid planting the drought-prone agricultural fields, or select more drought/heat resistant soy cultivars. Such forecasts are also relevant for (non)governmental institutions, commodity traders and crop insurance companies (Basso and Liu, 2019; Torreggiani et al., 2018). Forecasts based on surveys cannot provide farmers with this information at sufficient lead-times (Begueriá and Maneta, 2020; National Agricultural Statistics Service, 2012). To our best of knowledge, current long-range dynamical weather forecasts are also unable to inform farmers on the weather-related risks at these very long lead-times. Although a comparison is not within the scope of this paper, the North-American Multi-Model Ensemble re-forecast (1982-2009) shows negative skill for the JJA temperature in eastern US when initialized 1st of January (Kirtman et al., 2014). Similarly, the European dynamical seasonal forecast model (SEAS5) shows relatively low correlation coefficients (roughly 0.4) for summer temperature in our target region (Johnson et al., 2019). Another benefit of data-driven forecasts is that predictions of impact (e.g., poor harvest years) are as straightforward as predicting weather variability (e.g., hot-dry extremes), whereas predicting poor harvest

with plant simulation models is very difficult (Brown et al., 2018; Iizumi et al., 2018). For data-driven predictions, a proper verification is crucial, which is why we have used strict train-test splitting, multiple metrics (importance illustrated in Vijverberg et al. (2020)), and a clear benchmark forecast. Proper verification is unfortunately still an issue in statistical crop forecasting literature (Schauberger et al., 2020).

To achieve the high forecast skill, we found that the following steps are important: (1) clustering of target variable, (2) using the horseshoe Pacific region (extracted by the RGDR method) to identify the windows of predictability, (3) using multiple lags as input. First, the spatial clustering algorithm identifies spatial regions that behaved similarly in terms of harvests. The southern cluster (cluster 1 in Figure 5.2) encompasses the mid-to-southern producing region, which is sensitive to hot-dry extremes in summer, while the northern region is less sensitive to weather (Hamed et al., 2021; Schauburger et al., 2017a). We also obtain no forecast skill for the northern cluster 2 (result not shown). For the mid-to-southern region we obtain positive yet substantially less forecast skill for positive (>66th percentile) yield events (result not shown). This is expected given the strong non-linear temperature response of crop growth, i.e., temperatures exceeding 30°C are very harmful, while the sensitivity to temperature below the 30°C mark is much weaker (Schlenker and Roberts, 2009; Schauburger et al., 2017b). To summarize, optimizing the signal-to-noise ratio (SNR) of the target crop timeseries is a crucial step to improve skill on the subseasonal-to-seasonal timescales, as has also been shown for temperature forecasts (Vijverberg et al., 2020). The second step, i.e., identifying the window of predictability based on the state of the Pacific, also resulted in a systematic and substantial boost in skill, in line with previous results (Vijverberg and Coumou, 2022). As shown in Table 5.3, forecast skill is substantially increased by including the evolution of precursor patterns throughout the season, rather than just using a snapshot of their state at forecast time. Forecasting yield at the state-level reduces skill, which is expected as the effects of unobserved factors on a smaller spatial scale (such as crop management decisions, observational biases, or e.g., unpredictable local deep convection events) become more dominant. Nevertheless, the state-level predictions are still skillful in 6 states, which are responsible for ~31% of total US soy production.

The reduced set of precursors by applying conditional independence testing (i.e., ~10 precursors instead of ~40) enables a better physical interpretation (see section physical interpretation below). Moreover, this precursor selection step reduces the risk of overfitting. Figure 5.B.3 shows the precursor regions that were removed by the selection step. Based upon expert judgement, these regions (often small and located close to the coastline) are likely not causally linked to our target. Hence, although the selection step can slightly degrade skill metrics (Table 5.9), the trustworthiness is increased. However, we did notice that for the OSA-25 CV (with less training data), the selection step was less robust.

In general, there are still opportunities to further improve forecast skill. Here we used DBSCAN to cluster coherent precursor regions. We assume that gridcells that (1) correlate with the same sign and (2) are located close to each other carry the same signal. The spatial mean that we calculate for each precursor region reduces noise, thereby improving the S/N ratio, yet particularly for large precursor regions, there are likely spatial differences in the signal strength. For this reason, the spatial mean is weighted with the correlation value, but one might want to model this with more detail. Furthermore, since crop yield is also affected by other factors than weather, it will also be insightful to directly predict

hot-dry conditions. Furthermore, the selection step can be further fine-tuned by relying more on expert knowledge.

5.4.1 Physical interpretation

Our dimensionality reduction and precursor selection method based on conditional independence tests has identified the horseshoe Pacific SST region and the soil moisture patterns as the most robust precursors. The very long lead-time signal exists due to the low-frequency dynamics in the Pacific, which is often summarized using the Pacific Decadal Oscillation (PDO) index, although the PDO is not driven by a single mechanism (Newman et al., 2016). The persistence of the (horseshoe Pacific) SST anomaly is crucial to force a barotropic Rossby wave-like response, as supported by modeling experiments (Ferreira and Frankignoul, 2005) and observations (Vijverberg and Coumou, 2022). We argue that this is the physical reason behind the increased predictive skill when using multi-lagged input features, i.e., having information on multiple lags informs on the persistence and momentum of the SST precursor. Our window of opportunity (strong signal) years are therefore characterized as years with a high persistence (by calculating mean over all lags) and subsequently selecting anomalous states.

Besides the local effect of soil moisture on crop yield (Hamed et al., 2021), we found that the soil moisture pattern over the continental north-America reflect the dominant circulation patterns that are present prior to the summer season. The circulation associated with the soil moisture patterns is, based on previous research, expected to strengthen the horseshoe Pacific SSTs (Vijverberg and Coumou, 2022). Causal discovery analyses on observations showed that the slowly varying horseshoe Pacific region promotes the occurrence of this arcing Rossby wave and that, in turn, the horseshoe SST pattern is also strengthened by the RW. Both low-frequency ocean dynamics in the Pacific and atmosphere-to-ocean forcing (the latter captured indirectly via soil moisture) determine the development of a strong ocean-to-atmosphere boundary forcing. Soil moisture has a much longer memory, thereby making it better suited for forecasting purposes, i.e., it is much less noisy compared to using circulation directly. In June-July, the circulation pattern (Figure 5.B.1) also projects onto the concomitant robust SST Atlantic signal. However, the fact that the Atlantic SST signal is not conditionally independent when regressing out the influence of the horseshoe Pacific pattern, suggests another cause could be present. We suspect that dominant periods of the wavenumber 6 Rossby wave, which is a known summer mode of variability (Branstator and Teng, 2017; Kornhuber et al., 2017a; Vijverberg and Coumou, 2022), also plays a role in driving both yield variability as well as Atlantic SST variability.

5.5 Conclusion

We show that a good physical understanding in combination with innovative data-driven techniques that aim at optimizing the signal-to-noise ratio can achieve high forecast skill at unprecedented lead-times. Not all spatial regions and years have the same level of intrinsic predictability (Mariotti et al., 2020). To detect regions and periods with high predictability (i.e. windows of predictability), both target and input features need to be

carefully selected and optimized (Table 5.3), see also Vijverberg et al. (2020) and Vijverberg and Coumou (2022). To do so, we apply a clustering technique for the target, and we use a response-guided dimensionality reduction method and a feature selection based on causal inference. We target the 1-in-3 poor soy harvest years within an aggregated spatial domain (Figure 5.2) showing a homogeneous sensitivity to hot-dry extremes (Hamed et al., 2021). Our operational-like forecast system can predict poor harvest years with high forecast skill and high confidence already on the 1st of February, i.e., ~ 3 months prior to sowing (Figure 5.9).

This forecast can be released eight months prior to the harvest period, whereas the current operational forecast system, although released on a higher spatial resolution, is released only in August (Schnepf, 2017). If we forecast a poor harvest, during a year with an anomalous horseshoe Pacific SST state, the probability of a poor harvest released in February increases from the normal 33% to $>65\%$ (Figure 5.9). The high Brier Skill Score (>0.5) shows that the forecast is both reliable and confident in terms of its assigned probability. Most importantly, our forecast is released 3 months prior to sowing, which allows farmers to take anticipatory action, i.e., change to more drought resistant cultivars or change planting management. Our approach can be tuned to specific needs of stakeholders, e.g., focus on specific sub-regions, adapting the threshold that defines a poor harvest year, and making additional forecasts for hot-dry weather to better isolate the weather-induced risk.

Acknowledgements

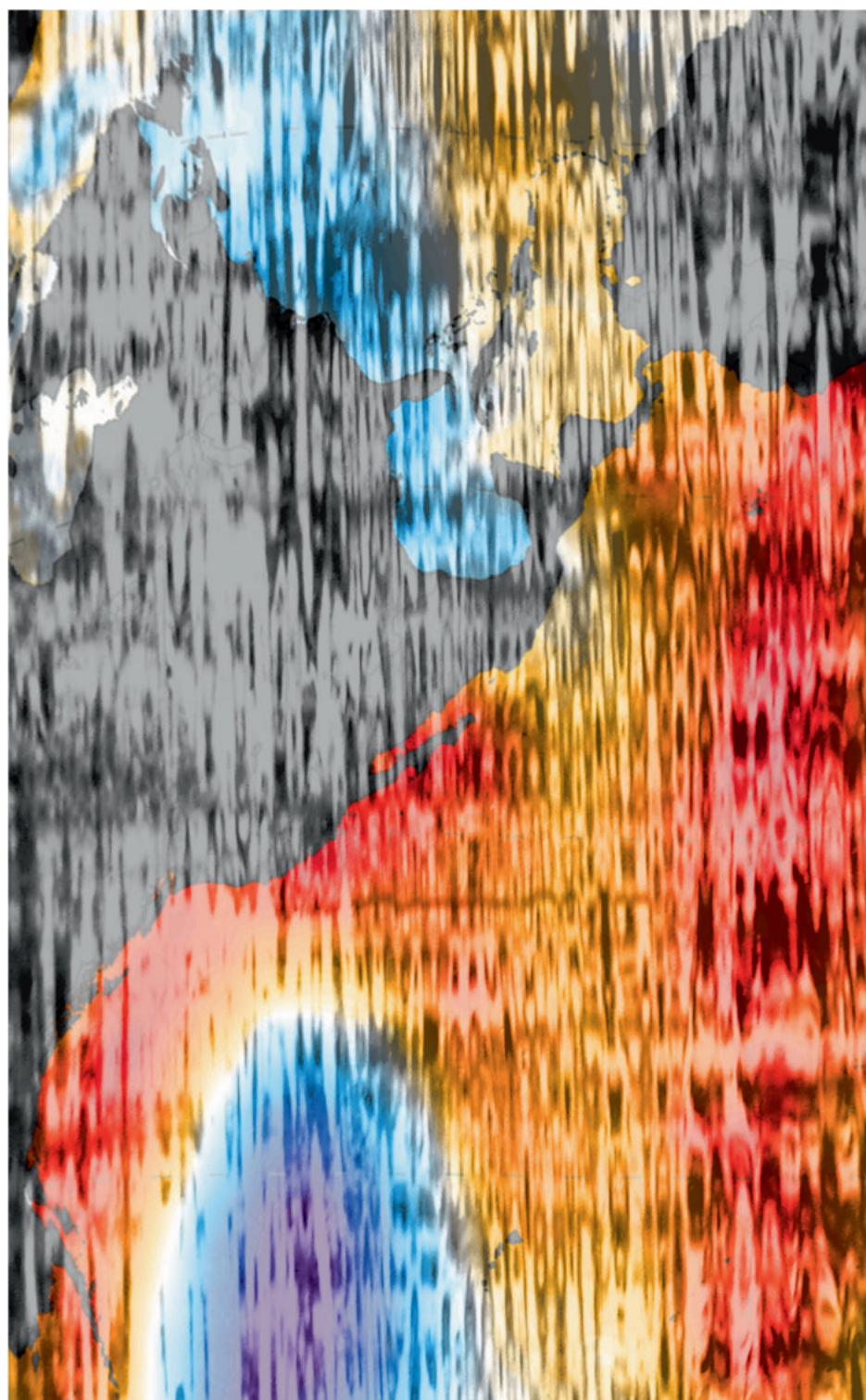
This research was supported by the Dutch Research Council under the grant agreement 016.Vidi.171011 (Vidi project: Persistent Summer Extremes), by the European Union's Horizon 2020 research and innovation programme under grant agreement No 820712 (project RECEIPT, REmote Climate Effects and their Impact on European sustainability, Policy and Trade) and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003469 (project XAIDA, eXtreme events: Artificial Intelligence for Detection and Attribution).

5.5.1 Data Availability Statement

Table 5.4 shows the links to the publicly available data sources. The Python code and gridded USDA soy yield data (as described in section 5.2.1) are available at: <https://doi.org/10.5281/zenodo.7498927>.

Table 5.4: All data used in this study are openly available at the following data repositories.

ERA-5 sea surface temperature and soil water volume layer 1 (1979-2019)	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview
ERA-5 sea surface temperature and soil water volume layer 1 (1950-1978)	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means-preliminary-back-extension?tab=overview
US Department of Agriculture (USDA) National Agriculture Statistics Survey (NASS) Quick Stats database	https://quickstats.nass.usda.gov



6

Next steps for S2S forecasting: Scientific and Societal valorization

This chapter describes my personal vision, motivation, challenges, and opportunities for the next steps in S2S forecasting. During my PhD study, I have initiated the start of two valorization projects that aim to accelerate innovations and improve applications within the subseasonal-to-seasonal (S2S) weather prediction domain. The ambition of both projects is to have a broader societal impact. Prior to introducing these projects, I share my view on the current deficiencies that hamper the wider uptake of S2S applications and efficiency of S2S research. Firstly, there are technological barriers that hamper innovation and there is ambiguity with regard to best practices for data-driven S2S forecasting. To help tackle these issues, I propose to develop an open source software community. Secondly, the academic research domain focusses on new knowledge and innovations by presenting evidence. However, scaling these innovations and bringing them to society will require additional steps by a commercial partner, which is why I am pursuing the creation of a spin-off company.

This Chapter is based on the following two valorization activities:

- Open-innovation eScience project - AI4S2S: A high-level python package integrating expert knowledge and artificial intelligence to boost (sub) seasonal forecasting
- Creation of spin-off company

6.1 Introduction

It is my personal ambition to help "strengthening climate resilience of societies around the world by providing accurate long-term weather information". The advent of machine learning techniques shows promising results by breaking through the traditionally expected predictability limits (e.g., chapter 5). New ML innovations are likely to revolutionize the field of subseasonal-to-seasonal (S2S) weather forecasting in the coming decade. ML can be used as a stand-alone tool or ML can be used as an advanced post-processing method to debias numerical model output (sometimes referred to as 'hybrid forecasting') (van Straaten et al., 2022; Scheuerer et al., 2020). Academic research will continue to fulfill a crucial role for (high risk) innovations, yet to derive socio-economic benefits from that academic knowledge, we will need to transfer this knowledge to society by creating new valuable products and services (Van de Burgwal et al., 2018).

Hence, we enter the domain of *knowledge valorization*, defined as: 'transferring knowledge or technology to actors with an industrial or societal perspective and the concept of commercializing knowledge by adapting and developing the knowledge in order to yield socio-economic benefits' (Van de Burgwal et al., 2018). Because knowledge valorization consists of many subprocesses, which are often transcending different disciplines, the valorization process is prone to failure. Active engagement of the research expert is an important asset during such a process (van de Burgwal et al., 2019).

The motivation to start with valorization was triggered after observing, in my view, two fundamental deficiencies, which are summarized as follows:

- As the data-driven pipelines are becoming increasingly complex, transparency, reproducibility, and efficiency reduces. To successfully operationalize ML applications and build trust with stakeholders, developers must adhere to best practices, reproducibility, skill assessment, and model understanding. Such best practices may be violated to save time or because the required knowledge of ML or the climate system is lacking.
- There is a high potential societal value in using ML techniques to improve weather forecasts on S2S timescales, yet currently it remains largely an academic exercise. This largely prevents real-world adoption of these innovations.

6.2 Opportunities, pitfalls, and challenges of S2S weather forecasting using statistical models

6.2.1 Opportunities of machine learning

Reducing the noise

The S2S prediction problem is plagued by a very low signal-to-noise ratio (Krishnamurthy, 2019). The typical precursor (i.e., signal) that we are interested in, evolves slowly through time. Since time and space are correlated (Galfi et al., 2018), the typical precursor also modulates variability over large spatial domains. A physical consequence is that the

spatial extent of the response to this precursor variability is generally much larger than that of a single gridcell. The implication is that the definition of our target variable can drastically influence the signal-to-noise ratio (S/N), and thus forecast skill. The simple solution is to aggregate over the spatial (and temporal) dimension, although the spatial and temporal extent to aggregate over varies and depends on the underlying physics. Hence, without detailed prior knowledge on a S2S prediction problem, it remains difficult to know which spatial domain and what timescale will maximize the S/N ratio. Machine learning clustering techniques or mode decompositions can help to find such spatially coherently behaving regions/modes (McKinnon et al., 2016; Vijverberg et al., 2020). However, the algorithms are generally purely statistical tools and should therefore be interpreted with care.

For real-world applications, a trade-off may exist between the needs of the stakeholder and optimizing the signal-to-noise ratio. Such a trade-off might arise when a stakeholder is interested in very extreme or local events, yet the very low S/N inhibits reliable predictions of such events. One can decide to predict less-extreme events to improve the signal-to-noise ratio, or apply a window probability (i.e., define the occurrence of extremes within a time window). Chapter 3 showed that such changes to the target improves forecast skill, yet at the cost of making your forecast target less precise/extreme. This highlights that stakeholder involvement is needed to define an optimal forecast target for real-world solutions.

Improving the signal

Once a proper forecast target has been identified, one may decide to reduce the dimensionality of the precursor datasets (time, latitude, longitude). A dimensionality reduction method tries to find the relevant low-dimensional signal that exists within your high dimensional dataset, with the dimension referring to the amount of samples. This is an important concept for making reliable ML models. When all samples with a low signal-to-noise ratio are fed to an ML model, the chance that the ML model learns a spurious relationship increases. Dimensionality reduction helps with finding a few low dimensional features with a higher signal-to-noise ratio, thereby reducing increasing the chance that the model learns something that is generalizable into the future.

How the dimensionality reduction is done determines if the method is able to find the features with the higher signal-to-noise ratio. Chapter 3 & 4 show that when using a response-guided dimensionality reduction (RGDR) approach, the forecast skill can strongly increase compared to the more standard climate indices. The RGDR approach encompasses methods that reduce dimensionality of the precursor field based on a relationship to a target, instead of using some statistic of the precursor field (e.g., maximizing the explained variance). Besides this approach, there are several other dimensionality reduction approaches that show great potential for S2S forecasting (Székely et al., 2016; Deo et al., 2017; Scheuerer et al., 2020; Ham et al., 2019; Yuan et al., 2019; Bueso et al., 2020; Fery et al., 2021).

Generally, the target is capturing the statistics of synoptic weather variability (e.g., a two-week mean temperature), while the precursors capture slow components with low-frequency variability up to interannual and decadal timescales. To statistically describe a relationship between such high and low-frequency variables, long-term memory effects

can be important to consider (Miao et al., 2019; Switanek et al., 2020). This can simply be achieved by not only considering the most recent state of the low-frequency variable, but also past states. This can improve skill as the 'evolution' of the slow component can inform on the future state of the precursor (Vijverberg and Coumou, 2022). For example, if a strong El Niño state is evolving towards a neutral state in recent months, the probability that it will transition towards a La Niña state has increased, and this has implications for the expected future atmospheric response (Anderson et al., 2017). Further, the signal-to-noise ratio may be higher or re-emerge when looking further into the past. For example, Switanek et al. (2020) correlated the ENSO3.4 timeseries versus precipitation across the US. While the signal initially weakens as function of lead-time, it re-emerges in months 9 to 16, indicating increased/decreased rainfall 1-1.5 years following an El Niño/La Niña event.

The signal-to-noise ratio can also be improved by searching for windows of opportunity (Mariotti et al., 2020). Windows of opportunity are periods when the predictability for a specific region and timescale is higher than average. In principle, one could also extrapolate a similar reasoning towards the spatial dimension since certain spatial domains are more predictable than others. By aggregating over the 'correct' spatial domains (as discussed in the first paragraph), one similarly identifies a coherent region of enhanced predictability. Nevertheless, the definition of a window of opportunity is generally reserved for specific time periods. The time periods could be identified from physical (Marshall et al., 2017; Mayer and Barnes, 2019; Vijverberg and Coumou, 2022) or from purely statistical arguments (Bloomfield et al., 2021; Mayer and Barnes, 2021). These windows of opportunities can be important for the communication of confidence in a forecast. Novel ML techniques may be able to identify such time periods in an explicit way via 'controlled abstention neural networks' (Barnes and Barnes, 2021).

Impact-based forecasting

To support decision making, the potential impact of the prediction needs to be clear (Giuliani et al., 2020; Coughlan de Perez et al., 2017). Process-based models for specific domains exist that can translate meteorological variables to impact variables (e.g., crop (Brown et al., 2018; Iizumi et al., 2018) or flood risk (De Perez et al., 2016) simulation models), yet it is generally challenging to make such impact predictions reliable. Process-based models generally require outputs from dynamical models as input (i.e., requiring daily/hourly weather information). This format is poorly suited for the ML approach, which tends to focus on predicting statistics of weather to improve performance (section 5.1.1). Still, forecasting impacts can also be tackled by ML by estimating the relationship between the predictand (e.g., 'August mean temperature above the 66th percentile') and the impact (e.g., crop losses or energy demand). Thus, as done in chapter 5, ML can be used to directly predict the impact using the input features.

Besides translating weather information to impact, understanding the specific needs of the stakeholder is very important to support decision making (Giuliani et al., 2020). Farmers need information on the risk of crop harvest failure already prior to sowing in order to make informed decisions on their most effective early-action measures to mitigate crop failure (Crane et al., 2010). For example, farmers can choose to adjust their crop cultivar, adjust planting management, or buying more drought resistant seeds. In chapter 5, I

present reliable forecasts for predicting poor soy yield from ~4 months up to 8 months prior to the soy harvesting period, which is ~3 months prior to the sowing period.

6.2.2 Pitfalls of machine learning

Modelling & Verification Pitfalls (MVP)

As climate scientists are eager to explore the possibilities of ML, common technical pitfalls are easily encountered (Li et al., 2020; Vijverberg et al., 2020; García-Serrano and Frankignoul, 2014; Schauburger et al., 2020). Best practices are often computationally and coding-wise demanding to implement. This introduces technological barriers that might obstruct their implementation, especially since scientists must manage their limited time. Rather shockingly, a systematic review of 362 papers published between 2004 and 2019 on crop forecasting found that more than half did not report an out-of-sample performance (Schauburger et al., 2020). Out-of-sample verification is a fundamental pillar of machine learning. The challenge with ML is that - with sufficient degrees of freedom - models can fit its parameters such that the prediction in the training set is perfect, but this is often purely a statistical artifact and does not relate to real-world relationships. Whether the model is able to learn true relationships that can extrapolate to new unseen data is called 'generalization'. Similarly, using multiple and proper verification metrics is needed to get a complete overview of the forecast quality. By incorrect cross-validation or forecast verification one can overstate the performance of the model, thereby harming the trustworthiness of statistical methods by end-users. To give an overview and thereby hopefully facilitate better dialogues on these issues, I here attempt to categorize different modelling & verification pitfalls in S2S predictability research (Table 6.1).

Model understanding

Model evaluation is only complete once its behavior can be understood (Haupt et al., 2021). The following example where a model is trained to classify huskies and wolfs illustrates this necessity: the model misclassified wolfs for huskies if there was grass in background, and misclassified huskies as wolfs when there was a snowy background (Ribeiro et al., 2016). The model used an incorrect feature, i.e., the background, to classify wolfs versus huskies. In the same manner, we can use our physical understanding of the climate system to evaluate if we can truly trust a model (McGovern et al., 2019).

Obtaining model understanding is relatively straightforward for regression models, but for more complex models we need specialized techniques. Model interpretation and visualization methods (also known as explainable AI techniques) enables practitioners to evaluate what the model has learned (McGovern et al., 2019). This allows for physical plausibility checks, which can help with gaining trust. It can also be used to diagnose potential model errors (Lagerquist et al., 2020). This is valuable information to steer model improvements, but also for adjusting the expected reliability during real-time decision making. If the model is confronted with a situation in which it has often performed badly in the testing phase, it will likely do so again in an operational setting.

Causal inference techniques offer a powerful framework to increase the physical plausibility of statistical models (Runge et al., 2019; Runge, 2018). Selecting the features based on

Table 6.1: Modelling & Verification Pitfalls (MVP)

Label	Description
CV-0	No cross-validation is performed at all. Especially models with many degrees of freedom compared to the observational datapoints will likely overfit on the data. Consequently, the model may have learned to partly fit variability that is unique in the training sample (i.e., noise), but does not represent a real-world relationship. Hence, such models cannot be used for prediction or interpretability.
CV-1	Train and test data are separated after the dimensionality reduction and/or feature selection step has been performed.
CV-2	Train and test data are separated after the pre-processing. De-seasonalizing, standardizing, and detrending may lead to information leakage. If a strong trend is present, users need to restrain from using 'future' information by only using training data from the past to pre-process and finally predict the future.
CV-3	Information leakage may occur between adjacent train and test periods due to high autocorrelation in the timeseries.
Verification-1	Different metrics measure different aspects of the forecast quality (Wilks, 2011). For example, some metrics are insensitive to the resolution of the forecast and can therefore overstate the performance for extreme events (Vijverberg et al., 2020). Multiple metrics are needed for proper verification.
Verification-2	Reliable quantification of forecast skill requires sufficient independent datapoints and for S2S timescales, these are often limited. The number of datapoints needed depends on the timescale of interest. For seasonal timescales, roughly 30 years would be best, but this often leaves too little observational data to properly train a statistical model. More data-efficient cross-validation approaches (such as leave-one-out) are often used. Depending on the timescales, the computed skill may suffer from a sampling bias and must therefore be interpreted as an estimated skill.
Verification-3	Process understanding remains important to check the validity of the model (Toms et al., 2019; McGovern et al., 2019; Li et al., 2020). Predictability by itself is insufficient; trust in a forecast and understanding of the physical processes, is essential to enable early action (Diffenbaugh et al., 2015; McGovern et al., 2017). The performance in an operational setting (i.e., generalizability) may be compromised if the model has learned relationships that existed in the training dataset, but are not causally related.

causal theory mitigates overfitting and thus can improve generalizability (Kretschmer et al., 2017a; Di Capua et al., 2019a). The framework also allows users to diagnose the model by calculating causal effects of features (Kretschmer et al., 2021). In practice, the objective of causal modelling may interfere with the objective of obtaining high forecast skill. For example, causal discovery attempts to eliminate indirect links, although these indirect links might be better predictors compared to the direct links. Chapter 5 illustrates the latter, as we have used winter soil moisture as a predictor for poor harvest years, whereas, the winter soil moisture pattern is not directly detrimental to crop growth in spring and summer. The winter soil moisture pattern captures the presence of dominant winter circulation, which is known to amplify the north-Pacific sea surface temperature pattern¹. Due to the higher memory of soil moisture, it acts as an integrator of circulation states during winter, which increases its signal-to-noise ratio. Conversely, circulation is much more chaotic which reduces the signal-to-noise ratio. In this scenario, it is better to use the indirect soil moisture precursor due to its higher signal-to-noise ratio. Another limiting factor of some causal inference or discovery methods is that it may require to specify a certain timescale of interest (Runge, 2018), whereas S2S problems often involve a combinations of features with different processes/timescales. This may lead to the exclusion of variables that may be important for predictability. To prevent the exclusion of important precursor variables, we applied a less-strict causal inference-based feature selection in chapter 5.

Reproducibility and transparency

With the ever-increasing complexity of data-driven software pipelines, it is becoming more difficult to reproduce and evaluate the quality of the results. For this reason, there is a need for more transparent analyses. This can be achieved by setting community standards for software documentation and implement reporting-standards that enable the detection of potential 'model and verification pitfalls', that may be present in the analysis (Table 6.1). Open-source software plays an important role here, as it can simplify the technical task, improve readability of the software pipeline, and enhance reproducibility via higher quality documentation.

6.2.3 Limitations and challenges

As shown in this thesis, we can infer the (causal) drivers of a target of interest purely from data. However, the techniques used in this work have their limitations and underlying assumptions. The theoretical assumptions of causal discovery are discussed in Runge (2018), whereas in Table 6.2 I list specific assumptions taken that are encountered in this thesis work. Our ability to learn from data is of course also limited by the number of independent datapoints we have available, which makes the detection of synergistic and conditional effects discussed in 6.2 more difficult. I believe the assumption that a signal propagates at a specific lag and timescale are important aspects to consider. Both these limitations can be dealt with by e.g., considering multiple lags, or by analyzing

¹As shown in Chapter 4, the horseshoe-shaped north-Pacific sea surface temperature pattern is a causal driver of crop growth as it promotes the occurrence and frequency of high-pressure systems over the crop planting domain in the mid-to-eastern US which leads to enhanced risk of detrimental hot-dry conditions.

the dependencies on multiple timescales, yet a more elegant solution to find the optimal number of lags and timescale(s) is (to my best of knowledge) still missing.

6.3 Way forward

To avoid the pitfalls of machine learning for S2S forecasting (section 6.2.2), we need clear community-guidelines on best practices and lower technological barriers that might obstruct the implementation of these best practices. In section 6.3.1, I argue that high-level open-source software, that is broadly supported by the S2S community, can help with this challenge. Besides, there is also an untapped potential for applying machine learning techniques. In section 6.3.2, I argue that a spin-off company, with strong links to academic research, can help to bring innovative ML-based S2S forecasting to society.

6.3.1 Scientific valorization: High-level community software

Data-driven S2S-forecasting needs to become trustworthy to have a lasting impact on science and society. To gain such trust, it will be essential to establish guidelines and standard terminology together with the global S2S community to enable:

- Transparent analyses that can efficiently be reproduced, preferably across different big-volume datasets without the need to download and store data locally.
- Implementation of community-wide accepted best practices on cross-validation and skill assessment using common benchmarks and evaluation metrics to enable quantitative comparisons between studies.
- Understanding of the sources of predictability and underlying physical mechanisms.

To achieve those goals, I initiated together with colleagues, the eScience Center, and the wider community, the 3-year AI4S2S project: 'A high-level python package integrating expert knowledge and artificial intelligence to boost (sub) seasonal forecasting' (<https://github.com/AI4S2S>), which started in spring 2022. The goal of the project is to promote the usage of our software that supports the S2S community with best practices for ML in a well-documented and user-friendly manner. At the start of the project, we made a questionnaire to get a better understanding of the needs of the S2S community. In this questionnaire we asked climate/data scientists "How important is it that the AI4S2S community promotes and defines a set of best practices for ML in S2S research?". 13 out of the 14 participants answered 4 or higher on a scale of 1 to 5, with 1 being 'not important at all', to 5 being 'very important'. This clearly shows the widely felt need for such guidelines. In terms of software needs, it was clear that the package should be modular, and integrate well with other commonly used ML packages. For that reason, we are currently following the widely used scikit-learn code structure, definitions, and API (Pedregosa et al., 2011). The packages will be developed such that they can easily leverage resources of high-performance computing (HPC) systems (Google Earth Engine, Planetary Computer by Microsoft or internal HPCs). The development will be done together with research software engineers from the eScience center. The objectives are to:

Table 6.2: List of assumptions and associated limitations/consequences of the data-driven methods that were used in this thesis to learn the causal drivers of a target from data.

Assumptions	Physical examples (limitations/consequences of assumption)
Linearity	Relationship between low soil moisture and surface temperature is often better described by a non-linear relationship. Surface temperature only responds to low soil moisture levels once the balance between available heat and available moisture is imbalanced (Seneviratne et al., 2010).
No synergistic effects	The existence of synergistic effects could mask or underestimate a relationship, for example, heat is more damaging to crops during drought conditions (Hamed et al., 2021).
No conditional effects	Modelling experiments showed that the atmospheric response to a sea surface temperature anomaly depends on the strength of the jet stream aloft (Thomson and Vallis, 2018).
Signal propagating from a specific lag (Markov process assumption)	Our approach in chapter 4 assumes the mean state of a precursor carries the signal, which is not always true. For example, transitional El Niño Southern Oscillation states have a different imprint compared to the classical bimodal definition (Anderson et al., 2017). A signal might emerge (more strongly) when considering a certain evolution of the climate system, as was also shown in chapter 4 when fitting a model with multiple lags. This evolutionary effect was not yet taken into account by the causal discovery framework in a satisfying manner. Another related consequence of the Markov process assumption is that features can be falsely rejected when analyzing the causal link between a low-frequency driver (x) and a higher frequency target (y). Due to the high autocorrelation in the low-frequency driver, two consecutive timesteps are more likely to be very similar (not independent). Consequently, computing the following conditional independence test $parcorr(x_{t-1}, y_t x_{t-2})$, will likely result in the conclusion that x_{t-1} is independent from y_t . Similarly, $parcorr(x_{t-2}, y_t x_{t-1})$ will likely result in the conclusion that x_{t-2} is independent, while in fact, y_t might be causally influence by x_{t-1} and x_{t-2} .
Signal propagating at a specific timescale	Focusing on a single timescale can lead to an incomplete set of drivers, as multiple drivers with different timescales could be important for variability in the target and therefore affect the probability of extremes. However, the causal discovery algorithm used is designed to analyze the causal links at a specific timescale.

1. Streamline the workflow by optimal handling of massive amounts of data that is scalable to any HPC, a requirement for it to become long-term sustainable and community-driven.
2. Provide a robust framework that avoids duplication of common core tasks and is tailored towards the unique aspects of climate data and specific needs of the S2S community (including proper verification metrics and community-approved benchmark forecasts) to foster sharing and collaboration.
3. Remain sufficiently versatile by creating modular packages that deliver micro-services forming the skeleton of the high-level S2SAIpy software, adopting commonly used input/output standards, and enabling scientists to implement their own pipelines.
4. Provide an efficient interface between AI and expert knowledge to co-develop skillful forecasts and physical understanding, enabling an expert-knowledge driven search for 'windows of predictability'.

Lowering the technical barriers to scale-up analysis will also be an important building block for more advanced AI techniques that can leverage the deluge of climate model data that is available. Such computationally high-demanding activities, such as training complex AI models (partly) on climate model data, has already proven to be a valuable approach (Ham et al., 2019; Gibson et al., 2021).

We hope that in the future, our package will be so useful that it will be widely used by the S2S community. By building upon the skeleton of the S2S package, reproducibility is enhanced, and following best practices is greatly simplified and promoted. Users can build upon ML techniques that are supported by the package, or they may choose to implement their own method(s), giving them all flexibility for new method development.

6.3.2 Societal valorization: Spin-off company

For data-driven S2S forecasting to have a lasting impact on society, commercialization of scientific innovation is helpful. The motivation to start a spin-off company is four-fold. Firstly, I believe that recent innovations for data-driven S2S forecasts now have a competitive advantage against the widely used traditional approach based on dynamical models. Secondly, there is a need for a valorization vehicle of scientific innovations in this field. Scientific results may be left untouched after research grants stop. Thirdly, scientific results are often focused on generalizability and/or serve as a proof-of-principle. The societal impact of academic research is therefore hampered because it tends to compromise at the cost of usability. In other words, findings are not tailored to the needs of a specific stakeholder. Fourth, the academic proposal-research-publishing cycle is not suited for the next phase of scaling-up and speeding-up S2S innovations. A substantial part of a researcher's time is spend on writing proposals and scientific communication (papers, presentations, etcetera), while for scaling-up emphasize shifts toward building the software infrastructure and finetune S2S products.

To explore this valorization option, I joined the Demonstrator Lab (www.demonstratorlab.nl) at the Vrije Universiteit Amsterdam, which offers an entrepreneurial laboratory for academics. Together with Jannes van Ingen and Dim Coumou, we are starting the spin-off

company 'Beyond Weather' (www.beyond-weather.com). Via demonstrator lab, Beyond Weather has worked on branding themselves by participating in the Amsterdam Sustainability & Innovation Award. Furthermore, Beyond Weather developed its entrepreneurial strength by participating in the ACE incubation program (<https://ace-incubator.nl>) and via a first pilot study together with an Amsterdam-based energy trader. Further, Beyond Weather is working towards improved drought predictions in Mozambique in collaboration with the World Food Program (www.wfp.org).

Long-range weather forecasts can serve many different sectors, such as the agricultural, insurance, humanitarian aid, and the energy sector. However, focus is needed when entering a new market. We decided to focus on developing products for the energy sector. This is not without reason since we believe that mitigating climate change is the most urgent global challenge of our time. By improving weather predictions on S2S timescales we hope to avoid waste of resources while maintaining energy security, thereby speeding up the energy transition. The Beyond Weather team is and will be strongly rooted within academia and shares its ethical values. By implementing differential pricing, our revenue will be generated from profit-oriented stakeholders, thereby enabling us to serve non-profit organizations against reduced, non-profit, tariffs. Hence, Beyond Weather is pursuing the concept of 'shared value' creation, where activities will benefit both the company and society (Dembek et al., 2016).

Energy Transition & Climate Extremes

To reach the Paris agreement climate targets, the energy sector will have to go through a massive transition in the coming 10 years, with renewables becoming the dominant source of energy. In the Netherlands, the share of renewable electricity (wind, solar and hydro) will increase from 18% (in 2019) to a whopping 75% in 2030 (PBL et al., 2020). The Netherlands is an integral part of the western European power market, which includes the United Kingdom, France, and Germany - all countries that have set similar targets for renewable energy production. Electricity production will thus become more and more dependent on weather conditions, increasing the need for reliable long-range forecasts.

The 2022 Ukraine-Russian war has drastically changed the landscape of the gas and power market. With the diminished supply of natural gas from Russia, Europe is exposed to record-breaking gas and power prices, extreme price volatility, and high monetary inflation. To increase EU gas reserves, the European Commission has released a plan to reduce energy consumption by 15% (EU, 2022). Whether the upcoming winter will be colder or warmer than normal will drastically affect our gas reserves. However, heavy industries are still buying gas and power for the upcoming winter to ensure security of supply, yet at very high prices. These high costs are seeping through society, thereby increasing the risk of energy poverty in vulnerable communities. Hence, both from the political and industry arena, the interest in temperature forecasts for the upcoming winter has never been higher. Finally, climate change has already increased the frequency and/or intensity of extreme weather events that impact the energy sector, and this is projected to continue over the next decades (Birol, 2021). For example, heatwaves are strongly increasing with global warming (Robinson et al., 2021), and they put particular pressure on electricity systems by increasing demand (e.g. air-conditioning) and squeezing electricity supplies (Magagna et al., 2019). For example, in the summers of 2018 and 2022, several nuclear power plants






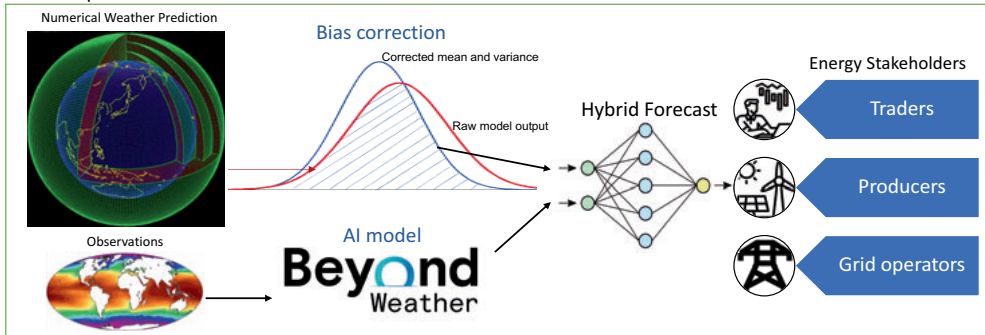
Cascading impact on stakeholders 				
	Economical		Sustainable/Societal	Societal
Weather events	 Energy producers	 Traders	 Transmission System Operators (TSO)	 Industry/Households
Decision making	<ul style="list-style-type: none"> ✓ Optimize planned maintenance ✓ Optimize hedging (selling future energy) ✓ Manage resources 	<ul style="list-style-type: none"> ✓ Trader buy/sell 	<ul style="list-style-type: none"> ✓ Plan maintenance to minimize instability risks ✓ Buy and manage dispatchable energy ✓ Prioritize clean energy 	<ul style="list-style-type: none"> ✓ Governmental legislation mitigating energy poverty. ✓ Manage financial risk ✓ Manage resources
Problems	<ul style="list-style-type: none"> ✗ Supply & demand and thus future prices are uncertain, posing huge economic risks. 		<ul style="list-style-type: none"> ✗ Elec. transport losses ✗ Damage to power lines ✗ Power plant cooling issues ✗ Large spatial disbalance in supply/demand 	<ul style="list-style-type: none"> ✗ Energy poverty ✗ Risk of shutting down big industry

Figure 6.1: Energy supply/demand discrepancies due to climate extremes can cause a cascade of problems for energy stakeholders. Skillful weather forecasts enable improved decision making to mitigate the economic and societal impact and make the energy system more resilient.

in France and Germany were temporarily enforced to reduce capacity due to the lack of cooling water and also hydropower output and stocks were reduced (Magagna et al., 2019; Paulson et al., 2022). The same hot-dry weather conditions caused a 'wind drought' with wind electricity 20% below expectation (Richard, 2018). Heatwaves therefore pose major risks for electrical grid failures, or 'black outs', which have already increased by over 60% in the United States over the last five years only (Stone et al., 2021). Also other seasons and extremes, e.g. cold winter spells, create major risks for the electricity sector (Coumou, 2021; Van Der Wiel et al., 2019).

The combined trends of a much stronger reliance on weather dependent renewable energy sources and more frequent weather extremes, results in more severe volatility on the electric grid, bringing enormous challenges for all stakeholders in the energy sector. Weather forecasts provide indispensable information for a safe, efficient, and sustainable operation of the power grid. However, the sub-seasonal to seasonal weather forecasts, as currently used in the energy sector, are not skillful enough to support decision making. Skillful forecasts several weeks to months in advance are urgently needed to support stakeholders in the energy sector to optimally plan production, maintenance, manage hydro-power reservoirs and distribution of electricity, so as to make our future energy systems resilient and sustainable. Figure 6.1 shows the current problems of energy stakeholders and the improved decision making that would be possible with skillful long-range predictions. For example, energy producers would be able to improve resource management, and Transmission System Operators (TSOs) may be able to better anticipate the amount of dispatchable energy (energy sources which are used to balance the grid during times of e.g., low wind production).

Scalable products



Stakeholder prediction products

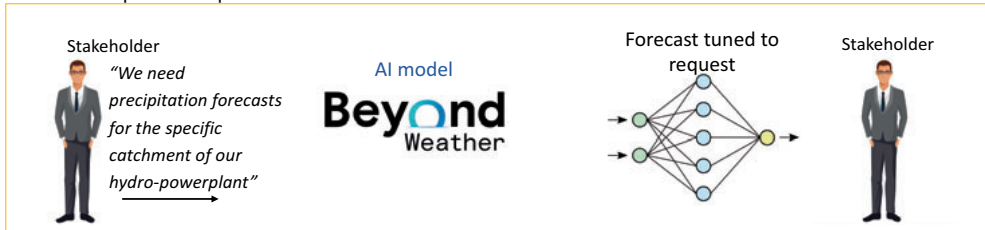


Figure 6.2: Schematic of Beyond Weather software that will allow for fast deployment to meet the needs of various stakeholders/sectors. *Scalable products* will merge numerical weather prediction model output with the Beyond Weather methods to benefit from the predictive signal in numerical models. The forecasts will be created for entire continents (e.g., European or US temperature) such that they are relevant for many different stakeholders, i.e., scalable. *Stakeholder predictions* is using the same software in a versatile manner to fine-tune towards a specific request of the stakeholder.

Products

Our AI-driven approach is fundamentally different and more computationally efficient compared to traditional weather forecasting models. The data-driven method can circumvent some of the errors made by traditional weather models. Especially the efficiency of data-driven methods to search for windows of predictability helps with finding predictability, with some periods being ostensibly more predictable than historically thought. We aim to develop a hybrid forecasting product that merges the output of dynamical models with our data-driven approach to ensure that our forecast will have at least the same skill as the dynamical model (Figure 6.2). This approach can for example be used to predict temperature on continental scales (Europe/North America). Since temperature remains one of the leading proxies for energy demand and production, we aim to sell this product to stakeholders across different domains within the energy sector (traders, energy producers and grid managers). Some stakeholders will have specific needs. For example, the water influx for hydro-powerplants will largely depend on the precipitation that falls within its catchment. In such cases, the performance may improve when tuning the model to predict rainfall within that specific catchment. The Beyond Weather software is developed in a scalable way such that these 'Stakeholder prediction products' can be optimized efficiently and without much additional effort. Our predictions could be disruptive for the current modus operandi of stakeholders in the energy sector.

S2S forecasting challenges	Open-Source Software	Spin-off company
<i>The signal-to-noise challenge</i>	✓	✓
<i>Statistical modelling of slow components</i>	✓	✓
<i>Support for decision making</i>	✗	⊕
<i>Modelling & Verification Pitfalls (MVP)</i>	⊕	✓
<i>Model fidelity and understanding</i>	⊕	✓
<i>Reproducibility</i>	⊕	✗
⊕ Focus of project ✓ Tightly connected to aims of project ✗ Not related to focus/aim of the project		

Figure 6.3: How the valorization projects contribute to the S2S challenges as introduced in section 6.2.2.

6.4 Conclusion

With the two parallel valorization activities described in the previous section, I hope to contribute to the challenges introduced above and listed in Figure 6.3. One is aimed at the scientific community, and one at the society at large. I believe both will be important for future innovations and uptake of S2S forecasts. The open-source software project will focus on lowering technological barriers, and standardization of best practices in data-driven S2S research. The spin-off company will act as a valorization vehicle of scientific research and make a societal impact by bringing the S2S predictions to stakeholders. Both projects have the overarching aim to speed-up AI innovations. The two activities can benefit from each other, but there are also potential conflicts that will need to be managed. For example, developing open-source software (essential for scientific community and thus future innovation) can conflict with the companies competitive edge. Although developing part of the companies software open-source increases visibility and trust among potential clients, it could weaken the companies technological advantage. Where to draw the line between open-source and commercial software?

Synthesis & Outlook

In this thesis, I tried to cover some major aspects that are needed to develop successful data-driven subseasonal-to-seasonal (S2S) forecasts. In chapter 2, I first reviewed evidence for dynamical (i.e., circulation) changes due to heterogeneous global warming. Gaining insight into the underlying dynamical processes, such as the interaction between jet stream dynamics and the ocean, is important for both improved regional projections as well as understanding sources of predictability on subseasonal-to-seasonal (S2S) timescales. In all subsequent chapters, I focused on improving the use of data-driven methods to gain new insights, with a particular focus on boosting subseasonal-to-seasonal (S2S) forecast skill and a better understanding on the role of mid-latitude ocean-atmosphere interaction. I developed statistical forecasts for the eastern United States, with the chapters focusing on different aspects, including: statistical pitfalls (chapter 3), the underlying physical processes (chapter 4), and impact-based forecasting (chapter 5). In chapter 6, we have identified opportunities and pitfalls of data-driven methods and have proposed two valorization activities to address these. Furthermore, we elaborate on these new activities, i.e., an open-source software project and a spinoff company. This Synthesis & Outlook summarizes the answers to my research questions stated in the introduction, and I share suggestions for future research.

7.1 Main findings

With the research presented in the preceding chapters, I have attempted to answer the research questions posed in the introduction. To facilitate a coherent synthesis, my research questions all focus on S2S dynamics and predictability, which we repeat here for convenience.

How can we improve the sub-seasonal forecast skill for eastern United States heatwaves?

Predicting extreme weather skillfully has been a major challenge of weather forecasting. Reliably, and precisely predicting very extreme events on subseasonal to seasonal timescales is still out of reach. Nevertheless, we can improve upon how we deal with the inherent chaos in the climate system. Although its chaotic nature will always impose a limit to predictability. The work in chapter 3 was inspired by earlier work by McKinnon et al. (2016), that hinted at remarkable forecast skill for daily temperature extremes in eastern United States (US) based on a specific sea surface temperature (SST) pattern in the north-Pacific. I introduced a new algorithm that can automatically extract robust SST patterns associated with temperature extremes. By rigorous subsampling of temperate extreme composites, the algorithm found a similar, but slightly, different SST pattern compared to the pattern that was manually identified by McKinnon et al. (2016). We reproduced the SST pattern by McKinnon et al. and compared the forecast skill against our algorithm.

First, we found that the verification of McKinnon et al. (2016) did not cover all steps needed to assess the quality of a forecast. After using a proper verification metric, we concluded there was in fact no forecast skill for *daily* temperature extremes. The earlier work only measured the discrimination of the forecast (Wilks, 2011), which quantifies whether the specific SST pattern was more often than not pronounced during heatwaves. However, there are also many situations in which the SST pattern was pronounced, without the occurrence of a heatwave. When quantifying the forecast reliability we found large inconsistencies (Wilks, 2011), i.e., the predicted probabilities from the statistical model did not match the observed frequency. Based on this, we conclude that, while there is a signal, we cannot reliably predict when *daily* extreme events will happen.

Next, I showed how one can develop skillful probabilistic forecasts for extreme events on S2S timescales for eastern US. Temporal and spatial aggregation, which is a common step to filter out the unpredictable high-frequency weather variability (i.e., noise on S2S timescales), enhances skill. However, the tail of the distribution was still not predictable using only information from sea surface temperature. Consequently, we adapted the event definition to allow for temporal flexibility and were able to skillfully predict moderate heatwaves (two days exceeding 1 standard deviation above climatology within a 3-day period) within a two-week window, with a lead-time up to 50 days. We additionally found that adding Soil moisture precursor timeseries can increase skill up to 30-days lead-time.

In summary, chapter 3 captures a variety of challenges that are associated with the S2S forecasting. These challenges are: extracting the relevant precursors, dealing with noise in the target timeseries, using a complex versus simple statistical model and dealing with the inherent uncertainty of the climate system. It also highlights that an adjustment of the

target variable might be necessary to enable the uptake of S2S forecasting. A stakeholder might be interested in the most extreme weather events, but it may be necessary to relax the temporal/spatial precision and/or extremity of the events to gain a more reliable prediction. Finding an optimal balance between these trade-offs requires a proper dialogue between stakeholder and forecast expert.

What is the role of ocean-atmosphere feedbacks in generating predictability for eastern US summer temperature?

The physical mechanisms that explain the long-lead predictability of eastern US temperature are not fully understood. The associated Rossby wave, with a high-pressure system over the eastern US, clearly plays an important role in promoting high temperatures. Such high-pressure systems are associated with enhanced incoming solar radiation, soil desiccation, and reduced advection.

However, the direction of causality between the Rossby wave (RW) and sea surface temperatures was unclear. Causal discovery, in combination with our proposed framework, presents a new way to investigate the causal direction and coupling strength between atmospheric circulation and the ocean. The framework can quantify the coupling strength for different modes of atmospheric circulation and infer a feedback or the directionality of the forcing (ocean forcing the atmosphere or atmosphere forcing the ocean). So far, quantifying the strength of the ocean-atmosphere interaction has mostly been done using climate model experiments (see review of Zhou (2019)), with some exceptions using statistical (correlation/regression) metrics on observations (Frankignoul and Sennéchaël, 2007; Liu et al., 2006). However, it is well known that the latter statistical metrics cannot be used to infer the causal direction (Runge et al., 2014). On the other hand, modelling experiments are generally used to study the (unidirectional) forced response to SST forcing. Using models to disentangling the two-way interaction requires expensive and complex experiment designs (such as Koster et al. (2006) and Wehrli et al. (2021)), which have not yet been performed to study the ocean-atmosphere coupling strength.

We found that the eastern Rossby wave is forced by low-frequency SST variability, and this variability is closely linked to the Pacific Decadal Oscillation (PDO). This insight led to a change in the design of our forecasting model, allowing the model to better consider the *evolution* of the low-frequency SST variability when making predictions. In addition, we were able to identify if the model would perform better or worse, depending on the past PDO state. If the winter-to-spring PDO state is pronounced and persistent, the summer will be much better predictable. Thus, by disentangling the physical mechanisms, we were able to improve upon the statistical model design and identify a window of enhanced predictability.

Can we predict harvest failure in the eastern US with sufficient lead-time to help farmers to take anticipatory action?

To mitigate the impact of detrimental weather, some anticipatory actions have strict windows of opportunity. Once such a window of opportunity has passed, the intervention can no longer be made. This is also true for soy farmers in the United States. Effective early-actions that local farmers can undertake are to better manage irrigation schedules (Villani et al., 2021), buy insurance against crop failure (Li et al., 2019), lower the sowing

density (Carter et al., 2018; Lobell et al., 2020), decide to only plant in lower (i.e., wetter) altitude areas (Crane et al., 2010), or decide to order more drought resistant crops or soy cultivars (Dong et al., 2019; Arya et al., 2021; Crane et al., 2010). The latter three interventions are no longer possible once the crops have been planted. Currently, there are no systems in place that can reliably predict the risk of crop failure so far in advance.

Soy yield is sensitive to hot-dry weather in summer (Hamed et al., 2021), and chapter 4 showed that Rossby waves - associated with a high-pressure system over the soy production region - are promoted by the low-frequency North-Pacific horseshoe-like SST pattern. These insights suggests that forecasts of crop failure at long lead-times may be possible. First, a clustering algorithm was used to separate the mid-to-southern US from the northern producing regions, which is ostensibly the result of the higher sensitivity of the mid-to-southern region to hot-dry weather. Focusing on the aggregated mid-to-southern US cluster, feature timeseries were extracted from both a sea surface temperature and soil moisture dataset using the Response-Guided Dimensionality Reduction (RGDR) method (see section 5.2.4). To gain trust in the forecast model, I applied a causal inference-based selection step to filter out spurious precursor timeseries. The selection step was adapted from the PC algorithm (named after its inventors Peter and Clark) (Runge et al., 2019; Sprites et al., 2001). In the implementation I excluded the possibility to test for conditional independence that would result from auto-correlation. I wanted to retain the highly auto-correlated features and use multiple lags to better capture the low-frequency variability of these features. The motivation is explained in more detail in chapter 5 and section 6.2.1. In addition, the selection step was purposefully made not very strict, as the machine learning model performs better when having a few additional spurious features (i.e., false positives) rather than missing important causal drivers (due to e.g., a sampling bias or violation of assumption(s)). During windows of opportunities, for some US states located in the mid-to-southern domain, the forecast models can reliably predict poor (1-in-3) soy harvest years already in February. On the larger aggregated mid-to-southern cluster spatial scale, the skill is substantially higher due to a better signal-to-noise ratio, achieving a precision of $\sim 75\%$ (versus a 33% precision of a random forecast) already in February, which is 3 months prior to sowing and 8 months prior to harvesting.

How can we promote new AI innovations for subseasonal-to-seasonal forecasting and how can we bring those to society?

Current potential applications hinge on the limited quality of S2S weather forecasting services. Many applications fail due to insufficient forecast skill (Coughlan de Perez, 2018; Vigo et al., 2019) and stakeholders suffer from the inability of weather services to clearly communicate on the skill of the forecast (Vigo et al., 2019). Machine learning (ML) techniques can be great tools to overcome these challenges, yet there are deficiencies in the current modus operandi. To provide more context to the reader, chapter 6 summarizes opportunities and pitfalls of using ML for statistical weather forecasting on S2S timescales. Machine learning pitfalls may originate from improper cross-validation, incomplete model verification, or low reproducibility due to the complexity of data-driven pipelines. Therefore, the AI4S2S project will focus on building a stronger data-driven S2S research community and open-source software to support best practices. By having good documentation, continuous maintenance, and an intuitive design, I believe open-source software can help lower the technological barrier to adopt best practices and improve upon transparency

and reproducibility. By using the open-source software as a skeleton for the pipelines we will develop within the spin-off company, we ensure future support for that software beyond the current AI4S2S project deadline.

As we increase our ability to incorporate the relevant low-frequency processes for S2S forecasting, machine learning opportunities are becoming more evident. The ability to (1) learn causal drivers, (2) gain deeper insights by using explainable AI techniques, and (3) search for windows of predictability are important assets of the data-driven approach. Our results suggest that the skill of the data-driven approach is at least competitive versus the traditional dynamical modelling approach. I, together with Jannes van Ingen & Dim Coumou, will launch a spin-off company to facilitate the operationalization of state-of-the-art ML solutions, ensuring that S2S innovations are brought to society.

7.2 Directions for future research

Within the domain of climate predictability, understanding, and projections, there is still a tremendous amount of research to be done. Here, I restrict myself to highlighting important research that could readily follow from the knowledge and software developed during this PhD project. Ongoing plans for data-driven S2S forecasts are outlined in chapter 6, here I touch upon other future opportunities and knowledge gaps.

7.2.1 Systematic evaluation of ocean-atmosphere interaction

The ocean is an important source of low-frequency variability in the atmosphere (Vijverberg et al., 2020; Vijverberg and Coumou, 2022; Li et al., 2016; Di Capua et al., 2021), yet the ocean-atmosphere coupling strength in many climate models appears to be underestimated (Vijverberg et al., 2020; Simpson et al., 2018; Li et al., 2016; Eade et al., 2014; Simpson et al., 2019; Sillmann et al., 2017; Cheung et al., 2017; Vannitsem and Ghil, 2017; Tsartsali et al., 2022). Climate models systematically underestimate the co-variability between low-frequency atmospheric modes of circulation, and the North Pacific and North Atlantic oceans, on decadal to multi-decadal timescales (Cheung et al., 2017; Vannitsem and Ghil, 2017). A growing body of evidence states that the real world is more predictable than currently suggested by dynamical models (Vijverberg et al., 2020; Simpson et al., 2018; Li et al., 2016; Eade et al., 2014; Simpson et al., 2019; Sillmann et al., 2017; Cronin et al., 2019; Smith et al., 2020), however, the exact reason for this is not fully understood. I expect that a deeper understanding of the role of ocean-atmosphere coupling in both observations and climate models will shed light on the underestimated predictability on longer timescales.

It is difficult to simulate and validate ocean-atmosphere coupling in dynamical models, as it emerges from the micro-scale air-sea interaction and small-scale atmospheric eddies (Robert et al., 2017). Both the small-scale eddies and air-sea interactions need to be parametrized. This is because (1) (unresolved) oceanic eddies with sharp SST gradients affect the air-sea fluxes and (2) small-scale (unresolved) atmospheric eddies are important for transferring these air-sea fluxes from the planetary boundary layer to the troposphere. These parametrizations suffer from biases (Demeyer and Vannitsem, 2018; Yu, 2019; Centurioni et al., 2019), yet the collection of parametrized processes are effectively deter-

mining the ocean-atmosphere coupling. A higher resolution can help reduce the portion of smaller-scale features and eddies that need to be parametrized (Tsartsali et al., 2022). For example, insufficient spatial resolution was the main reason of early modelling studies reporting a negligible effect of the ocean onto the atmosphere (Zhou, 2019) and it still appears to be a limiting factor in modern CMIP models (Van Der Linden et al., 2019; Haarsma et al., 2015; Tsartsali et al., 2022).

As the ocean-atmosphere coupling strength is important for predictability and future projections, there is a clear need to quantify and validate the coupling strength within dynamical models with greater detail. Current methods to study ocean-atmosphere coupling in observations cannot disentangle cause and effect (Cheung et al., 2017; Vannitsem and Ghil, 2017). Understanding the strength of ocean-to-atmosphere forcing is particularly important for predictability. Implementing a causal discovery algorithm on both observations and the latest CMIP models would shed light on this aspect.

7.2.2 Causality and machine learning methods for future impact and risk projections

The need for reliable regional climate projections is growing as governments, institutions, and non-governmental organizations are increasingly taking climate change into account when deciding their long-term strategy. In the past, a climate change signal was calculated using a 'model democracy' approach, i.e., give equal weight to each model. However, it is now more widely recognized that this is suboptimal due to the high interdependencies between models and due to model errors (Hegerl et al., 2021; Knutti, 2010).

Climate change signals could be either over- or underestimated due to model errors, and it is therefore important to understand the fidelity of the climate model before assessing its value. For example, studying the atmospheric response to sea ice loss requires models to have a reliable eddy feedback¹ (translating momentum of eddies in the planetary boundary layer to the free troposphere), which is generally underestimated by climate models (Smith et al., 2022). If models are constrained based on their eddy feedback, a robust (albeit weak) mid-latitude atmospheric response emerges due to the projected reduction in sea ice (Smith et al., 2022). Similarly, to study potential circulation changes over Europe, the models need to be able to simulate both the eddy feedback and the sea surface temperature (SST) changes associated with the weakening Atlantic Meridional Overturning Circulation (AMOC). Using a higher resolution model to more reliably simulate the atmospheric and ocean eddies and the concomitant SST gradients, Van Der Linden et al. (2019) found more pronounced circulation changes over the North Atlantic domain, leading to a more pronounced future drying over central-western Europe. The latter study supports the hypothesis that models tend to underestimate the ocean-atmosphere coupling strength, or that it is at least resolution dependent.

Future projections can be improved using a 'process-informed bias-correction' (or colloquially known as 'optimal weighting'), i.e., a correction that uses (an) observed relationship(s) and validates if the climate model simulates the same relationship(s) (Maraun et al., 2017;

¹The eddy feedback refers to the fact that the mean zonal flow can generate eddies, and these eddies can strengthen or weaken the core of the mean flow in such a way that more or less eddies will be generated by the mean flow, i.e., it can be either a positive or a negative feedback (Robert et al., 2017; Robinson, 2006).

Eyring et al., 2019). Subsequently, only the more reliable climate models can be selected (or weighted) to produce a constrained future projection. For example, models that show a too strong land-atmosphere coupling also show a stronger future warming, and vice versa (Vogel et al., 2018). Particularly models with a too strong land-atmosphere coupling simulate the strongest future warming (Vogel et al., 2018).

So far, constrained future projections are not yet tuned towards specific stakeholder needs. For example, (non-)governmental organizations might want to design adaptation measures against drought in the Netherlands or wildfire risk in the US. These two targets have widely different processes that should be considered for creating a constrained future projection. I propose to use causality and machine learning techniques to check whether climate models can simulate the relevant causal relationships accurately, tuned towards specific questions posed by industry or (non-)governmental organizations. The reliability of a constrained future projection can be tested by quantifying if the corrected climate model ensemble is better at capturing the already observed climate change trend (Maraun et al., 2017). However, such experiments are currently time-consuming, and thus expensive.

As we gain experience in learning the drivers of a target purely from data, I envision that in the next 5 years, we might be able to learn which models we can trust in a very time- and resource-efficient manner. This step-change in efficiency, in combination with a robust framework, could enable improved climate projections upon request that are tuned towards stakeholders' specific target.

7.2.3 Seasonal-to-Decadal predictions

Improving physical understanding and statistical and modelling tools is valuable for subseasonal-to-decadal predictions. In section 6.2, we already reflected on the opportunities, technological pitfalls, and challenges of data-driven techniques, focusing on subseasonal-to-seasonal dynamics and predictability. As introduced in section 7.2.1 and 7.2.2, learning the (causal) drivers from data opens-up many new opportunities. Here, we would like to highlight the opportunity for seasonal-to-decadal (S2D) predictions.

Especially tropical and sub-tropical regions are vulnerable for trends in droughts, heatwaves, wildfire risk, floods, energy demand, and risk of conflict in the next ~5 years (Hermanson et al., 2022). Societal vulnerability in tropical and sub-tropical regions is often large. Seasonal-to-decadal climate predictions will also be highly relevant for the extra-tropics, as here the role of internal climate variability versus the externally forced climate trends is relatively large (Collins et al., 2018; Xie et al., 2015; Deser et al., 2014; Shepherd, 2014). The predictability on these timescales stems mainly from the ocean, as well as external climate forcing (e.g., greenhouse gases, aerosols and solar variability) (Kushnir et al., 2019). The World Meteorological Organization (WMO), started to produce S2D predictions using an ensemble of different climate models (released every year) since 2017 (Hermanson et al., 2022). Currently, computationally heavy large ensembles, in combination with post-processing, are used to deal with the weak signals in dynamical models (Hermanson et al., 2022; Smith et al., 2019). Given the ocean variability on S2D timescales, the hypothesized underestimated ocean-atmosphere coupling strength is of particular importance for the predicted surface impact in the next decades (Cheung et al., 2017).

Climate models and advanced data-driven methods might be able to compensate each other's flaws. Data-driven methods can have trouble learning the annual-to-decadal

dynamics of low-frequency precursors given the short observational record. Dynamical models are likely better at this task, where bias-correction techniques can help with improving the forecast skill (WMO, 2020). On the other hand, dynamical models tend to underestimate particularly the remote impact of low-frequency processes, with potential implications for underestimated predictability of drought or surface temperature. This artifact was also found in chapter 3, where the predictability of eastern US heatwaves, based on sea surface temperature, within the climate model EC-Earth (with more data available) was lower than what was observed in reanalysis data. The hypothesis is that predicting the surface imprint of low-frequency processes is more successful when using data-driven methods (if sufficient training data is available). The proposed solution is to use the (better predictable) low-frequency precursors of the dynamical model and pass these to the data-driven model, which then predicts the variable or impact of interest.

Causal discovery can also be used to improve the sampling of already run climate models to produce a single large ensemble for S2D predictions, known as 'pseudo-initialization' (Hegerl et al., 2021). Since low-frequency drivers are the only source of memory on S2D timescales, they can be used to 'pseudo-initialize' an already simulated freely evolving climate model (Hegerl et al., 2021). This is done by searching for analog climate states based on the low-frequency drivers of the climate model. These analog states are then combined to create a large ensemble. If successful, this approach can generate very large ensembles at low computational costs, as we can sample these analog states from already completed simulations.

7.3 Concluding remarks

Due to technological innovations and the increasing amount of (observational) climate data, Artificial Intelligence (AI) techniques such as machine learning and causal learning are currently revolutionizing the field of weather forecasting and S2S research. As explained in chapter 6, I hope to make my modest contribution to the field of subseasonal-to-seasonal weather forecasting via high-level open-source software and a spin-off company that is strongly rooted within the academic community.

In this chapter, I highlighted three lines for future research that all leverage the advantages of causal learning from observational data. It is my hypothesis that mid-latitude ocean-atmosphere feedbacks are not well resolved in current numerical models. Because the ocean is an important source of teleconnections - and likely due to additional biases - the teleconnection strength is often too weak in numerical models (Scaife and Smith, 2018; Di Capua et al., 2021; Merryfield et al., 2020; Vijverberg et al., 2020; Williams et al., 2022). Consequently, the importance of ocean-atmosphere interaction requires further research (section 7.2.1). Luckily, there are many opportunities to use AI-techniques together with expert-knowledge to alleviate some of these errors, and thereby improve regional climate projections (section 7.2.2) and weather/climate forecasts on seasonal-to-decadal timescales (section 7.2.3).

Appendix

Appendix chapter 3

3.A Spatial clustering of heat extremes

A binary timeseries of extreme temperature event occurrences is calculated for each geographical location, the binary time series is 1 if the temperature exceeds the q^{th} percentile, 0 otherwise. The resulting strings of zeros and ones are the input for the clustering algorithm. The binary strings that are very similar, i.e. those that experience heat extremes simultaneously, are clustered together. To be consistent with McKinnon et al. (2016), we use the Hierarchical Agglomerative clustering algorithm (Murtagh and Contreras, 2012), with the ‘jaccard’ distance metric (Jaccard, 1912) and the linkage criterion is set to ‘average’, meaning that the average distance between the binary strings is minimized to create clusters. We tested for robustness of the clusters in ERA-5 (figure 3.A.1) and EC-Earth (figure 3.A.2) by varying the number of clusters ($n_clusters = [2, 3, 4, 5, 6, 7, 8]$) and percentile thresholds ($q = [80, 85, 90, 95]$) used to create the binary strings. Since there are slight differences between the datasets, we also observe only small differences in the boundaries of the clustering. Because of these small differences, we decided to not use the exact same parameters as used by McKinnon et al. (2016). In the original work, the threshold was fixed at the 95th percentile, and they choose $n_clusters = 5$. For ERA-5, the exact same settings render a similar clustering result. For EC-Earth we choose the clustering output ($n_clusters = 5, q = 90$) such that the eastern U.S. cluster is most similar to the original eastern U.S. cluster found by (McKinnon et al., 2016). The final clusters are shown in figure 5.2.

3.B Double cross-validation

To fit and validate a statistical model, we need a sufficient amount of independent datapoints. Particularly for dynamics on S2S timescales, this is challenging with only 40 years of data for ERA-5. As mentioned in section 23.2.1, we de-trend all data to avoid that we are fitting a spurious signal to a long-term trend. Using the response guided approach, we make choices drawn from data, which increases the danger of overfitting (Michaelsen, 1987). We can minimize this pitfall with (1) a strict train-test split throughout the whole analysis, (2) doing robustness tests such as testing different train-validation-test combinations (e.g. see figure 3.F.3). As depicted in figure 3.B.1, we use a stratified 10-fold cross validation to split training and test data. This means that the test years are not completely random, since the test set is forced to be a representative sample in terms of the amount of events. This helps to avoid train/test combinations that are by chance dominated by a certain phase of multi-annual or decadal variability and it allows us to

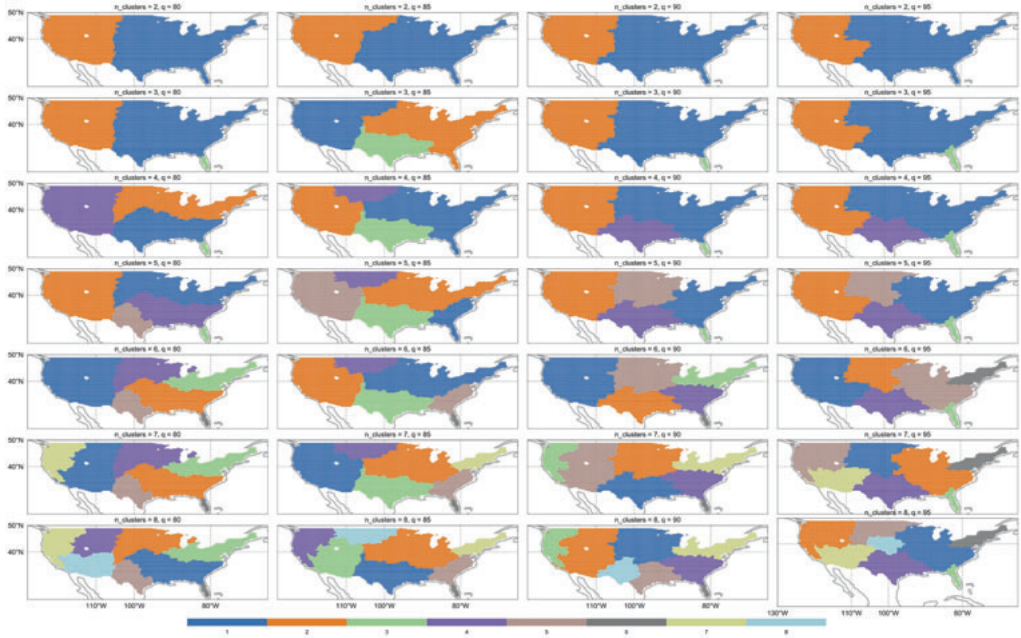


Figure 3.A.1: Parameter sweep spatial clustering results for ERA5 (.25°x.25°)

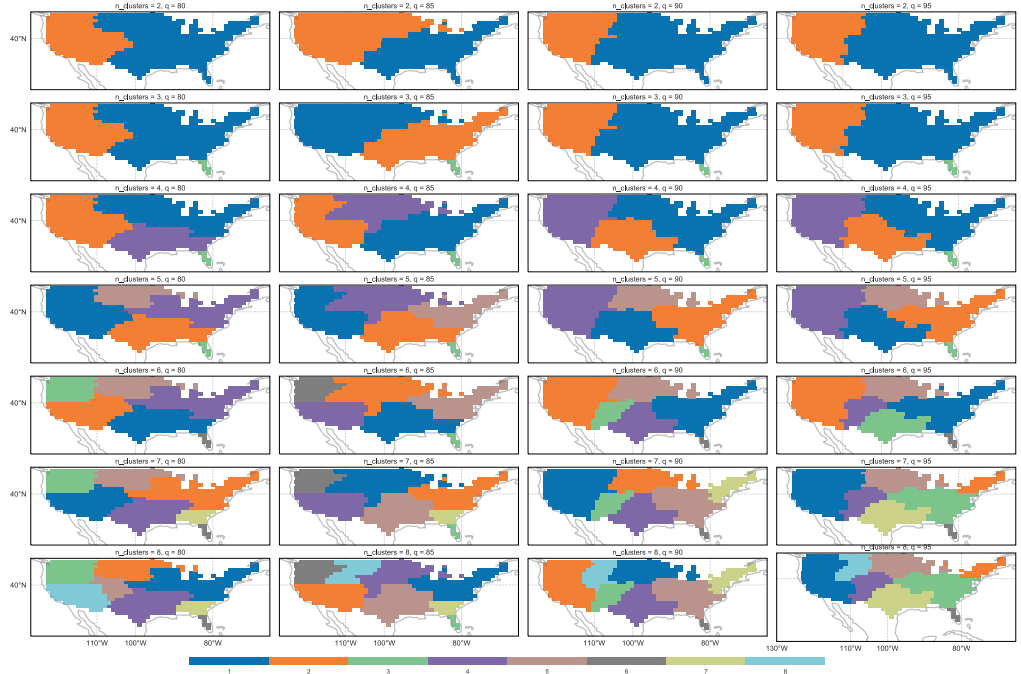


Figure 3.A.2: Parameter sweep spatial clustering results for EC-Earth (1.125°x1.125°)

validate with different train/test sets, which is not possible with e.g. the leave-one-year-out method. Because we cannot reliably estimate the skill based on only 4 years of test data, we repeat the CPPA algorithm and the subsequent model fitting 10 times. We then concatenate all forecasted test years and calculate our skill metrics based on all the years in the dataset (40 years for ERA-5). Thus we do not train a single statistical model, but 10 slightly different ones.

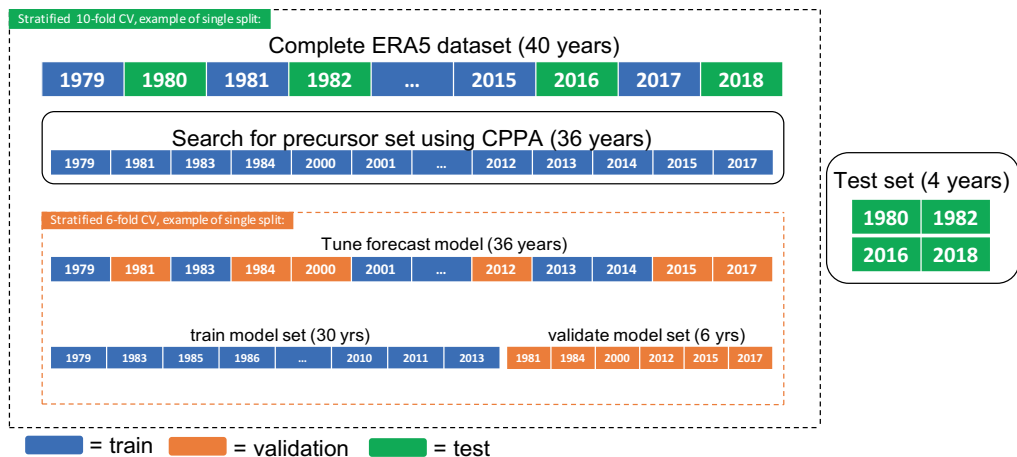


Figure 3.B.1: A complete overview of the 'double' stratified cross-validation procedure to enable a response-guided search for precursors and model tuning with a limited amount of data. This results in a forecast model for each 10 train-test splits and for each lag.

3.C CPPA vs. linear point correlation map approach

For the extracted precursors as shown in figure 3.4, we only show the mean over the training sets. However, as depicted in figure 3.B.1, we extract the precursor regions once for each training set (and for each lag), see figure 3.C.1. By looking at how robust the precursor region extraction was when using slightly different subsets of data, we can plot the robustness of the precursor regions (figure 3.C.2).

We also compare CPPA to the conventional point-wise correlation map approach. CPPA only looks at relatively extreme events (hot days) to learn the precursor regions. If the signal of the precursor only arises in the tail of the conditional temperature distribution, CPPA might enable detection of precursors showing a non-linear relationship with eastern U.S. temperature. When comparing the output of CPPA versus the correlation map approach shown in figure 3.C.3, we observe a qualitatively similar pattern. This shows that either (1) the correlation map approach was still able to detect a signal when the underlying signal was in reality non-linear, or (2) the SST relationship with temperature is by good approximation linear. We also note the correlation map shows a higher robustness compared to CPPA, which only learns from events versus non-events. The higher robustness is also the reason to use the correlation map approach to extract soil moisture timeseries.

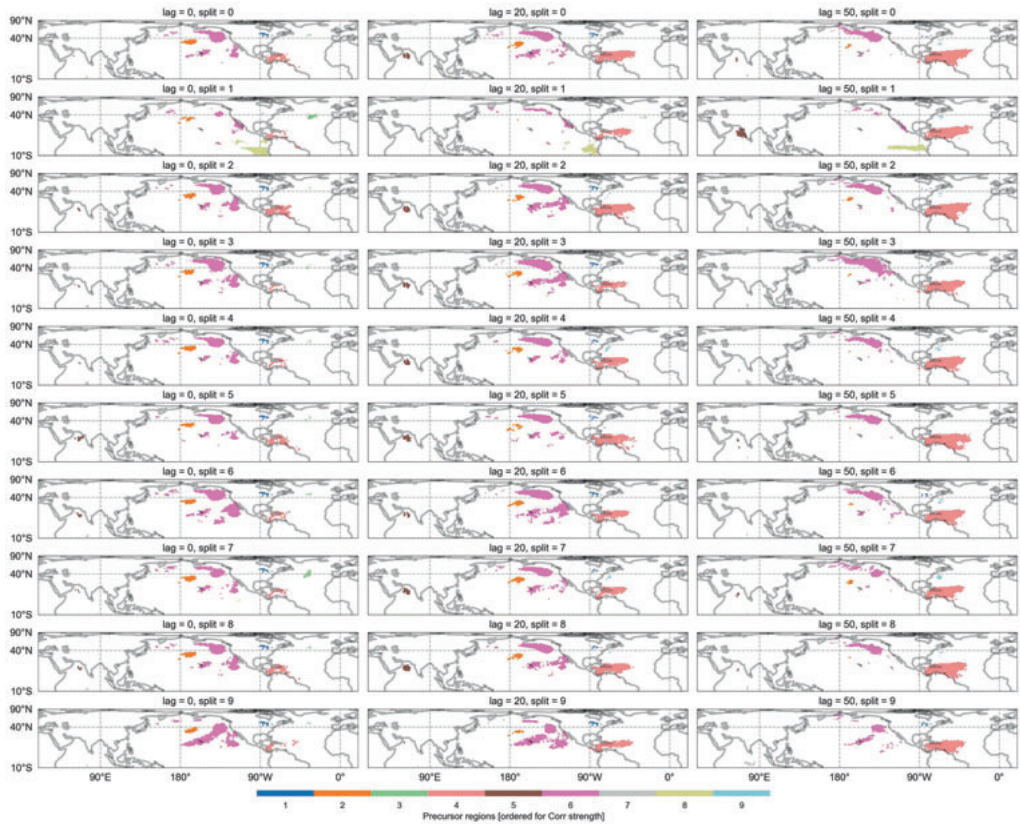


Figure 3.C.1: Sea surface temperature regions found by the CPPA algorithm using a single training set (36 years). Clusters should be at least 5 by 5 degrees big (defined at 45°N) to form a core sample, if they show a high density, they are more likely to include neighboring gridcells into the cluster. The radius at which core samples (initial clusters) search for neighboring gridcells is set by the *eps* parameter, in our case 500 [km]. We take into account the gridcell area by assigning weights to the samples, i.e. gridcells. Timeseries are calculated by taking daily means, weighted by gridcell area and the N-FSP, see section 3.2.3 and figure 3.1.

Although with CPPA, we were able to stay close to the analysis as done by McKinnon et al. (2016).

Because CPPA objectively searches for precursor regions based on training data that slightly differs for each train-test split, some precursor regions are not always extracted. Table 3.C.1 shows all the precursor regions (timeseries) that were extracted and the count denotes how many times it is present in the 10 training sets. The format of the labels is $\{lag\}..\{region\}..\{variable\}$. The labels correspond to the labels shown in figure 3.C.1. Note that the lag refers to the lag at which the precursors were retrieved. Thus, we did not change the precursors as function of lag as done by McKinnon et al. (2016), since we found that using the timeseries of lag=0, produced the best forecast skill. We expect this is due to the fact that the signal-to-noise ratio is largest at lag=0. The timeseries are subsequently shifted to match the lead-time on the x-axis of the verification figures.

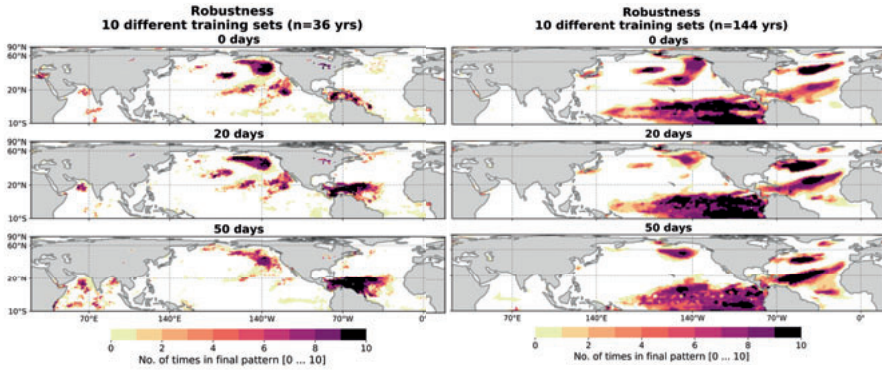


Figure 3.C.2: Robustness of gridcells for ERA-5 (left column) and EC-Earth (right column), see method section for details. Values equal to 10 means that the gridcell is extracted in all 10 different training sets. Gridcells which are consistently part of the Precursor Pattern are interpret as more robust.

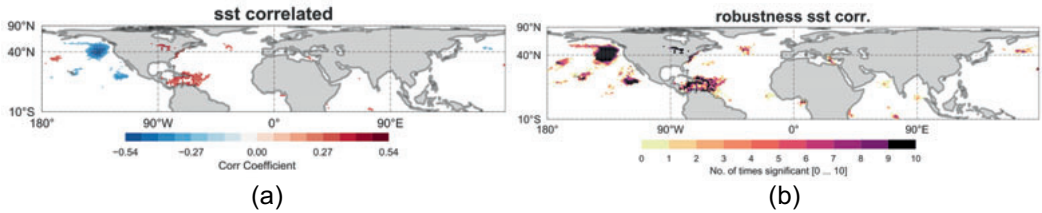


Figure 3.C.3: (a) SST correlation maps ($\alpha = 0.01$) for 15 day mean timeseries at lag = 0, (b) the robustness across different training sets. (a) the mean over training sets, gridcells are masked if they were not in 50% of the training sets.

3.D Soil moisture timeseries

For the final forecast we additionally add information from soil moisture layer 2 [7 - 28 cm] and layer 3 [28 - 72 cm]. We choose these two deeper layers because we expect that there is more memory in the deeper layers since there is less mixing with the atmosphere. We include soil-moisture using an existing framework as introduced by Kretschmer et al. (2017a) that is similar to CPPA. The soil moisture timeseries are retrieved by (1) calculating which grid-cells are significantly correlating with the T_{90_m} timeseries at lag=0, (2) subsequently clustering regions of same sign together in the same fashion as done for CPPA, and (3) calculating the area-weighted spatial mean timeseries for each cluster, results for this analysis are shown in figure 3.D.1.

3.E Climate indices

For our daily ENSO timeseries, we use the Nino-3.4 spatial region [5°S - 5°N and 170°-120°W] to calculate the area-weighted mean of the de-trended SSTA daily data (Deser and Trenberth, 2016). For the calculation of the PDO timeseries, we first aggregate the

Table 3.C.1: List of all SST precursor timeseries extracted by CPPA. The whole or a subset of the precursors are used for figures 3.5 to 3.11. Based on the ERA-5 dataset.

ERA-5 Precursor labels	count
CPPAsv	10
0..1..sst	10
0..2..sst	10
0..3..sst	10
0..4..sst	10
0..5..sst	5
0..6..sst	10
0..7..sst	7
0..8..sst	2

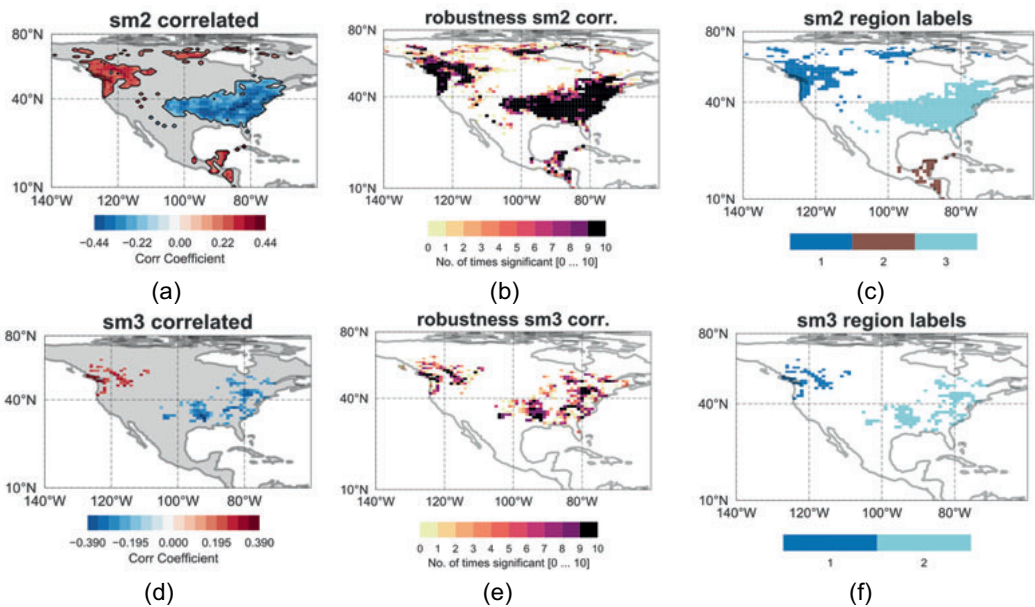


Figure 3.D.1: Same as figure 3.C.3, but for soil moisture.

de-trended SSTA daily data to monthly means. Based on the monthly mean area-weighted SSTA training data, we construct the first EOF (or loading pattern) of the North Pacific [20°-70°N and 115°E - 110°W] (Deser and Trenberth, 2016). Finally, the loading pattern is projected on the (daily) test data to obtain the daily principal component timeseries.

We calculate the PDO with the training data for each test set (as illustrated by figure 3.B.1) to obtain an out-of-sample timeseries of the PDO. See figure 3.E.1 for the (mean over training sets) PDO pattern and a composite mean of the El Nino phase.

Table 3.D.1: Same as Table 3.C.1, but for soil moisture precursors based on the ERA-5 dataset.

ERA-5 Precursor labels / names	count
0..1..sm2	10
0..2..sm2	10
0..3..sm2	10
0..1..sm3	10
0..2..sm3	10

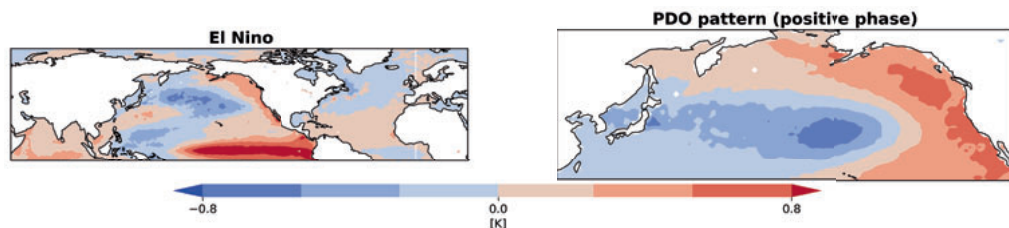


Figure 3.E.1: Left: El Nino phase of ENSO, found by taking a composite mean where the 5 months smoothed Nino3.4 timeseries exceeds 0.4 degrees. Right: PDO pattern (mean over training sets). Retrieved by calculating the first EOF (or loading pattern) for Pacific area-weighted SST between 20°N - 65°N and 115°E - 110°W. Timeseries are used for the computation of the cross-correlation matrix and for the forecasts (PDO+ENSO+sm).

3.F Supporting information forecasts

When we aggregate to 15-day means, without overlap in the windows, the lead-time can be defined in multiple ways. In order to make our forecast similar to an operational implementation, the lead time is defined such that we are predicting the centered date of a time-window, using only information from the past. Figure 3.F.1 shows a schematic illustration where we predict the centered date 2012-08-26. To select the precursor dates, we shift lag=25 and the additional 15 days back in time. Hence, the prediction is made on 2012-08-01, 25 day in advance, using information of 2012-08-01 and of the previous 14 days. Note, the exact summer dates that we originally forecast on daily timescale inevitably change from 06-24 to 08-22 to the centered dates 06-27 to 08-26, exactly 5 bins of 15 days.

In figure 3.F.2, we use a box-plot to convey the consistency between models that were learned on different training datasets. The corresponding precursor regions can be found in Appendix C and D. The spread in the logistic regression coefficients is generally small, indicating that overall, the models were similar. This supports that what the model learned was not a lucky fit that resulted in good skill scores on the test dataset, but rather, it re-learned the same associations when applying perturbations to the training data. We will not go into discussing the physical meaning of the coefficients, since a model that provides high forecast skill, does not necessarily inform about the underlying causal structure (Li et al., 2020; Runge et al., 2019).

Figure 3.F.3 show a robustness check for the forecast skill, where we tested the influence of using 3 different combinations of train-validation sets for the 'Tune forecast model' step in figure 3.B.1. Section 33.3.3 and 33.3.4 showed that, using CPPA or PEP as precursor(s), hot day events do not show predictive skill at long leads. Figure 3.F.4 shows the forecast

skill when keeping the same base-rate and aggregating over time, i.e. the hot 15-day mean events. ERA-5 does not show an increase in forecast skill compared to forecasting hot day events. EC-Earth, with many more datapoints, shows a small increase in skill compared to the daily events, but still not significant.

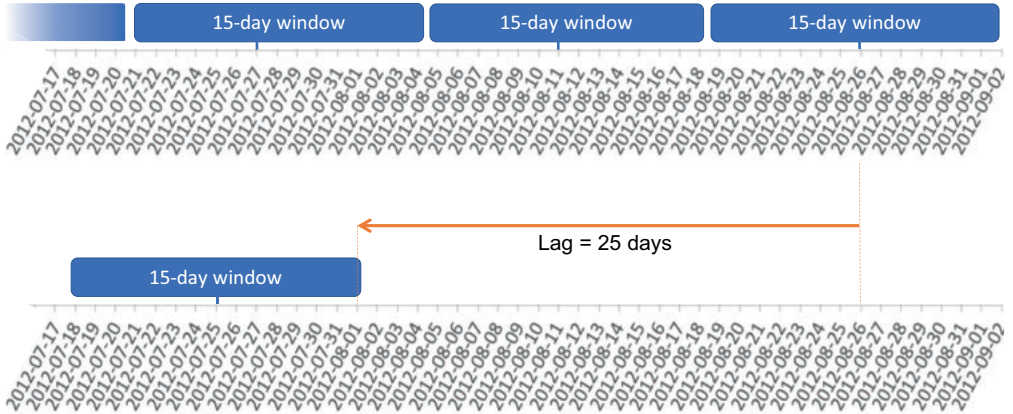


Figure 3.F.1: Schematic illustration of the temporal aggregation and how the lead times are defined. The upper dates represent the timeseries belonging to the target timeseries, while the second row of dates represent the timeseries of the precursors.

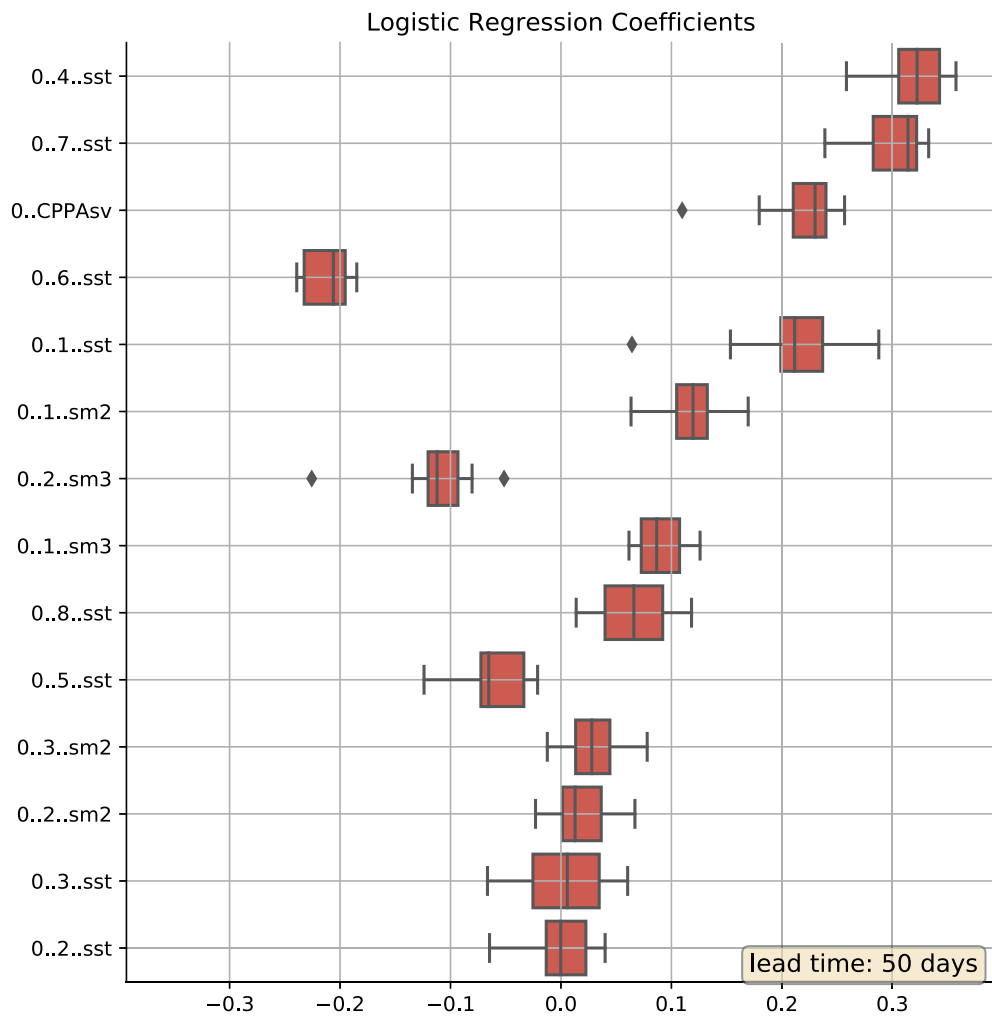


Figure 3.F.2: Box-plot of the logistic regression coefficients that were fitted using 10 different training sets with a lead-time of 50 days.

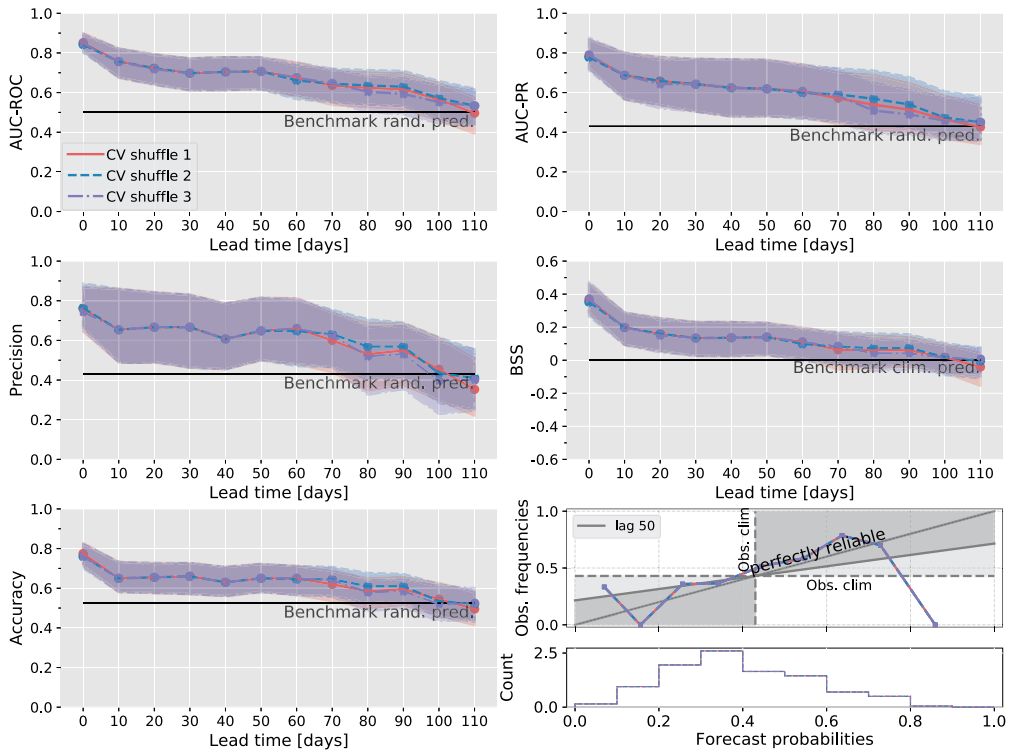


Figure 3.F.3: Forecast skill robustness test in which we used 3 different combinations of train-validation sets for the 'Tune forecast model' step (depicted in figure 3.B.1).

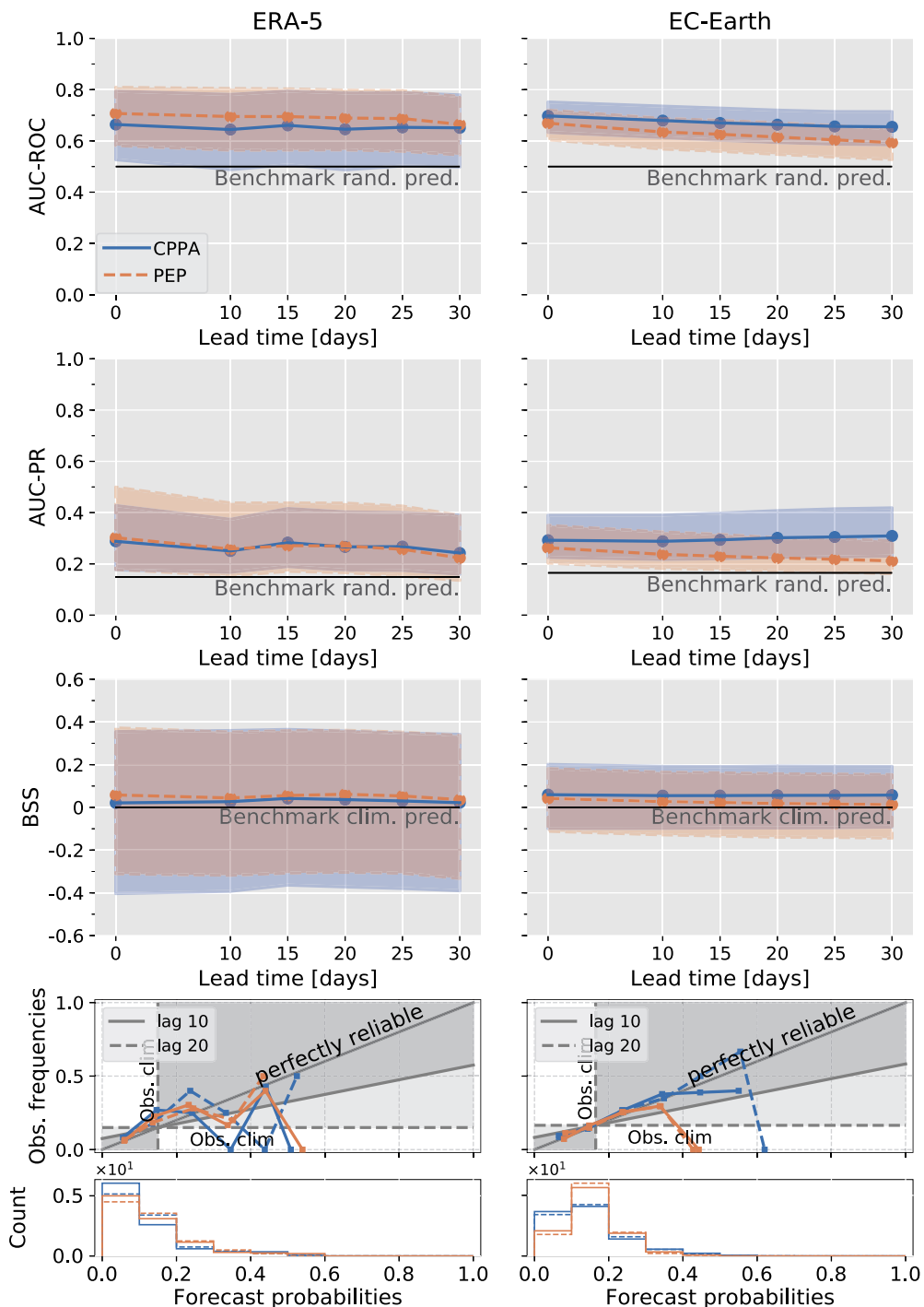


Figure 3.F.4: Forecast validation for 'hot 15 day mean events' using ERA-5 (40 years of data) and EC-Earth (160 years of data).

Appendix chapter 4

4.A clustering simultaneous warm temperature periods

The temperature clustering analysis shows that the western and eastern US are almost always well separated around the Rocky Mountains. The clustering results were evaluated by testing a range of temporal aggregations ($\text{tfreq} = [5,10,15,30]$) and number of clusters ($\text{n_clusters} = [4,5,6,7,8,9,10]$). We choose the clustering result based on 15-day mean data, as these clusters are similar to what has been found before (McKinnon et al., 2016; Vijverberg et al., 2020). We verified that the size of the clusters is appropriate by calculating one-point-correlation maps and checking if the length-scale of the spatial correlation falls within the clusters (Figure 4.A.2). The area-weighted spatial mean of cluster 1 is referred to as the western US temperature (T^W) and similarly, cluster 4 represents the eastern US temperature (T^E).

Note, that we focus on these two clusters since they exist approximately on the same latitude but show a clearly different Rossby wave pattern (Figure 4.A.1). Cluster 7, approx. in the north-western North American domain, shows the same RW as we find for the eastern US temperature, but of opposite polarity (Figure 4.A.3). This suggests that, similar to the mechanism leading to predictability in the eastern US, there is also predictability for that cluster.

4.B Defining the Rossy wave timeseries

Capturing (small-scale) Rossby waves into a 1-d timeseries can be difficult since other waves might project onto the (few) low/high pressure systems. A few slightly different approaches were tested, here we show our (best) approach used in the paper. We verified we can reconstruct the RW pattern accurately when correlating z500 with the target Rossby wave timeseries (RW_t^W and RW_t^E). Indicating that the timeseries are effective in capturing the target RW variability Figure 4.B.1.

4.C Comparison to atmospheric modes of variability

We compared the western and eastern RW to known summer modes of variability, of which we discuss two relevant modes in detail. All data is aggregated to 15-day means. We show the Northern hemispheric circumglobal [20-90°N] summer (JJA) mode of variability in Figure 4.C.2, defined as the first Empirical Orthogonal Function (EOF) of meridional wind (v) at 300 hPa (referred to as v300-EOF1). This pattern is similar (but of opposite

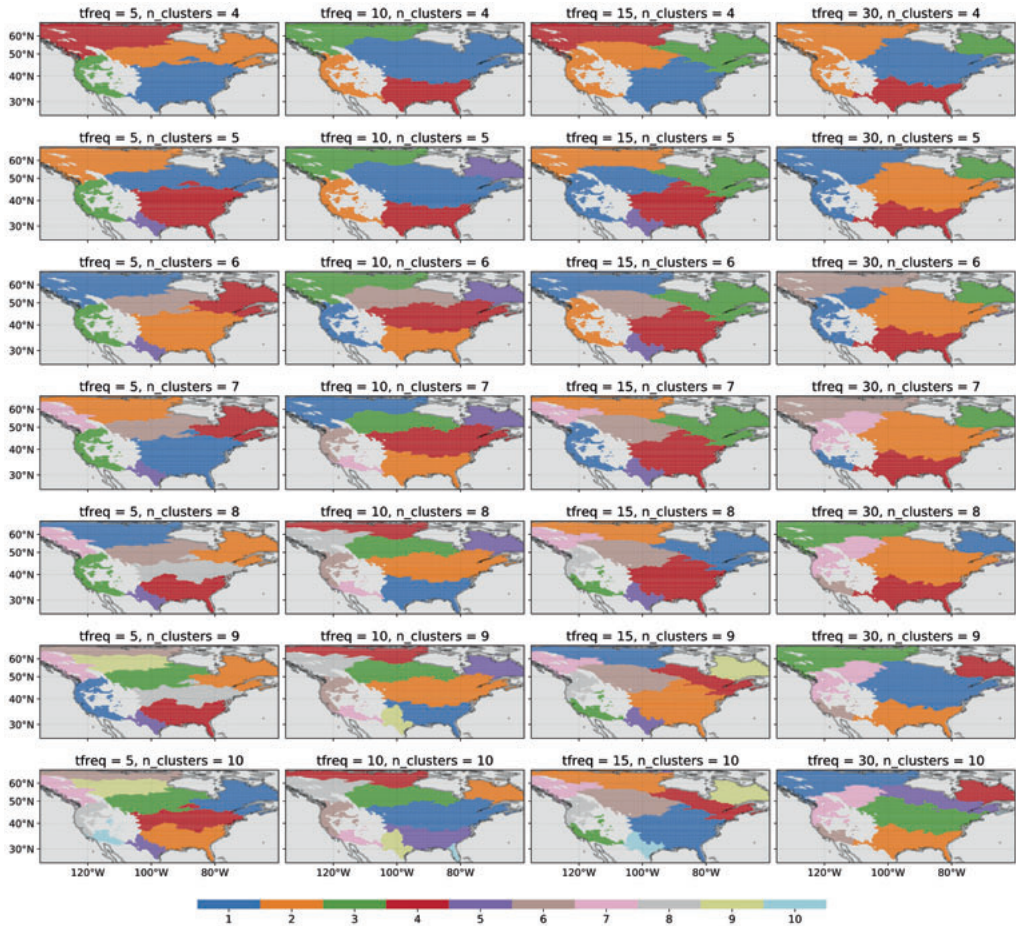


Figure 4.A.1: Hierarchical Aggl. clustering of binary temperature timeseries that are one when exceeding the temporal 66th percentile, and zero otherwise. Clustering performed on a range of temporal aggregations ($tfreq=[5,10,15,30]$) and number of clusters ($n_clusters = [4,5,6,7,8,9,10]$).

polarity) to Figure 4.3b from Branstator and Teng (2017), and strongly resembles the circumglobal wavenumber 6 pattern as identified by Kornhuber et al. (2017b). Note that the meridional wind associated with the western RW (4.C.2) fluctuates around the Rocky Mountains, suggesting that the Rocky Mountains play in role in existence of this atmospheric mode of variability (Kornhuber et al., 2017b; Hoskins and Karoly, 1981). The second relevant mode of variability is the summer PNA, based on the definition given by the NOAA Climate Prediction Center (referred to as PNA_{cpc}). It is based on a Rotated Principal Component Analysis applied to the 500 hPa geopotential height field between 20°N-90°N. The PNA_{cpc} was downloaded from the KNMI Climate Explorer (van Oldenborgh, 2020).

The overlap between the western RW and the v300-EOF1 is clearly visual (Figure 4.C.2). Similarly, the v300-EOF1 Principal Component (PC) timeseries correlates well with the RW_t^W (Figure 4.C.1). The RW_t^E does not appear to correlate well with circulation patterns

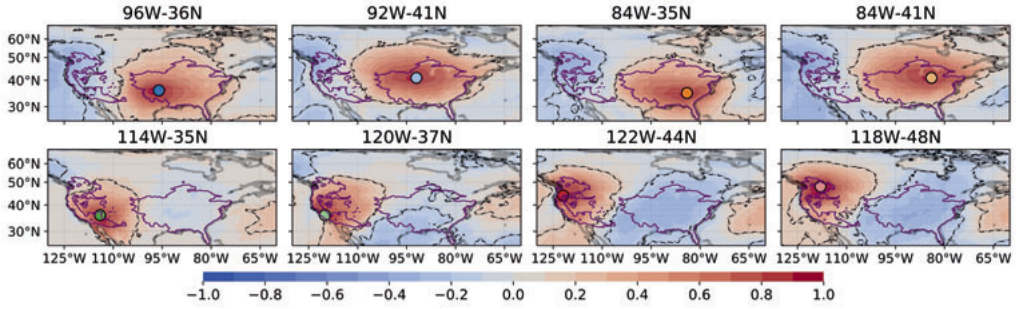


Figure 4.A.2: One-point-correlation maps of daily maximum mx2t using different lat-lon coordinates (depicted by the scatter points). Black dash-dot-dash contour lines shows significance ($\alpha_{\text{FDR}}=0.05$). Solid purple contour lines show the western and eastern US clusters (see Figure 4.A.1, tfreq=15, n_clusters=7)

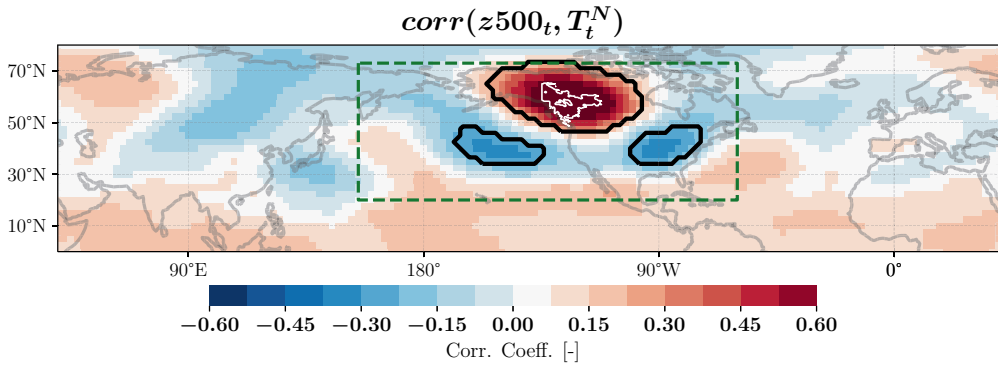


Figure 4.A.3: Same as Figure 4.2a and b of the main manuscript, but correlating $z500$ versus the spatial mean temperature in cluster 7 located in the north-west North American domain.

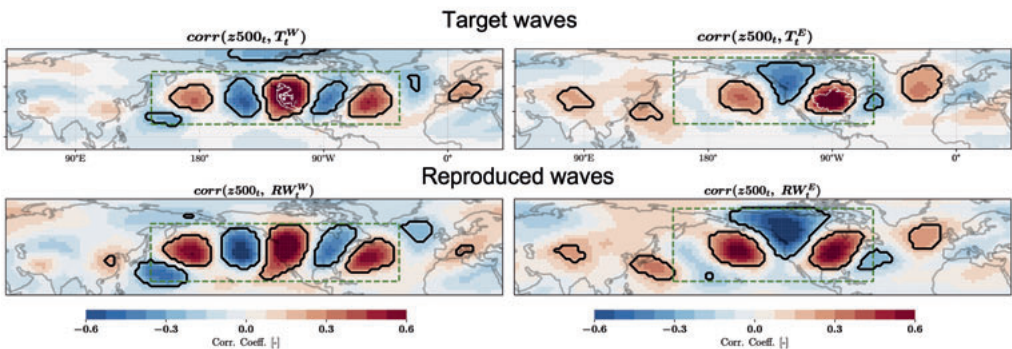


Figure 4.B.1: Correlation maps to verify that the RW timeseries we calculated, as explained in method section 4.2.3, is effective in capturing the RW variability of interest. The upper plots are copied from main manuscript Figure 4.2a and 4.2b for comparison. Black contour lines indicate significance in 60-out-of-70 training datasets ($\alpha_{\text{FDR}}=0.05$).

that explain a lot of variability, as indicated by the correlation with PNA_{cpc} and v300-EOF1 (Figure 4.C.1). Also, the eastern RW pattern also does not resemble the loading pattern of local z500. For the latter, we focus on variability patterns within the RW_t^E bounding box (by calculating the first EOF of z500, Figure 4.C.3. The latter analysis differs from the PNA_{cpc} definition, which attempts to explain high variability modes in z500 between 20°N-90°N using a Rotated-EOF. Upon visual inspection of the Pacific/North American z500 EOF loading patterns, there is no clear resemblance to the eastern RW pattern (Figure 4.2b of main manuscript). We have looked at other summer circumglobal waves (Kornhuber et al., 2017b) and other summer modes of variability that project onto the Pacific/North American region (Branstator and Teng, 2017; Ding et al., 2011). Based on visual inspection we did not find a summer mode of variability that resembles the eastern RW pattern. A circumglobal wave of wavenumber 8 (Figure 9d in Kornhuber et al. (2017b)) does match the meridional wind pattern over North America, however, the wave is already out-of-phase over the north Pacific. The correlating regions over Asia and the Atlantic might be a statistical artifact of the partial overlap with wavenumber 8. As is further supported by Figure 4.D.1, the centers of actions of the eastern RW project strongly an arcing RW that is very similar to the winter PNA pattern in its negative phase and the ENSO-forced atmospheric bridge that is forced by the La Nina state (Zhou, 2019).

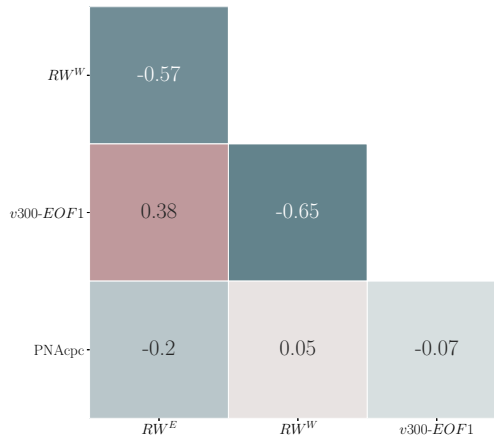


Figure 4.C.1: Correlation matrix showing the correlations between the summer east and western Rossby waves and two circulation modes of variability. v300-EOF1 is the northern hemispheric v-300 hPa mode of variability, defined by the first EOF [10°N-80°N]. PNA_{cpc} is the PNA timeseries as defined by the Climate Prediction Center [downloaded from van Oldenborgh (2020)]. Data is aggregated to 15-day means.

4.D SST-RW coupling in winter and spring

To obtain some additional insights into the SST-RW coupling in winter and spring, we repeated the experiment done in Figure 4.3 of the main manuscript for winter and spring. The Rossby wave timeseries still uses the eastern RW pattern as found in summer (Figure 4.2b of main manuscript). This way we ensure that the RW_t^E still captures the same

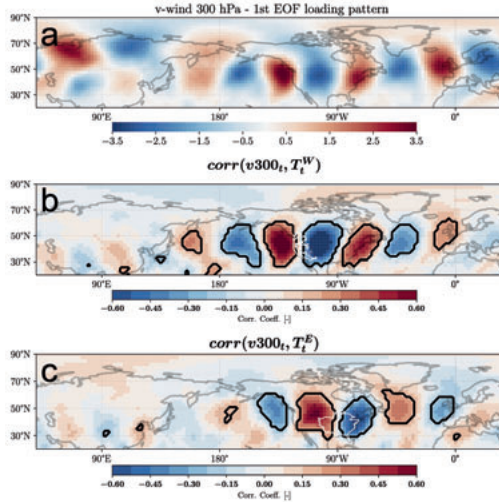


Figure 4.C.2: Comparison of RW patterns captured by meridional (v) wind at 300 hPa. Panel a: EOF-loading pattern of 15-day mean v-wind at 300 hPa (v300), 10-80N. Panel b: correlation map between v300 and western US temperature (black contour lines indicate significance at ($\alpha_{FDR}=0.05$). White contour line indicates the western US cluster. Panel c: same as panel b but for the eastern US temperature.

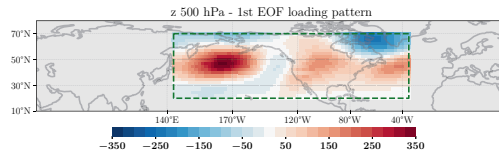


Figure 4.C.3: EOF-loading pattern of 15-day mean geopotential height (z) at 500 hPa within the green bounding box [155-300E, 20-73N], same as eastern US RW green bounding box shown Figure 4.2 of the main manuscript.

RW and not a different mode of variability that drives temperature in winter or spring. Figure 4.D.1 shows how the RW_t^E timeseries projects on winter and spring geopotential height in DJF and MAM. The correlating pattern shows the original eastern RW pattern and extends to the tropics, more clearly revealing the negative PNA pattern and the la Nina-related atmospheric bridge pattern (Figure 4.2b and d of Lopez and Kirtman (2019)). The latter two patterns are largely similar, the main difference is the mechanism; the PNA pattern can be intrinsically generated, while the atmospheric bridge is forced by ENSO variability.

In the CENs (Figure 4.D.2) we observe a stronger coupling in winter and spring compared to summer (in-line with existing literature, e.g., (Zhou, 2019)). During the winter seasons, we find a much stronger downward forcing Figure 4.D.2b-e. For spring, we also observe both a strong downward forcing and a two-way coupling on the 5-day mean timescale (Figure 4.D.2i).

Note that the CENs should be interpreted within the context of our research question. The summer eastern RW pattern was identified at the start of the analysis based upon

the fact that it drives (correlates with) eastern US summer temperature variability. We do not prove that this summer RW is most effective at capturing (i.e., maximizes) the coupling strength between the north-Pacific and the atmosphere. Instead, we are interested in the SST-RW interaction that is most important for eastern US summer temperature variability. The CEN analysis quantifies how the 'winter and spring geopotential height variability that projects onto our summer RW^E pattern' interacts with the SST and we show that the summer RW pattern projects strongly onto z500 variability in winter and spring (Figure 4.D.1). Figure 4.D.2a and Figure 4.D.2g confirms that the RW^E variability in winter and spring projects onto the mid- and eastern north-Pacific SST regions which are - in late spring - associated with the forcing of the RW^E in summer. Since these mid- and eastern north Pacific SST regions match the main features of the PDO pattern, Figure 4.D.2 shows that the RW^E strengthens the PDO pattern in winter and spring, although the relative importance for the development of PDO variability is unclear. However, the strong coupling between the PDO and our summer RW pattern is confirmed by the strong co-variability between the annual mean RW^E timeseries and the annual mean PDO timeseries (correlation of 0.71), as shown in Figure 4.D.3.

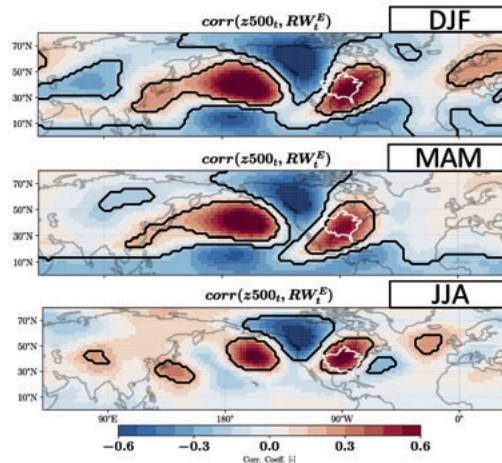


Figure 4.D.1: Correlation maps between geopotential height at 500 hPa ($z500$) and the eastern Rossby Wave timeseries for DJF, MAM and JJA periods using 15-day mean data. RW_t^E is calculated using the same Rossby wave pattern that is identified in Figure 4.3b, but projecting that pattern onto the geopotential height field in DJF and MAM. More details can be found in the method section 4.2.3.

4.E Seasonal dependence of temperature predictability

We show here that the western US July-August mean temperature is not predictable from SST at lag 1 Figure 4.E.2. As shown in Figure 4.5 of the main manuscript, we first calculate correlation maps at lag 1 for different target months of the bi-monthly mean western and eastern US temperature. To extract the precursor regions, we cluster the significantly correlating regions of the same sign (Figure 4.E.1). The clusters are used to

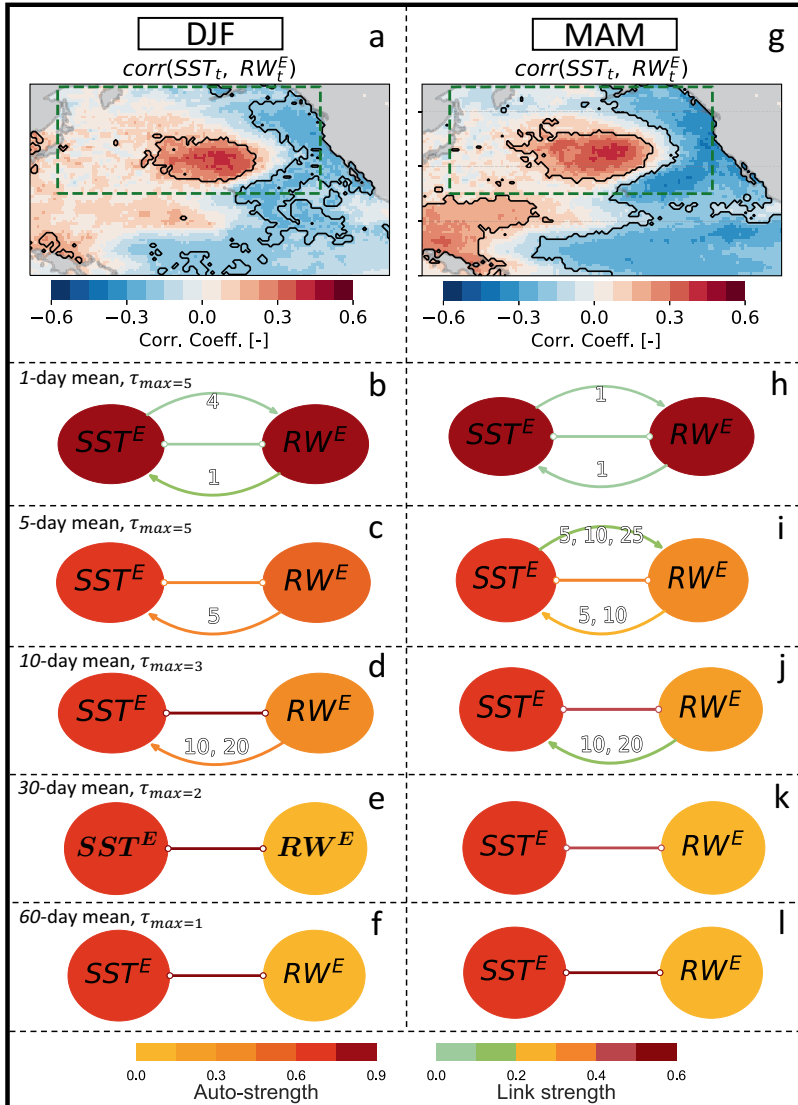


Figure 4.D.2: SST-RW coupling of the eastern RW in winter (DJF) and spring (MAM), see caption of Figure 4.3 of main manuscript for more information, all settings are identical.

calculate an area-weighted, correlation-value weighted timeseries. We explore the seasonal dependence of temperature forecasts skill by targeting different months to predict using a Ridge regression (Figure 4.E.2). The Ridge regression only uses the (standardized) lag 1 SST timeseries, the regularization coefficient ranges between $1E^{-3}$ up to 2 with 26 logarithmically spaced steps.

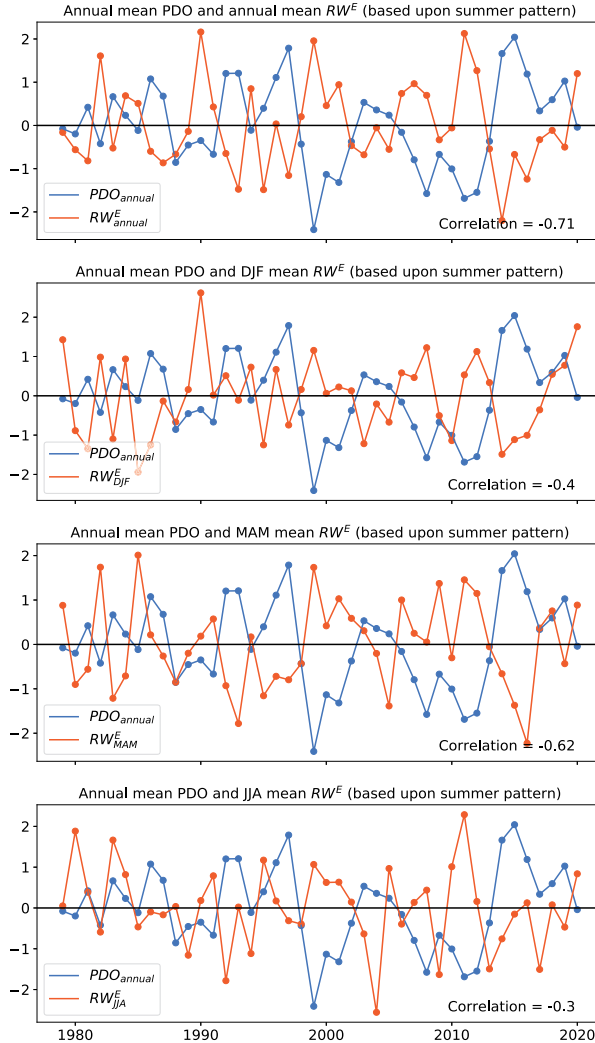


Figure 4.D.3: Timeseries plot of the annual mean PDO timeseries and the - from top to bottom - annual, DJF, MAM, JJA mean RW^E timeseries, respectively. The RW^E pattern used to calculate the spatial covariance timeseries is shown in Figure 4.2b of main manuscript.

4.F Window of Predictability emerging from PDO state

Here we test the robustness when splitting the forecasted July-August means into years with a strong vs. weak winter/spring PDO state. When making the seasonal forecast for the July-August mean temperature, we only have 42 forecast/observation pairs for verification in total and thus 21 datapoints per composite. Table 4.F.1 shows that results are robust when repeating the experiment 5 times using different training samples. Furthermore, in line with our hypothesis (i.e., that years with an anomalous winter/spring PDO state are more predictable), when we select the 30% instead of the 50% most anomalous PDO

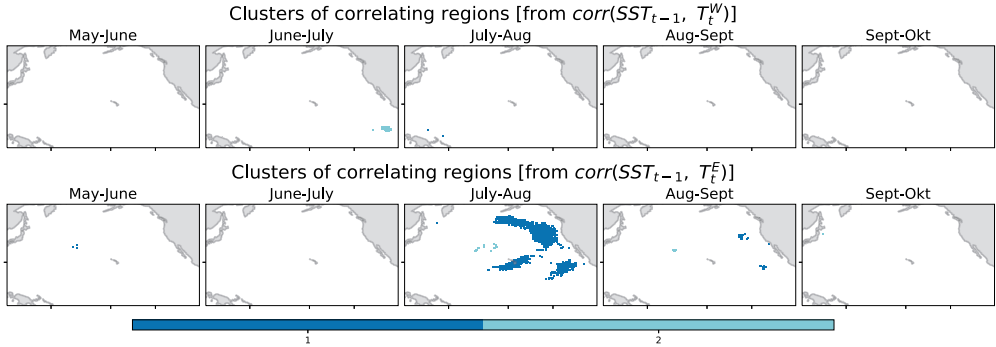


Figure 4.E.1: Precursor regions based on the correlation maps between SST at lag 1 and temperature (Figure 4.6 of main manuscript), (top row) showing clusters for western US, (bottom row) the eastern US temperature. DBSCAN is used for the clustering of the correlating regions. Only gridcell labels that are extract 9-out-of-10 training sets are shown.

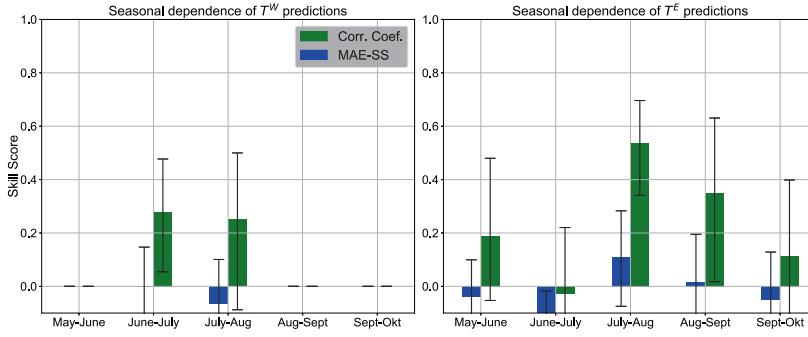


Figure 4.E.2: Forecast verification of western (left column) and eastern (right column) US bimonthly mean temperature. Forecast skill as function of target months using Pacific SST regions at lag 1 retrieved from out-of-sample correlation maps, see Figure 4.6 of main manuscript. Confidence intervals (significance level of 0.05) are calculated by bootstrapping ($n=2000$). Metrics used are the Correlation Coefficient and the Mean Absolute Error-Skill Score, where the benchmark forecast is the climatological mean of temperature ($MAE-SS = 1 - \frac{MAE_{forecast}}{MAE_{climatology}}$).

states, we observe a further increase in skill.

Note, the larger percentual increase that is observed for the MAE-SS compared to the decrease in the MAE in Figure 4.6 of the main manuscript means that the benchmark (predicting approx. an anomaly of 0) scored worse for the years with a strong winter/spring PDO. The benchmark scores worse for years with higher temperature, indicating that during strong winter/spring PDO years the temperature was also more anomalous.

Metric	Seed	strong 30%	weak 30%	strong 50%	weak 50%
Corr. Coef.	1	0,89	0,47	0,83	0,53
	2	0,85	0,51	0,79	0,60
	3	0,88	0,51	0,81	0,58
	4	0,88	0,57	0,82	0,62
	5	0,85	0,57	0,79	0,63
MAE-SS	1	0,57	0,12	0,40	0,08
	2	0,50	0,17	0,35	0,17
	3	0,53	0,12	0,35	0,10
	4	0,50	0,20	0,36	0,16
	5	0,48	0,20	0,32	0,17
mean absolute error	1	0,41	0,99	0,40	0,87
	2	0,48	0,94	0,44	0,80
	3	0,45	0,99	0,43	0,85
	4	0,48	0,90	0,42	0,80
	5	0,49	0,91	0,45	0,78

Table 4.F.1: July-august mean eastern US temperature forecasts divided based on strong vs weak winter/spring PDO states. Mean value of 2000 times bootstrapped composites. Strong (weak) 30% contain the 30% most anomalous (weak) winter/spring PDO states. Forecasts are based on using the mid and eastern Pacific precursor regions, using both the May-June (i.e., lag 1) and March-April (i.e., lag 2) mean timeseries.

Appendix chapter 5

5.A Pre-processing of crop yield data and cross-validation

After the clustering (as described in the method section) has been performed, we aggregate all data that is located within the southern cluster (label 1). We assume (as supported by literature), that the crop growth in each gridcell is negatively influenced by hot-dry conditions. However, there are clear differences within the cluster in terms of absolute productivity, interannual variability, long-term trend, and time period that the observations cover (Figure 5.A.1). Therefore, we first detrend (Figure 5.A.1, panel a) and standardize the timeseries (Figure 5.A.1, panel b), before calculating the spatial mean. Figure 5.A.1 panel b also shows that prior to the ~ 1975 the variability is relatively small (most gridcells vary between -1 and $+1\sigma$), while during the 1980-2019 period most gridcells vary between -3 and $+2\sigma$. Figure 5.A.3 shows the shift of producing regions towards lower latitudes around the 70s. The out-of-sample pre-processing is visualized in Figure 5.A.4.

The outer cross-validations are introduced in 25.2.3. For the tuning of hyperparameters, we apply another 'inner' cross-validation scheme (Vijverberg et al., 2020), meaning that each 'outer' training dataset is split into (inner) training and validation sets using a 10-fold CV approach. A schematic of the double cross-validation approach is shown in Figure 5.A.2.

5.B Response-guided dimensionality reduction and causal precursor selection

As discussed in section 5.3.3, Figure 5.B.1 shows that the soil moisture correlation patterns are associated with dominant circulation patterns that are known to strengthen the horseshoe Pacific SST state. More anomalous states (positive or negative) of the horseshoe Pacific are associated with a stronger boundary forcing for the atmosphere, and therefore a higher signal-to-noise ratio Vijverberg and Coumou, 2022.

Figure 5.B.3 shows the SST precursor regions that were filtered out by the precursor selection step after being identified by the RGDR method for the LTO cross-validation. Note that the spuriously correlating regions are often located close to the coast. Intuitively, SST variability close to the coast is much more affected by local small-scale dynamics. Due to the lack of large-scale spatial correlation, there are more independent realizations of SST variability which increase the probability that a timeseries strongly correlates by chance. Due to the high (spurious) significance, the Benjamin-Hochberg false discovery rate correction that is applied for the correlation maps is not sufficient to filter these out.

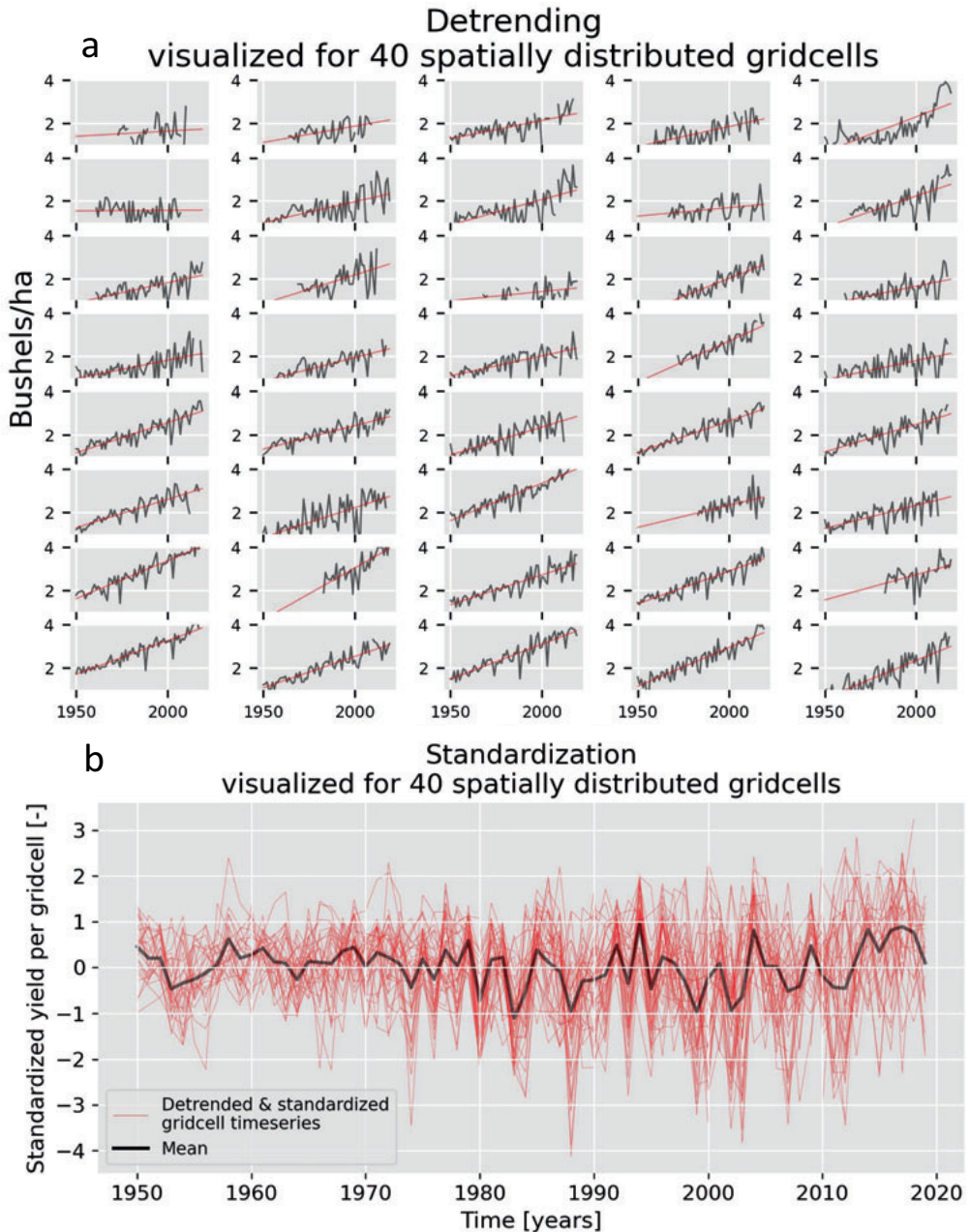


Figure 5.A.1: Panel a shows 40 gridcells timeseries of crop yield [bushels/hectare] (black line) and the fitted linear trend line (red line). The 40 gridcells are roughly evenly distributed over the mid-to-southern cluster. Panel b shows the 40 gridcell timeseries of crop yield (red lines) after detrending and standardizing. Black line shows the mean over all timeseries within the mid-to-southern cluster that contain more than 30 datapoints.

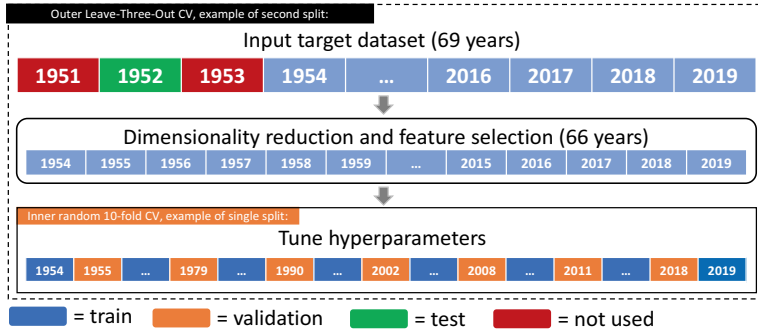


Figure 5.A.2: Schematic of double cross-validation, showing the second split of the Leave-Three-Out (LTO) outer CV. Note that we only predict a single year out of sample per training fold. Hence, this process is repeated until all years are predicted (i.e., there are 69 outer training folds). The inner CV is always random 10-fold CV.

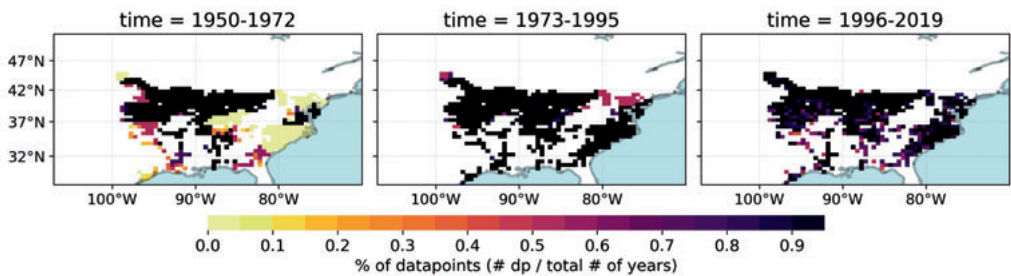


Figure 5.A.3: Coverage of datapoints in percentages per time-period of the mid-to-southern cluster (label 1 in Figure 2). The total # of years refers to the number of years within each time-period (23, 23 and 24 years).

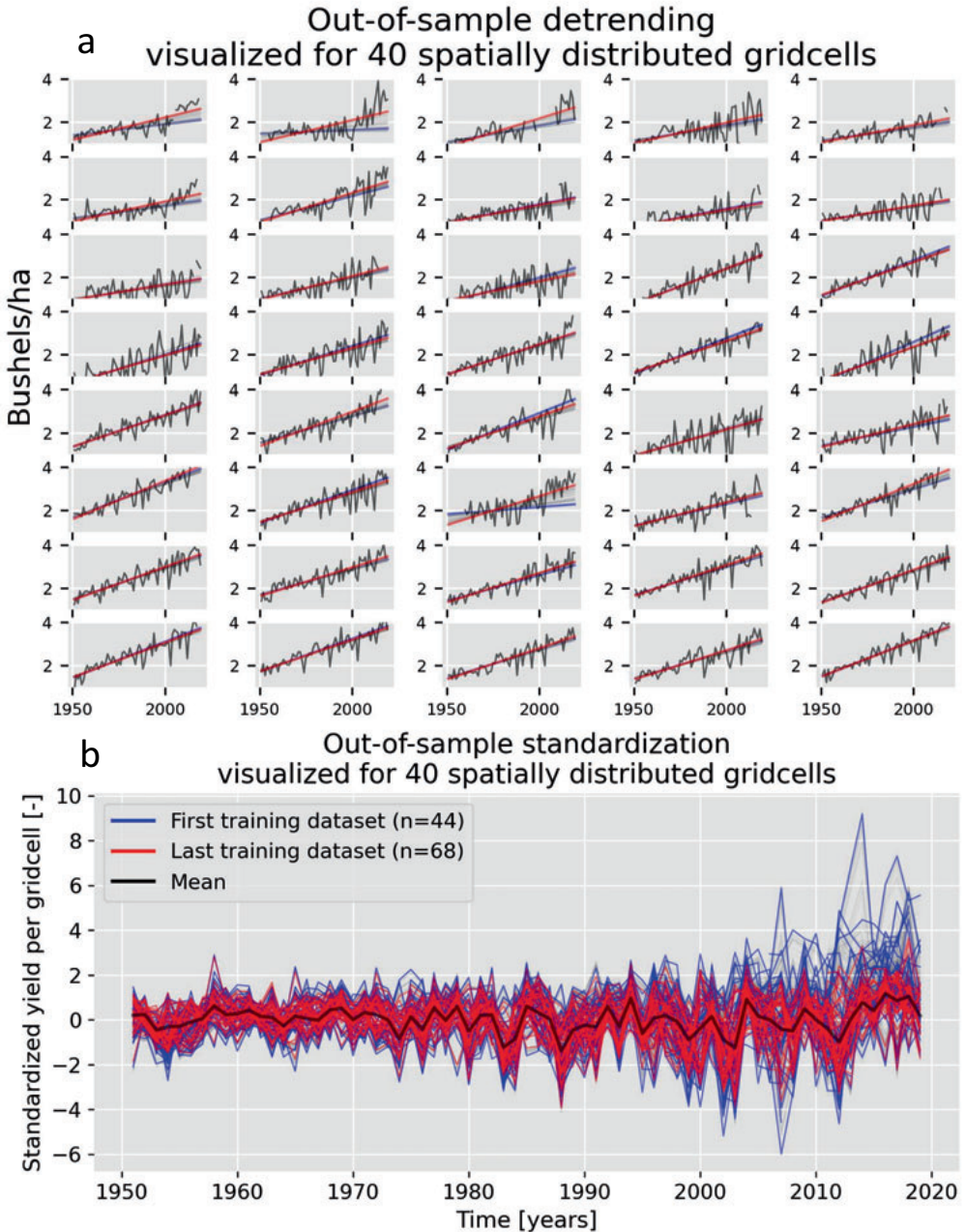


Figure 5.A.4: Same as Figure 5.A.1, but here done out-of-sample for the one-step-ahead-25 cross-validation. Blue indicates the first training dataset ($n=44$), red indicates the last training dataset ($n=68$). The trendline and standardized timeseries are here extrapolated to all remaining datapoints, yet in our final out-of-sample preprocessed timeseries, the extrapolation of the trend and standardization is done for a single year for each training dataset.

Figure 5.B.4 and Figure 5.B.5 show the robust SST regions and soil moisture patterns for the one-step-ahead cross-validation as discussed in section 5.2.3.

In DBSCAN, the radius (eps parameter) of 250 is chosen to define neighboring grid cells, which is found to produce regions of reasonable sizes and spatial separation. Well separated very small regions (size of ~ 1 gridcell) regions are automatically ignored. As explained in the method section, DBSCAN tends to create one single very large precursor region (the horseshoe Pacific region), while adjacent smaller regions are kept separate. This could lead to physically non-sensible partial correlation tests since adjacent regions are expected to correlate due to their spatial proximity. Hence, in a second step we find precursor regions which are close to each other, but before clustering them together, we validate if they are indeed sufficiently correlating (coefficient should be approx. > 0.4). In this second step, we calculate the inter-cluster haversine distance based on the center of each precursor region. Because we now use only the center latitude longitude location of the regions, the distances between the features are much larger compared to the first DBSCAN step (where gridcells were often right next to each other). Therefore, the radius (eps parameter) is now set to 2000, which clusters the regions together at a realistic distance, i.e., they are expected to be correlating due to their spatial proximity.

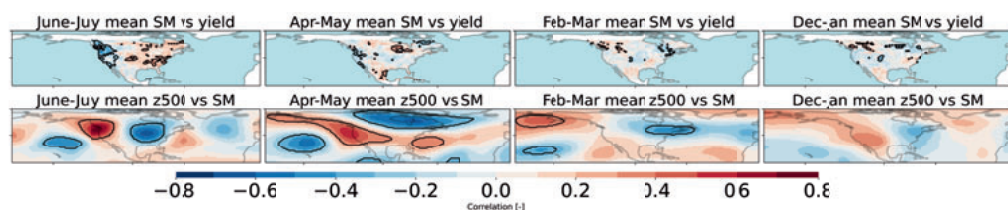


Figure 5.B.1: Top row: the correlation maps of soil moisture versus end-of-year crop yield at different lags. The soil moisture correlation maps correspond to the first column of Figure 5.7. Bottom row: the correlation between geopotential height at 500 hPa versus the spatial covariance timeseries of the soil moisture patterns (mean timeseries over the training samples of the LTO cross-validation).

5.C Forecast verification

The observed low yield events (Figure 5.8) shows some multi-year variability in the frequency of events, with less events occurring between 1955-1974 and 2013-2019 (16 events / 10 years), yet more events happening between 1975-2012 (47 events / 10 years). This decadal variability can be expected, given the importance of the extra-tropical Pacific SST variability in affecting the weather in eastern US. The Pacific is well known for its decadal variability associated with the Pacific Decadal Oscillation (Newman et al., 2016). Due to this decadal variability, the frequency of events (from here on called base-rate) was 40% in the recent 25 years. This deteriorates the benchmark forecast skill and thereby could lead to 'spurious' skill. However, changing the climatological benchmark to the true base-rate in the test set had a negligible effect on the BSS. On the other hand, the forecast model is challenged as it has to operate in a climate where the base-rate is different from what it was trained upon, the latter being 33%. To further investigate, we apply the one-step-ahead CV and concomitant pre-processing over different timespans

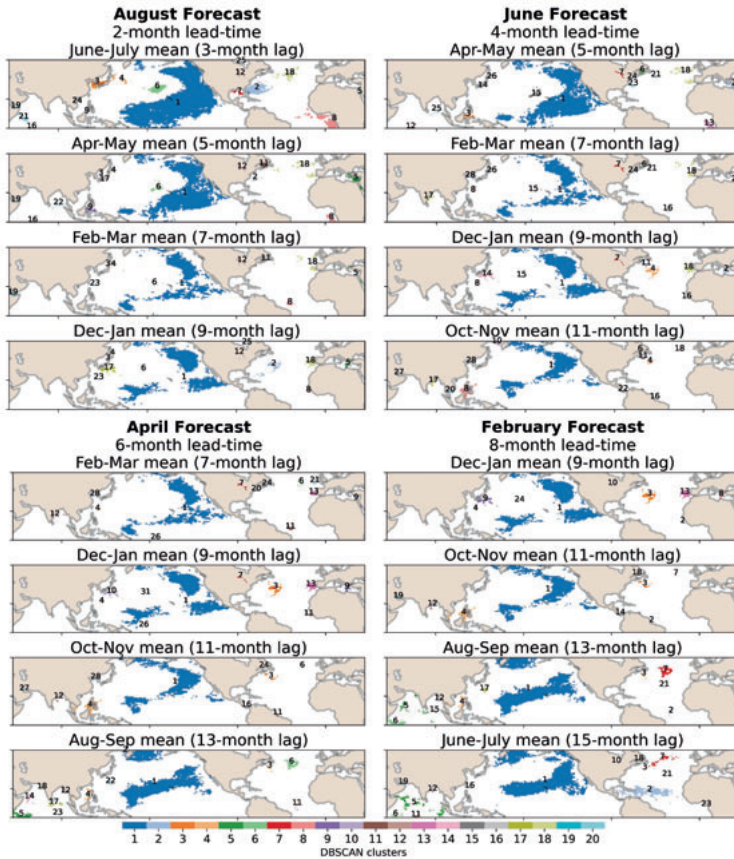


Figure 5.B.2: SST precursor regions clustered by the DBSCAN algorithm. Each label is used as a mask to calculate area-weighted and correlation-weighted timeseries.

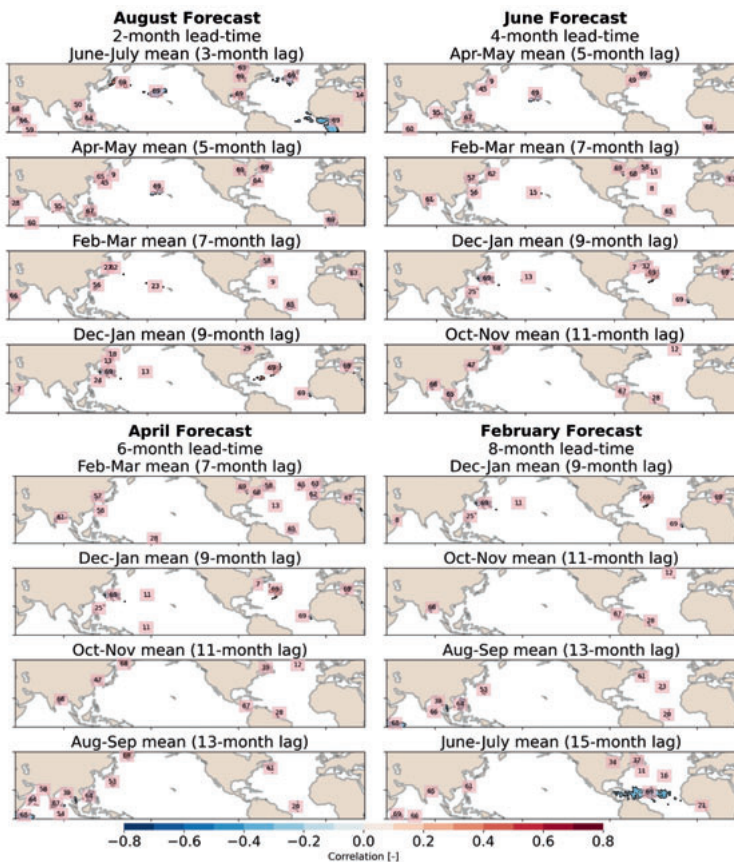


Figure 5.B.3: SST precursor regions that were identified by the RGDR method with the leave-three-out cross validation, but were filtered out by the precursor selection step, i.e., they were found conditionally independent given the influence of another precursor region timeseries. The integers denote the amount of training datasets in which the correlating region was extracted by the RGDR method.

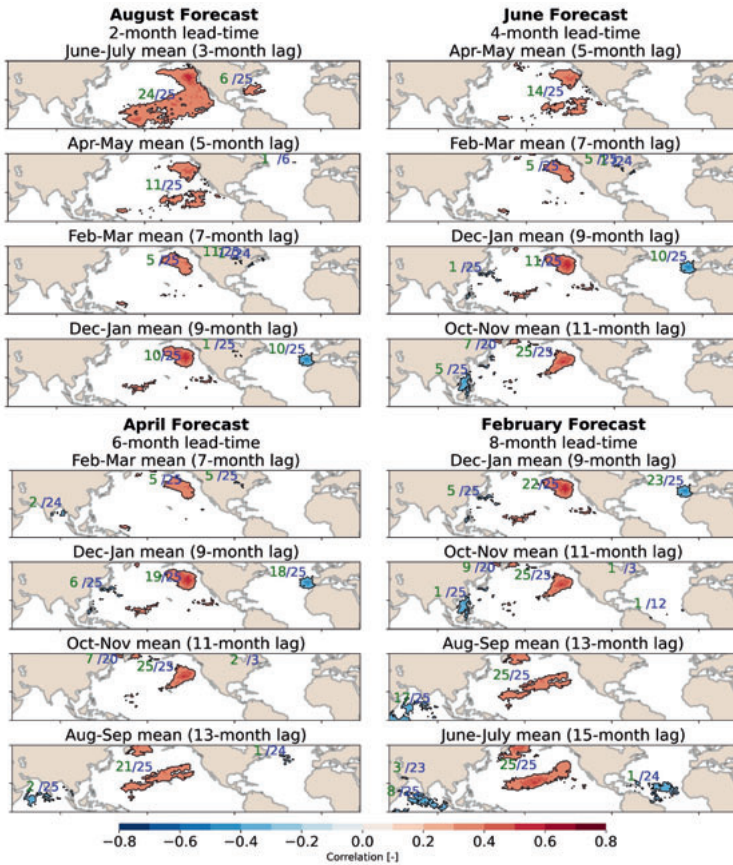


Figure 5.B.4: SST (2-month mean) correlation maps versus the crop yield variability in cluster 1 (see Figure 5.2) for each forecast month after the selection step. A correlation value is only shown if a gridcell is significantly correlating in one of the 25 training datasets. The green integers denote the number of training samples the precursors timeseries has passed all the conditional independence tests. The blue integers denote the number of training datasets the precursor timeseries is detected by the response-guided dimensionality reduction (RGDR). For clarity, we only show the regions which were conditionally dependent in at least 13 of the 25 training samples. Same Figure 5.6, but using the one-step-ahead 25 years cross-validation.

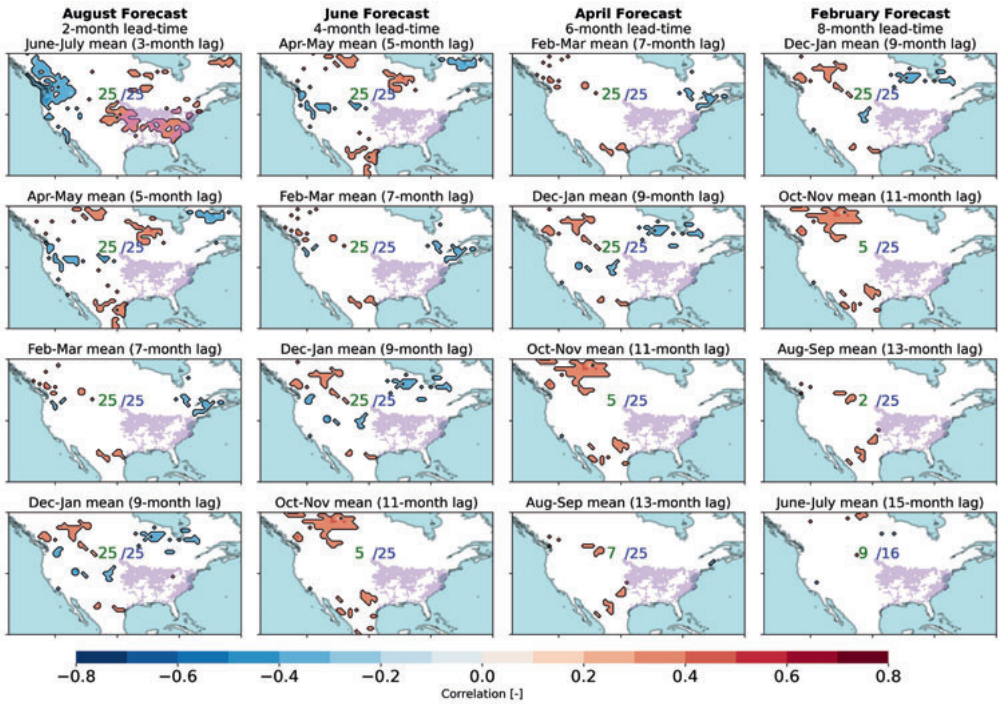


Figure 5.B.5: SM (SSI-2) correlation maps versus the crop yield variability in cluster 1 (see Figure 5.2) for each forecast month after the selection step. A correlation value is only shown if a gridcell is significantly correlating in one of the 69 training datasets. The SM precursor timeseries is based upon the spatial covariance of the (significant) correlation values. The ratio shows the conditional dependent/detected precursor timeseries, similar to Figure 5.B.4. If the SM timeseries is not conditionally independent in at least 13 of the 25 training samples, the SM correlation pattern is completely masked. The spatial domain of cluster 1 is shown in light pink.

(1990-2019 [30 years], 1995-2019 [25 years], and 2000-2019 [20 years]), with base-rates of 36%, 30%, and 36%. In general, we observe that the differences in skill during the window of predictability for the one-step-ahead 30, 25, and 20 fall within an expected sampling bias, with perhaps the one-step-ahead 30 performing slightly worse (Figure 5.C.1). The relatively small amount of training data (38 up to 43 years) for the first 5 years could have played a role in the small drop in skill. Figure 5.C.1 shows that the high forecast skill is robust against changing the verification period. This suggests that the forecast generalizes well, even when confronted with different base-rates in the test sets.

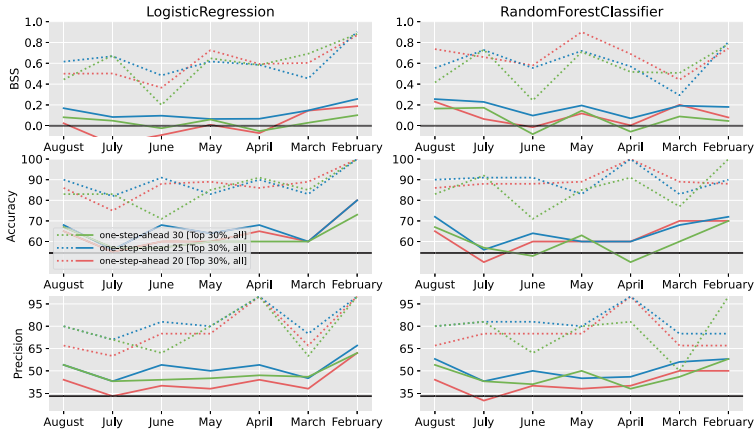


Figure 5.C.1: One-step-ahead forecast skill validated over different periods (the recent 30, 25 and 20 years) for the poor yield events in the mid-to-southern US cluster.

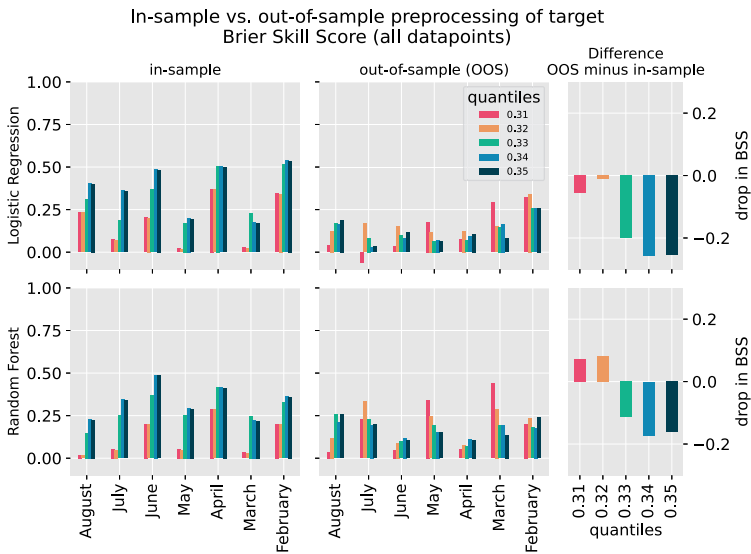


Figure 5.C.2: Visualizing the impact of in-sample versus out-of-sample pre-processing of the target variable (mid-to-southern US cluster) for the Brier Skill Score (BSS). Using the one-step-ahead cross-validation over the last 25 years. In addition, we test the skill when using different quantiles to observe the impact of minor changes in the binary event timeseries.

Publications

C. Raymond, D. Coumou, T. Foreman, A. King, K. Kornhuber, C. Lesk, C. Mora, S. Perkins-Kirkpatrick, S. Russo, and S. Vijverberg (2019). “Projections and Hazards of Future Extreme Heat”. In: *The Oxford Handbook of Planning for Climate Change Hazards*. Ed. by W. T. Pfeffer, J. B. Smith, and K. L. Ebi. December. Oxford University Press, pp. 1–43. DOI: [10.1093/oxfordhb/9780190455811.013.59](https://doi.org/10.1093/oxfordhb/9780190455811.013.59)

S. Vijverberg, M. Schmeits, K. van der Wiel, and D. Coumou (2020). “Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events”. In: *Monthly Weather Review* 148.12, pp. 4799–4822. DOI: [10.1175/MWR-D-19-0409.1](https://doi.org/10.1175/MWR-D-19-0409.1)

S. Vijverberg, D. Coumou, and M. Schmeits (2021). “Paper of note: Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events”. In: *Bulletin of the American Meteorological Society* 3.March, pp. 189–210. DOI: https://doi.org/10.1175/BAMS_1023_189-210_Nowcast

S. Vijverberg and D. Coumou (2022). “The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability”. In: *npj Climate and Atmospheric Science* 5.1, p. 18. DOI: [10.1038/s41612-022-00237-7](https://doi.org/10.1038/s41612-022-00237-7)

S. Vijverberg, R. Hamed, and D. Coumou (2022a). “Skilful US Soy-yield forecasts at pre-sowing lead-times”. In: *Artificial Intelligence for the Earth Systems* in review

R. Hamed, S. Vijverberg, A. Van Loon, and D. Coumou (2022). “Persistent La Nina conditions favour joint soybean production failures in North and South American regions”. In: *Environ. Res. Lett.* in review

Bibliography

- Alexander, M. A. et al. (2002). “The Atmospheric Bridge: The Influence of ENSO Teleconnections on Air–Sea Interaction over the Global Oceans”. In: *Journal of Climate* 15.16, pp. 2205–2231. DOI: 10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2.
- Alfaro, E. J., A. Gershunov, and D. Cayan (2006). “Prediction of summer maximum and minimum temperature over the central and western United States: The roles of soil moisture and sea surface temperature”. In: *Journal of Climate* 19.8, pp. 1407–1421. DOI: 10.1175/JCLI3665.1.
- Alley, R. B., K. A. Emanuel, and F. Zhang (2019). “Weather: Advances in weather prediction”. In: *Science* 363.6425, pp. 342–344. DOI: 10.1126/science.aav7274.
- Altenhoff, A. M. et al. (2008). “Linkage of atmospheric blocks and synoptic-scale Rossby waves: A climatological analysis”. In: *Tellus, Series A: Dynamic Meteorology and Oceanography* 60.5, pp. 1053–1063. DOI: 10.1111/j.1600-0870.2008.00354.x.
- Alvarez-Castro, M. C., D. Faranda, and P. Yiou (2018). “Atmospheric Dynamics Leading to West European Summer Hot Temperatures Since 1851”. In: *Complexity* 2018.2494509, pp. 1–10. DOI: 10.1155/2018/2494509.
- Anderson, W. et al. (2017). “Life cycles of agriculturally relevant <scp>ENSO</scp> teleconnections in North and South America”. In: *International Journal of Climatology* 37.8, pp. 3297–3318. DOI: 10.1002/joc.4916.
- Ardilouze, C. et al. (2017). “Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability”. In: *Climate Dynamics* 49.11-12, pp. 3959–3974. DOI: 10.1007/s00382-017-3555-7.
- Arya, H., M. B. Singh, and P. L. Bhalla (2021). “Towards Developing Drought-smart Soybeans”. In: *Frontiers in Plant Science* 12.October. DOI: 10.3389/fpls.2021.750664.
- Baker, H. S. et al. (2019). “Forced summer stationary waves: the opposing effects of direct radiative forcing and sea surface warming”. In: *Climate Dynamics* 53.7-8, pp. 4291–4309. DOI: 10.1007/s00382-019-04786-1.
- Barnes, E. A. and R. J. Barnes (2021). “Controlled Abstention Neural Networks for Identifying Skillful Predictions for Classification Problems”. In: *Journal of Advances in Modeling Earth Systems* 13.12, pp. 1–15. DOI: 10.1029/2021MS002573. arXiv: 2104.08281.
- Barnes, E. a. and D. L. Hartmann (2010). “Dynamical Feedbacks and the Persistence of the NAO”. In: *Journal of the Atmospheric Sciences* 67.3, pp. 851–865. DOI: 10.1175/2009JAS3193.1.

- Barnes, E. A. and D. L. Hartmann (2011). “Rossby Wave Scales, Propagation, and the Variability of Eddy-Driven Jets”. In: *Journal of the Atmospheric Sciences* 68.12, pp. 2893–2908. DOI: 10.1175/JAS-D-11-039.1.
- Barnes, E. A. and J. A. Screen (2015). “The impact of Arctic warming on the midlatitude jet-stream: Can it? Has it? Will it?” In: *WIREs Clim Change* 6, pp. 277–286. DOI: 10.1002/wcc.337.
- Barnes, E. A. et al. (2014). “Exploring recent trends in Northern Hemisphere blocking”. In: *Geophysical Research Letters* 2, pp. 638–644. DOI: 10.1002/2013GL058745. Received.
- Basso, B. and L. Liu (2019). “Seasonal crop yield forecast: Methods, applications, and accuracies”. In: *Advances in Agronomy* 154. January, pp. 201–255. DOI: 10.1016/bs.agron.2018.11.002.
- Bauer, P., A. Thorpe, and G. Brunet (2015). “The quiet revolution of numerical weather prediction”. In: *Nature* 525.7567, pp. 47–55. DOI: 10.1038/nature14956.
- Beguiería, S. and M. P. Maneta (2020). “Qualitative crop condition survey reveals spatiotemporal production patterns and allows early yield prediction”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.31, pp. 18317–18323. DOI: 10.1073/pnas.1917774117.
- Belgiu, M. and L. Drăgu (2016). “Random forest in remote sensing: A review of applications and future directions”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 114, pp. 24–31. DOI: 10.1016/j.isprsjprs.2016.01.011.
- Bello, G. A. et al. (2015). “Response-Guided Community Detection: Application to Climate Index Discovery”. In: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD*. Springer, pp. 736–751. DOI: 10.1007/978-3-319-23525-7_45.
- Benjamini, Y. and Y. Hochberg (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
- Bett, P. et al. (2018). *Assessment Of Seasonal Forecasting Skill For Energy Variables*. Tech. rep. D3.4.1, pp. 1–42.
- Bhardwaj, G., G. Gorton, and G. Rouwenhorst (2015). *Facts and Fantasies about commodity futures ten years later*. Tech. rep. National bureau of economic research.
- Birol, F. (2021). *The world’s electricity systems must be ready to counter the growing climate threat*.
- Bloomfield, H. C. et al. (2021). “Pattern-based conditioning enhances sub-seasonal prediction skill of European national energy variables”. In: *Meteorological Applications* 28.4, pp. 1–16. DOI: 10.1002/met.2018.
- Boschat, G. et al. (2016). “On the use of composite analyses to form physical hypotheses: An example from heat wave – SST associations”. In: *Scientific Reports* 6.1, p. 29599. DOI: 10.1038/srep29599.
- Branstator, G. (2002). “Circumglobal teleconnections, the jet stream waveguide, and the North Atlantic Oscillation”. In: *Journal of Climate* 15.14, pp. 1893–1910. DOI: 10.1175/1520-0442(2002)015<1893:CTTJSW>2.0.CO;2.
- Branstator, G. and H. Teng (2017). “Tropospheric Waveguide Teleconnections and Their Seasonality”. In: *J Atmos Sci* 74.5, pp. 1513–1532. DOI: 10.1175/JAS-D-16-0305.1.
- Brewer, M. C. and C. F. Mass (2016). “Projected changes in western U.S. large-scale summer synoptic circulations and variability in CMIP5 models”. In: *Journal of Climate* 29.16, pp. 5965–5978. DOI: 10.1175/JCLI-D-15-0598.1.

- Brown, J. N. et al. (2018). “Seasonal climate forecasts provide more definitive and accurate crop yield predictions”. In: *Agricultural and Forest Meteorology* 260-261, June, pp. 247–254. DOI: 10.1016/j.agrformet.2018.06.001.
- Bueso, D., M. Piles, and G. Camps-Valls (2020). “Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.8, pp. 5752–5763. DOI: 10.1109/TGRS.2020.2969813. arXiv: arXiv:2002.04539v1.
- Cai, M. et al. (2016). “Feeling the pulse of the stratosphere an emerging opportunity for predicting continental-scale cold-air outbreaks 1 month in advance”. In: *Bulletin of the American Meteorological Society* 97.8, pp. 1475–1489. DOI: 10.1175/BAMS-D-14-00287.1.
- Carter, E. K. et al. (2018). “Yield response to climate, management, and genotype: A large-scale observational analysis to identify climate-adaptive crop management practices in high-input maize systems”. In: *Environmental Research Letters* 13.11. DOI: 10.1088/1748-9326/aae7a8.
- Cattiaux, J., Y. Peings, and D. Saint-martin (2016). “Sinuosity of mid-latitude atmospheric flow in a warming world”. In: DOI: 10.1002/2016GL070309. Received.
- Centurioni, L. R. et al. (2019). “Global in situ Observations of Essential Climate and Ocean Variables at the Air–Sea Interface”. In: *Frontiers in Marine Science* 6, JUL, pp. 1–23. DOI: 10.3389/fmars.2019.00419.
- Chang, E. K. M., Y. Guo, and X. Xia (2012). “CMIP5 multimodel ensemble projection of storm track change under global warming”. In: *Journal of Geophysical Research Atmospheres* 117.23, pp. 1–19. DOI: 10.1029/2012JD018578.
- Cheung, A. H. et al. (2017). *Comparison of low-frequency internal climate variability in CMIP5 models and observations*. DOI: 10.1175/JCLI-D-17-0438.1.
- Chevallier, M. et al. (2019). “Chapter 10 - The Role of Sea Ice in Sub-seasonal Predictability”. In: *Sub-Seasonal to Seasonal Prediction*. Ed. by A. W. Robertson and F. Vitart. Elsevier, pp. 201–221. DOI: <https://doi.org/10.1016/B978-0-12-811714-9.00010-3>.
- Cohen, J. et al. (2018). “S2S Reboot: An Argument for Greater Inclusion of Machine Learning in Subseasonal to Seasonal (S2S) Forecasts”. In: *Wiley Interdisciplinary Reviews: Climate Change* 567, pp. 1–15. DOI: 10.1002/wcc.567.
- Collins, M. et al. (2018). “Challenges and opportunities for improved understanding of regional climate dynamics”. In: *Nature Climate Change* 8.2, pp. 101–108. DOI: 10.1038/s41558-017-0059-8.
- Copernicus Climate Change Service (C3S) (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate*.
- Coughlan de Perez, E. (2018). “Forecast-based financing: a scientific foundation for systematic early action”. PhD thesis. Vrije Universiteit Amsterdam, p. 121.
- Coughlan de Perez, E. et al. (2017). “Should seasonal rainfall forecasts be used for flood preparedness?” In: *Hydrology and Earth System Sciences* 21.9, pp. 4517–4524. DOI: 10.5194/hess-21-4517-2017.
- Coumou, D., J. Lehmann, and J. Beckmann (2015). “The weakening summer circulation in the Northern Hemisphere mid-latitudes”. In: *Science* 348.6232, pp. 324–327. DOI: 10.1126/science.1261768.

- Coumou, D. et al. (2014). “Quasi-resonant circulation regimes and hemispheric synchronization of extreme weather in boreal summer”. In: *Proceedings of the National Academy of Sciences* 111.34, pp. 12331–12336. DOI: 10.1073/pnas.1412797111. arXiv: arXiv:1408.1149.
- Coumou, D. et al. (2018). “The influence of Arctic amplification on mid-latitude summer circulation”. In: *Nature Communications* 9.1, p. 2959. DOI: 10.1038/s41467-018-05256-8.
- Coumou, D. (2021). “Stratospheric winds trigger cold spells”. In: *Science* 373.6559, pp. 1091–1091. DOI: 10.1126/science.ab19792.
- Coumou, D. et al. (2017). “Weakened Flow, Persistent Circulation, and Prolonged Weather Extremes in Boreal Summer”. In: *Climate Extremes: Patterns and Mechanisms, Geophysical Monograph 226*. Ed. by S.-Y. Wang et al. First edit, pp. 61–73. DOI: 10.1002/9781119068020.ch4.
- Crane, T. A. et al. (2010). “Forecast Skill and Farmers’ Skills: Seasonal Climate Forecasts and Agricultural Risk Management in the Southeastern United States”. In: *Weather, Climate, and Society* 2.1, pp. 44–59. DOI: 10.1175/2009WCAS1006.1.
- Cronin, M. F. et al. (2019). “Air-Sea Fluxes With a Focus on Heat and Momentum”. In: *Frontiers in Marine Science* 6.JUL. DOI: 10.3389/fmars.2019.00430.
- De Perez, E. C. et al. (2016). “Action-based flood forecasting for triggering humanitarian action”. In: *Hydrology and Earth System Sciences* 20.9, pp. 3549–3560. DOI: 10.5194/hess-20-3549-2016.
- Demaeyer, J. and S. Vannitsem (2018). *Advances in Nonlinear Geosciences*, pp. 55–85. DOI: 10.1007/978-3-319-58895-7.
- Dembek, K., P. Singh, and V. Bhakoo (2016). “Literature Review of Shared Value: A Theoretical Concept or a Management Buzzword?” In: *Journal of Business Ethics* 137.2, pp. 231–267. DOI: 10.1007/s10551-015-2554-z.
- Deng, K. et al. (2018). “Increased frequency of summer extreme heat waves over Texas Area Tied to the amplification of pacific zonal SST gradient”. In: *Journal of Climate* 31.14, pp. 5629–5647. DOI: 10.1175/JCLI-D-17-0554.1.
- Deo, R. C. et al. (2017). “Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model”. In: *Stochastic Environmental Research and Risk Assessment* 31.5, pp. 1211–1240. DOI: 10.1007/s00477-016-1265-z.
- Deser, C., M. A. Alexander, and M. S. Timlin (2003). “Understanding the persistence of sea surface temperature anomalies in midlatitudes”. In: *Journal of Climate* 16.1, pp. 57–72. DOI: 10.1175/1520-0442(2003)016<0057:UTPOSS>2.0.CO;2.
- Deser, C., R. A. Tomas, and S. Peng (2007). “The transient atmospheric circulation response to North Atlantic SST and sea ice anomalies”. In: *Journal of Climate* 20.18, pp. 4751–4767. DOI: 10.1175/JCLI4278.1.
- Deser, C. and K. Trenberth (2016). *The Climate Data Guide: Pacific Decadal Oscillation (PDO): Definition and Indices*.
- Deser, C. et al. (2010). *Sea Surface Temperature Variability: Patterns and Mechanisms*. Vol. 2. 1, pp. 115–143. DOI: 10.1146/annurev-marine-120408-151453.
- Deser, C. et al. (2014). “Projecting North American Climate over the Next 50 Years: Uncertainty due to Internal Variability*³”. In: *Journal of Climate* 27.6, pp. 2271–2296. DOI: 10.1175/JCLI-D-13-00451.1.

- Di Capua, G. et al. (2019a). “Long-Lead Statistical Forecasts of the Indian Summer Monsoon Rainfall Based on Causal Precursors”. In: *Weather and Forecasting* 34.5, pp. 1377–1394. DOI: 10.1175/WAF-D-19-0002.1.
- Di Capua, G. et al. (2021). “Drivers behind the summer 2010 wave train leading to Russian heatwave and Pakistan flooding”. In: *npj Climate and Atmospheric Science* 4.1. DOI: 10.1038/s41612-021-00211-9.
- Di Capua, G. et al. (2019b). “Tropical and mid-latitude teleconnections interacting with the Indian summer monsoon rainfall: A Theory-Guided Causal Effect Network approach”. In: *Earth System Dynamics Discussions*, pp. 1–27. DOI: 10.5194/esd-2019-11.
- Di Capua, G. et al. (2020). “Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales”. In: *Weather and Climate Dynamics* 1.2, pp. 519–539. DOI: 10.5194/wcd-1-519-2020.
- Diffenbaugh, N. S. et al. (2015). “Anthropogenic warming has increased drought risk in California”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.13, pp. 3931–3936. DOI: 10.1073/pnas.1422385112.
- Ding, Q. and B. Wang (2005). “Circumglobal teleconnection in the Northern Hemisphere summer”. In: *Journal of Climate* 18.17, pp. 3483–3505. DOI: 10.1175/JCLI3473.1.
- Ding, Q. et al. (2011). “Tropical-extratropical teleconnections in boreal summer: Observed interannual variability”. In: *Journal of Climate* 24.7, pp. 1878–1896. DOI: 10.1175/2011JCLI3621.1.
- Dirmeyer, P. A. (2003). “The role of the land surface background state in climate predictability”. In: *Journal of Hydrometeorology* 4.3, pp. 599–610. DOI: 10.1175/1525-7541(2003)004<0599:TR0TLS>2.0.CO;2.
- Dirmeyer, P. A. et al. (2013). “Trends in land-atmosphere interactions from CMIP5 simulations”. In: *Journal of Hydrometeorology* 14.3, pp. 829–849. DOI: 10.1175/JHM-D-12-0107.1.
- Doblas-Reyes, F. J. et al. (2013). “Seasonal climate predictability and forecasting: Status and prospects”. In: *Wiley Interdisciplinary Reviews: Climate Change* 4.4, pp. 245–268. DOI: 10.1002/wcc.217.
- Dobrynin, M. et al. (2018). “Improved teleconnection-based dynamical seasonal predictions of boreal winter”. In: *Geophysical Research Letters* 45, pp. 3605–3614. DOI: 10.1002/2018GL077209.
- Dong, S. et al. (2019). “A study on soybean responses to drought stress and rehydration”. In: *Saudi Journal of Biological Sciences* 26.8, pp. 2006–2017. DOI: 10.1016/j.sjbs.2019.08.005.
- Donges, J. F. et al. (2016). “Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks”. In: *European Physical Journal: Special Topics* 225.3, pp. 471–487. DOI: 10.1140/epjst/e2015-50233-y. arXiv: 1508.03534.
- Donkor, F. K. et al. (2019). “Climate Services and Communication for Development: The Role of Early Career Researchers in Advancing the Debate”. In: *Environmental Communication* 13.5, pp. 561–566. DOI: 10.1080/17524032.2019.1596145.
- Dorrington, J. et al. (2020). “Beyond skill scores: exploring sub-seasonal forecast value through a case-study of French month-ahead energy prediction”. In: *Quarterly Journal of the Royal Meteorological Society* 146.733, pp. 3623–3637. DOI: 10.1002/qj.3863. arXiv: 2002.01728.

- Duchez, A. et al. (2016). “Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 European heat wave”. In: *Environmental Research Letters* 11.7. DOI: 10.1088/1748-9326/11/7/074004.
- Eade, R. et al. (2014). “Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?”. In: *Geophysical Research Letters* 41.15, pp. 5620–5628. DOI: 10.1002/2014GL061146.
- Ebert-Uphoff, I. and Y. Deng (2012). “Causal discovery for climate research using graphical models”. In: *Journal of Climate* 25.17, pp. 5648–5665. DOI: 10.1175/JCLI-D-11-00387.1.
- Ester, M. et al. (1996). “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*. AAAI Press.
- EU (2022). *Save gas for a safe winter - Communication from the commission to the european, parliament, the council, the european economic and social committee and the committee of the regions*. Tech. rep., pp. 1–17.
- Eyring, V. et al. (2019). “Taking climate model evaluation to the next level”. In: *Nature Climate Change* 9.2, pp. 102–110. DOI: 10.1038/s41558-018-0355-y.
- Fawcett, T. (2006). “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8, pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.
- Fehlenberg, V. et al. (2017). “The role of soybean production as an underlying driver of deforestation in the South American Chaco”. In: *Global Environmental Change* 45. April, pp. 24–34. DOI: 10.1016/j.gloenvcha.2017.05.001.
- Ferreira, D. and C. Frankignoul (2005). “The transient atmospheric response to midlatitude SST anomalies”. In: *Journal of Climate* 18.7, pp. 1049–1067. DOI: 10.1175/JCLI-3313.1.
- Fery, L. et al. (2021). “Learning a weather dictionary of atmospheric patterns using Latent Dirichlet Allocation”. In: *HAL open science*. arXiv: hal-03258523.
- Finnis, J. et al. (2012). “Non-linear post-processing of numerical seasonal climate forecasts”. In: *Atmosphere - Ocean* 50.2, pp. 207–218. DOI: 10.1080/07055900.2012.667388.
- Frankignoul, C. (1985). “Sea surface temperature anomalies, planetary waves, and air-sea feedback in the middle latitudes”. In: *Reviews of Geophysics* 23.4, p. 357. DOI: 10.1029/RG023i004p00357.
- Frankignoul, C. and K. Hasselmann (1977). “Stochastic climate models, Part II Application to sea-surface temperature anomalies and thermocline variability”. In: *Tellus* 29.4, pp. 289–305. DOI: 10.3402/tellusa.v29i4.11362.
- Frankignoul, C. and N. Sennéchaël (2007). “Observed influence of North Pacific SST anomalies on the atmospheric circulation”. In: *Journal of Climate* 20.3, pp. 592–606. DOI: 10.1175/JCLI4021.1.
- Franzke, C. (2002). “Dynamics of Low-Frequency Variability: Barotropic Mode”. In: *Journal of the Atmospheric Sciences* 59.20, pp. 2897–2909. DOI: 10.1175/1520-0469(2002)059<2897:DOLFVB>2.0.CO;2.
- Galfi, V. M., V. Lucarini, and J. Wouters (2018). “A Large Deviation Theory-based Analysis of Heat Waves and Cold Spells in a Simplified Model of the General Circulation of the Atmosphere”. In: *arXiv*. DOI: 10.1088/1742-5468/ab02e8. arXiv: 1807.08261.

- García-Serrano, J. and C Frankignoul (2014). “Retraction Note: High predictability of the winter Euro–Atlantic climate from cryospheric variability”. In: *Nature Geoscience* 7.6, E2–E2. DOI: 10.1038/ngeo2164.
- Garfinkel, C. I. et al. (2020). “The Building Blocks of Northern Hemisphere Wintertime Stationary Waves”. In: *Journal of Climate* 33.13, pp. 5611–5633. DOI: 10.1175/JCLI-D-19-0181.1.
- Gerber, E. P. and G. K. Vallis (2007). “Eddy–Zonal Flow Interactions and the Persistence of the Zonal Index”. In: *Journal of the Atmospheric Sciences* 64.9, pp. 3296–3311. DOI: 10.1175/JAS4006.1.
- Gibson, P. B. et al. (2021). “Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts”. In: *Communications Earth & Environment* 2.1. DOI: 10.1038/s43247-021-00225-4.
- Giuliani, M. et al. (2020). “From skill to value: isolating the influence of end user behavior on seasonal forecast assessment”. In: *Hydrology and Earth System Sciences* 24.12, pp. 5891–5902. DOI: 10.5194/hess-24-5891-2020.
- Golding, B. (2022). *Towards the “Perfect” Weather Warning*. DOI: 10.1007/978-3-030-98989-7.
- Goodess, C. et al. (2019). “Advancing climate services for the European renewable energy sector through capacity building and user engagement”. In: *Climate Services* 16.December, p. 100139. DOI: 10.1016/j.cliser.2019.100139.
- Goulart, H. M. D. et al. (2021). “Storylines of weather-induced crop failure events under climate change”. In: *Earth System Dynamics* 12.4, pp. 1503–1527. DOI: 10.5194/esd-12-1503-2021.
- Guardian, T. (2021a). *Western US and Canada brace for another heatwave as wildfires spread*.
- (2021b). *‘Heat dome’ probably killed 1bn marine animals on Canada coast, experts say*.
- Guimarães Nobre, G (2019). “Floods, droughts and climate variability: From early warning to early action”. PhD thesis. Vrije Universiteit Amsterdam.
- Guimarães Nobre, G. et al. (2019). “Translating large-scale climate variability into crop production forecast in Europe”. In: *Scientific Reports* 9.1, p. 1277. DOI: 10.1038/s41598-018-38091-4.
- Haarsma, R. J., F. M. Selten, and S. S. Drijffhout (2015). “Decelerating Atlantic meridional overturning circulation main cause of future west European summer atmospheric circulation changes”. In: *Environmental Research Letters* 10. DOI: doi:10.1088/1748-9326/10/9/094007.
- Hall, R. J. et al. (2017). “Simple statistical probabilistic forecasts of the winter NAO”. In: *Weather and Forecasting* 32.4, pp. 1585–1601. DOI: 10.1175/WAF-D-16-0124.1.
- Ham, Y. G., J. H. Kim, and J. J. Luo (2019). “Deep learning for multi-year ENSO forecasts”. In: *Nature* 573.7775, pp. 568–572. DOI: 10.1038/s41586-019-1559-7.
- Hamed, R. et al. (2021). “Impacts of hot-dry compound extremes on US soybean yields”. In: *Earth System Dynamics Discussions* April, pp. 1–26. DOI: 10.5194/esd-2021-24.
- Hamed, R. et al. (2022). “Persistent La Nina conditions favour joint soybean production failures in North and South American regions”. In: *Environ. Res. Lett.* in review.
- Haqiqi, I. et al. (2020). “Quantifying the Impacts of Compound Extremes on Agriculture and Irrigation Water Demand”. In: *Hydrology and Earth System Sciences Discussions*, pp. 1–52. DOI: 10.5194/hess-2020-275.

- Haupt, S. E. et al. (2021). “Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194, p. 20200091. DOI: 10.1098/rsta.2020.0091.
- Hazeleger, W. et al. (2012). “EC-Earth V2.2: description and validation of a new seamless earth system prediction model”. In: *Climate Dynamics* 39.11, pp. 2611–2629. DOI: 10.1007/s00382-011-1228-5.
- Hegerl, G. C. et al. (2021). “Toward Consistent Observational Constraints in Climate Predictions and Projections”. In: *Frontiers in Climate* 3.June, pp. 1–22. DOI: 10.3389/fclim.2021.678109.
- Hermanson, L. et al. (2022). “WMO Global Annual to Decadal Climate Update: A Prediction for 2021–25”. In: *Bulletin of the American Meteorological Society* 103.4, E1117–E1129. DOI: 10.1175/BAMS-D-20-0311.1.
- Hessl, A. E., D. McKenzie, and R. Schellhaas (2004). “Drought and Pacific Decadal oscillation linked to fire occurrences in the inland Pacific northwest”. In: *America* 14.2, pp. 425–442. DOI: 10.1890/03-5019.
- Hewitt, H. T. et al. (2017). “Will high-resolution global ocean models benefit coupled predictions on short-range to climate timescales?” In: *Ocean Modelling* 120.November, pp. 120–136. DOI: 10.1016/j.ocemod.2017.11.002.
- Hodson, D. L. et al. (2010). “Climate impacts of recent multidecadal changes in Atlantic Ocean Sea Surface Temperature: A multimodel comparison”. In: *Climate Dynamics* 34.7, pp. 1041–1058. DOI: 10.1007/s00382-009-0571-2.
- Holton, J. R. (2004). *An introduction to dynamic meteorology*. Ed. by F. Cynar et al. 4th ed., p. 553.
- Horton, D. E. et al. (2015). “Contribution of changes in atmospheric circulation patterns to extreme temperature trends”. In: *Nature* 522.7557, pp. 465–469. DOI: 10.1038/nature14550.
- Horton, R. M. et al. (2016). “A Review of Recent Advances in Research on Extreme Heat Events”. In: *Current Climate Change Reports* 2.4, pp. 242–259. DOI: 10.1007/s40641-016-0042-x.
- Hoskins, B. and T. Woollings (2015). “Persistent Extratropical Regimes and Climate Extremes”. In: *Current Climate Change Reports* 1.3, pp. 115–124. DOI: 10.1007/s40641-015-0020-8.
- Hoskins, B. J. and T. Ambrizzi (1993). *Rosby Wave Propagation on a Realistic Longitudinally Varying Flow*. DOI: 10.1175/1520-0469(1993)050<1661:RWPOAR>2.0.CO;2.
- Hoskins, B. J. and D. J. Karoly (1981). “The Steady Linear Response of a Spherical Atmosphere to Thermal and Orographic Forcing”. In: *Journal of the Atmospheric Sciences* 38.6, pp. 1179–1196. DOI: 10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.
- Hoskins, B. J. and P. J. Valdes (1990). *On the Existence of Storm-Tracks*. DOI: 10.1175/1520-0469(1990)047<1854:OTEOST>2.0.CO;2.
- Iizumi, T. et al. (2018). “Global crop yield forecasting using seasonal climate information from a multi-model ensemble”. In: *Climate Services* 11.June, pp. 13–23. DOI: 10.1016/j.cliser.2018.06.003.
- Iizumi, T. et al. (2021). “Global within-season yield anomaly prediction for major crops derived using seasonal forecasts of large-scale climate indices and regional temperature

- and precipitation". In: *Weather and Forecasting* 36.1, pp. 285–299. DOI: 10.1175/WAF-D-20-0097.1.
- IPCC, (2021). *Summary for Policymakers*. Ed. by M. Delmotte et al.
- Jaccard, P. (1912). "the Distribution of the Flora in the Alpine Zone." In: *New Phytologist* 11.2, pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x.
- Jaiser, R. et al. (2012). "Impact of sea ice cover changes on the northern hemisphere atmospheric winter circulation". In: *Tellus, Series A: Dynamic Meteorology and Oceanography* 64.1, pp. 1–11. DOI: 10.3402/tellusa.v64i0.11595.
- Jézéquel, A., P. Yiou, and S. Radanovics (2017). "Role of circulation in European heatwaves using flow analogues". In: *Climate Dynamics* 50.April, pp. 1–15. DOI: 10.1007/s00382-017-3667-0.
- Jin, Z. et al. (2017). "The combined and separate impacts of climate extremes on the current and future US rainfed maize and soybean production under elevated CO₂". In: *Global Change Biology* 23.7, pp. 2687–2704. DOI: 10.1111/gcb.13617.
- Johnson, S. J. et al. (2019). "SEAS5: The new ECMWF seasonal forecast system". In: *Geoscientific Model Development* 12.3, pp. 1087–1117. DOI: 10.5194/gmd-12-1087-2019.
- Jong, B.-T. T., M. Ting, and R. Seager (2021). "Assessing ENSO Summer Teleconnections, Impacts, and Predictability in North America". In: *Journal of Climate* 34.9, pp. 1–47. DOI: 10.1175/jcli-d-20-0761.1.
- Kaczan, D. J. and J. Orgill-Meyer (2020). "The impact of climate change on migration: a synthesis of recent empirical insights". In: *Climatic Change* 158.3-4, pp. 281–300. DOI: 10.1007/s10584-019-02560-0.
- Kaspi, Y. and T. Schneider (2011). "Winter cold of eastern continental boundaries induced by warm ocean waters". In: *Nature* 471.7340, pp. 621–624. DOI: 10.1038/nature09924.
- Kennedy, D. et al. (2016). "The response of high-impact blocking weather systems to climate change". In: *Geophysical Research Letters* 43.13, pp. 7250–7258. DOI: 10.1002/2016GL069725.
- Kharin, V. V. and F. W. Zwiers (2003). "On the ROC score of probability forecasts". In: *Journal of Climate* 16.24, pp. 4145–4150. DOI: 10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.
- Kirtman, B. P. et al. (2014). "The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction". In: *Bulletin of the American Meteorological Society* 95.4, pp. 585–601. DOI: 10.1175/BAMS-D-12-00050.1.
- Klemm, T. and R. A. McPherson (2017). "The development of seasonal climate forecasting for agricultural producers". In: *Agricultural and Forest Meteorology* 232, pp. 384–399. DOI: 10.1016/j.agrformet.2016.09.005.
- Knutti, R. (2010). "The end of model democracy?" In: *Climatic Change* 102.3-4, pp. 395–404. DOI: 10.1007/s10584-010-9800-2.
- Kornhuber, K. et al. (2017a). "Evidence for wave resonance as a key mechanism for generating high-amplitude quasi-stationary waves in boreal summer". In: *Climate Dynamics* 49.5-6, pp. 1961–1979. DOI: 10.1007/s00382-016-3399-6.

- Kornhuber, K. et al. (2017b). “Summertime planetary wave resonance in the Northern and Southern hemispheres”. In: *Journal of Climate* 30.16, pp. 6133–6150. DOI: 10.1175/JCLI-D-16-0703.1.
- Kornhuber, K. et al. (2020). “Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions”. In: *Nature Climate Change* 10.1, pp. 48–53. DOI: 10.1038/s41558-019-0637-z.
- Koster, R. D. et al. (2006). “GLACE: The Global Land-Atmosphere Coupling Experiment. Part I: Overview”. In: *Journal of Hydrometeorology* 7.4, pp. 590–610. DOI: 10.1175/JHM510.1.
- Kretschmer, M., J. Runge, and D. Coumou (2017a). “Early prediction of extreme stratospheric polar vortex states based on causal precursors”. In: *Geophysical Research Letters* 44.16, pp. 8592–8600. DOI: 10.1002/2017GL074696.
- Kretschmer, M. et al. (2017b). “More-Persistent Weak Stratospheric Polar Vortex States Linked to Cold Extremes”. In: *Bulletin of the American Meteorological Society* 99.1, BAMS-D-16-0259.1. DOI: 10.1175/BAMS-D-16-0259.1.
- Kretschmer, M. et al. (2021). “Quantifying Causal Pathways of Teleconnections”. In: *Bulletin of the American Meteorological Society* 102.12, E2247–E2263. DOI: 10.1175/BAMS-D-20-0117.1.
- Krishnamurthy, V. (2019). “Predictability of Weather and Climate”. In: *Earth and Space Science* 6.7, pp. 1043–1056. DOI: 10.1029/2019EA000586.
- Kurtzman, D. and B. R. Scanlon (2007). “El Niño-Southern Oscillation and Pacific Decadal Oscillation impacts on precipitation in the southern and central United States: Evaluation of spatial distribution and predictions”. In: *Water Resources Research* 43.10, pp. 1–12. DOI: 10.1029/2007WR005863.
- Kushnir, Y. and N.-C. Lau (1992). “The General Circulation Model Response to a North Pacific SST Anomaly: Dependence on Time Scale and Pattern Polarity”. In: *Journal of Climate* 5.1, pp. 271–283.
- Kushnir, Y. et al. (2002). “Atmospheric GCM Response to Extratropical SST Anomalies: Synthesis and Evaluation*”. In: *Journal of Climate* 15.16, pp. 2233–2256. DOI: 10.1175/1520-0442(2002)015<2233:AGRTES>2.0.CO;2.
- Kushnir, Y. et al. (2019). “Towards operational predictions of the near-term climate”. In: *Nature Climate Change* 9.2, pp. 94–101. DOI: 10.1038/s41558-018-0359-7.
- Kusunose, Y. and R. Mahmood (2016). “Imperfect forecasts and decision making in agriculture”. In: *Agricultural Systems* 146, pp. 103–110. DOI: 10.1016/j.agsy.2016.04.006.
- Lagerquist, R. et al. (2020). “Deep learning on three-dimensional multiscale data for next-hour tornado prediction”. In: *Monthly Weather Review* 148.7, pp. 2837–2861. DOI: 10.1175/MWR-D-19-0372.1.
- Lau, N. C. and M. J. Nath (2001). “Impact of ENSO on SST variability in the North Pacific and North Atlantic: Seasonal dependence and role of extratropical sea-air coupling”. In: *Journal of Climate* 14.13, pp. 2846–2866. DOI: 10.1175/1520-0442(2001)014<2846:IOEOSV>2.0.CO;2.
- Lau, W. K. M. and K.-M. Kim (2012). “The 2010 Pakistan Flood and Russian Heat Wave: Teleconnection of Hydrometeorological Extremes”. In: *Journal of Hydrometeorology* 13.1, pp. 392–403. DOI: 10.1175/JHM-D-11-016.1.

- (2015). “Robust Hadley Circulation changes and increasing global dryness due to CO 2 warming from CMIP5 model projections”. In: *Proceedings of the National Academy of Sciences* 112.12, p. 201418682. DOI: 10.1073/pnas.1418682112.
- Lee, M. H. et al. (2017). “The recent increase in the occurrence of a boreal summer teleconnection and its relationship with temperature extremes”. In: *Journal of Climate* 30.18, pp. 7493–7504. DOI: 10.1175/JCLI-D-16-0094.1.
- Lehmann, J. and D. Coumou (2015). “The influence of mid-latitude storm tracks on hot, cold, dry, and wet extremes”. In: *Nature Scientific Reports* 5.17491.
- Lehmann, J. et al. (2020). “Potential for Early Forecast of Moroccan Wheat Yields Based on Climatic Drivers”. In: *Geophysical Research Letters* 47.12, pp. 1–10. DOI: 10.1029/2020GL087516.
- Lehmann, J. et al. (2014). “Future changes in extratropical storm tracks and baroclinicity under climate change”. In: *Environmental Research Letters* 9.8. DOI: 10.1088/1748-9326/9/8/084002.
- Lemos, M. C. et al. (2002). “The use of seasonal climate forecasting in policymaking: Lessons from Northeast Brazil”. In: *Climatic Change* 55.4, pp. 479–507. DOI: 10.1023/A:1020785826029.
- Lhotka, O., J. Kyselý, and A. Farda (2018). “Climate change scenarios of heat waves in Central Europe and their uncertainties”. In: *Theoretical and Applied Climatology* 131.3-4, pp. 1043–1054. DOI: 10.1007/s00704-016-2031-3.
- Lhotka, O., J. Kyselý, and E. Plavcová (2017). “Evaluation of major heat waves’ mechanisms in EURO-CORDEX RCMs over Central Europe”. In: *Climate Dynamics*, pp. 1–14. DOI: 10.1007/s00382-017-3873-9.
- Li, J. et al. (2020). “Accurate data-driven prediction does not mean high reproducibility”. In: *Nature Machine Intelligence* 2.1, pp. 13–15. DOI: 10.1038/s42256-019-0140-2.
- Li, L., R. W. Schmitt, and C. C. Ummenhofer (2022). “Skillful Long-Lead Prediction of Summertime Heavy Rainfall in the US Midwest From Sea Surface Salinity”. In: *Geophysical Research Letters* 49.13, pp. 1–10. DOI: 10.1029/2022GL098554.
- Li, L. et al. (2016). “Implications of North Atlantic sea surface salinity for summer precipitation over the U.S. Midwest: Mechanisms and predictive value”. In: *Journal of Climate* 29.9, pp. 3143–3159. DOI: 10.1175/JCLI-D-15-0520.1.
- Li, Y. et al. (2019). “Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States”. In: *Global Change Biology* 25.7, pp. 2325–2337. DOI: 10.1111/gcb.14628.
- Lin, H., R. Mo, and F. Vitart (2022). “The 2021 Western North American Heatwave and Its Subseasonal Predictions”. In: *Geophysical Research Letters* 49.6. DOI: 10.1029/2021GL097036.
- Liu, Q., N. Wen, and Z. Liu (2006). “An observational study of the impact of the North Pacific SST on the atmosphere”. In: *Geophysical Research Letters* 33.18, pp. 1–5. DOI: 10.1029/2006GL026082.
- Liu, Z. and E. Di Lorenzo (2018). “Mechanisms and Predictability of Pacific Decadal Variability”. In: *Current Climate Change Reports* 4.2, pp. 128–144. DOI: 10.1007/s40641-018-0090-5.
- Liu, Z. and L. Wu (2004). “Atmospheric response to North Pacific SST: The role of ocean-atmosphere coupling”. In: *Journal of Climate* 17.9, pp. 1859–1882. DOI: 10.1175/1520-0442(2004)017<1859:ARTNPS>2.0.CO;2.

- Liu, Z. et al. (2015). “Recent contrasting winter temperature changes over North America linked to enhanced positive Pacific-North American pattern”. In: *Geophysical Research Letters* 42.18, pp. 7750–7757. DOI: 10.1002/2015GL065656.
- Lobell, D. B., J. M. Deines, and S. D. Tommaso (2020). “Changes in the drought sensitivity of US maize yields”. In: *Nature Food* 1.11, pp. 729–735. DOI: 10.1038/s43016-020-00165-w.
- Lopez, H. and B. P. Kirtman (2019). “ENSO influence over the Pacific North American sector: uncertainty due to atmospheric internal variability”. In: *Climate Dynamics* 52.9-10, pp. 6149–6172. DOI: 10.1007/s00382-018-4500-0.
- Lorenz, D. J. and E. T. DeWeaver (2007). “Tropopause height and zonal wind response to global warming in the IPCC scenario integrations”. In: *Journal of Geophysical Research Atmospheres* 112.10, pp. 1–11. DOI: 10.1029/2006JD008087.
- Lorenz, D. J. and D. L. Hartmann (2003). “Eddy-zonal flow feedback in the Northern Hemisphere winter”. In: *Journal of Climate* 16.8, pp. 1212–1227. DOI: 10.1175/1520-0442(2003)16<1212:EFFITN>2.0.CO;2.
- Lorenz, E. N. (1969). “The predictability of a flow which possesses many scales of motion”. In: *Tellus* 21.3, pp. 289–307. DOI: 10.3402/tellusa.v21i3.10086.
- Luo, F. et al. (2022). “Summertime Rossby waves in climate models: substantial biases in surface imprint associated with small biases in upper-level circulation”. In: *Weather and Climate Dynamics* 3.3, pp. 905–935. DOI: 10.5194/wcd-3-905-2022.
- Luo, H. et al. (2020). “Ocean–atmosphere coupled Pacific Decadal variability simulated by a climate model”. In: *Climate Dynamics* 54.11-12, pp. 4759–4773. DOI: 10.1007/s00382-020-05248-9.
- Magagna, D et al. (2019). *Water - Energy Nexus in Europe*. Scientific analysis or review, Anticipation and foresight KJ-NA-29743-EN-N (online), KJ-NA-29743-EN-C (print), KJ-NA-29743-EN-E (ePub). Luxembourg (Luxembourg). DOI: 10.2760/968197 (online), 10.2760/285180 (print), 10.2760/541935 (ePub).
- Mahlstein, I. et al. (2012). “Changes in the odds of extreme events in the Atlantic basin depending on the position of the extratropical jet”. In: *Geophysical Research Letters* 39.22, n/a–n/a. DOI: 10.1029/2012GL053993.
- Mann, M. E. et al. (2017). “Influence of Anthropogenic Climate Change on Planetary Wave Resonance and Extreme Weather Events”. In: *Scientific Reports* 7. January. DOI: 10.1038/srep45242.
- Manola, I. et al. (2013). ““Waveguidability” of idealized jets”. In: *Journal of Geophysical Research Atmospheres* 118.18, pp. 10432–10440. DOI: 10.1002/jgrd.50758.
- Maraun, D. et al. (2017). “Towards process-informed bias correction of climate change simulations”. In: *Nature Climate Change* 7.11, pp. 764–773. DOI: 10.1038/nclimate3418.
- Mariotti, A. et al. (2020). “Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond”. In: *Bulletin of the American Meteorological Society* January 2020, pp. 608–625. DOI: 10.1175/bams-d-18-0326.1.
- Marshall, A. G. et al. (2017). “Impact of the quasi-biennial oscillation on predictability of the Madden–Julian oscillation”. In: *Climate Dynamics* 49.4, pp. 1365–1377. DOI: 10.1007/s00382-016-3392-0.

- Mason, S. J. and N. E. Graham (2002). “Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves”. In: *Quarterly Journal of the Royal Meteorological Society* 128, pp. 2145–2166.
- Mayer, K. J. and E. A. Barnes (2019). “Subseasonal Midlatitude Prediction Skill Following QBO-MJO Activity”. In: *Weather and Climate Dynamics* December.
- Mayer, K. J. and E. A. Barnes (2021). “Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network”. In: *Geophysical Research Letters* 48.10, pp. 1–9. DOI: 10.1029/2020GL092092.
- Mbow, C. et al. (2019). “Food Security”. In: *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. Ed. by P. Shukla et al. IPCC, pp. 437–520.
- McGovern, A. et al. (2017). “Using artificial intelligence to improve real-time decision-making for high-impact weather”. In: *Bulletin of the American Meteorological Society* 98.10, pp. 2073–2090. DOI: 10.1175/BAMS-D-16-0123.1.
- McGovern, A. et al. (2019). “Making the black box more transparent: Understanding the physical implications of machine learning”. In: *Bulletin of the American Meteorological Society* 100.11, pp. 2175–2199. DOI: 10.1175/BAMS-D-18-0195.1.
- McKee, T., N. Doesken, and J. Kleist (1993). “The relationship of drought frequency and duration to time scales”. In: *Eighth conference on applied climatology, American Meteorological Society*.
- McKinnon, K. A. et al. (2016). “Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures”. In: *Nature Geoscience* 9.5, pp. 389–394. DOI: 10.1038/ngeo2687.
- Merryfield, W. J. et al. (2020). “Current and emerging developments in subseasonal to decadal prediction”. In: *Bulletin of the American Meteorological Society* 101.6, E869–E896. DOI: 10.1175/BAMS-D-19-0037.1.
- Miao, Q. et al. (2019). “Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network”. In: *Water* 11.5, p. 977. DOI: 10.3390/w11050977.
- Michaelsen (1987). “Cross-Validation in Statistical Climate Forecast Models.pdf”. In: *American Meteorological Society* 26, pp. 1589–1600.
- Miralles, D. G. et al. (2014). “Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation”. In: *Nature Geoscience* 7.5, pp. 345–349. DOI: 10.1038/ngeo2141.
- Mitchell, D. et al. (2016). “Real-time extreme weather event attribution with forecast seasonal SSTs”. In: *Environmental Research Letters* 11.6, pp. 1–12. DOI: 10.1088/1748-9326/11/6/064006.
- Molnos, S. et al. (2017). “The sensitivity of the large-scale atmosphere circulation to changes in surface temperature gradients in the Northern Hemisphere”. In: *Earth System Dynamics Discussions* July, pp. 1–17.
- Murtagh, F. and P. Contreras (2012). “Algorithms for hierarchical clustering: An overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1, pp. 86–97. DOI: 10.1002/widm.53.
- Nakamura, N. and C. S. Y. Huang (2018). “Atmospheric blocking as a traffic jam in the jet stream”. In: *Science* 0721.May, eaat0721. DOI: 10.1126/science.aat0721.

- Namias, J. (1959). “Recent seasonal interactions between north Pacific waters and the overlying atmospheric circulation”. In: *Journal of Geophysical Research* 64.6, pp. 631–646. DOI: 10.1029/JZ064i006p00631.
- Namias, J. and R. M. Born (1970). “Temporal Coherence in North Pacific Sea-Surface Temperature Pattern”. In: *Journal of Geophysical Research* 75.30, pp. 5952–5955.
- Namias, J. and D. R. Cayan (1981). “Large-Scale Air-Sea Interactions and Short-Period Climatic Fluctuations”. In: *Science* 214.4523, pp. 869–876. DOI: 10.1126/science.214.4523.869.
- National Academies of Sciences (2016). *Next Generation Earth System Prediction*. Washington, D.C.: National Academies Press, pp. 1–335. DOI: 10.17226/21873.
- National Agricultural Statistics Service (2012). *The Yield Forecasting Program of NASS*.
- Newman, M. et al. (2003). “A Study of Subseasonal Predictability”. In: *Monthly Weather Review* 131.8, pp. 1715–1732. DOI: 10.1175//2558.1.
- Newman, M. et al. (2016). “The Pacific decadal oscillation, revisited”. In: *Journal of Climate* 29.12, pp. 4399–4427. DOI: 10.1175/JCLI-D-15-0508.1.
- Nie, Y. et al. (2016). “Delineating the Barotropic and Baroclinic Mechanisms in the Midlatitude Eddy-Driven Jet Response to Lower-Tropospheric Thermal Forcing”. In: *Journal of the Atmospheric Sciences* 73.1, pp. 429–448. DOI: 10.1175/jas-d-15-0090.1.
- O’Gorman, P. A. (2010). “Understanding the varied response of the extratropical storm tracks to climate change”. In: *Proceedings of the National Academy of Sciences* 107.45, pp. 19176–19180. DOI: 10.1073/pnas.1011547107.
- O’Reilly, C. H. et al. (2021). “Projections of northern hemisphere extratropical climate underestimate internal variability and associated uncertainty”. In: *Communications Earth & Environment* 2.1, p. 194. DOI: 10.1038/s43247-021-00268-7.
- Ortiz-Bobea, A. et al. (2019). “Unpacking the climatic drivers of US agricultural yields”. In: *Environmental Research Letters* 14.6. DOI: 10.1088/1748-9326/ab1e75.
- Oudar, T. et al. (2017). “Respective roles of direct GHG radiative forcing and induced Arctic sea ice loss on the Northern Hemisphere atmospheric circulation”. In: *Climate Dynamics* 49.11-12, pp. 3693–3713. DOI: 10.1007/s00382-017-3541-0.
- Paulson, L., P. A. Rocha, and T. Gillespie (2022). *French Nuclear Cuts Extend to Next Week as Temperatures Soar*.
- PBL et al. (2020). *Netherlands Climate and Energy Outlook 2020 - Summary*. Tech. rep.
- Pedregosa, F et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Peings, Y. et al. (2017). “Late twenty-first-century changes in the midlatitude atmospheric circulation in the CESM large ensemble”. In: *Journal of Climate* 30.15, pp. 5943–5960. DOI: 10.1175/JCLI-D-16-0340.1.
- Peng, S. and W. A. Robinson (2001). “Relationships between atmospheric internal variability and the responses to an extratropical SST anomaly”. In: *Journal of Climate* 14.13, pp. 2943–2959. DOI: 10.1175/1520-0442(2001)014<2943:RBAIVA>2.0.CO;2.
- Petoukhov, V. et al. (2013). “Quasiresonant amplification of planetary waves and recent Northern Hemisphere weather extremes”. In: *Proc. Natl. Acad. Sci.* 110.14, pp. 5336–5341. DOI: 10.1073/pnas.1222000110/-/DCSupplemental. www.pnas.org/cgi/doi/10.1073/pnas.1222000110. arXiv: arXiv:1408.1149.

- Petrie, R. E., L. C. Shaffrey, and R. T. Sutton (2015). “Atmospheric response in summer linked to recent Arctic sea ice loss”. In: *Quarterly Journal of the Royal Meteorological Society* 141.691, pp. 2070–2076. DOI: 10.1002/qj.2502.
- Pfahl, S. (2014). “Characterising the relationship between weather extremes in Europe and synoptic circulation features”. In: *Natural Hazards and Earth System Sciences* 14.6, pp. 1461–1475. DOI: 10.5194/nhess-14-1461-2014.
- Pfleiderer, P. and D. Coumou (2018). “Quantification of temperature persistence over the Northern Hemisphere land-area”. In: *Climate Dynamics* 51.1-2, pp. 627–637. DOI: 10.1007/s00382-017-3945-x.
- Philip, S. Y. et al. (2021). “Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021.” In: *World Weather Attribution* June, pp. 119–123.
- Pitcher, E. J. et al. (1988). “The Effect of North Pacific Sea Surface Temperature Anomalies on the January Climate of a General Circulation Model”. In: *Journal of the Atmospheric Sciences* 45.2, pp. 173–188. DOI: 10.1175/1520-0469(1988)045<0173:TEONPS>2.0.CO;2.
- Plavcová, E. and J. Kyselý (2016). “Overly persistent circulation in climate models contributes to overestimated frequency and duration of heat waves and cold spells”. In: *Climate Dynamics* 46.9-10, pp. 2805–2820. DOI: 10.1007/s00382-015-2733-8.
- Portmann, F. T., S. Siebert, and P. Döll (2010). “MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling”. In: *Global Biogeochemical Cycles* 24.1, n/a–n/a. DOI: 10.1029/2008gb003435.
- Putrasahan, D. A., A. J. Miller, and H. Seo (2013). “Isolating mesoscale coupled ocean-atmosphere interactions in the Kuroshio Extension region”. In: *Dynamics of Atmospheres and Oceans* 63, pp. 60–78. DOI: 10.1016/j.dynatmoce.2013.04.001.
- Qin, J. and W. A. Robinson (1993). “On the Rossby Wave Source and the Steady Linear Response to Tropical Forcing”. In: *Journal of the Atmospheric Sciences* 50.12, pp. 1819–1823. DOI: 10.1175/1520-0469(1993)050<1819:OTRWSA>2.0.CO;2.
- Ramírez-Rodriguez, M. A. et al. (2016). “The value of seasonal forecasts for irrigated, supplementary irrigated, And rainfed wheat cropping systems in northwest Mexico”. In: *Agricultural Systems* 147, pp. 76–86. DOI: 10.1016/j.agsy.2016.05.005.
- Raymond, C. et al. (2019). “Projections and Hazards of Future Extreme Heat”. In: *The Oxford Handbook of Planning for Climate Change Hazards*. Ed. by W. T. Pfeffer, J. B. Smith, and K. L. Ebi. December. Oxford University Press, pp. 1–43. DOI: 10.1093/oxfordhb/9780190455811.013.59.
- Ribeiro, M. T., S. Singh, and C. Guestrin (2016). “"Why Should I Trust You?"”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 1135–1144. DOI: 10.1145/2939672.2939778. arXiv: 1602.04938.
- Richard, G. (2018). *How Europe's wind drought and heat wave hit owners and operators*.
- Rivière, G., L. Robert, and F. Codron (2016). “A short-term negative eddy feedback on midlatitude jet variability due to planetary wave reflection”. In: *Journal of the Atmospheric Sciences* 73.11, pp. 4311–4328. DOI: 10.1175/JAS-D-16-0079.1.

- Robert, L., G. Rivière, and F. Codron (2017). “Positive and Negative Eddy Feedbacks Acting on Midlatitude Jet Variability in a Three-Level Quasigeostrophic Model”. In: *Journal of the Atmospheric Sciences* 74.5, pp. 1635–1649. DOI: 10.1175/JAS-D-16-0217.1.
- Robinson, A. et al. (2021). “Increasing heat and rainfall extremes now far outside the historical climate”. In: *npj Climate and Atmospheric Science* 4.1, p. 45. DOI: 10.1038/s41612-021-00202-w.
- Robinson, W. A. (2006). “On the self-maintenance of midlatitude jets”. In: *Journal of the Atmospheric Sciences* 63.8, pp. 2109–2122. DOI: 10.1175/JAS3732.1.
- Rodwell, M. J. and C. K. Folland (2002). “Atlantic air – sea interaction and seasonal predictability”. In: *Quarterly Journal of the Royal Meteorological Society* 128. February 2001, pp. 1413–1443.
- Röthlisberger, M., O. Martius, and H. Wernli (2018). “Northern Hemisphere Rossby Wave initiation events on the extratropical jet-A climatological analysis”. In: *Journal of Climate* 31.2, pp. 743–760. DOI: 10.1175/JCLI-D-17-0346.1.
- Röthlisberger, M. et al. (2019). “Recurrent synoptic-scale Rossby wave patterns and their effect on the persistence of cold and hot spells”. In: *Journal of Climate*, JCLI-D-18-0664.1. DOI: 10.1175/JCLI-D-18-0664.1.
- Rousi, E. et al. (2022). “Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia”. In: *Nature Communications* 13.1, p. 3851. DOI: 10.1038/s41467-022-31432-y.
- Runge, J. (2018). “Causal network reconstruction from time series: From theoretical assumptions to practical estimation”. In: *Chaos* 28.7. DOI: 10.1063/1.5025050.
- Runge, J., V. Petoukhov, and J. Kurths (2014). “Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models”. In: *Journal of Climate* 27.2, pp. 720–739. DOI: 10.1175/JCLI-D-13-00159.1.
- Runge, J. et al. (2012). “Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy”. In: *Physical Review E* 86.6, p. 061121. DOI: 10.1103/PhysRevE.86.061121.
- Runge, J. et al. (2015). “Identifying causal gateways and mediators in complex spatio-temporal systems”. In: *Nature Communications* 6, pp. 1–10. DOI: 10.1038/ncomms9502. arXiv: 1702.07007.
- Runge, J. et al. (2019). “Detecting and quantifying causal associations in large nonlinear time series datasets”. In: *Science Advances* 5.11. DOI: 10.1126/sciadv.aau4996. arXiv: 1702.07007.
- Saito, T. and M. Rehmsmeier (2015). “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”. In: *PLOS ONE* 10.3. Ed. by G. Brock, e0118432. DOI: 10.1371/journal.pone.0118432.
- Scaife, A. A. et al. (2016). “Seasonal winter forecasts and the stratosphere”. In: *Atmospheric Science Letters* 17.1, pp. 51–56. DOI: 10.1002/asl.598.
- Scaife, A. A. and D. Smith (2018). “A signal-to-noise paradox in climate science”. In: *npj Climate and Atmospheric Science* 1.1. DOI: 10.1038/s41612-018-0038-4.
- Scaife, A. A. et al. (2010). “Atmospheric blocking and mean biases in climate models”. In: *Journal of Climate* 23.23, pp. 6143–6152. DOI: 10.1175/2010JCLI3728.1.

- Schauberger, B., C. Gornott, and F. Wechsung (2017a). “Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting”. In: *Global Change Biology* 23.11, pp. 4750–4764. DOI: 10.1111/gcb.13738.
- Schauberger, B., J. Jägermeyr, and C. Gornott (2020). “A systematic review of local to regional yield forecasting approaches and frequently used data resources”. In: *European Journal of Agronomy* 120. June, p. 126153. DOI: 10.1016/j.eja.2020.126153.
- Schauberger, B. et al. (2017b). “Consistent negative response of US crops to high temperatures in observations and crop models”. In: *Nature Communications* 8. DOI: 10.1038/ncomms13931.
- Scheuerer, M. et al. (2020). “Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California”. In: *Monthly Weather Review* 148.8, pp. 3489–3506. DOI: 10.1175/mwr-d-20-0096.1.
- Schlenker, W. and M. J. Roberts (2009). “Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change”. In: *Proceedings of the National Academy of Sciences* 106.37, pp. 15594–15598. DOI: 10.1073/pnas.0906865106.
- Schnepf, R. (2017). “NASS and U.S. Crop Production Forecasts: Methods and Issues”. In: *Congressional Research Service*.
- Schubert, E. et al. (2017). “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN”. In: *ACM Transactions on Database Systems* 42.3, pp. 1–21. DOI: 10.1145/3068335.
- Schubert, S., H. Wang, and M. Suarez (2011). “Warm season subseasonal variability and climate extremes in the northern hemisphere: The role of stationary Rossby waves”. In: *Journal of Climate* 24.18, pp. 4773–4792. DOI: 10.1175/JCLI-D-10-05035.1.
- Screen, J. A. and I. Simmonds (2013). “Caution needed when linking weather extremes to amplified planetary waves”. In: *Proceedings of the National Academy of Sciences* 110.26, E2327–E2327. DOI: 10.1073/pnas.1304867110.
- Screen, J. A. and I. Simmonds (2014). “Amplified mid-latitude planetary waves favour particular regional weather extremes”. In: *Nature Climate Change* 4.8. DOI: 10.1038/nclimate2271.
- Screen, J. A. et al. (2013). “The atmospheric response to three decades of observed arctic sea ice loss”. In: *Journal of Climate* 26.4, pp. 1230–1248. DOI: 10.1175/JCLI-D-12-00063.1.
- Seneviratne, S. I. et al. (2010). “Earth-Science Reviews Investigating soil moisture – climate interactions in a changing climate : A review”. In: *Earth Science Reviews* 99.3-4, pp. 125–161. DOI: 10.1016/j.earscirev.2010.02.004.
- Seo, E. et al. (2019). “Impact of soil moisture initialization on boreal summer subseasonal forecasts: mid-latitude surface air temperature and heat wave events”. In: *Climate Dynamics* 52.3-4, pp. 1695–1709. DOI: 10.1007/s00382-018-4221-4.
- Shaw, T. A. and A. Voigt (2015). “Tug of war on summertime circulation between radiative forcing and sea surface warming”. In: *Nature Geoscience* 8.7, pp. 560–566. DOI: 10.1038/ngeo2449.
- Shaw, T. A. et al. (2016). “Storm track processes and the opposing influences of climate change”. In: *Nature Geoscience* 9.9, pp. 656–664. DOI: 10.1038/ngeo2783.
- Shepherd, T. G. (2014). “Atmospheric circulation as a source of uncertainty in climate change projections”. In: *Nature Geoscience* 7.10, pp. 703–708. DOI: 10.1038/ngeo2253.

- Sigmond, M., P. J. Kushner, and J. F. Scinocca (2007). “Discriminating robust and non-robust atmospheric circulation responses to global warming”. In: *Journal of Geophysical Research Atmospheres* 112.20, pp. 1–13. DOI: 10.1029/2006JD008270.
- Sillmann, J. et al. (2017). “Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities”. In: *Weather and Climate Extremes* 18.November, pp. 65–74. DOI: 10.1016/j.wace.2017.10.003.
- Simpson, I. R., T. A. Shaw, and R. Seager (2014). “A Diagnosis of the Seasonally and Longitudinally Varying Midlatitude Circulation Response to Global Warming”. In: *Journal of the Atmospheric Sciences* 71.7, pp. 2489–2515. DOI: 10.1175/JAS-D-13-0325.1.
- Simpson, I. R. et al. (2018). “Modeled and observed multidecadal variability in the North Atlantic jet stream and its connection to sea surface temperatures”. In: *Journal of Climate* 31.20, pp. 8313–8338. DOI: 10.1175/JCLI-D-18-0168.1.
- Simpson, I. R. et al. (2019). “Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection”. In: *Nature Geoscience* 12.8, pp. 613–619. DOI: 10.1038/s41561-019-0391-x.
- Smith, D. M. et al. (2019). “Robust skill of decadal climate predictions”. In: *npj Climate and Atmospheric Science* 2.1, p. 13. DOI: 10.1038/s41612-019-0071-y.
- Smith, D. M. et al. (2020). “North Atlantic climate far more predictable than models imply”. In: *Nature* 583.7818, pp. 796–800. DOI: 10.1038/s41586-020-2525-0.
- Smith, D. M. et al. (2022). “Robust but weak winter atmospheric circulation response to future Arctic sea ice loss”. In: *Nature Communications* 13.1, p. 727. DOI: 10.1038/s41467-022-28283-y.
- Sousa, P. M. et al. (2018). “European temperature responses to blocking and ridge regional patterns”. In: *Climate Dynamics* 50.1-2, pp. 457–477. DOI: 10.1007/s00382-017-3620-2.
- Sprites, P., C. Glymour, and R. Scheines (2001). *Causation, Prediction, and Search*. 2nd ed., p. 568.
- Steptoe, H., S. E. Jones, and H. Fox (2018). “Correlations Between Extreme Atmospheric Hazards and Global Teleconnections: Implications for Multihazard Resilience”. In: *Reviews of Geophysics* 56.1, pp. 50–78. DOI: 10.1002/2017RG000567.
- Stone, B. et al. (2021). “Compound Climate and Infrastructure Events: How Electrical Grid Failure Alters Heat Wave Risk”. In: *Environmental Science & Technology* 55.10, pp. 6957–6964. DOI: 10.1021/acs.est.1c00024.
- Strazzo, S. et al. (2019). “Application of a hybrid statistical-dynamical system to seasonal prediction of north american temperature and precipitation”. In: *Monthly Weather Review* 147.2, pp. 607–625. DOI: 10.1175/MWR-D-18-0156.1.
- Switanek, M. B. et al. (2020). “Present and Past Sea Surface Temperatures: A Recipe for Better Seasonal Climate Forecasts”. In: *Weather and Forecasting* 35.4, pp. 1221–1234. DOI: 10.1175/WAF-D-19-0241.1.
- Szegedy, C. et al. (2013). “Intriguing properties of neural networks”. In: *arXiv*, pp. 1–10. arXiv: 1312.6199.
- Székely, E., D. Giannakis, and A. J. Majda (2016). “Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data”. In: *Climate Dynamics* 46.5-6, pp. 1473–1502. DOI: 10.1007/s00382-015-2658-2.

- Templeton, S. R. et al. (2018). “Farmer interest in and uses of climate forecasts for Florida and the Carolinas: Conditional perspectives of extension personnel”. In: *Weather, Climate, and Society* 10.1, pp. 103–120. DOI: 10.1175/WCAS-D-16-0057.1.
- Teng, H. and G. Branstator (2019). “Amplification of Waveguide Teleconnections in the Boreal Summer”. In: *Current Climate Change Reports*, pp. 1–12.
- Teng, H. et al. (2013). “Probability of US heat waves affected by a subseasonal planetary wave pattern”. In: *Nature Geoscience* 6.12, pp. 1056–1061. DOI: 10.1038/ngeo1988.
- Teng, H. et al. (2019). “Circumglobal Response to Prescribed Soil Moisture over North America”. In: *Journal of Climate* 32.14, pp. 4525–4545. DOI: 10.1175/JCLI-D-18-0823.1.
- Thompson, D. W., M. P. Baldwin, and J. M. Wallace (2002). “Stratospheric connection to Northern Hemisphere wintertime weather: Implications for prediction”. In: *Journal of Climate* 15.12, pp. 1421–1428. DOI: 10.1175/1520-0442(2002)015<1421:SCTNHW>2.0.CO;2.
- Thomson, S. I. and G. K. Vallis (2018). “Atmospheric response to SST anomalies. Part II: Background-state dependence, teleconnections, and local effects in summer”. In: *Journal of the Atmospheric Sciences* 75.12, pp. 4125–4138. DOI: 10.1175/JAS-D-17-0298.1.
- Toms, B. A. et al. (2019). “Testing the Reliability of Interpretable Neural Networks in Geoscience Using the Madden-Julian Oscillation”. In: August, pp. 1–22. arXiv: 1902.04621.
- Torreggiani, S. et al. (2018). “Identifying the community structure of the food trade international multi-network”. In: *Environ. Res. Lett.* 13.054026.
- Totz, S. et al. (2017). “Winter precipitation forecast in the European and Mediterranean regions using cluster analysis”. In: *Geoph. Res. Lett.* DOI: 10.1002/2017GL075674.
- Trenberth, K. E. and J. T. Fasullo (2013). “An apparent hiatus in global warming?” In: *Earth’s Future* 1.1, pp. 19–32. DOI: 10.1002/2013EF000165.
- Trocchi, A. (2018). *Weather & Climate Services for the Energy Industry*. Ed. by A. Trocchi. Cham: Springer International Publishing, pp. 1–197. DOI: 10.1007/978-3-319-68418-5.
- Tsartsali, E. E. et al. (2022). “Impact of resolution on the atmosphere–ocean coupling along the Gulf Stream in global high resolution models”. In: *Climate Dynamics* 58.11, pp. 3317–3333. DOI: 10.1007/s00382-021-06098-9.
- Van Oldenborgh, G. J. (2020). *KNMI Climate Explorer*.
- Van Straaten, C. et al. (2020). “The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures”. In: *Quarterly Journal of the Royal Meteorological Society* 146.731, pp. 2654–2670. DOI: 10.1002/qj.3810.
- (2022). “Using Explainable Machine Learning Forecasts to Discover Subseasonal Drivers of High Summer Temperatures in Western and Central Europe”. In: *Monthly Weather Review* 150.5, pp. 1115–1134. DOI: 10.1175/mwr-d-21-0201.1.
- Van de Burgwal, L. H. M., A. Dias, and E. Claassen (2019). “Incentives for knowledge valorisation: a European benchmark”. In: *The Journal of Technology Transfer* 44.1, pp. 1–20. DOI: 10.1007/s10961-017-9594-8.
- Van de Burgwal, L., M. Van der Waal, and E. Claassen (2018). *Leveraging academic knowledge in the innovation ecosystem*, p. 72.

- Van der Wiel, K. et al. (2019). “Added Value of Large Ensemble Simulations for Assessing Extreme River Discharge in a 2 degree C Warmer World”. In: *Geophysical Research Letters* 46.4, pp. 2093–2102. DOI: 10.1029/2019GL081967.
- Van Der Linden, E. C., R. J. Haarsma, and G. Van Der Schrier (2019). “Impact of climate model resolution on soil moisture projections in central-western Europe”. In: *Hydrology and Earth System Sciences* 23.1, pp. 191–206. DOI: 10.5194/hess-23-191-2019.
- Van Der Wiel, K. et al. (2019). “The influence of weather regimes on European renewable energy production and demand”. In: *Environmental Research Letters* 14.9. DOI: 10.1088/1748-9326/ab38d3.
- Vannitsem, S. and M. Ghil (2017). “Evidence of coupling in ocean – atmosphere dynamics over the North Atlantic”. In: pp. 2016–2026. DOI: 10.1002/2016GL072229.
- Varoquaux, G. et al. (2015). “Scikit-learn”. In: *GetMobile: Mobile Computing and Communications* 19.1, pp. 29–33. DOI: 10.1145/2786984.2786995.
- Vigo, I. et al. (2019). *S2S4E D2.1 - User needs and decision-making processes that can benefit from S2S forecasts*. Tech. rep., p. 94.
- Vijverberg, S. (2020). *AMS-MWR-2020*. DOI: <https://doi.org/10.5281/zenodo.3856422>.
- Vijverberg, S., D. Coumou, and M. Schmeits (2021). “Paper of note: Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events”. In: *Bulletin of the American Meteorological Society* 3.March, pp. 189–210. DOI: https://doi.org/10.1175/BAMS_1023_189-210_Nowcast.
- Vijverberg, S. and D. Coumou (2022). “The role of the Pacific Decadal Oscillation and ocean-atmosphere interactions in driving US temperature predictability”. In: *npj Climate and Atmospheric Science* 5.1, p. 18. DOI: 10.1038/s41612-022-00237-7.
- Vijverberg, S., R. Hamed, and D. Coumou (2022a). “Skilful US Soy-yield forecasts at pre-sowing lead-times”. In: *Artificial Intelligence for the Earth Systems* in review.
- (2022b). “Skillful US Soy-yield Forecasts at Pre-sowing Lead-times”. In: *American Meteorological Society: Artificial Intelligence for the Earth Systems* Early Online Release, pp. 1–44. DOI: <https://doi.org/10.1175/AIES-D-21-0009.1>.
- Vijverberg, S. et al. (2020). “Subseasonal Statistical Forecasts of Eastern U.S. Hot Temperature Events”. In: *Monthly Weather Review* 148.12, pp. 4799–4822. DOI: 10.1175/MWR-D-19-0409.1.
- Villani, G. et al. (2021). “The iCOLT climate service: Seasonal predictions of irrigation for Emilia-Romagna, Italy”. In: *Meteorological Applications* 28.4, pp. 1–16. DOI: 10.1002/met.2007.
- Vitart, F. et al. (2017). “The subseasonal to seasonal (S2S) prediction project database”. In: *Bulletin of the American Meteorological Society* 98.1, pp. 163–173. DOI: 10.1175/BAMS-D-16-0017.1.
- Vitart, F. and A. W. Robertson (2018a). “The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events”. In: *npj Climate and Atmospheric Science* 1.1, p. 3. DOI: 10.1038/s41612-018-0013-0.
- Vitart, F. and A. Robertson (2018b). *Sub-seasonal to seasonal prediction - the gap between weather and climate forecasting*. 1st ed. Elsevier.
- Vitart, F. et al. (2019). “Sub-seasonal to Seasonal Prediction of Weather Extremes”. In: *Sub-Seasonal to Seasonal Prediction*, pp. 365–386. DOI: 10.1016/b978-0-12-811714-9.00017-6.

- Vogel, M. M., J. Zscheischler, and S. I. Seneviratne (2018). “Varying soil moisture-atmosphere feedbacks explain divergent temperature extremes and precipitation projections in Central Europe”. In: *Earth System Dynamics Discussions*, pp. 1–29. DOI: 10.5194/esd-2018-24.
- Walsh, J. E. (2005). “Sea Ice and Climate”. In: *Encyclopedia of World Climatology*. Springer Netherlands, pp. 639–641. DOI: 10.1007/1-4020-3266-8_178.
- Wang, H. et al. (2010). “The physical mechanisms by which the leading patterns of SST variability impact U.S. precipitation”. In: *Journal of Climate* 23.7, pp. 1815–1836. DOI: 10.1175/2009JCLI3188.1.
- Wang, H. et al. (2014). “On the role of SST forcing in the 2011 and 2012 extreme U.S. heat and drought: A study in contrasts”. In: *Journal of Hydrometeorology* 15.3, pp. 1255–1273. DOI: 10.1175/JHM-D-13-069.1.
- Wang, S. et al. (2015). “An intensified seasonal transition in the Central U.S. that enhances summer drought”. In: *Journal of Geophysical Research Atmospheres* 120, pp. 8804–8816. DOI: 10.1002/2014JD023013. Received.
- Watanabe, M. et al. (2020). “Enhanced future warming constrained by past trends in the equatorial Pacific sea surface temperature gradient”. In: *Nat. Clim. Chang.* 11. January, pp. 3–9. DOI: 10.1038/s41558-020-00933-3.
- Wehrli, K. et al. (2021). “The ExtremeX global climate model experiment: Investigating thermodynamic and dynamic processes contributing to weather and climate extremes”. In: *Earth System Dynamics Discussions* July, pp. 1–31.
- Weingärtner, L. and E. Wilkinson (2019). “Anticipatory Crisis Financing and Action: Concepts, Initiatives, and Evidence”. In: June, pp. 1–18.
- White, C. J. et al. (2021). “Advances in the application and utility of subseasonal-to-seasonal predictions”. In: *Bulletin of the American Meteorological Society*, pp. 1–57. DOI: 10.1175/BAMS-D-20-0224.1.
- Wikipedia (2021). *British Columbia Wildfires*.
- Wilks, D. S. (2006). “On “field significance” and the false discovery rate”. In: *Journal of Applied Meteorology and Climatology* 45.9, pp. 1181–1189. DOI: 10.1175/JAM2404.1.
- Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Ed. by R Dmowska, D Hartmann, and R. H. Thoman. 3rd ed. 100. Oxford: Elsevier, pp. 2–699.
- Wilks, Dan, S. (2016). “the Stippling Shows Statistically Significant Grid Points”. In: 97. December, pp. 2263–2274.
- Williams, N., A. Scaife, and J. Screen (2022). “Weak ENSO teleconnections contribute to the signal-to-noise paradox”. In: p. 5194.
- Winter, J. M. et al. (2015). “Uncertainty in modeled and observed climate change impacts on American Midwest hydrology”. In: *Water Resources Research* 51.5, pp. 3635–3646. DOI: 10.1002/2014WR016056.
- Wirth, V. et al. (2018). “Rossby Wave Packets on the Midlatitude Waveguide—A Review”. In: *Monthly Weather Review* 146.7, pp. 1965–2001. DOI: 10.1175/mwr-d-16-0483.1.
- WMO (2006). *WMO - Standardized Verification System (SVS) for Long Range Forecasts*. — (2017). *Annual Report: Climate Risk & Early Warning Systems*. Tech. rep. Climate Risk and Early Warning Systems, p. 23.
- WMO, W. M. O. (2020). *Guidance on Operational Practices for Objective Seasonal Forecasting 2020*. 1246, p. 106.

- WMO, W. M. O. (2021). *Weather-related disasters increase over past 50 years, causing more damage but fewer deaths*.
- Wolf, G. et al. (2020). “Connection between sea surface anomalies and atmospheric quasi-stationary waves”. In: *Journal of Climate* 33.1, pp. 201–212. DOI: 10.1175/JCLI-D-18-0751.1.
- Wolf, G. et al. (2018). “Quasi-stationary waves and their impact on European weather and extreme events”. In: *Quarterly Journal of the Royal Meteorological Society* 144.717, pp. 2431–2448. DOI: 10.1002/qj.3310.
- Woollings, T. (2010). “Dynamical influences on European climate: an uncertain future”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368.1924, pp. 3733–3756. DOI: 10.1098/rsta.2010.0040.
- Xie, S. P. et al. (2015). “Towards predictive understanding of regional climate change”. In: *Nature Climate Change* 5.10, pp. 921–930. DOI: 10.1038/nclimate2689.
- Yu, B. and F. W. Zwiers (2007). “The impact of combined ENSO and PDO on the PNA climate: A 1,000-year climate modeling study”. In: *Climate Dynamics* 29.7-8, pp. 837–851. DOI: 10.1007/s00382-007-0267-4.
- Yu, L. (2019). “Global Air–Sea Fluxes of Heat, Fresh Water, and Momentum: Energy Budget Closure and Unanswered Questions”. In: *Annual Review of Marine Science* 11.1, pp. 227–248. DOI: 10.1146/annurev-marine-010816-060704.
- Yuan, S. et al. (2019). “Prediction of north atlantic oscillation index with convolutional LSTM based on ensemble empirical mode decomposition”. In: *Atmosphere* 10.5. DOI: 10.3390/atmos10050252.
- Yuval, J. and Y. Kaspi (2020). “Eddy Activity Response to Global Warming–Like Temperature Changes”. In: *Journal of Climate* 33.4, pp. 1381–1404. DOI: 10.1175/JCLI-D-19-0190.1.
- Zaplotnik, Ž., M. Pikovnik, and L. Boljka (2022). “Recent Hadley circulation strengthening: a trend or multidecadal variability?” In: *Journal of Climate*, pp. 1–65. DOI: 10.1175/JCLI-D-21-0204.1.
- Zappa, G. et al. (2013). “A multimodel assessment of future projections of north atlantic and european extratropical cyclones in the CMIP5 climate models”. In: *Journal of Climate* 26.16, pp. 5846–5862. DOI: 10.1175/JCLI-D-12-00573.1.
- Zhang, L. and T. L. Delworth (2015). “Analysis of the characteristics and mechanisms of the pacific decadal oscillation in a suite of coupled models from the geophysical fluid dynamics laboratory”. In: *Journal of Climate* 28.19, pp. 7678–7701. DOI: 10.1175/JCLI-D-14-00647.1.
- Zhou, G. (2019). “Atmospheric Response to Sea Surface Temperature Anomalies in the Mid-latitude Oceans: A Brief Review”. In: *Atmosphere-Ocean* 57.5, pp. 319–328. DOI: 10.1080/07055900.2019.1702499.
- Zhou, G. et al. (2017). “State dependence of atmospheric response to extratropical North Pacific SST anomalies”. In: *Journal of Climate* 30.2, pp. 509–525. DOI: 10.1175/JCLI-D-15-0672.1.
- Zscheischler, J. and S. I. Seneviratne (2017). “Dependence of drivers affects risks associated with compound events”. In: *Science Advances* 3.6, pp. 1–11. DOI: 10.1126/sciadv.1700263.

Acknowledgements

Yes. I wrote a Phd thesis, and now I get the opportunity to thank all who helped shape my life positively and helped me with this kind of a weird process. Starting with trying to become a car mechanic (not a very talented one), I discovered at a late age that the world becomes so much more fascinating when you understand it slightly better. The enthusiasm of teachers can play quite a big role in this regard. René Enthoven is one those teachers. I regret that I can no longer visit him and tell him personally, but in terms of sparkling my curiosity and support me in taking chances, I want to thank him first.

Considering that without my friends, I would have had the spirit and energy of a lonely forgotten VGA cable in a dusty attic, I should definitely thank them! Starting with Barry, taoistische boswachter, thanks for all the 'pret', which was there in abundance on the many adventures. Ending up at sketchy parties with cheap beer and cold sausages with ketchup. Robin, so nice I met you there. If anyone ever needs a laugh, I recommend you call him (0620230303). Sander, beautiful person. Thanks for connecting in your very personal way. Annieee, thanks for your warm kindness and lovely vibe. Merel, part academic, part artist, part weird, please keep sharing your excitement and views. Hans, if I think of you I think of experiencing the quality of life. Jilles, off course they chew on their own eyes. Yannick, nice energy bro, you're gonna be a great dad!

Nicky L, bro's for life. Love that you're there on a weekly basis. Abel, please keep whistling through life, you're good at that. Duuu Sanne, I don't know when, but I would like to be in charge of the cash money again when we go out with the crew. Dennis, versatile guy, tackling tax evasion with a bucket, sometimes an automatic rifle (from CoD), and a data science master. Vinnie, although you sometimes make me feel otherwise, I actually do believe the nuclear safety of the Netherlands is in good hands.

Bert! Loved all the talks we had, music we made, (and off course the sailing trips)! Joost & Eef, you are great people, full of nice (mostly random) surprises! Moes and Kai, you both have evil humor, but somehow we appreciate tolerate it. Eke, loved all the sailing days, we should do that again soon! André, thanks for sharing your happy face, thoughts, and nice vibes.

Raed, I really enjoy your energy, you have mastered how to be super chill and funny at the same time. Tim, really love all the chats we have, which mostly involve laughing out loud because you simply tend to have that effect on people. Still happy to have survived the back ride to your home, was a great night. Timothy, such a social guy bringing in the good vibes, sorry to have missed your phd party man, still regret it on a hourly basis. Henrique, awesome to have met you! You're an endless source of energy and good stories. Anaïs, thanks for all the chats and being that infinite source of positive feedback. Marleen,

you kind of amaze me. A perfect mix of being professional and highly productive, as well as in control, cheerful, and in for a nice chat. Tamara & Chiem, loved the vibes we shared at EGU and Trieste! Kai, Giorgia, Jascha, Marlene, thanks for the warm welcome into Dim's group back in the Potsdam-to-Amsterdam transitioning days. Max, Pedja, Fei, Ileen, Judith, Alessia, Jens, Tadzio, Teun, Sanne, Hannah, Liselotte, Sophie, Eric, Lotte, Hans, Elco, Paolo, Hans, Rhoda, Tristian, Sem D., you are ALL wonderful people! Sorry if I'm still forgetting people. Thanks for making the phd-life sooooo much more fun! I would like to spent a sentence on my supervisor, Dim Coumou, it was truly fun to be supervised by you. You're a chill and kind person, which seeps through in your supervision style. Like when you decided to give me and Giorgia a specialized writing course, and the fact that your almost always approachable within days if not hours. Maurice, also thanks for all your time and valuable feedback, you were of great value for my progress. Jannes, thanks for all your effort, positive vibes, and for joining me on the exciting 'beyond weather' tour! Thijs Olsthoorn, so many good times to look back to. Thanks for all. Chris, always fun to talk about technical stuff. Thijs D, always happy to be involuntarily taught the difference between a Chamaecyparis and a Amelanchier. Merel, you have great humor, I'm jealous of your new English colleagues.

Moeders, bedankt voor al je interesse en steun, ondanks dat je weinig tot geen benul hebt van wat ik nou echt doe. Hoevaak jij hebt gezegd bij iedere iteratie "is je paper nu al klaar?" (terwijl het nog maanden werk zou zijn). Pap, bedankt voor al je harde werk en liefde, het is zo spijtig dat je er niet bij kan zijn. Sharon, je bent een lieve zus, fijn om te zien dat je steeds beter je weg aan vinden bent! Nicky, super fijn om jou als zusje te hebben! Janneke, je bent en zal altijd een geweldig persoon blijven met een speciaal plekje in m'n hart, thanks voor alles!

