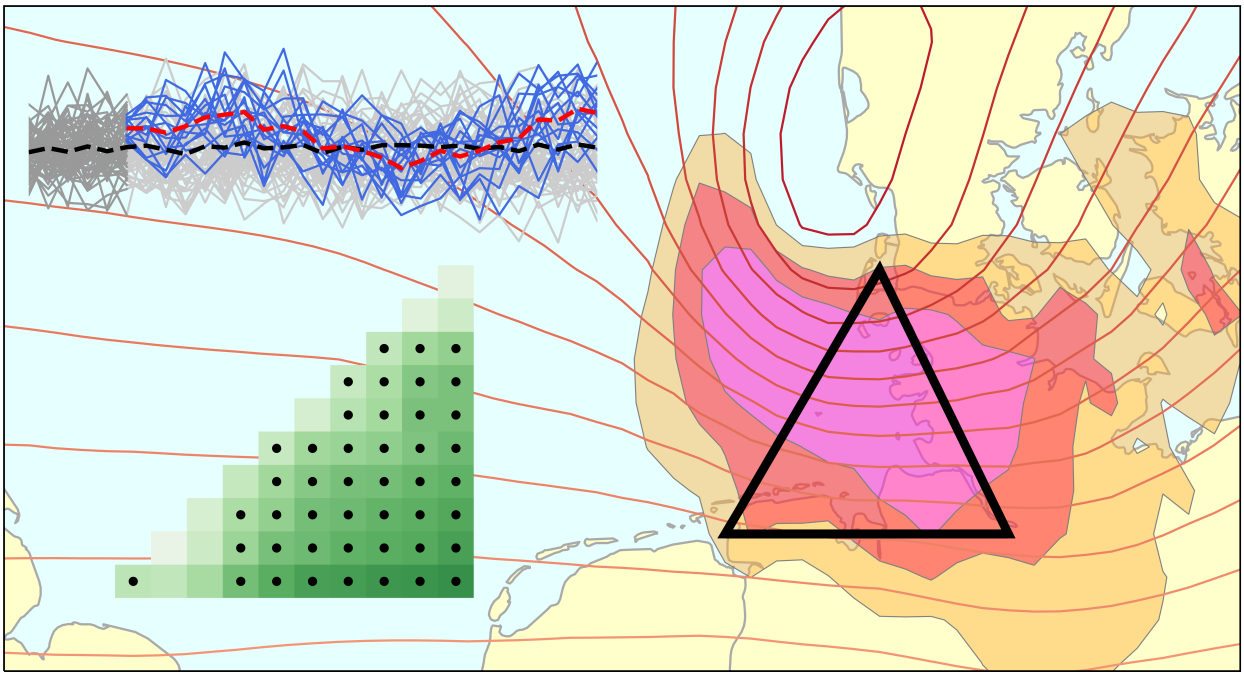
MAX-PLANCK-INSTITUT
FÜR METEOROLOGIE

Predictability of German Bight Storm Activity



Daniel Ulrich Ludwig Krieger

Hamburg 2023

Berichte zur Erdsystemforschung
Reports on Earth System Science

275
2023

Hinweis

Die Berichte zur Erdsystemforschung werden vom Max-Planck-Institut für Meteorologie in Hamburg in unregelmäßiger Abfolge herausgegeben.

Sie enthalten wissenschaftliche und technische Beiträge, inklusive Dissertationen.

Die Beiträge geben nicht notwendigerweise die Auffassung des Instituts wieder.

Die "Berichte zur Erdsystemforschung" führen die vorherigen Reihen "Reports" und "Examensarbeiten" weiter.

Anschrift / Address

Max-Planck-Institut für Meteorologie
Bundesstrasse 53
20146 Hamburg
Deutschland

Tel./Phone: +49 (0)40 4 11 73 - 0
Fax: +49 (0)40 4 11 73 - 298

name.surname@mpimet.mpg.de
www.mpimet.mpg.de

Notice

The Reports on Earth System Science are published by the Max Planck Institute for Meteorology in Hamburg. They appear in irregular intervals.

They contain scientific and technical contributions, including PhD theses.

The Reports do not necessarily reflect the opinion of the Institute.

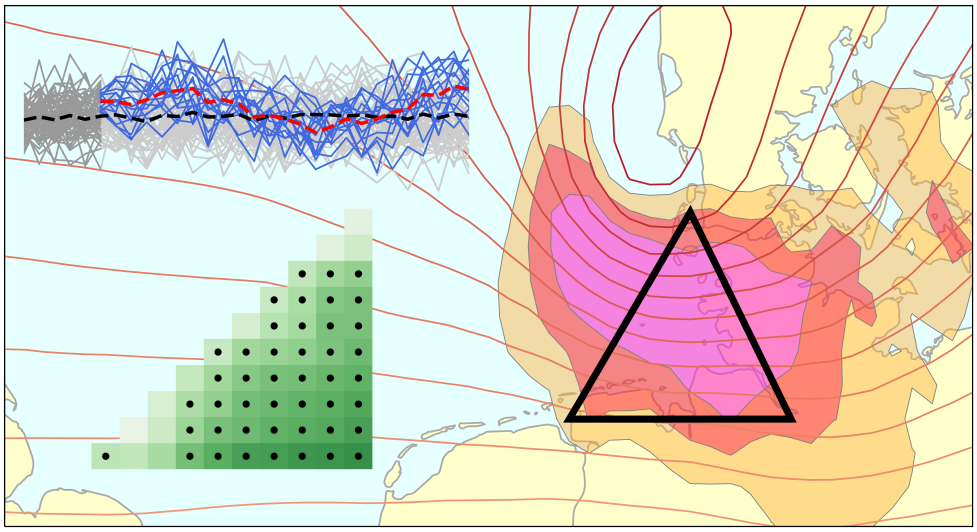
The "Reports on Earth System Science" continue the former "Reports" and "Examensarbeiten" of the Max Planck Institute.

Layout

Bettina Diallo and Norbert P. Noreiks
Communication

Copyright

Predictability of German Bight Storm Activity



Daniel Ulrich Ludwig Krieger

Hamburg 2023

Daniel Ulrich Ludwig Krieger

aus Neuburg an der Donau, Deutschland

Helmholtz-Zentrum Hereon
Institut für Küstensysteme – Analyse und Modellierung
Max-Planck-Straße 1
21502 Geesthacht

Max-Planck-Institut für Meteorologie
The International Max Planck Research School on Earth System Modelling
(IMPRS-ESM)
Bundesstrasse 53
20146 Hamburg

Universität Hamburg
Fachbereich Erdsystemwissenschaften
Institut für Meereskunde
Bundesstraße 53
20146 Hamburg

Tag der Disputation: 28. November 2023

Folgende Gutachter empfehlen die Annahme der Dissertation:
Dr. Ralf Weisse
Prof. Dr. Johanna Baehr

Vorsitzender des Promotionsausschusses:
Prof. Dr. Hermann Held

Dekan der MIN-Fakultät:
Prof. Dr.-Ing. Norbert Ritter

Titelgrafik: vom Autor erstellt

ISSN 1614-1199 - doi: 10.17617/2.3557092

Life is like a storm;
it may sweep you into darkness.
But in its rain you will find
what lets you blossom brightly
when the sun breaks through.

— ChatGPT

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The template is openly available at: https://bitbucket.org/amiede/classicthesis/.

ABSTRACT

Extreme events like storms and storm surges are an everlasting challenge for populated coastal areas such as the German Bight. These events pose a threat to critical infrastructure and property, and require constant planning, adaptation, and precautionary measures, which oftentimes take multiple years to be implemented. Therefore, coastal protection and regional stakeholders in the German Bight may greatly benefit from seasonal-to-decadal predictions of coastal hazards like storm and surge activity. Historical records of German Bight storm activity show a pronounced multidecadal variability, but no significant trend. The historical evolution of storm surges at the coast is mainly characterized by changes in the mean sea level and coastal engineering measures like damming and dredging. Apart from the further projected rise in sea level, however, current climate projections suggest low confidence in the response of regional storm and storm surge activity to global warming. Hence, there appears to be potential in initialized decadal prediction systems to provide forecasts of the local storm and surge climate on a seasonal-to-decadal scale.

In this thesis, I thus investigate the capabilities of a large-ensemble decadal prediction system based on the MPI-ESM-LR climate model to predict these climate extremes. In the first part, I evaluate the skill of the model for German Bight storm activity and winter mean sea-level pressure anomalies and find that the model is most skillful in predicting long averaging periods of more than five years. For shorter periods, such as the upcoming year, the model shows little to no forecast skill.

Subsequently, motivated by the lack of skill for shorter forecast lead times, I draw on physical predictors of winter storm activity and an already established ensemble subselection technique to prove that seasonal predictions of German Bight storm activity can be significantly improved. I also illustrate how this skill improvement is associated with a better representation of the large-scale circulation in the model.

Lastly, I build on the findings of the first part and show to what extent the skill for storm activity is utilizable for surge predictions. I introduce two approaches to derive surge statistics from storm-related parameters, since the model does not explicitly predict water levels. I demonstrate that these approaches provide a fair conversion from storms to surges, but the resulting prediction skill of the decadal hindcast system for surges is remarkably lower than for storm activity.

Overall, I provide in this thesis an overview of the limits and capabilities of a large-ensemble decadal prediction system in predicting the German Bight storm and storm surge climate on timescales ranging from several months up to ten years.

ZUSAMMENFASSUNG

Extremereignisse wie Stürme und Sturmfluten stellen eine fortwährende Herausforderung für besiedelte Küstengebiete wie die Deutsche Bucht dar. Diese Ereignisse bedrohen kritische Infrastruktur und Eigentum und erfordern ständige Planungs-, Anpassungs- und Vorsichtsmaßnahmen, welche oft mehrere Jahre in Anspruch nehmen. Aus diesem Grund könnten der Küstenschutz und lokale Akteure in der Deutschen Bucht erheblich von saisonalen bis hin zu dekadischen Vorhersagen von Küstengefahren wie Sturm- oder Sturmflutaktivität profitieren. Historische Beobachtungen von Sturmaktivität in der Deutschen Bucht zeigen ausgeprägte multidekadische Schwankungen, jedoch keinen signifikanten Trend. Die vergangene Entwicklung von küstennahen Sturmfluten ist überwiegend durch den Anstieg des mittleren Meeresspiegels und Küstenbaumaßnahmen wie Eindeichungen und Fahrrinnenvertiefungen geprägt. Abgesehen vom prognostizierten weiteren Anstieg des Meeresspiegels legen jedoch aktuelle Klimaprojektionen nahe, dass der Einfluss der globalen Erwärmung auf die regionale Sturm- und Sturmflutaktivität noch sehr unsicherheitsbehaftet ist. Deshalb sehen wir in initialisierten dekadischen Vorhersagesystemen das Potential, akkuratere Prognosen des lokalen Sturm- und Sturmflutklimas auf einer saisonalen bis dekadischen Zeitskala zu generieren.

In dieser Arbeit untersuche ich daher die Vorhersagegüte eines dekadischen Ensemble-Vorhersagesystems basierend auf dem Klimamodell MPI-ESM-LR für Vorhersagen von Sturm- und Sturmflutaktivität in der Deutschen Bucht. Im ersten Teil der Arbeit analysiere ich die Vorhersagegüte des Modells für die Sturmaktivität in der Deutschen Bucht und für Anomalien des Luftdrucks auf Meereshöhe im Winter. Ich komme zu dem Schluss, dass die Vorhersagen des Modells am besten für gemittelte Vorhersagen von mehr als fünf Jahren sind. Für kürzere Zeiträume wie das kommende Jahr zeigt das Modell nur geringe bis keine Vorhersagefähigkeit.

Motiviert durch die Schwäche des Modells in kürzeren Vorhersagezeiträumen greife ich anschließend auf physikalische Prädiktoren der Wintersturmaktivität und eine bereits etablierte Ensemble-Auswahltechnik zurück, um zu demonstrieren, dass saisonale Vorhersagen der Sturmaktivität in der Deutschen Bucht erheblich verbessert werden können. Ich erläutere auch, wie diese Verbesserung der Vorhersagegüte mit einer besseren Darstellung der großskaligen Zirkulation im Modell zusammenhängt.

Zuletzt baue ich auf den Erkenntnissen des ersten Teils auf und zeige, inwieweit die Vorhersagegüte für Sturmaktivität auf das Sturmflutklima übertragbar ist. Ich stelle zwei Ansätze vor, um Sturmflutstatistiken aus Sturmgrößen abzuleiten, da das Modell den Wasserstand nicht explizit vorhersagen kann. Ich zeige, dass mit diesen Ansätzen eine genügend gute Übersetzung von Sturm- zu Sturmflutstatistiken möglich ist, jedoch die resultierende Vorhersagbarkeit von Sturmfluten nennenswert geringer ist als jene von Sturmaktivität.

Insgesamt gebe ich in dieser Arbeit einen Überblick über die Grenzen und Möglichkeiten eines dekadischen Ensemble-Vorhersagesystems in der Vorhersage des Sturm- und Sturmflutklimas der Deutschen Bucht auf Zeitskalen von mehreren Monaten bis hin zu zehn Jahren.

ACKNOWLEDGMENTS

In one way or another, so many more people contributed to the successful writing of this thesis than are listed on the first page. It is therefore a great matter of concern to me to say a few words of thanks, with the risk of not adequately mentioning everyone.

First and foremost, I would like to express my deepest appreciation to my supervisors, Ralf Weisse and Johanna Baehr for being the scientific guiding light over the past three years. Despite all the challenges thrown our way, the supervision was superb and always spot on. The guidance and expertise, but also the given freedom to explore beyond a strictly confined path made this PhD a very enjoyable project.

I sincerely thank Sebastian Brune for providing me with the vast majority of data used in the dissertation, for being a real-life encyclopedia on modeling, a very constructive co-author, and of course for the countless amusing on- and off-topic chats over a coffee. I also thank Patrick Pieper for his invaluable contributions to the first paper and always ensuring *statistical correctness*.

I greatly thank Wolfgang Müller for his office as chair of the advisory panel and six panel meetings, which were crucial to keeping the project on track and my focus on the essential things.

I am thankful for the opportunity to be part of the IMPRS-ESM – working, researching, and learning among so many other great young scientists from widely different fields. Many thanks go to Connie, Michaela, and especially Antje for always spreading such a warm-hearted and supportive atmosphere for us PhD students, and for permanently helping us keep our spirits up, especially as times get tough.

I want to extend my sincere thanks to the Coastal Climate working group at Hereon and the Climate Modelling working group at UHH, in particular Björn, Céline, Edu, Goratz, Hongdou, Julianna, Lara, and Sebastian. Sharing this welcoming and encouraging work environment with you has been a blast, and I've been able to broaden my horizon by learning so many new things about a variety of topics from you.

I further want to offer my gratitude to the WAKOS project for providing the funding for my PhD, and for embedding my work into a bigger picture. The project meetings were great opportunities to zoom out a bit of the – at times – very specific research of this dissertation. Especially the trip to Norderney was a very welcome excursion into my actual study region.

I thank and applaud Alexandra Elbakyan for her unwavering commitment to uphold one of the most fundamental principles of science and humanity: the free and open access to knowledge for all people.

Special thanks go to Ina, Julia, Kai, Lara, and Moritz for proofreading and sanity-checking the thesis at various stages. Your comments were invaluable and immensely helpful for improving the quality when I could not see the forest for the trees.

Last but not least, it is my wish to express my heartfelt gratitude to the people who helped me ride out the storms that came with this adventure.

To my family, you provided unconditional support throughout the entire studies, tons of helpful advice, and numerous solutions to so many problems along the way, which made focusing on the dissertation a lot easier. Thank you! ♡

To the *Harmonic Quartet*, Philipp, Tobi, and Tobi, you guys have been there since the very beginning of this meteorological journey. We shared so many unforgettable memories together, and I will always consider you my special family. Thank you! ♡

To the *Hamburg-Fans*, Finn, Henning, Julia, Kathi, Kai, Lara, Luigi, Moritz, and Tobi, I still cannot believe how lucky I am to be a part of this semester. The past six years in Hamburg, Denmark, and beyond have been an absolute blast with you, and I am more than happy to have you in my life. Thank you! ♡

Finally, to Jannik and Max, and to Lara, words fail to describe how much it means to me to have you as my closest friends. You have always been my motivation to keep going when I was running on empty. You probably might not even realize how often you have carried me through thick and thin over the past years. I am certain that I wouldn't be here without you, and I am forever grateful for that. I know these two words don't cut it, but thank you! ♡

PUBLICATIONS

The following two publications were prepared as part of this dissertation.

The publications are included in the appendix:

Appendix A

> **Krieger, D.**, Brune, S., Pieper, P., Weisse, R., and Baehr, J. (2022): Skillful decadal prediction of German Bight storm activity. *Natural Hazards and Earth System Sciences*, 22, 3993-4009, DOI: 10.5194/nhess-22-3993-2022

Appendix B

> **Krieger, D.**, Brune, S., Baehr, J., and Weisse, R. (2023): Improving seasonal predictions of German Bight storm activity. *EGUsphere [preprint]*, DOI: 10.5194/egusphere-2023-2676

In addition to the two first-author papers above, I further contributed to the following publication as a co-author:

> Olonscheck, D., Suarez-Gutierrez, L., Milinski, S., and 14 co-authors (incl. **Krieger, D.**) (2023): The new Max Planck Institute Grand Ensemble with CMIP6 forcing and high-frequency model output. *Journal of Advances in Modeling Earth Systems*, DOI: 10.1029/2023MS003790

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACC	Anomaly correlation coefficient
BSS	Brier skill score
CMIP	Coupled Model Intercomparison Project
DJF	December–February
DPS	Decadal prediction system
ERA5	European Centre for Medium-Range Weather Forecasts Reanalysis v5
GBSA	German Bight storm activity
LR	Low-resolution
MiKlip	Mittelfristige Klimaprognosen
ML	Machine learning
MPI-ESM	Max Planck Institute Earth System Model
MPI-ESM-LR	Max Planck Institute Earth System Model in low-resolution mode
MSLP	Mean sea-level pressure
NAO	North Atlantic Oscillation
QBO	Quasi-Biennial Oscillation
RMSE	Root-mean-square error
SIWE	Storm-integrated wind speed exceedance
SSI	Storm Severity Index

Unifying Essay

1

INTRODUCTION

1.1 Storms and storm surges – an everlasting challenge

Coastal regions around the globe are frequently affected by various types of natural hazards, some of which include weather extremes like *storms* and *storm surges* (Kron, 2013). Storms, i. e., local events of abnormally high wind speeds caused by an atmospheric pressure gradient between areas of low and high pressure, have the capability to damage infrastructure and disrupt public life. Storm surges, i. e., local events of abnormally high water levels mainly driven by strong winds, can additionally cause coastal inundation, alter submarine sediment structures, and impact the coastal ecosystem. Both types of coastal hazards are impactful to society, and thus skillful predictions of these events are of great value to the coastal protection and management sectors.

The German Bight (Fig. 1.1) and the adjacent coastline is no exception in this regard. With its low-lying coastal marshlands, polders, the Wadden Sea tidal mudflats, and numerous small islands that are part of the Frisian Island archipelago, this part of the North Sea offers several features that are vulnerable to storms and storm surges. Therefore, the history of this region is filled with significant and impactful storm events (e. g., Lamb and Frydendahl, 1991; Gönnert et al., 2001). While storms and surges still cause damage on a regular basis in present times, the technical possibilities to protect property and livestock against the unrelenting forces of nature were much more limited centuries ago. Thus, storms and surges were oftentimes accompanied by significant losses and permanent changes, not only to man-made structures and settlements, but also to the shape and orography of the coastline and its offshore islands (Fig. 1.2, Hoffmann, 2004).

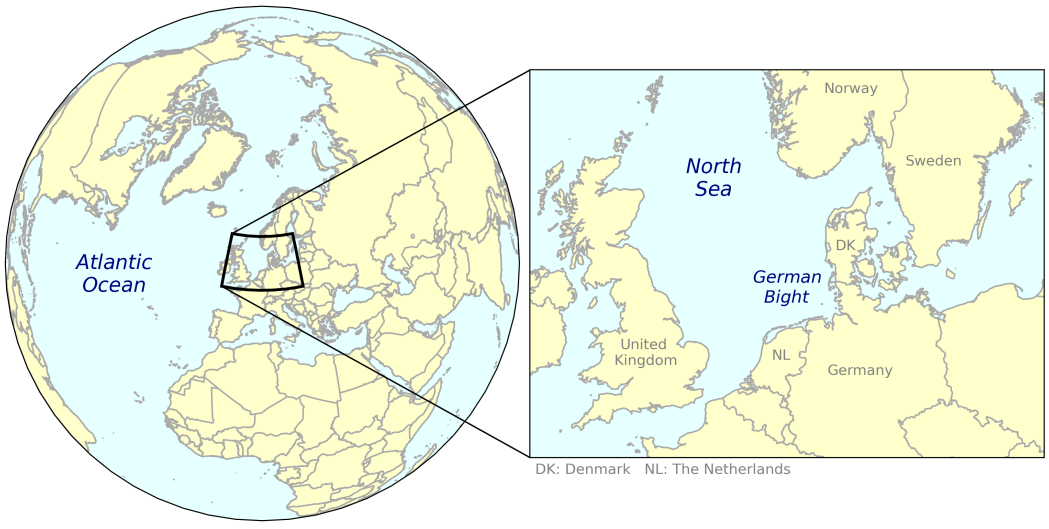


Figure 1.1: Locator map of the German Bight.

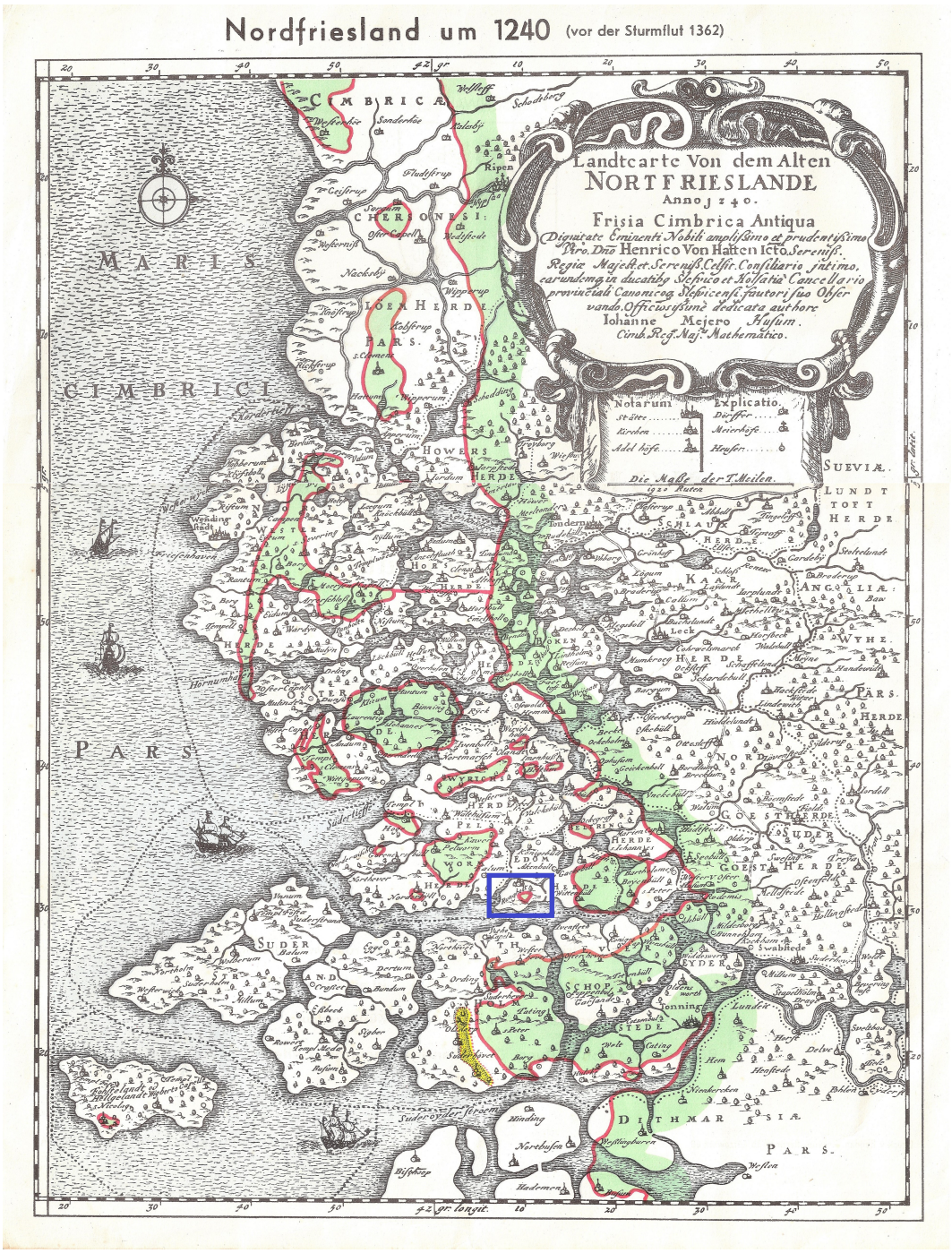


Figure 1.2: Reconstruction of the coastline and landscape of North Frisia around 1240. Originally published in Danckwerth and Mejer (1652) and reproduced in Strunk (1970). Red lines were added subsequently and mark the shape of the coastline around the time of reproduction, likely between 1970 and 1973. The blue rectangle indicates the location of Rungholt. Digital version obtained from Wikimedia Commons (2019).

One of the most prominent examples from historic times is the former island of *Strand* with its fabled harbor village *Rungholt*. The settlement was ravaged by multiple severe storms and storm surges, such as the *1. Grote Mandrenke* in 1362 (e. g., Hadler et al., 2018), before eventually being finally lost to the sea in 1634 after another severe storm surge that claimed the lives of thousands of humans and livestock (e. g., Heimreich,

1668; Kempe, 2006). A more recent and, arguably, the most infamous case of a devastating storm with far-reaching consequences is the windstorm *Vincinette*, which impacted the North Sea coast on February 16–17, 1962 (Huster, 1962). While not the most intense winter storm in terms of wind speed or central pressure, *Vincinette* caught many residents, agencies, and emergency managers off guard. Along the German Bight coast and far upstream several estuaries – especially the Elbe estuary – the nightly storm surge driven by *Vincinette*'s wind field led to failures of dikes and levees in multiple places (Fig. 1.3, Jochner et al., 2013). The large water masses subsequently devastated homes, eventually causing the death of 340 people, 315 of them in the city of Hamburg alone (de Guttry and Ratter, 2022). The timing of the storm and the associated surge, combined with overwhelmed crisis management, a lack of proper levee maintenance, and missing awareness for the existing threat – Germany experienced no levee breaches for more than 100 years prior to 1962 – are to blame for the perhaps worst natural disaster in German post-war history (de Guttry and Ratter, 2022).

Over time, the resilience of the local population has increased, and equipped with both new technical inventions and the vast experience handed down from previous generations, more advanced measures to protect the coastline were taken. Along the coast, dikes were reinforced and heightened, inland drainage systems were improved, and new buildings were constructed such that they would withstand stronger winds. Generally speaking, many extreme events were immediately followed by a response of the local population to improve protective measures against these types of hazards, also by learning from previous mistakes and underestimations (e. g., Siefert and Havnoe, 1988; von Storch and Woth, 2008; von Storch et al., 2008). In addition to the fortification of the coastline, many successful attempts were made to reclaim land once lost to the sea.

Figure 1.3: Flooded allotments in Hamburg-Wilhelmsburg during the 1962 storm surge (original German title: *Überflutete Kleingärten am Alten Bahnhof in Wilhelmsburg.*). © NDR, image from Edith Vasicek, licensed under CC BY-NC-ND 3.0 DE (https://creativecommons.org/licenses/by-nc-nd/3.0/de/). Obtained from NDR (2023).

Figure 1.4: Beach erosion on the East Frisian island of Langeoog, following a storm surge caused by the storm *Nadia* in February 2022 (original German title: *Das Bild zeigt abgetragene Sandflächen nach Sturm "Nadia" auf der ostfriesischen Insel Langeoog.*). © NDR, licensed under CC BY-NC-ND 3.0 DE (https://creativecommons.org/licenses/by-nc-nd/3.0/de/). Obtained from NDR (2022).

Nowadays, the state of coastal protection in the German Bight is impeccable compared to previous centuries and arguably one of the more sophisticated feats of coastal engineering in the world. Extreme events that would have caused catastrophic and irreversible damage a few hundred years ago are now merely a peak in statistical analyses. Still, certain elements of the German Bight coastline are still difficult to defend against storms and surges. For instance, dune damages and beach erosion are regularly reported after surge events, especially along the offshore islands of North and East Frisia (Fig. 1.4). Additionally, long periods of elevated water levels at the coastline caused by multiple consecutive storms can complicate the drainage of low-lying inland areas, leading to groundwater flooding and widespread inundation. Storm-related damages to these sensitive components of the coast usually prompt vast and expensive restoration efforts (Hanson et al., 2002; Post, 2005), which may be better planned and coordinated with the help of skillful storm activity predictions.

Furthermore, the size of the population and especially the amount and value of insured property along the coast has grown dramatically over the past century (e. g., Kron, 2013).The, in a historical context, quite recent emergence of renewable energy production, e. g., through offshore wind power plants, brings another term to the already complex equation of coastal extreme events and their impact on society. Not only does the offshore power generation heavily rely on accurate observations and predictions of the wind climate, but the sheer number of wind power plants already installed over the past few decades combined with vast areas approved for new wind farms also contribute to the insured capital at risk of suffering damages during extreme storm events (Buchana and McSharry, 2019). Consequently, studies are continuously spurred on evaluating past and modeling future storm-related insurance losses (e. g., Pinto et al., 2007; Schwierz et al., 2009; Donat et al., 2011; Gaslikova et al., 2011; Haylock, 2011; Karremann et al., 2014). As regards the accurate quantification of future socio-economic impacts, but also the estimation of

potential future renewable energy production, studies on skillful predictions of storm and storm surge activity are essential. To ensure cross-study intercomparability, these studies ideally entail consistent definitions of storms and storm surges.

1.2 What is a storm? What is a storm surge?

From a meteorological standpoint, the term *storm* describes sustained wind speeds that exceed a pre-determined threshold for a certain amount of time. The exact values of these thresholds differ between meteorological agencies and regions on Earth. The most common definitions rely on the Beaufort scale and speak of storms when the 10-minute-sustained wind speed 10 meters above ground reaches a Beaufort number of 8 (17.2 m/s) or 9 (20.8 m/s). The German national weather service (*Deutscher Wetterdienst*, DWD) refers to low-pressure systems as storm cyclones or storm lows once a Beaufort number of 8 can be observed or theoretically derived from the pressure field (DWD, 2023). The DWD, however, uses the term storm itself only for sustained winds of at least 9 Beaufort.

A *storm surge* is defined as a short-term extreme water level along the coast driven by strong surface winds. Such strong winds are usually found in storms, hence the term *storm* surge. The water level that is required for a high-water event to be classified as a storm surge again differs between hydrographic agencies around the globe, and is mostly based on the natural variability of the water level at the location. In regions with a dominant tidal cycle and thus a large naturally occurring amplitude of water levels, the threshold is usually a lot higher than in geographical areas with little or no tidal cycle. In the German Bight and the connected estuaries, the wind-induced component of storm surges can amount to several meters, which is on the same order of magnitude as the tidal cycle itself (Huthnance, 1991). The German federal hydrographic agency (*Bundesamt für Seeschifffahrt und Hydrographie*, BSH) therefore defines high water levels at the German Bight coastline as storm surges, severe storm surges, and very severe storm surges, once they rise 1.5, 2.5, and 3.5 meters above the mean tidal high water, respectively (BSH, 2023). A drawback of this definition is that places with a larger initial tidal amplitude, such as the Elbe estuary, will reach the storm surge threshold more often than locations in the open sea (e. g., Heligoland). Thus, on paper, it might appear as if offshore islands are less frequently impacted by storm surges, whereas, in reality, the severity of a surge for these islands already starts at a lower water level than, for instance, for Hamburg. An alternative approach which include these regional differences is to define storm surges via location-specific percentiles or return levels, that is, the surge height which is – on average – expected to be exceeded once within a certain period of time. This method is currently being used operationally in the Netherlands (2-year return levels; Rijkswaterstaat, 2023) and Denmark (20-year return levels; Danmarks Meteorologiske Institut, 2018).

Any investigation into the local storm and surge climate of the German Bight requires a sufficient representation and understanding of the larger- or *synoptic*-scale atmospheric features. In a synoptic view, the German Bight and, more generally, the entire North Sea lie well within the North Atlantic storm track (Blender et al., 1997; Dacre et al., 2012). The storm track, which roughly spans from the east coast of the United States and the Canadian Maritimes into northwestern Europe, describes a region experiencing increased occurrence frequency of extratropical cyclones and high wind speeds (Fig. 1.5). These cyclones either develop from disturbances in the flow, such as weak low-pressure systems, or move into the area from the subtropics

after undergoing extratropical transition. The intensification of extratropical cyclones into powerful storms that impact northwestern Europe is aided by strong baroclinicity, i. e., a misalignment of the local horizontal pressure and density gradients. Such baroclinic zones are usually associated with the location of the polar jet stream, which also meanders eastward across the North Atlantic, thereby governing the storm track. It follows from the physical description of the origin and development of storms and surges in the German Bight that long-term predictions of these phenomena can only be successful with models capable of simulating the connected larger-scale atmospheric dynamics reasonably well.

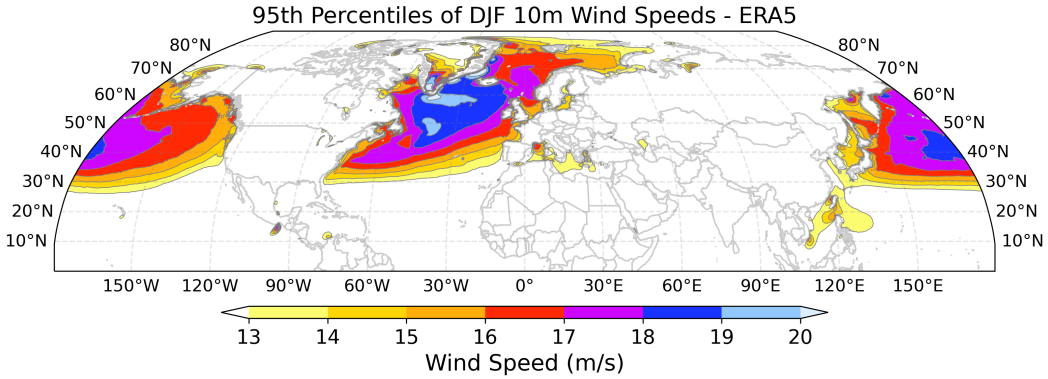


Figure 1.5: Northern Hemisphere extratropical storm track, displayed as long-term 95th percentiles of three-hourly 10 m wind speeds in winter (December–February, DJF) months. Based on data from the ERA5 reanalysis. Period 1959–2019.

1.3 How do we define storm activity?

Unlike storm or storm surge, the term *storm activity* is not strictly defined. Multiple studies have evaluated storm activity over the North Atlantic and Europe, relying on a variety of different proxies for quantification (Feser et al., 2015). These proxies range from event counts of low mean sea-level pressures (MSLP; e. g., Bärring and von Storch, 2004; Lehmann et al., 2011) or high wind speeds (e. g., Schiesser et al., 1997; Sweeney, 2000), individually tracked cyclones (e. g., Blender et al., 1997; Wang et al., 2006; Raible et al., 2008; Wang et al., 2012), and frequency analysis (e. g., Ciavola et al., 2011) to percentiles of geostrophic winds (e. g., Schmidt and von Storch, 1993; Alexandersson et al., 1998; Wang et al., 2009, 2011; Krueger et al., 2019; Krieger et al., 2021), principal component analysis of atmospheric fields (e. g., Leckebusch et al., 2008b; Bärring and Fortuniak, 2009; Gómez-Navarro and Zorita, 2013), and economic impacts and losses (e. g., Barredo, 2010; Pinto et al., 2012).

Since there exists no universal definition for storm activity, I have the freedom of choosing a suitable proxy to describe storm activity in the German Bight. This additional degree of freedom allows me to make use of the dense observational network and the shape of the German Bight coastline. With a primarily north-south oriented coastline in Schleswig-Holstein and a perpendicularly located, east-west oriented coastline in Lower Saxony, choosing sets of three observational stations along the German Bight coast that form triangles becomes possible. The triangular distribution of available air pressure observations facilitates the construction of

hypothetical near-surface *geostrophic wind* time series representative of the wind climate over the German Bight. In contrast to in-situ observations of actual wind speed which only represent the wind climate at a single location, the geostrophic wind approach gives an estimation of the average wind climate over a larger area, namely the entire German Bight, from triangles of observations that are spread along its boundaries (see Figs. 4.2 and A.1). I therefore construct indices for German Bight storm activity (GBSA) from upper seasonal and annual percentiles of geostrophic wind speeds over the German Bight, which I derive from triangles of pressure observations, as well as model and reanalysis data.

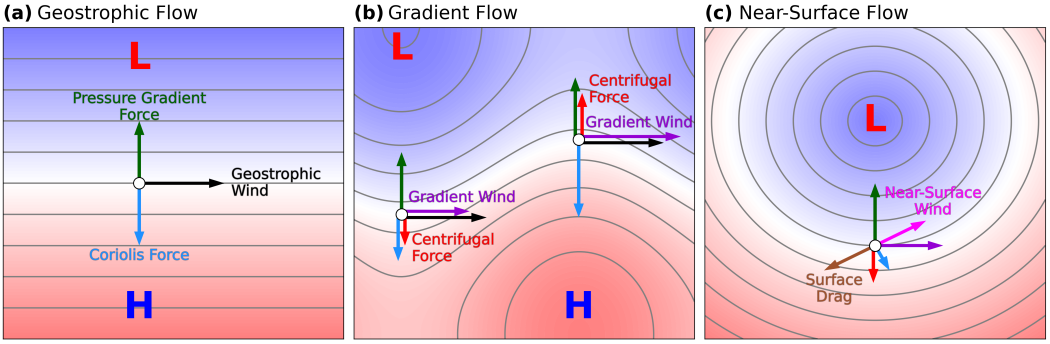


Figure 1.6: Schematic illustration of **(a)** the geostrophic wind force balance, **(b)** the gradient wind force balance, and **(c)** the near-surface wind force balance. Blue and red shadings, as well as **L** and **H** markers indicate areas of lower and higher atmospheric pressures. Isobars, i. e., lines of constant pressure, are marked as contours for illustrative purposes. Colored arrows indicate forces and the resulting winds.

The geostrophic wind is a rather conceptual horizontal wind which forms as a result of the geostrophic force balance, an equilibrium between the *pressure gradient force* $\vec{F}_p$ and the *Coriolis force* $\vec{F}_c$ (Fig. 1.6a):

$$\vec{F}_p + \vec{F}_c = 0. \tag{1.1}$$

In the Northern Hemisphere, the geostrophic wind follows the isobars, i. e., the lines of constant atmospheric pressure, in a way that higher (lower) pressures are located to the right (left) of the flow direction. The orientation of geostrophic flow is determined by the Coriolis force, an inertial force in rotating reference frames, such as the Earth, that deflects moving objects towards the right (left) in counterclockwise (clockwise) rotating systems. Geostrophic balance is only achieved in curvature-free flow. Curvature, which can for instance be induced by circularly-shaped high or low pressure systems, adds a *centrifugal force* vector $\vec{F}_\omega$ to the balance which has to be compensated by a reduction or acceleration of the flow away from geostrophy (Eq. 1.2; Fig. 1.6b). This so-called *gradient wind* is usually weaker than the geostrophic wind in cyclonically curved flow (around *low* pressure systems) and stronger than the geostrophic wind in anticyclonically curved flow (around *high* pressure systems).

$$\vec{F}_p + \vec{F}_c + \vec{F}_\omega = 0. \tag{1.2}$$

At lower levels, surface-induced drag exerts an additional *frictional force* $\vec{F}_f$ on the flow field (Fig. 1.6c):

$$\vec{F}_p + \vec{F}_c + \vec{F}_\omega + \vec{F}_f = 0. \tag{1.3}$$

This *near-surface wind* is slower than both the geostrophic and the gradient winds, and moves towards lower pressure. Thus, the assumption of geostrophic balance usually overestimates wind speeds near the surface, and creates a wind direction bias away from lower pressures. Typically, winds closer to geostrophy are found at higher altitudes, where surface-induced frictional forces are negligible. Nonetheless, Krueger and von Storch (2011) demonstrated that the geostrophic wind and its statistics can still be used to deduce assumptions about the statistics of the near-surface wind, especially in regions where frictional and orographic influences on the flow are minor.

There are some drawbacks of the geostrophic wind approach that need to be mentioned. First, because geostrophic winds are calculated from sets of air pressures that form triangles, these triangles need to be close to equilateral. As the geostrophic wind is based on the pressure gradient, strongly acute or obtuse triangles may amplify pressure errors (e. g., from measurement uncertainty) during the calculation of the horizontal pressure gradients. While the equilateral criterion can easily be satisfied on regular model grids, it could pose a challenge to observational networks. In the German Bight, we are lucky to be presented with a coastline that fosters the construction of almost equilateral triangles (Krieger et al., 2021).

Second, the geostrophic balance requires a sufficiently large contribution by the Coriolis term. This balance is therefore only observed in the extratropics. Statistical analyses of the climatology of tropical cyclones, for which the geostrophic assumption doesn't hold, can thus not be performed with the geostrophic wind approach. As the German Bight is located in the northern mid-latitudes, the necessary condition for the geostrophic assumption is met.

Since the aforementioned issues are of little concern in the German Bight, I consider the geostrophic wind approach to be appropriate for assessing German Bight storm activity. Hence, throughout this dissertation, my definition of storm activity is based on long-term statistics of the geostrophic wind $\vec{v}_g$. The geostrophic wind is derived from the horizontal pressure gradient $\vec{\nabla}_h\, p$ on a fixed height level with the vertical unit vector $\vec{k}$, a fixed air density ρ of 1.25 $\mathrm{kg/m^3}$, and the mean Coriolis parameter f of the latitudes of the three points used to determine the pressure gradient:

$$\vec{v}_g = \frac{1}{\rho f}\,\vec{k} \times \vec{\nabla}_h\, p. \tag{1.4}$$

This definition requires the availability of pressure information on a constant vertical level. From an observational standpoint, where the elevation of barometers differs between measurement sites, the comparability of measurements is usually ensured by converting the absolute station pressure to mean sea-level pressure (MSLP).

The choice of MSLP observations avoids a source of error that strongly impacts long records of wind speeds. Unlike MSLP – a slowly and steadily changing atmospheric quantity – wind speed is highly variable and fluctuates on a large range of timescales, making accurate measurements more difficult. Furthermore, the homogeneity of wind speed measurements suffers from station relocations, as well as vegetation and landscape changes upstream of the measurement site. Inhomogeneities are thus found in many time series of observed wind speeds and make these records unsuitable to detect subtle changes in the storm climate (e. g., Schmith et al., 1998). A further strength of the geostrophic wind approach is its independence of the surface characteristics. This can be particularly helpful with inter-model comparisons, where different parametrizations of the model surface would induce biases in near-surface wind speeds between different models.

1.4 How can we predict storm activity?

Skillful forecasts of the German Bight storm climate on many different timescales are desirable for the wide range of stakeholders along the coastline. When it comes to predictions of storm activity, however, it is imperative to be aware of the typical limits to prediction horizons. A good rule of thumb is that deterministic predictions of atmospheric phenomena, i. e., predictions of the exact state of the atmosphere at a certain point in time, may only be skillful as long as the forecast period is on the same order of magnitude as the average lifespan of the predicted event (Stull, 1985). In other words, individual storms, which usually last for several days to little over a week, can be skillfully predicted until about a week in advance by complex dynamical models. Storm surges, which are considered more or less direct consequences of storms, share the same potential prediction window. A reason for this constraint is that numerical weather prediction, which operates on daily-to-biweekly scales, is an *initial-value* problem. Besides understanding the physical laws that govern the evolution of the state of the atmosphere, the quality of weather prediction heavily depends on the knowledge of the initial state (Bjerknes, 1904). Small deviations from the "ground truth" at the time of initialization, that is, the start of the model run, can propagate into large errors and uncertainties in the forecast period. This error propagation leads to a point where the variability in the model, i. e., the range of different model outcomes for slightly varying initial model states, grows larger than the observed climatic variability, rendering any further prediction useless. As our knowledge of the current atmospheric state is limited by the spatial (and temporal) resolution of observational networks, there exists a natural limitation to how far out we can expect weather prediction models to show skill.

Generating reliable predictions for the long-term climate, on the other hand, is a *boundary-value* problem. Climate model simulations until the end of the century and beyond are more tailored towards accurately representing the internal variability of the Earth system, i. e., the change of climate variables like temperature on daily to yearly timescales, and how the long-term statistics of such variables develop under changing external boundary conditions. These boundary conditions are usually prescribed through radiative forcing, i. e., a change in Earth's energy balance, caused by a *projected* change in greenhouse gas concentrations, hence the name climate *projections*. As the response of the climate system to radiative forcing usually takes multiple years to establish, the climate projection models are factually independent of the initial conditions. Resulting from this independence, however, they lose the capability to deterministically predict the state of the Earth system at a fixed time, thereby decoupling temporally from the "real" world.

In this study, I am investigating the predictability of storm activity on a seasonal-to-decadal timescale, that is, from a few months up to ten years. The aforementioned rule of thumb for weather prediction prevents any attempts at sensibly predicting individual storms or storm surges months or years ahead. Yet, choosing the path of climate projections would yield too little confidence in the temporal determination of storm activity predictions, allowing for conclusions on the long-term evolution of storm activity but not the exact timing of high and low storm activity periods. Hence, rather than trying to do the impossible with either of the two presented approaches, I focus on the long-term statistics of storm and surge events, and investigate how well these time-aggregated statistics can be predicted on a seasonal-to-decadal timescale with an initialized dynamical model. Doing so, my analysis retains the information about the severity of a storm season by accounting for the number and intensity

of storms during a year or season in a combined metric, while simultaneously being independent of the exact occurrence of individual storms or surges. Also, the repeated initialization of the model, i. e., starting the model run again and again from continually updated observed atmospheric and oceanic fields, keeps the predictions connected to the actual state of the atmosphere and ocean and allows precise statements about the timing of periods of high or low storm activity.

1.5 Observing and modeling storm activity – where do we stand?

The Earth system is currently experiencing major and, on geological timescales, rapid changes, most of which can be attributed to anthropogenic influence and the global warming trend (Eyring et al., 2021). Not only does this trend affect the mean climate state in many places around the globe, it also influences the distribution and probability of extreme events. Climate projections indicate that many types of meteorological and hydrological extreme events will increase in likelihood over the next century. Especially for temperature- and precipitation-driven extremes, a connection between global warming and the increase in the frequency of these extremes has already been established through analysis of historical observations (e. g., Lehmann et al., 2015; Suarez-Gutierrez et al., 2020; Seneviratne et al., 2021), leading to a high confidence in the response of these extreme events to a warmer future climate.

Concerning the North Atlantic, different studies conclude that frequency and intensity of extratropical storms are highly variable, and do not follow a common trend across the entire region. Since the 1950s, there has been a poleward and eastward shift of North Atlantic storm activity, as well as an increase in storm intensity in higher latitudes (Zhang et al., 2004; Weisse et al., 2005; Wang et al., 2006; Raible et al., 2008). A general decrease in storm activity has been noted in the southern North Atlantic (Trigo, 2006; Wang et al., 2006; Raible et al., 2008). Tilinina et al. (2013) and Chang and Yau (2016) discovered a reduction in the number of deep cyclones in the North Atlantic between 1979 and 2010 in reanalyses, but also noticed that these changes are accompanied by high decadal variability. Other studies for the past century show that storm activity over the Northeast Atlantic and Europe does not exhibit any significant long-term trends, but instead is subject to a pronounced multidecadal variability (Schmidt and von Storch, 1993; Alexandersson et al., 1998; Bärring and von Storch, 2004; Matulla et al., 2008; Feser et al., 2015; Wang et al., 2016; Krueger et al., 2019; Varino et al., 2019; Krieger et al., 2021). A similar low-frequency variability was also detected in storm surge and sea-level records in the North Sea (Dangendorf et al., 2014; Frederikse and Gerkema, 2018) and traced back to the atmospheric forcing. Several studies suggest this variability to be linked to the North Atlantic Oscillation (NAO) (Trigo et al., 2002; Matulla et al., 2008; Donat et al., 2010; Feser et al., 2015), an atmospheric mode of variability emerging as a see-saw pressure pattern between the Azores high and the Icelandic low. However, the link between storm activity and the NAO phase depends on the investigated region and time periods, and fails to explain the variability in the early part of the 20th century (Matulla et al., 2008; Allan et al., 2009; Pinto et al., 2012; Raible et al., 2014). While Krueger et al. (2013), Tilinina et al. (2013), Chang and Yau (2016), and Wang et al. (2016) attribute the low confidence in historical trends to inhomogeneities among assimilated data, Ulbrich et al. (2009) argue that the high dissonance in cyclone related studies might result from the inherent diversity of applied methods, compared to other atmospheric variables which have historically been assigned a common

definition. Neu et al. (2013) however found that the choice of algorithms plays only a minor role in the tracking of intense cyclones, and disparities between methods are strongest for weak cyclones or during the development or dissipation stages.

The observed large variability and lack of significant long-term trends in historical storm activity impose a large uncertainty on the exact dependency between storm activity and greenhouse gas forcing, especially on a regional scale (NASEM, 2016; Vautard et al., 2019). Overall, generating robust emission-based predictions for regional storm activity over the next decades is often accompanied by low confidence. Studies on the effect of global warming on storm activity indicate a continuing poleward shift of the storm track (Lorenz and DeWeaver, 2007) in tandem with a change in storm frequency (Seiler and Zwiers, 2016; Chang, 2018), caused by a weaker low-level baroclinicity (Harvey et al., 2014; Seiler and Zwiers, 2016; Wang et al., 2017). Bengtsson et al. (2009) noted a small reduction in the number of cyclones, but no robust change in strength under a warming climate. Projections of a significant reduction of the number of extratropical storms were confirmed by Zappa et al. (2013) for the North Atlantic and Europe, and by Chang (2018) for the Northern Hemisphere. On the contrary, Yettella and Kay (2017) spot only little change in mean wind speed around extratropical cyclones between historical and future climates. Lang and Mikolajewicz (2020) observe an increase in the strength of westerly winds in the North Sea in a high-CO_2 climate, but no significant signal over the North Atlantic and Central Europe. Lang and Mikolajewicz (2020) also indicate that northwesterly storm tracks may become more intense with increasing greenhouse gas levels, but find no change in the relative predominance of storm track categories. The same study shows that, in high-emission scenarios, the increased prevalence of westerly winds may lead to an increase in frequency, duration, and height of future storm surges in the North Sea, regardless of sea-level rise. Mayer et al. (2022) later confirmed these findings. Harvey et al. (2020) discovered substantial biases in extratropical storm tracks across models of different Coupled Model Intercomparison Project (CMIP) generations, as well as a larger climate response in CMIP6 models. Throughout various studies, there is medium confidence that future changes in storm intensity are small. However, it is agreed upon that even small shifts in the storm tracks might result in large responses of extreme event frequencies and intensities in certain locations (Seneviratne et al., 2021).

As already established through a historical view on the German Bight, coastal protection, planning, and management can greatly benefit from forecasts of the storm and surge climate on a seasonal-to-decadal timescale. The aforementioned projection uncertainty however renders the use of climate scenarios obsolete for these types of forecasts. This apparent gap in predictability suggests a great potential for improvement in moving from uninitialized emission-based climate projections towards initialized seasonal-to-decadal near-term climate predictions (e. g., Kushnir et al., 2019).

In recent times, remarkable progress has been achieved in the field of decadal prediction, with studies revealing the predictability of several oceanic and atmospheric variables multiple years in advance. The research project MiKlip (*Mittelfristige Klimaprognosen*; Marotzke et al., 2016) developed a global decadal prediction system (DPS) based on the Max Planck Institute Earth System Model (MPI-ESM) under CMIP5 forcing. Investigations by Kruschke et al. (2014, 2016) within the MiKlip project found the MPI-ESM to exhibit positive forecast skill, i. e., high accuracy or goodness of the forecast, for cyclone frequency in specific North Atlantic regions

and prediction periods, even with smaller ensembles of ten members. While wind speed predictions showed promise, Haas et al. (2015) noted that skill decreased with longer lead times, that is, the temporal length between the start of the forecast and the time of the predicted event, especially over offshore regions. Moemken et al. (2021) confirmed the skill of the MPI-ESM for wind-related variables but highlighted the lower skill compared to temperature or precipitation-based predictions, with sensitivity to lead times and spatial variability. Athanasiadis et al. (2020) discovered significant predictive skill for the NAO and high-latitude blocking in the Community Earth System Model-Decadal Prediction Large Ensemble (CESM-DPLE; Yeager et al., 2018), and showed this skill to increase for longer averaging periods and larger ensemble sizes. Smith et al. (2020) demonstrated the potential for additional predictive skill extraction through output scaling of very large multi-model ensemble means.

The field of seasonal predictions and their improvement has similarly gained prominence. While subdecadal and decadal predictions may benefit from the representation of low-frequency oscillations driven by components of the Earth system with a long-term memory, such as the ocean, predictions on shorter timescales can be successful through the representation of atmospheric processes as well. Many studies have shown that current state-of-the-art climate model ensembles show noticeable prediction skill for various process of the Earth system that occur on a seasonal timescale, e. g., for the boreal winter climate (e. g., Fereday et al., 2012; Lockwood et al., 2022), and associated large-scale atmospheric modes (e. g., Scaife et al., 2014a; Athanasiadis et al., 2017). Renggli et al. (2011) found marginal predictability in their study on North Atlantic and Europe wintertime storm frequency. In a later study, Befort et al. (2018) revealed higher, but still only moderate predictive skill for direct predictions of winter storm activity. However, Befort et al. (2018) also noted that, through an indirect statistical prediction of winter storm activity via the NAO, the predictive skill could be improved in certain areas where a direct prediction failed. Degenhardt et al. (2022) built on these findings by considering multiple large-scale atmospheric modes as statistical predictors of winter storm activity. They found even higher skill than Befort et al. (2018) for the statistical approach, but no further improvement compared to direct forecasts. In addition to large-scale patterns in the boreal troposphere, Hansen et al. (2019) also underlined the importance of an accurate representation of the polar stratosphere, correct predictions of sudden stratospheric warmings, and also the state of the tropics for skillful predictions of storm activity in the Northern Hemisphere.

A strength of large-ensemble prediction systems is the possibility to select or discard individual ensemble members based on various criteria without losing too much of the internal ensemble variability. This technique is often referred to as ensemble *subsampling* (e. g., Dobrynin et al., 2018) or *subselection* (e. g., Neddermann et al., 2019). The subselection approach (illustrated in Fig. 1.7) is rooted in the hypothesis that a subset of a large climate model ensemble is able to predict components of the Earth system reasonably well, but this potential skill is hidden within the large uncertainty of the full ensemble. Therefore, the subselection rejects outliers that stray too far from the subjective truth, thereby reducing the noise and potentially improving the signal in the remaining ensemble. Ensemble subselection is considered to be a viable alternative to further inflating the ensemble size in order to keep computational costs at bay. Domeisen et al. (2015) took first steps towards the concept of subselecting members by demonstrating that a separation of ensemble members based on the simulation of sudden stratospheric warmings can improve European winter climate predictability. Further pioneering work on ensemble subselection has

been done by Dobrynin et al. (2018), who developed a statistical-dynamical approach to improve the predictability of the winter NAO. For this, Dobrynin et al. (2018) subselected a 30-member ensemble of seasonal hindcasts based on the MPI-ESM in mixed-resolution mode and CMIP5 forcing. As criteria for the subselection, Dobrynin et al. (2018) evaluated the state of four physical predictors, namely North Atlantic sea surface temperatures, Arctic sea ice volume, Eurasian snow depth, and stratospheric temperatures, in the fall months preceding the predicted winter. They generated a first-guess prediction of winter NAO based on the state of the four predictors and then selected ensemble members based on their proximity to those predictors. Doing so, they were able to significantly improve the seasonal prediction skill for the NAO and related atmospheric quantities (i. e., surface temperature, precipitation, and sea level pressure over Europe). The study by Dobrynin et al. (2018) constitutes a great example of how the potential that is hidden in large-ensemble prediction systems can be unlocked, thereby encouraging further research towards better predictability of various Earth system components (e. g., Neddermann et al., 2019; Polkova et al., 2021; Dobrynin et al., 2022; Heinrich-Mertsching et al., 2023).

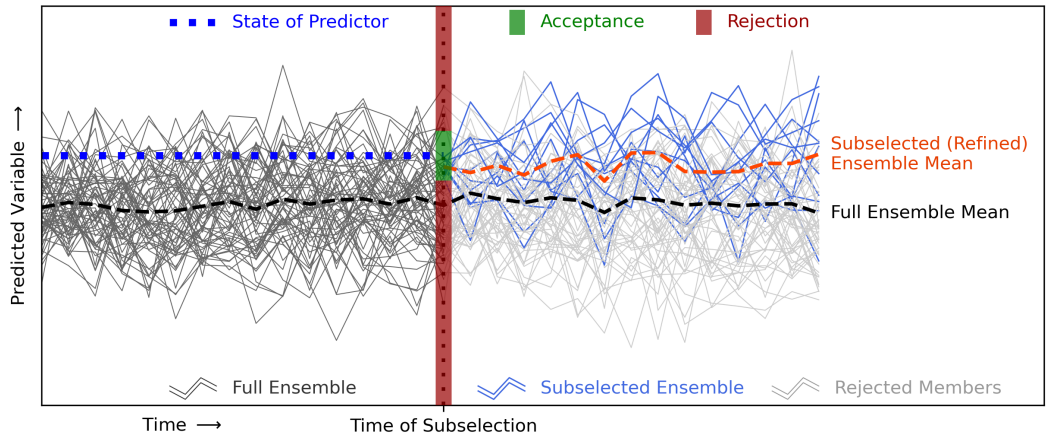


Figure 1.7: Schematic illustration of the ensemble subselection approach. An initial large ensemble (dark gray lines) is subselected at a certain time step in the model run (black dots), based on the state of one or multiple physical predictors (blue dots). Members closest to the state of the predictor are retained (green shading), while those further away are rejected (red shading). The remaining members form the new, smaller, subselected ensemble (blue lines), while the rejected members are discarded (light gray lines). The mean of the subselection (dashed orange line) is assumed to exhibit a better prediction skill than the full ensemble mean (dashed black line).

In summary, recent research shows encouraging advancements with DPSs, particularly for predicting temperature- and precipitation-related variables and the large-scale atmospheric patterns on a seasonal-to-decadal scale. However, challenges remain in predicting the small-scale wind climate and wind extremes and understanding the dependence of the forecast skill on the forecast lead time. Considering the rather low temporal resolution and ensemble size, these challenges warrant continued investigation and refinement of these models, especially since more and more sophisticated ensemble prediction systems are being developed.

1.6 How do I blend in?

Of the aforementioned studies, none had the possibility of investigating the seasonal-to-decadal predictability of a regional climate extreme with a prediction system that combines a large ensemble size and high temporal resolution. The results of both Kruschke et al. (2016) and Moemken et al. (2021) were obtained from model ensembles of ten or fewer members and daily output data. The study by Dobrynin et al. (2018) rested on subselecting a 30-member ensemble, whereas Athanasiadis et al. (2020) were able to employ a 40-member prediction system, but still relied on daily output data. Since then, further progress has been made in the development of DPSs. Based on these further improved model systems and motivated by the need for skillful seasonal-to-decadal predictions of the storm and storm surge climate in the German Bight, I aim at answering the following research questions:

1. **How well can a large-ensemble decadal prediction system predict German Bight storm activity on a decadal scale?**

2. **Can seasonal predictions of German Bight winter storm activity be improved through the use of physical predictors?**

3. **Can the decadal prediction skill for storm activity be exploited to generate skillful predictions of the storm surge climate at the German Bight coast?**

To find answers to these research question, I take advantage of a state-of-the-art single-model initial-condition large-ensemble DPS based on the Max Planck Institute Earth System Model in low-resolution mode (MPI-ESM-LR; Mauritsen et al., 2019). With 64 members that produce three-hourly output, the prediction system is unique in its large ensemble size and simultaneous high temporal resolution. Compared to previous versions of this DPS, the current version has been updated from CMIP5 to CMIP6 forcing. The DPS is initialized every 1 November from 1960 to 2019, generating predictions for the past, which are also called *hindcasts* or *reforecasts*. These hindcasts run for ten years and two months, each covering the period from November of the initialization year to December of the tenth year thereafter.

In Paper A (Chapter 2; Krieger et al., 2022), I evaluate the prediction skill of the aforementioned DPS for annual German Bight storm activity (GBSA) on a timescale of 1–10 years. Deterministic and probabilistic predictions of yearly GBSA are generated from the three-hourly MSLP output of the model. The modeled hindcast predictions are then compared to observed storm activity that has been thoroughly compiled in Krieger et al. (2021). I demonstrate how the model has difficulties predicting storm activity for single years, but shows increasing skill for longer-averaging periods in both deterministic and probabilistic prediction modes. I also point out how a differentiation of probabilistic predictions into three different categories (*high*, *moderate*, and *low* storm activity) is necessary to fully expose the model's predictive performance.

Inspired by the incapabilities of the model in predicting GBSA for single lead years and especially lead year 1, Paper B (Chapter 3; Krieger et al., 2023) aims at subselecting the large model ensemble to improve the predictability of GBSA for the subsequent winter. I apply a method developed by Dobrynin et al. (2018) that relies on physical predictors of winter storm activity. I determine tropical stratospheric temperatures in September and extratropical geopotential height anomalies in November as possible predictors for winter GBSA and generate first-guess GBSA predictions based on these predictors. By removing ensemble members that stray too far from the constraining

first guesses, I am able to significantly improve the predictability of winter GBSA in both deterministic and probabilistic prediction modes. The ensemble subselection process also enhances the predictability of large-scale atmospheric patterns, reinforcing the confidence in the physical connection between the predictors and GBSA.

As an excursion building on and going beyond the results of Paper A (Chapter 2; Krieger et al., 2022), I shift the focus to the storm surge climate of the German Bight in Chapter 4. I examine whether the windows of storm activity predictability can be transferred to also predict storm surge statistics reasonably well. Since the DPS does not provide any direct output related to sea-level heights, I demonstrate two approaches to translate model output to surge heights at the exemplary location of Cuxhaven. I first regress metrics of surge levels onto metrics of observed German Bight storms to build a regression model. I also train a machine learning (ML) model with MSLP data from the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) and recorded surge observations at the Cuxhaven tide gauge. Because the ML-based approach displays higher accuracy, I then apply the ML model to MSLP predictions of the DPS to generate surge predictions. Afterwards, I evaluate the decadal predictability of two surge-related metrics. I find that, while the ML-based translation from MSLP to surge heights produces sufficiently good results, some predictability is lost as a result of this conversion, leading to lower predictability of the storm surge climate in comparison to storm activity. Furthermore, the predictability decreases with increasing complexity of the considered surge metric, suggesting that surge predictions are pushing the limits of the DPS in its current state.

In Chapter 5, I wrap up the findings to all three research questions, give concluding remarks on the seasonal-to-decadal predictability of coastal hazards, and provide an outlook into the implications of the results presented in this dissertation.

2

DECADAL PREDICTIONS OF GERMAN BIGHT STORM ACTIVITY

Most studies in the field of decadal predictions assess the predictability of patterns and processes in the Earth system that occur on both long temporal and large spatial scales. Skillfully predicting the large-scale climate may aid in better comprehending long-term processes and how the current generation of models can represent and forecast them. The focus of this dissertation, however, lies on a comparably small-scale climate extreme, namely German Bight storm activity (GBSA). Skillful predictions of GBSA could serve as a proof-of-concept to open the door for more applied and targeted regional-scale climate forecasts. In Paper A, I therefore analyze the capability of a large-ensemble decadal prediction system (DPS) to skillfully predict the storm climate of the German Bight up to 10 years in advance, thereby tackling the following research question:

- **How well can a large-ensemble decadal prediction system predict German Bight storm activity on a decadal scale?**

Paper A addresses this research question with the help of decadal hindcasts simulated by the Max Planck Institute Earth System Model in low-resolution mode (MPI-ESM-LR) DPS. The prediction skill is assessed by comparing model-generated predictions of GBSA and winter MSLP up to ten years ahead with observed GBSA and reanalyzed MSLP fields from ERA5, each covering the period of 1960 to 2018. The choice of a model with low spatial resolution (T63 gaussian grid with 1.875° horizontal grid spacing) is a compromise that facilitates producing high-resolution temporal output and increasing the ensemble size. The number of ensemble members (64) and the temporal resolution of the model output (3-hourly resolution) are convincing characteristics that justify the use of the low-resolution (LR) version.

2.1 Predictability of MSLP

The aim of Paper A is to investigate the predictability of storm activity, which I calculate from geostrophic wind speeds derived from horizontal gradients of MSLP. This close connection between storm activity and MSLP warrants a look into the predictability of MSLP first. As annual GBSA is mainly driven by the winter months, I evaluate the prediction skill of the 64-member ensemble for winter-mean (December–February, DJF) anomalies of MSLP over the North Atlantic sector for both deterministic and probabilistic predictions. Deterministic predictions are generated via the ensemble mean, whereas probabilistic predictions take the distribution of the ensemble members into account. Probabilistic predictions are created for three distinct dichotomous forecast categories (strongly positive, near-average, and strongly negative anomalies) and consist of the fraction of all 64 ensemble members that predict MSLP anomalies above or below a pre-defined threshold. While the quality of the deterministic predictions is evaluated via anomaly correlation coefficients (ACCs) between the hindcast predictions and data from the ERA5 reanalysis, the probabilistic forecast

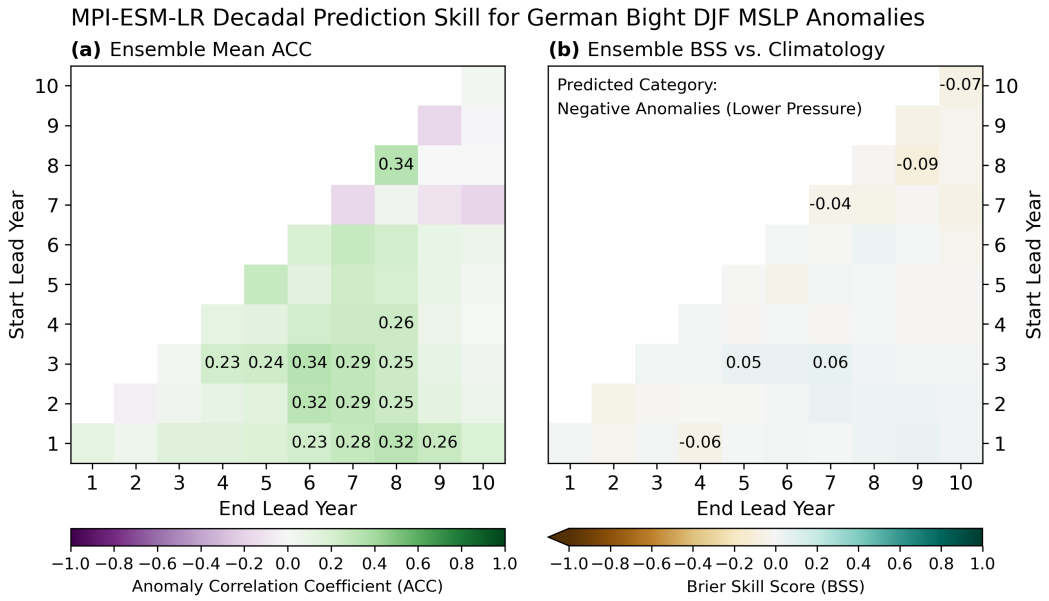


Figure 2.1: **(a)** Anomaly correlation coefficients (ACCs) between the deterministic DPS forecasts and observations of German Bight DJF MSLP anomalies, as well as **(b)** Brier skill scores (BSSs) of the DPS for negative DJF MSLP anomalies evaluated against a climatology-based prediction as a baseline. Skill scores are displayed for all combinations of start (y axis) and end lead years (x axis) for one model gridpoint at 55°N, 7.5°E near the center of the German Bight, i. e., the red marker in Figs. A.4, A.6, and A.7. For probabilistic predictions, the threshold for negative anomalies is set to one standard deviation below the long-term mean. Numbers in boxes indicate those skill scores that are significantly different from 0 ($p \leq 0.05$).

skill is assessed via the Brier skill score (BSS; Brier, 1950, Eqs. A.3 and A.4), using a persistence- and a climatology-based prediction as reference baselines.

Prediction skill is assessed for all combinations of start and end lead years. This includes single lead years (lead years 1 through 10, located along the main diagonal in Fig. 2.1 and all subsequent "matrix" plots) as well as multi-year averages with lengths of 2 to 10 years. Deterministic predictions of MSLP exhibit poor skill over the German Bight for most lead years and averaging periods (Figs. 2.1a and A.4). For a few averaging windows starting in lead years 1 to 3 and ending in lead years 6 to 8, the model presents significant skill over the German Bight, but unlike for GBSA a general increase in predictability towards longer averaging windows cannot be identified. On a larger spatial domain, longer averaging periods generally result in higher absolute correlations. However, the regions of significant skill for longer averaging periods are well removed from the German Bight, as the model performs best over the subtropical Atlantic Ocean, as well as the Subarctic and Arctic west of Greenland.

The probabilistic prediction skill for MSLP paints a similar picture. When compared against persistence, the model produces somewhat skillful predictions over the German Bight at single lead years, but fails to do so for longer averaging periods in all three forecast categories (Fig. A.6). Over the entire North Atlantic, the correlation patterns are patchy, regardless of the length of the averaging window and the forecast category. Largest skill deficits compared to persistence emerge for multi-year average predictions of negative anomalies over the Central North Atlantic. Nevertheless, the

model is able to outperform persistence-based predictions for a majority of lead times and in most regions. This supposedly good performance of the model is, however, mostly a result of the poor performance of the persistence-based forecasts. A comparison to climatology-based forecasts puts these results into perspective. Against climatology, the model shows almost no areas of skill improvement, irrespective of the averaging period or forecast category (Fig. A.7). In the German Bight, the quality of model predictions does not significantly deviate from climatology-based predictions either, regardless of lead time or forecast category. For negative anomaly predictions, which might arguably be the most relevant to the occurrence of storms, only spurious differences to climatology-based forecasts emerge (Fig. 2.1b). Just like the deterministic predictions, the probabilistic forecasts of positive and negative anomalies show vast deficiencies in an area west of the British Isles for longer averaging periods. In summary, the model performs poorly in predicting winter-mean MSLP anomalies over the German Bight and many parts of the larger surrounding regions.

2.2 Predictability of German Bight storm activity

Despite the apparent lack in prediction skill for MSLP, I again test the capabilities of the MPI-ESM-LR DPS, this time for deterministic and probabilistic predictions of GBSA. GBSA is defined as the standardized annual 95th percentiles of geostrophic wind speeds in the German Bight, which are derived from triangles of MSLP observations at 18 stations (for observed GBSA) and three model gridpoints (for modeled GBSA) along the German Bight coastline (Krieger et al., 2021, for details see Sect. 1.3 and). Since this definition relies on an accurate representation of short-term peaks in wind speeds and thus the MSLP gradients, and not on mean MSLP anomalies over a longer period of time, I presume that the model's skill for GBSA might be higher than – or at least independent of – that for simple MSLP anomalies.

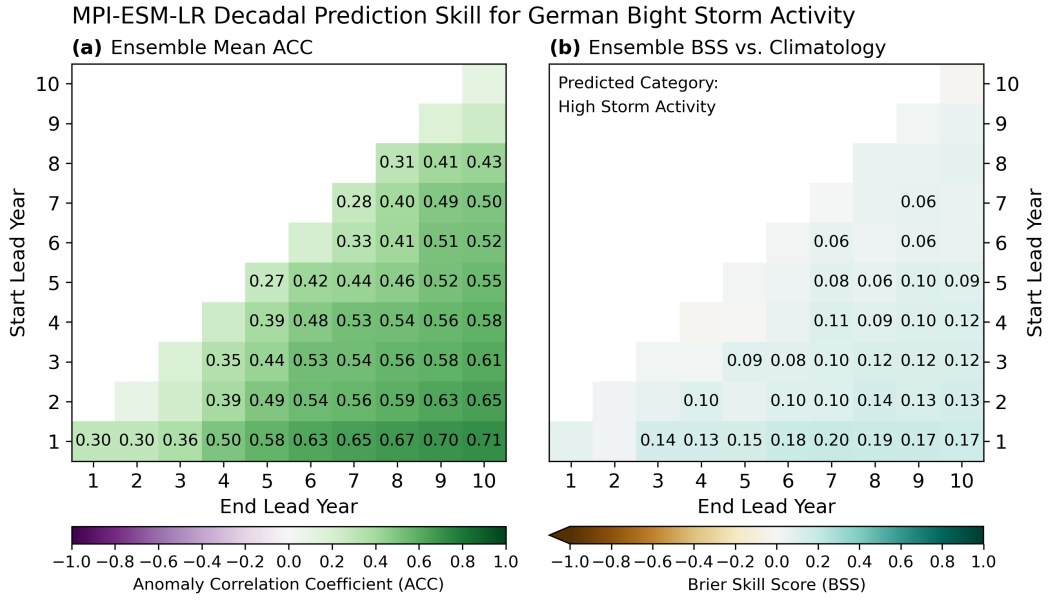


Figure 2.2: As Fig. 2.1, but for German Bight storm activity instead. The probabilistic prediction skill in **(b)** is shown for high storm activity. Adapted from Figs. A.5 and A.8b.

Again, deterministic and three separate probabilistic predictions are evaluated (high, medium, and low storm activity), with the latter ones using persistence and climatology as reference baselines. The deterministic predictions of GBSA show a clearly defined lead year dependency, with the highest ACCs emerging for the longest possible averaging period of 10 years. Most single lead years are not skillfully predictable (Fig. 2.2a). While previous studies also noted the presence of a lead year dependency (e. g., Kruschke et al., 2014, 2016; Moemken et al., 2021), they rather found that the skill decreased with increasing temporal distance from the initialization, not with decreasing averaging window length. The temporal skill pattern rather resembles the one found by Athanasiadis et al. (2020) for the predictability of the NAO. Regarding the high skill for long averaging periods of GBSA, I suspect that the filtering of high-frequency (i. e., interannual) variability that comes with longer averaging windows removes an unpredictable component, so that the model only has to correctly forecast the state of the underlying low-level variability, which is present in observational records of GBSA. The high deterministic skill for GBSA at long averaging periods is contrary to the lack of skill for German Bight MSLP at similar periods, suggesting that the long-term variability in the extremes of MSLP gradients is indeed better predictable than the mean state of MSLP.

The skill patterns of probabilistic GBSA predictions behave similarly to their MSLP counterparts. Higher skill scores emerge when using persistence as a baseline, which again are arguably attributable to persistence constituting a much less challenging reference prediction. Here, most lead year periods are more skillfully predictable by the model, except for high- and low-storm-activity forecasts at averaging window lengths of approximately 3–6 years (Fig. A.8a, A.8c, and A.8e). Notably, mainly the single lead years show up as the periods of highest skill increases, caused by a major relative underperformance of the persistence-based predictions for single lead years. Against climatology, most predictions by the model show no added skill, except for high storm activity forecasts at averaging periods of more than 5 years (Fig. 2.2b). This window of predictability is of particularly high interest, as the corresponding MSLP predictions showed no improvement over climatology, and especially predictions of high storm activity periods are beneficial for coastal planning and management. The high skill of the model that only emerges in forecasts of high storm activity also emphasizes the importance of assessing the probabilistic predictability separately for each forecast category, instead of employing a single integrated metric for all categories.

2.3 Answering the research question

In conclusion, the key findings of Paper A can be summarized as follows:

- Deterministic predictions of German Bight storm activity by the decadal prediction system based on the MPI-ESM-LR are skillful for most multi-year averaging periods, but poor for single lead years.
- Probabilistic predictions of high German Bight storm activity are more skillful than those derived from persistence and climatology for averaging periods of more than 5 years.
- The probabilistic prediction skill for high storm activity is exposed through differentiation between three forecast categories (high, medium, low), and benefits from the large ensemble size of the model.

2.4 A note on the predictability in different seasons

While Paper A only investigates annual storm activity and its predictability, I want to give a fuller picture in this dissertation by also bringing up the prediction skill for the four distinct seasons: spring (MAM), summer (JJA), fall (SON), and winter (DJF). Fig. 2.3 therefore displays the deterministic prediction skill of the DPS ensemble mean for seasonal GBSA, expressed as ACCs for all combinations of start and end lead years. It becomes immediately apparent that the prediction skill for summer (Fig. 2.3c) is lowest and insignificant across the entire lead-time domain. However, as the summer contains the fewest and also the weakest storms in the boreal extratropics, and wind speeds are generally lowest during the summer months, this season can be neglected in further analysis. If anything, the low ACC for summer reinforces the notion that the DPS performs particularly well for the parts of the year when storms are actually observed.

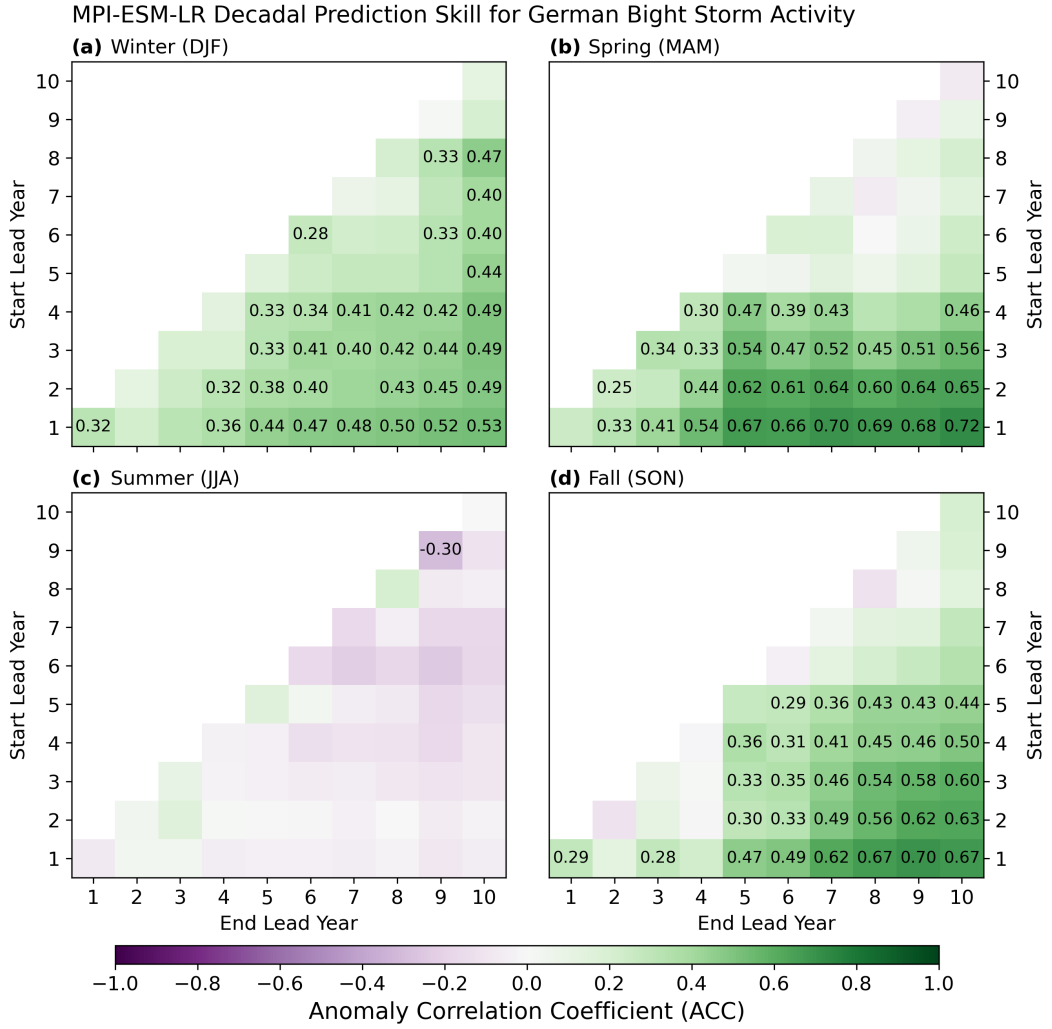


Figure 2.3: Anomaly correlation coefficients (ACCs) between the deterministic DPS forecasts and observations of seasonal GBSA for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes indicate those correlation coefficients that are significantly different from 0 ($p \leq 0.05$).

For the remaining three seasons – fall, winter, and spring – the prediction skill pattern generally follows a similar trend as for annual GBSA. The ACC increases in magnitude with increasing length of the averaging window. Absolute ACCs for long averaging-periods in winter are slightly lower than for annual GBSA, as well as for spring and fall. However, the skill for spring GBSA is insignificant for start years 5–10, while fall GBSA shows no predictability for most periods starting in lead years 6–10 or ending in lead years 1–4. It can be argued that the skill pattern of annual GBSA inherits several features from fall, winter, and spring, and that the high predictability for longer averaging periods arises not only from being able to predict the core storm season in the winter months, but also the slightly less active seasons fall and spring.

3

THE TRANSITION TO SEASONAL FORECASTS

Skillful decadal predictions of annual storm activity are of high value to stakeholders and coastal management agencies which operate on these longer timescales. However, they do not bring added value to shorter-term needs like, for instance, decisions regarding the upcoming storm season. To make matters worse, I show in Paper A and Fig. 2.3 that the skill for annual GBSA in the following year and for seasonal GBSA in the next winter is greatly reduced. To counteract this skill gap that lies between the timescales of conventional weather prediction models (up to a few weeks) and the aforementioned decadal predictions, I identify the need to improve predictability of storm activity on a seasonal scale. Drawing on the promising work by Dobrynin et al. (2018), and driven by the demand for skillful predictions of the upcoming winter storm season that is no less than that for decadal forecasts, I investigate the following research question in Paper B:

- **Can seasonal predictions of German Bight winter storm activity be improved through the use of physical predictors?**

To find compelling answers to this question, I identify physical predictors of German Bight winter storm activity on a seasonal scale. I investigate whether knowledge of the state of these predictors can be exploited to improve large-ensemble seasonal predictions of winter (DJF) GBSA. I also analyze how the predictor-based approach can add value to the predictability of the larger-scale atmospheric state. For this task, I again employ the large-ensemble MPI-ESM-LR DPS, with the same configuration as in Paper A. Contrary to most studies on seasonal prediction, which use explicit seasonal prediction systems tailored towards providing forecasts at lead times of months, I base my research on a decadal prediction system. The reasoning for this choice is twofold. Firstly, I motivate this study with the poor performance of the DPS in predicting the upcoming winter and year. Thus, a fair comparison can only be drawn through an improvement of the same prediction system that motivated the study. Secondly, and more importantly, the combination of ensemble size and temporal resolution of the DPS is hardly surpassed by any other single-model ensemble, including the MPI-ESM-based seasonal prediction system GCFS2.0 (Fröhlich et al., 2021).

3.1 Predictors of storm activity

Through a comprehensive analysis of atmospheric and oceanic parameters in the ERA5 reanalysis and their lagged correlation with observed winter GBSA, I identify September 70 hPa temperature (T_{70}) and November 500 hPa geopotential height (Z_{500}) anomalies as the two best-fitting predictors of winter GBSA on a seasonal scale. The highest correlations between T_{70} and DJF GBSA appear in a circumglobal band in tropical latitudes, roughly spanning from 20°N to 20°S (Fig. B.2). For Z_{500}, the highest correlations with DJF GBSA emerge over the Northeast Atlantic and the British Isles, as well as over the eastern part of Central Asia (Fig. B.3). A region of significant negative correlations is apparent in between, centered over the Ural Mountains.

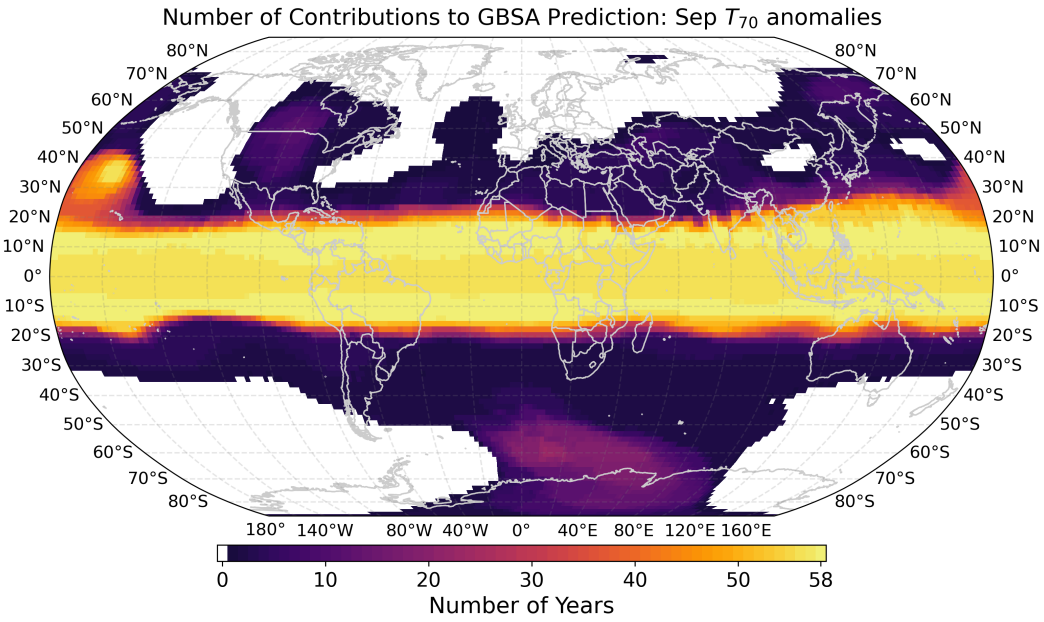


Figure 3.1: Number of model initialization for which T_{70} anomalies in ERA5 and observed winter (DJF) GBSA are significantly positively correlated from 1940/41 to the winter before the start of the model run.

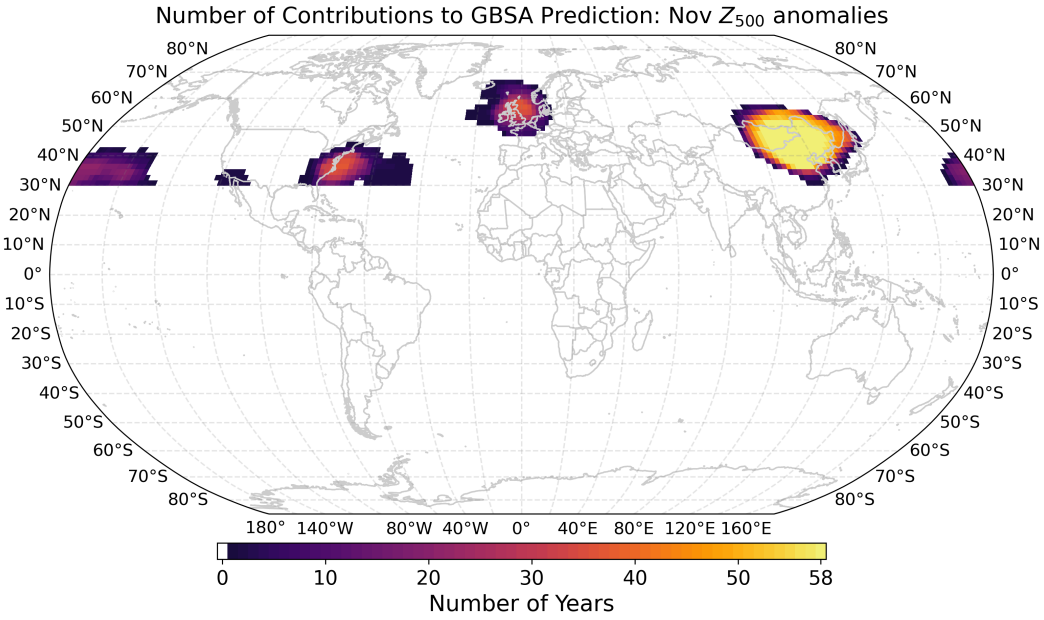


Figure 3.2: As Fig. 3.1, but for Z_{500} anomalies.

To generate a first-guess GBSA prediction from the state of each predictor, I calculate area-weighted averages of standardized anomalies of the predictor field over all areas in which the predictor is significantly positively correlated with DJF GBSA. For the correlation analysis, I take into account only the years up to one year before the start of the respective model run. With this practice, the number and location of the considered gridpoints varies from year to year. Figs. 3.1 and 3.2 show how often each gridpoint is selected and included in the calculation of the first guess of DJF GBSA. The patterns in Figs. 3.1 and 3.2 roughly correspond to the areas of high positive correlation over the entire time period (compare Figs. B.2 and B.3).

I associate the correlation patterns to two physical processes that I rely on in order to justify the ensemble subselection. The strong signal of T_{70} anomalies is attributable to the Holton-Tan effect (e. g., Ebdon, 1975; Holton and Tan, 1980), which connects the phase of the Quasi-Biennial Oscillation (QBO) to the state of the stratospheric polar vortex in the boreal winter months. The Holton-Tan effect links negative (positive) temperature anomalies in the lower tropical stratosphere to positive (negative) temperature anomalies in the polar stratosphere, effectively inducing a weakened (strengthened) polar vortex. Such a weakening (strengthening) or breakdown of the vortex further favors (disfavors) the occurrence of cold air outbreaks over Europe via prolonged atmospheric blocking patterns, effectively steering storms away from (towards) the German Bight and therefore causing below-average (above-average) storm activity. Thus, a connection between lower (higher) stratospheric temperatures in the tropics and lower (higher) storm activity in the German Bight can be drawn. The correlation pattern of Z_{500} anomalies, which I only consider in the boreal extratropics, i. e., north of 30°N, resembles that of a typical Rossby wave structure consisting of zonally juxtaposed quasi-circular areas of positive and negative correlations. The physical link to storm activity can be drawn from the occurrence (absence) of Ural blocking, which is associated with an increased likelihood of positive (negative) stratospheric winter temperature anomalies in the Arctic (Peings, 2019; Siew et al., 2020). Again, these positive (negative) anomalies are precursors to vortex breakdowns (strengthenings) and therefore also to lower (higher) winter storm activity in the German Bight. Since blocking over the Ural region can be physically connected to below-average DJF GBSA, the correlation is negative, while the regions east and west of the Ural Mountains, that would normally exhibit a troughing pattern during phases of Ural blocking, are positively correlated with winter GBSA.

3.2 Improving seasonal storm activity predictions

The predictor-based ensemble subselection approach developed by Dobrynin et al. (2018) requires the choice of a number of retained members per predictor. In their study, Dobrynin et al. (2018) selected 10 members for each of their four predictors from a 30-member ensemble, leading to a subselected ensemble size of somewhere between 10 (all predictors agree exactly) and 30 members (every member gets chosen through at least one predictor). They performed a sensitivity test by also conducting their analysis with 15 and 25 selected members per predictor, thereby noticing a change in skill depending on the number of members. Here, I too conduct a sensitivity analysis by calculating prediction skill scores for every possible number of chosen members per predictor between 1 and 64 (Fig. B.4). The optimal number for both deterministic (i. e., predicting a concrete value with the ensemble mean) and probabilistic (i. e., predicting a probability based on the ensemble distribution) predictions turns out to be 25 members per predictor. Choosing the 25 members with

a DJF GBSA closest to the initial state of T_{70} anomalies and those 25 closest to the initial state of Z_{500} anomalies yields an improvement in the predictability of winter GBSA in both deterministic and probabilistic prediction modes for the hindcasts initialized between 1960 and 2017. The deterministic skill, which is assessed via the temporal ACC and root-mean-square error (RMSE) between observed GBSA and subselected-ensemble-mean predictions, is significantly improved (Fig. 3.3). For roughly two thirds of all initialization years, the subselected prediction is closer to the observed state than the full-ensemble prediction (green markers in Fig. 3.3). The probabilistic skill for high-storm-activity predictions, which I assess via the BSS against a climatology-based prediction, is also significantly improved (Fig. B.5). I infer that the subselection process can successfully distill a high prediction skill for seasonal predictions of winter GBSA from the large ensemble spread, both for deterministic and probabilistic forecasts. The skill scores attained through the 25-member selection are higher than for most lead years on the decadal scale (compare Figs. A.5 and A.8).

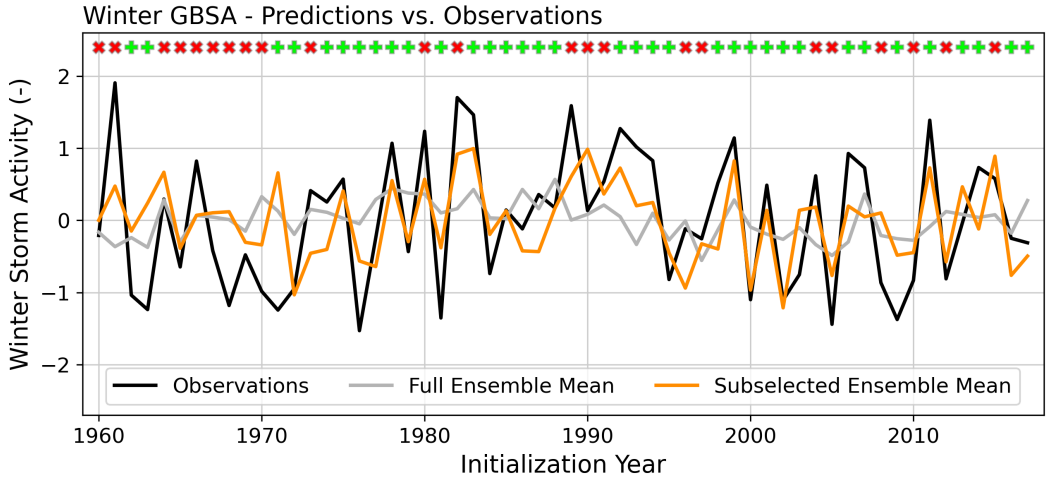


Figure 3.3: Deterministically predicted versus observed winter (DJF) GBSA (black), expressed as the mean of the full 64-member ensemble (gray) and the subselected ensemble after the predictor-based subselection (orange). Each datapoint represents one ensemble-mean forecast from a single model initialization. Green and red markers along the top denote years in which the subselection leads to an improvement or a deterioration of the prediction, respectively. Period 1960–2017 for hindcast initializations, 1960/61–2017/18 for winter GBSA. Adapted from Fig. B.5.

3.3 Improvements for the large-scale circulation

To confirm that the skill improvement through the ensemble subselection process is consistent with a better physical representation of the large-scale circulation in the remaining ensemble, I compare the ACC of the respective ensemble means for three different atmospheric variables that are associated with the state of the winter climate over Europe. The three chosen variables are winter-mean anomalies of 500 hPa geopotential height (Z_{500}), MSLP, and 200 hPa zonal wind (U_{200}). I hypothesize that an increase in skill for winter GBSA should also be reflected in a positive skill change for large-scale variables that are related to winter GBSA.

To evaluate the potential capabilities of the model, I perform a *perfect test* by repeating this analysis, but choosing the 25 ensemble members closest to the actually observed winter GBSA of each respective year instead. The perfect test acts as an upper boundary to the increase in skill that can be expected from the ensemble subselection. The increases in ACCs for MSLP, Z_{500}, and U_{200} obtained in this way are shown in Fig. 3.4. Both the subselection and the perfect test display similar spatial patterns of skill improvements, but differ in magnitude as expected. For U_{200}, which is related to the strength and location of the polar jet stream, the largest skill increase is visible close to the German Bight. Skill changes for MSLP and Z_{500}, however, are minor over the German Bight, but notably higher to the north and south. I draw the conclusion that, while not particularly better predictable over the German Bight itself, the predictability of the meridional gradient of MSLP and Z_{500} over the German Bight benefits from the subselection process. This is in line with the skill increase of U_{200}, a variable that is closely related to the gradient of geopotential heights.

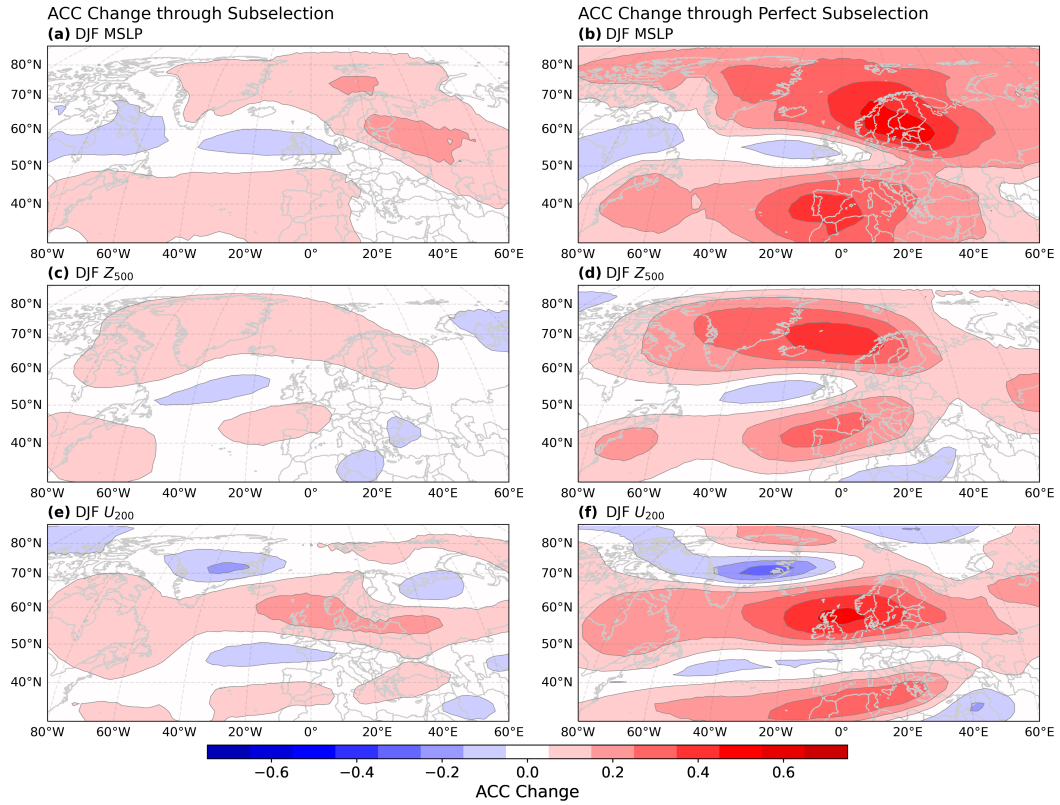


Figure 3.4: Change in ACCs between the full and subselected ensemble for winter-mean (DJF) MSLP anomalies (first row), 500 hPa geopotential height anomalies (Z_{500}, second row), and 200 hPa zonal wind anomalies (U_{200}, third row). The subselection is performed by choosing the 25 members closest to the predictor-based first guess of GBSA (left column), and for a perfect test which chooses the 25 members closest to the actually observed GBSA (right column). DJF anomalies are calculated by averaging monthly anomalies from December, January, and February. Period 1960/61–2017/18. Adapted from Figs. B.6 and B.7.

In summary, the subselection based on physical predictors of GBSA also improves the predictability of the large-scale atmospheric state of the European winter climate. The changes in predictability are subtle compared to those of a perfect test, but show similar spatial patterns, complementing the findings of Dobrynin et al. (2018). I therefore presume that, even though the perfect test suggests some unused potential contained within the large ensemble, the skill increase for seasonal predictions of winter GBSA occurs for physically consistent reasons.

3.4 Answering the research question

In conclusion, the key findings of Paper B can be summarized as follows:

- Tropical lower stratospheric temperature anomalies in September and boreal tropospheric geopotential height anomalies in November are connected to winter storm activity in the German Bight and therefore act as physical predictors.
- By using the two physical predictors and an ensemble subselection technique, both the deterministic and probabilistic prediction skill of the MPI-ESM-LR decadal prediction system for German Bight winter storm activity can be significantly increased.
- The resulting skill improvement for winter storm activity is linked to better predictability of the large-scale atmospheric patterns over Europe in general.

4

AN EXCURSION TOWARDS SURGE PREDICTIONS

Paper A demonstrated that the MPI-ESM-LR DPS shows potential to outperform conventional statistical forecasts in predicting German Bight storm activity on a decadal scale. Admittedly, storms are not the only coastal hazard that is observed in the German Bight. Due to its location and morphology, the coastline experiences a semi-diurnal tidal cycle with a tidal range of roughly two to four meters. These tides can be amplified through the large-scale wind field, leading to storm surges. Not only do these surges directly affect coastlines through inundation of land areas and coastal erosion (e. g., Kelletat, 1992), but their frequent occurrence can also lead to shifts in the submarine sedimentation process and the relocation of swash or inlet bars, which protect coastal islands from heavy surf through initiation of wave breaking (Niemeyer, 1986; Houser and Greenwood, 2007). While surge events at high tide are naturally perceived as an obvious threat, and therefore portrayed as the main concern for the immediate coastline by popular media, extremely high low tides also constitute a substantial danger to low-lying coastal regions. Due to the orographic profile of the German Bight coast, many regions near or below sea level that are protected by dikes require active draining of the inland drainage network. This draining can only take place as long as the water level of the sea is lower than inland. During periods of prolonged extremely high low tides, multiple tidal cycles may pass without the possibility to open the drainage gates, putting large inhabited areas at risk of inundation.

As already established in Chapter 1, the confidence in the evolution of the atmospheric contribution to surges in the German Bight is limited. This uncertainty is overlaid by the certain rise of the mean sea level, which, according to Steffelbauer et al. (2022) and Keizer et al. (2023), has been accelerating in the North Sea over the past decades. Irrespective of the exact pathway that the German Bight storm surge climate may take in the coming decades, the positive sea level trend and the associated shortening of drainage windows issues a challenge to coastal management agencies, who are forced to come up with new solutions to protect the coastline. Implementing such new solutions in turn requires a level of planning certainty, also in terms of expected storm surge activity, on a timescale of multiple years, which neither conventional weather forecasts nor climate projections are currently able to provide.

In this chapter, which is to be read more as a study on the applicability of previously gained knowledge rather than a simple paper summary, I will therefore explore the potential of the MPI-ESM-LR DPS to fill this gap by answering the following research question:

- **Can the prediction skill for storm activity be exploited to generate skillful predictions of the storm surge climate at the German Bight coast?**

The core quantity to evaluate when it comes to storm surge predictions is the local height of the sea surface or the *water level*. In a morphologically complex and small-scale region like the German Bight, the water level is highly dependent on the location, as it is affected by the mean sea level, the tidal cycle, the wind field, the wind direction relative to the coast, local funneling effects, the water depth, and the external surge forcing from outside the North Sea. Modeling the exact water level is therefore a challenging and computationally costly feat, which is usually accomplished with special ocean and wave models in a nested high-resolution setup with external boundary conditions. Thus, we would not expect a large-ensemble global low-resolution DPS to produce meaningful water level output. Accordingly, water level is not contained in the standard output of the DPS.

In order to still generate predictions of the storm surge climate, I therefore have to derive time series of storm surges from the model output that is available. The following sections will present two distinct approaches and their advantages and drawbacks, as well as an evaluation of the decadal prediction skill of the DPS for the German Bight storm surge climate at the exemplary location of Cuxhaven.

4.1 Extracting surges from water level records

The water level at Cuxhaven and, more generally speaking, in the German Bight is subject to a tidal cycle with a dominant period of slightly more than twelve hours and additional lower-frequency oscillations. This tidal signal is the strongest signal in the time series of observed water levels. Other influences on water levels like wind forcing and long-term mean sea-level rise are superimposed onto the tidal cycle. Since I am interested in developing a translation from storm metrics to storm surges at the coast, I first have to disassemble the water level time series in order to filter out the contribution of the wind.

For this, I use hourly astronomical tides for Cuxhaven which aim to estimate the tidal oscillation of the water level without any external atmospheric forcing (Fig. 4.1a). The astronomical tides are estimated from observed water levels (Fig. 4.1b) with the *UTide* Matlab package (Codiga, 2011) and already include the effect of the large-scale mean sea-level rise caused by the anthropogenic climate change. Therefore, I do not need to manually detrend the residual time series. The observed water levels were compiled from the Global Extreme Sea Level Analysis (GESLA) (01.01.1918–01.11.2020; Haigh et al., 2023), the Hereon Storm Surge Monitor (02.11.2020–31.12.2021; Hereon, 2023b), and the Federal Waterways and Shipping Administration (*Wasserstraßen- und Schifffahrtsverwaltung des Bundes*, WSV) (29.09.2021–30.09.2021; WSV, 2021). By subtracting the astronomical tides from hourly observed water levels, I obtain the residual water level which I attribute to the prevalent winds. I define this residual as *wind surge*, or simply *surge* (Fig. 4.1c).

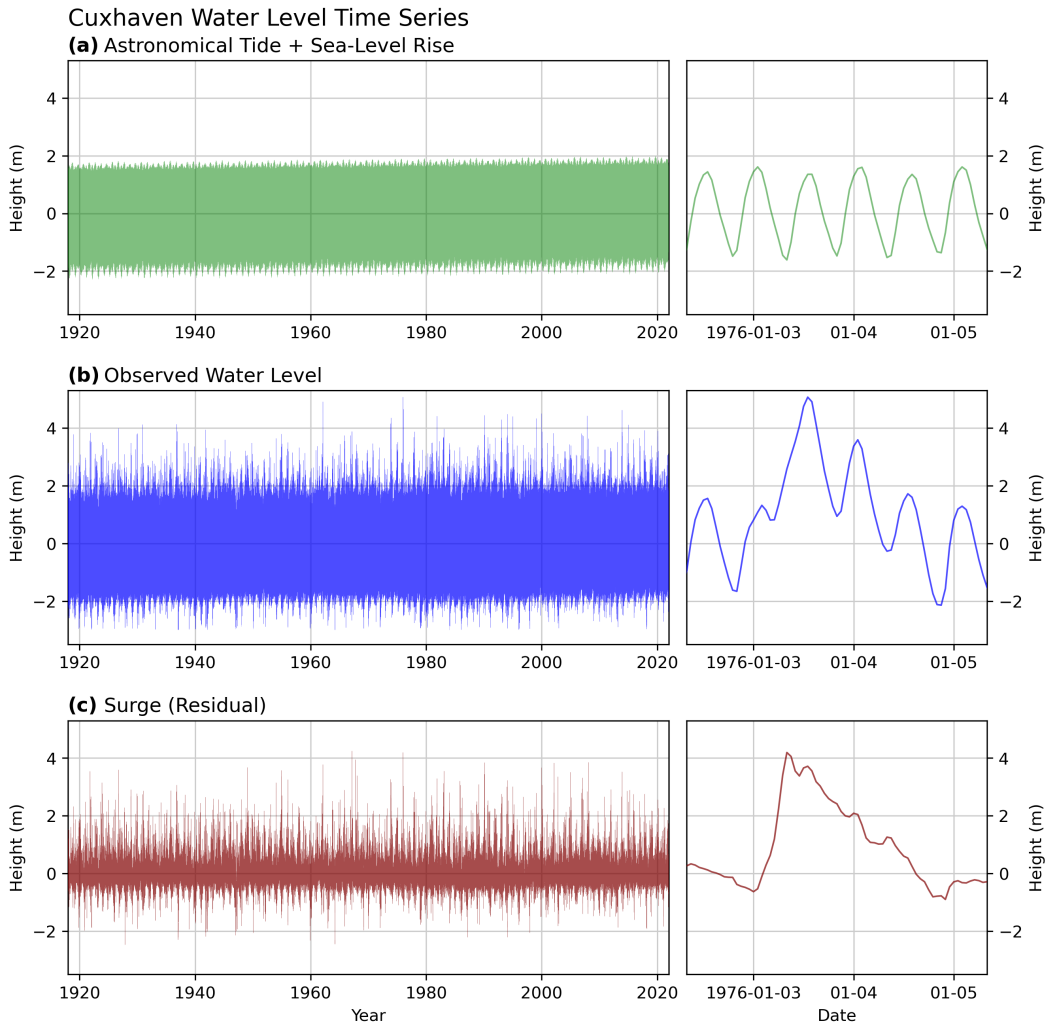


Figure 4.1: Time series of **(a)** estimated astronomical tides, **(b)** observed water levels, and **(c)** residual surge (observed minus astronomical) at the Cuxhaven tide gauge. Period 1918–2021. Right panels show exemplary zoomed-in snapshots of the time series during the January 3, 1976, storm surge.

4.2 Matching surge and storm events

German Bight surges are mostly wind-driven. Thus, the most obvious approach to translate storm activity into storm surge activity lies in associating extreme surge events with intense storm events. This association is accomplished by matching peaks in observed surge heights at the Cuxhaven tide gauge and storms derived through the geostrophic wind approach. Thereafter, surge heights are regressed onto various storm event metrics to establish a statistical relation between surges and storms.

4.2.1 The effective wind

A typical and widely-used metric for storm intensity is the maximum sustained absolute wind speed over a certain period of time. For pure storm activity considerations, the direction of the wind does only play a minor role. Storm surges, on the other hand, are strongly dependent on the wind direction relative to the coast, as the wind-driven surge is greatly enhanced by onshore flow and reduced by offshore flow.

A computationally efficient method to include both wind speed and direction is to use the so-called *effective wind* (Ganske et al., 2018). The effective wind v_{eff} is obtained by orthogonally projecting the observed horizontal wind v_{h} onto the horizontal wind direction d_{eff} that causes the highest surge at a specific coastal point. The projection is computed with the direction of the observed horizontal wind d_{h} as follows:

$$v_{\mathrm{eff}} = v_h * \cos(d_{\mathrm{h}} - d_{\mathrm{eff}}). \tag{4.1}$$

The most effective wind direction for a certain point at the coastline can be determined through comparison of wind and surge records. In the remainder of this section, I will use v instead of v_{eff} to denote the effective wind for reasons of brevity. Note that, nevertheless, all following computations in this chapter use the coast-relative effective wind and not the absolute wind speed.

To determine the most effective wind direction for Cuxhaven, I first calculate three-hourly time series of effective geostrophic wind speeds for different effective wind directions over the German Bight. I derive the geostrophic wind from three-hourly MSLP data from the ERA5 reanalysis at the gridpoints shown in Fig. 4.2. Here, I select reanalysis data over observations as the observational storm activity record from Krieger et al. (2021) does not provide a single three-hourly geostrophic wind speed record for the German Bight. Krieger et al. (2021) instead derived 18 individual time series of geostrophic wind speeds and individually converted them to a standardized storm activity index, before averaging over all 18 time series. Rather than relying on one of these 18 time series, which would be arbitrary and could potentially induce a bias, I resort to the ERA5 dataset instead.

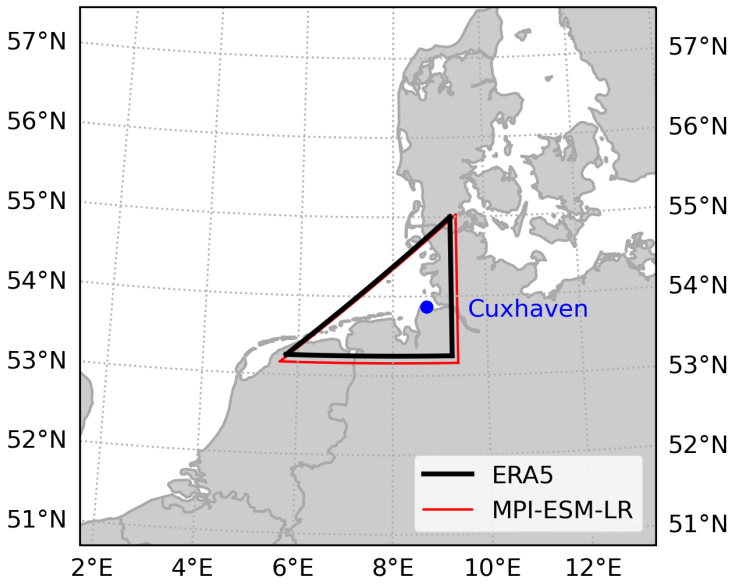


Figure 4.2: Map of the German Bight, showing the location of the triangles used for geostrophic wind calculations in ERA5 (black) and MPI-ESM-LR (red), as well as the location of the Cuxhaven tide gauge (blue).

I then extract individual storm events from the respective time series, based on criteria that closely follow the ones used in the *Hereon Storm Monitor* (Hereon, 2023a). A period of high wind speeds is hereby classified as a storm event once the wind speed exceeds the long-term 98th percentile. Individual storm events are required to be separated by at least 48 hours, measured by the time of the maximum effective geostrophic wind speed. The wind speed is also required to drop below three quarters of the 98th percentile for two peaks to be considered two individual storm events. This requirement ensures that most storms count as single events despite temporary lulls during the passage of a cyclone. An exemplary time series of effective geostrophic wind speeds and the resulting storm event classification is given in Fig. 4.3.

To detect individual surge events, I apply a fixed-threshold-based definition to the time series of surge levels at Cuxhaven. I define peaks above 1.50 m with a prominence of 0.50 m as surge events. Individual events are required to be separated by at least 24 hours, measured by the time of the maximum surge height. This definition results in 656 detected surge events between 1918 and 2021, which is slightly higher than the amount of observed storm surges in Cuxhaven based on the definition by the BSH (compare Hereon, 2023b). The discrepancy is mainly caused by my detection algorithm ignoring the tidal phase, so that some of the surges that I detect occurred during low tide and therefore did not lead to an extremely high water level.

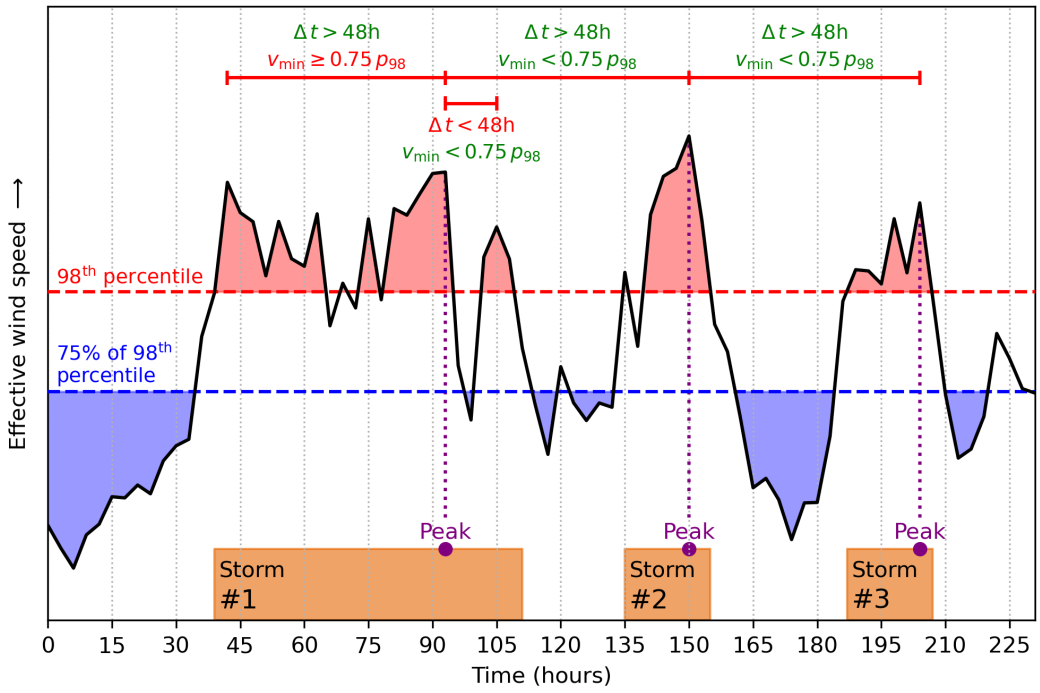


Figure 4.3: Schematic illustration of the storm event definition for an exemplary time series of effective wind speeds (solid black line). Red and blue shadings mark times when the effective wind speed is above the long-term 98th percentile (dashed red line) and below 75 % of the long-term 98th percentile (dashed blue line), respectively. Orange boxes denote individual storms and their duration, with purple dots and lines marking their respective peaks and peak intensities.

Afterwards, I correlate the annual number of storms obtained for different effective wind directions with the number of observed surge events (Fig. 4.4a). In addition, I calculate three different metrics for all storm events and subsequently correlate these metrics with observed surge heights at Cuxhaven (Fig. 4.4b, 4.4c, and 4.4d). The three metrics consist of the storm intensity, the Storm Severity Index (SSI; Leckebusch et al., 2008a), and the 12-hour-mean effective wind speed centered around the peak of each surge. I then subjectively define the most effective wind direction d_{eff} for Cuxhaven as the wind direction that results in the highest correlations across the three metrics.

The intensity of a storm event is defined by its maximum effective wind speed. For storm events in this chapter, the maximum effective wind speed does not necessarily have to be equal to the maximum absolute geostrophic wind speed, since the effective wind is modified by the wind direction relative to the coastline. The maximum effective wind speed can therefore occur at a different time than the maximum geostrophic wind speed.

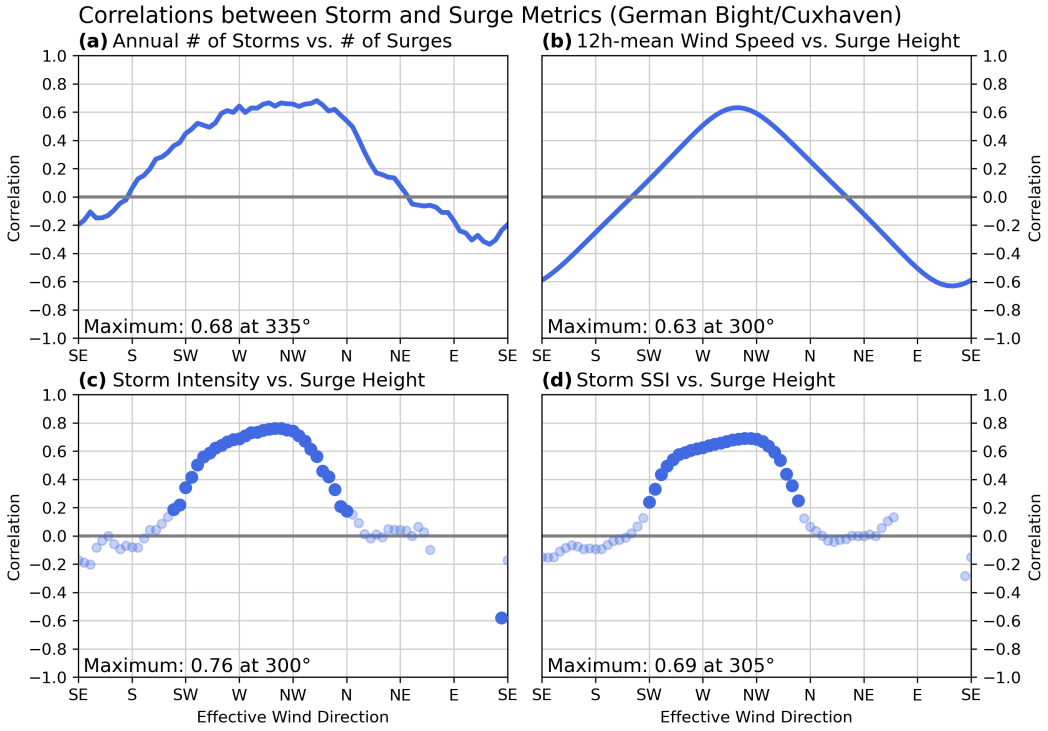


Figure 4.4: Correlation coefficients between metrics of German Bight storm activity and Cuxhaven surge events for different effective wind directions. **(a)** Annual number of storm events in the German Bight and annual number of surges at Cuxhaven. **(b)** 12-hour-mean effective wind speed in the German Bight centered around the peak of a surge event and surge heights at Cuxhaven. **(c)** Storm intensity, i.e, maximum effective wind speed, of German Bight storm events and surge heights at Cuxhaven. **(d)** SSI of German Bight storm events and surge heights at Cuxhaven. Storm events have been calculated from effective geostrophic wind speeds in ERA5 for effective wind directions between 0° and 355° in 5° intervals. Dark blue dots in **(c)** and **(d)** indicate significant correlations ($p \leq 0.05$), effective wind directions without a dot result in less than 10 total coincidences of storm and surge events. Period 1940–2019.

The SSI is an objective metric to quantify both duration and strength of a storm event. I use the area-independent definition of the SSI, which only takes the effective wind speed in the German Bight into account. The definition of the SSI follows

$$\mathrm{SSI} = \sum_{t}^{T} \max\left(0, \frac{v_t}{v_{98}} - 1\right)^3, \tag{4.2}$$

with the effective wind speed v_t evaluated at every three-hourly time step t for all time steps during a storm event T and the long-term 98th percentile of effective wind speeds v_{98}. The SSI serves as a reasonable indicator for the destructive potential of storm events, as it only yields non-zero values for speeds above a certain threshold. Furthermore, the cube of the wind speed is roughly proportional to the wind power density (e. g., Hennessey, 1977).

All four correlation tests show the highest positive correlations between the respective storm and surge metrics for effective wind directions in the westerly-to-northwesterly sector. The correlation between the annual number of storms and the annual number of Cuxhaven surge events peaks for 335°, with secondary peaks at 315° and 295° (Fig. 4.4a). Both storm intensity (Fig. 4.4c) and storm SSI (Fig. 4.4d) show the highest correlation with surge height at 305°. The correlation between surge heights and the mean effective wind speed in a 12-hour window centered around the surge peak finds its maximum for 295° (Fig. 4.4b). Based on these findings, I determine the most effective wind direction for Cuxhaven to be around 305°. Thus, in the remainder of this chapter, all storm-related metrics relate the effective wind at Cuxhaven to a flow direction of 305°.

4.2.2 Regressing surge metrics onto storm metrics

To find a fitting model for surge heights based on storm event metrics, I perform bivariate and multivariate linear regressions with one to four input metrics and surge heights as the output. I only use tuples of storms and surges if the peak of the surge occurs between 6 hours prior to and 24 hours after the peak of the storm. For the training period of 1940–2019, applying this constraint yields a total of 375 pairs of simultaneous storm and surge events at Cuxhaven. It should be noted that this training period extends further back than the period covered by the decadal hindcasts. However, I use data from 1940 onward in this approach to maximize the rather scarce number of coincidences of storm and surge events (326 pairs during 1960–2019). As training and testing data, I split the dataset of 375 event pairs into groups of 300 (80 %) and 75 (20 %) randomly selected pairs, respectively. I train the regression model on the training data and generate out-of-sample predictions from the testing data. The regression is repeated 1000 times for each combination of input metrics and subsequently assessed through evaluation of the out-of-sample predictions.

The storm metrics that act as inputs for the linear regression are the previously defined storm intensity and SSI, the duration of the storm event, as well as the storm-integrated wind speed exceedance (SIWE). The SIWE is similar to the SSI, however, instead of with its cube it scales linearly with the effective wind speed, so that:

$$\mathrm{SIWE} = \sum_{t}^{T} \max\left(0, \frac{v_t}{v_{98}} - 1\right). \tag{4.3}$$

While the SSI especially rewards high wind speeds through the third-order dependency, the SIWE is more sensitive to fluctuations in the storms near the threshold of the 98th percentile.

For the bivariate regressions, i. e., those involving only one input metric, the one based on storm intensity generates the most accurate predictions of surge height (Tab. 4.1). Models based on SSI or SIWE perform slightly worse, but still reasonably well. The storm-duration-based regression, however, reveals no connection between the length of a storm event and the resulting surge height. The multivariate regression models are all clustered within a narrow range, with skill scores similar to those of the bivariate models (except duration). While the addition of duration as an input metric does not appear to have a large negative impact on the model performance, the combination of multiple input variables also adds no significant benefit, either. No multivariate regression model is able to replicate the observed surge heights significantly better than the simple intensity-based model. Therefore, I conclude that a sufficiently good representation of surge heights can already be achieved by regressing storm intensity (i. e., the maximum effective wind speed) onto observed surges. Any further reduction of errors would require additional information beyond storm event metrics.

Table 4.1: Correlation coefficients, standard deviations (σ), and RMSEs of predicted versus observed surge heights at Cuxhaven, tested for 15 combinations of input variables. For every combination, the linear regression is repeated with 1000 different randomly selected training and testing datasets.

Input Variables	Correlation	σ (m)	RMSE (m)
Intensity	0.75	0.35	0.27
SSI	0.67	0.40	0.30
SIWE	0.65	0.41	0.30
Duration	0.15	0.54	0.40
Intensity+SSI	0.77	0.35	0.26
Intensity+SIWE	0.77	0.35	0.27
Intensity+Duration	0.76	0.35	0.27
SSI+SIWE	0.72	0.37	0.28
SSI+Duration	0.67	0.40	0.30
SIWE+Duration	0.70	0.38	0.29
Intensity+SSI+SIWE	0.77	0.34	0.26
Intensity+SSI+Duration	0.76	0.35	0.26
Intensity+SIWE+Duration	0.77	0.34	0.26
SSI+SIWE+Duration	0.74	0.37	0.27
Intensity+SSI+SIWE+Duration	0.77	0.34	0.26

4.3 Estimating local surge from atmospheric patterns

Another approach to the previously described regression of surge peaks onto storms is the application of an artificial neural network. This computationally costlier alternative to simple linear regressions allows for a direct conversion from spatial fields like those of MSLP to a target variable such as surge height. A recent study by Tiggeloven et al. (2021) demonstrated how deep learning can be successfully used to train a statistical model to predict coastal residual surge from atmospheric fields of MSLP, MSLP gradients, and wind speed. A drawback of such data-driven models is that they require an extensive set of training data. A mere database of storm events of the past century or so, such as the one used in Sect. 4.2, is insufficient for this task. Therefore, the translation from storm-related variables to surge metrics has to rely on temporally dense reanalysis data and observations. In this section, I thus show how a neural network can be trained on reanalyzed MSLP data to translate fields of MSLP, which are contained in the output of MPI-ESM-LR, to surge heights in Cuxhaven.

I use a sequential artificial neural network to predict surge heights at Cuxhaven from MSLP fields over the German Bight and the adjacent sector of the Northeast Atlantic and Europe, bounded by 30°N, 75°N, 25°W, and 40°E. I train the model with hourly fields of MSLP from ERA5 and hourly observations of surge heights at Cuxhaven from 1960 to 2021. As the effect of the MSLP and thus also the wind field over the German Bight on coastal tides and surge is not instantaneous but happens at a time lag of multiple hours, I shift the observed surge at Cuxhaven by a time lag which I determine through correlation analysis of observed surge and MSLP from reanalysis data (Fig. 4.5). The correlation analysis reveals a maximum absolute correlation of 0.715 at a time lag of six hours, indicating that the effect of the MSLP field over the German Bight has the biggest impact on surge heights in Cuxhaven six hours later.

The artificial neural network consists of an input layer, a dense hidden layer, and an output layer. The input layer of the neural network flattens the two-dimensional MSLP input grid (25×35) into a one-dimensional vector. The flattened array is then fed into a 32-neuron dense hidden layer, which is activated with the Rectified Linear Unit (ReLU) activation function (Wani et al., 2020). The output layer consists of a single neuron, which predicts the surge height. This layer is activated by a linear activation function to ensure continuous predictions of surge heights. The neural network uses the mean squared error (MSE) as a loss function with an adaptive moment estimation (*Adam*) optimizer (Kingma and Ba, 2014). In total, I train the neural network with a batch size of 64 over 20 epochs.

The entire dataset surge heights and MSLP fields (62 years at hourly resolution) is split into a training, a validation, and a testing subset. The training dataset consists of randomly selected datapoints encompassing 64 % of the full dataset and is used to train the neural network. The validation dataset (16 %) is used during the learning process to optimize the neural network. The testing dataset is made up of the remaining 20 % and provides unseen data to evaluate the final fitted model.

Testing the final fitted model with unseen MSLP data reveals a high correlation of 0.891 between predicted and observed surge at Cuxhaven (Fig. 4.6). The largest differences between model predictions and observations are found in both tails of the distribution, where the model underestimates surge extremes.

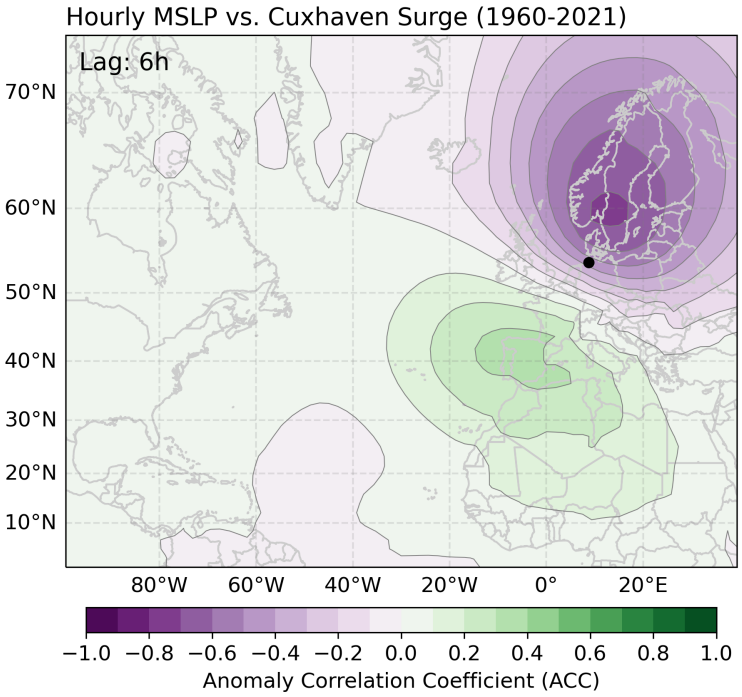


Figure 4.5: Anomaly correlation coefficients (ACCs) between hourly MSLP anomalies over the North Sea and Northeast Atlantic from ERA5 and hourly observed surge heights at Cuxhaven (indicated by black dot). MSLP and surge are lagged by six hours. Period 1960–2021.

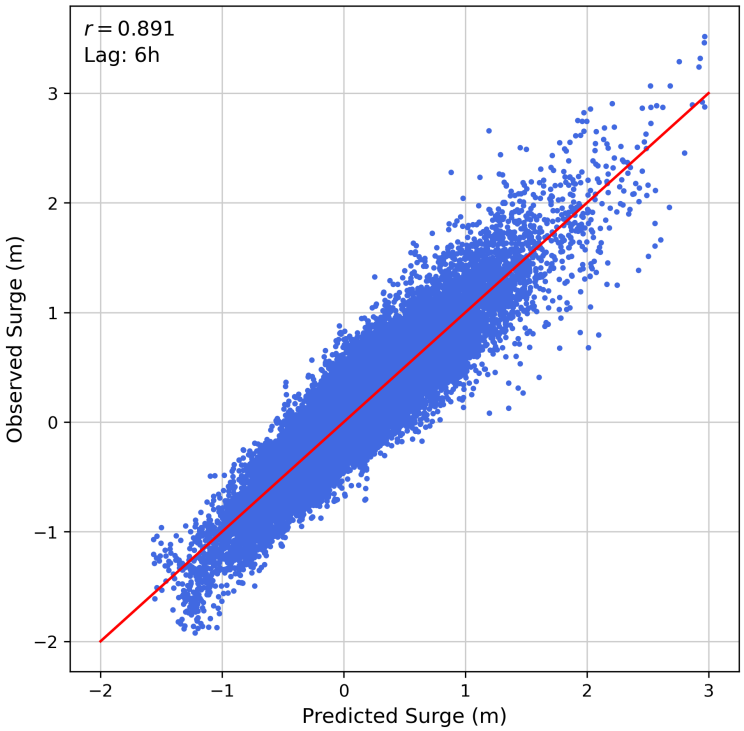


Figure 4.6: Predicted (x axis) and associated observed (y axis) hourly surge at Cuxhaven for the testing dataset, i. e., randomly selected 20 % of the full 1960–2021 period. MSLP and surge are lagged by six hours. The red diagonal indicates the target of $y = x$. Correlation between predictions and observations: 0.891.

4.4 Applying the model to hindcast output

Overall, the correlation of 0.891 for the ML-based model indicates a better and more robust fit between surge and MSLP than the multivariate linear model derived from extreme events in Sect. 4.2. Therefore, I conclude that using a ML approach is the better option to generate three-hourly hindcast predictions of surges at Cuxhaven from the three-hourly MSLP output data of the DPS. Still, Fig. 4.6 raises awareness for a small systematic error between ML-based surge height predictions and observed surge levels in both tails of the distribution.

Additionally, I account for potential MSLP biases and differences in the variability of MSLP between the DPS and the ERA5 training data. Fig. 4.7 compares the mean annual SSI and SIWE in ERA5 and the DPS. The annual SSI in ERA5 is about 60 % higher than in the ensemble mean, and about 10 % above the uppermost outliers in the model. The high SSI in ERA5 is likely a result of the ability of the reanalysis to capture extreme wind speeds better than the low-resolution hindcast, as the third-order dependency of the SSI on wind speed rewards high absolute speeds. Contrary to the SSI, the SIWE represents a metric for which the ERA5 mean lies within the ensemble spread of the DPS, despite still being above the ensemble mean for every lead year. The large discrepancy in SSI but simultaneous agreement in SIWE indicate that the strongest storms are significantly underrepresented in the model, leading to an underestimation of annual SSI. Based on this apparent underrepresentation of very high geostrophic wind speeds in the DPS, it is advised against performing direct comparisons of absolute surge heights. Thus, I map the generated hindcast time series of surge heights to quantiles for every ensemble member separately.

By using member-specific surge quantiles as reference values for the eventual calculation of surge statistic, I can assure that certain ensemble members are not over- or underrepresented in the number of storm surges just by virtue of showing an above- or below-average number of extreme MSLP gradient patterns.

To identify surge events, I apply a peak detection algorithm to the three-hourly time series of surge quantiles to each ensemble member separately. I define individual surge events as peaks above the 0.98-quantile that are both at least 24 h apart from each other and characterized by a quantile prominence of 0.03, meaning that the surge height has to drop below the 0.95-quantile between two peaks in order for them to be considered two separate events.

It should be noted that the conversion from MSLP patterns to surge heights assumes independence of the phase of the tidal cycle. This could mean that a high theoretical surge event might in practice coincide with low tide and thus not be of any immediate danger to the coastline, whereas other lower surge events that occur during high tide could potentially cause more coastal inundation, but appear as less severe events in the dataset. However, I argue that in reality the occurrence of storms is not linked to the tidal cycle, so that roughly the same number of storms occur during high and low tides on average. With a sufficiently large dataset, such as our hindcast, I thus conjecture that the erroneous interpretation of surge heights and surge events arising from the missing knowledge of the tidal phase tends towards negligibility.

Another assumption made by this ML-based conversion is the independence of the potential of the wind to generate surge on the initial water level. In reality, the topography of the sea floor leads to a growth in basin area with height, so that the same increase in water level requires more water volume at high tide than at low

tide. In other words, the same MSLP pattern might cause a higher surge at low tide than at high tide, introducing an initial training error to the ML training process. This error is subsequently handed down to the generated DPS surge event record through the application of the ML model to the hindcast data.

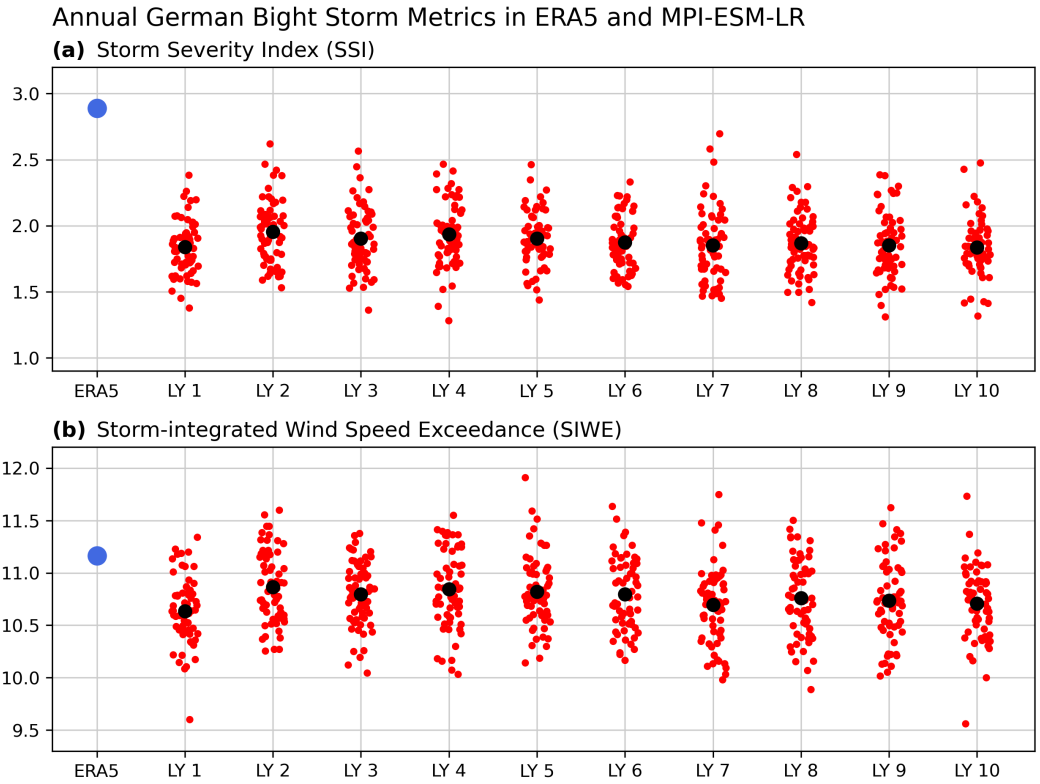


Figure 4.7: Mean annual **(a)** SSI and **(b)** SIWE over the German Bight for the ensemble mean (black dots) and individual members (red dots) of the hindcast for lead years 1–10, as well as calculated from ERA5 (blue dot in first column). Period 1940–2019 for ERA5, hindcast runs initialized 1960–2019.

4.5 Prediction skill for different surge metrics

With the ML-based generated time series of Cuxhaven surge heights in the DPS and observed surge heights at the Cuxhaven tide gauge, I now evaluate the decadal prediction skill for two differently complex surge metrics. The metrics I consider are (1) the DJF 98th percentiles of surge heights and (2) the more complex annual number of long-lasting surges, defined as events that exceed the long-term 75th surge height percentile by at least 24 consecutive hours. The skill evaluation follows the workflow of Krieger et al. (2022) such that I again take into account all lead years and averaging windows lengths. I investigate both deterministic predictions based on the ensemble mean of the hindcast and probabilistic predictions for high surge activity. High activity is hereby defined via a threshold of one standard deviation above the long-term mean of the respective metric. The probabilistic predictions are evaluated against a climatology-based prediction which assumes the underlying statistic to follow a Gaussian normal distribution (see Sect. A.2.5 for details). To ensure a fair comparison, I also map the absolute observed surge heights at Cuxhaven to quantiles,

and apply the same peak detection algorithm to the resulting quantile time series as for the hindcast data.

The deterministic prediction skill of the DPS for the DJF 98th percentiles of surge heights at Cuxhaven (Fig. 4.8a) bears a rather strong resemblance to the skill matrix for annual and winter GBSA (compare Figs. 2.3 and A.5) with slightly lower absolute correlations. The similarity hints at a fair preservation of predictability throughout the conversion from MSLP patterns to surge heights. Some reduction of skill is to be expected, since the ML model injects a certain level of error into the conversion, which is inevitably reflected in the resulting ACCs. Contrary to the skill matrices for annual and winter GBSA, however, there is no skill for the first three lead years. For probabilistic predictions of higher-than-normal 98th percentiles (Fig. 4.8b), some similarities to the annual GBSA skill matrix (Fig. 2.2b) are also apparent.

While short-lived high peaks in the storm surge record may correspond to actual high storm surges when occurring at high tide, these events are not the only threat to the coastal zone. Besides these events that may cause short-term inundation, the low-lying coast is also threatened by longer-lasting high water levels that prevent effective drainage from the interior land through tide gates. Exterior water levels far above the regular low tide for a prolonged time, i. e., across multiple tidal cycles, has the capability to increase interior water levels through a lack of release of water into the sea. To estimate the prediction skill of the DPS for these kinds of events, surge events capable of producing interior flooding events are defined as periods where the surge height does not drop below the 75th percentile for at least 24 consecutive hours, corresponding to at least two full tidal cycles. The deterministic prediction skill for the annual number of such events (Fig. 4.9a) is particularly low, especially when compared to the previously analyzed storm surge metrics. There appears to be a small window of opportunity around the lead year 3–6 mark. However, most of the lead year ranges exhibit no significant skill at all, hinting at a possible spuriosity of the few remaining positive significant correlations. The probabilistic skill matrix confirms this notion, as most lead year periods do not exhibit any skill improvement over a climatological prediction (Fig. 4.9b). For some lead times, the model even performs significantly worse than climatology.

The comparison of Figs. 4.8 and 4.9 shows that an increase in the subjective complexity of the considered surge metric notably lowers the predictive capability of the DPS. While simple percentiles of surges can be predicted reasonably well in deterministic mode, albeit a little worse than the corresponding percentiles of storm activity, the more complex long-term surge height exceedances show almost no significant skill anymore. A comparable picture is painted for probabilistic predictions of above-average surge activity, where the skill gain over climatology from employing the DPS is mostly negligibly small and completely absent in predictions of annual surge and long surge numbers. For more complex metrics and the skillful prediction thereof, one would have to improve the conversion from existing model output to surge heights by, for instance, employing a more sophisticated model or explicitly simulating water levels.

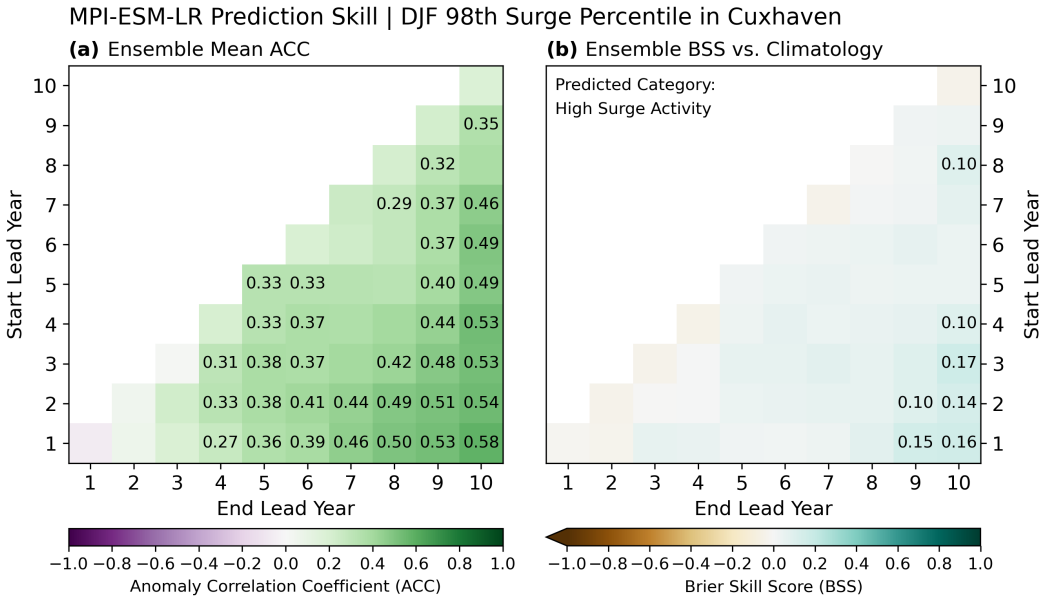


Figure 4.8: **(a)** Anomaly correlation coefficients (ACCs) between the deterministic DPS forecasts and observations of winter (DJF) 98th percentiles of surge heights in Cuxhaven, as well as **(b)** Brier skill scores (BSSs) of the DPS for high surge activity evaluated against a climatology-based prediction as a baseline. Skill scores are displayed for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes indicate those skill scores that are significantly different from 0 ($p \leq 0.05$). For probabilistic predictions, the threshold for high activity is set to one standard deviation above the long-term mean. Numbers in boxes indicate those skill scores that are significantly different from 0 ($p \leq 0.05$).

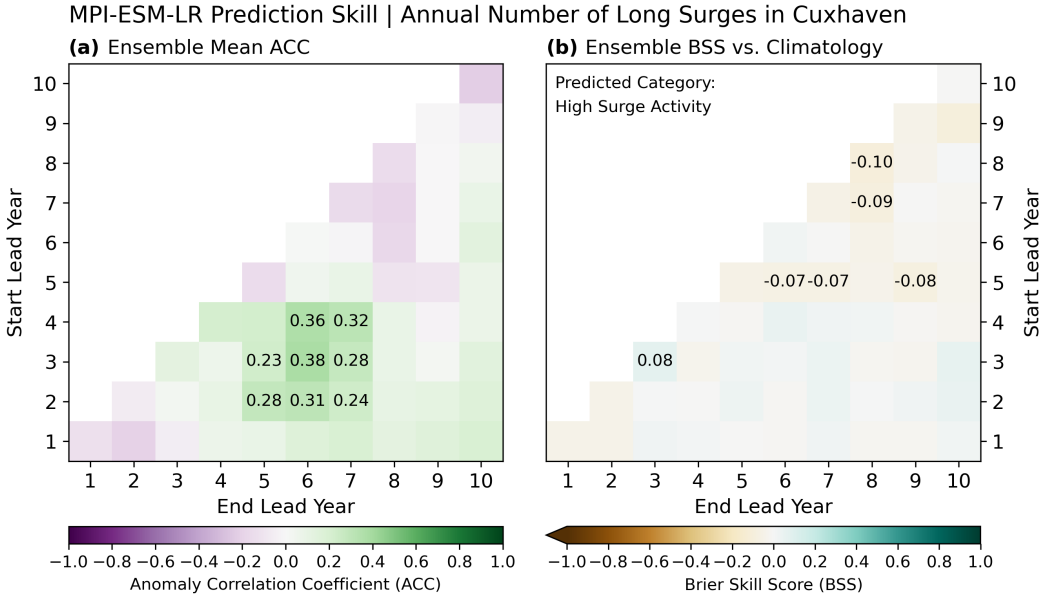


Figure 4.9: As Fig. 4.8, but for the annual number of surge events that exceeded the long-term 75th percentile of surge heights by at least 24 consecutive hours.

4.6 Was the excursion successful?

In this chapter, I showed two methods to derive surge statistics from available model output of the MPI-ESM-LR DPS. First, I converted geostrophic wind speeds to the direction-dependent effective wind to associate surge-relevant storm events with surges at the Cuxhaven tide gauge. I then used the pairs of matching extreme events to find the best-fitting multivariate regression between surge and storm event statistics. Overall, storm intensity, i. e., the maximum effective geostrophic wind speed during a storm, proved to be the best fit to estimate peak surge heights. Unfortunately, the correlation of the regression turned out to be too low for further application on the hindcast output, also owing to the limited number of surge-storm pairs available for training a regression model. Inspired by the findings of Tiggeloven et al. (2021), I then trained a sequential artificial neural network on hourly fields of MSLP from ERA5 and hourly surge heights in Cuxhaven to predict surge from two-dimensional patterns of atmospheric pressure. At a time lag of six hours (MSLP leads surge), the neural network achieved a correlation coefficient of almost 0.9 for predicting surge heights from previously unseen MSLP fields.

I then applied the neural network to modeled MSLP to generate hindcast time series of surge heights in Cuxhaven. To account for differences in the distribution and magnitudes of MSLP between training and model data, I converted surge heights to quantiles in all datasets. I evaluated two differently complex surge activity metrics, both for deterministic ensemble mean predictions and probabilistic predictions of high activity, defined as one standard deviation above the mean. The probabilistic predictions were evaluated against climatology. Overall, the DPS showed lower skill for the surge metrics than for storm activity, probably caused by the ML-based translation from MSLP to surge heights. While the model still showed fair predictability for annual percentiles of surge heights, the number of long surges per year turned out to not be predictable on a decadal scale anymore. The findings of this chapter indicate that decadal predictions of complex small-scale surge metrics which require a translation from other modeled variables might be pushing beyond the current capabilities of the decadal prediction system.

4.7 Answering the research question?

In summary, the key findings of this chapter can be outlined as follows:

- By using a statistical machine-learning approach, mean sea-level pressure model output can be converted to storm surge heights at Cuxhaven with reasonable accuracy.
- The prediction skill of the MPI-ESM-LR decadal prediction system for storm surge statistics at Cuxhaven shows similar patterns in temporal space to that of German Bight storm activity, although the skill is lower in general. The difference in skill can most likely be attributed to a partial loss of information during the conversion to surge heights.
- The predictability for different surge-related metrics decreases further with increasing subjective complexity of the considered metric.

5

CONCLUSIONS AND OUTLOOK

During the three storm seasons that were spent creating this dissertation, 38 storms impacted the German Bight, leading to 24 storm surges in Hamburg and 11 in Cuxhaven (Liu et al., 2022; Hereon, 2023a,b).

While these coastal extreme events repeatedly challenged coastal protection agencies, emergency management, and the general population along the German Bight coastline, I strove to answer the research questions posed in Chapter 1 in order to improve our understanding of the predictability of these extremes and their long-term statistics. The answers to the research questions are presented in Sect. 5.1. Sect. 5.2 then gives concluding remarks on the findings and provides a short outlook on the implications of these results.

5.1 Summary of the results

- **How well can a large-ensemble decadal prediction system predict German Bight storm activity on a decadal scale?**

To answer the first research question, I evaluated the prediction skill of the decadal prediction system (DPS) based on the Max Planck Institute Earth System Model in low-resolution mode (MPI-ESM-LR) for the geostrophic-wind-based German Bight storm activity (GBSA). I generated deterministic predictions based on the ensemble mean, and exploited the large ensemble size of 64 members to produce probabilistic predictions based on the ensemble distribution. I then compared deterministically and probabilistically predicted GBSA and winter mean sea-level pressure (MSLP) anomalies to observations to determine the prediction quality of the model. I found that the model produces poor deterministic predictions of GBSA and winter MSLP anomalies for individual years but fair predictions for longer averaging periods. A similar but smaller skill difference between short and long averaging periods emerged for probabilistic predictions of high storm activity. At averaging periods longer than 5 years, I showed that the model is more skillful than both persistence- and climatology-based predictions. For shorter aggregation periods (less than 5 years), the model is still superior to persistence-based predictions, but shows no additional skill over climatological predictions. I thus concluded that, for the German Bight, skillful decadal predictions of high storm activity are possible for long averaging periods with a carefully chosen approach and a large ensemble. Notably, a differentiation between probabilistic forecast categories (in this case: high, moderate, and low storm activity) is necessary to expose this skill. These findings are presented and further elaborated on in Paper A (Chapter 2; Krieger et al., 2022).

Based on the findings of the first research question, I turned to predictions of GBSA on the shorter seasonal timescale. Since the full ensemble can barely generate skillful predictions for winter GBSA in the first lead year (Fig. 2.3), the question arose whether these predictions can be enhanced through ensemble subselection.

- **Can seasonal predictions of German Bight winter storm activity be improved through the use of physical predictors?**

In Paper B (Chapter 3; Krieger et al., 2023), I showed that an improvement of the predictability of winter GBSA is possible through the use of two physical predictors, namely September 70 hPa temperature anomalies and November 500 hPa geopotential height anomalies. These predictors influence the winter storm climate over the German Bight through stratospheric pathways via the stratospheric polar vortex. I used the state of the predictors to generate a first-guess prediction of winter GBSA and removed ensemble members from the DPS that deviated too far from the first-guess prediction. I found that the remaining subselected ensemble showed a significantly increased forecast skill for both deterministic and probabilistic predictions of winter GBSA and an associated increase in the predictability of storm activity-related atmospheric fields over Europe.

Motivated by the promising findings on the decadal predictability of storm activity, I investigated whether the capability of the DPS to skillfully predict GBSA can be translated into skillful predictions of storm surges at the German Bight coastline.

- **Can the skill for storm activity be exploited to generate skillful predictions of the storm surge climate at the German Bight coast?**

Answering the third research question required a translation from storms to surges, as surge heights are not predicted by the model directly. I therefore regressed observational surge data at the exemplary location of Cuxhaven onto storm activity metrics to link surge heights to storm intensity. In order to minimize the loss of predictive skill that emerges from inferring surge statistics from storm activity statistics and to increase the sample size for the regression, I used a machine learning approach and trained a convolutional neural network to estimate surge height directly from hourly MSLP patterns over the Northeast Atlantic region, a method that has been previously used by Tiggeloven et al. (2021). I then applied the neural network to MSLP fields from the DPS to create predictions of surge heights, from which I evaluated the prediction skill of the surge climate, similar to the workflow in Paper A. Unfortunately, the predictive skill for the surge climate turned out to be slightly lower than that for German Bight storm activity and to deteriorate with increasing complexity of the chosen metric. While simple percentile-based surge indices are somewhat similarly predictable, I discovered that generating skillful forecasts of more advanced metrics that include consecutive surges or periods of higher-than-normal low tides already pushes beyond the limitations of the DPS. These results are presented in Chapter 4.

5.2 A look ahead: stormy times or smooth sailing?

In the pursuit of assessing the predictability of small-scale climate extremes, this study has unraveled where we stand in terms of current state-of-the-art modeling capabilities. The complexity of predicting such spatially limited phenomena on timescales much longer than their typical lifespan has long been acknowledged as a formidable challenge, and remains a challenging task to this day. Nevertheless, I have demonstrated in this dissertation that, in the temporal no-man's-land that is the seasonal-to-decadal timescale, predictability patterns emerge at lead time windows that one would not expect intuitively. Complementing the findings of, for instance, Kruschke et al. (2016) and Moemken et al. (2021), I illustrated that German Bight storm activity is indeed predictable on multiannual scales with a large ensemble and

high temporal resolution. I also underscored that probabilistic predictions for high storm activity can be skillful with a sufficiently fine segmentation of the prediction categories. Especially for a phenomenon whose response to climate change we can not yet confidently estimate, my findings disclosed that we are still able to draw on the internal low-frequency climate variability to produce forecasts of a higher quality than by using simple climatological statistics. On the seasonal timescale, where the full ensemble exposes deficiencies, I presented the viability of a statistical approach to distill additional prediction skill. These new insights add onto the efforts of Befort et al. (2018) and Degenhardt et al. (2022) and unveiled how both dynamical and statistical viewpoints can be utilized in synergy to enhance the quality of forecasting climate extremes.

While the findings on storm activity sound very promising, the limitations of the model world quickly approach as the focus turns toward storm surges. In an effort to derive surge heights with a simple machine-learning approach from model output, I revealed that the inherent loss of information and the complexity required for impact-relevant prediction metrics is likely one step too far for the prediction system. Despite a surprisingly fair reproducibility of actual surge heights from modeled pressure patterns, which follows the work of Tiggeloven et al. (2021) but requires far less input parameters, the previously opened predictability windows close again, once the predicted statistic becomes too “complex”, subjectively speaking. While these revelations might at first glance appear as a cul-de-sac, I would still like to incentivize more research into this direction, as I firmly believe that more sophisticated techniques could push the boundaries even further. Also, the profound socio-economic aspect of surge impact predictions along the coast might be of even more elevated relevance than that of storms alone.

The new possibilities in forecasting storm – and to a certain extent also surge – activity beyond the temporal barriers of conventional weather prediction are a contribution not only in a scientific way, but also to society in the coastal areas of the German Bight. The proof-of-concept that is the skillful prediction of storm activity on both a seasonal and a decadal timescale may pave the way for the eventual issuance of operational forecast products, targeting actual stakeholders along the coast that operate on such timescales. By reducing forecast error, providing more reliable predictions, and being able to better quantify and thus communicate probabilities and uncertainties, coastal planning and management agencies will be able to focus their preparation and mitigation efforts more effectively. In addition, local stakeholders that are impacted by coastal hazards or rely on predictions of the storm climate like the energy, infrastructure, and insurance sectors, will benefit from skillful predictions through minimization of damages and losses, and opportunities to better adapt to the climatic variability. All in all, my dissertation on the predictability of storm and surge activity reveals an emerging tool that can aid the society along the German Bight coastline in its endeavor to become more resilient to an ever-changing climate.

Publications

The following two publications were prepared over the course of the doctoral studies.

A

SKILLFUL DECADAL PREDICTION OF GERMAN BIGHT STORM ACTIVITY

The work in this chapter has been published as:

Krieger, D., Brune, S., Pieper, P., Weisse, R., and Baehr, J. (2022): Skillful decadal prediction of German Bight storm activity. In: *Natural Hazards and Earth System Sciences*, 22, 3993-4009, DOI: 10.5194/nhess-22-3993-2022

SKILLFUL DECADAL PREDICTION OF GERMAN BIGHT STORM ACTIVITY

Daniel Krieger[1,2], Sebastian Brune[3], Patrick Pieper[4], Ralf Weisse[1], Johanna Baehr[3]

[1]Institute of Coastal Systems – Analysis and Modeling, Helmholtz-Zentrum Hereon, Geesthacht, Germany
[2]International Max Planck Research School on Earth System Modelling, Hamburg, Germany
[3]Institute of Oceanography, Universität Hamburg, Hamburg, Germany
[4]Institute of Meteorology, Freie Universität Berlin, Berlin, Germany

Abstract

We evaluate the prediction skill of the Max Planck Institute Earth System Model (MPI-ESM) decadal hindcast system for German Bight storm activity (GBSA) on a multiannual to decadal scale. We define GBSA every year via the most extreme 3-hourly geostrophic wind speeds, which are derived from mean sea-level pressure (MSLP) data. Our 64-member ensemble of annually initialized hindcast simulations spans the time period 1960–2018. For this period, we compare deterministically and probabilistically predicted winter MSLP anomalies and annual GBSA with a lead time of up to 10 years against observations. The model produces poor deterministic predictions of GBSA and winter MSLP anomalies for individual years, but fair predictions for longer averaging periods. A similar but smaller skill difference between short and long averaging periods also emerges for probabilistic predictions of high storm activity. At long averaging periods (longer than 5 years), the model is more skillful than persistence- and climatology-based predictions. For short aggregation periods (4 years and less), probabilistic predictions are more skillful than persistence but insignificantly differ from climatological predictions. We therefore conclude that, for the German Bight, probabilistic decadal predictions (based on a large ensemble) of high storm activity are skillful for averaging periods longer than 5 years. Notably, a differentiation between low, moderate, and high storm activity is necessary to expose this skill.

A.1 Introduction

In low-lying coastal areas that are affected by mid-latitude storms, coastal protection and management may greatly benefit from predictions of storm activity on a decadal timescale. Decadal predictions bridge the gap between seasonal predictions and climate projections and may for example aid the planning of construction and maintenance projects along the coast. The German Bight in the southern North Sea

represents an example of such an area, where the coastlines are heavily and frequently affected by mid-latitude storms.

Climate projections suggest that many components of the Earth system undergo changes that can be attributed to the anthropogenic global warming (IPCC, 2021). For certain types of extreme events, like heavy precipitation or heat extremes, a link between the frequency of occurrence and the change in Earth's temperature has already been established (e. g., Lehmann et al., 2015; Suarez-Gutierrez et al., 2020; Seneviratne et al., 2021). For storm activity, studies for the past century showed a lack of significant long-term trends over the northeast Atlantic in general and the German Bight in particular. Instead, storm activity in this region is subject to a pronounced multidecadal variability (Schmidt and von Storch, 1993; Alexandersson et al., 1998; Bärring and von Storch, 2004; Matulla et al., 2008; Feser et al., 2015; Wang et al., 2016; Krueger et al., 2019; Varino et al., 2019; Krieger et al., 2021). This dominant internal variability suggests a great potential for improved predictability in moving from uninitialized emission-based climate projections towards initialized climate predictions. In this study, we demonstrate that initialized climate predictions are useful to predict German Bight storm activity (GBSA) on a multiannual to decadal timescale.

There have been considerable advancements in the field of decadal predictions of climate extremes in recent years. For example, the research project *MiKlip* (*Mittelfristige Klimaprognosen*; Marotzke et al., 2016) focused on the development of a global decadal prediction system based on the Max Planck Institute Earth System Model (MPI-ESM) under CMIP5 forcing. Using experiments from the MiKlip project, Kruschke et al. (2014) and Kruschke et al. (2016) found significant positive prediction skill for cyclone frequency in certain regions of the North Atlantic sector and for certain prediction periods, even for ensembles of 10 or fewer members. While Kruschke et al. (2016) used a probabilistic approach to categorize cyclone frequency into tercile-based categories, they did not explicitly assess the skill of the model for each category separately. Haas et al. (2015) found significant skill in MPI-ESM for upper quantiles of wind speeds at lead times of 1–4 years but also noted that the skill decreases with lead time and is lower over the North Sea than over the adjacent land areas of Denmark, Germany, and the Netherlands. Moemken et al. (2021) confirmed the capability of a dynamically downscaled component of the MiKlip prediction system for additional wind-related variables, such as winter season wind speed and a simplified winter season storm severity index (e. g., Pinto et al., 2012). However, Moemken et al. (2021) noted that wind-based indices are usually less skillful than variables based on temperature or precipitation, and are also heavily lead-time-dependent (Reyers et al., 2019). Furthermore, the prediction skill of wind-based indices shows strong spatial variability, which prevents any generalization of the current state of prediction capabilities for regionally confined climate extremes.

In addition to the high variability in the decadal prediction skill for wind-based indices, the depiction of near-surface wind in models strongly depends on the selected parameterization. Therefore, we circumvent the use of a wind-based index for evaluating the prediction skill for regional storm activity, and focus on a proxy that is based on horizontal differences of mean sea-level pressure (MSLP) and the resulting mean geostrophic wind speed instead. The index was first proposed by Schmidt and von Storch (1993) to avoid the use of long-term wind speed records, which oftentimes show inhomogeneities due to changes in the surroundings of the measurement site, and has already been used to reconstruct historical storm activity in the German Bight

(e. g., Schmidt and von Storch, 1993; Krieger et al., 2021). The geostrophic storm activity index is based on the assumption that the statistics of the geostrophic wind represent the statistics of the near-surface wind, which was confirmed by Krueger and von Storch (2011). The validity of the assumption is especially given over flat surfaces, like the open sea, where disturbances from friction are negligible. We therefore draw on the finding that the geostrophic wind-based index represents a suitable proxy for near-surface storm activity and can be used to derive some of the most relevant statistics of storm activity in the German Bight. Furthermore, the index is particularly well suited for small regions, since calculating the MSLP gradient over a small area allows for the detection of small-scale variability in the pressure field, which is crucial for estimating geostrophic wind statistics.

Besides the choice of variables, the ensemble size also plays an important role in decadal prediction systems. The experiments performed in MiKlip consisted of up to 10 members in the first two model generations and 30 members in the third generation (Marotzke et al., 2016). Sienz et al. (2016) showed that larger ensembles generally result in better predictability, especially in areas with low signal-to-noise ratios. However, Sienz et al. (2016) also noted the number of ensemble members alone does not compensate for other potential shortcomings of the model. In a more recent study, Athanasiadis et al. (2020) found that larger ensemble sizes increase the decadal prediction skill for the North Atlantic Oscillation and high-latitude blocking. Furthermore, the use of a large ensemble increases the reliability of probabilistic predictions. The concept of a probabilistic approach is the presumption that a change in the shape of the ensemble distribution can be used to predict likelihoods of actual changes in climatic variables. In contrast to deterministic predictions, probabilistic predictions are also able to provide uncertainty information. With increasing ensemble size and a resulting higher count of members in the tails of the predictive distribution, probabilistic predictions for extreme events, i. e., periods with very high or low storm activity, become feasible (e. g., Richardson, 2001; Mullen and Buizza, 2002). Therefore, we build on these findings by increasing the ensemble size in this study to a total of 64 members.

In this study, we assess the prediction skill for GBSA of a 64-member ensemble of yearly initialized decadal hindcasts, i. e., forecasts for the past, based on the MPI-ESM. Since GBSA is connected to the large-scale circulation (Krieger et al., 2021), we first analyze the ability of the decadal prediction system (DPS) to deterministically predict large-scale MSLP in the North Atlantic by comparing model ensemble mean output to data from the ERA5 reanalysis (Hersbach et al., 2020) (Sect. A.3.1.1). In the German Bight, most of the annual storm activity can be attributed to the winter season. Therefore, we focus on the winter (December–February, DJF) mean MSLP and quantify the quality of deterministic predictions by correlating time series of predictions (ensemble mean) and observations. We show how positive correlations emerge in predictions of both winter MSLP and GBSA (Sect. A.3.1.2). We then evaluate the skill of the DPS for probabilistic predictions of MSLP and GBSA (Sect. A.3.2.1 and A.3.2.2), expressed via the Brier skill score (BSS; Brier, 1950), and discuss the advantages and limits of our approach (Sect. A.3.3). Concluding remarks are given in Sect. A.4.

A.2 Methods and data

A.2.1 The observational reference

We use the time series of annual GBSA from Krieger et al. (2021) as an observational reference for the evaluation of prediction skill. The time series is based on standardized annual 95th percentiles of geostrophic wind speeds over the German Bight. The geostrophic winds are derived from triplets of 3-hourly MSLP observations at eight measurement stations at or near the North Sea coast in Germany, Denmark, and The Netherlands. MSLP measurements are provided by the International Surface Pressure Databank (ISPD) version 3 (Compo et al., 2015; Cram et al., 2015), as well as the national weather services of Germany (Deutscher Wetterdienst; DWD, 2019), Denmark (Danmarks Meteorologiske Institut; Cappelen et al., 2019), and the Netherlands (Koninklijk Nederlands Meteorologisch Instituut; KNMI, 2019). The time series of German Bight storm activity derived from observations covers the period 1897–2018.

Furthermore, we employ data from the ERA5 reanalysis (Hersbach et al., 2020), which has recently been extended backwards to 1950. The reanalysis data enables the prediction skill assessment over areas where in situ observations are incomplete or too infrequent, for example over the North Atlantic Ocean.

A.2.2 MPI-ESM-LR decadal hindcasts

We investigate the decadal hindcasts of the MPI-ESM coupled climate model in version 1.2 (Mauritsen et al., 2019), run in low-resolution (LR) mode. The MPI-ESM-LR consists of coupled models for ocean and sea ice (MPI-OM; Jungclaus et al., 2013), atmosphere (ECHAM6; Stevens et al., 2013), land surface (JSBACH; Reick et al., 2013; Schneck et al., 2013), and ocean biogeochemistry (HAMOCC; Ilyina et al., 2013). As we investigate the predictability of storm activity, which is derived from mean sea-level pressure, we focus on the atmospheric output given by the atmospheric component ECHAM6. The LR mode of ECHAM6 has a horizontal resolution of 1.875° (T63 grid), as well as 47 vertical levels between 0.1 hPa and the surface (Stevens et al., 2013). The horizontal extent of the grid boxes is approximately 210 km × 210 km at the Equator and 125 km × 210 km over the German Bight, which is still fine enough for the German Bight to cover multiple grid points. The model is forced by external radiative boundary conditions, which correspond to the historical CMIP6 forcing until 2014, and the SSP2–4.5 scenario starting in 2015 (contrary to CMIP5 and the RCP4.5 scenario used in the MiKlip experiments).

The ensemble members are initialized every 1 November from 1960 to 2019. The initialization and ensemble generation scheme is based on a system developed and tested within MiKlip (the “EnKF” system in Polkova et al. (2019)). For our study it has been updated from CMIP5 to CMIP6 external forcing and extended from 16 to 80 ensemble members. The basis of this scheme is formed by a 16-member ensemble assimilation, which from 1958 to 2019 assimilates the observed oceanic and atmospheric state into the model (Brune and Baehr, 2020). In particular, an oceanic ensemble Kalman filter is used with an implementation of the Parallel Data Assimilation Framework (Nerger and Hiller, 2013), and atmospheric nudging is applied. All 80 ensemble members of the predictions are directly initialized from the 16-ensemble member assimilation, with five different perturbations applied to the horizontal diffusion coefficient in the upper stratosphere to generate the

total amount of $5 \times 16 = 80$ ensemble members. For example, hindcast members 1, 17, 33, 49, 65 are all initialized from assimilation member 1, but with different perturbation in the upper stratosphere (no perturbation for member 1, four different non-zero perturbations for the other members). Since we require 3-hourly output (see Sect. A.2.2.2), which is not available for the first 16 members of the 80-member ensemble, we constrict our analysis to the remaining 64 members. In the following, we will refer to these members as members 1–64. Due to the observational time series of German Bight storm activity from Krieger et al. (2021) ending in 2018, we only evaluate hindcast predictions until 2018. For example, the last run considered in the evaluation for lead year 10 predictions is the one initialized in 2008, whereas the lead year 1 evaluation takes all runs initialized until 2017 into account.

A.2.2.1 Definition of lead times

All hindcast runs are integrated for 10 years and 2 months, each covering a time span from November of the initialization year (lead year 0) to December of the 10th following year (lead year 10). For consistency, we only consider full calendar years for the comparison, leaving us with 10 complete years per intialization year and ensemble member. The 10 individual prediction years are hereinafter defined as lead year i, with i denoting the difference in calendar years between the prediction and the initialization. By this definition, lead year 1 covers months 3–14 of each integration, lead year 2 covers months 15–26, and so on. Lead year ranges are defined as time averages of multiple subsequent lead years i through j within a model run and are called lead years i–j in this study. To compare hindcast predictions for certain lead year ranges to observations, we average annual observations over the same time period (see Supplement for more details).

It should be noted that winter (DJF) means are always labeled by the year that contains the months of January and February. A DJF prediction for lead year 4 therefore contains the December from lead year 3 plus the January and February from lead year 4. Likewise, a DJF prediction for lead years 4–10 contains every December from lead years 3 through 9, as well as every January and February from lead years 4 through 10.

In this study, we aim at drawing general conclusions about the prediction skill for North Atlantic MSLP anomalies for long and short averaging periods. Therefore, we focus on lead years 4–10, as well as lead year 7, as examples for long and short averaging periods for the prediction skill for MSLP anomalies, respectively. The choice of lead years 4–10 is based on selecting a sufficiently long averaging period that is representative of the characteristics of multi-year averages. Lead year 7 is chosen as it marks the center year within the lead year 4–10 period. We would like to note that the choice of lead years 4–10 and 7 is arbitrary, but we also analyze other comparable lead year periods (e. g., 2–8 and 5) to ensure sufficient robustness of our conclusions. However, we refrain from explicitly showing results for every lead time for reasons of brevity. For German Bight storm activity, which does not contain spatial information, we show the skill for all combinations of lead year ranges.

A.2.2.2 Geostrophic wind and German Bight storm activity

For our analysis, we use 3-hourly MSLP over the North Atlantic basin, including the German Bight. As 3-hourly MSLP is only available as an output variable for the ensemble members 33–64, but not for 1–32, we use surface pressure p, surface geopotential Φ and surface temperature T output from the model and apply a height

correction. Following Alexandersson et al. (1998) and Krueger et al. (2019), the equation for the reduction of p to the MSLP p_0 reads

$$p_0 = p \cdot \left(1 - \frac{\Gamma \frac{\Phi}{g}}{T}\right)^{\frac{M \cdot g}{R \cdot \Gamma}}, \tag{A.1}$$

with the Earth's gravitational acceleration g =9.80665 m/s², the assumed wet-adiabatic lapse rate Γ =0.0065 K/m, the molar mass of air M =28.9647 g/mol and the gas constant of air R =8.3145 J/mol K. A consistency check between ensemble members 1–32 (manually reduced to sea level) and 33–64 (MSLP available as model output) resulted in negligible differences in MSLP (not shown). Therefore, we assume that the pressure reduction does not significantly influence our results and treat the entire 64-member ensemble as a homogeneous entity.

We generate time series of German Bight storm activity (GBSA) in the MPI-ESM-LR hindcast runs. Owing to the low resolution of the model, we choose the 3 closest grid points that span a triangle encompassing the German Bight (Fig. A.1). The coordinates of the selected grid points are specified in Table A.1. The grid points are selected so that the resulting triangle is sufficiently close to an equilateral triangle. This requirement is necessary to avoid a large error propagation of pressure uncertainties, which would cause a shift in the wind direction towards the main axis of the triangle (Krieger et al., 2021). We use 3-hourly MSLP data from the decadal hindcast ensemble at the three corner points of the triangle and derive geostrophic winds from the MSLP gradient on a plane through these three points, following Alexandersson et al. (1998).

Table A.1: Coordinates of the three grid points used for storm activity calculation in the model.

Grid point	Latitude (°N)	Longitude (°E)
North	55.02	9.38
West	53.16	5.63
Southeast	53.16	9.38

GBSA is defined as the standardized annual 95th percentiles of 3-hourly geostrophic wind speeds. For each combination of ensemble member, initialization year, and forecast lead year, we determine the 95th percentile of geostrophic wind speed (exemplarily shown for one combination in Fig. A.2). The percentile-based approach incorporates both the number and the strength of storms, thereby ensuring that both years with many weaker storms and years with fewer but stronger storms are represented as high-activity years. However, the proxy is not able to differentiate whether high storm activity is caused by a large number of storms or by their high wind speed. The annual 95th percentiles of geostrophic wind speed take on values between 18 and 29 m/s with an average of 22.87 m/s (Fig. A.3), which is close to the observational average of 22.19 m/s derived by Krieger et al. (2021) for the period 1897–2018.

We accomplish the standardization by first calculating the mean and standard deviation of annual 95th percentiles of geostrophic wind speeds from the runs initialized in 1960–2009 for lead year 1 and each member. We then subtract the means from the annual 95th percentiles, and divide by the standard deviations. Since the lead year 1

predictions started in 1960–2009 cover the period of 1961–2010, our standardization period matches the reference time frame used for storm activity calculation in Krieger et al. (2021). The resulting time series of lead year 4–10 and 7 ensemble mean predictions of GBSA, as well as the corresponding time series of observed GBSA, are shown exemplarily in Fig. A.9.

While the analysis of GBSA only uses MSLP data from three grid points in the German Bight, we also analyze the prediction skill for MSLP anomalies over the entire North Atlantic.

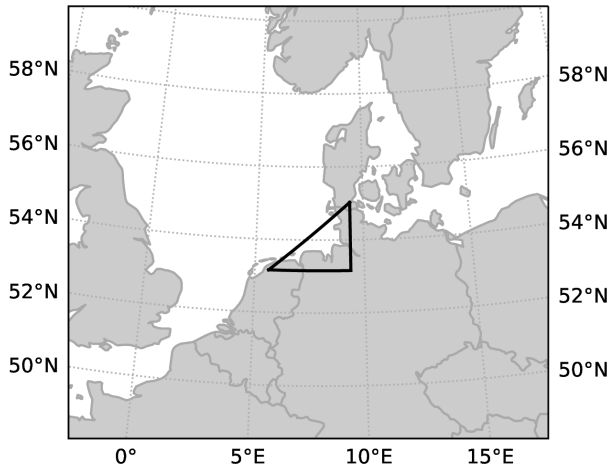


Figure A.1: Map of northwestern Europe, showing the location of the German Bight triangle.

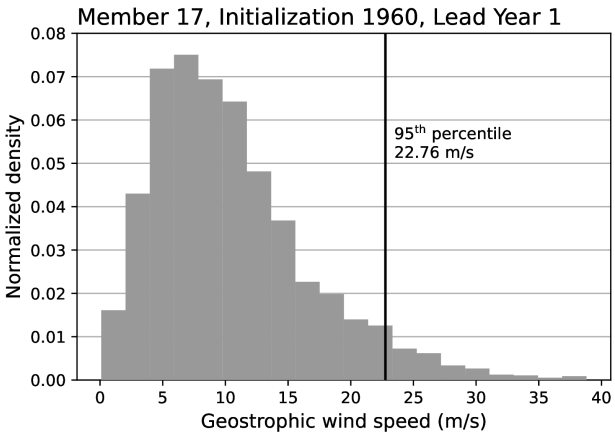


Figure A.2: Exemplary distribution of predicted 3-hourly geostrophic wind speeds for lead year 1 from member 17, initialized in 1960. The vertical line marks the 95th percentile, which is used in the calculation of storm activity.

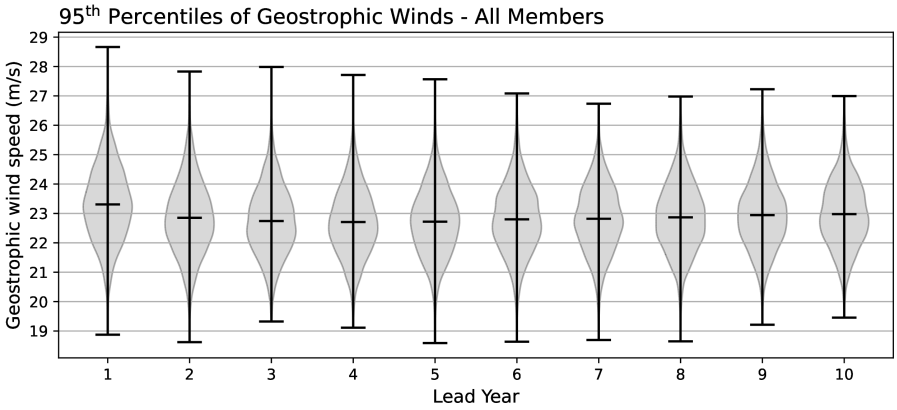


Figure A.3: Violin plot of the distribution of annual 95th percentiles of geostrophic wind speeds from all members and all initializations, separated by lead year. Lead years increase from left to right along the x axis. The width of the violin indicates the normalized density for a certain wind speed. Horizontal dashes mark maxima, means, and minima for each lead year.

A.2.3 Evaluation of model performance

In this study, we evaluate the model's performance for both deterministic and probabilistic predictions. First, we evaluate deterministic predictions to quantify the ability of the model to capture the variability in GBSA. Second, we analyze probabilistic predictions to examine whether the large ensemble is able to skillfully differentiate between extremes and non-extremes. These two prediction types require different evaluation metrics.

A.2.3.1 Anomaly correlation

For deterministic predictions, we calculate Pearson's anomaly correlation coefficient (ACC) between predicted and observed quantities:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N}(f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{N}(f_i - \bar{f})^2 \sum_{i=1}^{N}(o_i - \bar{o})^2}}, \tag{A.2}$$

with the predicted and observed quantities f_i and o_i, as well as the long-term averages of predictions and observations $\bar{f}$ and $\bar{o}$. The ACC can take on values from 1 to -1, with 1 indicating a perfect correlation, 0 equating to no correlation, and -1 showing a perfect anticorrelation. The statistical significance of the ACC is determined through a 1000-fold moving block bootstrapping with replacement (Kunsch, 1989; Liu and Singh, 1992), where the 0.025 and 0.975 quantiles of bootstrapped correlations define the range of the 95 % confidence interval. The block length is set to $k = 4$, following the suggestion of $k = \mathcal{O}(N^{\frac{1}{3}})$ (Lahiri, 2003) for a number of data points N between 50 and 60, depending on the variable and the length of the averaging period. The mean ACC is calculated by applying a Fisher z transformation (Fisher, 1915) to the bootstrapped correlations, averaging over all values in z space, and transforming the average back to the original space. The transformation of correlations to z scores z and its inverse are defined as $z = \mathrm{arctanh}(\mathrm{ACC})$ and $\mathrm{ACC} = \tanh(z)$, where $\tanh$ and $\mathrm{arctanh}$ are the hyperbolic tangent function and its inverse, respectively.

A.2.3.2 Brier skill score

Probabilistic predictions are evaluated against a reference prediction (see Sect. A.2.5) by employing the strictly proper Brier skill score (BSS, Brier, 1950). The BSS is a skill metric for dichotomous predictions and is defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{A.3}$$

where BS and $\mathrm{BS}_{\mathrm{ref}}$ denote the Brier Scores of the probabilistic model prediction and a reference prediction, respectively. This definition results in positive BSS values whenever the model performs better than the chosen reference and negative values when the reference outperforms the model. A perfect prediction would score a BSS of 1. The statistical significance of the BSS is calculated through a 1000-fold bootstrapping with replacement. We perform the bootstrapping in temporal space by selecting random blocks with replacement but do not bootstrap across the ensemble space. In this study, we use a significance level of 5 % to test whether model performance is significantly different from the reference.

The Brier Score BS is defined as

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2, \tag{A.4}$$

with the number of predictions N, the predicted probability of an event F_i and the event occurrence O_i. The predicted probability F_i is determined by the number of ensemble members that predict the event divided by the total ensemble size of 64. Note that O_i always takes on a value of either 1 or 0, depending on whether the event happened or not. Because the BS is calculated as the normalized mean square error in the probability space, it is negatively oriented with a range of 0 to 1, i. e., better predictions score lower BS values. A prediction based on flipping a two-sided coin ($F_i = 0.5$) would score a BS of 0.25.

We are interested in the skill of probabilistic predictions of periods of high, moderate, and low storm activity, as well as high, moderate, and low winter MSLP anomalies. To differentiate between events and non-events, the BS needs thresholds, which we set to 1 and -1. We define high-activity periods as time steps above 1, low-activity periods as time steps below -1, and moderate-activity periods as the remaining time steps. Since the BSS can only assess the skill of dichotomous predictions, we evaluate each of the three respective categories (high, moderate, low) separately. This methodology differs from Kruschke et al. (2016), as we do not evaluate one three-category forecast but three two-category forecasts instead.

A.2.4 Re-standardization of multi-year averages

Winter MSLP anomalies and GBSA time series are standardized before the analysis. To keep the evaluation of multi-year averaging periods consistent with that of single lead years, we re-standardize all time series after applying the moving average. We do this since the thresholds of our probabilistic prediction categories require the underlying data to be normally distributed with a mean of 0 and a standard deviation of 1 by definition. For spatial fields, we perform the standardizations and skill calculations grid-point-wise. As GBSA is based on the mean MSLP gradient of a plane through three grid points, we treat its spatial information like that of a single grid point and calculate skill metrics only once for the entire plane.

A.2.5 Reference forecasts

The BSS evaluates the skill of probabilistic predictions against a reference prediction. In this study, we use both a deterministic persistence prediction and a probabilistic climatological random prediction as a baseline against which we test the prediction skill of the MPI-ESM-LR, which is a common practice in climate model evaluation (e. g., Murphy, 1992).

The deterministic persistence prediction of storm activity is generated by taking the average observed storm activity of n years before the initialization year of the model run. n is defined to be equal to the length of the predicted lead year range. For example, a lead year 4–10 prediction ($n = 7$) initialized in 1980 is compared to the persistence prediction based on the observed average of the years 1973–1979, whereas a lead year 7 prediction ($n = 1$) from the same initialization is compared to the persistence prediction based on the observed storm activity of 1979. Persistence predictions of winter MSLP are generated likewise but use ERA5 reanalysis data instead of direct observations. We note that since the persistence prediction is not probabilistic, it can either be correct or incorrect in a given year, which corresponds to the term $(F_i - O_i)$ in Eq. A.4 taking on a value of either 0 (correct) or 1 (incorrect).

The probabilistic climatological random prediction uses the climatological frequencies of observed events (e. g., Wilks, 2011). As our time series of winter MSLP anomalies and GBSA are normally distributed by definition, the climatological frequencies can be derived from the Gaussian normal distribution. For instance, a climatological random prediction for high storm activity, which is defined via a threshold of 1 standard deviation above the mean, would always predict a fixed occurrence probability of $F_i = 1 - \Phi(1) = 0.1587$. Here, $\Phi(x)$ describes the cumulative distribution function of the normal distribution. $\Phi(x)$ gives the probability that a sample drawn from the Gaussian normal distribution at random is smaller or equal to $\mu + x\sigma$, with μ and σ denoting the mean and standard deviation of the distribution, respectively.

A.3 Results and discussion

A.3.1 Deterministic predictions

A.3.1.1 Mean sea-level pressure

Since geostrophic storm activity is an MSLP-based index, we first investigate the correlation between the model's deterministic predictions of winter (DJF) MSLP and data from the ERA5 reanalysis product, expressed as the grid-point-wise anomaly correlation coefficient (ACC). For lead year 4–10 winter MSLP anomalies, the ACCs are positive over larger parts of the subtropical Atlantic, as well as northeastern Canada and Greenland (Fig. A.4a). Negative ACCs emerge in a circular area west of the British Isles. Over the German Bight, however, the ACC for winter MSLP anomalies is insignificant. The pattern over the subtropical Atlantic Ocean agrees with the multi-model study by Smith et al. (2019), who found significant skill for winter MSLP in similar regions at lead years 2–9. Smith et al. (2019), however, also found skill over Scandinavia, where our DPS fails to provide any evidence of skill for long averaging periods. The ACC pattern of lead years 4–10 is also present for most other lead year ranges with averaging periods of 5 or more years (not shown).

For the single lead year 7, the ACC is negative over Scandinavia. Across the rest of the spatial domain, the absolute values of the ACC are lower for lead year 7 (Fig. A.4b) than for lead year 4–10, but the pattern shows some similarity. Again, the ACC is insignificant over the German Bight, indicating an insufficient skill to properly predict winter MSLP anomalies. The characteristics of the ACC distribution in Fig. A.4b also hold for other single lead years, suggesting that longer averaging periods generally result in higher absolute correlations, for regions with both positive and negative correlation values.

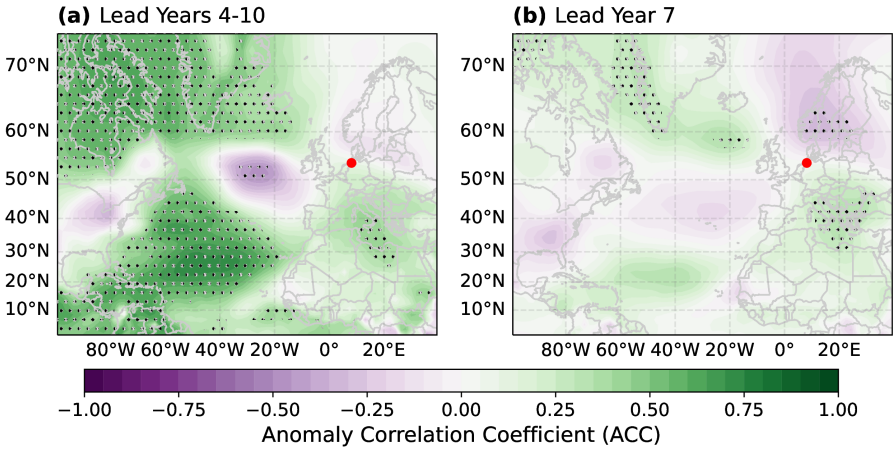


Figure A.4: Grid-point-wise anomaly correlation coefficient (ACC) between the deterministic hindcast ensemble mean prediction of winter-mean (DJF) MSLP anomalies and ERA5 reanalysis data for lead years 4–10 **(a)** and lead year 7 **(b)**. The German Bight is marked by a red dot. Anomalies are calculated for each member individually and averaged over the entire ensemble afterwards. Stippling indicates significant correlations ($p \leq 0.05$).

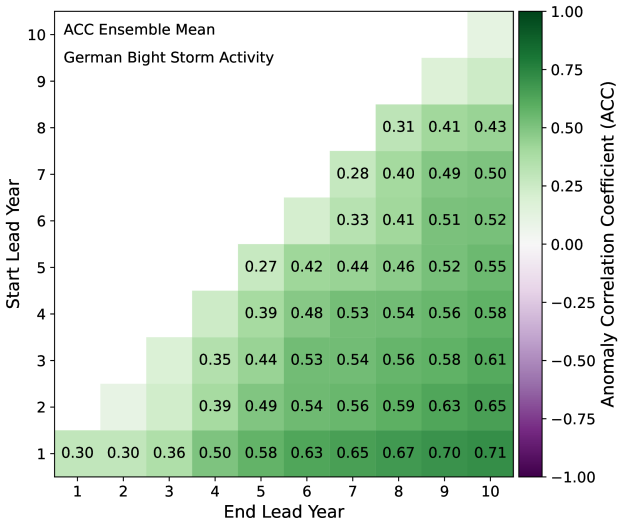


Figure A.5: Anomaly correlation coefficients between the deterministic DPS forecasts and observations of German Bight storm activity for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes indicate those correlation coefficients that are significantly different from 0 ($p \leq 0.05$).

A.3.1.2 Storm activity

We find that the ACC between ERA5 and DPS predictions for winter MSLP is significantly positive in certain regions of the North Atlantic, especially when averaged over multiple prediction years, but falls short of being significant over the German Bight. Still, the general predictive capabilities of the DPS for winter MSLP motivates the investigation of GBSA predictability. Fig. A.5 shows the deterministic predictability of GBSA, expressed as the ACC between the model ensemble mean and observations for all possible lead time combinations. Here, single lead years are displayed along the diagonal, while the length of the averaging period increases towards the bottom right corner. The ACC for GBSA is insignificant for most single prediction years (except for lead years 1, 5, 7, and 8), but it increases towards longer averaging periods. The ACC exhibits a clear dependence on the length of the averaging period, with lead years 1–10 showing the highest overall ACC among all lead year ranges ($r = 0.71$). Apart from lead years 2–3 and 9–10, the ensemble mean tends to become more skillful with longer averaging periods, and shows significant positive ACCs for all multi-year prediction periods. This stands in clear contrast to the results for winter MSLP predictions, where the model failed to produce significant ACCs for both short and long averaging periods in the German Bight (compare Fig. A.4).

Similar to the predictability of winter MSLP (Sect. A.3.1.1), we find a dependency of GBSA predictability on the length of the averaging window. Again, we argue that this may be caused by smoothing out the short-term variability that is apparent in reconstructed time series of annual GBSA (Krieger et al., 2021). However, the ACC is notably independent of the lead time. We would expect a deterioration of the ACC with increasing temporal distance from the initialization, i. e., along the diagonal in Fig. A.5. Instead, we observe a relative hotspot of predictability for lead year ranges of 2 to 4 years that start at lead year 3 and 4 (i. e., lead years 3–4 till 3–6 and 4–5 till 4–7). These ranges demonstrate higher predictability than comparable ranges closer to the present.

A.3.2 Probabilistic predictions

Since the deterministic predictions investigated so far are based on the ensemble mean, they do not take the ensemble spread into account. Therefore, we now make use of the large ensemble size to also generate probabilistic predictions for high-, moderate-, and low-storm-activity events, as well as high, moderate, and low winter MSLP anomaly events. We expect the DPS to be skillful in predicting probabilities since the large ensemble size allows us to detect changes in the shape of the ensemble distribution.

A.3.2.1 Mean sea-level pressure

When predicting positive winter MSLP anomalies (Fig. A.6a and A.6b), the DPS significantly outperforms persistence ($\mathrm{BSS} > 0$) over large parts of the central North Atlantic and Europe for both lead years 4–10 and 7. Over the North Sea, however, the BSS of the model is indistinguishable from 0 for lead years 4–10, indicating very limited skill to correctly predict positive winter MSLP anomalies. For lead year 7 predictions of positive winter MSLP anomalies, the BSS is slightly higher over the North Sea, with a higher model skill than that of persistence for most of the grid points. A similar pattern is found in predictions of negative anomalies (Fig. A.6c and A.6d), where the DPS does not show any additional skill compared to persistence

over the North Sea for lead years 4–10 but improves for lead year 7. Most notably, the DPS outperforms persistence in the far North Atlantic for lead years 4–10 but fails to do so in the subtropical North Atlantic.

Predictions of moderate winter MSLP anomalies (Fig. A.6e and A.6f) are skillful compared to persistence over most of the spatial domain. Still, a region of poor skill emerges over the German Bight and adjacent areas for lead year 4–10 predictions, while lead year 7 predictions show a BSS significantly higher than 0. The high BSS values of moderate anomaly predictions, however, are caused by poor performance of

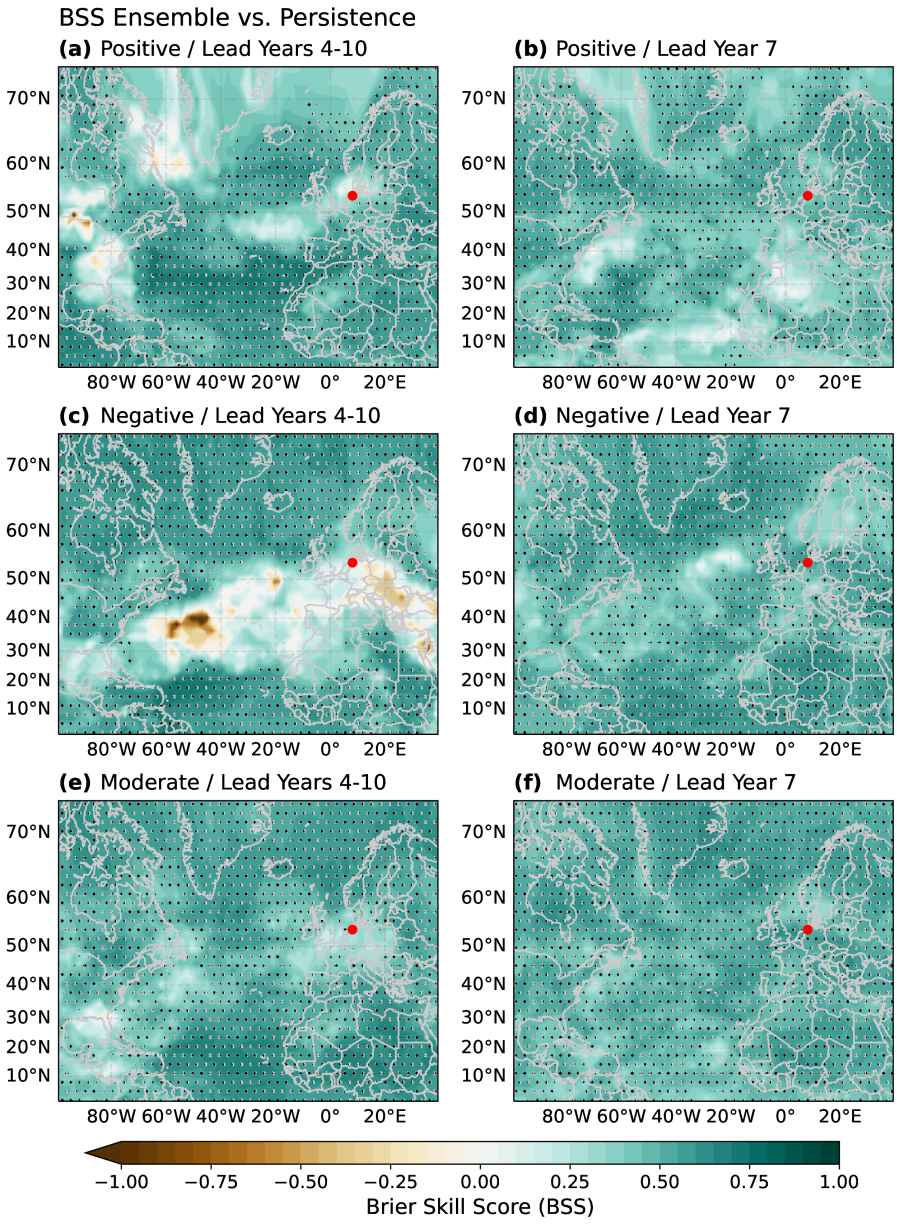


Figure A.6: Prediction skill of probabilistic forecasts of positive **(a,b)**, negative **(c,d)**, and moderate **(e,f)** winter mean (DJF) MSLP anomalies, expressed as the Brier skill score (BSS) of the 64-member ensemble evaluated against a persistence prediction as a baseline for lead years 4–10 **(a,c,e)** and lead year 7 **(b,d,f)**. Thresholds for event detection are set to -1 and 1. The German Bight is marked by a red dot. Stippling marks areas with a BSS significantly different from 0 ($p \leq 0.05$).

the persistence prediction serving as a reference. The BS of this reference prediction is significantly higher than 0.25 (not shown), demonstrating that persistence predictions are less skillful than a coin-flip-based prediction which assumes an occurrence probability of 50 % for every year. Hence, the BSS against persistence alone should not be used to infer the skill of the DPS for winter MSLP anomaly events.

Therefore, we additionally test the skill of the model for winter MSLP anomalies against that of a climatology-based prediction (Fig. A.7). The model BSS compared to climatology is mostly indistinguishable from 0 for both lead years 4–10 (Fig. A.7a, A.7c, and A.7e) and 7 (Fig. A.7b, A.7d, and A.7f), indicating a very limited potential of the DPS to outperform climatology over vast parts of the North Atlantic sector. Large patches of positive BSS values are found in lead year 4–10 predictions of negative winter MSLP anomalies over the tropical Atlantic (Fig. A.7c), whereas negative BSS values emerge over the polar North Atlantic for lead year 4–10 predictions of positive

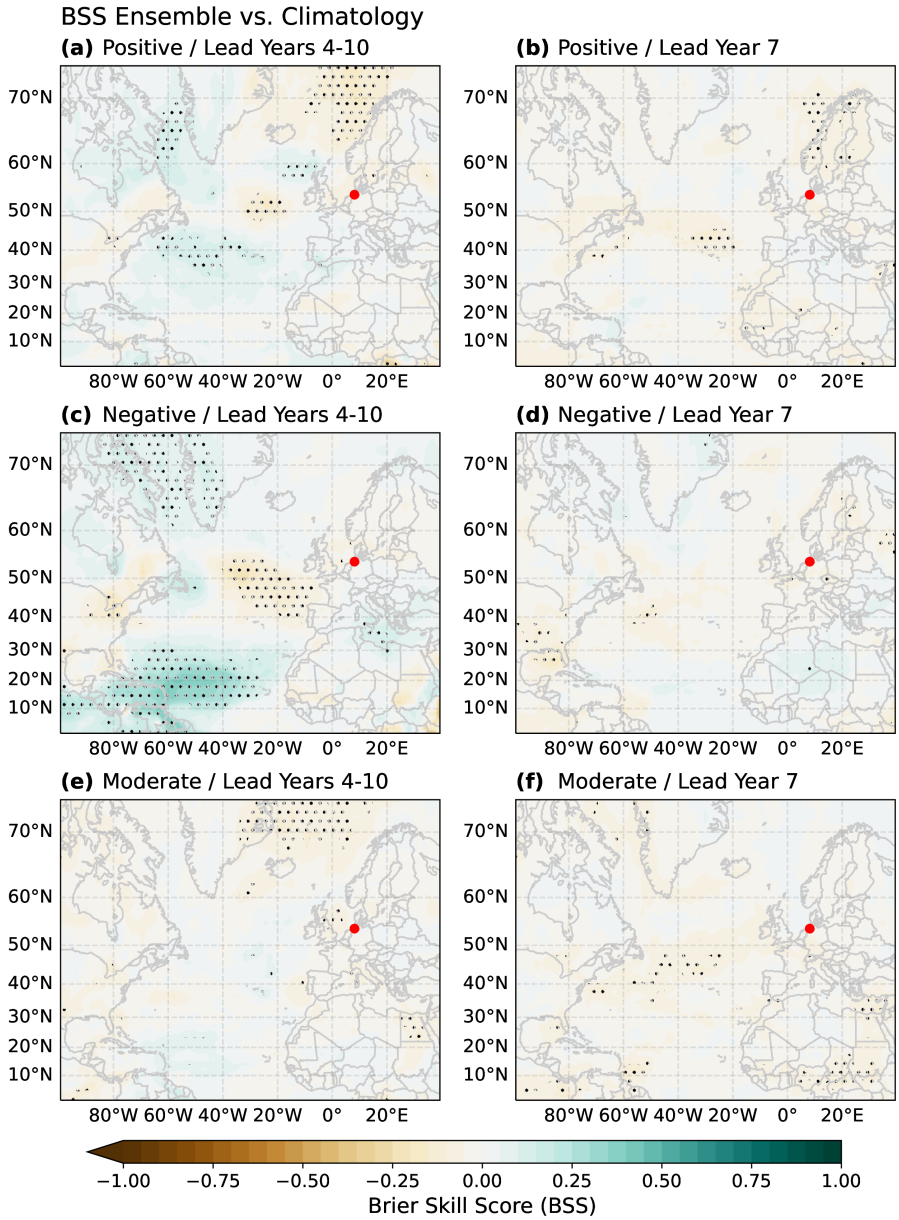


Figure A.7: Like Fig. A.6, but evaluated against a climatology-based prediction as a baseline.

and moderate winter MSLP anomalies (Fig. A.7a and A.7e), as well as over the central North Atlantic for lead year 4–10 predictions of negative winter MSLP anomalies (Fig. A.7c).

Overall, the DPS appears to predict positive and negative German Bight winter MSLP anomalies better than persistence for short averaging periods, while it fails to significantly outperform persistence for longer averaging periods. In addition, the DPS fails to consistently outperform climatology over large parts of the North Atlantic region for both short (lead year 7) and long (lead year 4–10) averaging periods. The comparison to climatology indicates that the high skill of the model when tested against persistence is caused by poor performance of the persistence prediction, rather than the prediction quality of the model. Nevertheless, the model shows some potential to bring additional value to the decadal predictability of winter MSLP anomalies.

A.3.2.2 Storm activity

The skill evaluation of probabilistic winter MSLP predictions shows that the BSS of the DPS for positive and negative anomalies are significantly better than those of persistence for large parts of the spatial domain. However, for long averaging periods, we do not observe a significant difference in skill between the DPS and persistence over the German Bight. Also, the model fails to outperform climatology for most parts of the North Atlantic sector. We now investigate the skill of probabilistic predictions of high-, moderate-, and low-storm-activity events, again using persistence and climatology as our baselines.

For high-storm-activity predictions, the BSS against persistence is positive for all lead year combinations, indicating a better performance of the DPS than persistence (Fig. A.8a). The BSS is significantly positive for most 1–2-year averaging windows, as well as for very long averaging windows (7 years or more). When testing the model's high-storm-activity predictions against a climatology-based forecast (Fig. A.8b), we find that the model exhibits significant skill for most averaging periods with a length of 4 or more years but shows no skill for short averaging periods. The distribution of significant BSS values among the lead year combinations against climatology differs strongly from the one obtained through testing against persistence (compare Fig. A.8a) and rather resembles the distribution of anomaly correlation coefficients between the deterministic predictions and observations (see Fig. A.5). Furthermore, the BSS against climatology is lower than against persistence for most lead year periods, indicating that climatology generally poses a tougher challenge for the model than persistence.

For low-storm-activity prediction (Fig. A.8c), the BSS is again positive for all lead year combinations. The BSS is significantly different from 0 for single-year and 3-year range predictions except for lead year 2 and lowest for averaging periods of 5–7 years. The higher BSS for single years than for periods of 5–7 years indicates that the model is valuable at predicting short periods. This behavior agrees with the findings in Sect. A.3.2, which significantly demonstrated positive skill for German Bight winter MSLP anomalies for a short period (lead year 7), but not for a multi-year average (lead years 4–10). However, the model only outperforms climatology (Fig. A.8d) for lead years 3–10, while all other lead years show insignificant BSS values. This suggests that while the model is able to beat a persistence-based prediction, it does not present any additional skill compared to climatology.

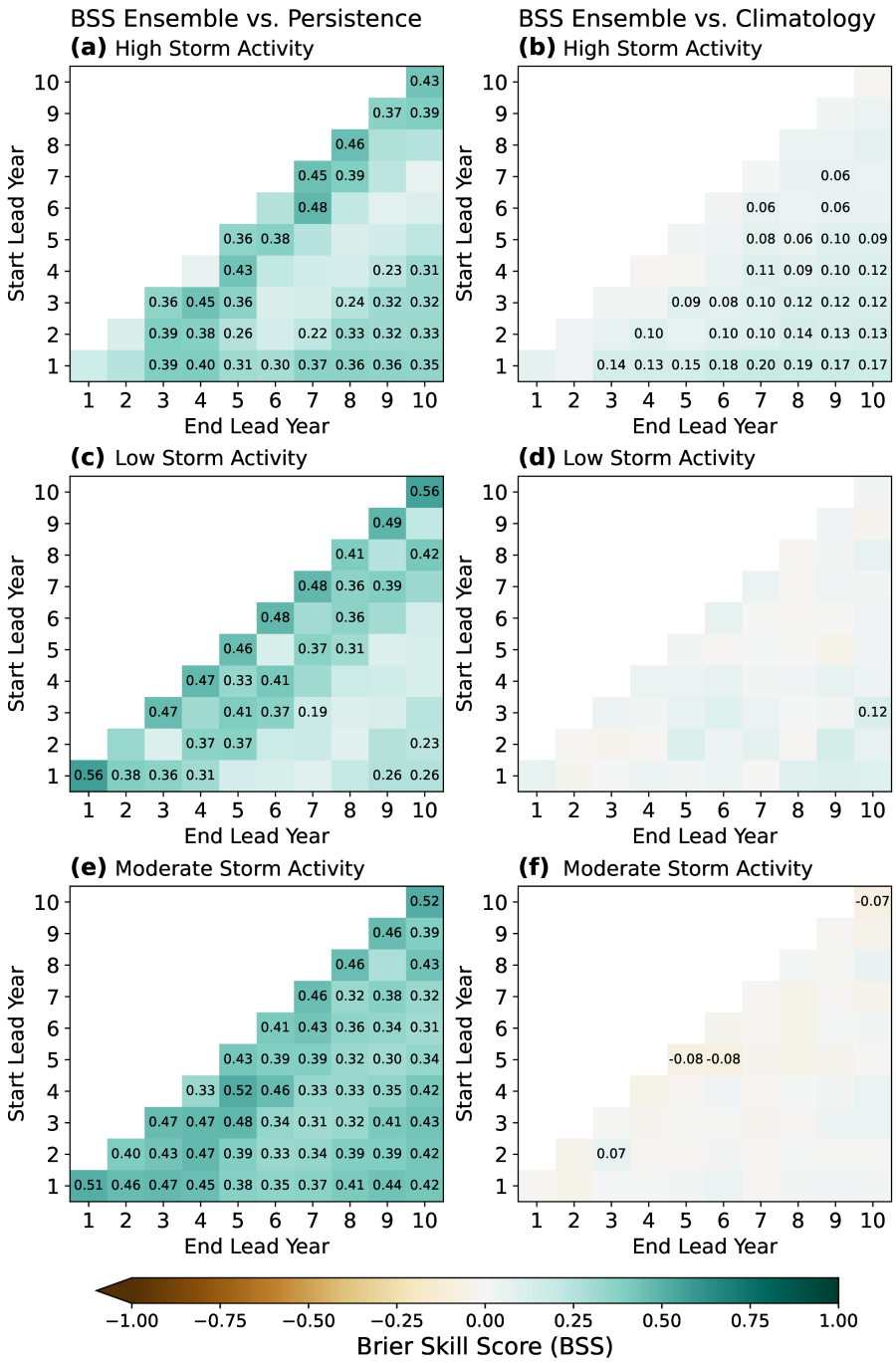


Figure A.8: Brier skill score (BSS) of the 64-member ensemble for high **(a,b)**, low **(c,d)**, and moderate **(e,f)** storm activity evaluated against both a persistence-based **(a,c,e)** and a climatology-based **(b,d,f)** prediction as a baseline, shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSSs that are significantly different from 0 ($p \leq 0.05$). Storm activity levels of 1 and -1 are used to differentiate between high, moderate, and low storm activity.

Moderate-storm-activity predictions (Fig. A.8e) also exhibit positive BSS values for all lead year ranges compared to persistence, and are significantly different from 0 except for lead years 8–9. However, this apparent high skill compared to persistence is once again only caused by the relative underperformance of the persistence prediction. A comparison with climatology (Fig. A.8f) confirms that the model significantly outperforms climatology for lead year 2–3 only and shows a reduced skill for lead years 5, 5–6, and 10, while it does not differ in skill for all remaining lead years.

Overall, the skill of the probabilistic forecast mostly depends on the choice of reference. While the model outperforms persistence over the majority of lead times in all three categories (high, moderate, low), it only outperforms climatology in predicting high storm activity for longer averaging windows. For probabilistic predictions of moderate and low storm activity, the model does not outperform climatology. Predictions of high storm activity with an averaging window of 6 or more years are the only ones where the model outperforms both climatology and persistence.

A.3.3 Discussion

We find that the ACC between deterministic predictions and observations of winter MSLP anomalies over large parts of the North Atlantic and GBSA is positive and significantly different from 0 for most multi-year averaging periods. Over the German Bight, however, ACCs for winter MSLP anomaly predictions are insignificant. We hypothesize that while the model is unable to deterministically predict winter MSLP anomalies over the German Bight, it is able to predict the annual upper percentiles of MSLP gradients sufficiently well for the ACCs of GBSA to become significant. This might be due to the model showing some predictive capabilities for sufficiently large deviations from the mean, but not for fluctuations around the mean.

The general lead year dependence of the magnitude of the ACC agrees with previous findings of Kruschke et al. (2014), Kruschke et al. (2016), and Moemken et al. (2021) for other storm activity-related variables. In our study, the correlation between reanalysis and prediction mainly depends on the length of the lead time window rather than the lead time (i. e., the temporal distance between the predicted point in time and the model initialization). We hypothesize that this dependency might be attributable to the filtering of high-frequency variability by the longer averaging windows, in combination with the model's ability to better predict the underlying low-frequency oscillation in the large-scale circulation. While our model is unable to deterministically predict the short-term variability within records of GBSA, these year-to-year fluctuations are smoothed out in predictions of multi-year averages, resulting in a higher ACC. Additionally, we would like to note that temporal autocorrelation might account for a part of these high ACC values. Smoothing that results from the multi-year averaging process introduces dependence to the time series which may lead to artificially inflated ACCs compared to non-smoothed time series.

The lack of a dependency of the ACC on the temporal distance from the initialization, however, cannot be explained by multi-year averaging. The relative hotspot of predictability for lead year ranges of 2 to 4 years starting at lead year 3 and 4 is counter-intuitive, especially due to the insignificant ACCs for lead years 2, 3, 4, and 2–3. These insignificant ACCs between GBSA observations and deterministic predictions hint at a possible initialization shock influencing the model performance. In fact, the average geostrophic wind speed for lead years 2, 3, and 4 is lower than for lead year 1 (Fig. A.3), supporting the hypothesis. Since all annual percentiles are

standardized using lead year 1 as a reference, we expect the resulting standardized storm activity for lead years 2, 3, and 4 to be slightly lower than for lead year 1. However, the average geostrophic wind speeds for lead years 5 through 10 are also lower than for lead year 1, yet the ACCs for these lead years are significant again. In addition, we tested whether standardizing each lead year with its respective mean and standard deviation (instead of always using lead year 1) has a notable effect on the ACC. We find that the ACC between model and observation is almost unaffected by the choice of our standardization reference (not shown). Hence, we rule out an initialization shock as the main reason for the low ACCs for lead years 2, 3, and 4. Beyond that, we are unable to come up with a convincing explanation for this behavior at this point. Thus, further studies are needed to investigate why the ACC does not steadily decline with increasing lead times.

For probabilistic predictions, the choice of reference plays a crucial role in the evaluation of the DPS. Since we test the performance of the model against that of persistence- and climatology-based predictions, the BSS not only depends on the prediction skill of the model but also on the skill of the reference. Most likely, a significant BSS is less a result of exceptional model performance but rather indicates the limits of persistence. This dependence becomes overtly apparent during the analysis of moderate GBSA predictability. Moderate GBSA predictability is skillful when evaluated against a persistent reference prediction. However, this significant prediction skill turns mostly insignificant when evaluated against a climatology-based prediction. On the contrary, we also find certain lead times where high-storm-activity predictions by the DPS beat climatology but fail to beat persistence.

The performance of persistence also contributes to the inverse dependency of the probabilistic skill on the length of the averaging window (i. e., a higher skill for shorter periods) that emerges in predictions of German Bight MSLP anomalies when tested against persistence. Here, the DPS exceeds the skill of persistence for short averaging periods but fails to do so for long averaging periods. This contradicts the assumption of the capability of the DPS to skillfully predict the underlying low-frequency variability (see Sect. A.3.1). However, the inverse dependency is more likely a result of better performance by the persistence prediction for longer averaging periods, which in turn challenges our model more than for short averaging periods. When evaluating probabilistic predictions of high GBSA against climatology, we find a similar dependency of the skill on the length of the averaging window as within deterministic predictions (i. e., a higher skill for longer periods), further confirming that the inverse dependency is an artifact of the performance of persistence.

Despite the aforementioned potential deficiencies, both persistence and climatology still range among the most appropriate references predictions to evaluate extreme GBSA predictability. We therefore conclude that our DPS is particularly valuable at lead times during which the reference forecasts are sufficiently poor. Vice versa, the benefits of a DPS are negligible at lead times during which the skill of the reference forecast is sufficiently fair. Naturally, we cannot determine in advance which of the two reference predictions will be more skillful at predicting GBSA. For most lead year periods, however, climatology poses a tougher challenge for the model than persistence, so we argue that outperforming climatology is an indication that the model can bring added value to GBSA predictability.

The separation of the probabilistic predictions into three categories also demonstrates the necessity to evaluate the skill for each prediction category individually. By individually assessing the skill for each forecast category, we find that the model is more

skillful than both persistence and climatology in predicting high-storm-activity periods for averaging windows longer than 5 years. We emphasize that evaluating three separate two-category forecasts is not as challenging to the model as incorporating all three categories into one aggregated skill measure (e. g., the ranked probability skill score, or RPSS; Epstein, 1969; Murphy, 1969, 1971). Yet, our analysis allows us to detect that our model shows skill in regions where previous studies that used a combined probabilistic skill score did not find any skill for storm-related quantities (e. g., Kruschke et al., 2016), a conclusion which would have not been possible to draw by evaluating a single three-category prediction.

Our results for probabilistic predictions suggest that our approach of employing a large ensemble notably aids the model's prediction skill. Contrary to previous studies on the decadal predictability of wind-related quantities, we find significant skill for high storm activity in the German Bight, especially for long averaging periods, where the model outperforms both persistence and climatology. The size of the ensemble might contribute to this skill, as similar analyses with smaller subsets of the DPS ensemble resulted in a slightly lower prediction skill (not shown), confirming the findings of Sienz et al. (2016) and Athanasiadis et al. (2020). However, the impact on prediction skill by a further increase in the number of members has yet to be investigated.

As this study is based on a single earth system model, the inherent properties of the MPI-ESM-LR might impact our findings. Thus, our conclusions drawn from these findings are only valid for this model. Model intercomparison studies for the decadal predictability of regional storm activity might eliminate the influence of possible model biases and errors. These intercomparisons will become possible once additional large-ensemble DPS products based on other earth system models are released.

It seems noteworthy that this study assumes annual storm activity and winter MSLP anomalies to be normally distributed, since the standardization process in the calculation of storm activity and winter MSLP anomalies fits a normal distribution to the data. Other distributions (e. g., a generalized extreme value distribution) might also be suited for a similar analysis and could provide an additional opportunity to enhance the description of storm activity and, thus, further improve the probabilistic prediction skill in the future.

A.4 Summary and conclusions

In this study, we evaluated the capabilities of a decadal prediction system (DPS) based on the MPI-ESM-LR to predict winter MSLP anomalies over the North Atlantic region and German Bight storm activity (GBSA), both for deterministic and probabilistic predictions. The deterministic predictions are based on the ensemble mean, whereas the probabilistic predictions evaluate the distribution of the 64 ensemble members. We assessed the anomaly correlation coefficient (ACC) between deterministic predictions and observations or reanalysis data, evaluated probabilistic predictions for three different forecast categories with the Brier skill score (BSS), and tested the probabilistic predictions of GBSA against both a persistence- and a climatology-based prediction.

Through comparison with data from the ERA5 reanalysis, we found that the DPS produces poor deterministic predictions of winter MSLP anomalies over the German Bight. Over the North Atlantic, certain regions with higher correlations emerge,

but the magnitude of the ACC is heavily dependent on the length of the averaging window. In general, longer averaging periods result in higher absolute correlations. The predictability for GBSA also depicts this same dependency on the averaging period, where ACCs are only significant for most averaging periods larger than 1 year.

Probabilistic predictions of winter MSLP anomalies over the North Atlantic are mostly skillful with respect to persistence, but do generally not show additional skill compared to climatology. For the German Bight in particular, only predictions for short lead year ranges are skillful with respect to persistence, while predictions for longer averaging periods exhibit poor skill.

For probabilistic predictions of high storm activity, averaging windows of 6 or more years are more skillfully predicted by the DPS than by both persistence and climatology. This study demonstrates that the model does bring an improvement to predictability of GBSA, and that a separation into multiple prediction categories is essential to detecting hotspots of predictability in the DPS which would have gone unnoticed in a more aggregated skill evaluation. Furthermore, we want to emphasize the ability of the DPS to especially issue reliable predictions for high storm activity, as this is arguably the most important category for which we could hope to achieve any prediction skill.

The high skill of probabilistic predictions for high storm activity, combined with the advantage of large-ensemble decadal predictions, can be expected to bring benefits to stakeholders, operators, and the society in affected areas by improving coastal management and adaptation strategies. By employing a large-ensemble DPS and carefully selecting a fitting prediction category, even regional climate extremes like GBSA can be skillfully predicted on multiannual to decadal timescales. With ongoing progress in the research field of decadal predictions and advancements in model development, we are therefore confident that this approach opens up new possibilities for research and application, including the decadal prediction of other regional climate extremes.

Appendix

Comparison of multi-year averages

In order to compare hindcast predictions for different lead year ranges to observations, we average hindcast predictions and observations over the same time periods. For example, a hindcast for lead years 4–10, which by definition is formed by averaging over a 7-year period, is always compared to a 7-year running mean of an observational dataset. The point-wise comparison of time series is performed in such a way that the predicted time frame matches the observational time frame. In other words, the lead year 4–10 prediction from a run initialized in 1960, which covers the years 1964–1970, is compared to the observational mean of 1964–1970. To form a time series from the model runs, the predictions from subsequent runs are concatenated. Thus, the predicted lead year 4–10 time series consists of a concatenation of predictions from the runs initialized in 1960, 1961, 1962, 1963, . . . , covering the years 1964–1970, 1965–1971, 1966–1972, 1967–1973, ...

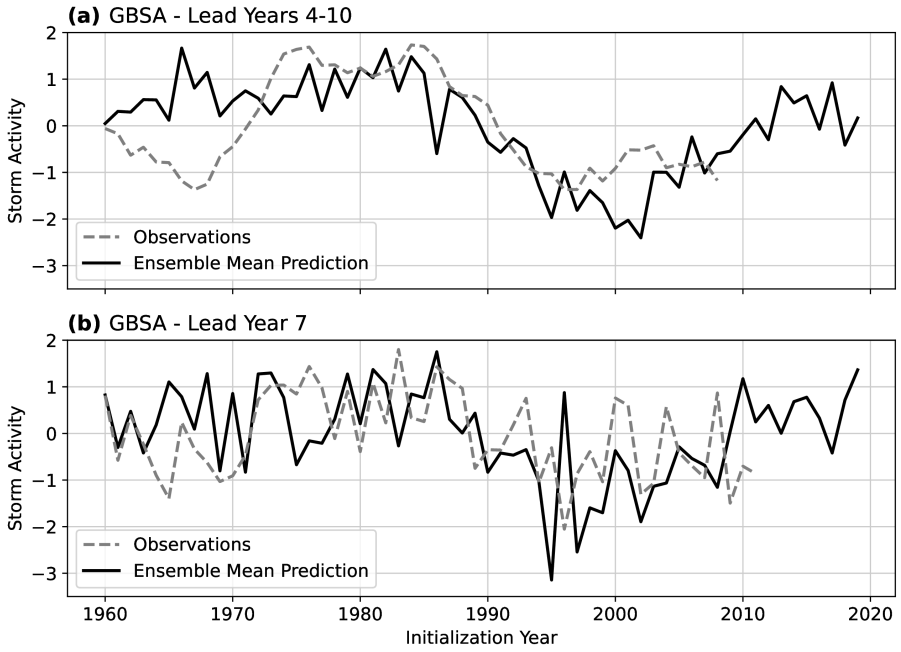


Figure A.9: Exemplary time series of ensemble mean predictions (black, solid) and corresponding observations (grey, dashed) of German Bight storm activity (GBSA) for lead years 4–10 **(a)** and lead year 7 **(b)**.

Acknowledgements

This work has been developed in the project WAKOS – Wasser an den Küsten Ostfrieslands. WAKOS is financed with funding provided by the German Federal Ministry of Education and Research (BMBF; Förderkennzeichen 01LR2003A). JB and PP were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2037 'CLICCS – Climate, Climatic Change, and Society' – Project Number: 390683824, contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg. JB and SB were supported by Copernicus Climate Change Service, funded by the EU, under contracts C3S-330, C3S2-370. We thank the German Computing Center (DKRZ) for providing their computing resources. The authors declare that they have no conflict of interest.

B

IMPROVING SEASONAL PREDICTIONS OF GERMAN BIGHT STORM ACTIVITY

The work in this chapter is currently under review and available as a preprint as:

Krieger, D., Brune, S., Baehr, J., and Weisse, R. (2023): Improving seasonal predictions of German Bight storm activity. *EGUsphere [preprint]*, DOI: 10.5194/egusphere-2023-2676

IMPROVING SEASONAL PREDICTIONS OF GERMAN BIGHT STORM ACTIVITY

Daniel Krieger[1,2], Sebastian Brune[3], Johanna Baehr[3], Ralf Weisse[1]

[1]Institute of Coastal Systems – Analysis and Modeling, Helmholtz-Zentrum Hereon, Geesthacht, Germany
[2]International Max Planck Research School on Earth System Modelling, Hamburg, Germany
[3]Institute of Oceanography, Universität Hamburg, Hamburg, Germany

Abstract

Extratropical storms are one of the major coastal hazards along the coastline of the German Bight, the southeastern part of the North Sea, and a major driver of coastal protection efforts. However, the predictability of these regional extreme events on a seasonal scale is still limited. We therefore improve the seasonal prediction skill of the Max-Planck-Institute Earth System Model (MPI-ESM) large-ensemble decadal hindcast system for German Bight storm activity (GBSA) in winter. We define GBSA as the 95th percentiles of three-hourly geostrophic wind speeds in winter, which we derive from mean sea-level pressure (MSLP) data. The hindcast system consists of an ensemble of 64 members, which are initialized annually in November and cover the winters of 1960/61–2017/18. We consider both deterministic and probabilistic predictions of GBSA, for both of which the full ensemble produces poor predictions in the first winter. To improve the skill, we observe the state of two physical predictors of GBSA, namely 70 hPa temperature anomalies in September, as well as 500 hPa geopotential height anomalies in November, in areas where these two predictors are correlated with winter GBSA. We translate the state of these predictors into a first guess of GBSA and remove ensemble members with a GBSA prediction too far away from this first guess. The resulting subselected ensemble exhibits a significantly improved skill in both deterministic and probabilistic predictions of winter GBSA. We also show how this skill increase is associated with better predictability of large-scale atmospheric patterns.

B.1 Introduction

The coastline of the German Bight, which is shared by the neighboring countries of Germany, Denmark, and the Netherlands, is frequently affected by strong extratropical cyclones and their accompanying hazards, such as storm surges. These extreme events repeatedly issue challenges to coastal protection agencies, emergency management, and other interests in the region. Therefore, local actors and stakeholders may

benefit from skillful predictions of these events on a seasonal-to-decadal scale. Still, skillful predictions of storm activity on a regional scale are a challenging task, even with today's state-of-the-art modeling capabilities. A recent study on the decadal predictability of German Bight storm activity (GBSA) has indicated that, with a carefully chosen approach, a large model ensemble, and an evaluation of different forecast categories, probabilistic predictions of high storm activity can be skillful for averaging periods longer than 5 years (Krieger et al., 2022). Krieger et al. (2022) also showed, however, that the predictive skill for single lead years in general and the next year in particular is often low and barely statistically significant, even when using a large-ensemble decadal prediction system. While Krieger et al. (2022) did not explicitly investigate the predictability of GBSA on a seasonal scale, the low skill for lead year 1 warrants an investigation into the seasonal predictability and its potential for improvement.

Previous studies have demonstrated that, on seasonal timescales, predictions for the state of large-scale modes of atmospheric variability like the North Atlantic Oscillation (NAO) can be improved through the use of known atmospheric and oceanic teleconnections (e. g., Dobrynin et al., 2018). These studies identified physical predictors that precede the desired predictand, and used first-guess predictions based on the state of the predictors to refine large model ensembles and thereby reduce model spread. Similar ensemble subselection techniques have also been used to increase the predictability of the European summer climate (Neddermann et al., 2019) and European winter temperatures (Dalelane et al., 2020). This technique, however, has not been applied to small-scale climate extremes like storm activity yet.

The storm climate of Central Europe, and in particular the German Bight, is connected to the large-scale atmospheric circulation in the Northern Hemisphere. GBSA has shown to correlate positively with the NAO, however the strength of this connection is subject to large fluctuations on a multidecadal scale. Other atmospheric phenomena during the winter season, such as the widely studied sudden stratospheric warmings, also play a role for the extratropical storm climate, since they influence the tropospheric weather regimes (e. g., Baldwin and Dunkerton, 2001; Song and Robinson, 2004; Domeisen et al., 2013, 2015) and are able to suppress or shift surface weather patterns in the mid-latitudes, sometimes even in a way that is contrary to the state of the NAO (Domeisen et al., 2020).

Peings (2019) found that a blocking pattern over the Ural region in November can be used to identify an increased likelihood of stratospheric warmings in the subsequent winter, which in turn favor blocking setups and thus lower-than-usual storm activity over Central Europe. Siew et al. (2020) confirmed this connection to be part of a troposphere-stratosphere causal link chain with a typical timescale of 2–3 months. The results of Peings (2019) and Siew et al. (2020) suggest that the status of the Rossby wave pattern in November might be usable as a predictor for the German Bight storm climate in the subsequent winter season.

The state of the stratospheric polar vortex in winter has also been linked to the Quasi-Biennial Oscillation (QBO) via the Holton-Tan effect (e. g., Ebdon, 1975; Holton and Tan, 1980). The Holton-Tan effect proposes a connection between easterly QBO phases, which are characterized by easterly wind and negative temperature anomalies in the lower stratosphere, and a weakened stratospheric polar vortex and thus positive stratospheric temperature anomalies in the polar Northern Hemisphere. The mechanism behind this effect has been widely studied and confirmed, e. g., by Lu et al. (2014). While some studies have already looked into the simultaneous occurrence of

QBO anomalies and shifts in the European winter climate and associated windows of opportunity for better predictability (e. g., Boer and Hamilton, 2008; Marshall and Scaife, 2009; Scaife et al., 2014b; Wang et al., 2018), the state of the tropical stratosphere has not been used as a predictor for the upcoming winter storm climate in Central Europe yet.

In this paper, we thus show that the predictability of German Bight storm activity on a seasonal scale is inherently low, but can be significantly improved through the combined use of tropospheric and stratospheric physical predictors. We use temperature anomalies in the lower tropical stratosphere in September, as well as extratropical geopotential height anomalies in the middle troposphere in November as predictors for GBSA. We generate first guesses of GBSA from these predictors and select members from our ensemble based on their proximity to the first guesses. From the large-ensemble prediction system with 64 members we generate both deterministic and probabilistic predictions of winter GBSA, both for the full and the subselected ensemble, and analyze the improvement of GBSA predictability through the subselection process. We demonstrate how, compared to the low prediction skill of the full ensemble, the subselection technique significantly increases the prediction skill. The large size of the ensemble also enables a thorough sensitivity analysis of the dependency of the skill on the subselection size.

B.2 Methods and data

B.2.1 Storm activity observations

As an observational reference for storm activity in the German Bight, we make use of the time series of winter GBSA from Krieger et al. (2021). The GBSA proxy in Krieger et al. (2021) is defined as the standardized 95th seasonal (December–February, DJF) percentiles of geostrophic winds. These geostrophic wind speeds were originally calculated from three-hourly observations of mean sea-level pressure along the German Bight coast in Denmark, Germany, and the Netherlands, and cover the period of 1897/98–2017/18.

B.2.2 MPI-ESM-LR decadal hindcasts

In this study, we employ the extended large-ensemble decadal hindcast system based on the Max Planck Institute Earth System Model (MPI-ESM) in low-resolution (LR) mode (Mauritsen et al., 2019; Hövel et al., 2022; Krieger et al., 2022). Even though this study focuses on the seasonal timescale, we choose decadal hindcasts over any seasonal prediction systems, as the already available MPI-ESM decadal hindcast system provides us with a large 64-member ensemble. At the time of this study, we are not aware of any single-model seasonal prediction system of this ensemble size and with three-hourly MSLP output available.

The MPI-ESM is a coupled climate model with individual components for the atmosphere (ECHAM6; Stevens et al., 2013), ocean and sea ice (MPI-OM; Jungclaus et al., 2013), land surface (JSBACH; Reick et al., 2013; Schneck et al., 2013), and ocean biogeochemistry (HAMOCC; Ilyina et al., 2013). Here, we only use the atmospheric output from the ECHAM6 component, which provides us with data at a temporal resolution of three hours, a horizontal resolution of 1.875°, as well as a vertical resolution of 47 levels between 0.1 hPa and the surface (Stevens et al., 2013). We use

all hindcast runs initialized between 1960 and 2017 as the observational reference time series of winter GBSA ends in 2017/18.

Since winter GBSA is not directly available as an output variable of the hindcast system, we derive it from the three-hourly MSLP output (Krieger and Brune, 2022). We calculate winter GBSA as the standardized seasonal (December–February) 95th percentiles of three-hourly geostrophic winds over the German Bight. The calculation follows the methodology of Krieger et al. (2022), however it uses seasonal instead of annual 95th percentiles. Doing so, we ensure that the calculation of GBSA in the hindcast is consistent with the derivation of observed GBSA in Krieger et al. (2021). We perform the GBSA calculations individually for every member of the hindcast.

B.2.3 Predictors of GBSA

In this study, we aim to increase the predictability of winter GBSA by refining a large ensemble by selecting individual members that are closest to a first-guess prediction of winter GBSA. To achieve this, we first need to define predictors and the generation of first guesses.

We use fields of September 70 hPa temperature (T_{70}) and 500 hPa geopotential height (Z_{500}) anomalies as our predictors for GBSA. The data for these fields are taken from the ERA5 reanalysis (Hersbach et al., 2020), which in its current state dates back to the year 1940. Anomalies are calculated by subtracting the 1940–2017 mean from the time series. We ensure that there are regions where the correlation coefficient between the predictor and GBSA is significantly different from zero over the whole investigation period (1940–2017 for predictors, winters 1940/41–2017/18 for GBSA).

In every prediction year, we generate a first guess of winter GBSA from the state of our chosen predictors. For each predictor x_p, we first analyze which gridpoints show a significant positive correlation with GBSA for all years from 1940 to the year before the initialization ($p \leq 0.05$). The statistical significance of the correlation is determined through a gridpoint-wise 1000-fold bootstrapping with replacement (Kunsch, 1989; Liu and Singh, 1992), where the 0.025 and 0.975 quantiles of bootstrapped correlations define the range of the 95 % confidence interval. If the 95 % confidence interval excludes a value of $r = 0$, we consider the correlation for this gridpoint significant and that gridpoint is taken into account for the generation of a first guess. As both the anomalies of the predictors and the index of winter GBSA are defined as standardized anomalies following a Gaussian normal distribution with a mean of 0 and a standard deviation of 1, we can directly translate the state of each predictor into a first guess of our predictand GBSA. Therefore, we compute the first guess of the predictand (GBSA) as an area-weighted average y_p of the state of the predictor x_p for those gridpoints (i, j) that are significantly positively correlated with GBSA, following Eq. B.1.

$$y_p = \frac{\sum_{i=1,j=1}^{sig} x_n(i,j)\cos\Phi_j}{\sum_{i=1,j=1}^{sig}\cos\Phi_j}. \tag{B.1}$$

In Eq. B.1, $\cos\Phi_j$ denotes the cosine of the latitude of each gridpoint used as a weighting factor. For geopotential height anomalies, we constrain the region that can contribute to the first guess to the boreal extratropics between 30°N-90°N, as the pattern of geopotential height in this region describes the Rossby wave train which strongly governs the extratropical winter storm climate. We make sure that each predictor always contributes significantly positively correlated gridpoints in every prediction year, as the correlation strength and location of the significant correlations may vary from year to year.

For every model run and predictor, we choose a number n of ensemble members in our forecast ensemble with a GBSA closest to the state of the respective predictor in that model run. Because we select n members twice in every run, i. e., once for every predictor, and the two selections of members might overlap, the size of this resulting subselection can vary between n and $2n$ members per run, depending on the distribution of the two respective first guesses derived from the two predictors. We then calculate deterministic and probabilistic GBSA predictions in the ensemble subselection. A schematic overview of the predictor-based subselection is given in Fig. B.1. It should be noted that members are weighted equally in all following computations, even though some of them might have been selected by multiple predictors. Deterministic predictions are computed by averaging the GBSA predictions over all members in the subselection. For probabilistic predictions, we calculate the fraction of members within the subselection that exceed a defined threshold for high storm activity of 1 standard deviation above the long-term mean.

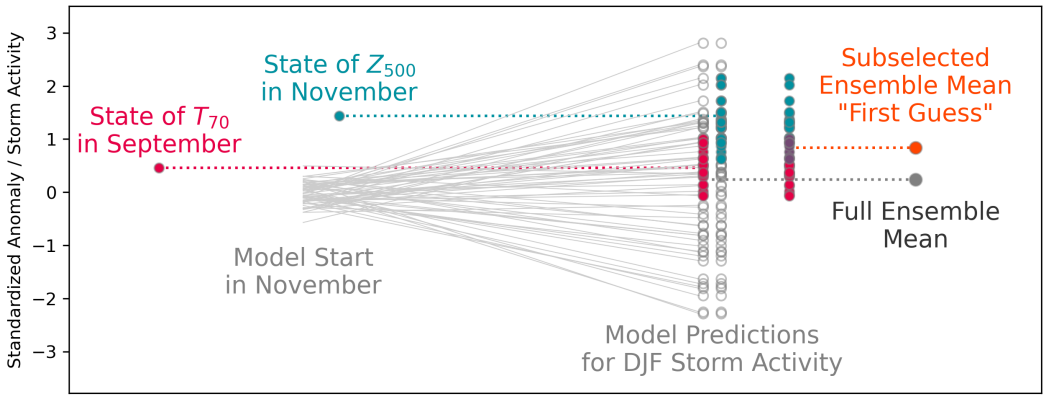


Figure B.1: Schematic depiction of the predictor-based subselection workflow, adapted from Dobrynin et al. (2018).

B.2.4 Skill metrics

To evaluate the improvement of prediction skill for winter GBSA, we first define separate skill metrics for deterministic and probabilistic model predictions.

We measure the skill of deterministic predictions with Pearson's anomaly correlation coefficient (ACC) and the root-mean-square error (RMSE) between predicted and observed quantities. The ACC is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N}(f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{N}(f_i - \bar{f})^2 \sum_{i=1}^{N}(o_i - \bar{o})^2}}, \tag{B.2}$$

where f_i and o_i denote predictions and observations at a time step i, and $\bar{f}$ and $\bar{o}$ mark the long-term averages of predictions and observations. ACC values of 1, 0, and -1 indicate a perfect correlation, no correlation, and a perfect anticorrelation, respectively. The statistical significance of the ACC is again determined through a 1000-fold bootstrapping with replacement and a significance criterion of $p \leq 0.05$.

The RMSE is calculated from the predicted and observed quantities f_i and o_i by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2}. \tag{B.3}$$

Probabilistic predictions of high storm activity are tested against a climatology-based reference prediction and evaluated with the strictly proper Brier skill score (BSS; Brier, 1950). The climatology-based reference prediction is constructed from the climatological frequencies of observed GBSA (e. g., Wilks, 2011). Here, we draw on the definition of GBSA from Krieger et al. (2021) which assumes an underlying Gaussian normal distribution.

We calculate the BSS as follows:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{cli}}}. \tag{B.4}$$

BS and $\mathrm{BS_{cli}}$ indicate the Brier Scores of the probabilistic model prediction and the fixed climatological reference prediction, respectively. Positive values show that the model predictions perform better than the climatology-based predictions and vice versa. A BSS of 1 would indicate a perfect model prediction, i. e., all members of the ensemble predicting the occurrence or absence of a high-storm-activity event correctly in every year.

The individual Brier Scores BS are defined as

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2, \tag{B.5}$$

where F_i and O_i denote predictions and observations at a time step i. In the model, we calculate the predicted probability F_i from the fraction of ensemble members that predict a high-storm-activity event. For the climatology-based prediction, F_i is a fixed value. As high storm activity is defined via a threshold of one standard deviation above the mean state, we calculate the climatological probability of a high-storm-activity event occuring to be $F_i = 1 - \Phi(1) = 0.1587$, where $\Phi(x)$ is the

cumulative distribution function of the Gaussian normal distribution. This means that the probability of a random sample from a Gaussian normal distribution with a mean of μ and a standard deviation of σ being larger than $\mu + 1\sigma$ is slightly less than 16 %. The observed probability O_i always takes on a value of either 1 or 0, depending on whether the event happened or not.

B.2.5 Training and hindcast periods

The recent backward extension of the ERA5 reanalysis extends the dataset back to 1940. Because the predictions of GBSA are based on predictors that are derived from regions where the predictor and GBSA correlate significantly, we require a sufficiently long training period to identify these regions. Hence, we classify the first two decades (1940–1959), for which only ERA5 and observational GBSA data are available, as the training-only period, and start the actual predictor-based first guesses of GBSA in the year 1960. Doing so, we can ensure that we only use data to predict GBSA that was already available at the starting point of the hindcast, but still use the full range of hindcasts which start in 1960. The hindcast period, i. e., the period in which we predict GBSA and assess the skill of the model and the subselection, is thus confined to a total of 58 winters from 1960/61 to 2017/18.

B.2.6 Composites

To check whether our prediction mechanism is also physically represented in the hindcast, we calculate composites of T_{70} and Z_{500} in the years with highest and lowest modeled DJF GBSA, respectively. We use all initialization years (1960–2019), all members (17–80) and all lead years except the first one after the initialization (2–10), leaving us with 34560 model years. From these 34560 years, we select the 100 highest and lowest GBSA winters, compute composite mean fields of both predictors in the respective years preceding these winters, and calculate the difference between the composites of high and low GBSA. We then analyze the patterns of the composite differences to determine whether they resemble the correlation patterns between the predictors in ERA5 and observed DJF GBSA.

B.3 Results

B.3.1 Correlations of predictor fields with winter storm activity

We identify T_{70} and Z_{500} anomalies as physical predictors for winter GBSA. To illustrate the connection between the global fields of these two predictors and storm activity, and to demonstrate which regions mainly contribute to the first-guess predictions, we correlate gridpoint-wise time series of T_{70} and Z_{500} anomalies with observed winter GBSA for the entire time period of 1940–2017.

The highest correlations between GBSA and T_{70} anomalies are found in the tropics in a circumglobal band between roughly 15°N and 15°S, with values as high as 0.5–0.6 (Fig. B.2). Notably, correlations are slightly lower directly at the equator than a few degrees north and south of it. Over Europe, a smaller region with slightly negative correlations is present, surrounded by slightly positive correlations to the northeast and northwest. Over the Southern Ocean, a signal of slightly negative correlations

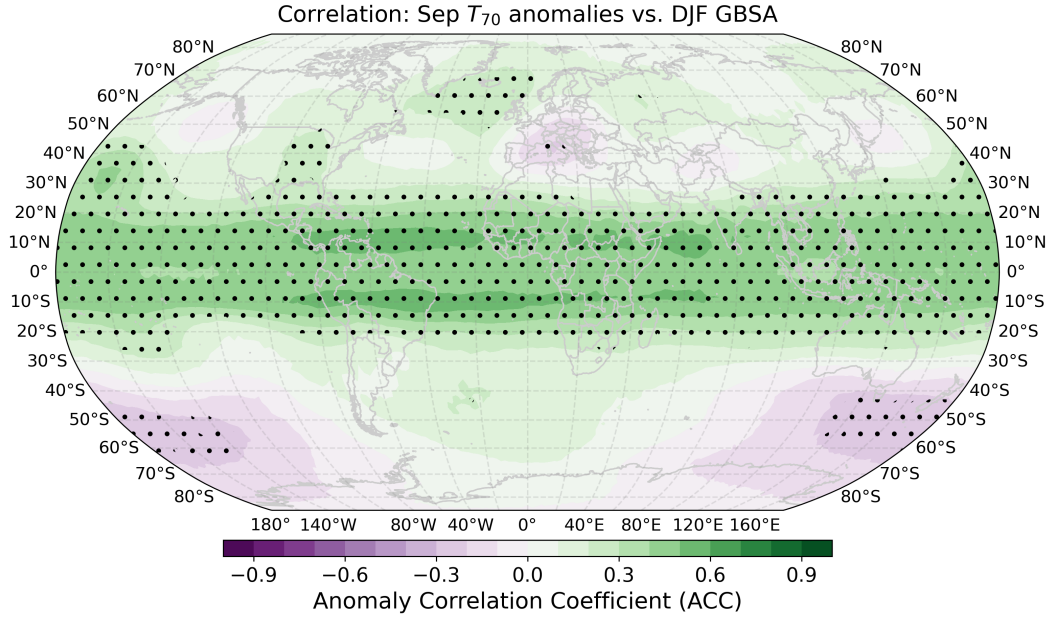


Figure B.2: Gridpoint-wise correlation coefficients between global T_{70} anomalies in ERA5 and observed winter (DJF) German Bight storm activity. Period 1940–2017 for temperature anomalies, 1940/41–2017/18 for storm activity. Stippling indicates statistical significance ($p \leq 0.05$) determined through 1000-fold bootstrapping.

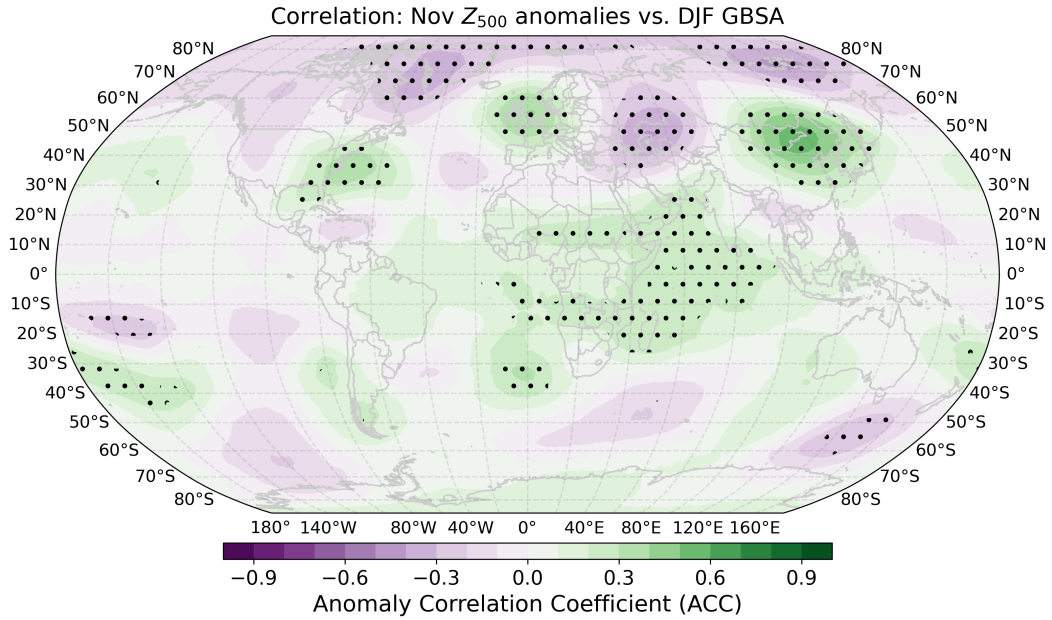


Figure B.3: Gridpoint-wise correlation coefficients between global Z_{500} anomalies in ERA5 and observed winter (DJF) German Bight storm activity. Period 1940–2017 for geopotential height anomalies, 1940/41–2017/18 for storm activity. Stippling indicates statistical significance ($p \leq 0.05$) determined through 1000-fold bootstrapping.

emerges as well. However, none of the regions outside of the tropics correlate with DJF GBSA as high as the tropics themselves.

For Z_{500} anomalies, the strongest positive correlations with winter GBSA are found over the British Isles and the adjacent Northeast Atlantic, as well as over East-central Asia and the North American East Coast with peaks around 0.4 (Fig. B.3). The strongest negative correlations emerge over East-central Europe, Greenland, and northeastern Siberia, reaching as low as -0.4. The correlation pattern in the boreal extratropics is in line with the findings of Peings (2019) and Siew et al. (2020) in a way that troughing (i. e., the opposite of ridging) over the Ural region and thus a reduced likelihood of stratospheric warmings in the following winter season is connected to higher-than-usual storm activity in the German Bight. Across the subtropical and tropical latitudes, some areas of slightly positive correlations can be found over the Indian Ocean. In the Southern Hemisphere, small patches of slightly positive and negative correlations are distributed circumglobally. However, the absolute correlations of the aforementioned regions in the tropics and the Southern Hemisphere are much lower than those in the northern extratropics, indicating that these correlations might be coincidental and spurious.

B.3.2 Improvement of GBSA predictability

We use the established connection between T_{70} and Z_{500} anomalies and DJF German Bight storm activity to predict the storm activity of the upcoming winter season at the end of November for the hindcast period of 1960–2017. We use latitude-weighted field means of T_{70} and Z_{500} in ERA5 as our initial guess for DJF storm activity. Since both the time series of temperature and geopotential height anomalies and those of GBSA are standardized, we do not need to apply a scaling factor to translate the field means of temperature and geopotential height anomalies to GBSA. We only use information from data between 1940 and the year of the start of the forecast. Thus, the amount and distribution of gridpoints that are included in the calculation of the first-guess prediction can vary from year to year. To generate first-guess predictions of winter GBSA, we need to select a certain number of ensemble members closest to the initial guess for each predictor.

One degree of freedom in this process is the sampling size, i. e., the number of members selected for each predictor. The choice of this sampling size has an effect on the skill metrics of the subselected ensemble predictions. To illustrate the dependency of the model skill on the sample size, we test the correlation, RMSE, and high-activity BSS against climatology for all sample sizes between 1 and 64 (Fig. B.4a) for the hindcast period of 1960–2017. Furthermore, we perform these sensitivity studies for both predictors individually to show how the combined use of both predictors changes the skill compared to just using one of the two (Fig. B.4b and B.4c).

The sensitivity analysis for the combined use of both predictors (Fig. B.4a) shows a strong increase in correlation to above 0.6 for up to roughly 50 members. This indicates that removing only about one sixth of all members per predictor is sufficient to increase the correlation between the deterministic prediction and observations significantly. The optimal sample size is found at 25 members per predictor ($r = 0.64$). For the RMSE, smaller sample sizes between 10 and 40 members yield the biggest improvement, with an optimum at 25 members ($\mathrm{RMSE} = 0.70$). The BSS can be maximized by selecting 25 members for each predictor as well ($\mathrm{BSS} = 0.28$), and shows a similar window of opportunity as the RMSE between 10 and 40 members.

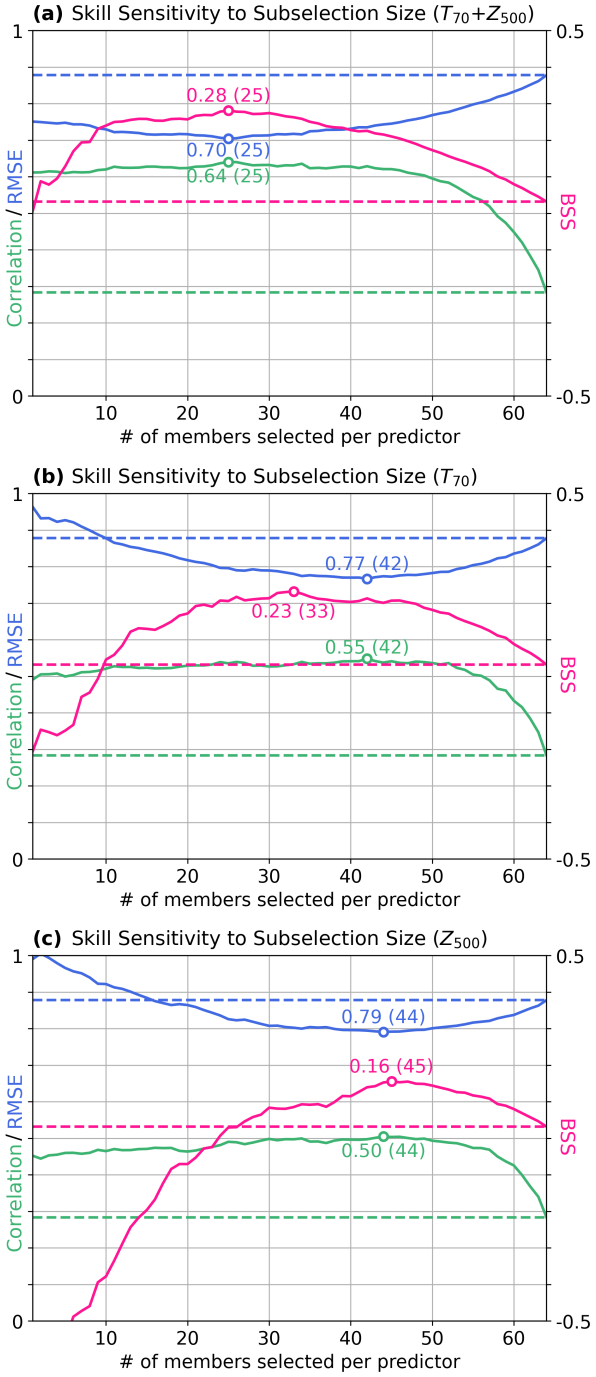


Figure B.4: Dependency of various skill scores (ACC (green), RMSE (blue), and BSS for high storm activity against climatology (pink)) of model ensemble predictions of DJF GBSA on the sample size chosen for each predictor during the subselection. The subselection is performed based on **(a)** both predictors, **(b)** only T_{70}, and **(c)** only Z_{500}. Dashed baselines show the respective skill scores of the full 64-member ensemble. Optimal skill scores (highest ACC and BSS, lowest RMSE) are displayed as annotated dots, together with the optimal sample size in brackets.

The sensitivity analysis for T_{70} alone (Fig. B.4b) reveals a slightly lower potential for probabilistic skill improvements. Here, the BSS can be increased to 0.23 with a sample size of 33 members, but a deterioration of the BSS compared to the full ensemble occurs below 10 members. Similarly, choosing $Z_{500,\mathrm{Sep}}$ alone (Fig. B.4c) only improves probabilistic forecasts when selecting more than 25 members with a maximum of 0.16 at 45 members.

The deterministic skill metrics also show similar windows of opportunity for both predictors. While correlation and RMSE for Z_{500} are maximized at sample sizes of 44 members ($r = 0.5$, $\mathrm{RMSE} = 0.79$) the optimum for T_{70} is located at 42 members ($r = 0.55$, $\mathrm{RMSE} = 0.77$). It should be noted that it is purely coincidental that, for both predictors, the optimal sample sizes for RMSE and correlation are equal. Just like for the BSS, the individual contributions of the predictors to correlation and RMSE are smaller than the combined effect, manifesting the need to combine multiple predictors in the subselection to achieve the best possible skill increase.

From the sensitivity study, we find that sample sizes of 20–30 members constitute a fair compromise between the optimal sample sizes of deterministic and probabilistic predictions. Therefore, we exemplarily analyze the prediction of winter GBSA in the hindcast period for a subselection size of 25 members per predictor in greater detail (Fig. B.5).

Over the forecast period, the first-guess estimates obtained from combining T_{70} and Z_{500} anomalies and observed winter GBSA correlate well (0.64), an improvement of 0.36 from the deterministic full-ensemble model prediction. The subselected ensemble captures the variability in DJF GBSA much better than the full 64-member ensemble. High agreements between first-guess predictions and observations are found in the late 1970s, the 1980s, as well as between the mid-1990s and the mid-2000s. With an RMSE of 0.70, the subselection-based prediction shows a slightly lower error than the full ensemble (0.88). Furthermore, the BSS against climatology of the reduced ensemble for high storm activity predictions is greatly increased to 0.28, compared to 0.03 for the full 64-member ensemble. In 39 out of the 58 individual predictions (67 %), the subselection leads to an improvement in the prediction as measured by the absolute difference between ensemble mean and observations.

Overall, all three metrics show a significant improvement for the first-guess-based reduced ensemble, revealing that both deterministic and probabilistic storm activity predictions can be significantly improved by the combined inclusion of T_{70} and Z_{500} as physical predictors.

B.3.2.1 Skill increase for large-scale atmospheric variables

In order to determine on a physical basis why the subselected ensemble shows a higher prediction skill for GBSA in both deterministic and probabilistic modes, we analyze the change in ACC between the full ensemble mean and the mean of the subselected ensemble for three atmospheric variables that can be associated with the state of the winter climate over Europe (Fig. B.6). We choose one variable that we also use for the ensemble subselection, winter-mean 500 hPa geopotential height (Z_{500}), as well as two variables that are not included in the ensemble subselection, namely winter-mean MSLP, and 200 hPa zonal wind (U_{200}). Variations in MSLP indicate the prevalent distribution of high and low pressure areas, which directly influence the near-surface wind speed and can be indicative of the mean wind climate during winter. The field of Z_{500} provides insight into the state of the Rossby wave pattern

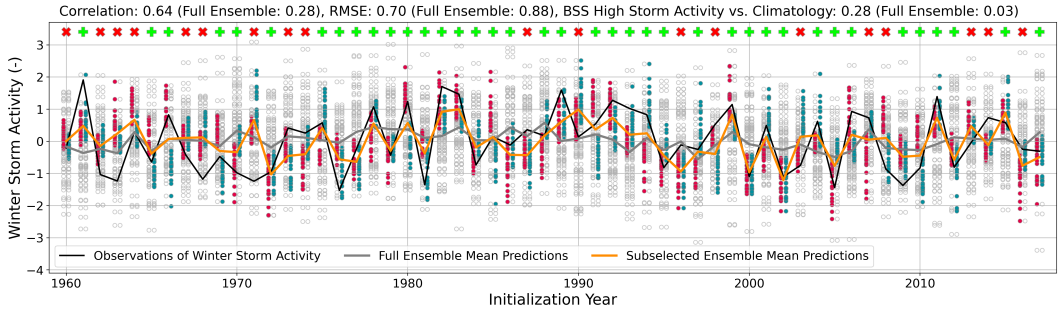


Figure B.5: Predictions of DJF GBSA by the 64-member ensemble mean (gray line), the subselected ensemble mean (orange line), as well as observed DJF GBSA (black line). Period 1960–2017 for model initializations, 1960/61–2017/18 for storm activity observations. Circles indicate GBSA predictions of individual members, colored circles indicate the selected 25 members closest to the first-guess predictions based on T_{70} (red), and Z_{500} (teal). Green plus signs and red "x" markers denote forecasts where the subselection is closer to or further away from the observation than the full ensemble.

in winter and whether the large-scale mid-tropospheric flow diverts storms away from or towards the German Bight. The location and strength of the polar jet stream, expressed as U_{200}, governs the lower tropospheric setup and can enhance or suppress the formation of storms.

We find that the full ensemble shows significant skill for deterministic winter MSLP forecasts north of 60°N, as well as for winter Z_{500} south of 45°N, but limited skill for both MSLP and Z_{500} over Central Europe and the adjacent region of the North Atlantic Ocean (Fig. B.6a and B.6d). The subselected ensemble shows a slightly higher skill for MSLP over Scandinavia and the Iberian Peninsula, but not over the German Bight and more generally Central Europe (Fig. B.6b and B.6c). The skill of the subselection for Z_{500} is also slightly improved from Greenland to northern Scandinavia (Fig. B.6e and B.6f). Despite not showing an improvement over the German Bight, higher skill north and south of the German Bight indicates an increase in the predictability of the meridional gradient of MSLP and Z_{500}, which is crucial to more accurately predict the wind climate in the German Bight. For U_{200}, the full ensemble shows significant skill in a mostly zonally-oriented band spanning from the North Atlantic around 55°N into west-central Europe (Fig. B.6g). Notably, positive correlations are located closer to the German Bight than for MSLP and Z_{500}. The subselected ensemble mostly retains this correlation pattern, but extends the significant skill across the German Bight into east-central Europe (Fig. B.6h and B.6i). The improvement in predictability of U_{200}, which is associated with the strength and location of the jet stream, is in accordance with the improvement in GBSA prediction skill, as the jet stream governs the formation and intensification of extratropical cyclones.

B.3.2.2 Potential capabilities of the model (perfect test)

To determine the theoretical maximum of skill improvement that the model could achieve, we perform a perfect test by selecting those 25 members in each forecast that are closest to the actual observed winter GBSA, and again analyze the change in ACC for MSLP, U_{200}, and Z_{500} (Fig. B.7). Note that this test includes information from the future and can therefore not be replicated operationally. Again, we find that the greatest skill increases occur in regions where the full ensemble already showed significant skill. For MSLP and Z_{500}, the skill north and south of the German Bight and therefore the predictability of the meridional gradient is significantly improved, while the skill in a region near and slightly west of the German Bight is almost unaffected by a perfect subselection (Fig. B.7b, B.7c, B.7e, and B.7f). Even with knowledge of the future German Bight storm activity, the ensemble is not able to significantly improve predictions of MSLP and Z_{500} in the same area. The perfect test also improves U_{200} predictability over regions where the full ensemble already showed skill, i. e., mostly between 50°N and 65°N (Fig. B.7h and B.7i).

Generally, the patterns of skill increase through ensemble subselection are similar for the non-cheating hindcast and the perfect test. The major difference between the two modes is that the increase in predictability of MSLP, Z_{500}, and U_{200} is much larger in the perfect test, which is to be expected as the model is able to use information from the future. From the similarity of the skill improvement patterns, however, we construe that the improvement of GBSA prediction skill through subselecting members is consistent with the physical mechanisms behind the extratropical winter storm climate and their predictability. The stark contrast in the magnitude of skill improvement points out the potential of the ensemble for even better predictions of the extratropical winter climate. However, additional research into more sophisticated ensemble refinement techniques is required to make use of this potential.

B.3.3 Representation of the mechanisms in the model

Figs. B.8 and B.9 show differences in composite mean modeled T_{70} and Z_{500} fields between years prior to modeled high- and low-storm-activity winters. The patterns of T_{70} differences (Fig. B.8) barely resemble the observed correlation patterns that are apparent between reanalyzed T_{70} fields and DJF GBSA observations (see Fig. B.2). Differences in the tropics, where observed correlations are highest, hardly exceed 0.3 K. In contrast, negative differences of up to -2 K, i. e., lower T_{70} preceding high DJF GBSA, emerge in the austral extratropics, where slightly negative correlations can also be found in the observations. Overall, the model appears to be incapable of reproducing the pathway from stratospheric temperature anomalies in September to changes in the extratropical winter storm climate in the German Bight.

The patterns in the composite differences of Z_{500} (Fig. B.9), however, demonstrate a fair agreement with observed correlation patterns (see Fig. B.3). Before high-storm-activity winters, geopotential heights in the model are up 30 gpm higher over the US East Coast, west-central Europe and northeast Asia than before low-storm-activity winters. Similarly, up to 30 gpm lower geopotential heights are modeled over Canada, Greenland, the Ural region and the Arctic in years prior to high-storm-activity winters. These regions of largest geopotential height differences match the regions of significant correlations between Z_{500} in ERA5 and observed DJF GBSA. We thus conclude that the physical link between November geopotential height anomalies and subsequent DJF GBSA is very well modeled by the hindcast system.

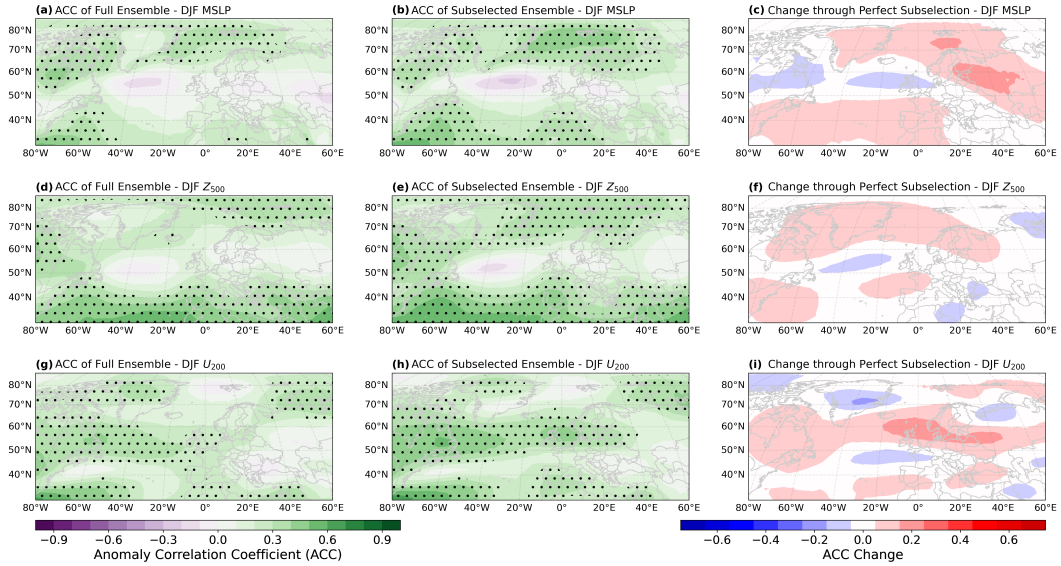


Figure B.6: Anomaly correlation coefficients (ACC) for ensemble mean predictions of the full 64-member ensemble (left column), the 25-member subselection (middle column), and the change in ACC between the full and subselected ensemble (right column) for winter-mean (DJF) MSLP anomalies (first row), 500 hPa geopotential height anomalies (Z_{500}, second row), and 200 hPa zonal wind anomalies (U_{200}, third row). Winter-mean anomalies are calculated by averaging monthly anomalies from December, January, and February. Period 1960/61–2017/18. Stippling indicates statistical significance ($p \leq 0.05$) determined through 1000-fold bootstrapping.

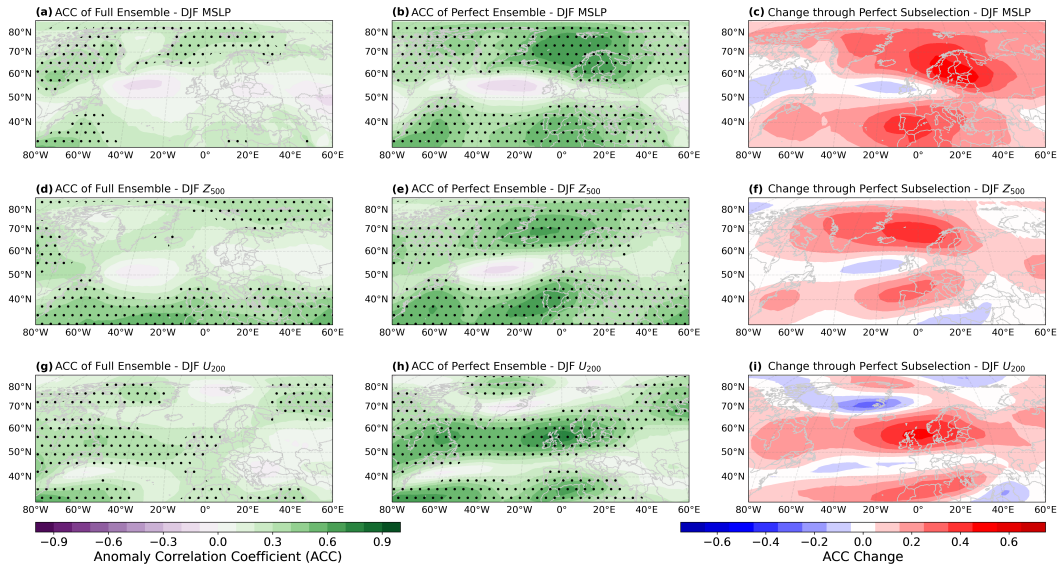


Figure B.7: Like Fig. B.6, but for a perfect test, i. e., the 25 members closest to the actually observed GBSA are selected.

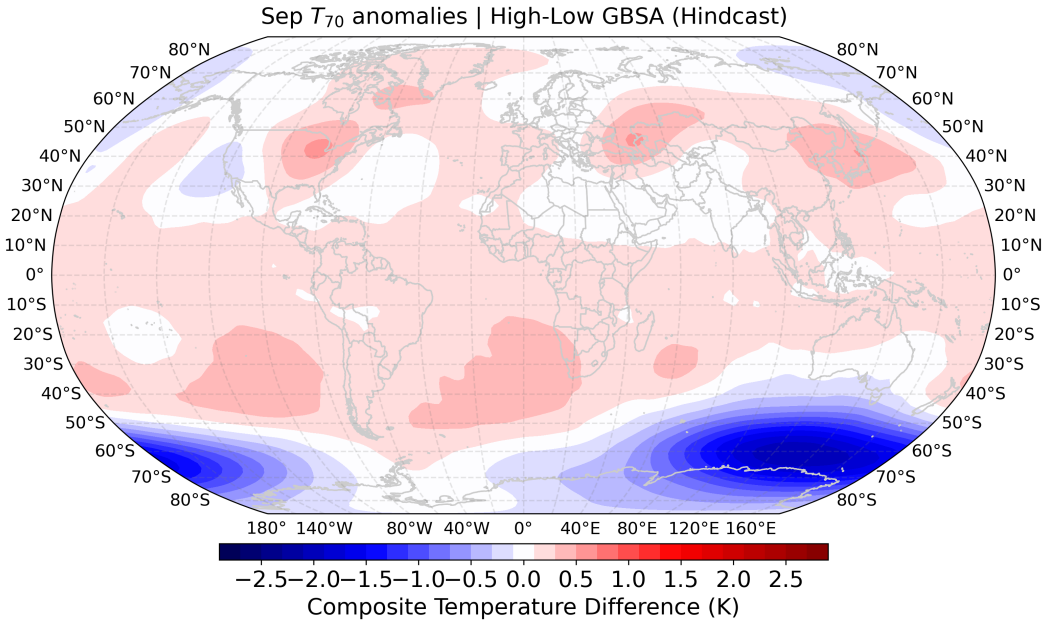


Figure B.8: Composite mean T_{70} of 100 model years with the highest subsequent DJF GBSA minus composite mean T_{70} of 100 model years with the lowest subsequent DJF GBSA in MPI-ESM-LR decadal hindcast runs. Data are taken from all initializations, all members, and all lead years except for the first year after initialization.

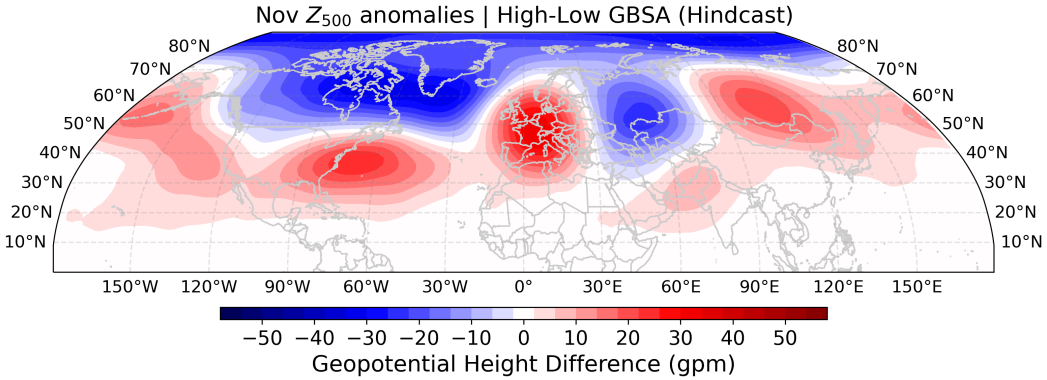


Figure B.9: Composite mean Z_{500} of 100 model years with the highest subsequent DJF GBSA minus composite mean Z_{500} of 100 model years with the lowest subsequent DJF GBSA in MPI-ESM-LR decadal hindcast runs. Data are taken from all initializations, all members, and all lead years except for the first year after initialization.

B.4 Discussion

We use a decadal prediction system for seasonal predictions because we want to make use of the large ensemble size and the high temporal resolution of the model output. While a seasonal prediction system would be sufficient for this analysis, we are not aware of any available seasonal single-model initialized large ensembles with 64 members and three-hourly MSLP output. In addition, the use of the MPI-ESM-LR decadal prediction system allows us to directly compare the predictability for the

first winter to the results from Krieger et al. (2022). We find that the full-ensemble prediction skill for winter GBSA ($r = 0.28$, $\mathrm{BSS} = 0.03$) is close to what Krieger et al. (2022) found for lead-year 1 predictions of annual GBSA.

Furthermore, the MPI-ESM-LR decadal hindcast offers a total of about 60 initialization years, while the corresponding seasonal prediction system based on the MPI-ESM-LR only covers about 40 initialization years. Additionally, the backward extension of ERA5 to 1940 allows us to define the training period as two decades which fully precede the decadal hindcast. Thus, we are able to generate predictor-based first guesses for almost six decades of hindcast initializations to test the skill of the model, while the seasonal system (in Dobrynin et al., 2018) only allowed for a hindcast period of two decades.

While our subselection increases the skill quite notably, there is still room for more improvement. This becomes especially apparent in the perfect test plots, where the potential perfect ACC increase for associated physical parameters like Z_{500}, U_{200}, and MSLP is a lot larger compared to our predictor-based ensemble subselection. A possible method to further improve the predictability and to rely more on the model physics would be checking which members actually predicted the observed patterns in November correctly and subselect those members. However, the ensemble spread in November (i. e., directly after the initialization) is too low to objectively distinguish good from bad members. We find that selecting the best members based on pattern correlations with observed fields in November does not increase the skill metrics as much as just using geopotential height anomaly-based first guesses from November. The method of refining the ensemble based on the predictions of observed patterns would become possible if the ensemble was initialized earlier than in November.

The correlation between temperature anomalies in the tropical stratosphere and GBSA is notably higher than the correlation between the same predictors and the North Atlantic Oscillation (NAO) index, a climate mode representative for the larger-scale atmospheric circulation over the North Atlantic (not shown). We argue that the increased correlation with GBSA is caused by the strong multidecadal signal within both the tropical stratosphere and GBSA which appears to be in phase over the investigated period. While GBSA is also connected to the NAO to a certain degree, this connection has been shown to fluctuate over time (e. g., Krieger et al., 2021). In the 1960s, the running correlation between GBSA and the NAO index reached its minimum at values below 0.2, indicating that the decadal to multidecadal signals in both time series appear to move out of phase at times.

Since this is a single-model study based on the MPI-ESM-LR, our findings are model-specific. Therefore, the conclusions we draw are true for this model and the associated model physics. However, because the subselection process is purely based on the statistical relationship between reanalysis data and observations, it could also work in other large model ensembles, as long as the internal variability in the ensemble encompasses the natural variability of the predicted quantity (GBSA).

We confirmed the connection between GBSA and the two chosen predictors through correlation analysis based on the ERA5 reanalysis. To ensure that the choice of reanalysis does not bias our results, we performed the correlation analysis between the predictor fields and GBSA in the NCEP-NCAR reanalysis (Kalnay et al., 1996) for the winters 1948/49–2017/18 and found similar patterns of correlations (not shown).

Despite having increased the predictability for the first winter on a seasonal scale, the decadal skill matrix for annual GBSA in Krieger et al. (2022) presents more lead times with poor predictability between lead year 1 and longer averaging periods. Using tropospheric patterns as predictors for longer lead times than the first winter is unphysical given the short memory of the troposphere. Therefore, new predictors (e. g., sea surface temperature) would need to be tested and used for an improvement of the GBSA prediction skill beyond the first winter. Alternatively, the model could be optimized to skillfully predict the state of the tropical stratosphere beyond the first year, for example via an accurate representation of the QBO. Such a prediction would then still require a statistical approach to link the QBO to GBSA, since we showed that the pathway from the tropical stratosphere to the extratropics in the boreal winter is poorly represented in the model (Fig. B.8). Any further analysis in this direction, however, is beyond the scope of this study.

B.5 Conclusions

We showed that the ensemble subselection technique first proposed by Dobrynin et al. (2018) can be applied to large-ensemble predictions of small-scale climate extremes. Using September T_{70} and November Z_{500} anomalies as predictors, we were able to increase the prediction skill of the MPI-ESM-LR large-ensemble decadal prediction system for winter GBSA for both deterministic and probabilistic predictions over a hindcast period of 58 winters. Compared to the inherently low prediction skill of the full ensemble, the subselection adds value to the seasonal predictability of GBSA by improving the ACC from 0.28 to 0.64, RMSE from 0.88 to 0.70, and BSS for high storm activity against climatology from 0.03 to 0.28. The sensitivity analysis showed that the improvement of skill metrics depends on the size of the subselection and on the combination of predictors. We also showed that the skill gain can be explained through physical mechanisms, as the subselected ensemble also displays a higher ACC for deterministic predictions of winter-mean U_{200} over the German Bight, as well as for the meridional gradient of MSLP and Z_{500} over north-central Europe, all of which are closely related to the European winter storm climate.

Bibliography

BIBLIOGRAPHY

Alexandersson, H. et al. (1998). "Long-term variations of the storm climate over NW Europe." *The Global Atmosphere and Ocean System* 6.2, pp. 97–120.

Allan, R. et al. (2009). "Fluctuations in autumn-winter severe storms over the British Isles: 1920 to present." *International Journal of Climatology* 29.3, pp. 357–371. DOI: 10.1002/joc.1765.

Athanasiadis, P. J. et al. (2017). "A Multisystem View of Wintertime NAO Seasonal Predictions." *Journal of Climate* 30.4, pp. 1461–1475. DOI: 10.1175/jcli-d-16-0153.1.

Athanasiadis, P. J. et al. (2020). "Decadal predictability of North Atlantic blocking and the NAO." *npj Climate and Atmospheric Science* 3.20. DOI: 10.1038/s41612-020-0120-6.

Baldwin, M. P. and T. J. Dunkerton (2001). "Stratospheric Harbingers of Anomalous Weather Regimes." *Science* 294.5542, pp. 581–584. DOI: 10.1126/science.1063315.

Barredo, J. I. (2010). "No upward trend in normalised windstorm losses in Europe: 1970–2008." *Natural Hazards and Earth System Sciences* 10.1, pp. 97–104. DOI: 10.5194/nhess-10-97-2010.

Bärring, L. and K. Fortuniak (2009). "Multi-indices analysis of southern Scandinavian storminess 1780–2005 and links to interdecadal variations in the NW Europe–North Sea region." *International Journal of Climatology* 29.3, pp. 373–384. DOI: 10.1002/joc.1842.

Bärring, L. and H. von Storch (2004). "Scandinavian storminess since about 1800." *Geophysical Research Letters* 31.20, p. L20202. DOI: 10.1029/2004GL020441.

Befort, D. J. et al. (2018). "Seasonal forecast skill for extratropical cyclones and windstorms." *Quarterly Journal of the Royal Meteorological Society* 145.718, pp. 92–104. DOI: 10.1002/qj.3406.

Bengtsson, L. et al. (2009). "Will Extratropical Storms Intensify in a Warmer Climate?" *Journal of Climate* 22.9, pp. 2276–2301. DOI: 10.1175/2008JCLI2678.1.

Bjerknes, V. (1904). "Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik." *Meteorol. Z.* 21, pp. 1–7.

Blender, R. et al. (1997). "Identification of cyclone-track regimes in the North Atlantic." *Quarterly Journal of the Royal Meteorological Society* 123.539, pp. 727–741. DOI: 10.1002/qj.49712353910.

Boer, G. J. and K. Hamilton (2008). "QBO influence on extratropical predictive skill." *Climate Dynamics* 31, pp. 987–1000. DOI: 10.1007/s00382-008-0379-5.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brune, S. and J. Baehr (2020). "Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions." *Wiley Interdisciplinary Reviews: Climate Change* 11.3, e637. DOI: 10.1002/wcc.637.

Buchana, P. and P. E. McSharry (2019). "Windstorm risk assessment for offshore wind farms in the North Sea." *Wind Energy* 22.9, pp. 1219–1229. DOI: 10.1002/we.2351.

Bundesamt für Seeschifffahrt und Hydrographie (2023). *Sturmfluten*. URL: https://www.bsh.de/DE/THEMEN/Wasserstand_und_Gezeiten/Sturmfluten/sturmfluten_node.html (visited on 08/09/2023).

Cappelen, J. et al. (2019). *DMI Report 19-02 Denmark - DMI Historical Climate Data Collection 1768-2018*. Tech. rep. tr19-02. Danish Meteorological Institute. URL: https://www.dmi.dk/fileadmin/user_upload/Rapporter/TR/2019/DMIRep19-02.pdf (visited on 05/19/2019).

Chang, E. K. M. (2018). "CMIP5 Projected Change in Northern Hemisphere Winter Cyclones with Associated Extreme Winds." *Journal of Climate* 31.16, pp. 6527–6542. DOI: 10.1175/jcli-d-17-0899.1.

Chang, E. K. M. and A. M. W. Yau (2016). "Northern Hemisphere winter storm track trends since 1959 derived from multiple reanalysis datasets." *Climate Dynamics* 47, pp. 1435–1454. DOI: 10.1007/s00382-015-2911-8.

Ciavola, P. et al. (2011). "Storm impacts along European coastlines. Part 1: The joint effort of the MICORE and ConHaz Projects." *Environmental Science & Policy* 14.7, pp. 912–923. DOI: 10.1016/j.envsci.2011.05.011.

Codiga, D. L. (2011). *Unified tidal analysis and prediction using the UTide Matlab functions*. en. Tech. rep. URI/GSO Technical Report 2011-01. Graduate School of Oceanography, University of Rhode Island. DOI: 10.13140/RG.2.1.3761.2008.

Compo, G. P. et al. (2015). *The International Surface Pressure Databank version 3*. Boulder, CO (United States): Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. DOI: 10.5065/D6D50K29. (Visited on 05/05/2018).

Cram, T. A. et al. (2015). "The International Surface Pressure Databank version 2." *Geoscience Data Journal* 2.1, pp. 31–46. DOI: 10.1002/gdj3.25.

Dacre, H. F. et al. (2012). "An Extratropical Cyclone Atlas: A Tool for Illustrating Cyclone Structure and Evolution Characteristics." *Bulletin of the American Meteorological Society* 93.10, pp. 1497–1502. DOI: 10.1175/bams-d-11-00164.1.

Dalelane, C. et al. (2020). "Seasonal Forecasts of Winter Temperature Improved by Higher-Order Modes of Mean Sea Level Pressure Variability in the North Atlantic Sector." *Geophysical Research Letters* 47.16, e2020GL088717. DOI: 10.1029/2020gl088717.

Danckwerth, C. and J. Mejer (1652). "Landtcarte Von dem Alten Nortfrieslande. Anno 1240." *Newe Landesbeschreibung Der Zwey Hertzogthümer Schleswich und Holstein: Zusambt Vielen dabey gehörigen Newen LandCarten*. Husum, pp. 89–92. URL: http://resolver.staatsbibliothek-berlin.de/SBB000187C200000000.

Dangendorf, S. et al. (2014). "North Sea Storminess from a Novel Storm Surge Record since AD 1843." *Journal of Climate* 27.10, pp. 3582–3595. DOI: 10.1175/jcli-d-13-00427.1.

Danmarks Meteorologiske Institut (2018). *Hvad er stormflod?* URL: https://www.dmi.dk/hav-og-is/temaforside-stormflod/danske-farvande-og-kyster/ (visited on 08/09/2023).

Degenhardt, L. et al. (2022). "Large-scale circulation patterns and their influence on European winter windstorm predictions." *Climate Dynamics* 60, pp. 3597–3611. DOI: 10.1007/s00382-022-06455-2.

De Guttry, C. and B. Ratter (2022). "Expiry date of a disaster: Memory anchoring and the storm surge 1962 in Hamburg, Germany." *International Journal of Disaster Risk Reduction* 70, p. 102719. DOI: 10.1016/j.ijdrr.2021.102719.

Deutscher Wetterdienst (2019). *Climate Data Center*. URL: https://opendata.dwd.de/climate_environment/CDC/ (visited on 03/11/2019).

– (2023). *Wetter- und Klimalexikon*. URL: https://www.dwd.de/lexikon (visited on 08/09/2023).

Dobrynin, M. et al. (2018). “Improved Teleconnection-Based Dynamical Seasonal Predictions of Boreal Winter.” *Geophysical Research Letters* 45.8, pp. 3605–3614. DOI: 10.1002/2018gl077209.

Dobrynin, M. et al. (2022). “Hidden Potential in Predicting Wintertime Temperature Anomalies in the Northern Hemisphere.” *Geophysical Research Letters* 49.20, e2021GL095063. DOI: 10.1029/2021GL095063.

Domeisen, D. I. V. et al. (2013). “The role of synoptic eddies in the tropospheric response to stratospheric variability.” *Geophysical Research Letters* 40.18, pp. 4933–4937. DOI: 10.1002/grl.50943.

Domeisen, D. I. V. et al. (2015). “Seasonal Predictability over Europe Arising from El Niño and Stratospheric Variability in the MPI-ESM Seasonal Prediction System.” *Journal of Climate* 28.1, pp. 256–271. DOI: 10.1175/JCLI-D-14-00207.1.

Domeisen, D. I. V. et al. (2020). “The role of North Atlantic–European weather regimes in the surface impact of sudden stratospheric warming events.” *Weather and Climate Dynamics* 1.2, pp. 373–388. DOI: 10.5194/wcd-1-373-2020.

Donat, M. G. et al. (2011). “Future changes in European winter storm losses and extreme wind speeds inferred from GCM and RCM multi-model simulations.” *Natural Hazards and Earth System Sciences* 11.5, pp. 1351–1370. DOI: 10.5194/nhess-11-1351-2011.

Donat, M. G. et al. (2010). “Examination of wind storms over Central Europe with respect to circulation weather types and NAO phases.” *International Journal of Climatology* 30.9, pp. 1289–1300. DOI: 10.1002/joc.1982.

Ebdon, R. (1975). “The Quasi-Biennial Oscillation and its association with tropospheric circulation patterns.” *Meteorol. Mag.* 104, pp. 282–297.

Epstein, E. S. (1969). “A Scoring System for Probability Forecasts of Ranked Categories.” *Journal of Applied Meteorology (1962-1982)* 8.6, pp. 985–987. URL: http://www.jstor.org/stable/26174707 (visited on 11/16/2022).

Eyring, V. et al. (2021). “Human Influence on the Climate System.” *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by IPCC. Cambridge University Press. DOI: 10.1017/9781009157896.005.

Fereday, D. R. et al. (2012). “Seasonal forecasts of northern hemisphere winter 2009/10.” *Environmental Research Letters* 7.3, p. 034031. DOI: 10.1088/1748-9326/7/3/034031.

Feser, F. et al. (2015). “Storminess over the North Atlantic and northwestern Europe-A review.” *Quarterly Journal of the Royal Meteorological Society* 141.687, pp. 350–382. DOI: 10.1002/qj.2364.

Fisher, R. A. (1915). “Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population.” *Biometrika* 10.4, pp. 507–521. DOI: 10.2307/2331838.

Frederikse, T. and T. Gerkema (2018). “Multi-decadal variability in seasonal mean sea level along the North Sea coast.” *Ocean Science* 14.6, pp. 1491–1501. DOI: 10.5194/os-14-1491-2018.

Fröhlich, K. et al. (2021). “The German Climate Forecast System: GCFS.” *Journal of Advances in Modeling Earth Systems* 13.2, e2020MS002101. DOI: 10.1029/2020ms002101.

Ganske, A. et al. (2018). “Identification of extreme storm surges with high-impact potential along the German North Sea coastline.” *Ocean Dynamics* 68, pp. 1371–1382. DOI: 10.1007/s10236-018-1190-4.

Gaslikova, L. et al. (2011). “Future storm surge impacts on insurable losses for the North Sea region.” *Natural Hazards and Earth System Sciences* 11.4, pp. 1205–1216. DOI: 10.5194/nhess-11-1205-2011.

Gómez-Navarro, J. J. and E. Zorita (2013). "Atmospheric annular modes in simulations over the past millennium: No long-term response to external forcing." *Geophysical Research Letters* 40.12, pp. 3232–3236. DOI: 10.1002/grl.50628.

Gönnert, G. et al. (2001). "7. Storm Surges generated by Extra-Tropical Cyclones - Case Studies." *Die Küste* 63. Ed. by Kuratorium für Forschung im Küsteningenieurwesen (KFKI), pp. 455–546. URL: https://hdl.handle.net/20.500.11970/101447.

Haas, R. et al. (2015). "Decadal predictability of regional-scale peak winds over Europe using the Earth System Model of the Max-Planck-Institute for Meteorology." *Meteorologische Zeitschrift* 25.6, pp. 739–752. DOI: 10.1127/metz/2015/0583.

Hadler, H. et al. (2018). "Geoarchaeological evidence of marshland destruction in the area of Rungholt, present-day Wadden Sea around Hallig Südfall (North Frisia, Germany), by the Grote Mandrenke in 1362 AD." *Quaternary International* 473.A, pp. 37–54. DOI: 10.1016/j.quaint.2017.09.013.

Haigh, I. D. et al. (2023). "GESLA Version 3: A major update to the global higher-frequency sea-level dataset." *Geoscience Data Journal* 10.3, pp. 293–314. DOI: 10.1002/gdj3.174. (Visited on 12/15/2021).

Hansen, F. et al. (2019). "Factors Influencing the Seasonal Predictability of Northern Hemisphere Severe Winter Storms." *Geophysical Research Letters* 46.1, pp. 365–373. DOI: 10.1029/2018gl079415.

Hanson, H. et al. (2002). "Beach nourishment projects, practices, and objectives—a European overview." *Coastal Engineering* 47.2, pp. 81–111. DOI: 10.1016/s0378-3839(02)00122-9.

Harvey, B. J. et al. (2014). "Equator-to-pole temperature differences and the extra-tropical storm track responses of the CMIP5 climate models." *Climate Dynamics* 43, pp. 1171–1182. DOI: 10.1007/s00382-013-1883-9.

Harvey, B. J. et al. (2020). "The Response of the Northern Hemisphere Storm Tracks and Jet Streams to Climate Change in the CMIP3, CMIP5, and CMIP6 Climate Models." *Journal of Geophysical Research: Atmospheres* 125.23, e2020JD032701. DOI: 10.1029/2020JD032701.

Haylock, M. R. (2011). "European extra-tropical storm damage risk from a multi-model ensemble of dynamically-downscaled global climate models." *Natural Hazards and Earth System Sciences* 11.10, pp. 2847–2857. DOI: 10.5194/nhess-11-2847-2011.

Heimreich, A. (1668). "Das XIII. Capitel. Von der An. 1634 ergangenen landverderblichen Sündenfluth." *M. Antoni Heimreichs Ernewrete Nordfresische Chronick*. Schleswig, pp. 355–368. URL: http://digital.slub-dresden.de/id363119175 (visited on 07/31/2023).

Heinrich-Mertsching, C. et al. (2023). "Subselection of seasonal ensemble precipitation predictions for East Africa." *Quarterly Journal of the Royal Meteorological Society* 149.755, pp. 2634–2653. DOI: 10.1002/qj.4525.

Helmholtz-Zentrum Hereon (2023a). *Hereon Storm Monitor*. URL: https://sturm-monitor.de (visited on 07/28/2023).

– (2023b). *Hereon Storm Surge Monitor*. URL: https://sturmflutmonitor.de (visited on 07/28/2023).

Hennessey, J. P. (1977). "Some Aspects of Wind Power Statistics." *Journal of Applied Meteorology* 16.2, pp. 119–128. DOI: 10.1175/1520-0450(1977)016<0119:saowps>2.0.co;2.

Hersbach, H. et al. (2020). "The ERA5 global reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: 10.1002/qj.3803.

Hoffmann, D. (2004). “Holocene landscape development in the marshes of the West Coast of Schleswig-Holstein, Germany.” *Quaternary International* 112.1, pp. 29–36. DOI: 10.1016/s1040-6182(03)00063-6.

Holton, J. R. and H.-C. Tan (1980). “The Influence of the Equatorial Quasi-Biennial Oscillation on the Global Circulation at 50 mb.” *Journal of the Atmospheric Sciences* 37.10, pp. 2200–2208. DOI: 10.1175/1520-0469(1980)037<2200:tioteq>2.0.co;2.

Houser, C. and B. Greenwood (2007). “Onshore Migration of a Swash Bar During a Storm.” *Journal of Coastal Research* 231, pp. 1–14. DOI: 10.2112/03-0135.1.

Hövel, L. et al. (2022). “Decadal Prediction of Marine Heatwaves in MPI-ESM.” *Geophysical Research Letters* 49.15. DOI: 10.1029/2022gl099347.

Huster, H. (1962). *Die große Februarsturmflut 1962 an Elbe-, Weser- und Oste-Mündung. Sturmflut-Katastrophe am 16. und 17. Februar 1962*. Otterndorf/Cuxhaven, Germany: Niederelbe-Zeitung / Cuxhavener Allgemeine.

Huthnance, J. (1991). “Physical oceanography of the North Sea.” *Ocean and Shoreline Management* 16.3–4. North Sea: Environment and Sea Use Planning, pp. 199–231. DOI: 10.1016/0951-8312(91)90005-M.

Ilyina, T. et al. (2013). “Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI–Earth system model in different CMIP5 experimental realizations.” *Journal of Advances in Modeling Earth Systems* 5.2, pp. 287–315. DOI: 10.1029/2012MS000178.

IPCC, ed. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. DOI: 10.1017/9781009157896.

Jochner, M. et al. (2013). “Reanalysis of the Hamburg Storm Surge of 1962.” *Weather extremes during the past 140 years* G89, pp. 19–26. DOI: 10.4480/GB2013.G89.02.

Jungclaus, J. H. et al. (2013). “Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI–Earth system model.” *Journal of Advances in Modeling Earth Systems* 5.2, pp. 422–446. DOI: 10.1002/jame.20023.

Kalnay, E. et al. (1996). “The NCEP/NCAR 40-Year Reanalysis Project.” *Bulletin of the American Meteorological Society* 77.3, pp. 437–471. DOI: 10.1175/1520-0477(1996)077<0437:tnyrp>2.0.co;2.

Karremann, M. K. et al. (2014). “Return periods of losses associated with European windstorm series in a changing climate.” *Environmental Research Letters* 9.12, p. 124016. DOI: 10.1088/1748-9326/9/12/124016.

Keizer, I. et al. (2023). “The acceleration of sea-level rise along the coast of the Netherlands started in the 1960s.” *Ocean Science* 19.4, pp. 991–1007. DOI: 10.5194/os-19-991-2023.

Kelletat, D. (1992). “Coastal Erosion and Protection Measures at the German North Sea Coast.” *Journal of Coastal Research* 8.3, pp. 699–711. URL: http://www.jstor.org/stable/4298018 (visited on 08/10/2023).

Kempe, M. (2006). “‘Mind the Next Flood!’ Memories of Natural Disasters in Northern Germany from the Sixteenth Century to the Present.” *The Medieval History Journal* 10.1-2, pp. 327–354. DOI: 10.1177/097194580701000212.

Kingma, D. P. and J. Ba (2014). *Adam: A Method for Stochastic Optimization*. arXiv. DOI: 10.48550/ARXIV.1412.6980.

KNMI (2019). *KNMI Data Centre*. De Bilt, the Netherlands: KNMI, Royal Netherlands Meteorological Institute. URL: https://dataplatform.knmi.nl (visited on 08/20/2019).

Krieger, D. et al. (2022). “Skillful decadal prediction of German Bight storm activity.” *Natural Hazards and Earth System Sciences* 22.12, pp. 3993–4009. DOI: 10.5194/nhess-22-3993-2022.
Krieger, D. et al. (2023). “Improving seasonal predictions of German Bight storm activity.” *In preparation*.
Krieger, D. and S. Brune (2022). *MPI-ESM-LR1.2 decadal hindcast ensemble 3-hourly German Bight MSLP*. DOKU at DKRZ. URL: http://hdl.handle.net/21.14106/04bc4cb2c0871f37433a73ee38189690955e1f90.
Krieger, D. et al. (2021). “German Bight storm activity, 1897—2018.” *International Journal of Climatology* 41.S1, E2159–E2177. DOI: 10.1002/joc.6837.
Kron, W. (2013). “Coasts: the high-risk areas of the world.” *Natural Hazards* 66, pp. 1363–1382. DOI: 10.1007/s11069-012-0215-4.
Krueger, O. and H. von Storch (2011). “Evaluation of an Air Pressure–Based Proxy for Storm Activity.” *Journal of Climate* 24.10, pp. 2612–2619. DOI: 10.1175/2011JCLI3913.1.
Krueger, O. et al. (2013). “Inconsistencies between Long-Term Trends in Storminess Derived from the 20CR Reanalysis and Observations.” *Journal of Climate* 26.3, pp. 868–874. DOI: 10.1175/JCLI-D-12-00309.1.
Krueger, O. et al. (2019). “Northeast Atlantic Storm Activity and Its Uncertainty from the Late Nineteenth to the Twenty-First Century.” *Journal of Climate* 32.6, pp. 1919–1931. DOI: 10.1175/JCLI-D-18-0505.1.
Kruschke, T. et al. (2014). “Evaluating decadal predictions of northern hemispheric cyclone frequencies.” *Tellus A: Dynamic Meteorology and Oceanography* 66.1, p. 22830. DOI: 10.3402/tellusa.v66.22830.
Kruschke, T. et al. (2016). “Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms.” *Meteorologische Zeitschrift* 25.6, pp. 721–738. DOI: 10.1127/metz/2015/0641.
Kunsch, H. R. (1989). “The Jackknife and the Bootstrap for General Stationary Observations.” *The Annals of Statistics* 17.3, pp. 1217–1241. DOI: 10.1214/aos/1176347265.
Kushnir, Y. et al. (2019). “Towards operational predictions of the near-term climate.” *Nature Climate Change* 9, pp. 94–101. DOI: 10.1038/s41558-018-0359-7.
Lahiri, S. N. (2003). “Empirical Choice of the Block Size.” *Resampling Methods for Dependent Data*. New York, NY: Springer New York, pp. 175–197. DOI: 10.1007/978-1-4757-3803-2_7.
Lamb, H. and K. Frydendahl (1991). *Historic Storms of the North Sea, British Isles and Northwest Europe*. Cambridge University Press.
Lang, A. and U. Mikolajewicz (2020). “Rising extreme sea levels in the German Bight under enhanced CO2 levels: a regionalized large ensemble approach for the North Sea.” *Climate Dynamics* 55, pp. 1829–1842. DOI: 10.1007/s00382-020-05357-5.
Leckebusch, G. C. et al. (2008a). “Development and application of an objective storm severity measure for the Northeast Atlantic region.” *Meteorologische Zeitschrift* 17.5, pp. 575–587. DOI: 10.1127/0941-2948/2008/0323.
Leckebusch, G. C. et al. (2008b). “Extreme wind storms over Europe in present and future climate: a cluster analysis approach.” *Meteorologische Zeitschrift* 17.1, pp. 67–82. DOI: 10.1127/0941-2948/2008/0266.
Lehmann, A. et al. (2011). “Detailed assessment of climate variability in the Baltic Sea area for the period 1958 to 2009.” *Climate Research* 46.2, pp. 185–196. DOI: 10.3354/cr00876.
Lehmann, J. et al. (2015). “Increased record-breaking precipitation events under global warming.” *Climatic Change* 132.4, pp. 501–515. DOI: 10.1007/s10584-015-1434-y.

Liu, R. Y. and K. Singh (1992). “Moving blocks jackknife and bootstrap capture weak dependence.” *Exploring the Limits of Bootstrap*. Ed. by R. LePage and L. Billard. Wiley, pp. 225–248.

Liu, X. et al. (2022). “Still normal? Near-real-time evaluation of storm surge events in the context of climate change.” *Natural Hazards and Earth System Sciences* 22.1, pp. 97–116. DOI: 10.5194/nhess-22-97-2022.

Lockwood, J. F. et al. (2022). “Predictability of European winter 2020/2021: Influence of a mid-winter sudden stratospheric warming.” *Atmospheric Science Letters* 23.12, e1126. DOI: 10.1002/asl.1126.

Lorenz, D. J. and E. T. DeWeaver (2007). “Tropopause height and zonal wind response to global warming in the IPCC scenario integrations.” *Journal of Geophysical Research* 112.D10, p. D10119. DOI: 10.1029/2006JD008087.

Lu, H. et al. (2014). “Mechanisms for the Holton-Tan relationship and its decadal variation.” *Journal of Geophysical Research: Atmospheres* 119.6, pp. 2811–2830. DOI: 10.1002/2013jd021352.

Marotzke, J. et al. (2016). “MiKlip: A National Research Project on Decadal Climate Prediction.” *Bulletin of the American Meteorological Society* 97.12, pp. 2379–2394. DOI: 10.1175/BAMS-D-15-00184.1.

Marshall, A. G. and A. A. Scaife (2009). “Impact of the QBO on surface winter climate.” *Journal of Geophysical Research* 114.D18, p. D18110. DOI: 10.1029/2009jd011737.

Matulla, C. et al. (2008). “European storminess: late nineteenth century to present.” *Climate Dynamics* 31, pp. 125–130. DOI: 10.1007/s00382-007-0333-y.

Mauritsen, T. et al. (2019). “Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2.” *Journal of Advances in Modeling Earth Systems* 11.4, pp. 998–1038. DOI: 10.1029/2018MS001400.

Mayer, B. et al. (2022). “RCP8.5-projected changes in German Bight storm surge characteristics from regionalized ensemble simulations for the end of the twenty-first century.” *Frontiers in Climate* 4, p. 992119. DOI: 10.3389/fclim.2022.992119.

Moemken, J. et al. (2021). “The regional MiKlip decadal prediction system for Europe: Hindcast skill for extremes and user-oriented variables.” *International Journal of Climatology* 41.S1, E1944–E1958. DOI: doi.org/10.1002/joc.6824.

Mullen, S. L. and R. Buizza (2002). “The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System.” *Weather and Forecasting* 17.2, pp. 173–191. DOI: 10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2.

Murphy, A. H. (1969). “On the “Ranked Probability Score”.” *Journal of Applied Meteorology and Climatology* 8.6, pp. 988–989. DOI: 10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2.

– (1971). “A Note on the Ranked Probability Score.” *Journal of Applied Meteorology and Climatology* 10.1, pp. 155–156. DOI: 10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

– (1992). “Climatology, Persistence, and Their Linear Combination as Standards of Reference in Skill Scores.” *Weather and Forecasting* 7.4, pp. 692–698. DOI: 10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2.

National Academies of Sciences, Engineering, and Medicine (2016). *Attribution of Extreme Weather Events in the Context of Climate Change*. Washington, D.C.: National Academies Press. DOI: 10.17226/21852.

Neddermann, N.-C. et al. (2019). “Seasonal predictability of European summer climate re-assessed.” *Climate Dynamics* 53, pp. 3039–3056. DOI: 10.1007/s00382-019-04678-4.

Nerger, L. and W. Hiller (2013). "Software for ensemble-based data assimilation systems—Implementation strategies and scalability." *Computers & Geosciences* 55, pp. 110–118. DOI: 10.1016/j.cageo.2012.03.026.

Neu, U. et al. (2013). "IMILAST: A Community Effort to Intercompare Extratropical Cyclone Detection and Tracking Algorithms." *Bulletin of the American Meteorological Society* 94.4, pp. 529–547. DOI: 10.1175/BAMS-D-11-00154.1.

Niemeyer, H. D. (1986). "Changing of Wave Climate due to Breaking on a Tidal Inlet Bar." *Coastal Engineering Proceedings* 1.20, pp. 1427–1443. DOI: 10.9753/icce.v20.105.

Norddeutscher Rundfunk (2022). *Wangerooge: Frau bei Sturz von Dünenkante schwer verletzt*. URL: https://www.ndr.de/nachrichten/niedersachsen/oldenburg_ostfriesland/Wangerooge-Frau-bei-Sturz-von-Duenenkante-schwer-verletzt,abbruchkante102.html (visited on 08/15/2023).

– (2023). *Eine Stadt steht unter Wasser*. URL: https://www.ndr.de/geschichte/chronologie/Sturmflut-1962-Eine-Stadt-steht-unter-Wasser,sturmflut229.html (visited on 08/15/2023).

Peings, Y. (2019). "Ural Blocking as a Driver of Early-Winter Stratospheric Warmings." *Geophysical Research Letters* 46.10, pp. 5460–5468. DOI: 10.1029/2019gl082097.

Pinto, J. G. et al. (2007). "Changing European storm loss potentials under modified climate conditions according to ensemble simulations of the ECHAM5/MPI-OM1 GCM." *Natural Hazards and Earth System Sciences* 7.1, pp. 165–175. DOI: 10.5194/nhess-7-165-2007.

Pinto, J. G. et al. (2012). "Loss potentials associated with European windstorms under future climate conditions." *Climate Research* 54.1, pp. 1–20. DOI: 10.3354/cr01111.

Polkova, I. et al. (2019). "Initialization and Ensemble Generation for Decadal Climate Predictions: A Comparison of Different Methods." *Journal of Advances in Modeling Earth Systems* 11.1, pp. 149–172. DOI: 10.1029/2018MS001439.

Polkova, I. et al. (2021). "Predictors and prediction skill for marine cold-air outbreaks over the Barents Sea." *Quarterly Journal of the Royal Meteorological Society* 147.738, pp. 2638–2656. DOI: 10.1002/qj.4038.

Post, V. E. A. (2005). "Fresh and saline groundwater interaction in coastal aquifers: Is our technology ready for the problems ahead?" *Hydrogeology Journal* 13, pp. 120–123. DOI: 10.1007/s10040-004-0417-2.

Raible, C. C. et al. (2008). "Northern Hemisphere Extratropical Cyclones: A Comparison of Detection and Tracking Methods and Different Reanalyses." *Monthly Weather Review* 136.3, pp. 880–897. DOI: 10.1175/2007MWR2143.1.

Raible, C. C. et al. (2014). "Changing correlation structures of the Northern Hemisphere atmospheric circulation from 1000 to 2100 AD." *Climate of the Past* 10.2, pp. 537–550. DOI: 10.5194/cp-10-537-2014.

Reick, C. H. et al. (2013). "Representation of natural and anthropogenic land cover change in MPI-ESM." *Journal of Advances in Modeling Earth Systems* 5.3, pp. 459–482. DOI: 10.1002/jame.20022.

Renggli, D. et al. (2011). "The Skill of Seasonal Ensemble Prediction Systems to Forecast Wintertime Windstorm Frequency over the North Atlantic and Europe." *Monthly Weather Review* 139.9, pp. 3052–3068. DOI: 10.1175/2011mwr3518.1.

Reyers, M. et al. (2019). "Development and prospects of the regional MiKlip decadal prediction system over Europe: predictive skill, added value of regionalization, and ensemble size dependency." *Earth System Dynamics* 10.1, pp. 171–187. DOI: 10.5194/esd-10-171-2019.

Richardson, D. S. (2001). "Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size." *Quarterly Journal of the*

Royal Meteorological Society 127.577, pp. 2473–2489. DOI: 10.1002/qj.49712757715.

Rijkswaterstaat (2023). *Stormvloed*. URL: https://www.rijkswaterstaat.nl/water/waterbeheer/bescherming-tegen-het-water/hoogwater/stormvloed (visited on 08/09/2023).

Scaife, A. A. et al. (2014a). "Skillful long-range prediction of European and North American winters." *Geophysical Research Letters* 41.7, pp. 2514–2519. DOI: 10.1002/2014gl059637.

Scaife, A. A. et al. (2014b). "Predictability of the quasi-biennial oscillation and its northern winter teleconnection on seasonal to decadal timescales." *Geophysical Research Letters* 41.5, pp. 1752–1758. DOI: 10.1002/2013gl059160.

Schiesser, H. H. et al. (1997). "Winter storms in Switzerland North of the Alps 1864/1865–1993/1994." *Theoretical and Applied Climatology* 58, pp. 1–19. DOI: 10.1007/bf00867428.

Schmidt, H. and H. von Storch (1993). "German Bight storms analysed." *Nature* 365.6449, p. 791. DOI: 10.1038/365791a0.

Schmith, T. et al. (1998). "Northeast Atlantic winter storminess 1875-1995 reanalysed." *Climate Dynamics* 14, pp. 529–536. DOI: 10.1007/s003820050239.

Schneck, R. et al. (2013). "Land contribution to natural CO2 variability on time scales of centuries." *Journal of Advances in Modeling Earth Systems* 5.2, pp. 354–365. DOI: 10.1002/jame.20029.

Schwierz, C. et al. (2009). "Modelling European winter wind storm losses in current and future climate." *Climatic Change* 101, pp. 485–514. DOI: 10.1007/s10584-009-9712-1.

Seiler, C. and F. W. Zwiers (2016). "How will climate change affect explosive cyclones in the extratropics of the Northern Hemisphere?" *Climate Dynamics* 46, pp. 3633–3644. DOI: 10.1007/s00382-015-2791-y.

Seneviratne, S. I. et al. (2021). "Weather and Climate Extreme Events in a Changing Climate." *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by IPCC. Cambridge University Press. DOI: 10.1017/9781009157896.013.

Siefert, W. and K. Havnoe (1988). "Einfluß von Baumaßnahmen in und an der Tideelbe auf die Höhe hoher Sturmfluten." ger. *Die Küste* 47. Ed. by Kuratorium für Forschung im Küsteningenieurwesen (KFKI), pp. 51–101. URL: https://hdl.handle.net/20.500.11970/101277.

Sienz, F. et al. (2016). "Ensemble size impact on the decadal predictive skill assessment." *Meteorologische Zeitschrift* 25.6, pp. 645–655. DOI: 10.1127/metz/2016/0670.

Siew, P. Y. F. et al. (2020). "Intermittency of Arctic–mid-latitude teleconnections: stratospheric pathway between autumn sea ice and the winter North Atlantic Oscillation." *Weather and Climate Dynamics* 1.1, pp. 261–275. DOI: 10.5194/wcd-1-261-2020.

Smith, D. M. et al. (2019). "Robust skill of decadal climate predictions." *npj Climate and Atmospheric Science* 2.13. DOI: 10.1038/s41612-019-0071-y.

Smith, D. M. et al. (2020). "North Atlantic climate far more predictable than models imply." *Nature* 583, pp. 796–800. DOI: 10.1038/s41586-020-2525-0.

Song, Y. and W. A. Robinson (2004). "Dynamical Mechanisms for Stratospheric Influences on the Troposphere." *Journal of the Atmospheric Sciences* 61.14, pp. 1711–1725. DOI: 10.1175/1520-0469(2004)061<1711:dmfsio>2.0.co;2.

Steffelbauer, D. B. et al. (2022). “Evidence of regional sea-level rise acceleration for the North Sea.” *Environmental Research Letters* 17.7, p. 074002. DOI: 10.1088/1748-9326/ac753a.

Stevens, B. et al. (2013). “Atmospheric component of the MPI–M Earth System Model: ECHAM6.” *Journal of Advances in Modeling Earth Systems* 5.2, pp. 146–172. DOI: 10.1002/jame.20015.

Strunk, A. (1970). *Die Heimat der Friesen, Teil 2 - Nordfriesland früher und heute*. Ingenieurbüro Alfred Strunk.

Stull, R. B. (1985). “Predictability and Scales of Motion.” *Bulletin of the American Meteorological Society* 66.4, pp. 432–436. DOI: 10.1175/1520-0477-66.4.432.

Suarez-Gutierrez, L. et al. (2020). “Dynamical and thermodynamical drivers of variability in European summer heat extremes.” *Climate Dynamics* 54, pp. 4351–4366. DOI: 10.1007/s00382-020-05233-2.

Sweeney, J. (2000). “A three-century storm climatology for Dublin 1715–2000.” *Irish Geography* 33.1, pp. 1–14. DOI: 10.1080/00750770009478595.

Tiggeloven, T. et al. (2021). “Exploring deep learning capabilities for surge predictions in coastal areas.” *Scientific Reports* 11, p. 17224. DOI: 10.1038/s41598-021-96674-0.

Tilinina, N. et al. (2013). “Comparing Cyclone Life Cycle Characteristics and Their Interannual Variability in Different Reanalyses.” *Journal of Climate* 26.17, pp. 6419–6438. DOI: 10.1175/JCLI-D-12-00777.1.

Trigo, I. F. (2006). “Climatology and interannual variability of storm-tracks in the Euro-Atlantic sector: a comparison between ERA-40 and NCEP/NCAR reanalyses.” *Climate Dynamics* 26, pp. 127–143. DOI: 10.1007/s00382-005-0065-9.

Trigo, R. M. et al. (2002). “The North Atlantic Oscillation influence on Europe: climate impacts and associated physical mechanisms.” *Climate Research* 20.1, pp. 9–17. DOI: 10.3354/cr020009.

Ulbrich, U. et al. (2009). “Extra-tropical cyclones in the present and future climate: a review.” *Theoretical and Applied Climatology* 96, pp. 117–131. DOI: 10.1007/s00704-008-0083-8.

Varino, F. et al. (2019). “Northern Hemisphere extratropical winter cyclones variability over the 20th century derived from ERA-20C reanalysis.” *Climate Dynamics* 52, pp. 1027–1048. DOI: 10.1007/s00382-018-4176-5.

Vautard, R. et al. (2019). “Human influence on European winter wind storms such as those of January 2018.” *Earth System Dynamics* 10.2, pp. 271–286. DOI: 10.5194/esd-10-271-2019.

Von Storch, H. and K. Woth (2008). “Storm surges: perspectives and options.” *Sustainability Science* 3, pp. 33–43. DOI: 10.1007/s11625-008-0044-2.

Von Storch, H. et al. (2008). “Storm surges—An option for Hamburg, Germany, to mitigate expected future aggravation of risk.” *Environmental Science & Policy* 11.8, pp. 735–742. DOI: 10.1016/j.envsci.2008.08.003.

Wang, J. et al. (2017). “Changes in Northern Hemisphere Winter Storm Tracks under the Background of Arctic Amplification.” *Journal of Climate* 30.10, pp. 3705–3724. DOI: 10.1175/jcli-d-16-0650.1.

– (2018). “Interannual Modulation of Northern Hemisphere Winter Storm Tracks by the QBO.” *Geophysical Research Letters* 45.6, pp. 2786–2794. DOI: 10.1002/2017GL076929.

Wang, X. L. et al. (2006). “Climatology and Changes of Extratropical Cyclone Activity: Comparison of ERA-40 with NCEP–NCAR Reanalysis for 1958–2001.” *Journal of Climate* 19.13, pp. 3145–3166. DOI: 10.1175/JCLI3781.1.

Wang, X. L. et al. (2009). "Trends and variability of storminess in the Northeast Atlantic region, 1874–2007." *Climate Dynamics* 33, pp. 1179–1195. DOI: 10.1007/s00382-008-0504-5.

Wang, X. L. et al. (2011). "Trends and low-frequency variability of storminess over western Europe, 1878–2007." *Climate Dynamics* 37, pp. 2355–2371. DOI: 10.1007/s00382-011-1107-0.

Wang, X. L. et al. (2012). "Trends and low frequency variability of extra-tropical cyclone activity in the ensemble of twentieth century reanalysis." *Climate Dynamics* 40, pp. 2775–2800. DOI: 10.1007/s00382-012-1450-9.

Wang, X. L. et al. (2016). "Inter-comparison of extra-tropical cyclone activity in nine reanalysis datasets." *Atmospheric Research* 181.10, pp. 133–153. DOI: 10.1016/j.atmosres.2016.06.010.

Wani, M. A. et al. (2020). *Advances in Deep Learning*. Springer Singapore. DOI: 10.1007/978-981-13-6794-6.

Wasserstraßen- und Schifffahrtsverwaltung des Bundes (2021). *Pegeldaten Wasserstand (Ganglinie) Mess-Station Cuxhaven-Steubenhöft*. URL: https://www.kuestendaten.de (visited on 12/15/2021).

Weisse, R. et al. (2005). "Northeast Atlantic and North Sea Storminess as Simulated by a Regional Climate Model during 1958–2001 and Comparison with Observations." *Journal of Climate* 18.3, pp. 465–479. DOI: 10.1175/JCLI-3281.1.

Wikimedia Commons (2019). *Nordfriesland-Karte von Johannes Mejer um 1240 (vor der Sturmflut 1362). Die roten Linien geben den Küstenverlauf zur Zeit der Berichtsvorlage an.* URL: https://commons.wikimedia.org/wiki/File:Nordfriesland_um_1240.jpg (visited on 08/15/2023).

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

Yeager, S. G. et al. (2018). "Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model." *Bulletin of the American Meteorological Society* 99.9, pp. 1867–1886. DOI: 10.1175/BAMS-D-17-0098.1.

Yettella, V. and J. E. Kay (2017). "How will precipitation change in extratropical cyclones as the planet warms? Insights from a large initial condition climate model ensemble." *Climate Dynamics* 49, pp. 1765–1781. DOI: 10.1007/s00382-016-3410-2.

Zappa, G. et al. (2013). "A Multimodel Assessment of Future Projections of North Atlantic and European Extratropical Cyclones in the CMIP5 Climate Models." *Journal of Climate* 26.16, pp. 5846–5862. DOI: 10.1175/JCLI-D-12-00573.1.

Zhang, X. et al. (2004). "Climatology and Interannual Variability of Arctic Cyclone Activity: 1948–2002." *Journal of Climate* 17.12, pp. 2300–2317. DOI: 10.1175/1520-0442(2004)017<2300:CAIVOA>2.0.CO;2.

EIDESSTATTLICHE VERSICHERUNG

Eidesstattliche Versicherung
Declaration on Oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated.

Hamburg, den 14.12.2023

Daniel Ulrich Ludwig Krieger