North Atlantic Circulation Variability and Teleconnections to Surface Climate Conditions in CMIP6 Models

Master's Thesis
MSc Climate Change

Anna Ida Katharina Kirchner
December 20^{th} 2022

Supervised by
Jens Hesselbjerg Christensen
and Ruth Mottram (Danish Meteorological Institute)

University of Copenhagen
Niels Bohr Institute - Physics of Ice, Climate and Earth

Abstract

Weather conditions and extreme events in Europe are strongly influenced by the state of the atmospheric circulation over the North Atlantic. Typical modes of circulation variability can be identified, with the North Atlantic Oscillation (NAO) being the dominant mode, driving surface conditions in the region and further away through teleconnections connected to changes between its positive and negative phase. There is a high interest in the scientific community in improving predictability of the NAO and other modes of circulation variability, especially in the context of anthropogenic climate change. The understanding of the mechanisms driving circulation variability can be improved by studying them under different climate conditions with the help of climate models or proxy-based reconstructions, that rely on the observed relationship between circulation variability and surface conditions. This work addresses the need to verify the correct representation of modes of variability in climate models and the robustness of their teleconnection patterns in space and time. It aims to investigate how state-of-the-art Global Circulation Models models represent different modes of North Atlantic circulation variability and their relationship with surface climate conditions under current climate conditions, and how these modes and teleconnections can be captured and compared. We use an Empirical Orthogonal Function analysis to identify different modes of circulation variability in the winter geopotential height field over the North Atlantic and determine their surface temperature and precipitation response patterns in the domain. We investigate the spatial patterns of the modes and their teleconnections in a subset of different CMIP6 models historical runs and explore approaches to evaluate them against reanalysis data. The models show a high ability to represent three distinct modes of variability and realistic corresponding temperature and precipitation responses. Notable differences between the models provide insights into the performance of different models, potential drivers of these differences, and the role of natural spatial variability of the modes and its impacts on teleconnections. A robust evaluation of the models is prevented by limitations of the used methods and the lacking consideration of uncertainties connected to decadal variability.

Acknowledgements

Large thanks go to my supervisor Jens for providing the inspiration and the approach for this work. Thank you for a long journey of introducing me to the analysis of atmospheric variability, for taking the time for many hours of supervision and discussion and teaching me so many things along the way. I thoroughly enjoyed the process and am grateful for what I learned from you. I would also like to thank Ruth for agreeing to be my co-supervisor and providing fresh perspectives and great enthusiasm for my research.

Much is owed to my friends and family for all their support both academically and emotionally, for supplying me with enough snacks to get through this and taking my mind of this thesis when I needed it. Thanks to Anna for being a good office mate and friend through the whole thesis process and the best company in long nights at the office. Lastly, thank you Julian for your invaluable support on all fronts and the time and patience you dedicated to supporting me and this project.

Contents

List of Figures

Acronyms

AR Atlantic Ridge

CMIP6 Coupled Model Intercomparison Project Phase 6

EOF Empirical Orthogonal Functions

GCMs Global Circulation Models

GHG Greenhouse Gas

IPCC Intergovernmental Panel on Climate Change

NAO North Atlantic Oscillation

PC Principal Component

PCR Principal Component Regression

SB Scandinavian Blocking

SVD Singular Value Decomposition

Z500 geopotential height at 500 hPa

1 Introduction

The weather conditions experienced in Europe are strongly influenced by the state of the atmosphere over the North Atlantic, intensifying, diverting or blocking the westerly flow and thus the transport of humidity and temperatures across the ocean. The most important mode of circulation variability in this region is the North Atlantic Oscillation (NAO). The swings between its positive and negative phase, characterised by a strengthening and weakening of the north-south pressure gradient over the North Atlantic, can explain the largest part of climate variability in the North Atlantic region in all seasons. This relationship applies both to average climate conditions and extreme events. Extreme states of the NAO have been linked to heatwaves and cold spells over Europe (Beniston 2019; Magnusson et al. 2022), as well as storms (Magnusson et al. 2022) and droughts (Seneviratne et al. 2021). For example, the winter of 2009/10 was characterized by an extremely strong and persistent negative phase of the NAO, causing severe cold spells in northern and western Europe (Cattiaux et al. 2010). This connection increases in relevance in the context of anthropogenic climate change, which is expected to increase the intensity and frequency of extreme events across the globe (Seneviratne et al. 2021). Other recurrent patterns of atmospheric variability influencing climate conditions and extreme events in the region are blocking events, characterised by persistent quasi-stationary anticyclones blocking and diverting the westerly flow (Davini et al. 2012).

Despite their importance for European weather, the dominant patterns of atmospheric circulation variability are hard to predict and their connections not entirely understood (Hurrell 2015). For the sake of improving predictability of weather and extreme events, especially under future climate changes, there is a high interest in the scientific community to further understanding of North Atlantic atmospheric circulation variability and its relationships with climate variables on the continents around the North Atlantic basin Hurrell et al. 2003.

This can be achieved by studying these mechanisms under past, current and future climate changes. The GreenPlanning project, a research collaboration between the Niels-Bohr-Institute, University of Copenhagen and the Geological Survey of Denmark and Greenland, with further cooperation with the Danish Meteorological Institute (DMI) and the Universities of Bergen, Norway and Reykjavik, Iceland, aims to contribute to this by reconstructing a time series of North Atlantic climate variability over the past 2000 years. It aims to improve seasonal to decadal predictions of European climate by studying the relationships between atmospheric circulation variability, melt water from the Greenland ice sheet and ocean circulation in the past (Christensen n.d.). Reconstructions like this are based on proxy records found on land, e.g. ice cores, tree rings or speleotherms, and exploit the known teleconnections between weather conditions and atmospheric variability. Therefore, they are connected to a need to explore the robustness of these teleconnections under current and changing climate conditions (Pinto et al. 2012). Besides through reconstructions, modes of circulation variability, their teleconnections and their behavior under climate change can also be studied with the help of climate models, simulating the response of the climate system under stable and changing conditions, like increased greenhouse gas forcing. Connected to this, there is a need to verify that these models are capturing the dynamics related to these modes of variability correctly under current climate conditions.

This work is based on these research needs. It aims to assess the representation of the dominant modes of North Atlantic circulation variability and their teleconnections to surface climate conditions in state of the art Global Circulation Models (GCMs). To this end, we use an Empirical Orthogonal Functions (EOF) analysis of the geopotential height field in the extended winter season over the North Atlantic region to identify the most important modes of atmospheric variability. A regression of temperature and precipitation variability onto the Principal Component (PC) time series of these modes of variability then allows to characterize the typical teleconnections associated with each mode. This analysis is applied both to the ERA5 reanalysis dataset covering the period from 1959 to 2014 and historical runs of different GCMs from the Coupled Model Intercomparison

Project Phase 6 (CMIP6). The models are evaluated based on differences in their spatial pattern of variability and teleconnections to the reanalysis data.
This analysis is carried out with the aim of answering two questions. First, different CMIP6 models are compared and evaluated against reanalysis data to investigate *how modes of winter North Atlantic circulation variability and their teleconnections to surface temperature and precipitation are represented in CMIP6 models*. We hope to gain insights into the suitability of these models for research relating to modes of variability and the robustness of their relationship to surface climate conditions. Second, the used methods are critically evaluated in order to determine *what methods are suitable to capture the representation of modes of variability and teleconnections and compare them between models*.

2 Background

Within the atmospheric circulation, large-scale modes of variability can be found, that are defined by recurrent spatial patterns and unique time scales of variability (IPCC 2021). They are the result of internal variability in the atmosphere and interactions with other components of the climate system, especially the ocean circulation, and influence climate variability and surface conditions locally and remotely, through teleconnections. Thus, the climate variability in a particular region, especially on seasonal to multi-decadal time scales, can largely be explained by the states of one or the combination of several typical modes of climate variability in this region (IPCC 2021). This work is interested in this relationship between surface climate conditions and modes of circulation variability in the North Atlantic sector. Thus, the following chapter serves as an introduction into the known modes of variability and their teleconnection patterns over the North Atlantic region.
The dominant mode of variability in the North Atlantic region is the North Atlantic Oscillation (NAO). It is characterized by variations in sea level pressure between a low pressure center in the high latitudes around Iceland and subtropical high pressure, typically located close to the Azores. It oscillates on irregular timescales between a higher than usual pressure difference caused by an anomalously low Icelandic Low and strengthened Azores High, denoted as its positive phase, and a weaker pressure gradient, representing the negative phase. The strength of this pressure gradient influences the strength and position of the jet stream and North Atlantic storm track and thus directly links the phases of the NAO to typical surface climate conditions on the surrounding continents. A positive phase of the NAO for instance is associated with mild and wet conditions in north western Europe, brought in by the westerly flow that is directed along the pressure gradient towards the British isles and Scandinavia, leading to cold and dry anomalies south of it. During a negative phase of the NAO, the westerly flow is displaced southwards, bringing moisture and storminess towards southern Europe and allowing the influence of cold and dry easterly winds in northern Europe Thus, the phases of the NAO control a large fraction of wind, storminess, temperature and precipitation variability over the North Atlantic and the surrounding continents, with this influence being strongest in boreal winter (Hurrell et al. 2003; IPCC 2021). These teleconnection patterns have been observed for centuries, making the NAO one of the oldest known modes of variability. For example, Danish missionary Hans Egede noted in 1785 the inverse relationship between the severity of winters in Denmark and Greenland, where anomalously cold winters in Denmark correlated with rather mild conditions in Greenland and vice versa (Stephenson et al. 2003).
The temperature and precipitation teleconnections of the NAO over the whole North Atlantic region in winter, which we aim to investigate in this work, can be summarised as follows: The temperature response to the winter NAO shows a quadripolar pattern. A positive phase of the NAO is connected to warm temperature anomalies over northwestern Europe and the eastern United States, and cold temperature anomalies around the Mediterranean and the Labrador sea (Hurrell et al. 2003; IPCC 2021). The precipitation response is dominated by the north-south displacement of the storm track, where a positive NAO phase leads to increased precipitation from Iceland to northwestern Europe and negative precipitation anomalies in southern Europe, the Mediterranean region and around the Labrador sea (Hurrell et al. 2003). The precipitation pattern is subject to local orographic conditions, so that the mountainous coasts of Scotland and Norway show the strongest precipitation response and locations in the lee of these mountains can experience an opposite signal (Burt et al. 2013; Uvo 2003).
Despite being an important driver of climate variability, the NAO does not explain the full range of observed variations in climate conditions around the North Atlantic basin. Another atmospheric phenomenon impacting the westerly flow across the North Atlantic and thus climate conditions around it, is atmospheric blocking. Blocking events are characterized as quasi-stationary anticyclones persisting over unusually long timescales, that block the westerly flow and lead to persisting weather conditions (Davini et al. 2012). Blocking has been linked to extreme weather events, for

example by Buehler et al. (2011), showing a connection between an increased amount of blocking events over the winter North Atlantic and cold and dry spells over central to eastern Europe in ERA-40 reanalysis data. Atmospheric blocking, especially over Greenland, is closely related to the negative phase of the NAO, shown by a strong inverse correlation between Greenland blocking frequency and the NAO index (Woollings et al. 2008; Davini et al. 2012).

Considering these impacts, understanding North Atlantic atmospheric circulation variability, particularly the NAO, and its teleconnection patterns continues to be of high interest to the scientific community. There is an extensive body of literature aimed at defining the NAO, characterizing its relationship to climate conditions through teleconnections, improving its predictability and exploring its response to present and future climate changes (Hurrell et al. 2003).

The NAO is usually defined based on the difference in sea level pressure or geopotential height between its northern and southern center of action. The simplest approach is calculating a station-based index, indicating the gradient between locations in Iceland and on the Azores or in Portugal at different points in time. An alternative approach is an Empirical Orthogonal Functions (EOF) analysis, reducing the pressure or geopotential height field to its dominant modes of variability (IPCC 2021). Taking the full gridded field into account, this method additionally allows an exploration of the spatial pattern of the variability (Hurrell et al. 2003). This approach allows an exploration of the circulation variability beyond the NAO, as it also returns further patterns of variability that explain smaller fractions of the total observed variability. Studies considering EOF analysis of the pressure/geopotential height field in the North Atlantic sector in boreal winter consistently find the leading EOF pattern to correspond to the NAO pattern, explaining the largest fraction of the observed variance (Hurrell et al. 2003; IPCC 2021). Further EOF patterns, accounting for a smaller share in the explained variance, have been identified as East Atlantic or Atlantic ridge pattern with a positive pressure/geopotential height anomaly over the central North Atlantic in its positive phase, and Scandinavian blocking, with a positive anomaly over northwestern Europe in its positive phase (Hurrell et al. 2003; Ruggieri et al. 2020). However, the physical interpretability of EOF patterns is limited by the mathematical definition of EOFs, making them uncorrelated and orthogonal to each other (Hannachi et al. 2007). Similar pattern-based statistical techniques, such as rotated EOF analysis and cluster analysis have been used to alleviate some of the shortcomings of the EOF analysis (Michelangeli et al. 1995; Hannachi et al. 2007; IPCC 2021). However, the application of those techniques to the North Atlantic region tends to identify the same patterns of variability in boreal winter, i.e., the positive and negative state of the NAO, the Atlantic ridge pattern and the Scandinavian blocking pattern (Strommen et al. 2019; Ferranti et al. 2015; Delgado-Torres et al. 2022). Figure 1 shows the spatial structure of these four patterns identified as the four dominant quasi-persistent weather regimes over the Euro-Atlantic domain with k-means clustering to ERA-Interim reanalysis data by Strommen et al. (2019).

The teleconnections of the NAO can be explored based on the relationship between a time series of the variable in question and the NAO index (Hurrell 2015). This can be done for specific locations by means of correlation between the NAO index and observed conditions at weather stations (e.g. Hurrell 1995), or at every grid point of a whole field (e.g. Uvo 2003; Hurrell 2015).

The knowledge on teleconnections of the NAO can be used to improve understanding of the variability of the NAO itself. The known links between the phases of the NAO and its influence on the conditions on the continents around the North Atlantic allow reconstructions of the NAO index based on proxy data found in different archives, such as ice cores, tree rings or speleotherms (e.g. Cook 2003; Pinto et al. 2012). Based on this information on past NAO variability, researchers hope to improve their understanding of the underlying mechanisms controlling the variability of the NAO. It has been suggested to be linked to ocean circulation, coupling with the cryosphere, internal variability in the atmosphere, such as Rossby wave breaking, teleconnections to processes in other parts of the globe, or external forcings (Pinto et al. 2012; Hurrell 2015). The goal of understanding the driving processes behind NAO variability is to develop the ability of predicting the NAO and thus the climate impacts associated with it (e.g. Smith et al. 2020). In recent years,

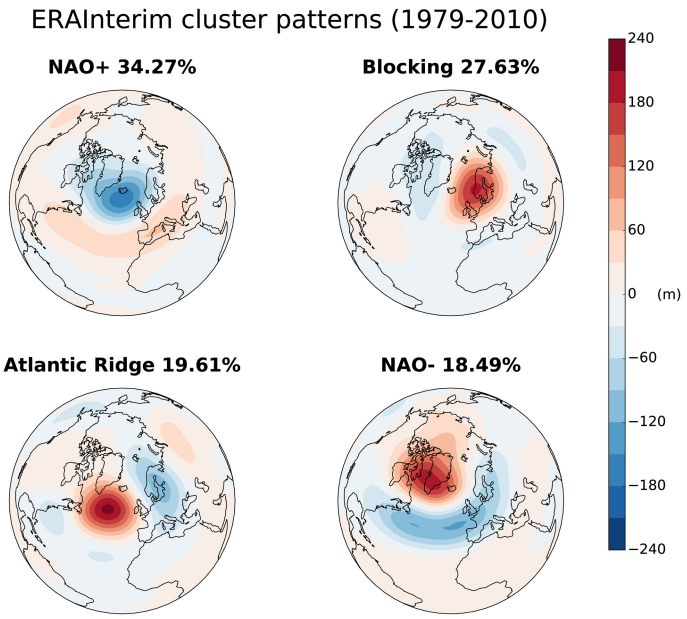


Figure 1: **Example of the spatial structures of the dominant patterns of variability found over the North Atlantic domain in winter.** Figure taken from Strommen et al. (2019): "Spatial patterns of the four regimes defined by the cluster centroids for ERA-Interim (1979–2010). Obtained by applying k-means clustering to the geopotential height anomalies at 500 hPa, restricted to the Euro-Atlantic region. The percentages indicate the frequency of occurrence of that regime during the entire time period."

the influence of anthropogenic factors, such as global warming due to increased greenhouse gas concentrations, on the NAO has gained importance in the scientific debate. Significant interactions between the observed warming and trends in the NAO have been observed, and research is focusing on understanding the contributions of the NAO on observed climate changes in the North Atlantic region and projecting future changes (e.g. Kjellström et al. 2013; Eyring et al. 2021). These investigations are usually conducted with the help of Global Circulation Models (GCMs). Due to their central role in understanding the NAO and other modes of variability and their changes under future climate change, it is essential to ensure that climate models correctly represent these modes of variability and the dynamics related to them. Therefore, the NAO and other modes of variability are used as a common quality check for GCMs (e.g. Döscher et al. 2022). A comprehensive overview on the evaluation of the representation of the NAO and other modes of variability in state-of-the-art climate models can be found in the latest assessment report of the Intergovernmental Panel on Climate Change (IPCC) (Eyring et al. 2021).

The representation of modes of variability in climate models is usually evaluated by comparing both the spatial structure and temporal variability of the mode in question to climate observations. However, this work focuses on the spatial structure of the modes of variability in the North Atlantic region, as represented in the newest generation of GCMs of the Coupled Model Intercomparison Project Phase 6 (CMIP6). The representation of the NAO and other modes of variability in CMIP6 models has been investigated in a number of studies, e.g. by Fasullo et al. (2020) and Lee et al. (2021), comparing CMIP3, CMIP5 and CMIP6 models. In contrast, this work focuses on CMIP6 models only, examining the spatial patterns of modes of variability in the North Atlantic region and adding an analysis of the temperature and teleconnection patterns associated with these modes of variability. A similar study has been undertaken by Rousi et al. (2020), investigating the spatial variability of the NAO and its impact on European temperature and precipitation. However, they consider only one GCM, in great details, while our study focuses on the comparison of several models.

3 Data and Methods

3.1 Data

This work aims to compare different Global Circulation Models (GCMs) from the Coupled Model Intercomparison Project Phase 6 (CMIP6) of the World Climate Research Programme (WCRP) w.r.t. their representation of atmospheric circulation variability over the north Atlantic and its impact on weather conditions on the surrounding continents. CMIP6 consists of around 130 different climate models, developed by 49 different modelling groups following a prescribed set of standards to allow for comparability between the models (WCRP-CMIP n.d.) (Eyring et al. 2016). Due to time and data processing constraints, we have chosen a subset of seven models for the comparison. The choice of models is presented in Table 1. Note that the selection is dominated by models developed in Europe, with two models developed in France, one in Germany, one in the United Kingdom, one by a European collaboration effort, one in the US and one in Japan. To evaluate the models w.r.t. their representation of variability under current climate conditions we consider historical simulations. In these historical simulations, the models are driven by external forcing conditions equivalent to the natural and anthropogenic forcing observed in the years 1850-2014, including Greenhouse Gas (GHG) emissions and concentrations, land use changes, solar forcing and volcanic aerosols (Eyring et al. 2016). Due to the chaotic dynamics of the climate system, these simulations will not replicate historically observed climate conditions. However, they can be compared to observations in terms of climatology, i.e. long term means and variability. To this end, we use methods like an EOF analysis to extract mean patterns of variability and compare them to reanalysis data. In this work, each climate model is represented by only one run, usually ensemble member *r1i1p1f1*, if available. The 'ripf' indices identify individual members of an ensemble of simulations by their characteristics, where r indicates the realization (i.e. initial conditions), i the initialization method, p differences in model physics and f the forcing data used (Taylor et al. 2018). For two models, *r1i1p1f1* was not available, and *r1i1p1f2* is used instead, denoting that a different forcing data set was used. The different realizations r are runs of the same model under identical settings, but starting at different initial conditions. Thus, they represent the internal climate variability of the model. Due to time and computational constraints, this work is not able to include ensembles consisting of a number of realizations of each model, that would allow to take this variability into account. Therefore, it should be kept in mind that even though we will refer to the data by their model names, only one realization of each model is considered. We aim to address the question of the relationship between the variability within and between models in a limited scope by including two realizations of one of the models in the analysis. The EC-Earth3 model is represented by the *r1i1p1f1* and *r10i1p1f1* ensemble members. They will be referred to as EC-Earth3 realization 1 (EC-Earth3 r1) and EC-Earth3 r10 and treated the same way as the distinct models in the analysis.
Model output data in monthly resolution is obtained from the Earth System Grid Federation (ESGF)[1]. We consider the variables geopotential height at 500 hPa (Z500) to represent the atmospheric circulation and near surface (2 meter) air temperature and precipitation flux, hereafter referred to as temperature and precipitation to represent teleconnections to surface conditions in the North Atlantic region.

The historical runs model output is compared against the ERA5 reanalysis dataset (Hersbach et al. 2020) and its back extension (Bell et al. 2021), currently available in its final form covering the period from 1959 to present. ERA5 is the newest reanalysis product of the European Centre for Medium-Range Weather Forecasts (ECMWF), providing a detailed, globally complete and consistent record of the state of the global atmosphere, land and ocean surface by combining observations with weather forecasting models through a process called data assimilation (Hersbach et al.

[1] https://esgf-data.dkrz.de/search/cmip6-dkrz/

2020). Monthly averaged values of geopotential, 2 meter air temperature and total precipitation are obtained from the Copernicus Climate Data Store[2] and subsequently transformed to match the variables obtained for the CMIP6 models.

Model / Source ID	Institution	Variant	Timeframe	Nom. Res.	Reference
ERA5 Reanalysis	ECMWF	-	1959-2022	30 km	(Hersbach et al. 2020)
CESM2	NCAR	r1i1p1f1	1850-2014	100 km	(Danabasoglu 2022)
CNRM-ESM2-1	CNRM-CERFACS	r1i1p1f2	1850-2014	250 km	(Seferian 2022)
EC-Earth3	EC-Earth-Consortium	r1i1p1f1	1850-2014	100 km	(EC-Earth Consortium 2022)
EC-Earth3	EC-Earth-Consortium	r10i1p1f1	1850-2014	100 km	(EC-Earth Consortium 2022)
IPSL-CM6A-LR	IPSL	r1i1p1f1	1850-2014	250 km	(Boucher et al. 2022)
MIROC6	MIROC	r1i1p1f1	1850-2014	250 km	(Tatebe et al. 2022)
MPI-ESM1-2-HR	MPI-M DWD DKRZ	r1i1p1f1	1850-2014	100 km	(Jungclaus et al. 2022)
UKESM1-0-LL	MOHC NERC NIMS-KMA NIWA	r1i1p1f2	1850-2014	250 km	(Tang et al. 2022)

Table 1: Overview of the used data, containing information on model name, modelling group, ensemble member, available timeframe, nominal resolution and further references.

3.2 Data Preprocessing

As the analysis is focused on the North Atlantic region, the global model output is cropped to the region between 80° West to 40° East and 20° to 85° North, covering the North Atlantic basin, including Greenland, parts of the North American East coast, Europe and northern Africa. To make the data comparable between models, they are interpolated to a common grid, using the 2.5° (longitude) x 1.27° (latitude) grid of IPSL-CM6A-LR, as it is one of the models with lowest resolution. All three variables are detrended using a linear regression in an attempt to remove the climate change signal from the data. These steps of data preprocessing are executed with the help of the Climate Data Operator (CDO) command line operators (Schulzweida 2022). The following steps of the analysis are conducted in Python, for which the code is made available online[3].
The following EOF and PCR analysis is conducted based on seasonal mean values of Z500, temperature and precipitation. The winter season is chosen for this analysis because it has shown to be most dynamically active and characterized by a clearer dominance of few patterns, with the NAO dominating and exerting the largest influence on surface temperature and precipitation (Hurrell et al. 2003). After comparing the results for different months, the extended winter season from November to March is selected. The CMIP6 data is available for a considerably longer time frame than the ERA5 dataset, but a pre-analysis showed that the analysis results are most comparable between both if using a common time frame. Thus the analysis is conducted over the model years 1959-2014.

3.3 EOF Analysis

An Empirical Orthogonal Functions (EOF) analysis is used to determine the dominant patterns of variability in geopotential height over the studied domain and time frame. EOF analysis is an alternative name for Principal Component Analysis (PCA), frequently used in atmospheric sciences. It is a common tool for dimensionality reduction of large spatio-temporal datasets, that has provided valuable insights to understanding the processes in the atmosphere (Hannachi et al. 2007). It divides the data into spatial patterns of variability (here referred to as EOFs) and associated time series (Principal Component (PC) time series) that specify the contribution of each pattern to the observed state of the data at each time step. Thus, the geopotential height field at every time step can be understood as a combination of the EOF patterns of variability, with varying relative contributions. The EOFs are ordered by the fraction of overall variance they explain and consequently the first few EOF patterns usually account for the majority of the variance observed in the data. Thus, considering the data in terms of the first few EOF

[2]https://cds.climate.copernicus.eu/
[3]https://github.com/Anna-Ida/masters_thesis/

patterns and their variance is a simplified way of understanding the variance present in the data, while still conserving the majority of the information. EOF analysis is able to reduce complexity by removing redundant information in the form of correlations present in the input dataset and replace it with uncorrelated linear combinations of the original variables, the EOFs. To conduct an EOF analysis, the eigenvectors and eigenvalues of the covariance matrix of the data are taken, where the eigenvectors and eigenvalues correspond to the EOF patterns and the explained variance associated with each pattern, respectively. The associated time series of amplitude (PC time series) can be obtained by projecting the EOF patterns onto the data field at each time step. Geometrically, the leading eigenvector, associated with the largest eigenvalue and hereafter referred to as EOF1, can be understood as pointing in the direction of largest variance present in the data. As the covariance matrix is symmetric, the second order eigenvector, associated with the second largest eigenvalue, is orthogonal to the first and follows the direction with the next highest variance. This is true for all following eigenvectors as well. This orthogonality presents the largest limitation to the interpretability of the second and following EOF patterns as physical patterns, as they are found under the condition of being uncorrelated to the first pattern, which may not reflect the physical reality (Wilks 2005; Feldstein et al. 2017).

Here, the EOF analysis is conducted with the help of the Python library *eofs* (Dawson 2016) to ensure correctness and reproducability of the results. Input is a dataset of geopotential height at 500 hPa, with values at each gridpoint over the North Atlantic region and for each winter season over the studied timeframe from 1959-2014. Before conducting the EOF analysis, the temporal mean is removed from the data, so that the analysis is based on the anomalies. The anomalies can be understood as arranged in an anomaly matrix A with dimensions M x N, where M denotes the amount of spatial locations and N the amount of time steps. Each row of the matrix represents a map of anomalies at all grid points, and each column is a time series of anomalies at one point.

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}$$

To compensate for the differences in grid cell area due to the convergence of the meridians, the data is weighted with $\sqrt{cos(latitude)}$ (Wilks 2005). Instead of computing the covariance matrix of A, the *eofs* package calculates the eigenvectors and eigenvalues based on Singular Value Decomposition (SVD), which is computationally more efficient. The SVD is given by:

$$\mathrm{SVD}(\mathbf{A}) = \mathbf{U}\mathit{\Gamma}\mathbf{V}^T \tag{1}$$

where the columns of U and V are the singular vectors, and the singular values are on the leading diagonal of Γ. From the SVD of A, the EOFs can be obtained as the right singular vectors, their associated variances as the singular values and the standardized PC time series as the left singular vectors (Dawson n.d.). For the comparability of the SVD and the covariance matrix approach see Dawson (n.d.). The method returns the EOFs, eigenvalues and the PC time series, scaled to unit variance. The EOFs are sorted by decreasing order of eigenvalues. We only consider the leading four EOFs.

The resulting EOF patterns are dimensionless and their sign is arbitrary. Their centers of action indicate centers of high/low geopotential height anomalies and at the same time the regions of strongest variability (Hurrell et al. 2003). The associated PC time series indicate the sign and relative importance of the EOF pattern at each time step. The patterns can be expressed in the units of the original variable, here geopotential height in meters, by regressing the anomaly time series of the original variable upon the respective principal component time series, as described in the following section (Hurrell et al. 2003).

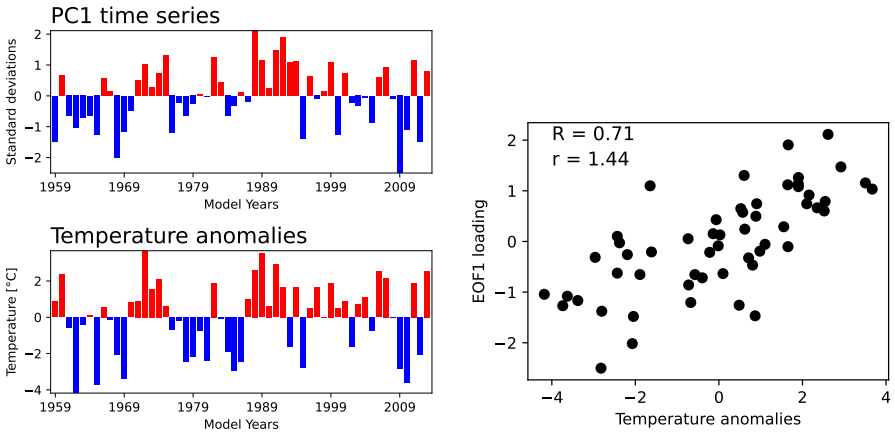


Figure 2: **Illustration of the Principal Component Regression method.** Left: PC time series of EOF1 and time series of temperature anomalies at 59.6°N, 12.5°E of the ERA5 data. Right: Scatterplot of both time series, indicating correlation coefficient R and linear regression coefficient r.

Next to the orthogonality of the eigenvectors, a further problem with physical interpretation of the EOF patterns can arise if the EOFs are not well defined. If two or more EOFs are associated with similar eigenvectors, it can be an indication that their sampling distributions are entangled, resulting in the true population counterparts (i.e. the true patterns of variance) being nearly arbitrarily mixed between the sample eigenvectors (i.e. the obtained EOF patterns). These groups of eigenvalues lying within one or two $\delta\lambda$ of each other, are called '*effectively degenerate multiplets*' and their associated eigenvectors, i.e., EOF patterns present a challenge for physical interpretation, as they usually display a combination of different patterns. Whether an EOF pattern is well defined can be determined by considering the distance of its eigenvalue to the closest eigenvalues, taking their sampling errors into account. The sampling error δ of each eigenvalue λ can be approximated as $\delta\lambda \sim \lambda(2/N)^{1/2}$ where N is the number of observations. Following the rule of thumb suggested by North et al. (1982), an EOF is considered well separated from following EOFs, and can thus be attempted to be interpreted as a physical pattern, if the sampling error of its eigenvalue is smaller than the distance to a nearby eigenvalue. In practice, we consider an EOF well separated from neighboring patterns if the sampling error of its eigenvalue does not overlap with sampling errors of neighboring eigenvalues (North et al. 1982; Wilks 2005).

3.4 Principle Component Regression

Based on the results of the EOF analysis, a regression analysis is used to explore the connection between the modes of variability and surface climate conditions. In this work, the dependent variables of interest are precipitation and temperature. At each grid point, the time series of anomalies of this variable can be compared against each PC time series. If they vary in connection with each other, we can hypothesize that the surface conditions at this location are strongly influenced by this particular mode of variability and even make a statement about the direction of this relationship. An example of this can be seen in Figure 2, showcasing the strong positive relationship between the NAO time series and surface temperatures close to Oslo, Norway in the ERA5 data. We calculate both correlation and linear regression coefficient between the anomaly time series of both dependent variables and each of the four leading PC time series, normalized to standard variance, at each grid point. While Pearson's correlation coefficient R, see Equation 2, indicates the strength of the relationship, the value of the linear regression coefficient r indicates the change in the variable at question associated with one standard deviation change along the respective PC time series axis. Therefore, it is also useful to regress the original variable of the EOF analysis, geopotential height, back on to the PC time series, to obtain meaningful units

representing the magnitude of variation of each spatial mode. The correlation and regression values can be combined to display the results in a meaningful way, by showing the regression value for each grid point that reached a correlation value above a certain threshold, 0.3 in this work, to only show robust results.

3.5 Model Comparison

The above described EOF analysis and PC regression is applied both to the ERA5 reanalysis dataset and the CMIP6 model data. The results of the analysis of ERA5 and model data can be qualitatively compared through visual inspection. To facilitate comparison of a large number of datasets, we include quantitative methods to support the results of the qualitative inspection which we present in the following.

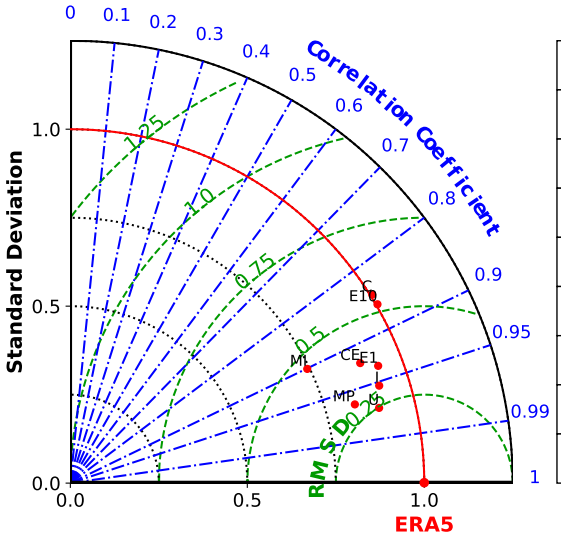
In order to compare how different models represent the relationship between weather conditions and different modes of variability, it is necessary to first compare how those patterns of variability differ between models. As the EOF and PCR patterns are based on deviations from the mean state, their differences might be due to differences in the mean state between models. Hence, the comparison methodology contains approaches to compare the results of the PC regression, the EOF analysis, which is represented by regressing Z500 onto the EOFs, and the mean state of the three variables Z500, temperature and precipitation. Note that in order to facilitate the comparison, as an initial step we inspect the four leading EOF patterns of each model and if necessary, invert the signs inside the patterns to correspond to ERA5. We also associate the patterns with known physical modes of variability and, if a pattern is associated with a different eigenvalue rank than in ERA5, change the order of the patterns. This 'EOF swapping' ensures that the comparison captures differences in the spatial patterns and their teleconnections compared to ERA5 and not just the fact that the eigenvalues associated with each EOF pattern vary between the models. This different relative importance of modes of variability does not play a significant role when investigating the differences in their teleconnections to weather conditions. However, it is an expression of an underlying difference in atmospheric variability and will therefore be accounted for by comparing the eigenvalue spectra of the models to ERA5.

3.5.1 Principle Component Regression Pattern Comparison: Taylor Diagram and Skill Score

To compare the spatial patterns obtained from the EOF and PCR analysis we follow the approach proposed by Taylor (2001), who devised a method and diagram that allows for a well structured comparison of a large number of spatial patterns. The Taylor diagram is well-suited to compare a number of models to a reanalysis dataset in a concise overview. In the here present case, the compared patterns are the patterns of Z500, temperature and precipitation PC regression for each EOF separately. Similarity between a model and ERA5 PCR pattern is measured by three metrics: pattern correlation, mean squared difference and standard deviation. The pattern correlation R is calculated as the Pearson correlation coefficient between the regression values at all N grid points of the ERA5 data, X and a CMIP6 model, Y, as seen in Equation 2,

$$R = \frac{1}{N} \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y} \tag{2}$$

where x_i and y_i are the values at each gridpoint i of the ERA5 and model data, $\overline{x}$ and $\overline{y}$ the spatial mean regression values and σ_x and σ_y the standard deviations (Taylor 2001). Equation 2 is altered to account for unequal sizes of grid cells by weighting values with the cosine of their respective latitude. Thus, $\overline{x}$ and $\overline{y}$ are replaced with the weighted average, σ_x and σ_y with the weighted standard deviation, the product is multiplied with the weight corresponding to the respective latitude and the result is divided by the sum of all weights instead of N. The formula for the



Model	Skill Score
U = UKESM1-0-LL	0.934
I = IPSL-CM6A-LR	0.903
MP = MPI-ESM1-2-HR	0.9
E1 = EC-Earth3 r1	0.871
CE = CESM2	0.844
E10 = EC-Earth r10	0.755
MI = MIROC6	0.748
C = CNRM-ESM2-1	0.729

Figure 3: **Example of a Taylor diagram and corresponding Taylor skill score ranking.** Left: Taylor diagram, comparing model patterns (red points) to reference pattern (ERA5) by correlation coefficient (blue), normalized standard deviation (black) and centered RMSD (green). Right: Ranking of the models by their Taylor skill score, calculated based on correlation coefficient and standard deviation difference (ref. Equation 7).

weighted average $\overline{x}^*$ is given in Equation 3, where w_i is the weight, given by $\cos(latitude)$ at each grid point.

$$\overline{x}^* = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} \tag{3}$$

The second measure of pattern similarity is the centered root mean square difference $CRMSD$ between the patterns, shown in Equation 4. It is based on the root mean square difference, which is centered by removing the field means $\overline{x}$ and $\overline{y}$. The area weights are included the same way as explained above, by considering the weighted average, weighing each difference pair and dividing by the sum of weights.

$$CRMSD = \sqrt{\frac{\sum_{i=1}^{N} w_i((x_i - \overline{x}^*) - (y_i - \overline{y}^*))^2}{\sum_{i=1}^{N} w_i}} \tag{4}$$

Lastly, the spread in the patterns can be compared by calculating the area weighted standard deviation σ^* for each pattern, as shown in equation 5.

$$\sigma^* = \frac{\sum_{i=1}^{N} w_i (x_i - \overline{x}^*)^2}{\sum_{i=1}^{N} w_i} \tag{5}$$

The difference between models and ERA5 as expressed by these three metrics can then be visualised in a Taylor diagram, where the standard deviation is plotted as the radial distance to the origin, the correlation coefficients as the azimuthal position and the $CRMSD$ in concentric circles (Taylor 2001). For better comparability of the diagrams, the standard deviation is normalized by the observed ERA5 standard deviation. An example of this can be see in Figure 3. The reference dataset ERA5 is plotted on the x-axis, as it correlates perfectly with itself and the models can be evaluated by their distance along the correlation coefficient axis, to the reference standard deviation and by their CRMSD. This analysis focuses mainly on the former two metrics.

Next to this visual comparison, the pattern similarity can be quantified for each model by a skill

score S following Taylor (2001). Its calculation is shown in Equation 6,

$$S = \frac{4(1+R)^4}{(\hat{\sigma} + \frac{1}{\hat{\sigma}})^2(1+R_0)^4} \tag{6}$$

where $\hat{\sigma}$ is the ratio $\frac{\sigma_y}{\sigma_x}$ between the model and reference standard deviation and R_0 is the maximum attainable correlation value. Setting R_0 to 1 simplifies Equation 6 to Equation 7 (Hirota et al. 2011).

$$S = \frac{(1+R)^4}{4(\hat{\sigma} + \frac{1}{\hat{\sigma}})^2} \tag{7}$$

This skill score sums up each model patterns similarity to the ERA5 pattern, both with regards to the pattern correlation and the spread, quantified by σ^*. The table on the right of Figure 3 shows the skill scores for all models displayed in the Taylor diagram. Models that are located close to the reference dataset in the diagram obtain a higher skill score than those far away.
This comparison approach is applied to the patterns of temperature, precipitation and geopotential height PC regression. The latter is used to evaluate the EOF patterns, as the Z500 PCR results in the exact same patterns as the EOF analysis but with meaningful units. The Taylor skill score allows a ranking of models taking both their similarity in spatial pattern, measured by the correlation coefficient, and in magnitude, measured by the standard deviation, to ERA5 into account. In order to ensure that the pattern comparison focuses on the centers of action of the EOF patterns and the most important regression locations, we apply the same comparison to the PCR fields where grid cells with low correlation coefficients between the PC time series and the dependent variable are masked out. We set the correlation threshold to 0.3, and hence consider all locations with a correlation coefficient above 0.3 to have a robust connection between the PC time series and dependent variable. The Taylor metrics are then calculated only over those grid cells. At locations, where only either ERA5 or the CMIP6 model have a robust value, the missing value is set to zero.

3.5.2 Mean Field Comparison

As mentioned above, differences in the EOF patterns between ERA5 and model data can indicate differences in the circulation variability, but can also be due to differences in the mean state of the atmosphere. To investigate whether a low skill score on the Z500 PCR shows that the model simulates atmospheric modes of variability differently, or whether shifted centers of action or different magnitudes are the result of a bias in the mean state, a few comparison approaches are applied to the mean field of Z500, temperature and precipitation. This mean field is obtained as the arithmetic mean over the observed time frame at each grid point.
To get an insight into the difference between the ERA5 and model mean state, they are subtracted on grid point level (model-ERA5). Calculating the area weighted arithmetic mean (compensating for different grid cell sizes) over the difference field can give a quantitative indication of the model bias. However, the impact on the circulation strongly depends on where the difference is located in relation to the mean flow field. Therefore, the difference is visualized together with the ERA5 mean field of geopotential height at 500 hPa. We evaluate whether the difference is statistically significant at each grid point with a two-sided z-test, comparing the mean of the variable under investigation at each grid point of ERA5 and the respective model, assuming normal distribution in both. The null hypothesis is that the two distributions are equal, and the z-score is calculated as in Equation 8. If the corresponding p-value is smaller than 0.05, the null hypothesis is rejected and the difference between ERA5 and the model is deemed significant.

$$z_i = \frac{(x_i - y_i)}{\sqrt{\sigma_{x_i}^2 + \sigma_{y_i}^2}} \tag{8}$$

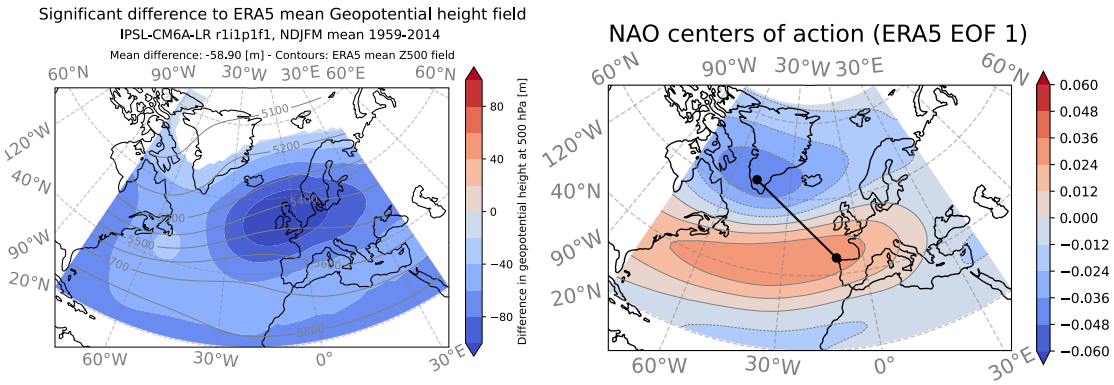


Figure 4: **Example of a difference plot and illustration of the NAO gradient calculation.** Left: Difference pattern between IPSL-CM6A-LR and ERA5 mean state of Z500, showing only locations with statistically significant difference. Contour lines show the ERA5 mean Z500 field. Right: Locations to calculate the NAO gradient, marked as black points on top of the ERA5 EOF1 pattern of Z500.

Using only those significantly different grid points helps making both the difference plot easier to evaluate and the mean difference value more robust. An example of the significant difference plot together with the ERA5 mean flow field can be seen in the left panel of Figure 4.

A second approach aims at specifically evaluating the mean geopotential height field with regards to the NAO. It compares the mean gradient between the two centers of the NAO in each CMIP6 model to the one observed in the ERA5 data. For this, the northern low geopotential height and southern high geopotential height center are localised in the EOF1 pattern of the ERA5 data, as shown in the right panel of Figure 4. They are set at the southern tip of Greenland and before the north-western coast of the Iberian Peninsula. The gradient between the mean geopotential height at both locations is then compared between ERA5 and CMIP6 model data. Taking interannual variability into account, a z-test is again used to determine if the difference between the two is statistically significant at the 5% level. If $p < 0.05$, they are outside the 2σ range, corresponding to a 95% confidence interval and thus deemed significantly different. This measure helps to evaluate both whether the north south gradient that is the driving force of the circulation is of comparable size to the reanalysis data in the mean state of the model and whether the NAO centers of action are located close to where they are found in ERA5.

4 Results

This chapter first presents the modes of atmospheric circulation variability identified in the North Atlantic region extended winter season by applying an EOF analysis to the ERA5 geopotential height data, and their associated temperature and precipitation teleconnections, identified with a Principal Component Regression Analysis. Section 4.2 will then compare these variability and teleconnection patterns to the ones found with the same analysis in different CMIP6 models historical runs.

4.1 ERA5

Figure 5 shows the mean states over the extended winter seasons 1959-2014 of the three variables that are part of the analysis. The geopotential height at 500 hPa (Z500) field shows the mean state of the atmosphere, with a north-south gradient of geopotential height. The wave shape of the westerly flow along this gradient, diverging over Europe, is well visible. The Z500 pattern is closely linked to the temperature pattern, showing a similar shape of the north-south gradient. The precipitation field follows the wavy patterns of the mean flow field, with its highest values in the western North Atlantic, the southeastern coast of Greenland, Iceland, the coasts of Scotland and Norway, the westcoast of the Iberian Peninsula and the eastern Mediterranean region.

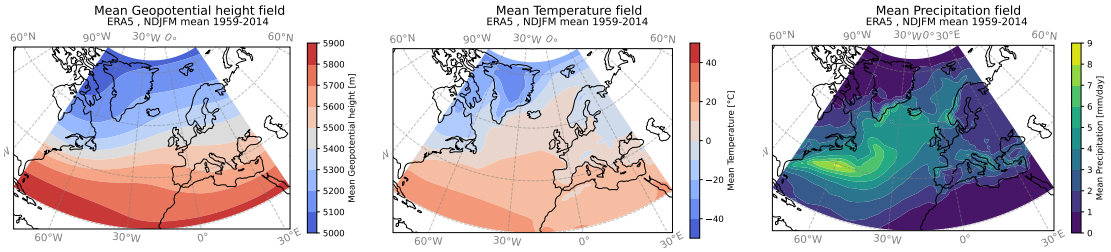


Figure 5: **Mean state of the variables in ERA5**. Mean state of geopotential height at 500 hPa (left), 2m air temperature (center) and precipitation flux (right) in the ERA5 dataset over the extended winter season (NDJFM) for the years 1959-2014.

4.1.1 Modes of Variability

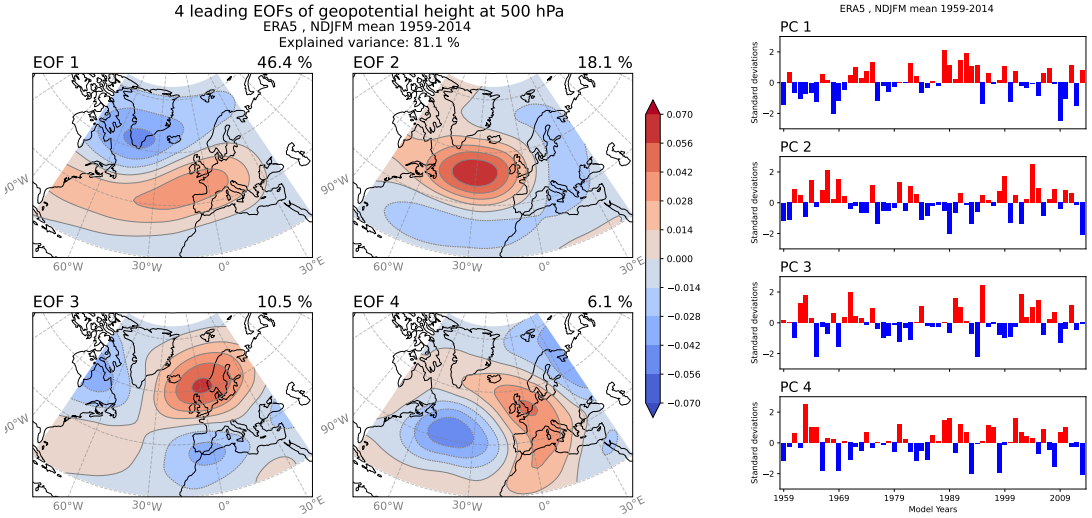


Figure 6: **Results of the EOF analysis on ERA5 Z500 data.** Left: Spatial patterns of the four leading EOFs of ERA5 geopotential height at 500 hPa for the extended winter season 1959-2014 and their explained variance. Right: Associated Principal Component time series, scaled to unit variance.

Figure 6 shows the leading four EOF patterns and their corresponding PC time series that have been obtained based on the deviations from the ERA5 mean geopotential height field. Together, the four leading EOFs explain 81% of the total variance observed in the geopotential height field, out of which the first EOF pattern (EOF1) alone explains almost half, with 46%. The explained variance and associated uncertainty, calculated by dividing the eigenvalues and their sampling errors by the sum of all eigenvalues and sampling errors, respectively (see Section 3.3), are depicted for the leading 40 eigenvalues in Figure 7. The leading EOF is most distinct from the others. The second and third eigenvalues are also well separated, following the criterion established in Section 3.3, but starting with the fourth eigenvalue, the error bars of neighboring eigenvalues are overlapping. Following the rule of thumb by North et al. (1982), this observation suggests that only the first three EOFs should be considered independent patterns that can be physically interpreted. Thus, the following analysis will focus on the three leading EOFs, with respective weight based on their explained variance.

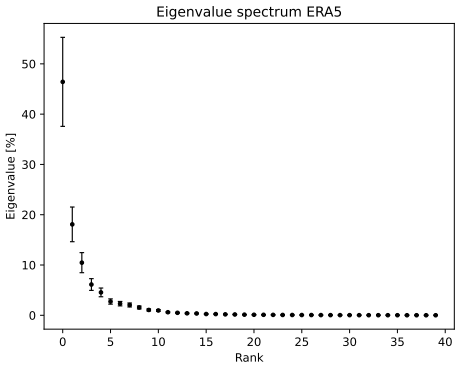


Figure 7: **Eigenvalue spectrum of ERA5**. Explained variance and associated uncertainty of the 40 leading eigenvalues of ERA5 geopotential height at 500 hPa over the extended winter season 1959-2014.

The patterns in Figure 6 show that EOF1 is characterized by a dipole pattern with one center over the southern tip of Greenland and a southern band with opposite sign, with its center stretching from the center of the North Atlantic into central Europe. This corresponds with the well-known spatial pattern of the NAO. As mentioned above, the first EOF, associated with the NAO pattern explains 46% of the total observed variance. The corresponding time series, visible in the right panel of Figure 6, can thus be understood as the NAO index, defining the positive and negative phases of the NAO.

The spatial pattern of EOF2 is dominated by a center of action over the north-eastern part of the North Atlantic well between the land masses of North America, Greenland and Europe, surrounded by a belt with opposite sign stretching from northeastern Europe over the Mediterranean into the southern North Atlantic. In correspondence with the literature (i.e. Strommen et al. 2019; Ruggieri et al. 2020), we will refer to this pattern as Atlantic Ridge (AR) pattern. With 18% it explains a significantly smaller fraction of the observed variance, but enough to explain the variance equivalent to one out of the five months of each extended winter season on average.

EOF3 shows a quadripole pattern with a strong center over northwestern Europe and weaker centers with opposite sign over the Hudson bay area and the straight of Gibraltar and lastly a very weak center in the southwestern North Atlantic. This corresponds to what Ruggieri et al. (2020) identify as Scandinavian Blocking (SB) pattern, characterised by a positive geopotential height anomaly over Scandinavia in its positive phase. It explains 10.5% of the observed variance.

Lastly, EOF4 shows a tripolar pattern, with one center of action over the central North Atlantic, followed by a center of opposite sign to its east centered over western Europe and another center of the same sign as the first to the northeast of it. However, as it only explains 6% of the total variance and it is not well separated from the following EOFs, according to the eigenvalue spectrum depicted in Figure 7, we will not attempt to associate it with a known physical pattern of variability.

The right panel of Figure 6 shows the scaled PC time series, indicating the relative loading of each EOF pattern in each winter season of the analysis time frame. This will be used for the Principal Component Regression (PCR) analysis, relating temperature and precipitation anomalies at each grid point to the variations in loading of the EOF patterns. The interplay between the loadings is

interesting to see, where for example in the 1960s a persistently negative NAO seems to coincide with a positive state of the Atlantic Ridge pattern and vice versa in the 1990s. The trends to a dominant negative phase of the NAO in the 1960s and positive phase in the 1990s are in line with observation-based NAO indices (Cropper et al. 2015; Hurrell 2015). The PC time series also matches observations w.r.t to single extreme NAO events, for example the strong negative NAO in the winter 2009/10 (Cattiaux et al. 2010).

4.1.2 Teleconnections

The precipitation and temperature teleconnection patterns of the four leading EOFs of ERA5 in the wintertime Nort Atlantic region are displayed in Figure 8. At every grid point, the value of the linear correlation coefficient between the anomaly time series of the variable in question and the PC time series of the respective EOF pattern is displayed, but only if the correlation coefficient between the two time series exceeds 0.3. Thus, only locations with a robust relationship are shown. For the full PCR field see Figure A.1. These PCR maps show the local temperature/precipitation response to a one standard deviation excursion in the positive direction of each respective EOF pattern.

The temperature regression with the NAO index (PC1 time series) shows the typical quadripolar pattern described in Chapter 2, with a temperature seesaw between northern Europe (around the Baltic sea) and the region around the Labrador sea. At its strongest locations, the temperature change associated with one positive standard deviation change of the EOF pattern is +1.4°C at the Swedish border close to Oslo, Norway (59.6°N, 12.5°E) and −2.5°C in the Davis Strait between Nuuk, Greenland and Baffin Island, Canada (63.4°N, 57.5°W). Furthermore, a weak positive response leads from the US east coast across the North Atlantic towards northwestern Europe, while south of it there is a band of negative regression values, stretching over northern Africa and the eastern Mediterranean region. This corresponds to the well known temperature teleconnection pattern of the NAO, where a positive state of the NAO leads to mild winters in northern Europe and cold winters in Greenland and the opposite response further south, and a negative state to the opposite response. It is also in line with the expected circulation resulting from the EOF pattern, with the westerly flow being directed towards northern Europe in the positive state.

The precipitation response of the NAO pattern, visible in the upper left panel of the lower half of Figure 8, shows a positive response in the northeastern part of the domain, from Iceland to the west coasts of Scotland and Norway and a negative response on the Iberian peninsula and the northern Mediterranean region, as well as a weaker negative response around Baffin Bay. The locations with the strongest relationships are close to Bergen, Norway (60.85°N, 5°E), where a one standard deviation positive NAO leads to a positive precipitation anomaly of 1.7 mm/day and at the northern border between Portugal and Spain (41.8°N, 7.5°W), where the same state is associated with 1.3 mm/day less than usual. Overall, the temperature and precipitation responses have a very similar spatial pattern, but the precipitation response pattern is located further north.

EOF2, the Atlantic Ridge pattern, shows an associated temperature response pattern that resembles the EOF pattern, with a positive response just west of the center of action over the North Atlantic and a negative response all around. Similarly, the precipitation response is marked by an inverse relationship at the European west coast and the Greenlandic southeast coast and a positive relationship around, with robust locations at the southwest coast of Greenland, north of Norway, around the Black sea, over northern Africa and in the southeastern North Atlantic. Interesting are locations of opposite sign in relative spatial proximity, both in southern Greenland and in southern Norway. These dipole responses in the same region can serve as robust identifiers for example in the context of reconstructing atmospheric variability from proxy records. In the positive phase of the AR pattern, the temperature and precipitation response can be associated with the blocking characteristics of a high pressure center over the Atlantic and the diversion of the flow around it.

The variability of EOF3, with a center of high geopotential height center over Scandinavia in its

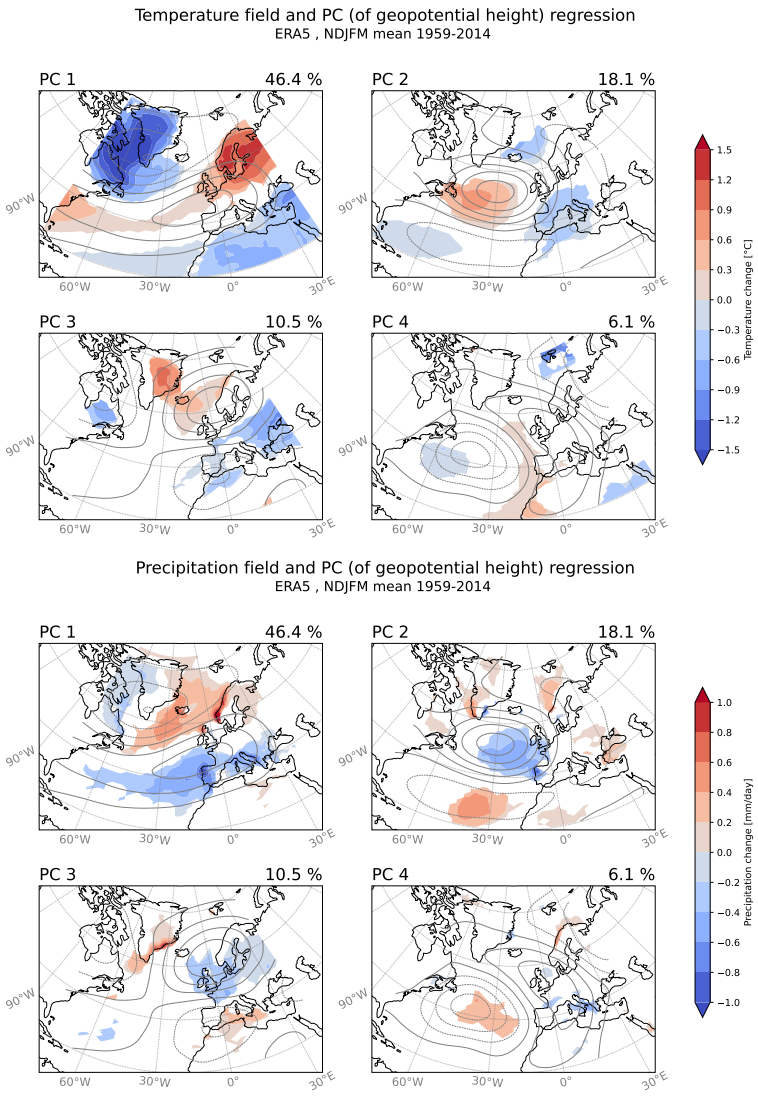


Figure 8: **Temperature and precipitation response patterns of the four leading EOFs of ERA5 Z500**. Regression of temperature (upper) and precipitation (lower) anomaly time series on the four leading PC time series of ERA5 Z500. Colors: linear regression coefficient at locations where linear correlation coefficient 0.3. Contour lines: EOF patterns of 500 hPa geopotential height.

positive phase, and opposite centers over the Hudson bay area and Gibraltar, is associated with a temperature response with a very similar pattern, with the strongest positive response in eastern Greenland and Iceland and the strongest inverse relationship over Quebec and around the Black Sea. The precipitation regression pattern with PC3 is quite similar to that of PC2, but, like the center of action in the EOF pattern, shifted northeast. The relationship to the PC3 time series is inverse over the British isles, northwestern Europe and southern Scandinavia, and positive at the Greenlandic southeast coast and over the Mediterranean, which could again physically be explained by a diversion of the westerly flow around the (blocking) high pressure center that is indicated by the geopotential height anomaly in the positive state of the EOF pattern.

Lastly, the comparison of temperature and precipitation anomalies against the PC4 time series shows very few locations with robust correlations. This is in line with the small fraction of the total variability that is explained by this pattern. However, the faintly visible responses are located close to the EOF centers of action.

4.2 Model Comparison

The following section presents the results of applying the EOF and PCR analysis to the used CMIP6 model data and compares them against the above presented results obtained with the ERA5 data. Figure 9 gives an insight into the EOF analysis results of the models by showing the leading 40 eigenvalues and their associated sampling errors for all evaluated model realizations and ERA5. Like ERA5, all models show a well separated leading eigenvalue that explains a large fraction of the total variance. However, for none of the models, EOF1 can explain as much of the variance as in the ERA5 data (46.6%). EC-Earth3 r10 gets closest, with 45.4% explained variance on the leading EOF pattern, while PC1 of MIROC6 shows the lowest explained total variance with 32.7%. Also the sum of the leading four eigenvectors, as shown by the number in the upper right corner of each plot, accounts for less of the variance than in ERA5, with EC-Earth r10 again getting closest to ERA5 (81.1%) with 78.3% and MIROC6 furthest away with 69% of variance explained. This is the result of more entanglement, or 'effectively degenerate multiplets' between the eigenvalues than seen in ERA5. Notably, only two models (IPSL-CM6A-LR and MPI-ESM1-2-HR) show more than one well-separated EOF, which in both cases is due to the second eigenvalue explaining a higher fraction of variance than in ERA5, separating it from the third one. The fact that in most models, the EOF analysis can only obtain one well separated pattern should be kept in mind for the following analysis. Nevertheless, Figure 9, showing the three leading ERA5 eigenvalues and their error bars as grey shading in the model plots, demonstrates a relatively high overall agreement between the models and ERA5 in terms of explained variance, as the leading three eigenvalues agree with ERA5 within their error bars for all models.

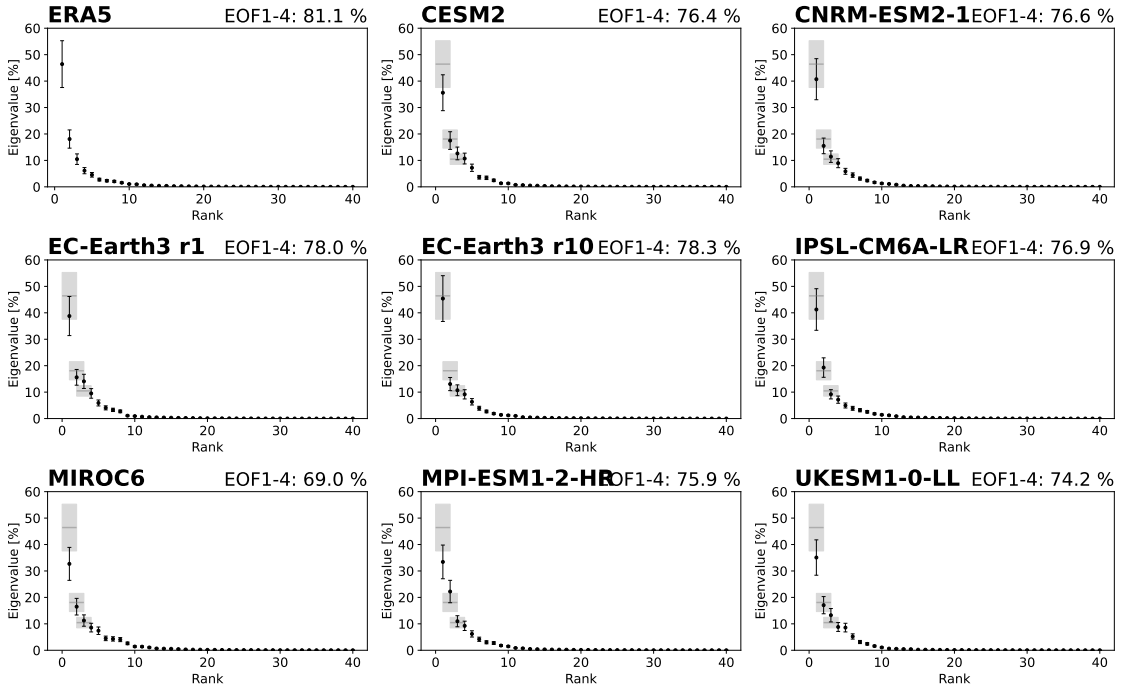


Figure 9: **Eigenvalue spectra of all models**. Shown are the leading 40 eigenvalues, scaled to explained variance in % by dividing by the sum of all eigenvalues, and their sampling errors, scaled by dividing by the sum of all errors. Grey shading indicates the leading three ERA5 eigenvalues and errors.

4.2.1 EOF1: North Atlantic Oscillation

Figure 10 shows the leading EOF pattern of each of the eight analysed model realizations and ERA5, expressed in terms of geopotential height change as the result of regressing Z500 anomalies onto the PC1 time series. A visual comparison shows that all models show the expected NAO

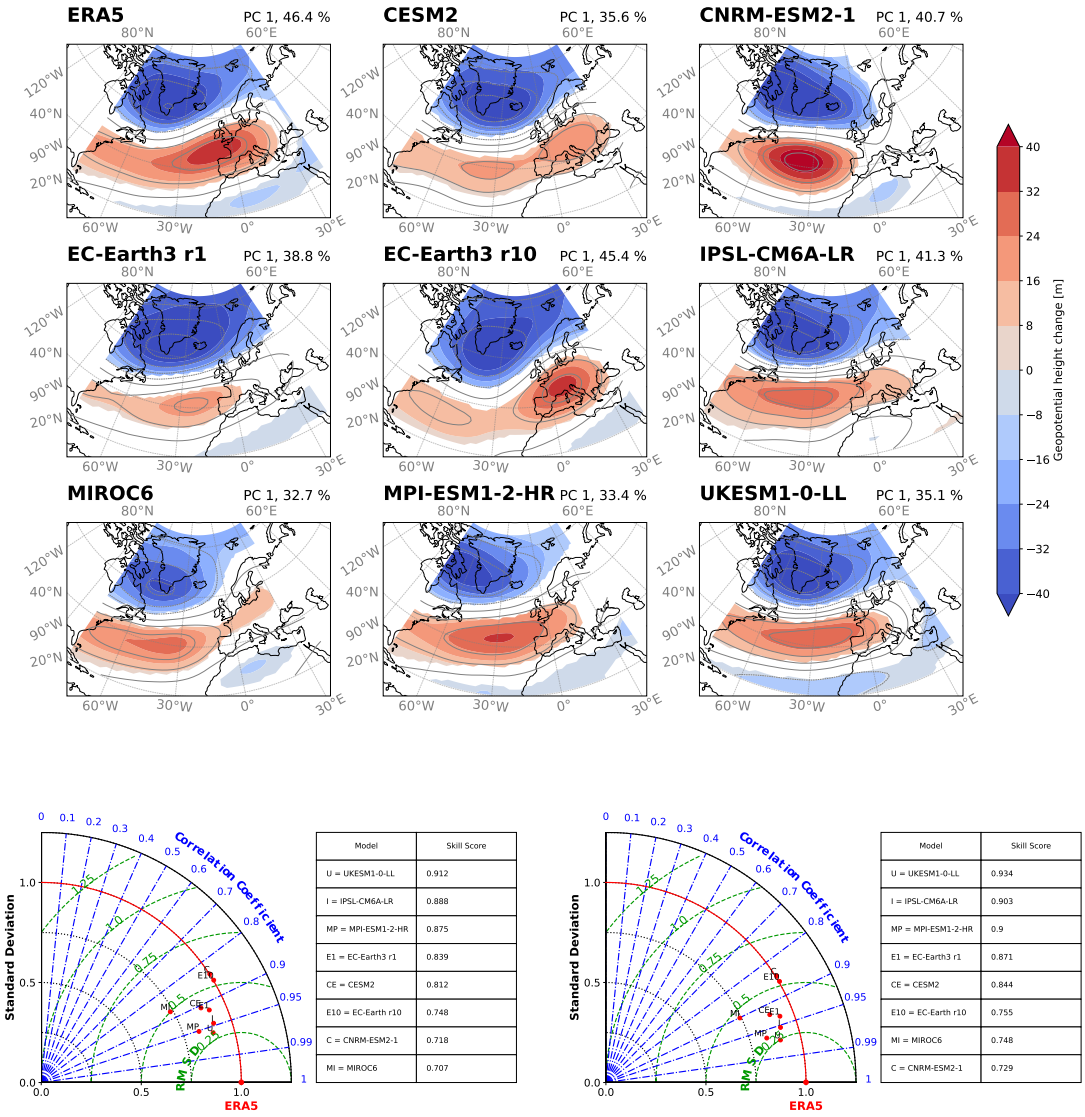


Figure 10: **Comparison of EOF patterns associated with NAO in all models compared to ERA5.** Upper: Spatial patterns of the EOF pattern identified as NAO for ERA5 and all investigated CMIP6 model realizations. Expressed in units of geopotential height change associated with one standard deviation excursion by regressing Z500 anomalies onto the normalized PC1 time series. Only locations where both time series correlate with a coefficient > 0.3 are shown. The signs of the patterns are flipped if necessary to correspond to ERA5. Contour lines show the EOF patterns. Lower: Taylor diagram and skill score ranking for high correlation locations (left) and full field (right). Taylor diagram shows pattern correlation coefficient, centered RMSD and standard deviation, normalized by the ERA5 standard deviation $\sigma^* = 21$ m.

pattern as EOF1. However, in direct comparison to the ERA5 pattern, some models show weaker amplitudes of geopotential height change (e.g. EC-Earth3 r1 or CESM2), while in others the centers of action are shifted (e.g. CNRM-ESM2-1 or MIROC6). These differences in pattern compared to ERA5 are summarized by the metrics shown in the Taylor diagram in the lower left panel of Figure 10. It shows that with one exception (CNRM-ESM2-1), all models have a lower standard deviation than ERA5, underestimating the magnitude of geopotential height variability across the domain. MIROC6 shows the lowest magnitude of geopotential height change associated with EOF1, resulting in a normalized standard deviation of $\sigma^*_{norm} = 0.74$. The pattern correlation with ERA5 is generally high, with coefficients between $R = 0.96$ (UKESM1-0-LL) and $R = 0.84$ (CNRM-ESM2-1). The table next to the Taylor diagram combines the models' performance with regards to variance and pattern correlation by ranking them after their skill score (ref. Equation 7). Low correlation and high difference in standard deviation lead to low skill scores. Thus, MIROC6 obtains a lower skill score than CNRM-ESM2-1 and EC-Earth3 r10, despite a slightly higher pattern correlation coefficient ($R = 0.88$), due to its much higher difference in standard deviation. This figure shows the Taylor diagram and skill score ranking for both the full field (right) and only robust locations (left). Both comparisons lead to very similar results, with the full field approach showing slightly higher correlation coefficients. The skill score ranking is also almost identical. In both cases, the UKESM1-0-LL model performs best, and EC-Earth3 r10, CNRM-ESM2-1 and MIROC6 have the lowest scores, due to their low correlation (all three) and high standard deviation difference (MIROC6 only). In the robust locations comparison, the highest represented skill score is 0.91 (UKESM1-0-LL) and the lowest is 0.71 (MIROC6), in the full field comparison the spread is the same, but the scores are slightly higher with maximum and minimum scores 0.93 (UKESM1-0-LL) and 0.73 (CNRM-ESM2-1), respectively.
Considering the Z500 PCR patterns (upper panel of Figure 10) largely justifies these scores, where the two lowest performing models CNRM-ESM2-1 and MIROC6 show a very clear westward shift in their southern center of action compared to ERA5, barely extending over Europe. Furthermore, in the CNRM-ESM2-1 data, the northern center of action is tilted the opposite way from ERA5 and the rest of the models, extending southeast towards Europe instead of further northeast over Svalbard. The reason for EC-Earth3 r10's low score might be connected to the model showing a rather tripolar pattern, with two positive anomaly centers in the south of the domain and a stronger bent in the southern band of high geopotential height connecting them. Furthermore, the whole pattern is shifted south-east. Visual inspection shows that the other EC-Earth3 realization, r1, deviates from the ERA5 pattern in terms of magnitude, overestimating the northern and strongly underestimating the southern center of action. These shortcomings seem insufficiently reflected in the Taylor diagram and skill score ranking. The full field of geopotential height regression patterns can be seen in Figure A.2.

4.2.2 Model biases

Before evaluating the models' representation of the NAO teleconnection patterns, potential model biases are investigated by examining differences in the mean field of geopotential height, temperature and precipitation. Table (a) in Figure 11 ranks the models after the average difference of their mean Z500 field to the ERA5 field, calculated as the area weighted mean over the difference plot, taking only locations with differences that are statistically significant at the 5% level into account (see Section 3.5.2). It is notable that with the exception of CESM2, all models seem to underestimate geopotential height, indicated by the negative sign of the difference. The model realization with the smallest difference in the mean Z500 field to ERA5 is EC-Earth3 r1, with -12 m on average, while IPSL-CM6A-LR is furthest away, underestimating the Z500 field by 59 m on average. The models with the highest differences after that are CNRM-ESM2-1 (-44 m) and MIROC6 (-37 m). Figure 12 allows an investigation of the spatial pattern of the deviation. Note that the two EC-Earth3 realizations show almost identical difference patterns, with a center

Z500 mean field difference

Model	Difference [m]
EC-Earth3 r1	-11.89
UKESM1-0-LL	-13.33
CESM2	14.89
MPI-ESM1-2-HR	-15.81
EC-Earth3 r10	-18.48
MIROC6	-37.44
CNRM-ESM2-1	-44.09
IPSL-CM6A-LR	-58.9

(a)

NAO gradient

Model	Gradient [m]
CESM2	520.88
EC-Earth3 r10	447.6
EC-Earth3 r1	434.72
ERA5	421.61
UKESM1-0-LL	415.7
MPI-ESM1-2-HR	402.07
IPSL-CM6A-LR	371.25
CNRM-ESM2-1	358.08
MIROC6	315.06

(b)

Temperature mean field difference

Model	Difference [°C]
CESM2	0.2
MPI-ESM1-2-HR	-0.63
MIROC6	0.76
CNRM-ESM2-1	-0.92
IPSL-CM6A-LR	-1.6
UKESM1-0-LL	-2.17
EC-Earth3 r1	-2.24
EC-Earth3 r10	-3.23

(c)

Figure 11: **Ranking of model biases.** (a) Mean difference in Z500 compared to ERA5, calculated as the weighted area mean over the significant difference plot. Models are ranked by their absolute difference. (b) Difference between the northern and southern center of action of the NAO, located at the southern tip of Greenland and the north-western tip of the Iberian Peninsula. Green shading indicates that the difference to ERA5 is not of statistical significance, according to a z-test. (c) Mean difference in temperature, calculated identical to Z500 mean difference.

between Iceland and the British isles, but stronger in magnitude for r10. Both CNRM-ESM2-1 and IPSL-CM6A-LR show a well-spread overall bias of underestimating geopotential height. The largest differences, up to 100 m, are located before the coast of and over western Europe, where the mean flow pattern diverges. Therefore, the differences to ERA5 do not seem to be in contradiction to the general pattern, potentially even enhancing the mean flow. This seems to be the case for most of the models. Two models, however, show a bipolar difference pattern, that could imply changes to the mean flow. CESM2 shows lower geopotential height than ERA5 over Greenland and Iceland and higher than ERA5 over the southern North Atlantic and most of Europe, thereby enhancing the north-south gradient compared to ERA5. MIROC6 shows the opposite pattern, overestimating Z500 in the north and underestimating it in the south, thereby reducing the gradient. This difference is relevant in that it interacts with the mean pattern, but the metric of mean difference is not very suitable to capture it, as positive and negative differences cancel each other out. However, the shape of the bipolar difference notably resembles the NAO pattern. Therefore, this difference can be captured by comparing the gradient between the two NAO centers of action to the same gradient observed in ERA5. This NAO gradient comparison is listed in Table (b) in Figure 11, where green shading indicates that the difference to the gradient observed in ERA5 data is not of statistical significance. The two models with the bipolar difference pattern, CESM2 and MIROC6, show up at the extreme ends of this comparison, with CESM2 largely overestimating the difference between the centers of action in southern Greenland and the Iberian peninsula and MIROC6 underestimating it, both by around 100 m compared to ERA5. Furthermore, also CNRM-ESM2-1 and IPSL-CM6A-LR show up as significantly underestimating the difference in geopotential height, which can be attributed to their underestimation of Z500 around southwestern Europe, but not over Greenland. EC-Earth3 r10 significantly overestimates the gradient, due to its underestimation of Z500 around Iceland, strengthening the Icelandic low. The connections between biases in the mean field and the EOF patterns will be discussed in Section 5.

As previously mentioned, the geopotential height field is closely linked to temperature. The dif-

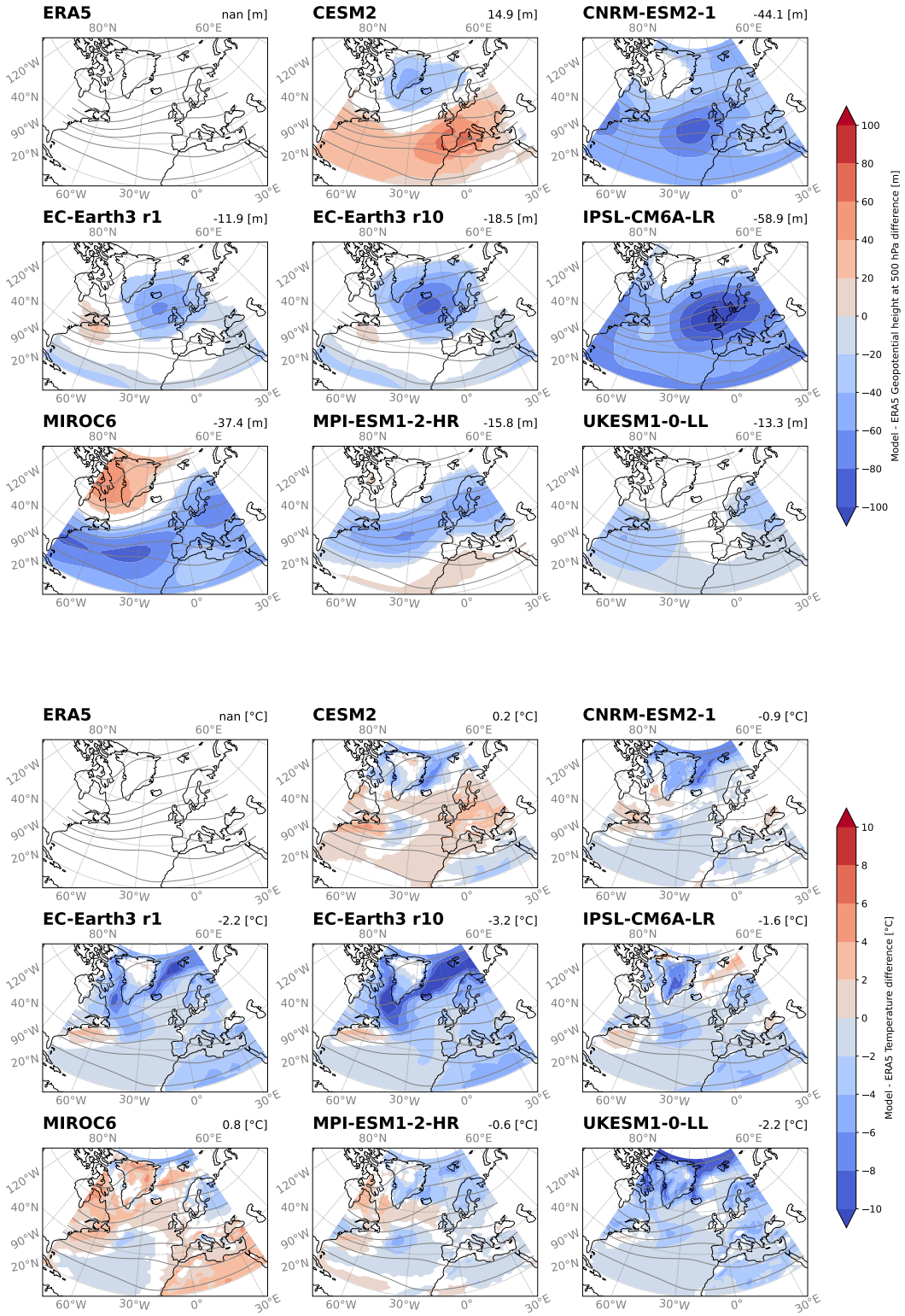
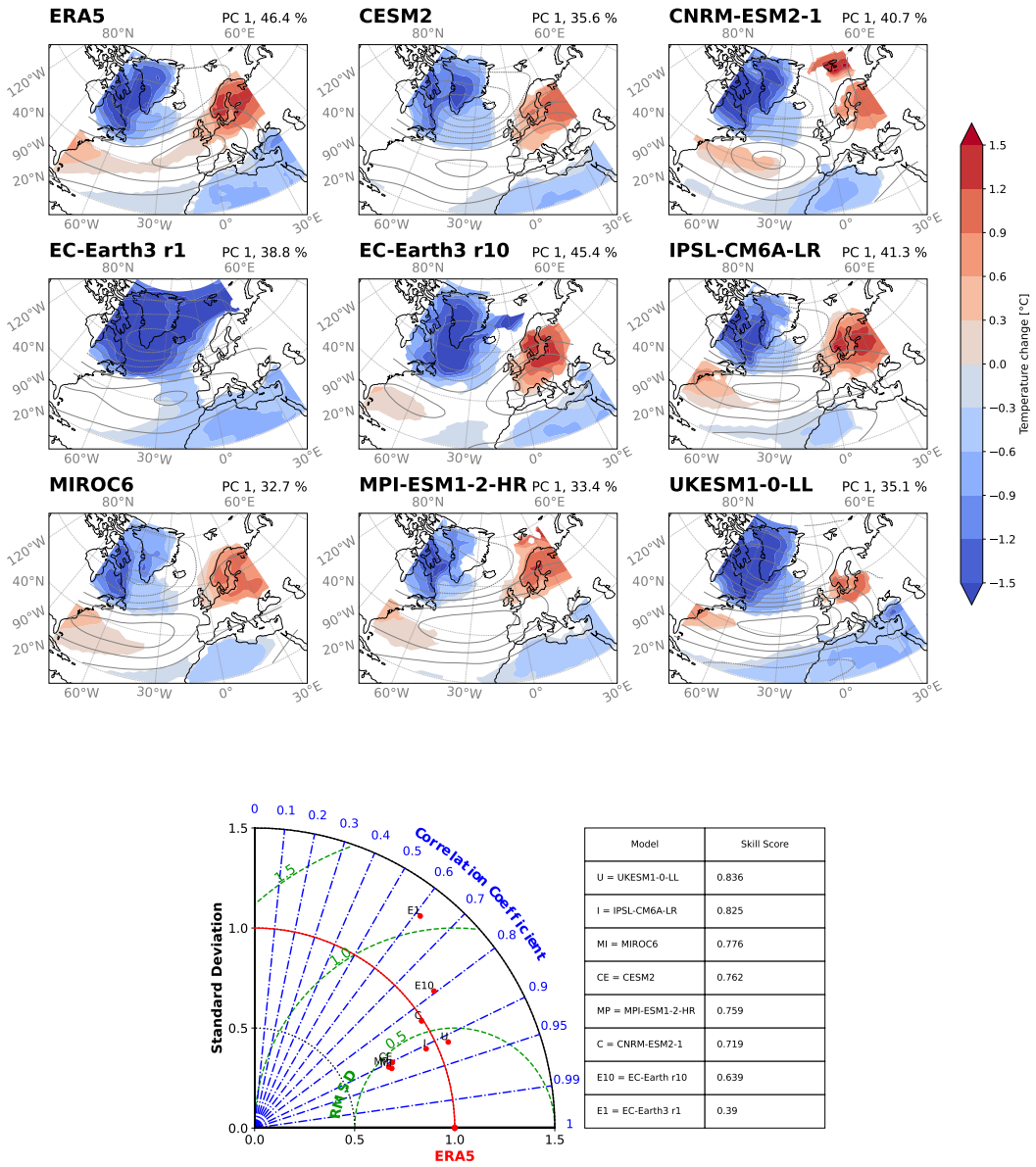


Figure 12: **Spatial patterns of model biases.** Difference in the mean Z500 (upper) and temperature (lower) field to ERA5, calculated by subtracting the ERA5 mean from the model mean at every grid point. Only locations where the difference is statistically significant at 5% level, determined with a z-test (see Section 3.5), are shown. The number in the upper right corner of each plot denotes the area weighted mean over the whole field. Contour lines show the ERA5 mean Z500 field.

ference maps of temperature are shown in the lower panel of Figure 12, while the right table in Figure 11 ranks the models after their mean temperature difference. Most models underestimate temperature, by up to 3.2°C on average (EC-Earth3 r10). The only models with a net positive temperature bias are CESM2 (0.2°C) and MIROC6 (0.8°C), whose temperature difference patterns show spatial resemblance to the bipolar patterns of difference in geopotential height. For the other models, not much resemblance between the Z500 and temperature difference patterns is visible. Note that several models show the largest magnitude of differences up to over 10°C close to the northern boundary of the domain. In the two EC-Earth3 ensemble members, again showing a very similar pattern in spatial extent, but with stronger magnitude in r10, this difference is strong over large parts of the northern north Atlantic, but notably not present over the Greenlandic land mass. A potential explanation for this observation, as well as potential connections between the temperature biases of the models and their modes of variability and teleconnections will be discussed in Section 5.
The differences in the precipitation mean field are more scattered and do not show clear patterns like in the case of Z500 and temperature. They are shown in Figure A.5 in the Appendix.

4.2.3 NAO Teleconnections

Temperature
The following section compares the NAO teleconnections, as captured by a regression of temperature and precipitation anomalies against PC1, of the CMIP6 models to ERA5. Beginning with temperature, Figure 13 shows the spatial patterns of temperature PC regression at robust locations and the corresponding Taylor diagram and skill score ranking. Five out of eight models show a high pattern correlation with coefficients around 0.9. Out of these, the best skill scores are awarded to UKESM1-0-LL ($S = 0.84$) and IPSL-CM6A-LR ($S = 0.83$) for lying closest to the ERA5 standard deviation ($\sigma^*_{norm} = 1.06$ and 0.94, respectively), while MIROC6, CESM2 and MPI-ESM1-2-HR are clustered close together at lower standard deviations (around $\sigma^*_{norm} = 0.75$, resulting in $S = 0.78, 0.76, 0.76$, respectively). Like in the case of Z500 PCR, CNRM-ESM2-1 matches ERA5 closely in terms of standard deviation ($\sigma^*_{norm} = 0.99$), but shows a lower pattern correlation than most models ($R = 0.84, S = 0.72$). The two EC-Earth3 ensemble members show least similarity to the ERA5 pattern. With a correlation coefficient of 0.8 and a higher standard deviation than ERA5 ($\sigma^*_{norm} = 1.13$), EC-Earth3 r10 earns a skill score of 0.64, while EC-Earth3 r1 is a true outlier with $R = 0.62$, $\sigma^*_{norm} = 1.34$ and a skill score of only 0.4.
Like in the case of Z500 PC regression, the full field comparison results in a very similar Taylor diagram pattern and thus skill score ranking, with slightly higher correlations. The most notable difference is CNRM-ESM2-1 lying closer to the best performing models in terms of correlation, improving its overall ranking from rank 6 to 4. The Taylor diagram and skill score table for the full field comparison can be seen in the Appendix (Figure A.3).
Looking at the spatial regression patterns in Figure 13 confirms that most models do reasonably well at reproducing the quadripolar NAO temperature response. EC-Earth3 r1 stands in strong contrast to the rest, as it is the only model that shows no positive temperature response at all. A look at the full regression field (see Appendix Figure A.3) shows that it does have a very weak positive relationship in Scandinavia and the US east coast, but with too low correlation values to appear in the masked plot. Interestingly, this is not the case for EC-Earth3 r10, which, along with the rest of the models, shows a quite similar pattern to ERA5. Despite the overall similarity to the ERA5 pattern, each of the models shows smaller deviations in some locations. CESM2 is missing the positive response over the US East coast and over northern Norway, while generally showing lower temperature change amplitudes. CNRM-ESM2-1 also has too little response over northern Norway and the British isles, but a strong positive response over Svalbard that is not present in the ERA5 data. EC-Earth3 r10 looks very similar to ERA5, but with a too strong negative response



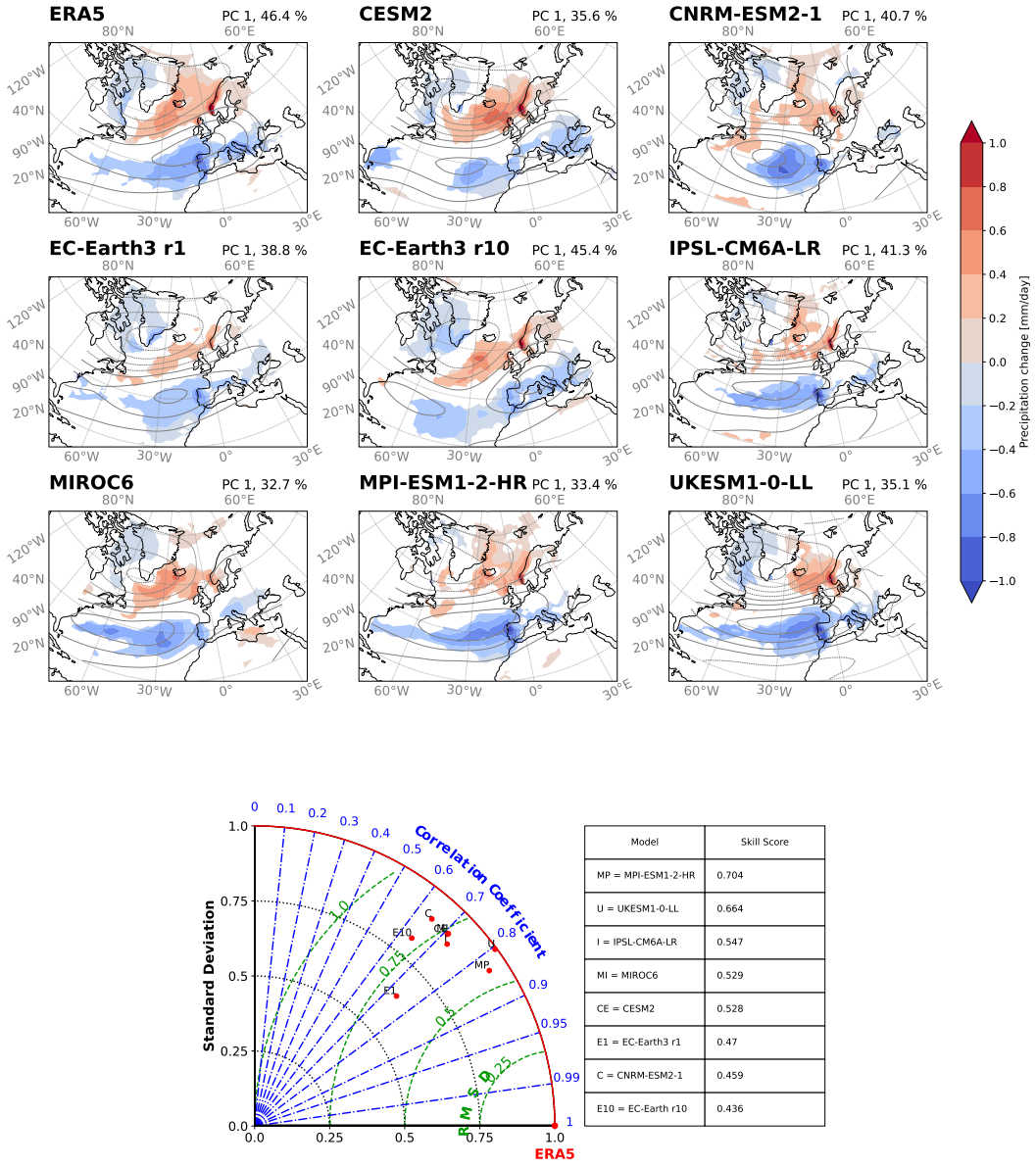
Model	Skill Score
U = UKESM1-0-LL	0.836
I = IPSL-CM6A-LR	0.825
MI = MIROC6	0.776
CE = CESM2	0.762
MP = MPI-ESM1-2-HR	0.759
C = CNRM-ESM2-1	0.719
E10 = EC-Earth r10	0.639
E1 = EC-Earth3 r1	0.39

Figure 13: **Comparison of NAO temperature response in all models against ERA5.** Upper: Spatial patterns of regression of temperature anomalies against the PC time series associated with NAO. Colors indicate the linear regression coefficient, displayed at robust locations (correlation coefficient > 0.3). Z500 EOF patterns associated with NAO as contour lines. Lower: Taylor diagram and skill score ranking. Standard deviation normalized with ERA5 $\sigma^* = 0.62$°C.

around Greenland and also missing some response in northern Scandinavia. IPSL-CM6A-LR is missing the negative response over large parts of northern Africa. MIROC6 shows a response that is largely matching ERA5 in its spatial extent, but too weak in amplitude, MPI-ESM1-2-HR is underestimating the negative relationship around Greenland and finally UKESM1-0-LL seems to match the ERA5 response very well in terms of location and amplitude in most places, but is missing the positive response over large parts of northern Scandinavia and northwestern Europe. This is noteworthy because despite being awarded the highest skill score, UKESM1-0-LL is not able to capture the complete temperature response pattern either.
In conclusion, the relationships that show up in all models (except EC-Earth3 r1) are the negative response around the Labrador sea, the positive response around the Baltic sea and the negative response over northeast Africa. Additionally it can be said that visually the difference between the best and worst performing models, except for EC-Earth3 r1, is not as large, suggesting that any pattern correlation coefficient of 0.8 or above seems sufficient to indicate a good agreement with ERA5.

Precipitation

The first notable observation when considering the Taylor diagram of the precipitation PC regression model comparison in Figure 14 is the considerably lower pattern correlation coefficients compared to Z500 and temperature regression. Only two models obtain a pattern correlation coefficient above 0.8, namely MPI-ESM1-2-HR ($R = 0.83$) and UKESM1-0-LL ($R = 0.81$), making those two the models with the highest skill scores ($S = 0.7$ and 0.66, respectively). UKESM1-0-LL also matches ERA5 most closely in terms of standard deviation, with $\sigma^*_{norm} = 0.99$. IPSL-CM6A-LR, MIROC6 and CESM2 lie close together in the diagram, with correlation coefficients of around 0.71 and normalized standard deviations of around 0.9. EC-Earth3 r1 stands out again with the highest difference in standard deviation ($\sigma^*_{norm} = 0.64$), but shows one of the highest correlation coefficients with ERA5 ($R = 0.73$). CNRM-ESM2-1 and EC-Earth3 r10 obtain the lowest skill scores because of their low pattern correlation coefficients of 0.65 and 0.64, respectively. In contrast to the temperature comparison, the Taylor diagram does not show one model to clearly be worst at matching the ERA5 precipitation response pattern, but rather that the models are grouped relatively close together and seem to all not be able to capture the exact same pattern as ERA5. Once again, the full field comparison (ref. Figure A.4) shows very comparable results and is thus disregarded here.
Considering the spatial patterns in the upper panel in Figure 14 confirms that the precipitation response pattern seems to be harder to capture than temperature. While all models show the general pattern of a positive response in the north and a negative response in the south of the domain, the patterns are scattered, showing only few locations with robust relationships, and of weak amplitude. The strongest response locations at the southwestern coast of Norway and the westcoast of the Iberian peninsula are present in all models, but many other locations are missing the precipitation response in some models, e.g. Iceland in both EC-Earth3 members, Scotland in most models except CESM2, northern and eastern Scandinavia in most models except CESM2 and EC-Earth3 r10 (and MPI-ESM1-2-HR), southeastern Europe in CNRM-ESM2-1 and MIROC6, or northwestern Canada in more than half of the models (CNRM-ESM2-1, EC-Earth3 r1, IPSL-CM6A-LR, MIROC6, MPI-ESM1-2-HR). Interestingly, five models show a negative response at the southeast coast of Greenland, that is not present (or robust) in the ERA5 data. It is also notable that even the model with the highest pattern correlation, MPI-ESM1-2-HR is not capturing the full precipitation response, as it is missing almost all of the negative response around Baffin Bay. Similarly, it is hard to say whether the lowest ranking model realization, EC-Earth3 r10 does indeed show the precipitation response pattern that is most different from ERA5. In line with that, it is hard to confirm the Taylor skill score ranking with visual inspection, as the precipitation response is scattered in most models and all show some shortcomings compared to ERA5.



Model	Skill Score
MP = MPI-ESM1-2-HR	0.704
U = UKESM1-0-LL	0.664
I = IPSL-CM6A-LR	0.547
MI = MIROC6	0.529
CE = CESM2	0.528
E1 = EC-Earth3 r1	0.47
C = CNRM-ESM2-1	0.459
E10 = EC-Earth r10	0.436

Figure 14: **Comparison of NAO precipitation response in all models compared to ERA5.** Equivalent to Figure 13 for precipitation. Standard deviation normalized by ERA5 standard deviation $\sigma^* = 0.23$ mm/day.

In summary, comparing the results of the model comparison with regards to the NAO pattern captured by EOF1 and both temperature and precipitation teleconnections, it can be said that most models show relatively high agreement with ERA5 with regards to the EOF1 NAO pattern, a little lower agreement with regards to its temperature regression, and clearly lower agreement when it comes to the NAO precipitation response. The model ranking is very similar for the temperature and precipitation regression, with UKESM1-0-LL and IPSL-CM6A-LR with high agreement with ERA5, MIROC6 and CESM2 agreeing slightly less and CNRM-ESM2-1, and the two EC-Earth3 realizations showing least agreement with ERA5. Only MPI-ESM1-2-HR performs best for precipitation but slightly disqualifies itself in the temperature regression due to a high standard deviation difference to ERA5. More interesting, however, is the connection between the models performance w.r.t. the NAO EOF pattern and their representation of its teleconnections. The models with the teleconnection patterns matching those of ERA5 best, UKESM1-0-LL, IPSL-CM6A-LR and MPI-ESM1-2-HR, are also the ones ranking highest in the comparison of EOF1 patterns. The model with the lowest pattern correlation of EOF1 is CNRM-ESM2-1, which seems to directly translate into low correlation coefficients of the NAO teleconnections. They can be associated with the shortcomings of the EOF1 pattern, as the tilt of the northern center of action towards northern Europe results in missing temperature and precipitation responses in northern Scandinavia and instead responses around Svalbard that do not exist in the ERA5 data, and the westward shift of the southern center of action results in the southern temperature and precipitation responses not extending as far east as in ERA5. EC-Earth3 r10 is a similar case of low agreement in all three PC regression cases, having a slightly higher Z500 regression pattern correlation, a lower temperature regression pattern correlation and a similar precipitation regression pattern correlation than CNRM-ESM2-1. In contrast to this, MIROC6, CESM2 and EC-Earth3 r1 have similar EOF1 pattern correlations, with EC-Earth3 r1 agreeing most with ERA5 out of these three, but while MIROC6 and CESM2 do reasonably well at matching ERA5's teleconnections, with correlation coefficients around 0.9 for temperature and 0.7 for precipitation, EC-Earth3 r1 is not able to reproduce the teleconnection patterns seen in ERA5, showing too weak precipitation responses and large spatial differences and a missing positive response in the temperature pattern. The left table in Figure 15 ranks the models after their mean skill score for all three PC1 regressions combined. UKESM1-0-LL is ranked first, with a combined skill score of 0.81, closely followed by MPI-ESM1-2-HR (0.78) and IPSL-CM6A-LR (0.75). Ranked last are the two EC-Earth3 members, with scores of 0.61 (r10) and 0.57 (r1). Overall, it is notable that the models skill scores are all relatively close together.

Combined NAO Score

Model	Skill Score
UKESM1-0-LL	0.804
MPI-ESM1-2-HR	0.779
IPSL-CM6A-LR	0.753
CESM2	0.7
MIROC6	0.671
CNRM-ESM2-1	0.632
EC-Earth3 r10	0.608
EC-Earth3 r1	0.567

Combined AR Score

Model	Skill Score
MIROC6	0.631
IPSL-CM6A-LR	0.608
MPI-ESM1-2-HR	0.587
CNRM-ESM2-1	0.511
CESM2	0.508
EC-Earth3 r1	0.436
UKESM1-0-LL	0.401
EC-Earth3 r10	0.211

Combined SB Score

Model	Skill Score
MPI-ESM1-2-HR	0.608
IPSL-CM6A-LR	0.493
CESM2	0.373
UKESM1-0-LL	0.342
CNRM-ESM2-1	0.336
EC-Earth3 r10	0.32
MIROC6	0.301
EC-Earth3 r1	0.122

Figure 15: **Combined model skill score ranking.** Models ranked after their mean skill score for Z500, temperature and precipitation PC regression (robust correlation locations) for the NAO (left), AR (center) and Scandinavian blocking pattern (right).

4.2.4 EOF2: Atlantic Ridge

After evaluating the models ability to represent the NAO and its teleconnections with the EOF approach, the following EOF patterns and their temperature and precipitation responses are investigated. Figure 16 shows the second EOF pattern, that has been identified to resemble the Atlantic Ridge (AR) pattern in ERA5. Most models have a similar proportion of explained variance on PC2 to ERA5 (18.1%), but both EC-Earth3 ensemble members and UKESM1-0-LL stand out with lower explained variances. In the case of EC-Earth3 r1 and UKESM1-0-LL, this is due to the fact that in a pre-evaluation, the EOF pattern of PC3 has been identified to resemble the Atlantic Ridge pattern more closely than PC2, which is why PC3 will be used for the comparison of those two models.

The Taylor diagram and skill score ranking in the lower panel of Figure 16 show that five of the models, namely IPSL-CM6A-LR, MPI-ESM1-2-HR, CNRM-ESM2-1, MIROC6 and UKESM1-0-LL, show a similar pattern to ERA5, with pattern correlation coefficients around 0.9 (except UKESM1-0-LL, with $R = 0.85$). They all show a center of strong geopotential height anomalies over the central North Atlantic and a band of opposite sign around it, stretching from the northern to the western boundary of the domain. However, in IPSL-CM6A-LR, MIROC6 and MPI-ESM1-2-HR this band does not cover the same extent as in ERA5, while CNRM-ESM2-1 and especially UKESM1-0-LL instead show a too weak and small center over the Atlantic and a pronounced band of opposite sign extending too high into the Arctic ocean. CESM2 and EC-Earth3 r1 are awarded lower scores, due to lower pattern correlations ($R = 0.77$ and 0.74, respectively). In the CESM2 data, the Atlantic center is shifted east and the band very weak, while EC-Earth3 r1 is characterized by a westward shift of the pattern. Lastly, EC-Earth3 r10 obtains the lowest score, due to low correlation ($R = 0.57$) and a lower standard deviation than ERA5 ($\sigma^*_{norm} = 0.74$). Its EOF2 pattern does not show much resemblance to the Atlantic Ridge pattern, additionally to the Atlantic center displaying a second center of the same sign over northeastern Europe and a weak center of opposite sign over Greenland.

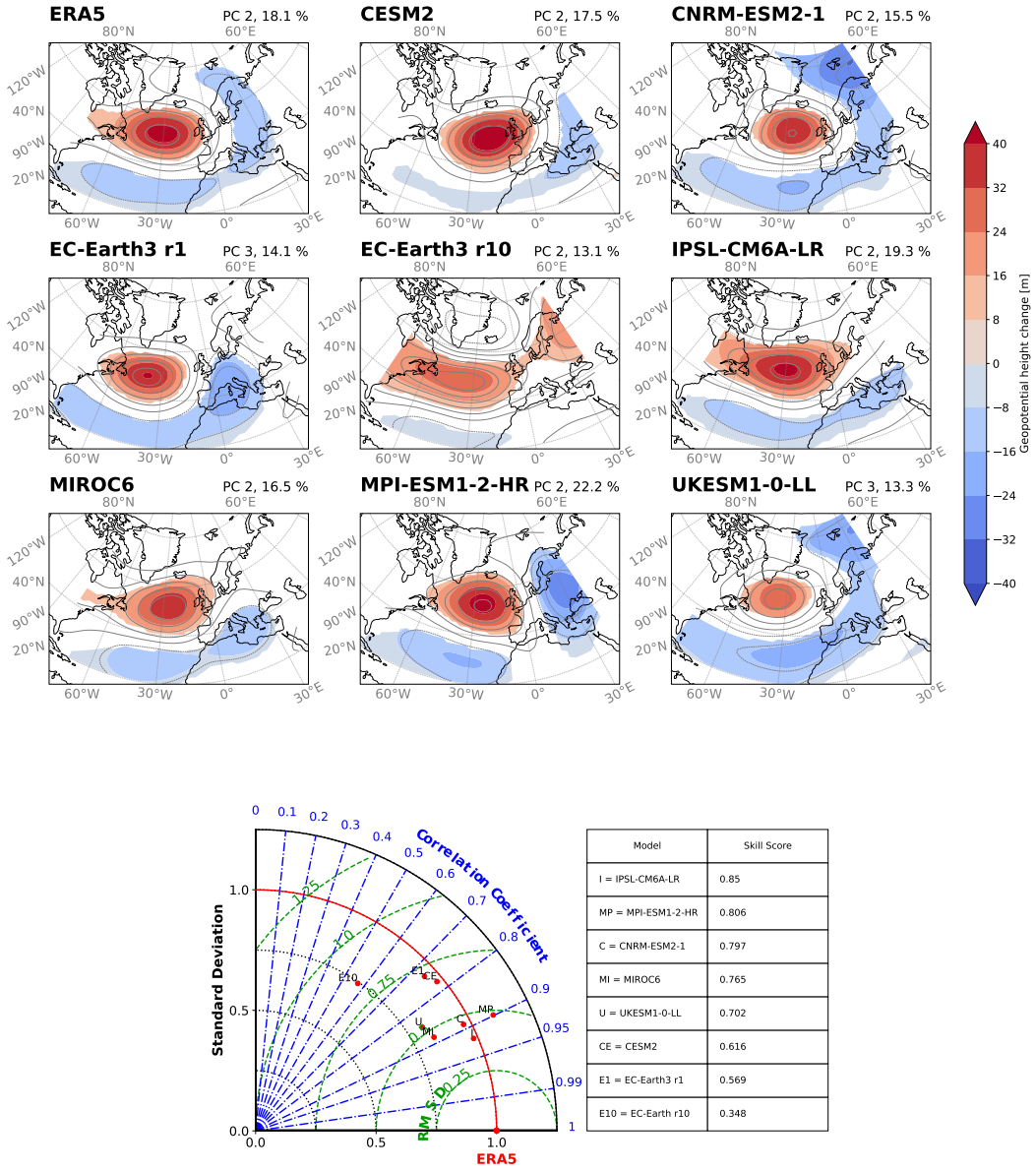
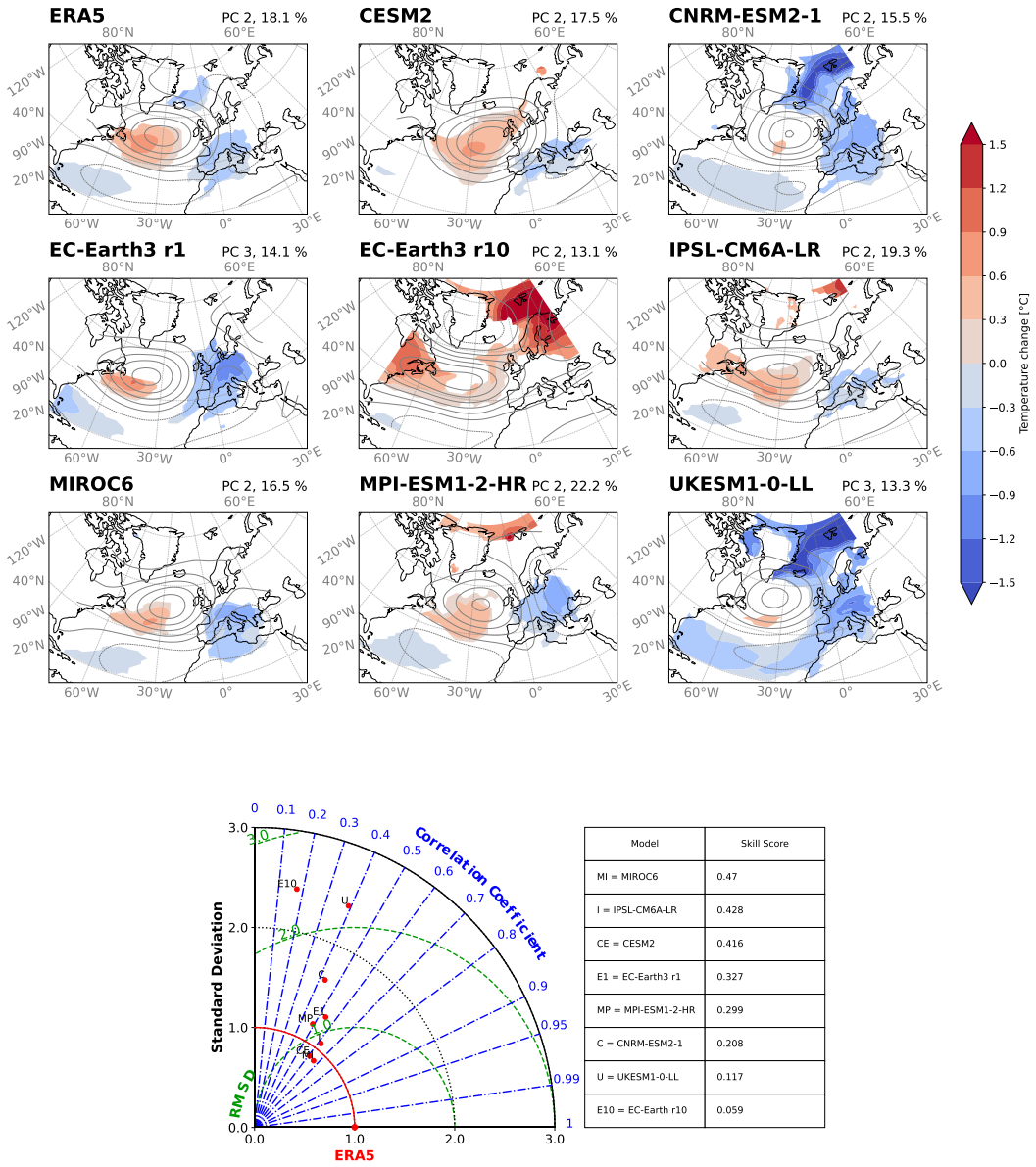


Figure 16: **Comparison of EOF patterns associated with AR in all models compared to ERA5.** Equivalent to Figure 10 for AR pattern. Note that the rank order of the used PC time series is shown in the upper right corner of each plot, indicating when EOF swapping was deemed necessary. Standard deviation normalized by ERA5 standard deviation $\sigma^* = 12.72$ m.



Model	Skill Score
MI = MIROC6	0.47
I = IPSL-CM6A-LR	0.428
CE = CESM2	0.416
E1 = EC-Earth3 r1	0.327
MP = MPI-ESM1-2-HR	0.299
C = CNRM-ESM2-1	0.208
U = UKESM1-0-LL	0.117
E10 = EC-Earth r10	0.059

Figure 17: **Comparison of Atlantic Ridge temperature response in all models compared to ERA5.** Equivalent to Figure 13 for the AR pattern. Standard deviation normalized by ERA5 standard deviation $\sigma^* = 0.2$°C.

Atlantic Ridge Temperature Teleconnections

The first insight when looking at the temperature regressions with the PC time series associated with the AR pattern is the much lower agreement with ERA5 than seen in the previous comparisons. The Taylor diagram in the lower panel of Figure 17 shows correlation coefficients ranging from $R = 0.66$ (MIROC6) to as low as $R = 0.17$ (EC-Earth3 r10) and normalized standard deviations up to $\sigma^*_{norm} = 2.4$ (also EC-Earth3 r10). Following its low agreement with the ERA5 EOF2 pattern, EC-Earth3 r10 thus also shows a temperature response that has little to do with the Atlantic Ridge teleconnections seen in ERA5, dominated by a strong positive response over Svalbard, the Greenland and Barents Sea and north-eastern Scandinavia, and the North American east coast. The temperature responses of UKESM1-0-LL and CNRM-ESM2-1 can be connected to the shortcomings of their EOF patterns, with a strong negative temperature response almost all around the domain, particularly over the Arctic ocean, but notably not Greenland, but showing almost no positive response in connection with the weak center of action in the center of the domain. The remaining five model realizations show higher resemblance to the ERA5 temperature

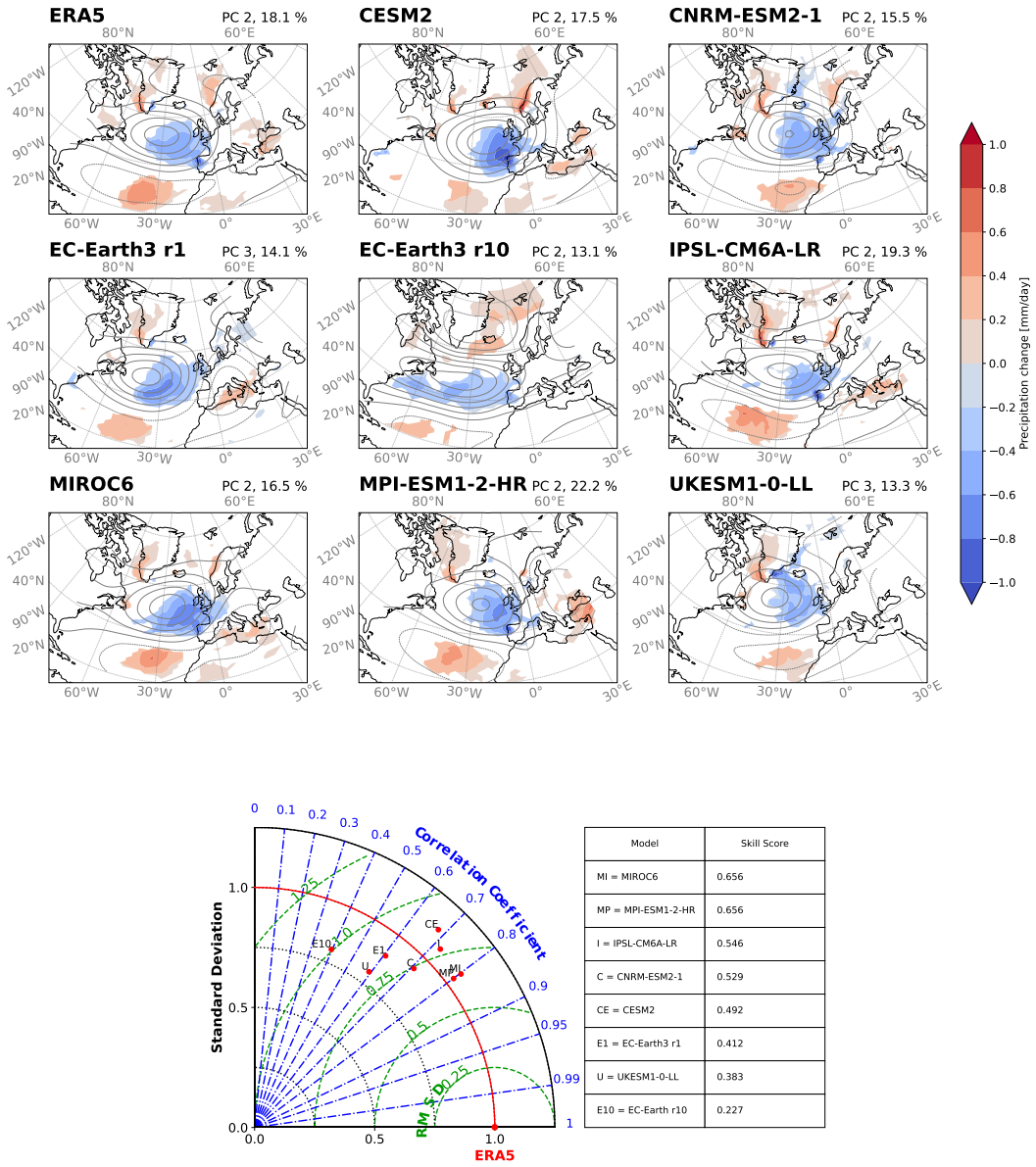


Figure 18: **Comparison of Atlantic Ridge precipitation response in all models compared to ERA5.** Equivalent to Figure 17 for precipitation. Standard deviation normalized by ERA5 standard deviation $\sigma^* = 0.17$ mm/day.

response pattern, with a positive response in the center of the domain and a negative response around, mainly over southwestern Europe. Note that MPI-ESM1-2-HR additionally shows a positive temperature response at the northern edge of the domain. This temperature response over the northern North Atlantic is also found - even more pronounced - in UKESM1-0-LL, CNRM-ESM2-1 and EC-Earth3 r10.

Atlantic Ridge Precipitation Teleconnections

The precipitation regression on the PC time series associated with the AR pattern shows, contrary to what was seen for the NAO teleconnections, higher agreement with ERA5 than the temperature regression. The Taylor diagram in Figure 18 shows the models spread between correlation coefficients of $R = 0.8$ (MIROC6 and MPI-ESM1-2-HR) and $R = 0.4$ (EC-Earth3 r10) and normalized standard deviations of $\sigma^*_{norm} = 1.1$ (MIROC6) to $\sigma^*_{norm} = 0.8$ (UKESM1-0-LL). All models reproduce the basic temperature response pattern seen in ERA5, with a negative response at the eastern side of the positive center of action and a positive response all around it. However, as pointed

out in Section 4.1, the interesting aspects of these response pattern are the several precipitation dipoles, between the southwestern and southeastern tip of Greenland, weakly in southern Norway and Sweden and more large-scale between the European west coast and the region around the Black Sea. None of the investigated CMIP6 models are able to reproduce all of these. As in the case of temperature, EC-Earth3 r10 clearly shows least resemblance to the ERA5 pattern, agreeing only at the European west coast. UKESM1-0-LL and EC-Earth3 r1 are ranked next lowest, agreeing only over the British Isles and southern Greenland, but notably showing the precipitation dipole there. The other five models capture the overall pattern well, but only IPSL-CM6A-LR shows the precipitation dipole over southern Greenland and only CESM2 and MPI-ESM1-2-HR show the positive response over the Black Sea.

Table (b) in Figure 15 shows the combined model skill scores for the AR Z500, temperature and precipitation regression. Following the previously presented results, EC-Earth3 r10 obtains the lowest skill score, followed by UKESM1-0-LL. Comparing this to the overall ranking for NAO (Table (a)), it is noticeable that the scores are generally considerably lower for the AR pattern, with a maximum score of 0.6 (MIROC6) and a minimum of 0.2 (EC-Earth3 r10), compared to 0.8 and 0.6 for NAO. Furthermore, it is interesting to see that UKESM1-0-LL shows greatest resemblance to ERA5 with regards to NAO, but almost least in the case of the Atlantic Ridge pattern, according to the skill score ranking.

4.2.5 EOF3: Scandinavian Blocking

After swapping the order of EOFs in five out of eight cases, the pattern called Scandinavian blocking in EOF3 of the ERA5 data can be identified in most models, shown in Figure 19. Most models show elements of the almost quadripolar pattern observed in ERA5, with a dipole between a center of action over northwestern Europe and around Gibraltar, and an opposite dipole at the western side of the domain, of which only the northern center of action over northeastern Canada shows a robust relationship, reducing it to a tripolar pattern. However, the differences to ERA5 are large, with many models showing east- or northwards shifts in the centers of action and only half of the models reproducing all three robust centers of action. This is reflected in the Taylor diagram showing low correlation coefficients and low overall skill scores. MPI-ESM1-2-HR and IPSL-CM6A-LR are awarded the highest skill scores due to correlation coefficients above 0.8, because they show all three centers of action with small shifts in location compared to ERA5. However, the third highest skill score is awarded to CNRM-ESM2-1, which visually seems like one of the least similar patterns to ERA5, reproducing only the Scandinavian center of action. EC-Earth3 r10 and UKESM1-0-LL are ranked slightly lower, along with CESM2, but reproduce the tripolar pattern in greater similarity to ERA5. MIROC6 and EC-Earth3 r1 show the lowest pattern correlations, around 0.4, due to a north and eastward shift of the centers of action in MIROC6 and a strong dominance of the Scandinavian center of action in EC-Earth3 r1, which is much larger in extent, covering the whole of Europe and extending westward over the North Atlantic until the North American east coast. This pattern shows little resemblance to the Scandinavian blocking pattern observed in ERA5 and the other models. The fraction of explained variance associated with the respective patterns of variability is similar to ERA5 (10.5%) in most cases, despite the frequent EOF swapping necessary. EC-Earth3 r1 show the largest difference in explained variance (15.6%).

Scandinavian Blocking Temperature Regression

The temperature regression pattern with the PC time series associated with the Scandinavian blocking pattern shows similarly low agreement with the ERA5 pattern as seen for the AR teleconnections. The Taylor diagram shown in Figure 20 shows correlation coefficients below 0.8, more specifically between 0.75 (MPI-ESM1-2-HR) and 0.3 (EC-Earth3 r10). EC-Earth3 r1 appears as an extreme outlier in the diagram, because of an extreme difference in standard deviation ($\sigma_{norm} = 3.5$) and no correlation ($R = 0$) with ERA5. Its temperature response pattern consequently shows no similarities to the ERA5 pattern, with a positive temperature response covering almost the entire domain, notably excluding Greenland, that is strongest over the northern North Atlantic and eastern and northern Europe. The rest of the model realizations show few locations with a robust temperature response, that show parts of the positive response over Greenland, Iceland and northwestern Europe and the negative response over large parts of Europe and northeastern Canada observed in ERA5. MPI-ESM1-2-HR matches the ERA5 pattern best.

Scandinavian Blocking Precipitation Regression

The precipitation regression analysis shows even fewer locations and lower comparison scores than the temperature analysis. The temperature response patterns recorded for the models are weak and scattered, but correspond to the ERA5 pattern, with a negative response over northwestern Europe and a positive response around, at the Greenlandic southeast coast and over the Mediterranean sea. It is notable that the response patterns of the models correspond well to the centers of action of their EOF patterns, shown as grey contours. EC-Earth3 r1 again receives the lowest score, due to it showing a positive precipitation response over large parts of the northern North Atlantic, but almost no negative response.

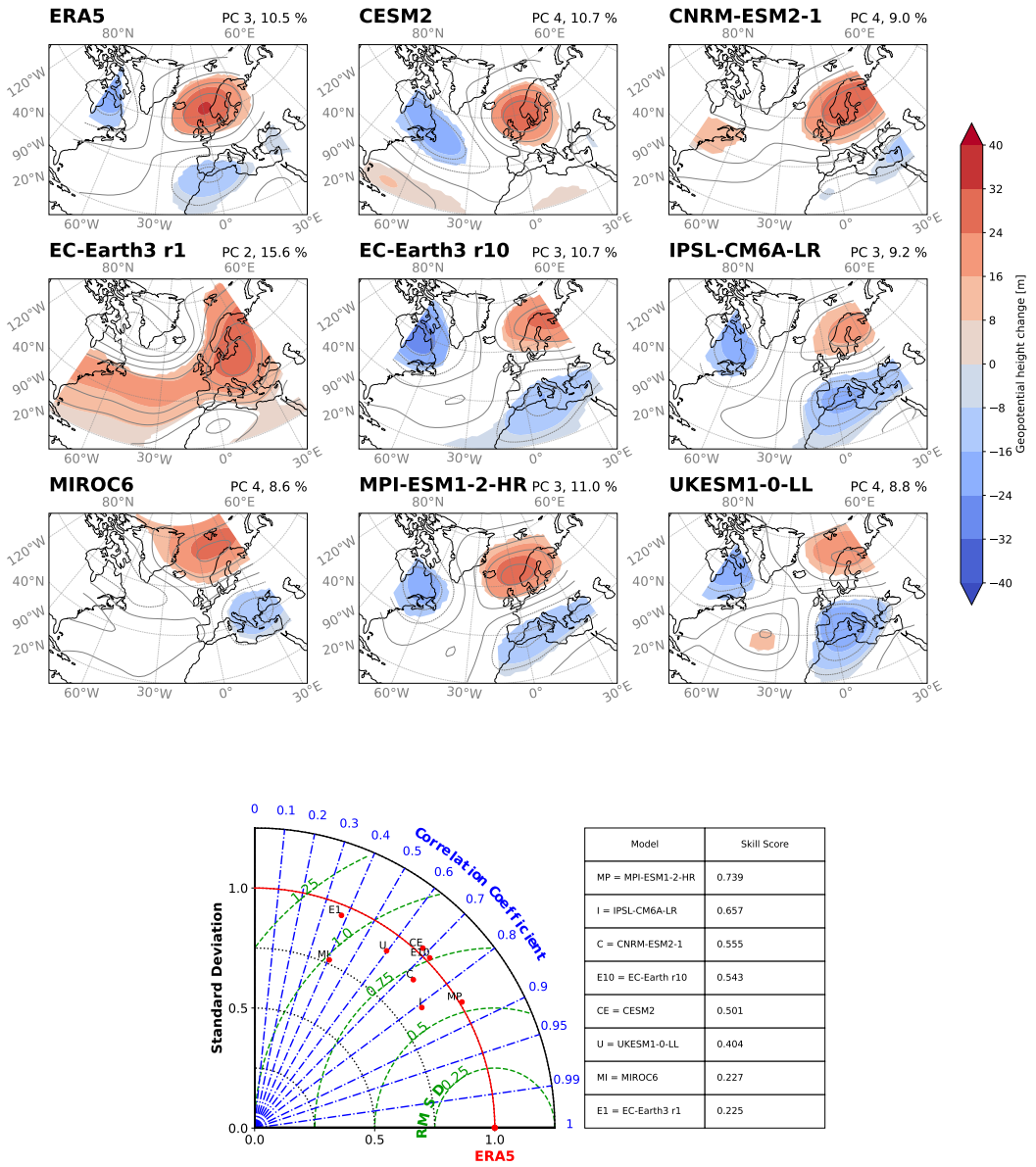
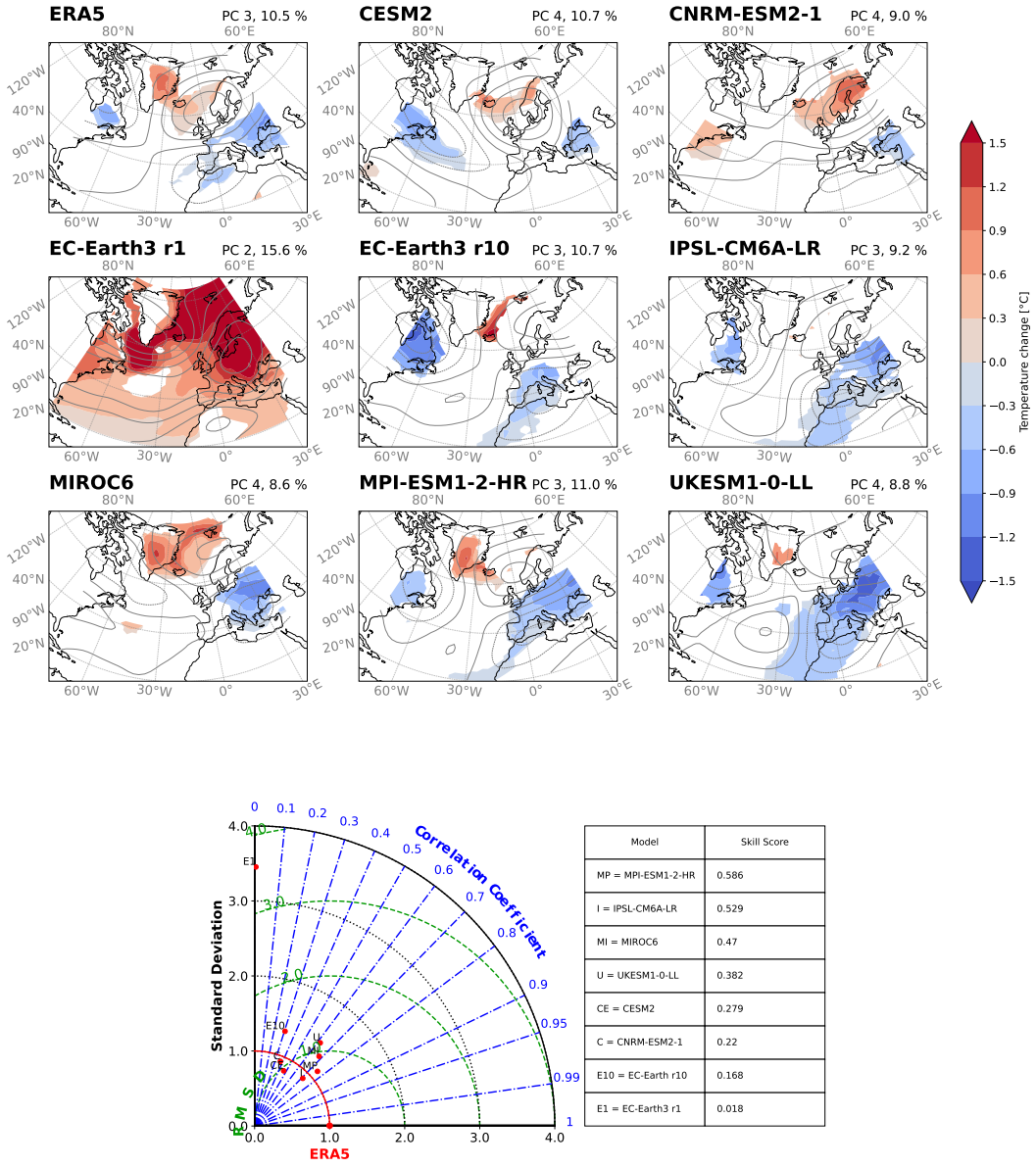
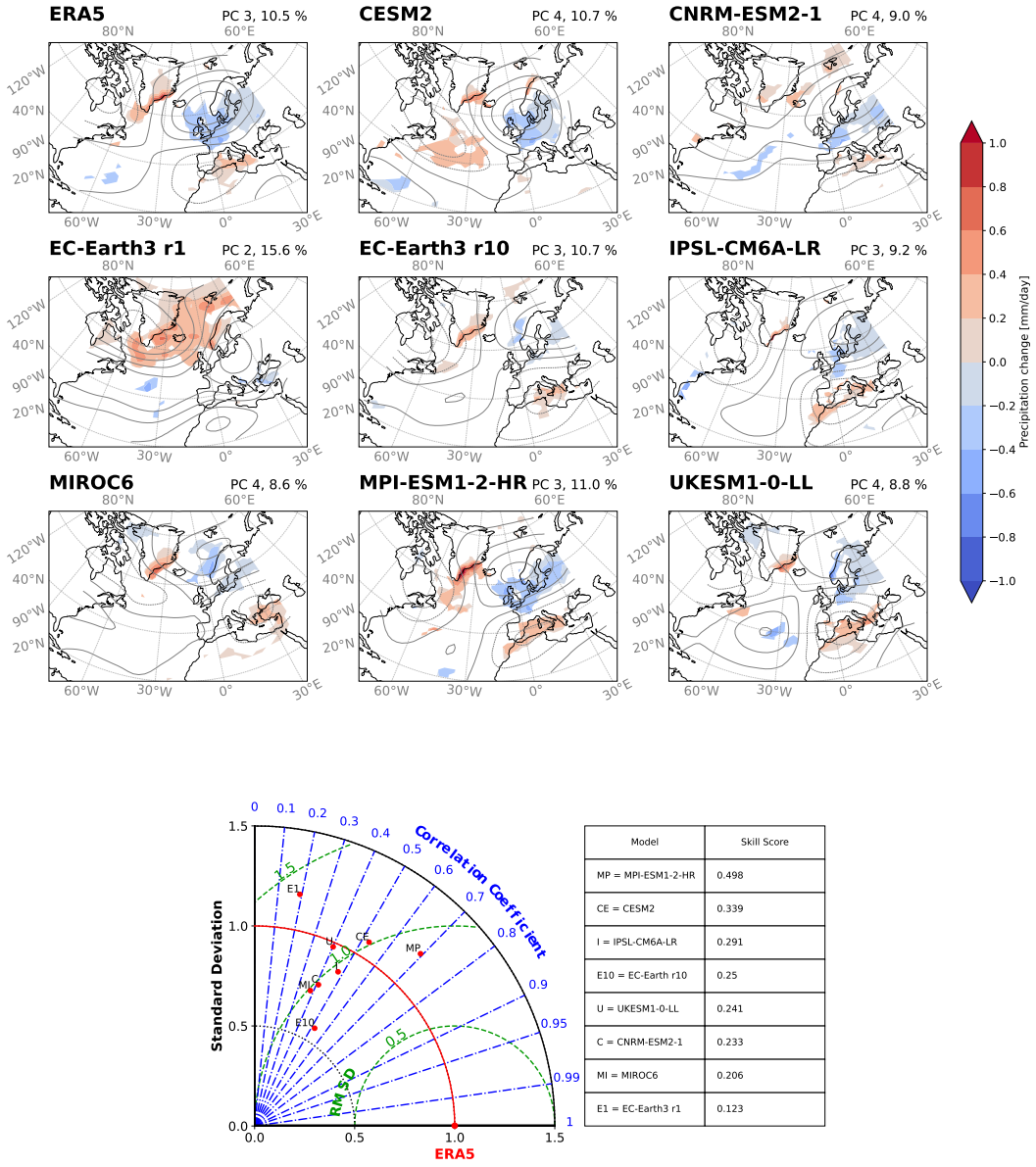


Figure 19: **Comparison of EOF patterns associated with SB in all models compared to ERA5.** Equivalent to Figure 10 for SB pattern. Note that the rank order of the used PC time series is shown in the upper right corner of each plot, indicating when EOF swapping was deemed necessary. Taylor diagram: Standard deviation normalized by ERA5 standard deviation $\sigma^* = 9.26$ m.



Model	Skill Score
MP = MPI-ESM1-2-HR	0.586
I = IPSL-CM6A-LR	0.529
MI = MIROC6	0.47
U = UKESM1-0-LL	0.382
CE = CESM2	0.279
C = CNRM-ESM2-1	0.22
E10 = EC-Earth r10	0.168
E1 = EC-Earth3 r1	0.018

Figure 20: **Comparison of Scandinavian blocking temperature response in all models compared to ERA5.** Equivalent to Figure 13 for Scandinavian blocking pattern. Taylor diagram: Standard deviation normalized by ERA5 standard deviation $\sigma^* = 0.24$°C.



Model	Skill Score
MP = MPI-ESM1-2-HR	0.498
CE = CESM2	0.339
I = IPSL-CM6A-LR	0.291
E10 = EC-Earth r10	0.25
U = UKESM1-0-LL	0.241
C = CNRM-ESM2-1	0.233
MI = MIROC6	0.206
E1 = EC-Earth3 r1	0.123

Figure 21: **Comparison of Scandinavian blocking precipitation response in all models compared to ERA5.** Equivalent to Figure 20 for precipitation. Taylor diagram: Standard deviation normalized by ERA5 standard deviation $\sigma^* = 0.11$ mm/day.

5 Discussion

This section discusses the results of the analysis. First, we evaluate the representation of modes of variability and their teleconnections in CMIP6 data, by putting the findings of the comparison analysis into context with each other and the literature, attempting to explain and interpret observed differences and finally summing up the results of the model comparison. Section 5.2 then critically evaluates the used methods for defining the modes and teleconnections and evaluating them against reanalysis data.

5.1 Discussion of results: How are modes of variability and their teleconnections represented in CMIP6 models?

5.1.1 Modes of variability

NAO

The analysis shows that in all investigated CMIP6 models the leading mode of variability, as characterized by an EOF analysis of the winter geopotential height field at 500hPa over the North Atlantic domain, can be identified as the NAO by its typical spatial pattern. The explained variance associated with it is lower than in the reanalysis data for all models, but always within the range of uncertainty of ERA5 (ref. Figure 9). The spatial patterns of the models EOF1 closely resemble the pattern observed in the ERA5 data, with high correlation coefficients, low centered RMSD and low difference in amplitude (standard deviation) (ref. Figure 10). This observation is in line with multiple studies evaluating the representation of NAO in CMIP models. Fasullo et al. (2020) and Lee et al. (2021) evaluate modes of variability, including the NAO, in CMIP3, CMIP5 and CMIP6 models. Both studies find high pattern correlations for NAO with observations and reanalysis data and improvements for CMIP6 compared to the older CMIP generations. These studies use different methods to complement the EOF approach and compensate for its shortcomings. Thus, it is important to keep in mind that the method used here is limited and has its shortcomings, which will be discussed later on in this chapter.
Despite the high overall agreement, the models show differences to the ERA5 EOF patterns in magnitude, location, shape and orientation of the centers of action. This is in line with findings by Fasullo et al. (2020). In the case of EOF1, especially the southern center of action tends to be underestimated in magnitude and shifted westward compared to ERA5. This westward displacement is also observed by Rousi et al. (2020) in their evaluation of the ECHAM5/MPI GCM. The northern center of action seems more robust, but is subject to underestimation and tilt in some models. MIROC6 stands out by underestimating the eastward extension and magnitude of both centers of action. CNRM-ESM2-1 shows a tilt of the whole pattern, resulting in the westward flow being directed southeast instead of northeast over Europe. The two EC-Earth3 ensemble members allow interesting insights into the origin of these differences. They show different EOF1 patterns, with r1 severely underestimating the magnitude and extent of the southern and overestimating the northern center of action, while r10 is characterized by a stronger wave shape of the pattern. Considering that both EC-Earth3 members are based on the same model and differ only in their initial conditions (Döscher et al. 2022), these differences in EOF pattern can be linked to decadal variability rather than differences in the underlying mechanisms.

The origins of these differences in EOF patterns can be investigated by considering model biases in the mean state of geopotential height and temperature. Due to the construction of the EOF patterns based on deviations from the mean state, observed differences in patterns of variability do not necessarily point to differences in the actual representation of variability, as they can also be caused by differences in the mean state (Davini et al. 2013). Thus, an evaluation of mean model biases can help to explain and attribute observed differences in EOF patterns. For this purpose, both differences in the mean field of Z500 and temperature to ERA5 are evaluated,

as well as the mean geopotential height gradient between the centers of action of the NAO pattern considered (ref. Figure 11 and 12). In the case of EC-Earth3, both realizations show the same shape of difference pattern to ERA5 for the mean fields of Z500 and temperature, but the differences are of stronger magnitude for r10. These patterns can partly explain the shape of the EOF1 pattern for r10, where the location of the Z500 bias south of Iceland may be associated with the southward bend of the southern center of action, while the shape of strong underestimation of temperature around the Arctic ocean matches the shape of the northern center of action. The latter connection between temperature bias and northern center of action of EOF1 can be confirmed for r1 as well, but the weak southern center of action can not be explained by these biases. It seems that in this case, differences in the mean state of the variables can only partly explain the observed EOF1 pattern, especially because they are unable to explain the differences between the two ensemble members of the same model. Also beyond EC-Earth3, differences in the mean geopotential height field seem to be linked to differences in the patterns of variability. The two models with the lowest Taylor skill score for EOF1, CNRM-ESM2-1 and MIROC6 are among the models with the highest mean Z500 difference and difference in NAO gradient. Our approach does not allow a conclusion on whether this means that the differences in the EOF1 pattern are largely caused by mean field biases instead of different representations of variability (Davini et al. 2013), but it calls for caution when judging models based on low similarity in their NAO patterns. In contrast, IPSL-CM6A-LR shows that a high difference in the mean state does not always lead to a high difference to the ERA5 EOF pattern. However, here, the high difference stems from a well-spread bias across the domain. If the overall bias was removed, the difference pattern would be small and focused around the British isles, enhancing the mean flow. Cases like this explain why the difference in geopotential height between the locations of the NAO centers of action seems to be better at explaining differences in NAO (EOF1) patterns than the mean difference over the whole region. Figures 10 and 11 show a high agreement between a low Taylor skill score for the NAO pattern and a significant over- or underestimation of the NAO gradient in the mean field. A strong deviation from the NAO gradient in either direction (CESM2 and MIROC6) seems to be connected to an underestimation of the magnitude of the NAO pattern, as both models show weaker centers of action than ERA5. However, a causal relationship between the NAO gradient in the mean field and the magnitude in the EOF1 pattern can not be established based on this observation, as a low magnitude in the centers of action is observed for models with a strong overestimation of the gradient (CESM2), a strong underestimation of the gradient (MIROC6) and no significant difference in the gradient compared to ERA5 (EC-Earth3 r1).

As geopotential height is strongly related to temperature, observed differences in the Z500 EOF patterns may also be due to model biases in temperature, as pointed out above in the case of EC-Earth3. Several studies have identified an eastward shift of the NAO, both in connection with global warming at the end of the 20th century (Jung et al. 2003), in simulations of future global warming (Ulbrich et al. 1999; Hu et al. 2004) and as a seasonal shift in the summer months (Portis et al. 2001). The global warming trend is removed in the here analysed data, but it could be hypothesized that an overall warm bias in a model may have the same effect on the NAO. However, the comparison of temperature differences in Figure 12 shows that most models actually underestimate temperature compared to ERA5, at least after linear detrending. The two models that do show a warm bias at least over parts of the domain (CESM2 and MIROC6) show a westward shift in their NAO pattern instead. Thus, this hypothesis does not apply to the here analysed data. Several studies also challenge this hypothesis, linking the observed eastward displacement of the NAO centers of action to a shift in trend towards a predominantly positive phase of the NAO (Peterson et al. 2003; Hurrell 2015) and a decrease in Greenland blocking (Davini et al. 2012) instead. This connection can not be investigated further with our results. A notable finding with respect to temperature biases is that several models show large temperature differences to ERA5 in the Arctic ocean (EC-Earth3, UKESM1-0-LL, CNRM-ESM2-1). The fact that they are not as strong over the Greenlandic land mass leads to the hypothesis that these differences are due

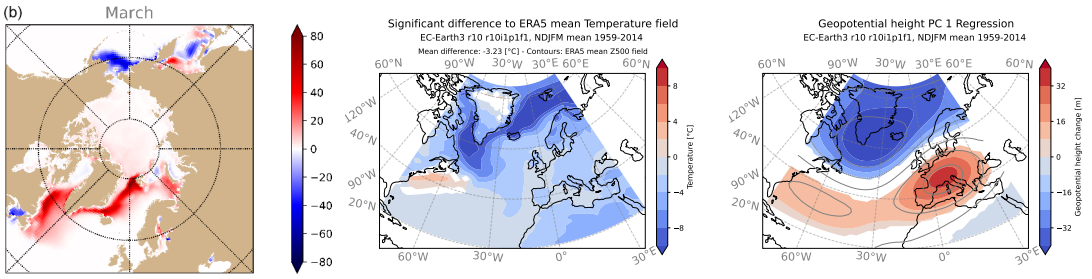


Figure 22: **Connection between observed temperature bias and NAO pattern in EC-Earth3 and overestimation of Arctic sea ice.** Left: Difference in Arctic sea ice concentration in percent between the ensemble mean of EC-Earth3 and OSI SAF observations in March, averaged over 1980–2010. Figure taken from Döscher et al. (2022). Center: Difference in mean temperature NDJFM 1959-2014 to ERA5 for EC-Earth3 r10. Right: NAO pattern, as represented by regression of Z500 anomalies onto the leading PC of Z500 for EC-Earth3 r10.

to differences in the representation of sea ice, compared to reanalysis data. This is confirmed for EC-Earth3 by Döscher et al. (2022), who find an overestimation of Arctic sea ice concentration in the EC-Earth3 historical runs ensemble in March 1980-2010 compared to satellite observations that matches the here observed strong temperature bias in its spatial extent, as shown in Figure 22. These northern temperature biases clearly have an impact on the observed circulation variability. In the two EC-Earth3 members, the shape of the northern center of action of the NAO corresponds well to the shape of the temperature bias (as shown for EC-Earth3 r10 in Figure 22). In the case of CNRM-ESM2-1 and UKESM1-0-LL there seems to be a inverse relationship, where the northern center of action of the NAO seems to be limited by the temperature bias along the northern edge of the domain and Svalbard (ref. Figure 10). This sensitivity of the NAO pattern to sea ice differences between the models provides an interesting insight in the context of changes in the NAO under future climate change, which is expected to be connected to rapid and large-scale changes of Arctic sea ice.

Other modes of variability

As the NAO pattern explains the largest fraction of variance in reanalysis and CMIP6 model data, the analysis mainly focuses on the evaluation of the representation of the NAO and its teleconnection patterns in the models. However, the applied methods also allow for an exploration of further modes of variability, offering additional insights into the variability mechanisms of the models. The second and third EOF pattern were identified as Atlantic Ridge and Scandinavian Blocking patterns in the ERA5 data. They are also reasonably well represented in most models, with lower skill scores than the NAO pattern. The spatial pattern identified as AR in the CMIP6 model data resembles the EOF2 pattern in ERA5 less than was the case for the NAO, with frequent underestimation of the center of action over the central North Atlantic and especially of the belt of opposite sign around it (ref. Figure 16). Two models (CNRM-ESM2-1 and UKESM1-0-LL) instead overestimate this belt, making it stretch further north over the ocean around Svalbard than observed in ERA5. This can be linked to the temperature bias potentially caused by sea ice differences, present in both models. It is interesting to see that the northern temperature bias linked to sea ice differences present in four models seems to interact mainly with the NAO pattern for two of the models (the two EC-Earth3 realizations) and shows a higher influence on the AR patterns for the other two models.

The third EOF pattern, identified as Scandinavian blocking in ERA5, is represented even more differently in the models, resulting in lower skill scores yet again. However, the majority of the models still shows the same basic characteristics of the pattern. Most differences are due to a northward shift of the pattern, or a domination of the Scandinavian center of action and its extension across the North Atlantic in the models compared to ERA5 (ref. Figure 19).

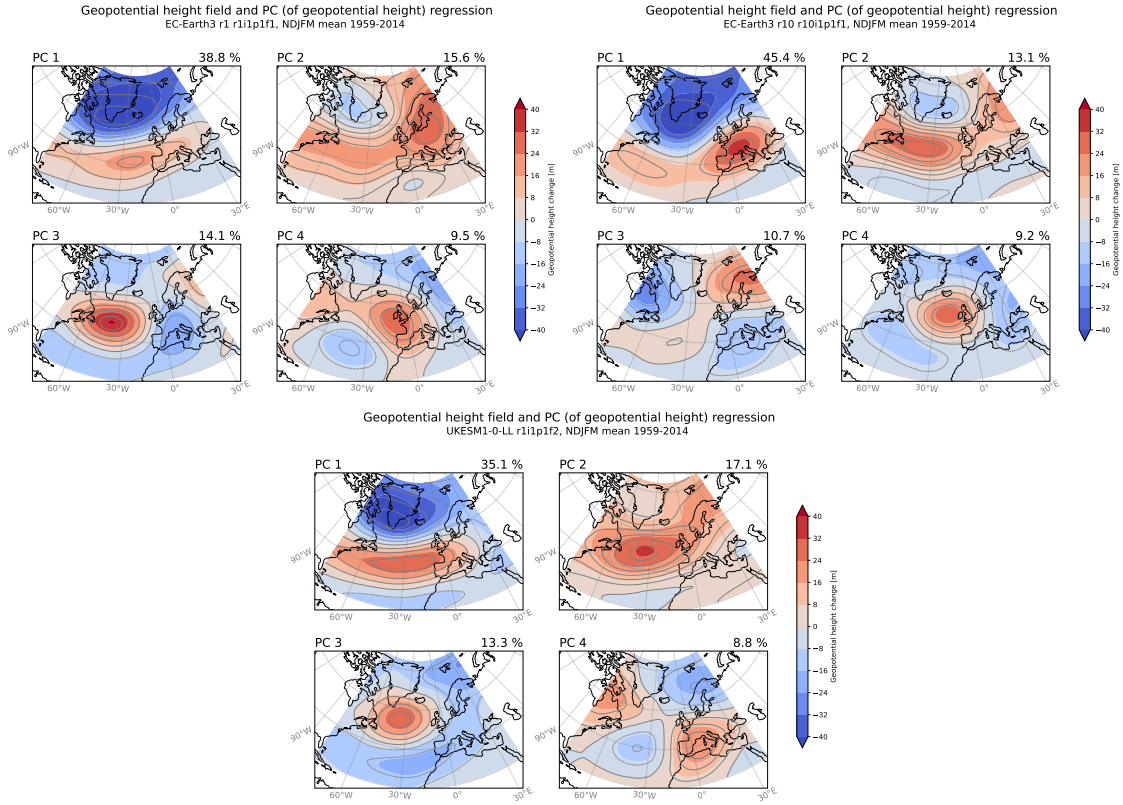


Figure 23: **Illustration of overlaps between EOF patterns**. Four leading EOFs of EC-Earth3 r1 (UL), EC-Earth3 r10 (UR) and UKESM1-0-LL (L), displayed as Z500 PCR showing the full field of regression coefficients. EOFs are in their original order (no swapping) and signs of the patterns are arbitrary.

However, despite being present in most models, the patterns are not always associated with the same EOF ranking as in ERA5. Two out of eight models show the AR pattern as EOF3 instead of EOF2, and for the Scandinavian blocking pattern, EOF swapping was necessary in five out of eight cases. This results in large differences in explained variance compared to ERA5. But more importantly, it introduces an element of uncertainty to the evaluation, as EOFs are swapped based on subjective identification of the patterns. However, a clear identification of the patterns is not always possible. This becomes evident in the case of the two EC-Earth3 members, that stand out with the least similar patterns to ERA5 in comparison both of the Atlantic Ridge and the Scandinavian blocking pattern. EOF2 of ensemble member r10 has been identified as the AR pattern, but shows distinct differences to the pattern in ERA5, by missing the belt of opposite sign around the Atlantic center of action and instead showing a second center of action of the same sign over northeastern Europe. EOF2 of ensemble member r1 was identified to resemble the Scandinavian blocking pattern most, but differs from ERA5 by showing a strongly enhanced center of action over Scandinavia, that extends over most of Europe and westward across the Atlantic. The difficulties with pattern identification are demonstrated by Figure 23, showing the leading four EOF patterns for both EC-Earth3 members and UKESM1-0-LL. They are displayed as Z500 PC regression, showing the full field instead of only locations with correlation coefficients above 0.3, to make identification of patterns easier. Note that the EOFs are displayed in their original order, and that the sign of the patterns is arbitrary. I shows that EOF2 shows the same characteristics in EC-Earth3 r1 and r10, displaying elements of both the AR and the Scandinavian blocking patterns. For r1, the Scandinavian blocking characteristics seem more dominant, with a stronger and larger center of action over northeastern Europe, while for r10, EOF2 shows more AR characteristics, with a stronger center of action over the Atlantic. Comparing the leading four EOFs, it becomes apparent that other EOFs show elements of these patterns as well. This is especially true for

ensemble member r10, where EOF3 shows a similar pattern as EOF2, but with more emphasis on the Scandinavian blocking characteristics, and EOF4 resembles the AR pattern, but with an eastward shifted Atlantic center of action and differences in the belt around it. Additionally, EOF2 could also be identified as a second NAO pattern, showing the clear characteristics of a northern center of action over Iceland and Greenland and a southern center of action over the Atlantic, which is shifted northwest compared to the ERA5 pattern. A similar dynamic of mixing between the patterns is observable in the four leading EOF patterns of UKESM1-0-LL. It received low scores in the AR comparison as well, due to its EOF3 pattern resembling the AR pattern, but underestimating the center of action over the Atlantic and overestimating the belt around. This comparison shows that its EOF2 pattern resembles the AR pattern as well, but, as in the case of EC-Earth3, missing the belt of opposite sign around the Atlantic and instead extending its center of action northeast over Scandinavia.

These observations explain the poor performance of UKESM1-0-LL and EC-Earth3 r10 in the AR comparison and of EC-Earth3 r1 in the Scandinavian blocking comparison. These models show EOF patterns that are not as well separated as in ERA5, showing elements of several modes of variability in one EOF pattern and elements of one mode of variability in several EOF patterns. This relates back to the issue of 'effective degenerate multiplets' raised in Section 3.3. The eigenvalue spectra shown in Figure 9 show that for almost all considered CMIP6 models, the second and third eigenvalues overlap with their nearest eigenvalues within their confidence intervals, indicating that the corresponding EOF patterns can not be considered independent, as they may represent arbitrary mixtures of the true populations. This is the case for the just discussed models. On the other hand, the three models with the best separation of the second eigenvalue, IPSL-CM6A-LR, MPI-ESM1-2-HR and MIROC6 are showing the greatest resemblance to the AR pattern seen in ERA5 (ref. Figure 16).

This finding raises serious issues with regards to the suitability of the EOF analysis to identify well defined modes of variability beyond the NAO. With the used methods, the ability of the models to reproduce patterns of variability similar to those seen in ERA5 seems to strongly depend on the separation of their eigenvalues. Thus, it is hard to consider the obtained patterns as well defined physical modes of variability and evaluate the models on them. It might lead to the conclusion that only the leading EOF pattern, shown to be associated with a well separated eigenvalue in all cases, can be considered well defined and associated with the physical mode of variability of the NAO and consequently be used for model evaluation, which is in line with the focus of our analysis. However, even this hypothesis can be questioned by the findings of this analysis. The observed relationship at certain time steps between the PC1 and PC2 time series in the ERA5 data (ref. Figure 6) hints at a possible connection of EOF2 to the negative phase of the NAO. This is supported by the EOF patterns found in EC-Earth3 r10 (ref. Figure 23), where both EOF1 and EOF2 show clear NAO characteristics. This is in line with the fact that from a physical point of view, it is problematic to consider the opposite phases of the NAO as the same spatial pattern only distinguished by alternating signs, as they show a distinct spatial structure. Several studies have outlined asymmetries in the spatial patterns and temperature and precipitation responses between positive and negative NAO events (Schmith et al. 2022; Luo et al. 2018; Hurrell 2015). However, the definition of the EOF analysis method requires both phases of the NAO to be combined in one EOF pattern, as they are of course highly correlated. The suitability of the EOF analysis method will be discussed in length in Section 5.2. For now, these conclusions call for care when considering the identified EOF patterns as distinct physical modes, keeping in mind the overlaps between the modes. This also applies when considering their teleconnections. It might lead to questioning the reliability of proxy-based reconstructions of NAO time series based on these teleconnection mechanisms, as the separation between different modes of variability may be less clear than assumed for this method.

5.1.2 Teleconnections

The results of regressing temperature anomalies against the NAO time series show a high ability of the models to capture the quadripolar temperature response pattern to the NAO observed in ERA5. One notable exception is the ensemble member r1 of EC-Earth3, not capturing any positive temperature response to the positive phase of the NAO. This can be attempted to be explained by biases in the mean temperature field (ref. Section 4.2.2). EC-Earth3 r1 does indeed show a large mean difference of -2.2°C, underestimating temperatures especially in the northern North Atlantic and the Arctic ocean around Greenland, but notably EC-Earth3 r10 shows the same difference pattern with larger magnitude, resulting in a higher mean difference of -3.2°C. Since EC-Earth3 r10 is able to capture the temperature response a lot better than EC-Earth3 r1, the temperature difference pattern does not seem to be the cause for the missing temperature response in EC-Earth3 r1.
A better explanation for the differences in teleconnection patterns to ERA5 seems to lie in the associated EOF patterns, displayed as contour lines in all regression plots. In the case of EC-Earth3 r1, the shortcomings of the NAO temperature regression seem to be linked to its NAO EOF pattern. Not only is EC-Earth3 r1 the model with the weakest southern center of action, the northern center of action also extends further east over Scandinavia than in the other models. Due to this weaker and less wave shaped gradient, the model might not produce the strong westward flow directed towards northwestern Europe that is usually associated with a positive state of the NAO, explaining the missing temperature response in that region. This finding can be generalised by showing that the shape of the EOF1 pattern can explain the temperature response patterns of other models as well. For instance, the northeast shift of the positive response over Europe in CNRM-ESM2-1, including the relationship over Svalbard that is not recorded in ERA5, can be linked to the models tilted northern center of action in the NAO pattern. Similarly, CNRM-ESM2-1, IPSL-CM6A-LR and MIROC6 all show a southern negative temperature response that is not extending as far east over the southern Mediterranean region as in ERA5, which can be associated to their southern center of action of the NAO not extending as far east over Europe as in ERA5. Similarly, the differences of the precipitation response patterns to ERA5 can be partly explained by the shapes of the models EOF1 patterns. Again, EC-Earth3 r1 stands out, in this case with the weakest magnitude of the precipitation response, which can be connected to the weak southern center of action in its EOF1 pattern. Similarly, the missing positive response over northern Scandinavia in CNRM-ESM2-1 is likely due to the tilted northern center of action, not directing the westerly flow as much north, and its missing response in the Mediterranean region is connected to the westward shifted southern center of action.

Connecting the EOF and teleconnection patterns of the AR pattern confirms the observation that differences in teleconnection patterns to ERA5 can largely be explained by differences in the respective EOF pattern. The temperature responses of UKESM1-0-LL and CNRM-ESM2-1 (ref. Figure 17), with a strong negative temperature response almost all around the domain, particularly over the Arctic ocean, but notably not Greenland, but barely any positive response in the center of the domain, are connected to the shortcomings of their EOF patterns (ref. Figure 16), that show a weaker center of action over the Atlantic than ERA5, but a stronger sign over the northern ocean around Svalbard. Above, these shortcomings in the AR pattern have been linked to temperature biases, due to an overestimation of sea ice in the models. This relationship can be restated here, as the shape of the AR temperature response over the northern part of the domain matches the shape of the temperature bias in both models (ref. Figure 12). The other models show AR temperature responses similar to ERA5, and differences again correspond to differences in their corresponding EOF pattern. EC-Earth3 r10 stands out with the least similar temperature response, which can be explained by its EOF pattern, that, as described above, presents a combination of the NAO, AR and Scandinavian blocking patterns due to entanglement in the eigenvectors, which makes it

not suitable for this comparison.
The explanatory power of the EOF patterns for observed teleconnections is supported by the precipitation response to the AR pattern, as the precipitation response is visibly linked to the shape of the corresponding EOF pattern for all models. The negative precipitation response to the AR pattern is located in the southeastern corner of the Atlantic center of action, and the positive temperature responses around are shifted according to the shift in the central pattern of action as well.

Similar observations have been made for the teleconnection patterns related to the Scandinavian blocking pattern, where the temperature and precipitation responses strongly correspond to the respective EOF pattern in their spatial extent and location (ref. Figures 20 and 21).

In summary, the results show that deviations in the teleconnection pattern compared to ERA5 can in most cases be linked to deviations in the models related EOF pattern, usually linked to a shift or magnitude difference in the centers of action. These findings relate to the work of Rousi et al. (2020), that investigates spatial variability of the NAO within one GCM. They define different 'NAO flavors', characterised by shifts or tilts in the centers of action compared to a typical NAO spatial pattern and find differences in the effects on European temperature and precipitation patterns depending on the specific locations of the NAO centers of action. They conclude that the NAO is not a stationary pattern, but shows considerable spatial variability that has significant implications for its temperature and precipitation teleconnections. Comparing their results to reanalysis data, they find that the investigated model shows the NAO centers of action displaced to the west, which is in line with the observations made for several of our investigated models. Rousi et al. (2020) also investigate a future period, finding an eastward displacement of the NAO pattern. These findings have implications for our results, indicating that the observed differences between the models and reanalysis data might be at least partly an expression of natural variability of the NAO. They also highlight the importance of further investigation of this spatial variability, especially under future climate change.

The analysis of teleconnection patterns of the different modes of variability leads to the conclusion that the CMIP6 models show teleconnection patterns similar to those observed in the reanalysis data. Furthermore, deviations in those patterns can be explained by deviations in the EOF patterns compared to ERA5. Thus, the models show the same temperature and precipitation responses as observed in reality, even in cases where the actual response patterns differ from the ERA5 patterns. This is an important quality check of the models, showing that they model the same responses to circulation anomalies as seen in reality.
However, the benefit of comparing teleconnection patterns of the models goes beyond a simple quality check, as it offers an evaluation of the used comparison methods as well. When comparing the results of the EOF analysis and the temperature and precipitation PC regression, in many cases a model showing high agreement with the ERA5 EOF pattern will also show high agreement with the teleconnection patterns, while a spatially shifted or otherwise deviating EOF pattern leads to shifted temperature and precipitation responses. In these cases, a high Taylor skill score in the comparison of the EOF patterns seems to be an indicator of the models ability to represent both the respective mode of variability and its teleconnections in a realistic way that is comparable to reanalysis data. However, this conclusion can be question by exceptions, for example the case of EC-Earth3 r1 and its representation of the NAO and the NAO teleconnections. Here, we observe low agreement with ERA5 for temperature and precipitation teleconnections despite a high skill score for the EOF 1 comparison. This could be attributed to a misrepresentation of the response mechanisms in the model, but the fact that another ensemble member of the same model (EC-Earth3 r10) is able to produce similar response patterns to ERA5 does not allow this conclusion. Instead, it points to the EOF1 pattern being more different from ERA5 than the metrics are suggesting. Indeed, as discussed above, EC-Earth3 r1 shows a strong underestimation of the

southern center of action in its NAO pattern, that is responsible for the deviations in temperature and precipitation response. However, the Taylor comparison metrics do not "punish" this misrepresentation sufficiently, placing the model in the Taylor diagram and skill score ranking close to others that by visual inspection resemble the ERA5 pattern more closely. This finding questions the reliability of the used evaluation metrics. It also suggests that a combined metric considering the evaluation of the EOF pattern and its teleconnections, as provided in Figure 15, may provide a more accurate evaluation of the models representation of the respective mode of variability, as the teleconnection response patterns seem to emphasize relevant shortcomings in the models EOF pattern.

In conclusion, an answer to the question how well the evaluated CMIP6 models represent different modes of variability over the North Atlantic region, compared to reanalysis data, can be attempted to be provided based on the combined skill scores for Z500, temperature and precipitation regression for the investigated three modes of variability that is provided in Figure 15. Many models show consistent results across the three modes of variability. MPI-ESM1-2-HR and IPSL-CM6A-LR show high consistency with the ERA5 data for all three investigated modes of variability, consistently placing in the top three of the skill score ranking. CESM2 and CNRM-ESM2-1 are consistently ranked slightly lower, but with small actual differences in score to the two former models, indicating a high similarity to ERA5 as well. The UKESM1-0-LL and MIROC6 members show less consistent results across the modes of variability. UKESM1-0-LL receives the highest score for the NAO comparison, but a very low score for the AR pattern that has above been attributed to its temperature bias over the northern edge of the domain, and ranks in the middle of the field for the Scandinavian blocking pattern. MIROC6 ranks middle of the field for NAO, but receives the top rank for AR, while placing second last with a low skill score of 0.3 for the Scandinavian blocking pattern, due to a strong northward shift of its associated EOF pattern. Despite both models showing similarly mixed results, due to the high importance of the NAO, UKESM1-0-LL can be considered more similar to ERA5 than MIROC6. Both EC-Earth3 members receive consistently low skill scores relative to the other models, placing in the bottom three spot of the ranking for each mode of variability. While placing close together for the NAO, in the AR comparison r10 stands out with a low score (0.2) and in the Scandinavian blocking comparison r1 receives a considerably lower score than the rest of the models (0.1). This observation provides an interesting insight to the question whether the difference between the two EC-Earth3 members due to decadal variability is larger than their difference to other models due to model characteristics. Despite the difference between the two ensemble members being large in terms of the score, their scores, especially taken together, are also separating them from the other models. This leads to an important limitation of the model ranking method, as the skill scores provide a relative ranking of the models to each other, but we are lacking an interpretation of their meaning in absolute terms. It limits our ability to provide a qualitative model evaluation. This, along with several other limitations of the comparison method, that put the just presented results into perspective, are discussed in the following section.

5.2 Discussion of methods: How can the representation of modes of variability and teleconnections be captured and compared between models?

In this section, the results of the model evaluation are put into perspective by critically evaluating the suitability of the applied methods. It questions the suitability of the EOF analysis method to define several modes of variability over one domain and the applied methods to evaluate their representation in models and quantify their differences to reference data. Limitations of the methods and possibilities for future work are outlined, especially regarding the inclusion of uncertainties in the evaluation.

5.2.1 Defining modes of variability with EOF analysis

The application of an EOF analysis to identify modes of variability in the data is connected to several limitations. First, the interpretability of the EOFs is limited by definition, as due to the mathematical construction of EOFs as uncorrelated orthogonal functions, the patterns do not necessarily correspond to physical modes (Hannachi et al. 2007; Davini et al. 2013; Lee et al. 2021). Therefore, Davini et al. (2013) suggest to apply EOF analysis only to identify the dominant pattern of variability of a region, which would be the NAO in the North Atlantic winter geopotential height field, represented by the leading EOF.

We were able to associate the leading and some of the lower order EOF patterns with known physical modes of variability, but found several problems connected to identifying modes of variability with an EOF analysis. First, the comparison of models to reanalysis data was complicated by the fact that while the investigated patterns were present in most model data, they were not represented by the same order EOFs in all models. Thus, comparing the same order EOF patterns of the models with reanalysis data can lead to misleadingly low model evaluation (Lee et al. 2019). To avoid this, EOF swapping based on a subjective visual inspection was applied to improve comparability. Lee et al. (2019) address this problem in a comparison of modes of variability in CMIP5 models by introducing two different objective decision criteria for EOF swapping. They find that the model performance is significantly improved if EOF swapping is applied, and that the model ranking is sensitive to the choice of swapping method. Interestingly, while in our analysis EOF swapping was only necessary for the non-dominant patterns of variability, Lee et al. (2019), who inspected a larger ensemble of CMIP5 historical runs, found EOF swapping increasing correspondence of the modelled to the observed patterns of variability also for the leading EOFs.

The need for EOF swapping is related to a second, more severe limitation of the EOF method. We found that the modes identified with the EOF analysis are often not well defined. This is evident by several cases where a mode is represented in several of a models EOF patterns, or inversely, where an EOF pattern mixes characteristics of several modes, as discussed in Section 5.1. This poses a severe limitation for model evaluation. It can be linked to sampling uncertainties of the EOF method and the issue of 'effectively degenerate multiplets' established by North et al. (1982), stating that if a group of true eigenvalues lie too close to each other, the EOF method can not sample them separately, and the resulting EOF patterns present a mixture of the true patterns. Indeed, we showed that low resemblance of a models EOF pattern to the spatial pattern found in ERA5 is often due to the corresponding eigenvalues being not well separated following Norths rule of thumb. This represents a major limitation of using an EOF analysis to compare the modes of variability in CMIP6 model data to reanalysis data, in particular when considering not leading modes of variability.

However, the issue of inter-linkages between the modes of variability extends beyond the issue of overlapping eigenvalues. We found even the leading mode of variability, the NAO, whose eigenvalues are well separated in all considered cases, to show relations to the following modes. This is evident by a relationship between the phases of the NAO (PC1 time series) and the AR (PC2)

time series at some time points in the ERA5 data, but also the fact that the EOF2 pattern of EC-Earth3 r10 shows spatial resemblance to the NAO pattern.
This can be linked to another shortcoming of the EOF approach, the fact that both phases of a mode of variability are captured by the same spatial pattern that alternates between a positive and a negative phase by reversing the signs of the anomalies. However, the positive and negative state of the NAO have been identified by other approaches as distinct patterns with different spatial characteristics, as demonstrated for example by the results of the k-means clustering approach shown in Figure 1 (Strommen et al. 2019; Hurrell 2015). The two NAO phases show asymmetries in their spatial patterns, characterised by an east-west shift of the centers of action and differences in amplitude and persistence (Luo et al. 2018; Hurrell 2015), in their behavior and interactions with other phenomena such as blocking (Davini et al. 2012; Luo et al. 2018) and consequently in their teleconnections (Rousi et al. 2020; Schmith et al. 2022). Thus, the attempt to capture both phases of the NAO in one pattern of variability presents a further limitation and explains why the method is unable to capture the full dynamics of the NAO.
In the context of these considerations, it needs to be considered that the wish to separate the reanalysis and model data into well defined, independent modes of variability does to a certain degree conflict with the reality of circulation variability. The ERA5 PC time series show that the state of the atmosphere at every time step is a combination of several, if not all, identified modes (ref. Figure 6). Even in winters with strong loading on one EOF pattern, like the winter season 2009/2010 with a strong negative phase of the NAO, the other PC time series show a weak loading as well. A clearer definition of the typical patterns may be obtained by looking at the state of the field in seasons that are dominated by one mode of variability, but they can not be entirely isolated. This can partly also be attributed to the high temporal aggregation to winter seasons, resulting in loss of detail.
Several alternative approaches have been suggested to complement the EOF analysis method and alleviate some of its shortcomings. One of them is the Common Basis Function (CBF) approach used by Lee et al. (2019) and others. It improves comparability of models to a reference dataset by projecting model anomalies on the EOF of the reference dataset. Thus, it solves the problem of having to flip the signs of the patterns to correspond to the reference pattern, eliminates the need for EOF swapping, and reduces the ambiguities connected to not well separated EOFs. While Lee et al. (2019) observed improved scores of the models with respect to the reference data, they also find that the conventional EOF approach, if applying EOF swapping, leads to consistent results in most cases. An extension of the here presented work could therefore be to repeat the comparison utilizing the CBF approach to evaluate the sensitivity of our results to the shortcomings of the EOF analysis method.

However, our results show that a model comparison based on the EOF patterns is still insightful despite the shortcomings of the method. Most of the here considered model data represents the key characteristics of all three evaluated modes, and EOF swapping allows to evaluate them against the ERA5 data. Only in a few cases, e.g. the EC-Earth3 members, evaluation is hindered by strong entanglement of the modes in the EOF patterns. In these cases, the low comparison scores of the models need to be seen in that context.

5.2.2 Comparing modes of variability based on EOF patterns

A further limitation relates to the assumption that comparing the spatial pattern of a mode of variability identified in model data to the pattern found in a reference dataset, even if assuming that the EOF method captures the mode well, is enough to evaluate the full dynamic representation of this mode in the model. As Davini et al. (2013) point out in the case of the NAO, an evaluation of the spatial pattern of variability alone is not sufficient to investigate the models' representation of the full dynamics of the NAO. Differences in NAO patterns can be misleading, as patterns can

differ despite an adequate representation of the variability due to differences in the mean state, as discussed above, or patterns can appear similar, but be connected to different modes of variability. Thus, more characteristics need to be evaluated, if the whole NAO dynamic is to be captured. To this end, Davini et al. (2013) suggest to evaluate dynamic phenomena related to the NAO, such as the connection between blocking and the NAO phase. They find shortcomings in models that underestimate blocking over Greenland. We complement the comparison of EOF patterns with a comparison of their temperature and precipitation response patterns and the mean state of the investigated variables. However, as pointed out above, the teleconnection patterns are largely a reflection of the EOF patterns. Future work should consider comparing further characteristics, such as blocking frequency or fluctuations in jet stream location and strength.
Similarly, evaluating the representation of modes of variability based on the spatial pattern only disregards a second important aspect of these modes: their temporal variability. The strength, frequency and persistence of a mode of variability, as well as observed trends, carries a lot of information about its representation in a model, that is complementary to the information gained by investigating its spatial pattern. In the 6th Assessment Report of the IPCC, Eyring et al. (2021) suggest evaluating models representation of modes of variability with regards to their spatial structure and magnitude, as well as temporal trends and variability. Studies investigating the temporal variability of NAO in CMIP6 models find severe shortcomings in the models ability to simulate long term variability of the NAO, as they tend to underestimate multi-decadal variability compared to interannual variability (Bracegirdle 2022; Eyring et al. 2021). Furthermore, as mentioned above, including the PC time series in the analysis would also allow an investigation of the interactions of different modes within a model, like the relationship between the negative phase of the NAO and blocking (Davini et al. 2012; Luo et al. 2018). While it would be beneficial to address this temporal component in future work, the here presented analysis should be considered as an evaluation of the spatial structure of modes of variability and their teleconnections represented in the models only. Despite its limitations, this approach offers valuable insights into the models representation of modes of variability and their teleconnections. This is in line with Rousi et al. (2020), who encourage the exploration of the spatial variability of the NAO and its impacts on European climate.

5.2.3 Comparison metrics

The analysis uses a comparison approach based on Taylor (2001), that complements visual comparison of patterns by providing an objective measure of pattern similarity based on a few simple statistical measures. The final model ranking is based on a skill score taking both spatial correlation and magnitude of the patterns compared to ERA5 into account. But models are also evaluated based on a visual inspection of their patterns and their location in the Taylor diagram, which allows for a separation of pattern and amplitude errors, as suggested by Lee et al. (2019).
Model comparison has been carried out both based on the pattern over the full domain, and based on a pattern covering only locations with a correlation coefficient between the respective time series of above 0.3. This approach is not only more robust, it also aids visual evaluation by highlighting the most significant patterns. The results show that both the full field and the high correlation pattern lead to comparable results of the model comparison, showing a very similar pattern in the Taylor diagram but with higher correlation coefficients for the full field and an almost identical model ranking. Thus, they justify the use of the robust pattern view.
In many cases, visual comparison and the Taylor metrics agree on the level of similarity of the models. Low skill scores can often be justified with obvious shortcomings in the patterns and high skill scores correspond to a clear match in the pattern to ERA5. In some cases however, the findings of the comparison metrics can not be confirmed by visual evaluation. In the case of the NAO precipitation response pattern, the model ranking can not be justified by visual inspection, as patterns look similar and all show differences to ERA5 (ref. Figure 14). Here, it is difficult to

determine whether the comparison metrics add to the evaluation by quantifying differences not visible by eye and adding objectivity to the evaluation, or if their results have to be treated with care and the visual comparison should be relied on to evaluate the relevance of differences. Similarly, the skill score ranking can be misleading, as models often receive very similar scores, leading us to question the significance of their differences. Possible approaches to take uncertainties into account and determine significance will be discussed below. When evaluating models based on their skill score it is important to consider that these scores are only relative measures of similarity, and carry information only in relation to other models scores. The analysis offers no absolute measure of quality of the models. This could be addressed partly by including more, ideally all CMIP6 models into the analysis, allowing statements with respect to the whole ensemble. In other cases, the results of the comparison metrics can be directly challenged by visual evaluation, questioning their reliability. One example of this is the NAO pattern in EC-Earth3 r1, that has been discussed in length above. The underestimation of the southern center of action can be identified as a shortcoming of the pattern by visual inspection, but the Taylor metrics do not reflect this, ranking the pattern similar to other more ERA5-like patterns in the Taylor diagram and skill score ranking. This can be explained by the fact that the pattern magnitude is measured by the standard deviation ratio, which takes the variance over the whole field into account. As the northern center of action is slightly overestimated compared to ERA5, the pattern has an overall standard deviation close to that of ERA5. This flaw can not be circumvented if using measures that are calculated over the whole domain, but it is a reminder that the metrics are simple statistical measures that can be misleading. In the case of NAO in EC-Earth3 r1, a combined skill score based on the EOF pattern and both teleconnection patterns has proven to overcome these limitations and providing a more robust measure of evaluation.

5.2.4 Sampling uncertainties and the role of decadal variability

One of the most important limitations of the presented comparison analysis is that it does not include any measures of uncertainty. However, the results are subject to sampling uncertainty due to the influence of internal climate variability on decadal time scales. The two main sources of sampling uncertainty are the short length of the analysed time period and the limited amount of model data compared, considering only one realization (or ensemble member) per model and only a small selection of models. The large influence of sampling uncertainty on the results is illustrated by the differences observed between the patterns identified in the two realizations of the EC-Earth3 model. A third source is the sampling uncertainty related to the EOF patterns that has been discussed above.

Due to the limited availability of the ERA5 record, the analysis is performed over a timeframe of 55 (model) years. This is a short time frame to obtain robust results, especially considering the role of decadal variability in the NAO (Bracegirdle 2022). Schmith et al. (2022) are emphasizing the important role of sampling uncertainty due to decadal-scale variability in observational records of 75-150 years length, which is significantly longer than the timeframe used in this analysis. They point out that the patterns obtained from an analysis of this time frame are always subject to sampling uncertainty and only approximately equal to the true pattern. This limitation could easily be addressed for the CMIP6 data by extending the studied time frame, as historical runs are available for the time period of 1850-2014. However, the model comparison would especially benefit from an uncertainty range on the ERA5 patterns that the models are compared against. As the ERA5 backward extension is only available in its final form from 1959 on, such a measure of robustness would need to be obtained another way. Similar to the approach taken by Schmith et al. (2022), Monte Carlo methods could help to determine robust versions of the ERA5 EOF patterns connected with a probability distribution. To obtain these, an EOF analysis could be applied to a large number of subsets sampled from the data, for example consisting of different sets of 25 winter seasons, resulting in a large set of EOF patterns. This would allow a comparison

of models against a robust pattern representing the natural climate with larger confidence and an evaluation of the significance of the differences. Thus, uncertainty ranges could be added to the Taylor comparison metrics.

The second source of sampling uncertainty is connected to internal variability present in the model data. With one exception we are only considering one realization of each model. Thus, we are not able to identify what fraction of the differences observed between the models is due to actual differences in model behaviour and what fraction is due to internal variability. This limitation could be alleviated by using a longer time frame of model data, but pre-analyses have shown that the comparability with reanalysis data decreases if using different time scales. It could also be addressed, however, by considering an ensemble of realizations of each model, where each member differs only in the initial conditions the run is based on. This would allow to calculate an ensemble mean and uncertainty range for the comparison metrics and thus enable a robust comparison of different models representations of patterns of variability and teleconnections. A first insight into the relationship between intra- and inter-model variability is given by the two realizations of EC-Earth3 used in this analysis. The two members show the same patterns of mean model biases, but of different magnitude, confirming that they are based on the same mechanisms. They show considerable differences in some of their EOF and teleconnection patterns, pointing at the large role of decadal variability. In the final robust evaluation, based on the mean skill score of all three regression analyses combined for each mode of variability, their skill scores are consequently quite different, considering the range of scores obtained by all models (NAO: 0.57 and 0.61, AR: 0.44 and 0.21, Scand: 0.12 and 0.32 for r1 and r10, respectively; ref. Figure 15). However, in relation to the other models both members are ranked close to each other, suggesting that the differences between models are larger than the internal variability of the models. Lee et al. (2019) make a similar observation in their comparison of different modes of variability in CMIP5 models, finding that despite the influence of internal variability and the limited length of the measurement record the analysis is relatively insensitive to the choice of model realization compared, allowing a separation of the different models. However, this question can only conclusively be answered by evaluating a larger sample of CMIP6 model data.

6 Conclusion and Outlook

We have succeeded to identify a set of methods that allow to: 1. identify different modes of circulation variability in the winter geopotential height field over the North Atlantic region, 2. determine the impacts of those modes on surface temperature and precipitation variability in the domain, 3. compare the spatial patterns of both found in different CMIP6 models against reanalysis data bot qualitatively and quantitatively and 4. explore differences between the models that can explain their observed behavior in terms of circulation variability. Thus, we are able to answer both the questions of *how different modes of winter North Atlantic circulation variability and their temperature and precipitation teleconnections are represented in CMIP6 models*, and *what methods are suitable to identify these patterns and compare them against a reference dataset.*

To answer the first research question, we found that almost all investigated models show the central characteristics of the spatial patterns of the three modes of variability identified in the reanalysis data. This agreement is largest for the dominant mode of variability, the NAO. For the two other modes of variability, the Atlantic Ridge and Scandinavian Blocking pattern, the differences between models and reanalysis data are larger. In some cases, most notably the two investigated realizations of the EC-Earth3 model, the identified patterns are not well defined and sometimes not even distinct enough from each other to be clearly associated with one of the physical modes of variability at all. This mixing between the modes can be attributed to sampling uncertainties in the used EOF analysis method. Large parts of the pattern differences between model and reanalysis data however are limited to smaller shifts, tilts or changes in magnitude of the centers of action. These can partly be explained by differences in the mean state of the investigated variables in the models compared to reanalysis data, for example temperature biases that have been associated with different representations of sea ice. These differences give an insight in potential changes of the NAO under future climate change. Whether the remaining differences represent natural spatial variability of the NAO or biases in the model representation of circulation variability remains to be investigated. The comparison of two different realizations of one of the models hints at a large role of decadal variability. The investigation of teleconnections to surface temperature and precipitation provided a valuable addition to the model evaluation. Overall, the models capture the spatial patterns of temperature and precipitation response similar to what is observed in the reanalysis data, with the agreement again largest for the NAO. Deviations from the patterns seen in ERA5 can largely be linked to shifts, tilts, differences in magnitude or larger differences in the models corresponding EOF pattern. Thus, the investigated models pass the quality check of representing the response mechanisms in a realistic manner. Further, the investigation of teleconnection patterns provides an added value to the model evaluation by identifying relevant shortcomings in the EOF patterns that are not recognised by the comparison methods but lead to large differences in teleconnection patterns. The results provide insights in the relative performance of different models, as well as the role of natural spatial variability of the modes of variability and its impact on teleconnections. The large spatial variability and the observed overlaps between the modes of variability hint at an important limitation for proxy-based reconstructions of time series of the NAO and other modes of variability. Our results question the spatial stationarity of the NAO and its teleconnections on decadal time scales, as well as the degree of separation between different modes of variability that can be achieved.

With regards to the second research question, we have identified methods that are able to capture the spatial patterns of different modes of variability and their teleconnections, as well as compare and rank a large set of model simulations against a reference data set. However, we have also identified important limitations of the methods and potential for future work. The characterisation of modes of variability in the models is limited by shortcomings of the EOF analysis method, that is not always able to separate modes or account for the asymmetries in the two phases of the NAO.

Therefore, the analysis would benefit from complementing the EOF analysis with other methods such as cluster analysis or the Common Basis Functions approach, to evaluate the sensitivity of the results to the shortcomings of the EOF method and explore relationships between the different modes. Additionally, the representation of modes of variability in different models can not be evaluated via their spatial patterns only. The model evaluation could be made more robust and further insights into the behavior of the modes could be gained by evaluating other aspects of the variability as well, such as temporal variability and interactions with other atmospheric phenomena. The here presented evaluation of the representation of spatial patterns of variability and teleconnections in different models provides important insights already, but is severely limited by its lack of measures of uncertainty. Due to this, we are not able to determine what fraction of the observed differences is an expressions of natural variability of the NAO and other modes, and what actually points at differences between the models. To answer this question and allow for a true model evaluation, sampling uncertainties need to be taken into account by estimating a probability distribution of the patterns observed in the reanalysis data and by evaluating intra-model variability by adding several representations of each analysed model. Similarly, the model evaluation would to be strengthened by including of a larger set of CMIP6 models, which would allow the development of more meaningful and robust evaluation metrics. Finally, valuable information on the spread of natural variability and potential changes under future climate change could be gained by investigating the spatial patterns of variability and their temperature and precipitation responses under extreme conditions. This can be done by repeating the same analysis for months or seasons characterized by extreme conditions in a selected region. This would provide valuable information on atmospheric variability and the stability of teleconnection patterns under extreme conditions, which is relevant in the context of proxy-based reconstructions, which due to the nature of the proxy archives rely strongly on records of extreme events. Additionally, it is relevant for improving our understanding of the relationships between modes of variability and extreme events, as well as expected changes of variability and teleconnections under future climate change.

References

Bell, B., H. Hersbach, A. Simmons, P. Berrisford, P. Dahlgren, A. Horányi, J. Muñoz-Sabater, J. Nicolas, R. Radu, D. Schepers, C. Soci, S. Villaume, J.-R. Bidlot, L. Haimberger, J. Woollen, C. Buontempo, and J.-N. Thépaut (2021). "The ERA5 global reanalysis: Preliminary extension to 1950". In: *Quarterly Journal of the Royal Meteorological Society* 147.741, pp. 4186–4227. DOI: 10.1002/qj.4174.

Beniston, M. (2019). "Modulation of extreme temperatures in Europe under extreme values of the North Atlantic Oscillation Index." In: *Annals of the New York Academy of Sciences* 1436.1, pp. 174–183. DOI: 10.1111/nyas.13636.

Boucher, O., S. Denvil, G. Levavasseur, A. Cozic, A. Caubel, M.-A. Foujols, Y. Meurdesoif, P. Cadule, M. Devilliers, J. Ghattas, N. Lebas, T. Lurton, L. Mellul, I. Musat, J. Mignot, and F. Cheruy (2022). *IPCC DDC: IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP historical*. URL: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=C6CMIPICLhi.

Bracegirdle, T. J. (2022). "Early-to-Late Winter 20th Century North Atlantic Multidecadal Atmospheric Variability in Observations, CMIP5 and CMIP6". In: *Geophysical Research Letters* 49.11, e2022GL098212. DOI: 10.1029/2022GL098212.

Buehler, T., C. C. Raible, and T. F. Stocker (2011). "The relationship of winter season North Atlantic blocking frequencies to extreme cold or dry spells in the ERA-40". In: *Tellus A: Dynamic Meteorology and Oceanography* 63.2, pp. 174–187. DOI: 10.1111/j.1600-0870.2010.00492.x.

Burt, T. P. and N. J. K. Howden (2013). "North Atlantic Oscillation amplifies orographic precipitation and river flow in upland Britain". In: *Water Resources Research* 49.6, pp. 3504–3515. DOI: 10.1002/wrcr.20297.

Cattiaux, J., R. Vautard, C. Cassou, P. Yiou, V. Masson-Delmotte, and F. Codron (2010). "Winter 2010 in Europe: A cold extreme in a warming climate". In: *Geophysical Research Letters* 37.20. DOI: 10.1029/2010GL044613.

Christensen, J. H. (n.d.). "GreenPlanning – depends on reliable seasonal to decadal climate predictions".

Cook, E. R. (2003). "Multi-Proxy Reconstructions of the North Atlantic Oscillation (NAO) Index: A Critical Review and a New Well-Verified Winter NAO Index Reconstruction Back to AD 1400". In: *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*. American Geophysical Union (AGU), pp. 63–79. DOI: 10.1029/134GM04.

Cropper, T., E. Hanna, M. A. Valente, and T. Jónsson (2015). "A daily Azores–Iceland North Atlantic Oscillation index back to 1850". In: *Geoscience Data Journal* 2.1, pp. 12–24. DOI: 10.1002/gdj3.23.

Danabasoglu, G. (2022). *IPCC DDC: NCAR CESM2 model output prepared for CMIP6 CMIP historical*. URL: https://www.wdc-climate.de/ui/entry?acronym=C6CMNRCES2hi.

Davini, P. and C. Cagnazzo (2013). "On the misinterpretation of the North Atlantic Oscillation in CMIP5 models". In: *Climate Dynamics* 43. DOI: 10.1007/s00382-013-1970-y.

Davini, P., C. Cagnazzo, R. Neale, and J. Tribbia (2012). "Coupling between Greenland blocking and the North Atlantic Oscillation pattern". In: *Geophysical Research Letters* 39.14. DOI: 10.1029/2012GL052315.

Dawson, A. (2016). "eofs: A Library for EOF Analysis of Meteorological, Oceanographic, and Climate Data". In: *Journal of Open Research Software* 4.1. DOI: http://doi.org/10.5334/jors.122.

— (n.d.). *eofs User Guide*. Version v1.4.0. URL: https://ajdawson.github.io/eofs/latest/userguide/method.html.

Delgado-Torres, C., D. Verfaillie, E. Mohino, and M. G. Donat (2022). "Representation and Annual to Decadal Predictability of Euro-Atlantic Weather Regimes in the CMIP6 Version of the EC-

Earth Coupled Climate Model". In: *Journal of Geophysical Research: Atmospheres* 127.14, e2022JD036673. DOI: 10.1029/2022JD036673.

Döscher, R., M. Acosta, A. Alessandri, P. Anthoni, T. Arsouze, T. Bergman, R. Bernardello, S. Boussetta, L.-P. Caron, G. Carver, M. Castrillo, F. Catalano, I. Cvijanovic, P. Davini, E. Dekker, F. J. Doblas-Reyes, D. Docquier, P. Echevarria, U. Fladrich, R. Fuentes-Franco, M. Gröger, J. v. Hardenberg, J. Hieronymus, M. P. Karami, J.-P. Keskinen, T. Koenigk, R. Makkonen, F. Massonnet, M. Ménégoz, P. A. Miller, E. Moreno-Chamarro, L. Nieradzik, T. van Noije, P. Nolan, D. O'Donnell, P. Ollinaho, G. van den Oord, P. Ortega, O. T. Prims, A. Ramos, T. Reerink, C. Rousset, Y. Ruprich-Robert, P. Le Sager, T. Schmith, R. Schrödner, F. Serva, V. Sicardi, M. Sloth Madsen, B. Smith, T. Tian, E. Tourigny, P. Uotila, M. Vancoppenolle, S. Wang, D. Wårlind, U. Willén, K. Wyser, S. Yang, X. Yepes-Arbós, and Q. Zhang (2022). "The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6". In: *Geoscientific Model Development* 15.7, pp. 2973–3020. DOI: 10.5194/gmd-15-2973-2022.

EC-Earth Consortium (2022). *IPCC DDC: EC-Earth-Consortium EC-Earth-3-CC model output prepared for CMIP6 CMIP historical*. URL: https://www.wdc-climate.de/ui/entry?acronym=C6CMEEEEChi.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016). "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5, pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016.

Eyring, V., N. Gillett, K. A. Rao, R. Barimalala, M. B. Parrillo, N. Bellouin, C. Cassou, P. Durack, Y. Kosaka, S. McGregor, S. Min, O. Morgenstern, and Y. Sun (2021). "Human Influence on the Climate System". 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*. Ed. by V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 3, pp. 423–552. DOI: 10.1017/9781009157896.005.

Fasullo, J. T., A. S. Phillips, and C. Deser (2020). "Evaluation of Leading Modes of Climate Variability in the CMIP Archives." In: *Journal of Climate* 33.13, pp. 5527–5545.

Feldstein, S. and C. Franzke (2017). "Atmospheric Teleconnection Patterns". In: pp. 54–104. ISBN: 9781107118140. DOI: 10.1017/9781316339251.004.

Ferranti, L., S. Corti, and M. Janousek (2015). "Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector". In: *Quarterly Journal of the Royal Meteorological Society* 141.688, pp. 916–924. DOI: 10.1002/qj.2411.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007). "Empirical orthogonal functions and related techniques in atmospheric science: A review". In: *International Journal of Climatology* 27 (9), pp. 1119–1152. DOI: 10.1002/joc.1499.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: 10.1002/qj.3803.

Hirota, N., Y. N. Takayabu, M. Watanabe, and M. Kimoto (2011). "Precipitation Reproducibility over Tropical Oceans and Its Relationship to the Double ITCZ Problem in CMIP3 and MIROC5 Climate Models". In: *Journal of Climate* 24.18, pp. 4859–4873. DOI: 10.1175/2011JCLI4156.1.

Hu, Z.-Z. and Z. Wu (2004). "The intensification and shift of the annual North Atlantic Oscillation in a global warming scenario simulation". In: *Tellus A: Dynamic Meteorology and Oceanography* 56.2, pp. 112–124. DOI: 10.3402/tellusa.v56i2.14403.

Hurrell, J. W. (1995). "Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation". In: *Science* 269, pp. 676–679.

Hurrell, J. W., Y. Kushnir, G. Ottersen, and M. Visbeck (2003). "An Overview of the North Atlantic Oscillation". In: *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*. Ed. by J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. Washington: American Geophysical Union, pp. 1–35.

Hurrell, J. (2015). "CLIMATE AND CLIMATE CHANGE — Climate Variability: North Atlantic and Arctic Oscillation". In: *Encyclopedia of Atmospheric Sciences (Second Edition)*. Ed. by G. R. North, J. Pyle, and F. Zhang. Second Edition. Oxford: Academic Press, pp. 47–60. DOI: 10.1016/B978-0-12-382225-3.00109-2.

IPCC (2021). "Annex IV: Modes of Variability". 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*. Ed. by V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 2153–2192. DOI: 10.1017/9781009157896.018.

Jung, T., M. Hilmer, E. Ruprecht, S. Kleppek, S. K. Gulev, and O. Zolina (2003). "Characteristics of the Recent Eastward Shift of Interannual NAO Variability". In: *Journal of Climate* 16.20, pp. 3371–3382. DOI: 10.1175/1520-0442(2003)016<3371:COTRES>2.0.CO;2.

Jungclaus, J., M. Bittner, K.-H. Wieners, F. Wachsmann, M. Schupfner, S. Legutke, M. Giorgetta, C. Reick, V. Gayler, H. Haak, P. de Vrese, T. Raddatz, M. Esch, T. Mauritsen, J.-S. von Storch, J. Behrens, V. Brovkin, M. Claussen, T. Crueger, I. Fast, S. Fiedler, S. Hagemann, C. Hohenegger, T. Jahns, S. Kloster, S. Kinne, G. Lasslop, L. Kornblueh, J. Marotzke, D. Matei, K. Meraner, U. Mikolajewicz, K. Modali, W. Müller, J. Nabel, D. Notz, K. Peters-von Gehlen, R. Pincus, H. Pohlmann, J. Pongratz, S. Rast, H. Schmidt, R. Schnur, U. Schulzweida, K. Six, B. Stevens, A. Voigt, and E. Roeckner (2022). *IPCC DDC: MPI-M MPI-ESM1.2-HR model output prepared for CMIP6 CMIP historical*. URL: https://www.wdc-climate.de/ui/entry?acronym=C6CMMXME2hi.

Kjellström, E., P. Thejll, M. Rummukainen, J. H. Christensen, F. Boberg, O. B. Christensen, and C. F. Maule (2013). "Emerging regional climate change signals for Europe under varying large-scale circulation conditions". In: *Climate Research* 56.2, pp. 103–119. DOI: 10.3354/cr01146.

Lee, J., K. R. Sperber, P. J. Gleckler, C. J. Bonfils, and K. E. Taylor (2019). "Quantifying the agreement between observed and simulated extratropical modes of interannual variability". In: *Climate Dynamics* 52, pp. 4057–4089. DOI: 10.1007/s00382-018-4355-4.

Lee, J., K. R. Sperber, P. J. Gleckler, K. E. Taylor, and C. J. W. Bonfils (2021). "Benchmarking Performance Changes in the Simulation of Extratropical Modes of Variability across CMIP Generations." In: *Journal of Climate* 34.17, pp. 6945–6969.

Luo, D., X. Chen, and S. Feldstein (2018). "Linear and Nonlinear Dynamics of North Atlantic Oscillations: A New Thinking of Symmetry Breaking". In: *Journal of the Atmospheric Sciences* 75. DOI: 10.1175/JAS-D-17-0274.1.

Magnusson, L., C. Prudhomme, F. Di Giuseppe, C. Di Napoli, and F. Pappenberger (2022). "Chapter 2 - Operational multiscale predictions of hazardous events". In: *Extreme Weather Forecasting*. Ed. by M. Astitha and E. Nikolopoulos. Elsevier, pp. 87–129. DOI: 10.1016/B978-0-12-820124-4.00008-6.

Michelangeli, P.-A. and R. Vautard (1995). "Weather regimes: Recurrence and quasi stationarity." In: *Journal of the Atmospheric Sciences* 52.8, p. 1237.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng (1982). "Sampling Errors in the Estimation of Empirical Orthogonal Functions". In: *Monthly Weather Review* 110.7, pp. 699–706. DOI: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.

Peterson, K. A., J. Lu, and R. J. Greatbatch (2003). "Evidence of nonlinear dynamics in the eastward shift of the NAO". In: *Geophysical Research Letters* 30.2. DOI: 10.1029/2002GL015585.

Pinto, J. G. and C. C. Raible (2012). "Past and recent changes in the North Atlantic oscillation". In: *WIREs Climate Change* 3.1, pp. 79–90. DOI: 10.1002/wcc.150.

Portis, D. H., J. E. Walsh, M. E. Hamly, and P. J. Lamb (2001). "Seasonality of the North Atlantic Oscillation". In: *Journal of Climate* 14, pp. 2069–2078.

Rousi, E., H. Rust, U. Ulbrich, and C. Anagnostopoulou (2020). "Implications of Winter NAO Flavors on Present and Future European Climate". In: *Climate* 8.1. DOI: 10.3390/cli8010013.

Ruggieri, P., M. C. Alvarez-Castro, P. Athanasiadis, A. Bellucci, S. Materia, and S. Gualdi (2020). "North Atlantic Circulation Regimes and Heat Transport by Synoptic Eddies". In: *Journal of Climate* 33.11, pp. 4769–4785. DOI: 10.1175/JCLI-D-19-0498.1.

Schmith, T., S. M. Olsen, S. Yang, and J. H. Christensen (2022). "Asymmetries in Circulation Anomalies Related to the Phases of the North Atlantic Oscillation on Synoptic Time Scales". In: *Geophysical Research Letters* 49.11. DOI: 10.1029/2022GL098149.

Schulzweida, U. (2022). *CDO User Guide*. Version 2.1.0. DOI: 10.5281/zenodo.7112925.

Seferian, R. (2022). *IPCC DDC: CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP historical*. URL: https://www.wdc-climate.de/ui/entry?acronym=C6CMCECE1hi.

Seneviratne, S., X. Zhang, M. Adnan, W. Badi, C. Dereczynski, A. Di Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, F. Otto, I. Pinto, M. Satoh, S. Vicente-Serrano, M. Wehner, and B. Zhou (2021). "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*. Ed. by V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 1513–1766. DOI: 10.1017/9781009157896.013.

Smith, D. M., A. A. Scaife, R. Eade, P. J. Athanasiadis, A. Bellucci, I. Bethke, R. A. F. Bilbao, L. F. Borchert, L.-P. Caron, F. Counillon, G. Danabasoglu, T. L. Delworth, F. Doblas-Reyes, N. J. Dunstone, V. Estella-Perez, S. Flavoni, L. Hermanson, N. S. Keenlyside, V. V. Kharin, M. Kimoto, W. J. Merryfield, J. Mignot, T. Mochizuki, K. Modali, P.-A. Monerie, W. A. Müller, D. Nicolì, P. Ortega, K. Pankatz, H. Pohlmann, J. Robson, P. Ruggieri, R. Sospedra-Alfonso, D. Swingedouw, Y. Wang, S. Wild, S. G. Yeager, X. Yang, and L. Zhang (2020). "North Atlantic climate far more predictable than models imply". In: *Nature* 583, pp. 796–800. DOI: 10.1038/s41586-020-2525-0.

Stephenson, D. B., H. Wanner, S. Brönnimann, and J. Luterbacher (2003). "The History of Scientific Research on the North Atlantic Oscillation". In: *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*. American Geophysical Union (AGU), pp. 37–50. DOI: 10.1029/134GM02.

Strommen, K., I. Mavilia, S. Corti, M. Matsueda, P. Davini, J. von Hardenberg, P.-L. Vidale, and R. Mizuta (2019). "The Sensitivity of Euro-Atlantic Regimes to Model Horizontal Resolution". In: *Geophysical Research Letters* 46.13, pp. 7810–7818. DOI: 10.1029/2019GL082843.

Tang, Y., S. Rumbold, R. Ellis, D. Kelley, J. Mulcahy, A. Sellar, J. Walton, and C. Jones (2022). *IPCC DDC: MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP historical*. URL: https://www.wdc-climate.de/ui/entry?acronym=C6CMMOU0hi.

Tatebe, H. and M. Watanabe (2022). *IPCC DDC: MIROC MIROC6 model output prepared for CMIP6 CMIP historical*. URL: http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=C6CMMIMIhi.

Taylor, K. E., M. Juckes, V. Balaji, L. Cinquini, S. Denvil, P. J. Durack, M. Elkington, E. Guilyardi, S. Kharin, M. Lautenschlager, B. Lawrence, D. Nadeau, and M. Stockhause (2018). *CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's*. Version 6.2.7. URL: https://goo.gl/v1drZl.

Taylor, K. E. (2001). "Summarizing multiple aspects of model performance in a single diagram". In: *Journal of Geophysical Research: Atmospheres* 106.D7, pp. 7183–7192. DOI: 10.1029/2000JD900719.

Ulbrich, U. and M. Christoph (1999). "A shift of the NAO and increasing storm track activity over Europe due to anthropogenic greenhouse gas forcing". In: *Climate Dynamics* 15.7, pp. 551–559. DOI: 10.1007/s003820050299.

Uvo, C. B. (2003). "Analysis and regionalization of northern European winter precipitation based on its relationship with the North Atlantic oscillation". In: *International Journal of Climatology* 23, pp. 1185–1194. DOI: 10.1002/joc.930.

WCRP-CMIP (n.d.). *CMIP6 CVs - Controlled Vocabularies (CVs) for use in CMIP6*. Accessed: 2022-12-08. URL: https://wcrp-cmip.github.io/CMIP6_CVs/.

Wilks, D. S. (2005). "Statistical Methods in the Atmospheric Sciences: An Introduction". In: Burlington: Elsevier Science & Technology. Chap. 11, pp. 463–508.

Woollings, T., B. Hoskins, M. Blackburn, and P. Berrisford (2008). "A New Rossby Wave–Breaking Interpretation of the North Atlantic Oscillation." In: *Journal of the Atmospheric Sciences* 65.2, pp. 609–626.

A Appendix

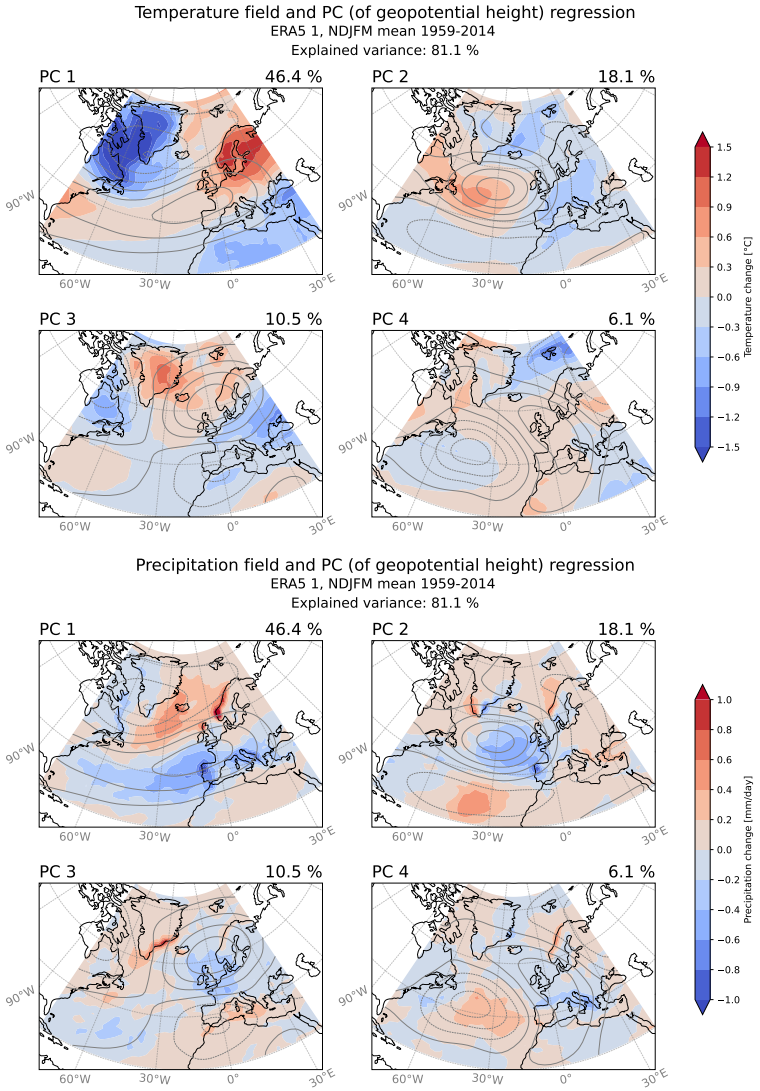


Figure A.1: Full field of ERA5 temperature (left) and precipitation (right) regression against the PC1 (NAO) time series. Colors indicate linear regression coefficient and grey contours show ERA5 Z500 EOF patterns.

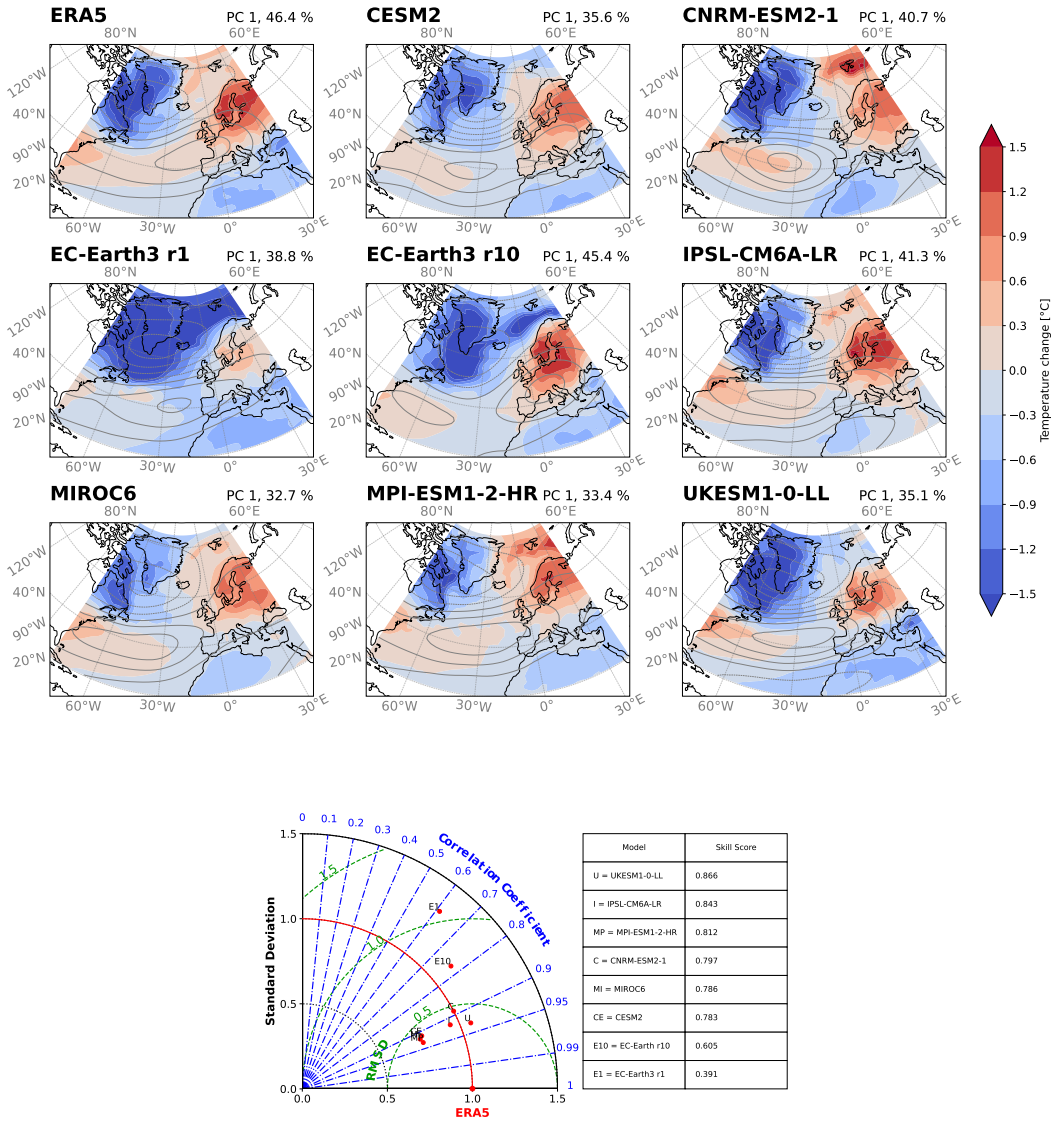


Figure A.2: **Comparison of NAO EOF pattern in all models against ERA5.** Equivalent to Figure 10, but for the full field of regression coefficients.

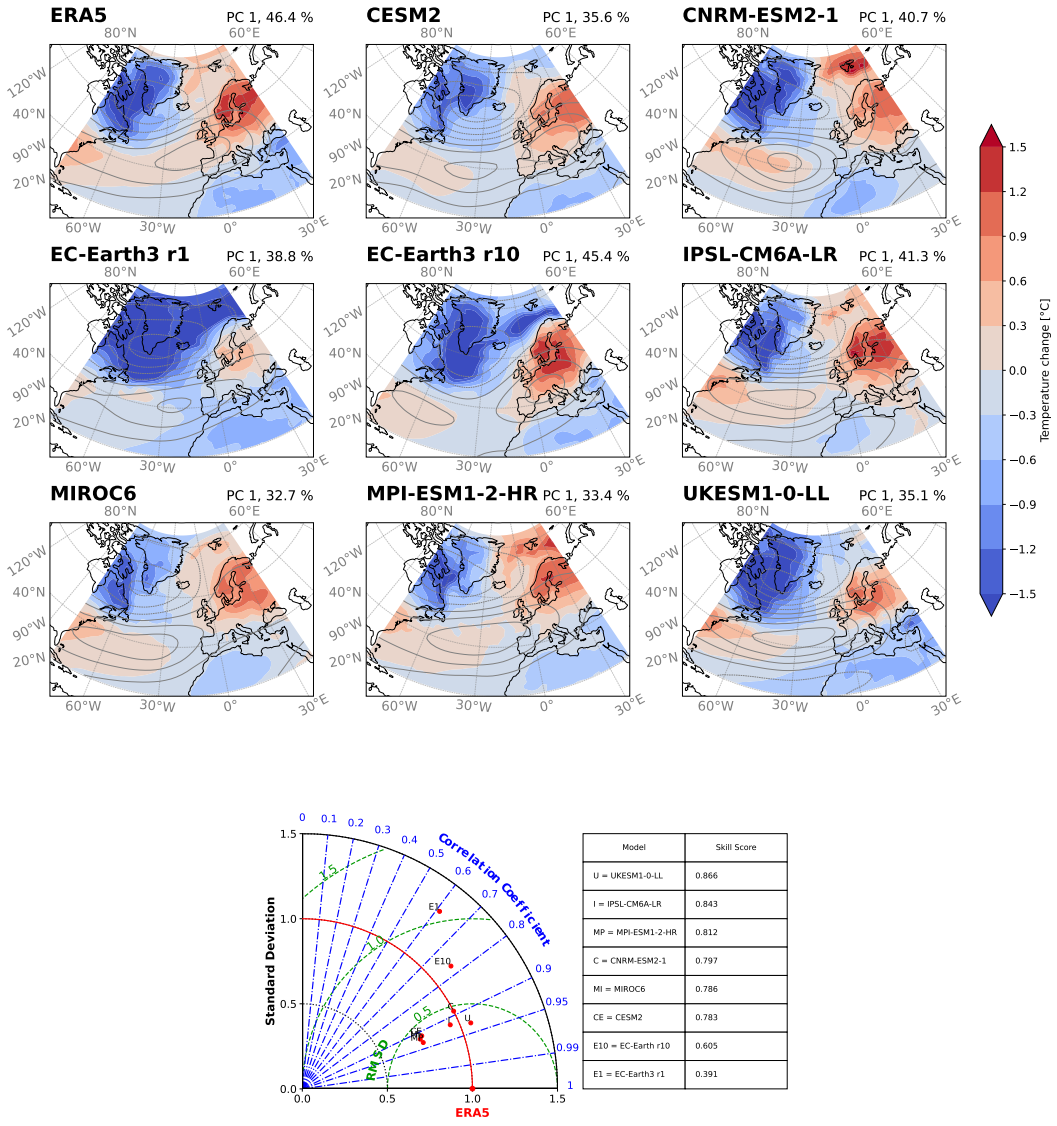


Figure A.3: **Comparison of NAO temperature response in all models against ERA5.** Equivalent to Figure 13, but for the full field of regression coefficients.

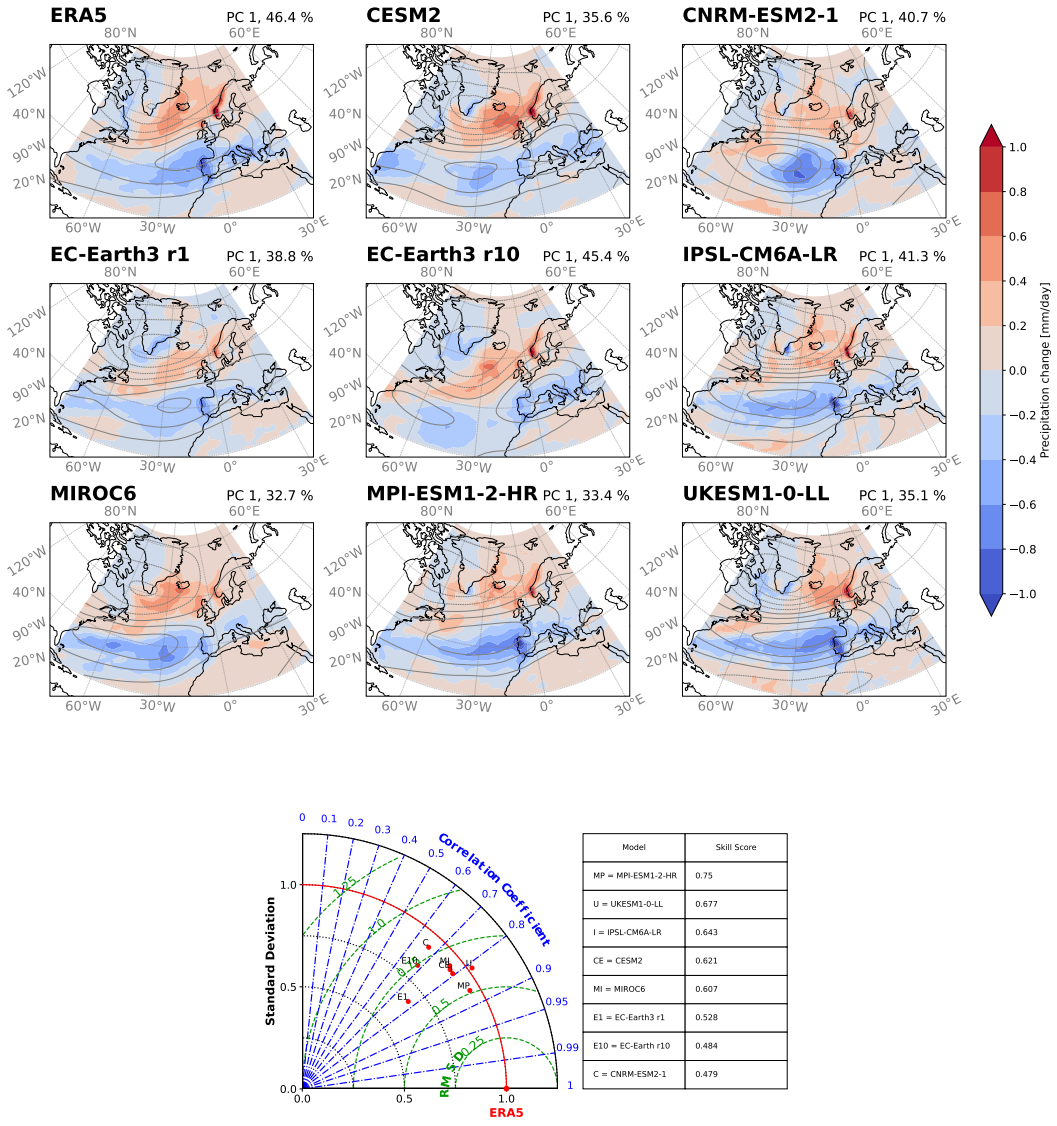
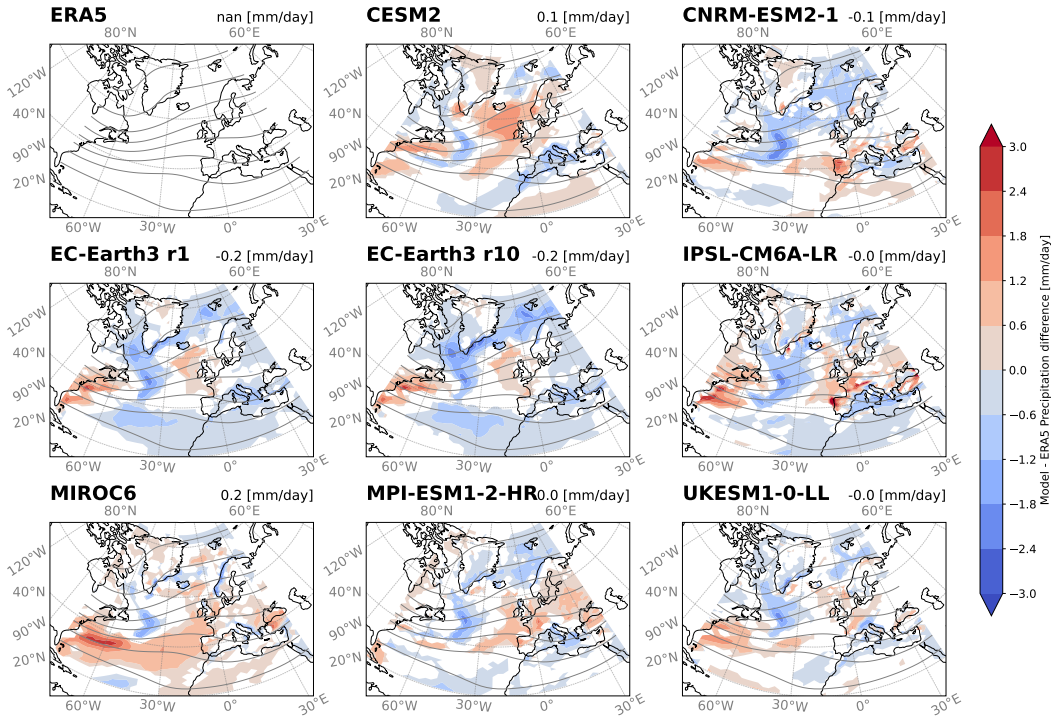


Figure A.4: **Comparison of NAO precipitation response in all models against ERA5.** Equivalent to Figure 14, but for the full field of regression coefficients.



Model	Difference [mm/day]
UKESM1-0-LL	-0.01
MPI-ESM1-2-HR	-0.02
IPSL-CM6A-LR	-0.05
CESM2	0.07
CNRM-ESM2-1	-0.08
EC-Earth3 r1	-0.16
MIROC6	0.2
EC-Earth3 r10	-0.21

Figure A.5: Spatial pattern (upper) and mean difference ranking (lower) for the difference between model and ERA5 mean precipitation field. Equivalent to Figure 12 (upper) and Figure 11 (lower) for precipitation.