Skillful Decadal Prediction of German Bight Storm Activity

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

Correspondence: Daniel Krieger (daniel.krieger@hereon.de)

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 three-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 ten years against observations. The model ~~shows limited deterministic skill for single prediction~~ produces poor deterministic predictions of GBSA and winter MSLP anomalies for individual years, but ~~significant positive deterministic skill for long~~ fair predictions for longer averaging periods. ~~For~~ A similar but smaller skill difference between short and long averaging periods also emerges for probabilistic predictions of high ~~and low storm activity~~storm activity. At long averaging periods (longer than 5 years), the model is ~~skillful at both short and long averaging periods, and outperforms persistence-based~~ more skillful than persistence- and climatology-based predictions. For short ~~lead years , the skill of the probabilistic prediction for high and low storm activity notably exceeds the deterministic skill~~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, ~~skillful decadal predictions of regional storm activity can be viable with a large ensemble and a carefully designed approach~~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, medium, and high storm activity is necessary to expose this skill.

1 Introduction

In low-lying coastal areas that are affected by mid-latitude storms, coastal protection ~~, planning,~~ 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~~. Here, the low-lying~~ , where the coastlines are heavily and frequently affected by ~~storm surges caused by~~ mid-latitude storms.

Climate projections suggest that many components of the Earth system undergo changes that can be attributed to the ~~anthropogenically caused global warming trend~~ 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). ~~However~~For storm activity, studies for the past century showed ~~that storm activity~~ a lack of significant long-term trends over the Northeast Atlantic in general and the German Bight in particular~~does not exhibit any significant long-term trends, but instead~~. 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., 2020). This dominant internal variability suggests ~~great potential value~~ 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 ~~(Marotzke et al., 2016)~~ (*MittelfristigeKlimaprognosen*, 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 ten 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 ~~separate category~~ 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 ~~the MPI-ESM decadal~~ 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, ~~? also~~ 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 of 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., 2020). 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, ~~an assumption which was shown by Krueger et al. (2019) to be valid~~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 ~~ageostrophic disturbances~~ disturbances from friction are negligible. We therefore ~~assume~~ 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 ~~averaging~~ calculating the MSLP gradient over a small area ~~preserves much of the spatial~~ allows for the detection of small-scale variability of the pressure field, which is crucial for estimating geostrophic wind statistics.

Besides the choice of ~~parameters~~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 ~~benefits the generation~~ increases the reliability of probabilistic predictions. The concept of a probabilistic approach is the presumption that a ~~shift in the~~ change in the shape of the ensemble distribution can be used to predict likelihoods of actual ~~shifts in~~ changes of 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 ~~prediction~~ predictive distribution, probabilistic predictions for extreme events, i.e. periods with very high or low storm activity, become ~~viable~~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-LR~~MPI-ESM. Since GBSA is connected to the large-scale circulation (Krieger et al., 2020), 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. 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 ~~show how a high deterministic skill for winter MSLP translates to a high deterministic skill of model system for GBSA (Sect. 3.1.2). The deterministic skill is quantified~~ 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. 3.1.2). We then evaluate the ~~probabilistic prediction~~ skill of the DPS for probabilistic predictions of MSLP and GBSA (Sect. 3.2.1 and 3.2.2), expressed via the Brier Skill Score

~~(BSS)~~(BSS, Brier, 1950), and discuss the advantages and limits of ~~the probabilistic approach .~~ our approach (Sect. 3.3). Concluding remarks are given in Sect. 4.

2 Methods and Data

2.1 The Observational Reference

We use the time series of annual GBSA from Krieger et al. (2020) 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 three-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 (Cram et al., 2015; Compo et al., 2015), as well as the national weather services of Germany ~~(Deutscher Wetterdienst; DWD) (DWD, 2019), Denmark (Danmarks Meteorologiske Institut; DMI) (Cappelen et al., 2019)~~(Deutscher Wetterdienst, DWD, 2019), Denmark (Danmarks Meteorologiske Institut, Cappelen et al., 2019) , and the Netherlands ~~(Koninklijk Nederlands Meteorologisch Instituut; KNMI) (KNMI, 2019). The thereby derived observational time series for~~ (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.

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 $0.1\,\mathrm{hPa}$ and the surface (Stevens et al., 2013). The horizontal extent of the grid boxes is approximately $210\,\mathrm{km}$ x $210\,\mathrm{km}$ at the Equator, and $125\,\mathrm{km}$ x $210\,\mathrm{km}$ over the German Bight, which is still fine enough for the German Bight to cover multiple gridpoints. 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 November 1st 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 ~~initial states are taken from~~ ensemble members. The basis of this scheme is formed by a 16-member ~~simulation assimilating both~~ ensemble assimilation, which from 1958 to 2019 assimilates the observed oceanic and atmospheric state ~~(Brune and Baehr, 2020). Here~~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. ~~In addition, four different perturbations are~~ 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 5x16=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 three-hourly output (see Sect. 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. (2020) 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.

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 tenth following year (lead year 10). For consistency, we only consider full calendar years for the comparison, leaving us with ten complete years per intialization year and ensemble member. The ten 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 Supplementary Material 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 analyse 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.

2.2.2 ~~Pressure Reduction and~~ Geostrophic Wind ~~Calculation~~and German Bight Storm Activity

~~Following Krieger et al. (2020)~~For our analysis, we use three-hourly MSLP ~~data from the decadal hindcast ensemble and derive geostrophic winds from the horizontal MSLP gradients~~over the North Atlantic basin, including the German Bight. As three-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. ~~The~~ Following Alexandersson et al. (1998) and Krueger et al. (2019), the equation for the reduction of p to the MSLP p_0 reads

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

with the Earth's gravitational acceleration $g = 9.80665\,\mathrm{m\,s^{-2}}$, the assumed wet-adiabatic lapse rate $\Gamma = 0.0065\,\mathrm{K\,m^{-1}}$, ~~and the assumed isentropic coefficient $\kappa = 1.235$.~~ the molar mass of air $M = 28.9647\,\mathrm{g\,mol^{-1}}$, and the gas constant of air $R = 8.3145\,\mathrm{J\,mol^{-1}\,K}^{\circ}{-1}$. 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 three closest gridpoints that span a triangle encompassing the German Bight ~~.~~ (Fig. 1). The coordinates of the selected gridpoints are specified in Table 1. The gridpoints 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 of the wind direction towards the main axis of the triangle (Krieger et al., 2020). We use three-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).

~~We generate time series of German Bight storm activity in the MPI-ESM-LR hindcast runs. According to Krieger et al. (2020), we define German Bight storm activity~~

GBSA is defined as the standardized annual 95th percentiles of three-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. 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$\,\mathrm{m\,s^{-1}}$ with an average of 22.87$\,\mathrm{m\,s^{-1}}$ (Fig. 3), which is close to the observational average of 22.19$\,\mathrm{m\,s^{-1}}$ derived by Krieger et al. (2020) 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. (2020).

Table 1. Coordinates of the three gridpoints used for storm activity calculation in the model.

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

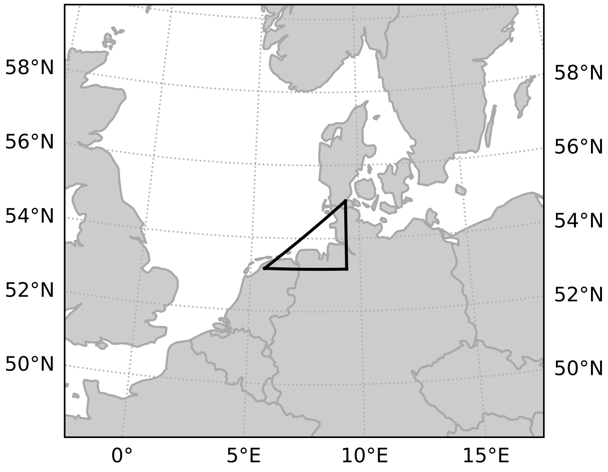

Figure 1. Map of Northwestern Europe, showing the location of the German Bight triangle.

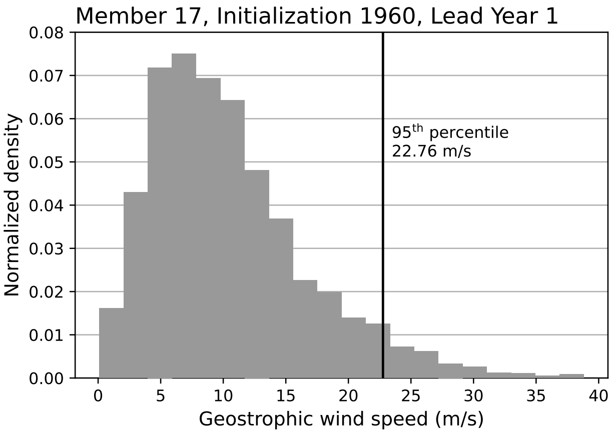

Figure 2. Exemplary distribution of predicted three-hourly geostrophic wind speeds for lead year 1 from member 17, initialized in 1960. The vertical line marks the 95th percentile, which is used in the calculation of storm activity.

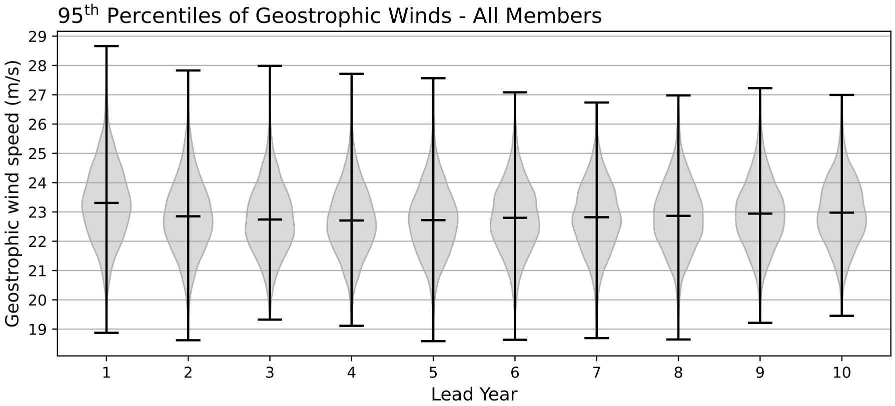

Figure 3. Violin plot of the distribution of annual 95th percentiles of geostrophic wind speeds from all members and all initializations, separated by lead year. Lead years increase from left to right along the x-axis. The width of the violin indicates the normalized density for a certain wind speed.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

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

2.3 Evaluation of ~~Prediction Skill~~Model Performance

In this study, we evaluate the model's ~~predictions skill~~ performance for both deterministic and probabilistic predictions. First, we evaluate deterministic predictions to quantify the ability of the model to capture the variability of GBSA. Second, we analyze probabilistic predictions to examine whether the large ensemble is able to skillfully differentiate between extremes and

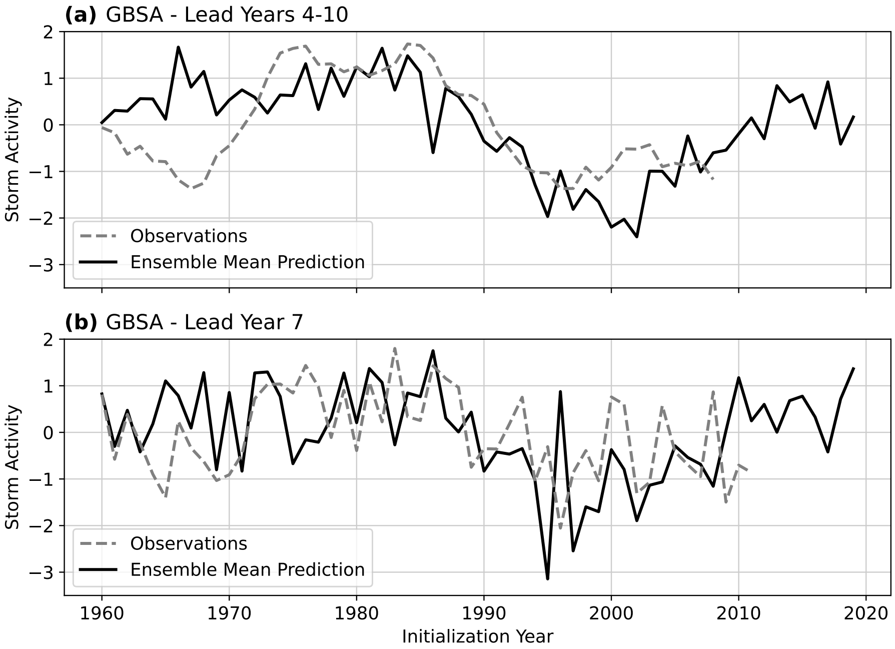

Figure 4. Time series of observations (grey, dashed) and ensemble mean predictions (black, solid) of German Bight storm activity (GBSA) for lead years 4-10 **(a)** and lead year 7 **(b)**.

non-extremes. These two prediction types require different evaluation metrics.

2.3.1 Anomaly Correlation

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

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

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, 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 datapoints 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 ~~correlations, computing the 95 % confidence intervals~~ bootstrapped correlations, averaging over all values in z-space, and transforming ~~them~~ the average back to the original space. The transformation of correlations ACC to z-scores z and its inverse are defined as $z = \mathrm{arctanh}(ACC)$ and $ACC = \tanh(z)$, where $\tanh$ and $\mathrm{arctanh}$ are the hyperbolic tangent function and its inverse, respectively.

2.3.2 Brier Skill Score

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

$$BSS = 1 - \frac{BS}{BS_{\mathrm{ref}}}, \tag{3}$$

where BS and BS_{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 ~~skill scores are~~ model performance is significantly different from the reference.

The ~~individual Brier Scores~~ Brier Score BS ~~are defined via~~ is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\underline{f}F_i - \underline{o}O_i)^2, \tag{4}$$

with the number of predictions N, the predicted probability of an event ~~f_i~~ F_i and the event occurrence ~~o_i. Note that o_i~~ 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 ~~predicted~~ event happened or not. Because the BS is ~~caulated~~ calculated as the normalized mean square error in the probability space, ~~a perfect prediction~~it is negatively oriented with a range of 0 to 1, i.e.~~a prediction that always predicts the outcome correctly, would score a~~, better predictions score lower BS ~~of 0, while a prediction that is always incorrect would score a 1.~~ values. A prediction based on ~~random guesses ($f_i = 0.5$~~flipping a two-sided coin ($F_i = 0.5$) would score a BS of 0.25.

We are interested in the ~~probabilistic prediction skills for~~ 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 ~~1σ and -1σ, with σ denoting the standard deviation of the underlying time series.~~ 1 and -1. We define high activity ~~/anomaly~~ periods as time steps above ~~1σ~~1, low activity ~~/anomaly~~ periods as time steps below ~~-1σ~~-1, and moderate activity ~~/anomaly~~ 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.

2.4 Re-standardization of Multi-year Averages

Winter MSLP anomalies and ~~storm activity~~ 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 gridpoint-wise. As GBSA is based on ~~spatially averaged MSLP gradients~~the mean MSLP gradient of a plane through three gridpoints, we treat its spatial information like that of a single gridpoint and calculate skill metrics only once for the entire ~~spatial average~~plane.

2.5 Reference Forecasts

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

The deterministic persistence prediction of storm activity is generated by taking the average observed storm activity of ~~a number~~ n ~~of~~ 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 with~~ 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. 4 taking on a value of either 0 (correct) or 1 (incorrect).

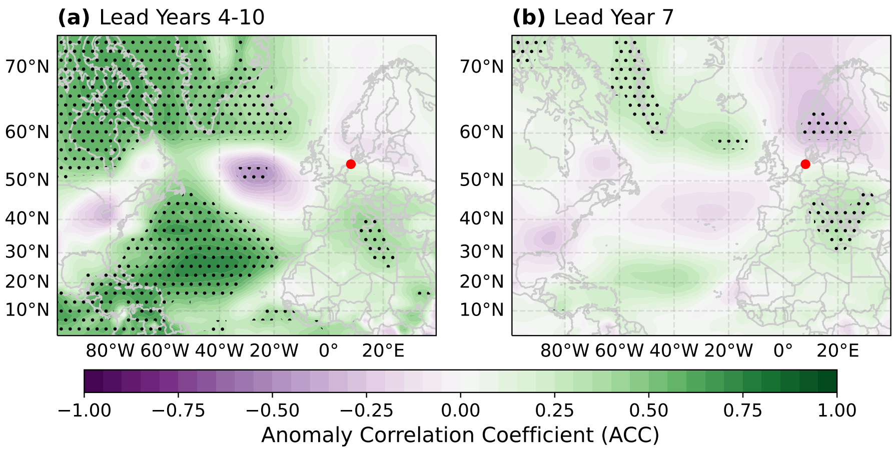

Figure 5. ~~Prediction skill for winter mean (DJF) MSLP anomalies, expressed as the gridpoint-wise~~ Gridpoint-wise anomaly correlation coefficient (ACC) between the deterministic hindcast ensemble mean prediction of winter mean (DJF) MSLP anomalies and ERA5 reanalysis data for lead years 4-10 **(a)** 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 ($p \leq 0.05$).

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 one 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.

3 Results and Discussion

3.1 Deterministic Predictions

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 ~~prediction skill for~~ predictions of winter (DJF) MSLP and data from the ERA5 reanalysis product, expressed as the gridpoint-wise anomaly correlation coefficient (ACC). For lead year 4-10 winter MSLP anomalies, the ~~model displays significant prediction skill~~ ACCs are positive over larger parts of the subtropical Atlantic, as well as Northeastern Canada and Greenland (Fig. 5a). ~~It also shows significant skill~~ Negative ACCs emerge in a circular area west of the British Isles. Over the German Bight, however, the ~~skill~~ 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. ~~But for~~ The *ACC* pattern of lead year 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, ~~our DPS displays significant skill~~ the *ACC* is negative over Scandinavia. ~~Anyhow, the overall magnitude~~ Across the rest of the spatial domain, the absolute values of the *ACC* ~~is~~ are lower for lead year 7 ~~, but the pattern shows some similarity compared to lead years 4-10~~ (Fig. 5b) ~~. In Scandinavia, a region of significant skill emerges, which is not present in the longer lead year range~~than for lead year 4-10, but the pattern shows some similarity. Again, ~~there is little to no skill for winter MSLP in~~ the *ACC* is insignificant over the German Bight~~. Over the majority of the spatial domain,~~ , indicating an insufficient skill to properly predict winter MSLP anomalies. The characteristics of the *ACC* distribution in Fig. 5b also hold for other single lead years, suggesting that longer averaging periods generally result in higher absolute correlations, both for regions with positive and negative correlation values.

~~The general lead-year dependence of the deterministic prediction skill agrees with previous findings of Kruschke et al. (2014), Kruschke et al. (2016), and **?** for other storm activity-related variables. In our study, the deterministic skill 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). This dependency on the window length implies that the deterministic predictions are unable to predict the short-term variability within winter MSLP. When applying longer averages, these year-to-year fluctuations are smoothed out, resulting in a higher prediction skill which likely arises from better predictable low-frequency variability of winter MSLP.~~

3.1.2 Storm Activity

We find that the ~~DPS shows some skill~~ *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 ~~providing skillful predictions~~ being significant over the German Bight. ~~Anyhow~~Still, the general ~~prediction skill of winter MSLP, in combination with similar prediction skill of winter MSLP gradients (not shown),~~ predictive capabilities of the DPS for winter MSLP motivates the investigation of GBSA predictability. ~~To investigate GBSA predictability , we calculate the ensemble mean GBSA. The deterministic prediction skill~~ Fig. 6 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), ~~whereas~~ but it increases towards longer averaging periods~~(Fig. 6). The skill~~. The *ACC* exhibits a clear dependence on the length of the averaging period, with lead years 1-10 showing the highest overall ~~skill~~ *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 ~~skill~~ *ACCs* for all multi-year prediction periods. This stands in clear contrast to the results for winter MSLP predictions~~in the~~

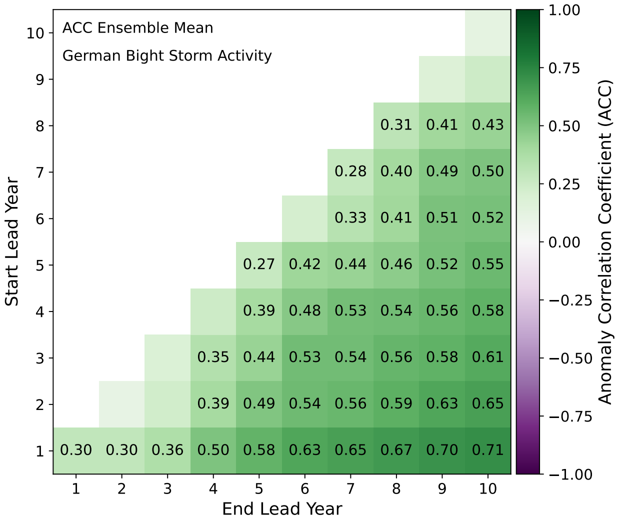

Figure 6. ~~Deterministic prediction skill of~~ Anomaly correlation coefficients between the deterministic DPS ~~for~~ forecasts and observations of German Bight storm activity for all combinations of start (y-axis) and end lead years (x-axis). Numbers in boxes indicate those correlation coefficients that are significantly different from 0 ($p \leq 0.05$).

~~German Bight~~, where the model failed to ~~be skillful~~ produce significant ACCs for both short and long averaging periods ~~.~~in the German Bight (compare Fig. 3.1.1).

Similar to the predictability of winter MSLP (Sect. 3.1.1), we ~~again~~ 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., 2020). ~~There is, however, a notable lack of a dependency of the deterministic skill~~ However, the ACC is notably independent on the lead time. We would expect a deterioration of the ~~deterministic skill~~ ACC with increasing temporal distance from the initialisation, i.e. along the ~~diagonals~~ diagonal in Fig. 6. 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~~, which is counter-intuitive. At this point, we are unable to come up with a convincing explanation for this behavior. Thus, further studies are needed to investigate why the prediction skill does not steadily decline with increasing lead times.~~ .

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 ~~shifts in the tails~~ changes in the shape of the ensemble distribution.

3.2.1 Mean Sea-Level Pressure

When predicting positive winter MSLP anomalies (Fig. 7a and 7b), the DPS significantly outperforms persistence ($BSS > 0$) over large parts of the Central North Atlantic and Europe for both ~~short and long lead year ranges.~~ 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 gridpoints. A similar pattern is found in predictions of negative anomalies (Fig. 7c and 7d), 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. 7e and 7f) 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 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 ~~even~~ less skillful than a ~~random guessing-based~~ 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 ~~absolute~~ skill of the DPS for ~~moderate~~ winter MSLP anomaly events.

Therefore, we additionally test the skill of the model for winter MSLP anomalies against ~~random guessing~~ that of a climatology-based prediction (Fig. ~~??~~8). The model ~~outperforms random guessing for both positive~~ BSS compared to climatology is mostly indistinguishable from 0 for both lead years 4-10 (Fig. ~~??aand ??b) and negative~~ 8a, 8c, and 8e) and 7 (Fig. ~~??c and ??d) winter MSLP anomalies, which is to be expected as extreme anomaly events occur much less frequently than the assumed probability of 50~~8b, 8d, and 8f), 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. ~~% by definition.However, the model~~ 8c), whereas negative BSS ~~for moderate~~ values emerge over the polar North Atlantic for lead year 4-10 predictions of positive and moderate winter MSLP anomalies (Fig. ~~??e and ??f)winter MSLP anomalies is mostly indistinguishable from 0, indicating a very limited potential of the DPS to predict moderate~~ 8a and 8e), as well as over the central North Atlantic for lead year 4-10 predictions of negative winter MSLP anomalies ~~better than randomly predicting them with a coin flip.~~(Fig. 8c).

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. ~~This inverted~~

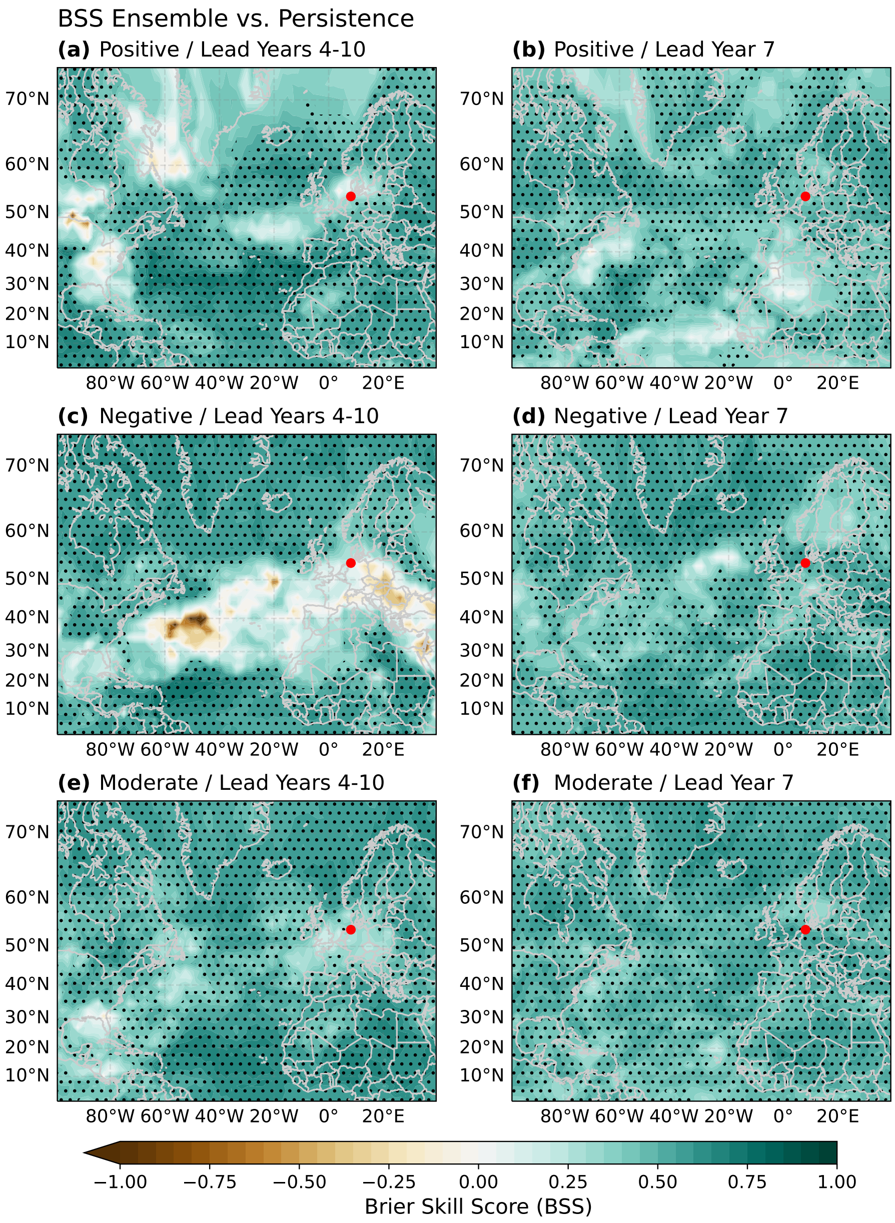
BSS Ensemble vs. Persistence
(a) Positive / Lead Years 4-10
(b) Positive / Lead Year 7
(c) Negative / Lead Years 4-10
(d) Negative / Lead Year 7
(e) Moderate / Lead Years 4-10
(f) Moderate / Lead Year 7
70°N
60°N
50°N
40°N
30°N
20°N
10°N
80°W 60°W 40°W 20°W 0° 20°E
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Brier Skill Score (BSS)

Figure 7. ~~Probabilistic prediction~~ Prediction skill ~~for~~ 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 ~~-1σ~~ -1 and ~~1σ~~ 1. The German Bight is marked by a red dot. Stippling marks areas with a BSS significantly different from 0 ($p \leq 0.05$).

~~dependency of the skill on the length of the averaging window (i.e., a higher skill for shorter periods) indicates that the~~

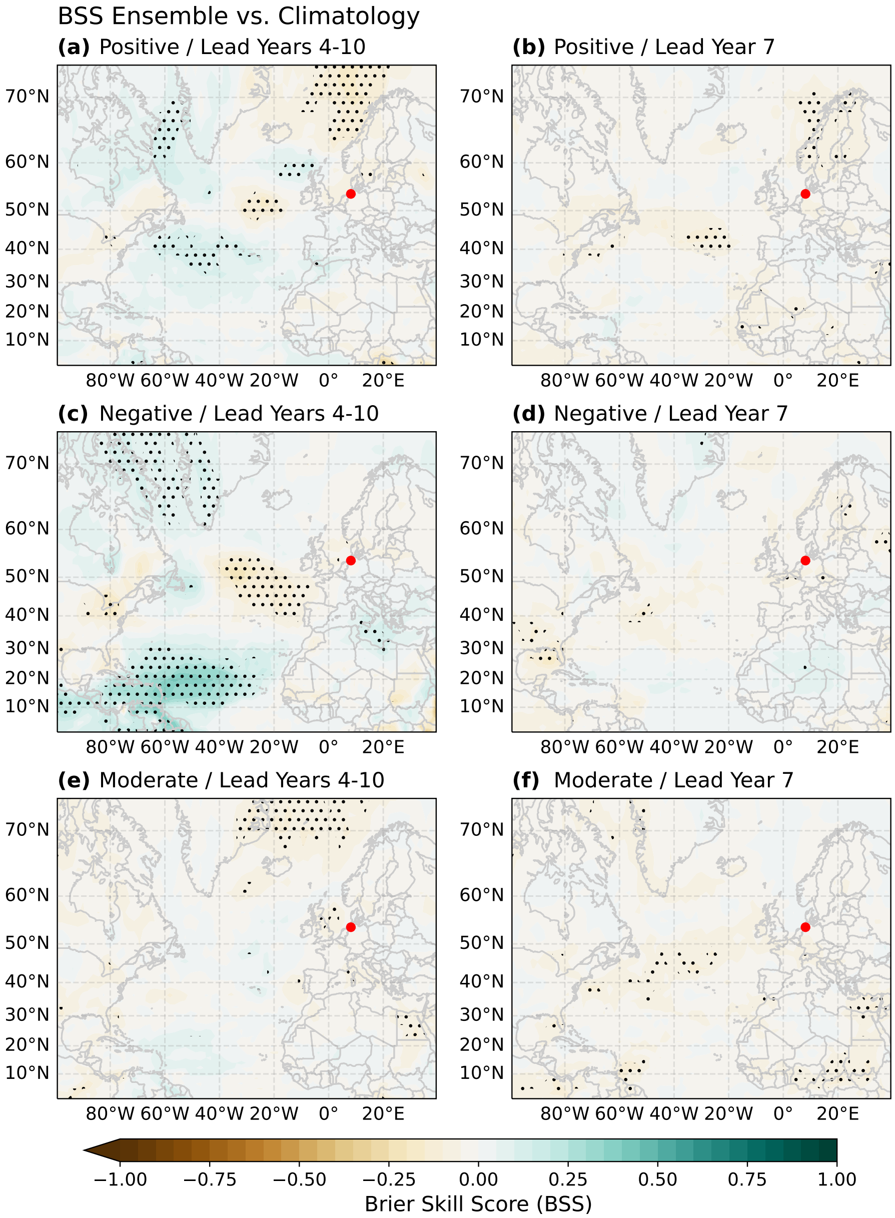
BSS Ensemble vs. Climatology
(a) Positive / Lead Years 4-10
(b) Positive / Lead Year 7
(c) Negative / Lead Years 4-10
(d) Negative / Lead Year 7
(e) Moderate / Lead Years 4-10
(f) Moderate / Lead Year 7
70°N
60°N
50°N
40°N
30°N
20°N
10°N
80°W 60°W 40°W 20°W 0° 20°E
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Brier Skill Score (BSS)

Figure 8. Like Fig. 7, but evaluated against a climatology-based prediction as a baseline.

~~assumption of a capability of the DPS to skillfully predict the underlying low-frequency variability is only valid for deterministic predictions (see Sect. 3.1) , but not for probabilistic predictions. Here, the DPS appears to be more skillful for probabilistic~~ 390 ~~predictions of short averaging periodsand thus the high-frequency variability of winter MSLP anomalies. The skill of probabilistic predictions of moderate winter MSLP anomalies significantly exceeds that of persistence , yet this~~ In addition, the DPS fails

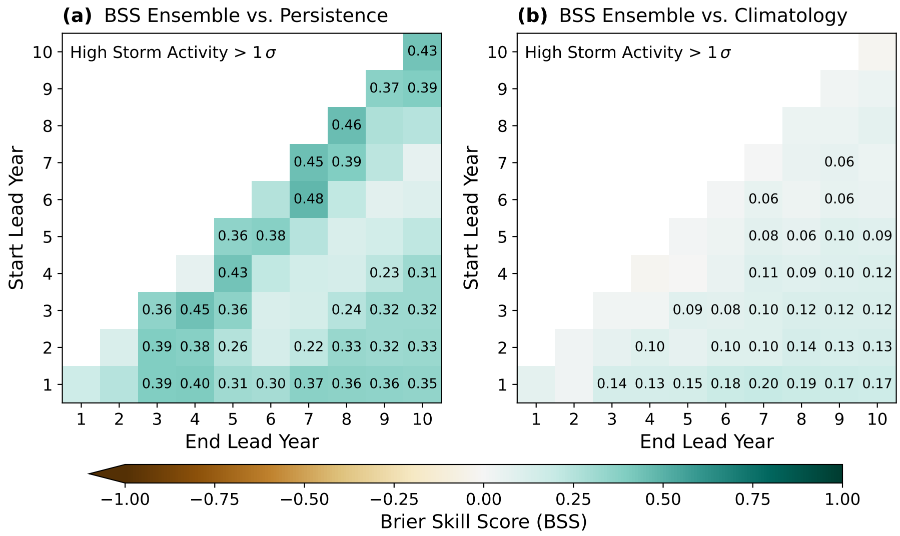

Figure 9. Brier Skill Score (BSS) of the 64 member ensemble for high storm activity evaluated against a persistence prediction as a baseline, shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A storm activity level of ~~1σ~~ 1 is used as a detection threshold for high activity.

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 ~~the low skill of persistence predictions rather than high skill of the DPS.~~ 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.

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 ~~as our baseline~~and climatology as our baselines.

For high storm activity predictions, the ~~ensemble~~ BSS against persistence is positive for all lead year combinations, indicating a better performance of the DPS than persistence (Fig. 9a). 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. 9b), 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

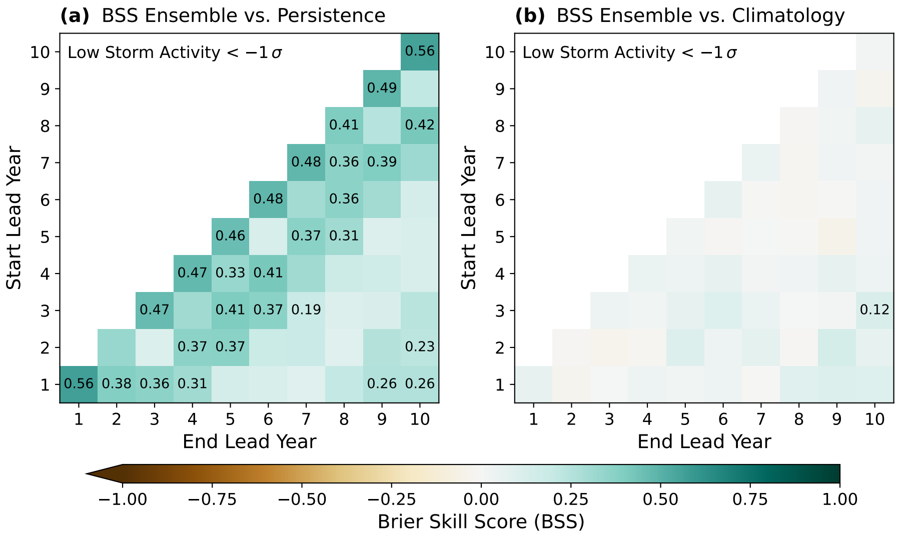

Figure 10. ~~Brier Skill Score (BSS) of the 64 member ensemble~~ Like Fig. 9, but for low storm activity~~evaluated against a persistence prediction as a baseline~~, ~~shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A~~ defined as storm activity ~~level of -1σ is used as a detection threshold for low activity.~~below -1.

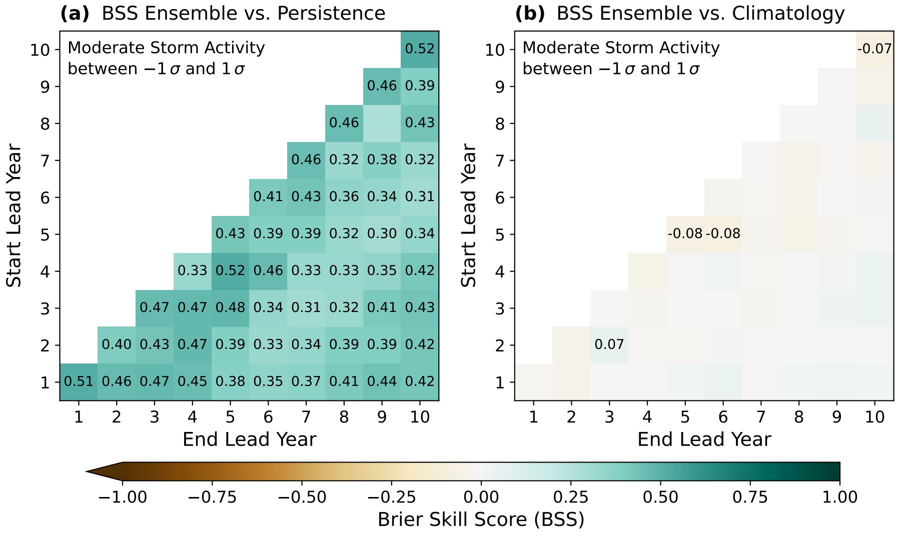

Figure 11. ~~Brier Skill Score (BSS) of the 64 member ensemble~~ Like Fig. 9, but for moderate storm activity~~evaluated against a persistence prediction as a baseline~~, ~~shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A~~ defined as storm activity ~~level of 1σ is used as a detection threshold for high activity.~~between 1 and -1.

BSS values among the lead year combinations against climatology differs strongly from the one obtained through testing against persistence (compare Fig. 9a), and much rather resembles the distribution of anomaly correlation coefficients between the deterministic predictions and observations (see Fig. 6). 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. 10a), 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. ~~There appears to be a higher skill difference between the DPS and persistence~~ The higher BSS for single years than for periods of 5-7 years ~~, indicating~~ indicates that the model is ~~most valuable at skillfully~~ valuable at predicting short periods. This behavior agrees with the findings in Sect. 3.2, which ~~demonstrated significantly~~ 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. 10b) for lead year 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.

Moderate storm activity predictions (Fig. 11a) 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 ~~persistent reference prediction. Similar to the evaluation of winter MSLP anomalies~~ persistence prediction. A comparison with climatology (Fig. ~~??) , we can challenge the model more honestly by replacing persistence with random guessing which assesses the model's prediction skill more realistically~~11b) 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 ~~BSS for both high (Fig. ??)and low (Fig. ??) storm activityagain outperform random guessing as expected, the comparison of predictions of moderate storm activity against random guessing (Fig. ??) reveals that all significantly positive BSS values vanish, and , for several lead year ranges, the model BSS even turns negative. Thus, we conclude that~~ 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 ~~probabilistic approach is not viable to skillfully predict moderate storm activity events.~~ 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.

~~Despite the inability of the DPS to skillfully predict moderate storm activity, the results suggest that our approach of employing a large ensemble notably aids the~~

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 ~~prediction skill. Contrary to previous studies on the decadal predictability of wind-related quantities, we find significant skill for extreme storm activity in the German Bight.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).~~ 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. 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 ~~impact on prediction skill by a further increase in the number of members is yet to be investigated.~~

~~Our separation of the probabilistic predictions also demonstrates the necessity to evaluate the skill for each prediction category individually. The 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)~~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.

~~Furthermore~~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~~ persistence- and climatology-based predictions, the BSS not only depends on the prediction skill of the model, but also on the skill of ~~persistence~~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 ~~overwhelmingly significant~~ skillful when evaluated against a persistent reference prediction. ~~Anyhow this overwhelmingly~~ However, this significant prediction skill turns ~~completely~~ mostly insignificant when evaluated against ~~random guessing as reference prediction. The signifcant BSS for extreme GBSA should, consequently, also be treated cautiously.As for moderate GBSA, signifcant BSS for extreme GBSA might turn out to be less~~ 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. 3.1). However, the inverse dependency is more likely a result of ~~exceptional model performance , but might rather indicate the limits of persistence forecasts. Unfortunately, random guessing is ill-suited as a reference prediction to evaluate extreme GBSA predictability.Therefore,~~ better performance by the persistence ~~still ranges~~ 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~~– despite the aforementioned potential deficiencies. Our~~. We therefore conclude that our DPS is particularly valuable at lead times during which ~~persistence~~ the reference forecasts are sufficiently poor. Vice-versa, the benefits of a DPS are negligible at lead times during which the skill of the ~~persistence~~ 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). 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 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 is 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.

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 ~~deterministic skill via the correlation coefficient ,~~ anomaly correlation coefficient (ACC) between deterministic predictions and observations or reanalysis data, respectively, evaluated probabilistic predictions for three different forecast categories with the Brier ~~Score~~Skill Score (BSS), and tested the probabilistic predictions of GBSA against ~~a persistence-based~~ both a persistence- and a climatology-based prediction.

Through comparison with data from the ERA5 reanalysis, we found that the DPS ~~shows poor skill for~~ produces poor deterministic predictions of winter MSLP anomalies over the German Bight. Over the North Atlantic, certain regions with ~~significant skill~~ higher correlations emerge, but the ~~skill~~ 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 ~~skill~~ predictability for GBSA also depicts this same dependency on the ~~lead range length, and is~~ averaging period, where ACCs are only significant for most ~~non-single yearlead times.We hypothesize that this lead time 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.~~averaging periods larger than 1 year.

~~In contrast to the limited deterministic skill, the DPS generates skillful probabilistic predictions for extreme low and high~~ Probabilistic predictions of winter MSLP anomalies over the North Atlantic ~~sector. This skill in predicting the extremes of the distribution is significant for both long and short averaging periods~~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.~~As this stands in contrast to the deterministic predictability of winter MSLP anomalies, we want to emphasize that we do not have a convincing explanation for this behavior and more research is needed.~~

~~This skill pattern for winter MSLP extremes translates to skillful predictions of extreme low and high GBSA, where the model consistently outperforms persistence . Most notably, the probabilistic prediction shows good GBSAprediction skill for single lead years, a time domain where deterministic predictions struggle to be skillful. The skill of probabilistic predictions is , however, limited to predictions of extreme activity. For periods with moderate storm activity, as well as moderate winter MSLP anomalies, the probabilistic predictions~~

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 ~~does outperform persistence, but fails to show a significantly higher skillthan random guessing~~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 ~~short lead-year periods~~ 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. ~~The high skill of probabilistic GBSA predictions facilitates the prediction of occurrence probabilities for different event categories, which might add to the applicability and usability of such predictions.~~

~~This study emphasizes the need to differentiate between event categories in the evaluation of GBSA predictability. Highly aggregated probabilistic skill scores, which aim at incorporating the model performance for various categories into one single value, might underestimate the capabilities to predict extremes, since poor performance in one event category could overshadow a higher prediction skill in other categories.~~

~~Additionally, the estimation of GBSA predictability heavily relies on the choice of a reference prediction. As it is difficult to find a single reference which properly evaluates both the tails and the center of a distribution correctly, there might be a risk of overestimating the capabilities of the DPS for certain event categories. However, further research is needed to investigate the prediction skill sensitivity to the choice of a reference, which is beyond the scope of this study.~~

~~The findings of this study highlight the advantage of large-ensemble decadal predictions.~~ By employing a large-ensemble DPS and ~~restricting the probabilistic prediction approach to positive and negative extreme events~~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 A: 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 ~~so~~ 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, ... ~~)~~.

Appendix B: ~~Probabilistic Skill against Random Guessing~~

~~Probabilistic prediction skill for 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 random guessing 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 $-1\,\sigma$ and $1\,\sigma$. Stippling marks areas with a BSS significantly different from 0 ($p \leq 0.05$).~~

~~Brier Skill Score (BSS) of the 64 member ensemble for high storm activity evaluated against random guessing as a baseline, shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A storm activity level of $1\,\sigma$ is used as a detection threshold for high activity.~~

~~Brier Skill Score (BSS) of the 64 member ensemble for low storm activity evaluated against random guessing as a baseline, shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A storm activity level of $-1\,\sigma$ is used as a detection threshold for low activity.~~

~~Brier Skill Score (BSS) of the 64 member ensemble for moderate storm activity evaluated against random guessing as a baseline, shown for all combinations of start (y axis) and end lead years (x axis). Numbers in boxes are those BSS that are significantly different from 0 ($p \leq 0.05$). A storm activity level between $-1\,\sigma$ and $1\,\sigma$ is used as a detection threshold for moderate activity.~~

Data availability. ERA5 reanalysis products that were used to support this study are available from the Copernicus Data Store under https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset. Three-hourly German Bight MSLP output data from the decadal prediction system are available under https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=DKRZ_LTA_1075_ds00011. Seasonal means of North Atlantic MSLP from the decadal prediction system are available under https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=DKRZ_LTA_1075_ds00012. Computed German Bight storm activity time series are available under https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=DKRZ_LTA_1075_ds00013. Selected global three-hourly atmospheric output from the decadal prediction system is available under https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=DKRZ_LTA_1075_ds00014.

Author contributions. DK, RW and JB conceived and designed the study. SB carried out the MPI-ESM hindcast experiments and contributed model data. DK, SB, PP, RW and JB analyzed and discussed the results. DK created the figures and wrote the manuscript with contribution from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

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.

References

Alexandersson, H., Schmith, T., Iden, K., and Tuomenvirta, H.: Long-term variations of the storm climate over NW Europe, The Global Atmosphere and Ocean System, 6, 1998.

Athanasiadis, P. J., Yeager, S., Kwon, Y.-O., Bellucci, A., Smith, D. W., and Tibaldi, S.: Decadal predictability of North Atlantic blocking and the NAO, npj Climate and Atmospheric Science, 3, 893, https://doi.org/10.1038/s41612-020-0120-6, 2020.

Bärring, L. and von Storch, H.: Scandinavian storminess since about 1800, Geophysical Research Letters, 31, 97, https://doi.org/10.1029/2004GL020441, 2004.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Brune, S. and Baehr, J.: Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions, Wiley Interdisciplinary Reviews: Climate Change, 11, 741, https://doi.org/10.1002/wcc.637, 2020.

Cappelen, J., Laursen, E. V., and Kern-Hansen, C.: DMI Report 19-02 Denmark - DMI Historical Climate Data Collection 1768-2018, Tech. Rep. tr19-02, Danish Meteorological Institute, 2019.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Allan, R. J., McColl, C., Yin, X., Vose, R. S., Matsui, N., Ashcroft, L., Auchmann, R., Benoy, M., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Cram, T. A., Crouthamel, R., Groisman, P. Y., Hersbach, H., Jones, P. D., Jonsson, T., Jourdain, S., Kelly, G., Knapp, K. R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S. J., Luterbacher, J., Marshall, G. J., Maugeri, M., Mock, C. J., Mok, H. Y., Nordli, O., Przybylak, R., Rodwell, M. J., Ross, T. F., Schuster, D., Srnec, L., Valente, M. A., Vizi, Z., Wang, X. L., Westcott, N., Woollen, J. S., and Worley, S. J.: The International Surface Pressure Databank version 3, https://doi.org/10.5065/D6D50K29, Accessed: 05 May 2018, 2015.

Cram, T. A., Compo, G. P., Yin, X., Allan, R. J., McColl, C., Vose, R. S., Whitaker, J. S., Matsui, N., Ashcroft, L., Auchmann, R., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, B. E., Groisman, P. Y., Hersbach, H., Jones, P. D., Jonsson, T., Jourdain, S., Kelly, G., Knapp, K. R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S. J., Luterbacher, J., Marshall, G. J., Maugeri, M., Mock, C. J., Mok, H. Y., Nordli, O., Rodwell, M. J., Ross, T. F., Schuster, D., Srnec, L., Valente, M. A., Vizi, Z., Wang, X. L., Westcott, N., Woollen, J. S., and Worley, S. J.: The International Surface Pressure Databank version 2, Geoscience Data Journal, 2, 31–46, https://doi.org/10.1002/gdj3.25, 2015.

DWD: Climate Data Center, https://opendata.dwd.de/climate_environment/CDC/, 2019.

Feser, F., Barcikowska, M., Krueger, O., Schenk, F., Weisse, R., and Xia, L.: Storminess over the North Atlantic and northwestern Europe-A review, Quarterly Journal of the Royal Meteorological Society, 141, 350–382, https://doi.org/10.1002/qj.2364, 2015.

Fisher, R. A.: Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, Biometrika, 10, 507, https://doi.org/10.2307/2331838, 1915.

Haas, R., Reyers, M., and Pinto, J. G.: Decadal predictability of regional-scale peak winds over Europe using the Earth System Model of the Max-Planck-Institute for Meteorology, Meteorologische Zeitschrift, 25, 739–752, https://doi.org/10.1127/metz/2015/0583, 2015.

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

N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: 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, 287–315, https://doi.org/10.1029/2012MS000178, 2013.

IPCC, ed.: 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, 2021.

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and Storch, J. S.: 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, 422–446, https://doi.org/10.1002/jame.20023, 2013.

KNMI: KNMI Data Centre, 2019.

Krieger, D., Krueger, O., Feser, F., Weisse, R., Tinz, B., and von Storch, H.: German Bight storm activity, 1897–2018, International Journal of Climatology, 41, E2159–E2177, https://doi.org/https://doi.org/10.1002/joc.6837, 2020.

Krueger, O. and von Storch, H.: Evaluation of an Air Pressure–Based Proxy for Storm Activity, Journal of Climate, 24, 2612 – 2619, https://doi.org/10.1175/2011JCLI3913.1, 2011.

Krueger, O., Feser, F., and Weisse, R.: Northeast Atlantic Storm Activity and Its Uncertainty from the Late Nineteenth to the Twenty-First Century, Journal of Climate, 32, 1919–1931, https://doi.org/10.1175/JCLI-D-18-0505.1, 2019.

Kruschke, T., Rust, H. W., Kadow, C., Leckebusch, G. C., and Ulbrich, U.: Evaluating decadal predictions of northern hemispheric cyclone frequencies, Tellus A: Dynamic Meteorology and Oceanography, 66, 22 830, https://doi.org/10.3402/tellusa.v66.22830, 2014.

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U.: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, Meteorologische Zeitschrift, 25, 721–738, https://doi.org/10.1127/metz/2015/0641, 2016.

Kunsch, H. R.: The Jackknife and the Bootstrap for General Stationary Observations, The Annals of Statistics, 17, 1217 – 1241, https://doi.org/10.1214/aos/1176347265, 1989.

Lahiri, S. N.: Empirical Choice of the Block Size, pp. 175–197, Springer New York, New York, NY, https://doi.org/10.1007/978-1-4757-3803-2_7, 2003.

Lehmann, J., Coumou, D., and Frieler, K.: Increased record-breaking precipitation events under global warming, Climatic Change, 132, 501–515, https://doi.org/10.1007/s10584-015-1434-y, 2015.

Liu, R. Y.: Moving blocks jackknife and bootstrap capture weak dependence, Exploring the limits of bootstrap, 1992.

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., Prömmel, K., Rust, H. W., Stammer, D., Ulbrich, U., Kadow, C., Köhl, A., Kröger, J., Kruschke, T., Pinto, J. G., Pohlmann, H., Reyers, M., Schröder, M., Sienz, F., Timmreck, C., and Ziese, M.: MiKlip: A National Research Project on Decadal Climate Prediction, Bulletin of the American Meteorological Society, 97, 2379 – 2394, https://doi.org/10.1175/BAMS-D-15-00184.1, 2016.

Matulla, C., Schöner, W., Alexandersson, H., von Storch, H., and Wang, X. L.: European storminess: late nineteenth century to present, Climate Dynamics, 31, 125–130, https://doi.org/10.1007/s00382-007-0333-y, 2008.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: 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, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Moemken, J., Feldmann, H., Pinto, J. G., Buldmann, B., Laube, N., Kadow, C., Paxian, A., Tiedje, B., Kottmeier, C., and Marotzke, J.: The regional MiKlip decadal prediction system for Europe: Hindcast skill for extremes and user–oriented variables, International Journal of Climatology, 27, 100 226, https://doi.org/10.1002/joc.6824, 2021.

Mullen, S. L. and Buizza, R.: The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System, Weather and Forecasting, 17, 173 – 191, https://doi.org/10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2, 2002.

Murphy, A. H.: Climatology, Persistence, and Their Linear Combination as Standards of Reference in Skill Scores, Weather and Forecasting, 7, 692–698, https://doi.org/10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2, 1992.

Nerger, L. and Hiller, W.: Software for ensemble-based data assimilation systems—Implementation strategies and scalability, Computers & Geosciences, 55, 110–118, https://doi.org/10.1016/j.cageo.2012.03.026, 2013.

Pinto, J. G., Karremann, M. K., Born, K., Della-Marta, P. M., and Klawa, M.: Loss potentials associated with European windstorms under future climate conditions, Climate Research, 54, 1–20, https://doi.org/10.3354/cr01111, 2012.

Polkova, I., Brune, S., Kadow, C., Romanova, V., Gollan, G., Baehr, J., Glowienka-Hense, R., Greatbatch, R. J., Hense, A., Illing, S., Köhl, A., Kröger, J., Müller, W. A., Pankatz, K., and Stammer, D.: Initialization and Ensemble Generation for Decadal Climate Predictions: A Comparison of Different Methods, Journal of Advances in Modeling Earth Systems, 11, 149–172, https://doi.org/https://doi.org/10.1029/2018MS001439, 2019.

Reick, C. H., Raddatz, T., Brovkin, V., and Gayler, V.: Representation of natural and anthropogenic land cover change in MPI-ESM, Journal of Advances in Modeling Earth Systems, 5, 459–482, https://doi.org/10.1002/jame.20022, 2013.

Reyers, M., Feldmann, H., Mieruch, S., Pinto, J. G., Uhlig, M., Ahrens, B., Früh, B., Modali, K., Laube, N., Moemken, J., Müller, W., Schädler, G., and Kottmeier, C.: 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, 171–187, https://doi.org/10.5194/esd-10-171-2019, 2019.

Richardson, D. S.: 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, 2473–2489, https://doi.org/https://doi.org/10.1002/qj.49712757715, 2001.

Schmidt, H. and von Storch, H.: German Bight storms analysed, Nature, 365, 791, https://doi.org/10.1038/365791a0, 1993.

Schneck, R., Reick, C. H., and Raddatz, T.: Land contribution to natural CO 2 variability on time scales of centuries, Journal of Advances in Modeling Earth Systems, 5, 354–365, https://doi.org/10.1002/jame.20029, 2013.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S. M., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events in a Changing Climate, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by IPCC, Cambridge University Press, 2021.

Sienz, F., Müller, W. A., and Pohlmann, H.: Ensemble size impact on the decadal predictive skill assessment, Meteorologische Zeitschrift, 25, 645–655, https://doi.org/10.1127/metz/2016/0670, 2016.

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Hermanson, L., Kharin, V., Kimoto, M., Merryfield, W. J., Mochizuki, T., Müller, W. A., Pohlmann, H., Yeager, S., and Yang, X.: Robust skill of decadal climate predictions, npj Climate and Atmospheric Science, 2, 1366, https://doi.org/10.1038/s41612-019-0071-y, 2019.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI–M Earth System Model: ECHAM6, Journal of Advances in Modeling Earth Systems, 5, 146–172, https://doi.org/10.1002/jame.20015, 2013.

Suarez-Gutierrez, L., Müller, W. A., Li, C., and Marotzke, J.: Dynamical and thermodynamical drivers of variability in European summer heat extremes, Climate Dynamics, 54, 4351–4366, https://doi.org/10.1007/s00382-020-05233-2, 2020.

Varino, F., Arbogast, P., Joly, B., Riviere, G., Fandeur, M.-L., Bovy, H., and Granier, J.-B.: Northern Hemisphere extratropical winter cyclones variability over the 20th century derived from ERA-20C reanalysis, Climate Dynamics, 52, 1027–1048, https://doi.org/10.1007/s00382-018-4176-5, 2019.

Wang, X. L., Feng, Y., Chan, R., and Isaac, V.: Inter-comparison of extra-tropical cyclone activity in nine reanalysis datasets, Atmospheric Research, 181, 133–153, https://doi.org/10.1016/j.atmosres.2016.06.010, 2016.

Wilks, D. S.: Chapter 8 - Forecast Verification, in: Statistical Methods in the Atmospheric Sciences, edited by Wilks, D. S., vol. 100 of *International Geophysics*, pp. 301–394, Academic Press, https://doi.org/10.1016/B978-0-12-385022-5.00008-7, 2011.