Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL107838

Decadal Predictability of Seasonal Temperature Distributions

André Düsterhus[1,2] and Sebastian Brune[3]

[1]National Centre for Climate Research (NCKF), Danish Meteorological Institute, Copenhagen, Denmark, [2]Department of Geography, Irish Climate Analysis and Research UnitS (ICARUS), Maynooth University, Maynooth, Ireland, [3]Institute of Oceanography, Center for Earth System Research and Sustainability, Universität Hamburg, Hamburg, Germany

Key Points:

- Potential of decadal prediction of temperature distributions
- Variability in prediction skill vary regionally over seasons
- The North Atlantic offers an important area where temperature distribution predictability is improved by initialization

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Düsterhus,
andu@dmi.dk

Citation:

Düsterhus, A., & Brune, S. (2024). Decadal predictability of seasonal temperature distributions. *Geophysical Research Letters*, *51*, e2023GL107838. https://doi.org/10.1029/2023GL107838

Received 20 DEC 2023
Accepted 15 MAY 2024

Author Contributions:

Conceptualization: André Düsterhus, Sebastian Brune
Data curation: Sebastian Brune
Formal analysis: André Düsterhus
Methodology: André Düsterhus, Sebastian Brune
Writing – original draft: André Düsterhus
Writing – review & editing: Sebastian Brune

Abstract Decadal predictions focus regularly on the predictability of single values, like means or extremes. In this study we investigate the prediction skill of the full underlying surface temperature distributions on global and European scales. We investigate initialized hindcast simulations of the Max Planck Institute Earth system model decadal prediction system and compare the distribution of seasonal daily temperatures with estimates of the climatology and uninitialized historical simulations. In the analysis we show that the initialized prediction system has advantages in particular in the North Atlantic area and allow so to make reliable predictions for the whole temperature spectrum for two to 10 years ahead. We also demonstrate that the capability of initialized climate predictions to predict the temperature distribution depends on the season.

Plain Language Summary The usual way to make statements about temperatures two to 10 years in advance is by using one value. This could be the average, minimum or maximum temperature over some time period. Nevertheless, this simplification hides that this represents only partial information about the full distribution of temperature values. We demonstrate that a climate model is in many areas, especially over and around the North Atlantic, better in predicting the temperature multiple years ahead than assuming a constant climate. We also show that in some areas a climate model, which is starting from a specific point of observations is better than one which does not do that. This shows that it is possible and useful to apply climate predictions to predict the future not only for averages, but for the whole distribution.

1. Introduction

Changes in distributions of temperature are a common theme in the analysis of effects of climate change (Meehl et al., 2000; Samset et al., 2019; Sherrer et al., 2005). Distributions and their changes are therein characterized by their averages and variability (Meehl et al., 2000), with considerable differences by region and seasons (Samset et al., 2019). Studies in decadal prediction regularly focus on the predictability and the prediction skill of single values, like means or percentiles of extremes. These studies have shown considerable success in the past years in predicting surface temperatures years in advance. Therein, it is common to compare initialized prediction systems with their uninitialized counter part, demonstrating significant increase in skill, in particular on regional scales (Marotzke et al., 2016; Meehl et al., 2009; Merryfield et al., 2020). Nevertheless, this advantage appeared reduced within the most recent CMIP6 model simulations (Borchert et al., 2021).

Long-term climate predictions are done for time spans of multiple years from which the single value is estimated with an average, extreme, percentile or variability measure (Boer et al., 2022). The underlying reason for this is that generally on these time scales only the prediction of statistical properties can be achieved (Schneider & Griffies, 1999). While single summarizing values offer a good first insight toward predictions, for some applications the full underlying distribution might be of interest for stakeholders from different sectors, for example, for the planning and adaptation of energy distribution, agricultural business, building activities, or health services. In the case of a normal distribution it is determined by two values (mean and variance) alone, while in the case of non-parametric distributions, stakeholder can profit from additional knowledge about the occurrence of single day values. So they might be able to benefit not only of the information on the amplitude of an extreme, but also their frequency.

We demonstrate with this study the general concept of an approach to predict and verify the full distribution of seasonal daily temperatures on decadal time scales. To compare the predictions of different simulations, the difference of two non-parametric distributions is estimated. For this, different measures exist and each measure

highlights different properties of the differences and with this show strength and weaknesses in comparing distributions depending on the application (Düsterhus & Hense, 2012; Thorarinsdottir et al., 2013). With such a difference measure it is possible to create scores for predictions, but also simply rank different simulations by their distance to a reference (Düsterhus, 2020).

In this study we investigate daily 2 m air temperature predictions with the Max Planck Institute Earth system model (MPI-ESM) for 10 lead years. After bias correcting the discrete distributions of temperature, anomalies are estimated for each 3 year period of any season. Initialized decadal prediction, uninitialized historical simulation and climatology are then compared to a reference created from the corresponding MPI-ESM assimilation simulation. For the comparison we apply a metric based on the integrated quadratic distance (IQD) (Düsterhus, 2020; Thorarinsdottir et al., 2013). The results demonstrate that prediction skill for distributions can be identified around the globe, especially in the North Atlantic sector. Additionally we show that while the annual patterns of predictability are generally similar, seasons show distinct differences especially in the northern latitudes.

2. Model Data

All simulations used in this study were done with version 1.2 of the in its low resolution setup (MPI-ESM-LR, Mauritsen et al. (2019)). It primarily consists of the atmospheric component ECHAM6 (Stevens et al., 2013) with a horizontal resolution of 1.85°, and 47 vertical levels between surface and 0.1 hPa, and the oceanic component MPIOM (Jungclaus et al., 2013) with a horizontal resolution of nominally 1.5°, and 40 vertical levels. All simulations use external forcing according to the Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al. (2016)), that is, historical forcing until 2014, and SSP2-45 scenario forcing after 2014. Both the ensemble assimilation and the initialized decadal retrospective forecasts ("hindcasts") are taken from the 80 member MPI-ESM large ensemble decadal prediction system (Hövel et al., 2022; Krieger et al., 2022), with hindcasts initialized November 1st every year from the ensemble assimilation. The system uses an oceanic ensemble Kalman filter (Brune et al., 2015) to assimilate temperature and salinity profiles from EN4 (Good et al., 2013) with the Parallel Data Assimilation Framework (Nerger & Hiller, 2013), and atmospheric nudging to ERA40, ERAInterim, ERA5 reanalyzes (Dee et al., 2011; Hersbach et al., 2020; Uppala et al., 2005). Due to this atmospheric nudging, the 2m air temperature in the ensemble assimilation very closely resembles the one from ERA reanalyzes. We therefore chose to use the assimilation simulation to create our references. The uninitialized simulations ("historical") used in this study are part of the MPI CMIP6 grand ensemble (MPI-GE CMIP6, Olonscheck et al. (2023)). For consistency reasons, we use the first 16 members of each simulation set, because the assimilation ensemble consists in total of 16 members. We investigate for all of the historical and hindcast simulations and lead years the time period 1971 until 2017 (for the boreal winter season DJF until February 2018).

3. Method

In a first step the daily data of the assimilation, the historical simulation, and for each lead year of the hindcast for each season (MAM, JJA, SON, DJF) are averaged over the time period of 1981–2010 to create the seasonal mean for each simulation. This is then subtracted from each corresponding data set to create anomalies and performing a bias correction. The distributions are created by a histogram with 200 classes between the anomalies of −20°C and +20°C (step size 0.2°C). This approach is applied for the assimilation simulation for each season between 1981 and 2010 to generate the climatology. For the hindcast simulation for three lead years, for example, lead years 2–4, 3–5, etc, the data is transformed to a distribution by the same approach. For the historical simulation, the distributions are created for the same 3 year windows. To create the reference for the comparisons, the same 3 year spans are used on the assimilation data. As a consequence, each distribution for a simulation in a specified season has the same number of entries (number of days in a season (DJF 90; MAM and JJA: 92; SON: 91) times number of years (3) times number of ensemble members (16)).

To compare two distributions we apply integrated quadratic distance (IQD) (Düsterhus, 2020; Thorarinsdottir et al., 2013). In this the distance of two discretized distributions (f and g) at given points v_i is measured by the squared distance of their cumulative distribution functions (cdfs, F and G). The distance is then given by

$$D_{IQD}(f,g) = \frac{1}{n_b}\sum_{i=1}^{n_b}(F(v_i) - G(v_i))^2. \quad (1)$$

In this study we will compare the climatological, hindcast and historical distributions for each grid point against the corresponding distribution from the reference. Thus a distance value is created for each 3 year period for each simulation at each grid point. Our evaluation is based on the comparison of two simulation for each grid point and for each 3 year time period. To evaluate over the whole time period for each grid point, two simulations are compared and the number of years counted where the distance of one simulation is lower to the reference distribution than the other. To calculate the significance by the 5% significance level the result is compared to a random walk (DelSole & Tippett, 2016). By this approach we prevent that a single outlier year will dominate our analysis and answer for each grid point the question how often a specific simulation is better than another over the evaluation period.

4. Results

4.1. Summer Prediction

To demonstrate the overall capability of the simulations, we investigate in Figure 1 the prediction skill of temperature distributions for the global scale for June-July-August (JJA). In this we compare in a first step the hindcast prediction for different lead times against climatology. For lead year 2–4 (Figure 1a) we see that the hindcast has advantages compared to the climatology in the North Atlantic area, the Indian Ocean area, and in the Western Pacific. We identify that especially in the Arctic region the constant climatology assumption is superior. For longer lead times (Figures 1b and 1c) the general patterns persist and only minor changes happen. When we look at the historical simulation (Figures 1d–1f), which is for all lead times in this setup the same, the patterns are comparable, but show distinctive differences in shape in some areas. Investigating the effect of initialization by comparing the hindcasts with historical simulations (Figures 1g–1i) shows an advantage of the hindcasts mainly in the Southern Ocean and in the North Atlantic, while the tropical areas are mainly indecisive. In the Arctic region, the historical simulation is first superior in lead years 2–4 (Figure 1g) before becoming neutral in lead years 5–7 (Figure 1h). On time scales of lead year 8–10 (Figure 1i), the hindcast is then significantly superior compared to the historical simulations. Areas on land where the hindcast is superior against the historical simulation are India, West Arabia and North-Eastern Brasil. We also can see some added skill in Northern Europe.

As the North Atlantic area and Europe are one of the prime areas of skill, we investigate this area in Figure 2 in more detail. For the hindcast (Figures 2a–2c) and historical (Figures 2d–2f) simulations we see vast areas being better predictable when compared to climatology. This is especially true for the subpolar gyre region and Central Europe. The climatology dominates especially in areas affected by the Gulf Stream and on the Western side of Northern Africa. When we compare the hindcast and the historical simulations, we identify for lead years 2–4 (Figure 2g) the Northern part of the North Atlantic and Northern Europe as areas where the initialization in particular shows clear advantages. Skill reduces over longer lead times and we can identify for lead years 5–7 (Figure 2h) an area in the North East Atlantic as predictable. The skill in Southern Scandinavia as well as the British and Irish Isles persist up to lead years 8–10 (Figure 2i).

4.2. Predictions for Different Seasons

After focusing on the boreal summer (JJA) we now investigate in a next step the skill for different seasons. Here we concentrate in Figure 3 on the comparison between hindcast and historical simulations for Europe. In the boreal spring (MAM, Figures 3a–3c), we see a vast area east of Greenland over the Denmark Strait where the initialization creates additional value, which can be attributed to sea ice. Also compared to the predictions for the boreal summer, the Eastern North Atlantic area is more predictable with the hindcast simulation for lead years 2–4 (Figure 3a). While this prediction skill reaches down to 35°N for the early lead years, it covers only to about 45°N for lead years 8–10 (Figure 3c). In contrast, the prediction skill for the hindcast simulation is much less pronounced on the land areas. Especially in Scandinavia there is much less area covered by significant skill than it was in the summer. Just like the boreal spring, the boreal autumn (SON, Figures 3d–3f), also demonstrates a pattern in the Denmark Strait, but for a much narrower band along the Greenland Coast. In contrast to the boreal spring, there is a clear indication that on the West side of Greenland over the Baffin Bay, the historical simulations show more skill than the hindcasts. Also in the lower latitudes the historical simulations show their advantages. The subpolar gyre region can be much better compared in its pattern toward the boreal summer. In Scandinavia none of the two simulations show any clear indication of significance. The boreal winter (DJF, Figures 3g–3i) is

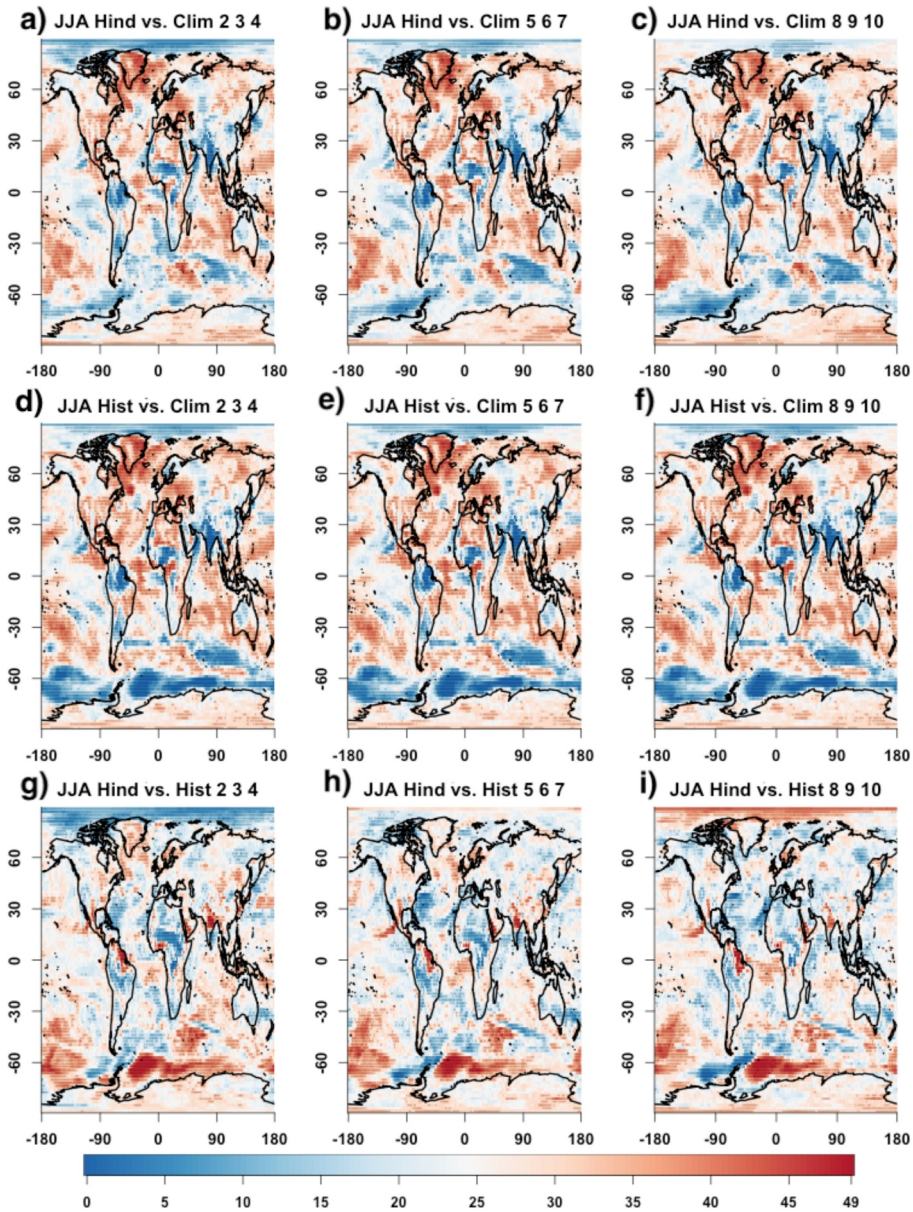

Figure 1. Global comparison between initialized decadal prediction and uninitialized historical simulation applying the 1d-continuous-IQD-score on daily 2 m temperature distributions in the boreal summer. Analyzed are the years between 1971 and 2017 over three different lead times (lead years 2–4 (1st column), 5–7 (2nd column) and 8–10 (3rd column)) and comparisons between hindcast and climate (1st row), historical and climate (2nd row) and hindcast and historical simulation are shown. The results show the number of years in which the score of the distribution of daily mean temperature anomaly was better compared to the assimilation simulation in the first two rows and in the last row the initialized decadal prediction system was better compared to uninitialized historical simulations. Black dots indicate significance by the 5% significance level estimated by comparison to a random walk process.

on the West side of Greenland more comparable toward the boreal autumn, while on the east side it resembles the pattern of the boreal spring. In the subpolar gyre region significant prediction skill is seen for all three seasons for all lead times, but with different patterns. For longer lead years, like lead years 8–10 (3 i), it is the only time that the historical simulations outweigh the hindcasts over Scandinavia. Overall it can be shown, that there is a seasonal cycle of the shown differences between hindcast and historical simulation over the North Atlantic sector. The results in the mentioned regions in the higher latitudes might be explained by differences within the

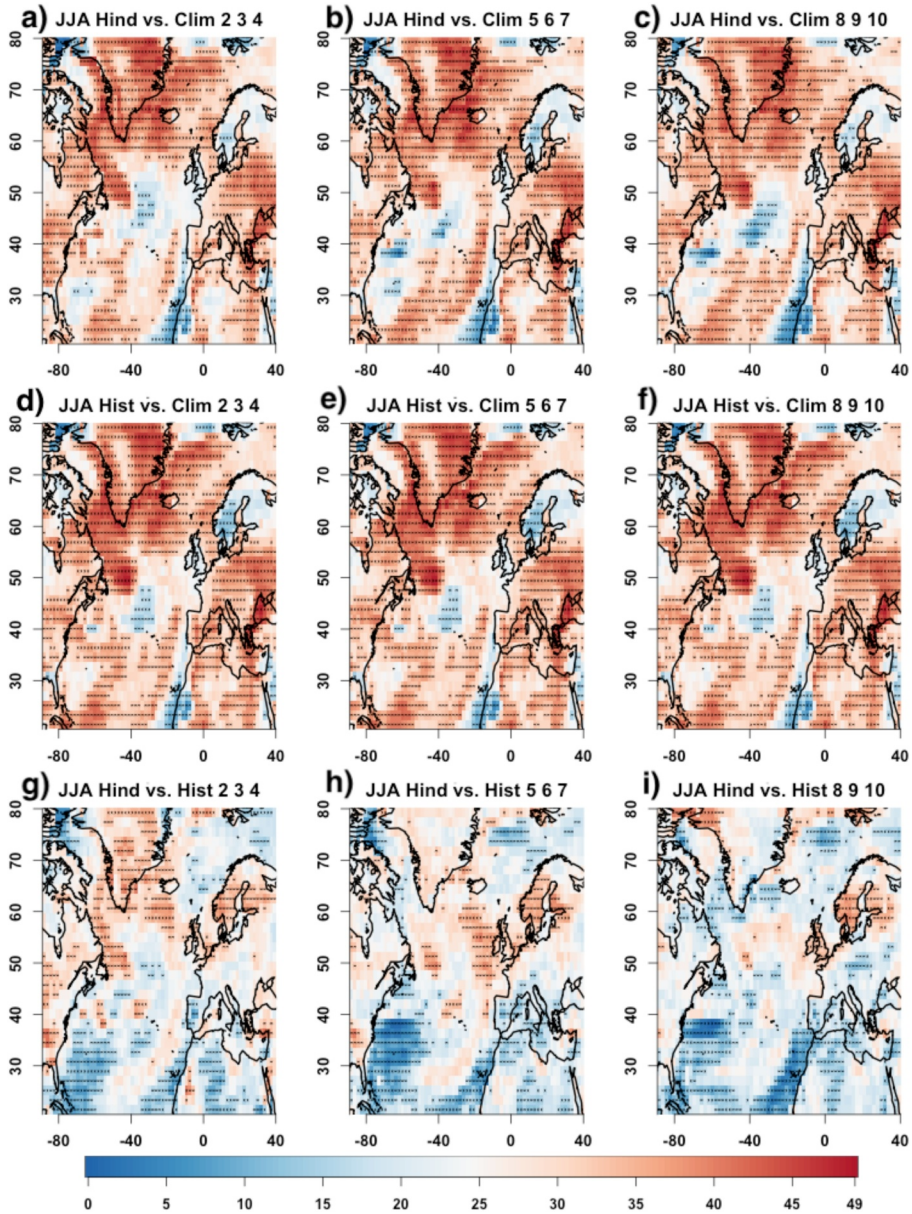

Figure 2. Comparison between decadal prediction and historical simulation over the North Atlantic sector applying the 1d-continuous-IQD-score on daily 2m temperature distributions in the boreal summer. Analyzed are the years between 1971 and 2017 over three different lead times (lead years 2–4 (1st column), 5–7 (2nd column) and 8–10 (3rd column)) and comparisons between hindcast and climate (1st row), historical and climate (2nd row) and hindcast and historical simulation are shown. The results show the number of years in which the score of the distribution of daily mean temperature anomaly was better compared to the observations in the decadal prediction system than in the historical simulation. Black dots indicate significance by the 5% significance level estimated by comparison to a random walk process.

simulations in its capacity to represent snow and sea ice, in particular during melting and freezing times throughout the year.

4.3. Example From the North Atlantic

To further dive into the comparison of the different prediction, we chose a grid point in the North East Atlantic (20°W, 60°N) from the boreal summer prediction. This point is taken from a region where the hindcast prediction is significantly better than the other two simulations for lead years 2–4. For each year we see in Figure 4 the distribution, with the inter-quantile range (25%–75%) shown by a block and the distance toward the 5th and 95th

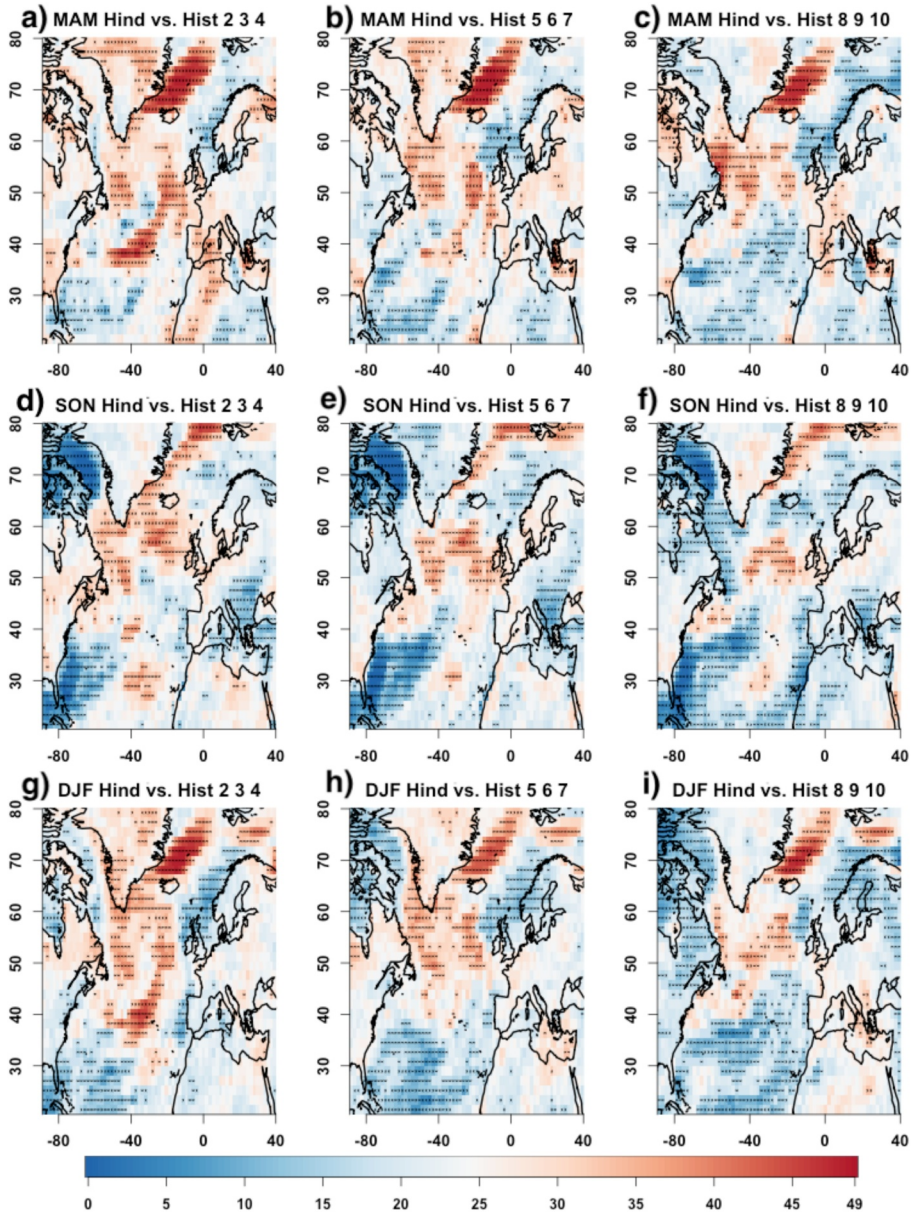

Figure 3. Comparison between decadal prediction and historical simulation over the North Atlantic sector applying the 1d-continuous-IQD-score on daily 2m temperature distributions for three seasons (MAM (1st row), SON (2nd row), DJF (3rd row)). Analyzed are the years between 1971 and 2017 over three different lead times (lead years 2–4 (1st column), 5–7 (2nd column) and 8–10 (3rd column)). The results show the number of years in which the score of the distribution of daily mean temperature anomaly compared to the assimilation in the initialized decadal prediction system was better than in the uninitialized historical simulation. Black dots indicate significance by the 5% significance level estimated by comparison to a random walk process.

percentile by whiskers. The reference assimilation shows generally an upward trajectory, but also demonstrates some variability over time. In the middle of the 1970 and 1990s, there are weak negative dents in the trend, while in the late 2000s and early 2010s a positive anomaly can be identified. For most times, e. g. highlighted especially around the year 2000, the assimilation exhibits that the lower whiskers are longer than the upper ones, indicating a non-symmetric distribution. In general, this evolution over time is better represented with the hindcasts than in the historical simulations. By analyzing in how many years one simulation is better than the other we get a good indication about the quality of a particular prediction. The analysis shows that hindcast (40–7) and historical (43–4) are clearly better than the climatology at this grid point. We also find that the hindcast shows better results than

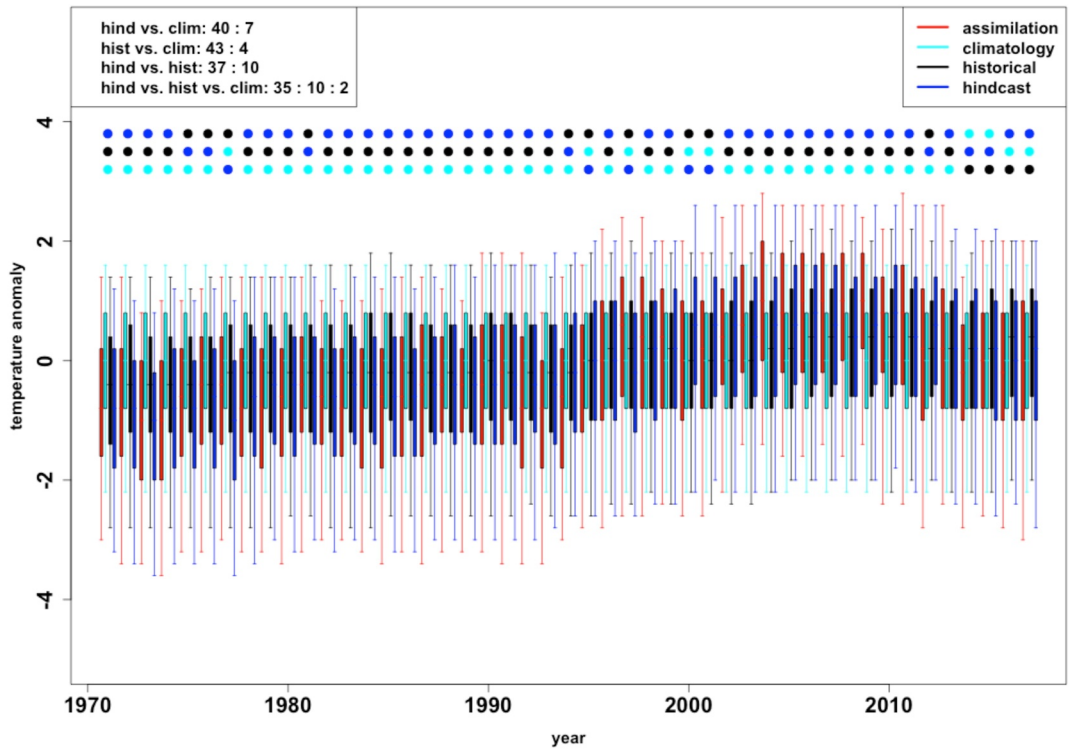

Figure 4. Comparison of the distributions of daily temperature data in boreal summer for a grid point in the North Atlantic (20°W; 40°N). Shown are for each year the distribution of the hindcast (blue) with lead times 2–4 years and historical simulation (black), climatology (cyan) and assimilation reference (red) between 1971 and 2017. Dots at the top demonstrate the ranking (top is higher) of hindcast (blue), historical (black) or climatology (cyan) for each year applying the 1d-continuous-IQD-score The upper left legend gives the statistics of the analysis.

the historical in the majority of the years (37–10). Comparing hindcast, historical, and climatology in a three-way-tie, we find that hindcast performs best in 35 years, historical performs best in 12 years, and climatology only in two years (35–10–2). This grid point example demonstrates a potential application, comparable to the common scatter plots, for the deeper understanding for correlation based measures. We like to highlight that analyzing single grid points without a physical context is not sufficient enough to understand the implications of predictability for a given area.

5. Discussion and Conclusion

This study demonstrates the ability of initialized decadal simulations to predict 2m air temperature reliably better than climatology could do, on a global scale, and in particular in the North Atlantic region. Furthermore it shows an improved prediction skill compared to uninitialized historical simulations in the Northern parts of the North Atlantic and Europe. This has already been shown with predecessor systems based on MPI-ESM (Marotzke et al. (2016); Polkova et al. (2019); Brune and Baehr (2020)). Compared to their investigations, which were based on the commonly used anomaly correlation (ACC), the bias and its correction plays an important role in our investigation. Our metric of the 1D-IQD takes the deviation from the mean into account for the evaluation, similar to, for example, an analysis of the root-mean-square error. By using a simple bias correction scheme as done in this study, we can focus on the prediction of the variability of the target variable. The correction might be challenging for variables with specified boundaries like precipitation on short time scales. The bias correction is also the main explanation when we see with increasing lead times increased relative skill of initialized over uninitialized simulations. Especially, such phenomena can be observed when bias correction is applied to data sets and regions with different trends or cycles. While generally such results are counter-intuitive, Düsterhus and Brune (2024) have recently demonstrated that the assumption that an initialized hindcast regresses back to the uninitialized historical simulation is not necessarily applicable, not even over longer lead times.

We note that for metrics based on the correlation the results can be dominated by single outliers. This is prevented in our approach by the yearly evaluation of the results, which is then aggregated by a sum. In situations of Gaussian or symmetrically distributed variables, it can be assumed that the analysis will reach similar results to an ACC analysis on seasonal mean values (see Figures S1A–S3 in Supporting Information S1). Nevertheless, we also find areas with considerable differences, for example, over Scandinavia the difference in prediction skill between initialized decadal predictions and uninitialized historical simulations based on 1D-IQD (Figures 2g–2i) is generally higher than the difference in skill based on ACC (Figures S1d–S1f in Supporting Information S1). We highlight that this is not only an effect due to the non-normality of the temperature distributions, but also the different skill measures. Furthermore, we have analyzed the skill on the grid point level. It is therefore important to state that statistical significance can only be a first step in the determination of predictability, while subsequent strains of argument have to be drawn by physical reasoning.

We apply the inter-quartile distance to investigate the difference between two distributions, which evaluates how much the probability needs to be shifted to transform one distribution into another. This approach allows us not only to focus on an ensemble mean, but all ensemble members. It is important to take into account that the assimilation simulation we use as a reference has by design a smaller variability than the other simulations, even when we use the same amount of ensemble members. Nevertheless, it is still a valid approach, as the inter-quartile distance investigates not each bin separately, but the transfer of probability (also known as binning-problem, Düsterhus and Hense (2012)). This is also an important advantage of the IQD compared with other metrics that compare distributions with each other. One major difference to the correlation based metrics is our evaluation procedure by giving the overall results for a time series by the number of three-year-periods one simulation is better than the other. As a consequence we do not receive a continuous result, but a categorical one, posing a challenge to existing controls for false alarm rates, for example, like Wilks (2016). Having tested our results with this test, we conclude that the outcomes can hardly be interpreted the same way as correlation based metrics are.

Daily mean 2 m air temperature shown here as a target variable is only a first step toward a better understanding of how well simulations, be it initialized decadal predictions or uninitialized historical simulations, are able to predict the full distribution of the climate system. For stakeholders, absolute temperature distributions on time scales shorter than a daily mean might be of interest, like currently provided as minimum or maximum temperature. Other variables could be of interest for stakeholder like the renewable sector, especially wind. The full advantage of this approach can be expected for generally non-normal distributed variables. The advantage of this non-parametric approach is that it can be applied to most variables predicted by the climate model and is only limited by the capacity of the evaluating computer system and the ability to properly bias-correct the simulations.

Data Availability Statement

The daily 2m air temperature of the decadal prediction system (Brune et al., 2021) is publicly available from the Long-term Archiving of Climate Model Data at WDC Climate and DKRZ. The daily 2m air temperature from the MPI-GE CMIP6 uninitialized simulations (historical and ssp245) are publicly available from DKRZ's ESGF server at https://esgf-data.dkrz.de/search/cmip6-dkrz/ by specifying the following search criteria: Source ID: MPI-ESM1-2-LR; Institution ID: MPI-M; Experiment ID: historical, ssp245; Variant Label: r1i1p1f1 to r16i1p1f1, Frequency: day, Variable: tas.

Acknowledgments

A.D. is supported by A4 (Aigéin, Aeráid, agus athrú Atlantaigh), funded by the Marine Institute (grant: PBA/CC/18/01) and received funding from the Danish state through the National Centre for Climate Research (NCKF). S.B. is supported by the German Ministry of Education and Research (BMBF) under Grant 01LP2327A (project Coming Decade), and by Copernicus Climate Change Service, funded by the EU, under contract C3S2-370. We thank the staff at the German Climate Computing Center (DKRZ), Hamburg and the Max Planck Institute for Meteorologie, Hamburg for their support. All simulations analyzed in this study have been run and processed on DKRZ high performance computers.

References

Boer, G. J., Sospedra-Alfonso, R., Martineau, P., & Kharin, V. V. (2022). Verification data and the skill of decadal predictions. *Frontiers in Climate*, *4*, 836817. https://doi.org/10.3389/fclim.2022.836817

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., & Mignot, J. (2021). Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophysical Research Letters*, *48*(3), e2020GL091307. https://doi.org/10.1029/2020GL091307

Brune, S., & Baehr, J. (2020). Preserving the coupled atmosphere-ocean feedback in initializations of decadal climate predictions. *WIREs Clim. Change*, *11*(3), e637. https://doi.org/10.1002/wcc.637

Brune, S., Nerger, L., & Baehr, J. (2015). Assimilation of oceanic observations in a global coupled Earth system model with the SEIK filter. *Ocean Modelling*, *96*(2), 254–264. https://doi.org/10.1016/j.ocemod.2015.09.011

Brune, S., Pohlmann, H., Müller, W. A., Nielsen, D. M., Hövel, L., & Baehr, J. (2021). MPI-ESM-LR_1.2.01p5 decadal predictions localEnKF: Daily mean values. [Dataset]. *DOKU at DKRZ*. Retrieved from http://hdl.handle.net/hdl:21.14106/f2fdc61b13828ed5284f4e4ab41e63f8a84c6e52

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

DelSole, T., & Tippett, M. K. (2016). Forecast comparison based on random walks. *Monthly Weather Review*, *144*(2), 615–626. https://doi.org/10.1175/MWR-D-15-0218.1

Düsterhus, A. (2020). Seasonal statistical–dynamical prediction of the North Atlantic oscillation by probabilistic post-processing and its evaluation. *Nonlinear Processes in Geophysics*, *27*(1), 121–131. https://doi.org/10.5194/npg-27-121-2020

Düsterhus, A., & Brune, S. (2024). The effect of initialisation on 20 year multi-decadal climate predictions. *Climate Dynamics*, *62*(2), 831–840. https://doi.org/10.1007/s00382-023-06941-1

Düsterhus, A., & Hense, A. (2012). Advanced information criterion for environmental data quality assurance. *Advances in Science and Research*, *8*(1), 99–104. https://doi.org/10.5194/asr-8-99-2012

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research*, *118*(12), 6704–6716. https://doi.org/10.1002/2013JC009067

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hövel, L., Brune, S., & Baehr, J. (2022). Decadal prediction of marine heatwaves in MPI-ESM. *Geophysical Research Letters*, *49*(15), e2022GL099347. https://doi.org/10.1029/2022GL099347

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., 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), 422–446. https://doi.org/10.1002/jame.20023

Krieger, D., Brune, S., Pieper, P., Weisse, R., & Baehr, J. (2022). Skillful decadal prediction of German Bight storm activity. *Natural Hazards and Earth System Sciences*, *22*(12), 3993–4009. https://doi.org/10.5194/nhess-22-3993-2022

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., et al. (2016). MIKLIP—A national research project on decadal climate prediction. *Bulletin of the American Meteorological Society*, *97*(12), 2379–2394. https://doi.org/10.1175/BAMS-D-15-00184.1

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., 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), 998–1038. https://doi.org/10.1029/2018MS001400

Meehl, G. A., Goddard, L., Murphy, J. M., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486. https://doi.org/10.1175/2009BAMS2778.1

Meehl, G. A., Karl, T., Easterling, D. R., Changnon, S., Pielke, Jr., R., Changnon, D., et al. (2000). An introduction to trends in extreme weather and climate events: Observations, socioeconomic impacts, terrestrial ecological impacts, and model projections. *Bulletin of the American Meteorological Society*, *81*(3), 413–416. https://doi.org/10.1175/1520-0477(2000)081⟨0413:AITTIE⟩2.3.CO;2

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Coelho, C. A. S., Danabasoglu, G., et al. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896. https://doi.org/10.1175/BAMS-D-19-0037.1

Nerger, L., & Hiller, W. (2013). 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

Olonscheck, D., Suarez-Gutierrez, L., Milinski, S., Beobide-Arsuaga, G., Baehr, J., Fröb, F., et al. (2023). The new Max Planck Institute grand ensemble with CMIP6 forcing and high-frequency model output. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003790. https://doi.org/10.1029/2023MS003790

Polkova, I., Brune, S., Kadow, C., Romanova, V., Gollan, G., Baehr, J., 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), 149–172. https://doi.org/10.1029/2018MS001439

Samset, B. H., Stjern, C. W., Lund, M. T., Mohr, C. W., Sand, M., & Daloz, A. S. (2019). How daily temperature and precipitation distributions evolve with global surface temperature. *Earth's Future*, *7*(12), 1323–1336. https://doi.org/10.1029/2019EF001160

Schneider, T., & Griffies, S. M. (1999). Conceptual framework for predictability studies; schneider and griffies. *Journal of Climate*, *12*(10), 3133–3155. https://doi.org/10.1175/1520-0442(1999)012⟨3133:ACFFPS⟩2.0.CO;2

Sherrer, S. C., Appenzeller, C., Liniger, M. A., & Schär, C. (2005). European temperature distribution changes in observations and climate change scenarios. *Geophysical Research Letters*, *32*(L19705). https://doi.org/10.1029/2005GL024108

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M Earth system model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534. https://doi.org/10.1137/130907550

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., et al. (2005). The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, *131*(612), 2961–3012. https://doi.org/10.1256/qj.04.176

Wilks, D. S. (2016). The stippling shows statistically significant grid points—How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1