UNIVERSITY OF READING

Identifying Signals of Atmospheric Circulation Regime Variability

Swinda Klaasje Jantine Falkena

A thesis submitted for the degree of Doctor of Philosophy

Department of Mathematics and Statistics

August 2022

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Swinda Klaasje Jantine Falkena

Acknowledgements

First of all I would like to thank Ted Shepherd, my main supervisor, for his guidance and support. Despite the difficulties faced during the covid pandemic, he always found time to look at my work or have a quick chat on a problem at hand. I always came out of our meetings full of new ideas for the possible directions of my research.

Also, I would like to thank my co-supervisors Antje Weisheimer and Jana de Wiljes for their contributions towards my research. I enjoyed the meetings the four of us had together and am grateful for your ideas and realism with regard to my work. Thank you for having me present at both ECMWF and Oxford, Antje, and Jana for welcoming me in Potsdam for some interesting discussions.

I would like to thank everyone in Ted's group throughout the past years for the interesting presentations and good discussions in the group meetings. It was great to learn about the work of everyone else and have a chat before or after.

A massive thank you to the MPE CDT, staff and students, for making me feel welcome in Reading and providing support and chats before, during and after the pandemic. Thanks for the opportunities to learn new skills and the company in the office and online. A special thank you to Elena for joining me on those early morning runs and welcoming me in Cambridge after her move.

Thank you to all the people who made me feel at home in Reading. Thanks to everyone at the Reading Roadrunners, especially Sahan, Ed and Chloe, for welcoming me and making me fall in love with cross country. Thanks to John for keeping me fit and sane throughout the pandemic and after, and to Richa for helping me out with all my niggles and injuries while having good conversations. Thanks to the people at Wokingham and Emmbrook FC and Wargrave Women and Girls FC for welcoming me in their teams and dragging me out of the university bubble.

A big thank you to my friends in the Netherlands, especially Monique, Elsa, Jennifer, Manon and Margriet, for staying in touch, hearing me out on the good, the bad and the ugly of the UK and supporting me throughout the pandemic.

A massive thanks to my parents for always supporting me, bringing me down to earth when I

needed it and listening to all the insignificant and not-so-insignificant stories about life in Reading. Thanks to my dad, Gerhard, for offering practical solutions and advice in every situation and to my mom, Rineke, for being a great listener and putting things in perspective.

Lastly, thank you Sibrand! Thank you for moving to the UK with me, simply because I had the crazy idea to pursue a PhD here. Thank you for all the runs and rides, and exploring this country together. Thank you for your love and support in every situation, for listening, for dealing with me being me and not running away, for just being there to give me a hug and for everything I cannot put into words.

The end is never easy, but sometimes life throws something at you that makes it nearly impossible. Oege (†) and Klaske, all my love and support. Sibrand, I always will be there; to listen, to give a hug and to support you.

Thank you mom and dad for thinking along with us and being so flexible. Thank you Monique for sharing all those photo's of Emily to cheer us up. Thank you John and Sahan for helping out with the practical things and being amazing friends in the final stressful weeks in the UK. Without all of you I do not think I would have made it to the end ♡.

Abstract

To describe the dynamics of the atmospheric circulation variability often the circulation is divided into a limited number of so-called atmospheric circulation regimes, which characterise the low-frequency variability in the dynamics. These recurrent and persistent circulation patterns can be identified in different regional domains and time periods, where the focus in this thesis is on the wintertime Euro-Atlantic sector. A central challenge here is to accurately capture the regime variability signal. To this end several novel methods are introduced and analysed in this study.

Existing methods mostly identify four circulation regimes over the Euro-Atlantic sector. The common approach to regime identification is to apply a k-means clustering algorithm to principal component (PC) data, often after applying a low-pass time filter. We find that using gridpoint data instead of PC data gives an optimal number of regimes of six instead of four. Furthermore, a time-regularised clustering algorithm is proposed to identify the persistent regime dynamics. This regularised approach increases the persistence of the regimes compared to a standard k-means clustering algorithm, with the dynamics being less affected by noise. The use of a low-pass filter leads to a bias in the regime frequencies, while the regularised method does not.

To study non-stationary regime dynamics signals on (sub-)seasonal and interannual timescales an ensemble-regularised k-means clustering algorithm is proposed. This approach couples the information within an ensemble of model hindcast data, allowing to identify a more pronounced non-stationary regime signal. On interannual timescales this signal is dominated by the El Niño Southern Oscillation (ENSO) and it is found to be predictable on seasonal timescales for two of the six regimes considered.

The two regularised clustering methods discussed require the selection of a constraint parameter, which can be made using for example information criteria. In the final part a Bayesian approach to regime assignment, which does not require such a parameter selection, is proposed. This method uses Bayes theorem to incorporate prior information to obtain a better informed probabilistic regime assignment. It leads to more persistent regime dynamics, even compared to the time-regularised clustering method, and helps to identify a pronounced interannual variability signal comparable to

that obtained using the ensemble-regularised clustering method.

The three introduced and investigated approaches help in identifying robust regime variability signals, significantly improving on existing methods. These techniques can be adapted to suit different regime variability signal questions and are not limited to circulation regimes in terms of applicability. The identified regime signals raise questions for future research directions.

Lay Summary

The weather can change a lot. Sometimes it is quite similar for a week, sometimes it changes completely from one day to the next. Such changes are well forecast up to a week in advance, but for predictability further ahead the uncertainty is still large. To improve predictability on these timescales it is important to understand the low-frequency dynamics of the atmospheric circulation. On timescales of one to two weeks the atmospheric dynamics can be represented by so-called atmospheric circulation regimes. These regimes are patterns in the atmospheric circulation, consisting for example of a high pressure area over Scandinavia or the central North Atlantic ocean. They are persistent, lasting longer than a couple of days, and occur repeatedly. The focus in this thesis is on regimes over Europe and the North Atlantic in winter, when the regime dynamics is most noticeable.

The aim in this study is to improve the understanding and determination of these atmospheric circulation regimes. Specifically, in it we study different methods to identify the regime variability signal. Small deviations in the atmospheric pressure patterns can have large impacts on the regime dynamics later on. Imagine you get out of a busy tube train. The time at which you exit the station depends strongly on whether you get out before that retired couple standing next to you, or get stuck behind them going up the stairs. Similarly, small changes in the atmospheric circulation can affect the regime dynamics in the future. It is important to separate the signal, the average time between getting out of the tube train and exiting the station, from the noise, caused by the people walking around you.

To identify the regime variability signal, in the methods studied we make use of prior knowledge on the regime dynamics. Firstly, we know the regimes are persistent and unlikely to switch from one to another and back to the first in one day. Limiting these types of transitions allows for identifying more persistent regime dynamics. Secondly, we know that the chance of a regime occurring is affected by external factors, such as the El Niño Southern Oscillation which is an oscillation in the tropical Pacific sea surface temperature impacting the atmospheric circulation around the world. The effect of such remote drivers can be identified using an ensemble of model realisations, where members of the ensemble are obtained using slightly different initial conditions. How long does it

take twenty different people from the same tube train to get out of the station when all barriers are open, or when work is done on half of them? The movement of those twenty people together gives more information than only considering them one by one. Using such ensemble information for the regimes helps to better identify the effect of external factors on the regime dynamics.

One of the more elegant ways of incorporating prior information in the computation of the regime probabilities is Bayes theorem. Imagine we know that people from the eastbound train take on average 5 minutes to exit the station, while those coming from the westbound train need 8 minutes on average. It took Sarah 7 minutes to exit the station. Did she arrive on the east- or westbound train? Based on the average times, Sarah probably took the westbound train, as 7 is closer to 8. However, if we know that there only is one westbound train an hour compared to six eastbound ones, this can shift the likelihood and indicate that it is more likely that Sarah arrived on an eastbound train. This is what Bayes theorem allows you to compute, using prior information and the observed likelihood. Applied to the circulation regime dynamics it yields persistent regime dynamics with pronounced non-stationary behaviour.

The application of these methods that make use of prior information allows us to identify more pronounced and informative regime dynamics. The regimes are found to be more persistent and the stronger interannual variability in regime frequency increases the regime predictability. The latter is linked to the effect of remote drivers on the regime dynamics. This thesis discusses the methods used in detail and studies the effect on the circulation regime dynamics.

Contents

Chapter 1

Introduction

The IPCC's Sixth Assessment Report states that "It is unequivocal that human influence has warmed the atmosphere, ocean and land. Widespread and rapid changes in the atmosphere, ocean, cryosphere and biosphere have occurred." (IPCC, 2021). Global warming affects all areas of the globe and it is important to understand the regional consequences. This regional impact is what people experience. These are the changes that affect their livelihoods, the production of food, the probability of extreme weather and more. One way to consider the regional effects is through a storyline approach (Zappa and Shepherd, 2017; Doblas-Reyes et al., 2021). Such an approach allows one to separate uncertainty due to lack of knowledge in the climate system response to forcing from uncertainty due to the climate only following one particular realisation from the many that are possible (Shepherd, 2019).

The first step in a storyline approach is to obtain the causal chain of events. For example, rising CO_2-levels lead to increasing temperatures which in turn affect the atmospheric circulation, where the eventual response can be uncertain. Then, conditional on a certain change in the atmospheric circulation, the probability of warm temperature extremes increases a little or a lot. The first part deals with the uncertainty due to lack of knowledge in the climate system response, while the second assesses the intrinsic uncertainty in the climate variability. Here, it is important to understand the non-stationary dynamics within the climate system to know which factors to include in the storyline. Under these non-stationary dynamics we consider all processes that break the system's stationarity. A stationary process is one whose statistics are invariant with respect to shifts in time (Doob, 1953), where non-stationarity destroys this time-invariance. In the context of the climate system such sources of non-stationarity can be time-dependent forces or influences that are external to the system considered.

1.1 Non-Stationarity in the Climate System

A first source of non-stationarity has already been mentioned; climate change. Within the climate system there are many other sources of non-stationarity, often linking atmospheric dynamics in far away regions. Teleconnection patterns, being recurring large-scale patterns of circulation anomalies, represent sources of non-stationary variability (e.g. Feldstein and Franzke, 2017). They connect the atmospheric circulation both regionally, e.g. over the North Pacific, and between regions, e.g. linking the tropical Pacific to the North Atlantic. The latter connections between regions can provide remote drivers for non-stationary atmospheric variability.

The regional teleconnection patterns represent the dominant modes of low-frequency variability. For example, the Pacific North American pattern (PNA), with positive geopotential height anomalies over the north-western part of North America and negative anomalies in the south-east (in its positive phase), and the North Atlantic Oscillation (NAO), with a negative anomaly on the northern side of the North Atlantic and a positive anomaly in the south, are two of the most prominent modes of variability in the northern extratropical atmospheric circulation (e.g. Walker and Bliss, 1932; Hurrell, 1995; Horel and Wallace, 1981; Wallace and Gutzler, David, 1981; Barnston and Livezey, 1987). In this thesis the region of the interest is the Euro-Atlantic sector covering the North Atlantic and Europe. Next to the dominant mode of variability in this region, the NAO (whose variability can be represented by an NAO-index (e.g. Hurrell, 1995)), the East Atlantic pattern (EA) is another prominent variability pattern. It has a negative anomaly just south of Greenland and a positive one south-east of it (e.g. Wallace and Gutzler, David, 1981; Barnston and Livezey, 1987).

There are several modes of variability that are known to impact the atmospheric circulation all around the world. The most prominent one is the El Niño Southern Oscillation (ENSO) in the tropical Pacific ocean (e.g. Rasmusson and Wallace, 1983; Philander, 1983, and references therein). ENSO is an irregular oscillation in the tropical Pacific Sea Surface Temperature (SST) between a warm (El Niño) and a cold anomaly (La Niña) with a period of 3-7 years. A schematic diagram of the different phases is shown in Figure 1.1. In normal conditions there are westward winds located over the equator blowing the warm surface waters towards the western side of the equatorial Pacific. In the region where those warm surface waters are located air rises as it is heated by the ocean, creating a convective circulation with air flowing eastward at higher elevation, cooling down and sinking to the surface near South America, enhancing the westward surface winds. In El Niño conditions the westward winds are weaker and the warm surface water extends further towards the east. On the other hand, during La Niña conditions the westward winds are stronger and the warm surface water

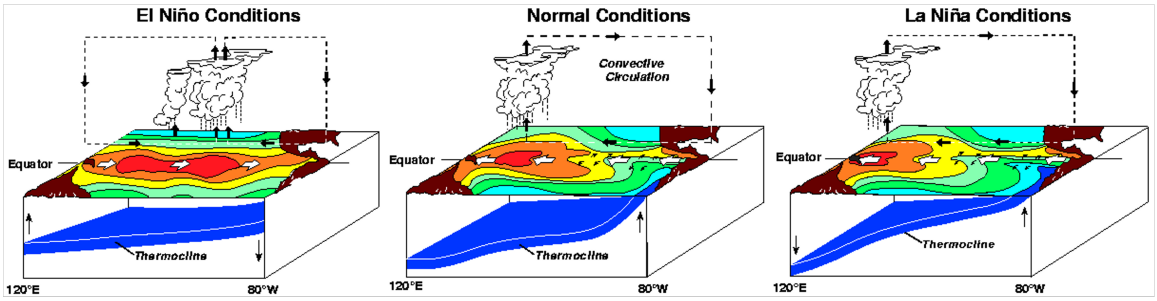


Figure 1.1: The oceanic and atmospheric conditions in the equatorial Pacific region during El Niño, La Niña and normal conditions. Figure courtesy of the Pacific Marine Environmental Laboratory/National Oceanic and Atmospheric Administration.

is located further westward, while the central equatorial Pacific is cooler.

In addition to ENSO, there are more remote drivers affecting the atmospheric circulation around the world and specifically over the Euro-Atlantic sector. Each of these drivers affects the non-stationary dynamics in their own way. Here, we mention some of the remote drivers that impact the northern extratropical atmospheric circulation. Firstly, there is the Quasi-Biennial Oscillation (QBO), which represents variability in the stratospheric winds over the equator with them changing direction around every 14 months (e.g. Holton and Tan, 1980; Baldwin and Dunkerton, 2001). The QBO is known to impact the jet stream over the North Atlantic by confining or dispersing wave activity at mid-to-high latitudes. Secondly, we have the Madden-Julian Oscillation (MJO), which represents tropical variability due to the interaction between the atmospheric circulation and convection with a period of 30 to 60 days (Madden and Julian, 1971). As for the QBO, links with the Northern Hemisphere extratropical circulation have been established for this variability mode (e.g. Hoskins and Karoly, 1981; Garfinkel et al., 2014). The last driver we mention is the stratospheric polar vortex (SPV), whose strength impacts the tropospheric circulation below. Especially sudden stratospheric warmings (SSWs), when the temperature in the stratosphere suddenly rises due to a weakening of the SPV, can strongly impact the Northern Hemisphere tropospheric circulation by shifting the jet stream (e.g. Baldwin and Dunkerton, 2001).

A thorough understanding of the remote drivers of non-stationary variability in the atmospheric circulation is essential to the storyline approach. One needs to know which processes are relevant and have to be included, and which can be left out. The drivers of non-stationarity mentioned here can interact as well, thus it is important to find ways to disentangle the effects of different drivers on the atmospheric circulation. A storyline approach can help here, aiding in developing a better understanding of the dynamics involved.

1.2 Atmospheric Circulation Regimes

To apply a storyline approach to changes in the atmospheric circulation it is useful to separate the dynamics into a limited set of states. This allows one to condition the atmospheric circulation onto each of these states and study its relation with e.g. temperature extremes or energy production by wind farms. When separating the atmospheric circulation into a limited number of states, often referred to as regimes, it is important to think about the criteria to use for this. For the storylines to be physically relevant, it is necessary that these states represent realistic flow patterns. Furthermore, the impact on other variables, such as rainfall, ideally is significantly different between the states. The first point has been assessed in numerous studies for several regions using clustering approaches to find the states, where also impacts on e.g. temperature and energy production have been quantified.

The first classifications of the atmospheric circulation were developed with weather forecasting in mind. During the 1940s and 1950s the German weather service developed the so-called "Grosswetterlagen", which form a classification of the synoptic circulation (Deutscher Wetterdienst, 2019). These 29 synoptic patterns are used up to this day to aid the weather prediction. By predicting the evolution of the weather patterns for the coming two weeks, they help to improve the weather forecast, as the surface impacts of each of the regimes are known. These first classifications of the atmospheric circulation were obtained manually by people with plenty of experience in weather forecasting. From around 1990 onward, the use of computers allowed for more automated and objective ways of identifying these patterns using different types of clustering methods (e.g. Mo and Ghil, 1987; Vautard, 1990; Michelangeli et al., 1995).

These clustering methods allow to identify the low-frequency atmospheric variability as represented by regimes. The concept behind atmospheric circulation regimes is that they are recurrent and persistent patterns within the circulation, representing quasi-stationary regions of the phase space (Hannachi et al., 2017). The idea is that of an atmosphere with multiple stable flow configurations, corresponding to the regimes, as e.g. found in a barotropic channel model (Charney and DeVore, 1979). The timescales on which the regime dynamics evolve is longer than that of baroclinic instability, but shorter than the seasonal variability. This makes them suitable for sub-seasonal predictability, with the rationale that if you can predict the regime you have a better idea of the surface conditions at that time. Despite the significant progress in numerical weather prediction (Bauer et al., 2015), the intrinsic predictability limit of the weather on synoptic timescales (Lorenz, 1969) means that forecasts over two weeks ahead still have at most moderate skill (White et al., 2017). A

better understanding of the circulation regimes can improve the skill of forecasts on these leadtimes, as well as help in improving the understanding of how the atmospheric dynamics responds to sources of non-stationarity. The response to remote drivers can also provide predictability beyond the deterministic predictability limit. The boundary conditions, which are affected by these non-stationary drivers such as ENSO, can inform skillfull predictions on seasonal timescales (e.g Shukla, 1998). Changes in the statistics of the circulation regimes can be affected by such remote drivers on these timescales and thus might be predictable. Note that we are concerned with non-stationarity in the dynamics of the circulation regimes, i.e. their occurrence rates and transition probabilities, and not with changes in the regime patterns themselves.

Initial studies focused on the existence and identification of regimes, primarily in the wintertime Northern Hemisphere. Starting from the idea of blocking as a persistent and stationary feature in the regional circulation (e.g. Rex, 1950; Dole and Gorden, 1983), it was inferred that the extratropical atmospheric dynamics could be considered as transitions between a select number of regimes (Vautard, 1990). For the Northern Hemisphere in winter the regimes obtained in the first regime studies broadly correspond to well known teleconnection patterns, such as the PNA and NAO patterns (Mo and Ghil, 1988; Molteni et al., 1990; Cheng and Wallace, 1993; Kimoto and Ghil, 1993). Following the correspondence with these teleconnection patterns, smaller sectors such as the North Pacific and Euro-Atlantic region were considered. For the North-Pacific sector three regimes were identified (Michelangeli et al., 1995; Jung et al., 2005), where later studies considered a region that also contained North America and obtained four regimes (Casola and Wallace, 2007; Straus et al., 2007). In the Euro-Atlantic region four regimes have been consistently identified (e.g. Vautard, 1990; Michelangeli et al., 1995; Kageyama et al., 1999; Yiou and Nogaj, 2004; Cassou, 2008). The Southern Hemisphere has been less well studied, although some results on regime dynamics exist (e.g. Kidson, 1988; O'Kane et al., 2013).

Some studies have questioned the physical existence and number of regimes, despite the extensive robustness analyses available in literature. For example, Stephenson et al. (2004) found no evidence for multimodality in the climate system and Christiansen (2007) concluded that there is only weak evidence for multiple regimes, making it nearly impossible to identify an optimal number, despite the clear non-Gaussianity of the atmospheric circulation. On the other hand, Hochman et al. (2021) established the physical existence of seven year-round weather regimes over the Euro-Atlantic sector using tools from dynamical systems analysis. Similarly, Dorrington and Strommen (2020) found a clear regime structure over the Euro-Atlantic sector after regressing out the jet speed. This discussion on the physical existence of regimes does not question the usefulness of separating

the atmospheric circulation into a discrete number of states exhibiting low-frequency dynamics for forecasting purposes and to improve the understanding of the circulation.

The knowledge on circulation regimes and their dynamics, both for the North-Pacific and Euro-Atlantic sector, has expanded rapidly in the past decades. After the initial studies establishing the regimes themselves, focus shifted to assessing the impact of these regimes on surface weather and links between the regimes and other modes of variability. A natural first question to ask is what the impact of these circulation regimes is on the surface temperature and precipitation. Liu et al. (1994) considered this for five Northern Hemispheric regimes, where later studies studied it for the Euro-Atlantic sector (e.g. Plaut and Simonnet, 2001; Ortizbeviá et al., 2011) and the North Pacific (e.g. Robertson and Ghil, 1999; Amini and Straus, 2018). Next to a canonical response of temperature and precipitation, also the dynamics of their extremes has been studied in relation to regimes (e.g. Cassou et al., 2005; Yiou and Nogaj, 2004; Yiou et al., 2008).

To better understand the regime dynamics it is important to know how they are connected to other known local atmospheric dynamics features, such as the jet stream and Rossby waves for the North Atlantic region. Over the North Atlantic the jet latitude has been shown to have a multimodal distribution, resulting in a distinction of three so-called jet regimes (Woollings et al., 2010; Franzke et al., 2011; Hannachi et al., 2012). Madonna et al. (2017) related these three regimes, representing a southern, central and northern jet, to the commonly used four Euro-Atlantic regimes by adding a split or tilted jet to the equation. In addition to the jet stream, studies have also considered the link between Rossby wave breaking and the regime dynamics (e.g. Michel and Rivière, 2011; Swenson and Straus, 2017) and investigated how the poleward heat flux is modulated by circulation regimes (Ruggieri et al., 2020).

In this thesis we are concerned with the regimes in the wintertime Euro-Atlantic sector. This is one of the most studied regions where regimes are considered and relevant for understanding the atmospheric dynamics affecting Europe. The winter months show more persistent dynamics compared to summer (e.g. van den Dool and Chervin, 1986) and therefore it is the season on which regime studies tend to focus. As mentioned, most studies have identified four regimes over the Euro-Atlantic region (e.g. Michelangeli et al., 1995; Yiou and Nogaj, 2004; Cassou, 2008), although some arrive at another suitable number of regimes (e.g. Fereday et al., 2008; Grams et al., 2017). These four circulation regimes are the two phases of the NAO (where the negative phase can also be associated with Greenland blocking), a ridge over the Atlantic ocean (Atlantic ridge) and a blocking region over Scandinavia (Scandinavian blocking), as shown in Figure 1.2. In the following sections of this chapter the focus is on this Euro-Atlantic region, with some notes on other regions where

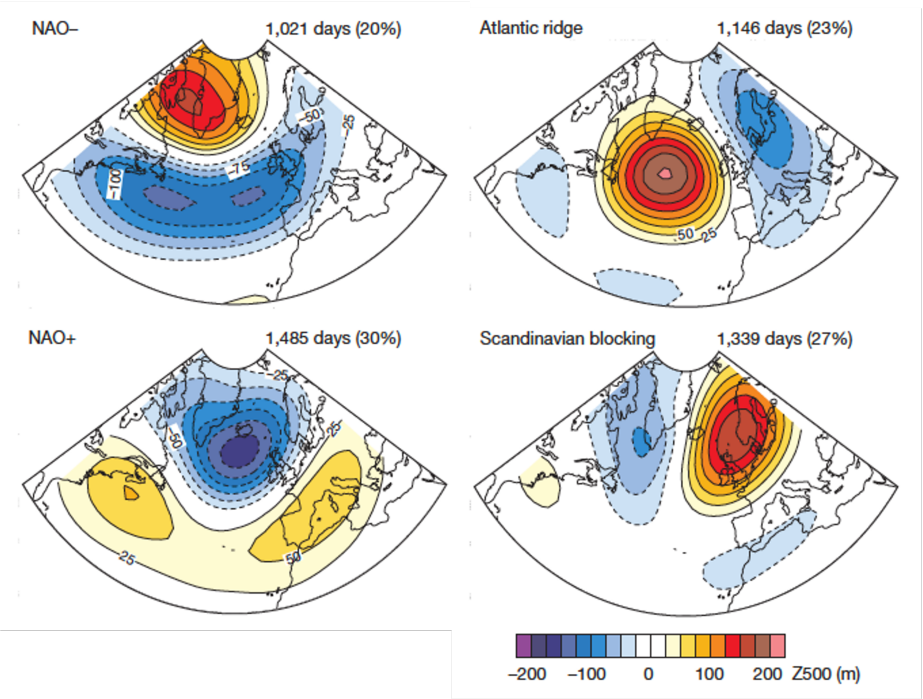


Figure 1.2: The 500 hPa geopotential height (z500) anomalies of the four regimes identified over the Euro-Atlantic sector. Figure by Cassou (2008).

circulation regimes have been studied.

1.3 Circulation Response to External Forcing

Next to local dynamics, external factors, or remote drivers, can also impact the circulation regimes. One of these factors is ENSO, for which the response of the circulation over the North Pacific is well understood. During El Niño conditions the warm SST in the central equatorial Pacific forces atmospheric Rossby waves that travel away from the equator and bend eastward, following a great-circle style path as shown in Figure 1.3 (Horel and Wallace, 1981; Hoskins and Karoly, 1981). These Rossby waves can be largely explained by linear wave theory (Hoskins and Karoly, 1981; Sardeshmukh and Hoskins, 1988) and form a tropospheric pathway by which ENSO affects the atmospheric circulation over the North Pacific and North America. The effect is approximately linear, with opposite responses to El Niño and La Niña, and projects quite well onto the PNA pattern (Horel and Wallace, 1981). These links are reflected in the response of the circulation regimes over North America and the Pacific (e.g. Casola and Wallace, 2007; Johnson and Feldstein, 2010; Riddle et al., 2013; Vigaud et al., 2018).

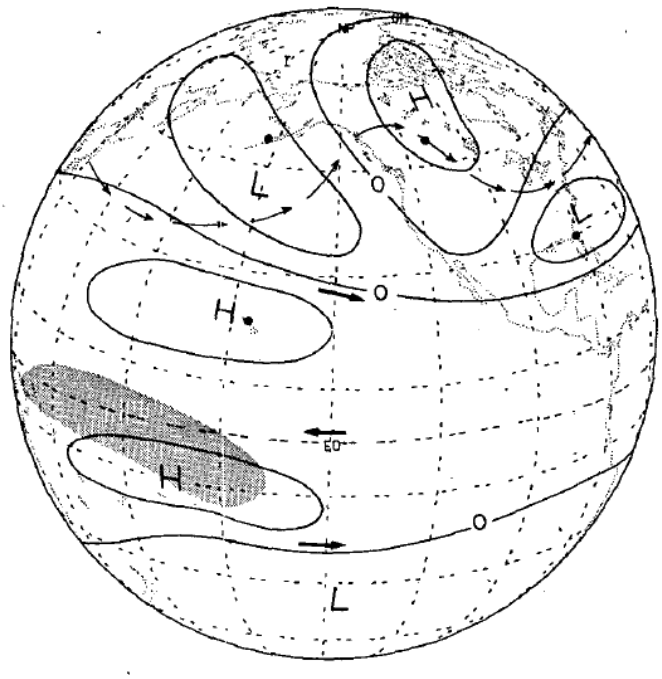


Figure 1.3: A schematic illustration of the Rossby wave response to El Niño indicated by the high (H) and low (L) pressure areas. The thick arrows show the strengthening of the subtropical jets in both hemispheres and the thin arrows indicate a mid-tropospheric streamline. Figure by Horel and Wallace (1981).

The response to ENSO over the North Atlantic and Europe is less well understood. The Rossby wave response to the equatorial SST heating extends to the North Atlantic where it projects onto the NAO (e.g. Trenberth et al., 1998). However, this effect is more difficult to detect due to the large interannual variability in this region (Brönnimann et al., 2007). The canonical response shows a negative NAO in reponse to El Niño events and a positive NAO during La Niña events (e.g. Li and Lau, 2012; Drouard and Cassou, 2019). Individual ENSO events can strongly deviate from this due to the sensitivity of the atmosphere to the specific SST conditions (Mathieu et al., 2004), e.g. it matters whether the highest SST is found in the central or eastern Pacific (Yeh et al., 2018). Also the strength of the temperature anomaly can modulate the atmospheric response over the North Atlantic and Europe (Toniazzo and Scaife, 2006).

In addition to this tropospheric pathway between the tropical Pacific SSTs, a stratospheric pathway has become topic of interest since initial studies indicated its relevance (e.g. Brönnimann et al., 2004; Ayarzaguena et al., 2018). The effect of El Niño and La Niña events on the stratosphere has been extensively studied, with relevant impacts in the Northern Hemisphere (see Domeisen et al., 2019, and references therein). Changes in the Northern Hemisphere troposphere lead to changes in

the polar stratosphere, for which a primary mechanism related to ENSO is the following: El Niño leads to a deepening of the Aleutian Low (a low pressure area over the Gulf of Alaska) in the troposphere, which strengthens the wave flux into the stratosphere by constructive interference with the climatological pattern (e.g. Garfinkel and Hartmann, 2008; Ineson and Scaife, 2009). These anomalous waves propagate upward into the stratosphere, where they break and weaken and warm the polar vortex. This increases the chance of SSWs (e.g. Polvani and Waugh, 2004; Bell et al., 2009; Polvani et al., 2017), which impacts the circulation in the North Atlantic and European sector by increasing the likelihood of a negative phase of the NAO (Hitchcock and Simpson, 2014). This effect is opposite during La Niña events when the polar vortex tends to be colder and stronger. The presence of both a tropospheric and stratospheric pathway makes it difficult to pin down the specific impact of ENSO on the Euro-Atlantic sector. The canonical negative NAO response can be associated with both pathways, although Butler et al. (2014) showed this response only occurred in winters where the stratosphere was severely perturbed.

The study of circulation regimes could aid in better understanding the subtleties of the atmospheric response to ENSO and in disentangling the impacts of the two pathways. Several studies have considered the effect of ENSO on Euro-Atlantic cirulation regimes. In line with the canonical ENSO signal, the response shows an increase of the NAO+ regime during La Niña years and an increase of NAO−, albeit less strong, in El Niño years (e.g. Fereday et al., 2008). Moron and Plaut (2003) noted a difference in the regime response between early (November-December) and late (January-March) winter, with early winter showing the canonical response, but late winter exhibiting contrasting dynamics. In addition to the relation with the standard Euro-Atlantic circulation regimes, also effects on Mediterranean regimes have been found (Giuntoli et al., 2022).

ENSO is not the only remote driver impacting Euro-Atlantic circulation regimes. Also the MJO and the stratosphere have been shown to impact the regime variability. Considering the MJO there exists an extensive literature on links to the NAO, with pathways through both the troposphere and stratosphere (e.g. Hoskins and Karoly, 1981; Garfinkel et al., 2014; Barnes et al., 2019). Furthermore, Henderson et al. (2016) found a link with blocking over the Euro-Atlantic sector with the strongest effects during phases 3,4,6 and 7. Using reanalysis data links between the MJO and Euro-Atlantic circulation regimes have been established, with an increased likelihood of NAO+ after phases 3 and 4 and of NAO− after phases 6 to 8 with a lag of around 10 days, being in line with the known NAO and blocking responses (Cassou, 2008). Lee et al. (2020) studied this effect on the evolution in more detail using 29 Grosswetterlagen regimes, describing more synoptic variability, while Lee et al. (2019) found that the teleconnection between the regimes and MJO is modulated by ENSO, with

the tropospheric pathway being enhanced in El Niño years and the stratospheric pathway during La Niña years. This MJO-link to the Euro-Atlantic circulation has been verified in modelling studies where MJO-like heating was added to the system (Straus et al., 2015; Yadav et al., 2019).

The effect of the stratosphere on the tropospheric circulation regimes, through which one pathway of the MJO to the regimes runs, has recently become a topic of interest. In accordance with the increased probability of the negative phase of the NAO in response to a higher likelihood of SSWs, Charlton-Perez et al. (2018) found that the NAO− regime is more likely after weak SPV conditions, but less likely following a strong SPV. The NAO+ regime showed the opposite response, being more likely following a strong SPV and less likely after a weak SPV, with the Atlantic Ridge regime exhibiting a similar but weaker relation. Following SSW events a Greenland blocking and Atlantic trough regime were found to be most likely when using seven circulation regimes instead of the common four (Domeisen et al., 2020).

A better understanding of the regime dynamics and their links to ENSO, the MJO and the stratosphere can help in improving model predictability on the subseasonal timescales on which the regimes involve. Higher resolution models and including stochastic physics parametrisations have been shown to improve the representation of regimes within models (e.g. Dawson et al., 2012; Dawson and Palmer, 2015; Strommen et al., 2019, Fabiano et al., 2020), which is important for this prediction skill. The forecast skill of models has been found to be flow-dependent, i.e. the regime at the initialization of the forecast affects the skill, with the highest skill found when initialized in the NAO− regime (Ferranti et al., 2015; Matsueda and Palmer, 2018) and forecast busts linked to a poor model representation of regime dynamics (Grams et al., 2018). For predicting the regimes themselves skill is found to be highest in winter (Cortesi et al., 2021). Using seven year-round regimes Büeler et al. (2021) found the skill horizon to be longest for a zonal and Greenland low regime and lowest for blocking regimes, with enhanced skill when the SPV is strong or following MJO phases 4 and 7.

With the increase of solar and wind power, a reliable and useful prediction on sub-seasonal timescales has become even more relevant. More and more research is done in this direction. The use of weather regimes has been shown to be useful for the energy sector on longer timescales, but falls short at short lead times (van der Wiel et al., 2019). This indicates that the common studied circulation regimes may not all be as relevant in predicting energy output, which motivates the introduction of targeted circulation types that also take into account the sensitivity of the energy system (Bloomfield et al., 2020). Since the deployment of wind turbines and solar panels is not equally distributed over Europe, it is important to focus on the effect of the regimes on the relevant

regions (Beerli and Grams, 2019). Grams et al. (2017) used seven year-round regimes to identify what would be a good distribution of wind power over Europe to avoid large variability in output. In relation to this it is important to know whether the regimes and their dynamics will change in a warming world. So far there is no evidence for the regimes themselves changing, but rather for changes in their occurrence rates (Corti et al., 1999). For example, Fabiano et al. (2020) found an increase in the NAO+ occurrence and a decrease of the Atlantic ridge frequency using CMIP6 model runs under projected climate change.

1.4 Predictability and the Signal-to-Noise Problem

When one is interested in sub-seasonal to seasonal prediction it is important to know how well the models do. That is, what is the predictable signal in the model and does this match the signal in the real world? To that end often an ensemble of model runs is considered (Palmer et al., 2004; Wang et al., 2009). The model is run a number of times with slightly different initial conditions, where the individual ensemble members give an idea of the uncertainty in the model prediction. To assess the skill of wintertime seasonal predictions often the ensemble is initialized on November 1st and run till April. After approximately three to four weeks the information of the initial atmospheric conditions is effectively lost. From then on the dynamics of the ensemble is primarily affected by predictable components on these timescales, such as ENSO, which in turn force the dynamics in other regions.

In general, models show moderate skill on seasonal timescales for e.g. the NAO or temperature over Europe (e.g. Baker et al., 2018; Weisheimer et al., 2017, 2019). However, the amplitude of the forecast signal is not always as large as that found in observations (Scaife et al., 2014; Eade et al., 2014; Siegert et al., 2016; Scaife and Smith, 2018). This so-called signal-to-noise paradox indicates that the model is better at predicting the observations than its own ensemble members. As a consequence most skill measures underestimate the real predictability. The forecast signal is determined from the ensemble mean, whereas the individual ensemble members usually do show similar variability to observations. The signal-to-noise ratio, which is the standard deviation of the ensemble mean divided by the standard deviation of all ensemble members (Kumar, 2009), tends to be small, while there still is a reasonable predictable signal. The Ratio of Predictable Components (RPC)

$$\mathrm{RPC} = \frac{PC_{\mathrm{observations}}}{PC_{\mathrm{model}}} \geq \frac{r}{\sqrt{\sigma^2_{\mathrm{signal}}/\sigma^2_{\mathrm{total}}}}, \tag{1.1}$$

being the ratio of the predictable component in observations $PC_{\mathrm{observations}}$ over that in the model

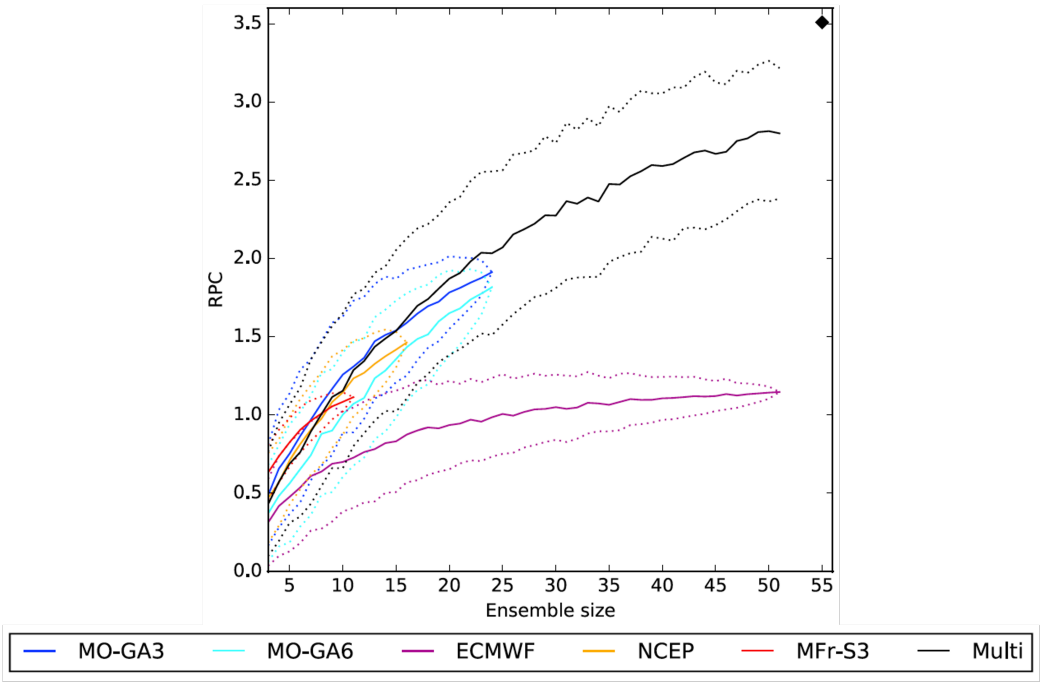


Figure 1.4: The RPC for an NAO-index for different models. The dotted lines give the ± one standard deviation range. Figure by Baker et al. (2018).

PC_{model}, is the common way of studying the signal-to-noise problem (Eade et al., 2014). Here, the predictable component is defined as the square root of the predictable fraction of the variance for respectively the observations and the model. Furthermore, r is the Pearson correlation between the observations and ensemble mean and σ^2 represent the variance of either the signal in the model ensemble mean or the total of the individual members. Ideally its value would be close to one, indicating that the predictable components in observations and model are of similar order. However, in most models the RPC for e.g. an NAO-index is found to be close to two, as shown in Figure 1.4, indicating the predictable component in observations is approximately twice as large as that in these models (Baker et al., 2018).

Highest RPC values on seasonal timescales are found in the mean sea level pressure over Iceland and the Azores, being the centres of action of the NAO (Eade et al., 2014). The Euro-Atlantic sector is the region where the signal-to-noise issue is most prominent and where most research has focused on. In a toy model it was shown that if one considers the NAO as a two-regime system an underestimation of the regime persistence results in both a high skill and high RPC, even with realistic noise levels (Strommen and Palmer, 2019). Similar results were obtained using a simple Markov model (Zhang and Kirtman, 2019). A too-weak teleconnection between the NAO and stratosphere, involving the QBO, has been linked to the small predictable signal in the model

(O'Reilly et al., 2019). However, Charlton-Perez et al. (2019) found no evidence of a signal-to-noise issue in the stratosphere, nor for forecasts initialized on December 1st. They related the low signal-to-noise ratio to a low amplitude of the model signal in early winter. These recent studies indicate that a better understanding of the links between the troposphere and stratosphere is important for better understanding this signal-to-noise issue.

1.5 Clustering Methods for Regime Identification

For understanding the relation between signal and noise it is important to have a suitable way of identifying the relevant signals. When atmospheric circulation regimes are considered this means understanding how the regime patterns and their dynamics are obtained. A number of different regime identification methods have been used towards this end. The first question to address in this regard is what the characteristics of the regimes are. They are recurrent, meaning certain areas of the phase-space are visited regularly, and persistent, indicating they are quasi-stationary and that the dynamics stays in those areas of phase-space for some time (Hannachi et al., 2017). Most studies have focused on recurrence for identifying the regimes, taking into account the desired quasi-stationarity.

To identify recurrent circulation patterns one looks for dense areas in the phase space, which then are associated with the regimes. Some initial studies directly studied the probability density function (pdf) of the data. They first project the gridpoint data onto the first few Principal Components (PCs), which represent the dominant modes of variability, to obtain a lower dimensional phase space. Probability density estimation methods are then used to identify dense regions (e.g. Hansen and Sutera, 1986; Mo and Ghil, 1988; Kimoto and Ghil, 1993). Despite their appeal, these methods are not widely used as they are computationally expensive and difficult to extend to higher dimensions, while there are other suitable methods, such as clustering, that achieve similar results.

Clustering methods have been developed to separate datasets into subsets which are similar within, but different between (Jain, 2010). That is, they identify sets of datapoints which lie close together in the phase space (a cluster), but further away from other sets in the phase space (other clusters). Many different clustering methods have been developed. The main ones that have been used to identify circulation regimes are k-means clustering, hierarchical clustering and mixture models. Of these methods k-means is most widely used, since it is relatively easy to apply and yields good results (e.g. Michelangeli et al., 1995; Plaut and Simonnet, 2001; Straus et al., 2007; Dawson et al.,

2012). With this clustering approach a regime pattern corresponds to the average over all datapoints assigned to a cluster, which is not guaranteed to be a realistic flow pattern. In practice this is not found to be an issue with the regimes corresponding to observed states. The k-means clustering method is discussed in detail in Chapter 2 and is used in this thesis to identify the atmospheric circulation regimes.

Hierarchical clustering, in the way it is used for regime identification, is a clustering method in which each datapoint starts in its own cluster (Johnson, 1967). These clusters are then merged pairwise to the closest cluster, where a new cluster centre is computed as the mean of the two merged clusters. This is repeated until all data is merged into one cluster, forming a hierarchical tree of clusters. One then has to decide on the desired clusters, with the cluster centres being the circulation regime patterns, using some criterion. For example, Cheng and Wallace (1993) used reproducability as a criterion for their three Northern Hemisphere regimes. This method has been used several times (e.g. Toth, 1993; Casola and Wallace, 2007), but is not widely applied for the identification of circulation regimes.

Where k-means and hierarchical clustering assign each datapoint to one regime only, mixture modelling yields a probabilistic approach. It fits a number of (usually Gaussian) distributions to the data, where the region of highest probability is then linked to the regime and each datapoint has a probability of belonging to each of the regimes (e.g. Smyth et al., 1999; Hannachi and O'Neill, 2001). Similar to the approaches using the pdf, this method has the drawback that it struggles to handle high-dimensional data, which is the main reason it is not widely applied. Nevertheless, such a probabilistic approach might be desirable over a categorical, hard, assignment of the data to the regimes and an alternative probabilistic regime assignment approach is considered in this thesis.

1.6 A Bayesian Approach to Hypothesis Testing

Once the regimes have been identified, it is the regime dynamics signal that is of interest. When considering a model ensemble, the ensemble mean response is taken as the signal and compared to observations. As the observations are impacted by noise there cannot be a perfect correspondence, but if the model is good the observations should fall within the ensemble spread. For seasonal forecasts the correlation between (anomalies of) the ensemble mean and observations is often considered as a skill measure (e.g. Weisheimer et al., 2019; Portal et al., 2022). In addition also the ensemble spread is studied to verify the model shows reasonable dynamics compared to the observations.

These are by no means the only or main aspects considered, but they are most relevant for the work discussed in this thesis.

Having computed the ensemble mean as the model signal, one needs a statistical test to verify its significance. The most common approach for this is null-hypothesis significance testing by computing the p-value. However, this approach is by no means ideal and has various problems (Nicholls, 2001; Ambaum, 2010). The p-value indicates the probability of finding (or exceeding) the observed correlation r_0 given the null-hypothesis H_0 of there being no relation $P(r > r_0|H_0)$ (Ambaum et al., 2001). Smaller p-values indicate it is unlikely that there is no relation between the studied variables, but this does not imply there is a relation. Thus the p-value does not give statistical evidence towards there being a relation H, i.e. towards $P(H|r > r_0)$ (Wagenmakers, 2007). One can relate the p-value to the probability of the hypothesis being true given exceedance of the observed correlation by using Bayes Theorem

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \tag{1.2}$$

where D represents the data (here $r > r_0$). Following the computation in Ambaum et al. (2001), one finds that in addition to the p-value, also the odds ratio, i.e. the ratio between the probability of the two hypotheses, is required. This shows that the p-value alone does not provide sufficient evidence towards the truth of a hypothesis.

Besides this fundamental issue with the information contained in a p-value, there is also the point of the threshold used to establish significance. Most studies use 0.05 as a level for significance, but a value of 0.06 is not necessarily much worse. Furthermore, the obtained p-value can strongly depend on the sample size, making it unsuitable as a test for many climate applications, where the timeseries can be relatively short (Nicholls, 2001). One suspects that in most studies the computation of the p-value is done out of habit, or because the editor or a reviewer asks for it, without actually thinking through what one wants to achieve with it (Gigerenzer, 2004; Shepherd, 2021).

An alternative for the p-value in testing hypotheses against data is to consider the Bayes factor (Kass and Raftery, 1995). It arises naturally from Bayes Theorem (1.2) by comparing two hypotheses H_1 and H_2:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)}\frac{P(H_1)}{P(H_2)} = BF\frac{P(H_1)}{P(H_2)}. \tag{1.3}$$

This represents the odds-version of Bayes Theorem, stating that the posterior odds equal the Bayes factor BF times the prior odds. This ratio between the probability of the data given two different hypothesis (e.g. the null hypothesis and its complement) indicates which of the two is more likely, with values above one indicating H_1 is more likely and conversely for H_2. The larger the Bayes

factor, the stronger the evidence towards H_1. In general, values of BF between 3 and 20 are said to constitute positive evidence towards H_1, while values over 20 provide strong evidence (Kass and Raftery, 1995). However, these bounds should by no means be considered as absolute.

1.7 Outline of Thesis

This thesis is concerned with identifying a robust atmospheric circulation regime signal, with a focus on the persistent dynamics and non-stationary variability. To that end we start with a chapter on k-means clustering, which is the most commonly used approach for the identification of circulation regimes. The clustering method is key to the regime dynamics identified and thus it is important to understand how these methods work. This chapter does not present any new results, but merely aims to explain the k-means clustering method used throughout this thesis. The three Chapters 3, 4 and 5 that follow apply this clustering method, and extensions to it, to identify the circulation regimes and their persistent and non-stationary dynamics.

In the study of atmospheric circulation regimes a standard approach has developed and Chapter 3 is concerned with revisiting this procedure. In this standard approach the data is first projected onto the leading Empirical Orthogonal Functions (EOFs), the dominant modes of variability, with their evolution in time represented by the corresponding PCs, to reduce the dimension. Then often a 10-day low-pass filter is applied to remove high frequency variability and focus on the low-frequency dynamics. Next, a k-means clustering algorithm is applied to this filtered data to identify (usually) four regimes over the Euro-Atlantic sector. In reviewing this approach, we aim to answer two questions:

1. What is the optimal number of circulation regimes over the wintertime Euro-Atlantic sector?
2. How persistent are the circulation regimes?

The first question is discussed by considering the common methods for identifying a suitable number of regimes, adding the use of information criteria to the mix. In this respect the input data to the clustering algorithm is considered, being PC data or gridpoint data. It is found that the data used leads to different optimal numbers, where optimal refers to the regimes accurately representing the data without overfitting. The standard four regimes are found to be optimal for PC data, while six regimes is best for gridpoint data. The second question is discussed in the context of the use of a low-pass filter to remove high frequency oscillations, which is contrasted to a time-regularised k-means

clustering algorithm. This regularised clustering approach enforces a level of persistence which, for a suitable constraint value, is found to not affect the regime occurrence rates, while increasing the persistence. This in contrast to the use of a low-pass filter, which does alter the regime occurrence rates.

In Chapter 4 we focus our analysis on the six regimes identified in Chapter 3 using gridpoint data. Here, we are primarily concerned with the identification of robust (sub-)seasonal and interannual regime variability signals, where the main questions being answered are:

1. Is the seasonal variability in regime frequency primarily due to background variability or do other factors play a role?

2. What is the interannual regime frequency variability signal? Is there a predictable signal on these timescales and does the signal-to-noise issue extend to the regime dynamics?

To answer these two questions a novel ensemble-regularised k-means clustering method is developed, enforcing a level of similarity between different ensemble members to identify a stronger non-stationary signal. This regularisation results in a more pronounced non-stationary signal, where information criteria are used to identify a suitable constraint value. The obtained seasonal variability signal is for the most part determined by the background variability. The interannual non-stationary signal is dominated by years with a very strong El Niño, when there is the strongest deviation in regime frequency. Also the signal-to-noise problem in relation to the predictable regime signal is discussed, where we find just as strong a predictable signal in reanalysis data as in the model. This contrasts the presence of a signal-to-noise issue in an NAO-index and we discuss the relation between this index and the regimes.

The persistence and non-stationarity results discussed in Chapters 3 and 4 have been obtained using regularised k-means clustering methods. For such approaches a constraint parameter has to be selected, where e.g. information criteria can be used to inform this choice. The aim of Chapter 5 is to study whether similar persistent and non-stationary results can be obtained using a sequential Bayesian probabilistic regime assignment approach. This approach has the benefit of not requiring any parameter selection, simply following the basic rules of probability. Furthermore, it allows for a better understanding of the regime dynamics by not having a hard, categorical, regime assignment, but a probabilistic one. In this chapter we again discuss some of the questions considered in Chapters 3 and 4, but using this novel regime assignment approach. Specifically, we focus our discussion on the persistent (question 2, Chapter 3) and interannual (question 2, Chapter 4) regime dynamics

obtained using the Bayesian regime assignment. The results are compared with those obtained in Chapters 3 and 4, finding a higher persistence and comparable interannual variability.

At the end of this thesis the results presented in Chapters 3, 4 and 5 are discussed. In Chapter 6 the conclusions from the different chapters are summarized and linked, and directions for future research are outlined.

1.8 Publications

The work discussed in this thesis has led to a number of journal articles. Most results discussed in Chapter 3 are based on a paper published in the Quarterly Journal of Royal Meteorological Society (QJRMS) (Falkena et al., 2020), building on the work presented in the MRes thesis by the thesis candidate (Falkena, 2019). Chapter 4 is based on a second paper published in QJRMS (Falkena et al., 2022). Finally, Chapter 5 is built from a manuscript submitted to the Journal of Climate.

The first draft of all papers, as well as the revisions and response to the reviewers comments, have been written by the thesis candidate. The supervisors provided feedback on these drafts to improve them, which were incorporated by the thesis candidate. It is estimated that the candidate contributed around 90% of the work presented in these papers.

Chapter 2

K-means Clustering

The main purpose of applying a clustering method to atmospheric circulation data is to identify an underlying structure, helping to better understand the atmospheric dynamics and improve predictability. Clustering algorithms partition a dataset in clusters of data that within each cluster are similar with respect to some measure, i.e. low within-cluster variance, but display clearly distinguishable features between clusters, i.e. high between-cluster variance. With e.g. data of handwritten characters it is relatively easy to determine what a "correct" partition is, with the measures that are used not being very sensitive to noise. For atmospheric data, which is strongly impacted by internal variability, the separation between clusters tends to be less clear. This makes it more difficult to identify the optimal set of clusters. The cluster structure also is more complex, which makes it harder to separate the signal from noise.

As mentioned in Section 1.5, k-means clustering is one of the most commonly used methods to identify atmospheric circulation regimes. K-means clustering has been used for a long time in all sorts of applications, from image segmentation (Jain and Flynn, 1996) to studying genome data (Baldi and Hatfield, 2002). It has been discovered multiple times in different contexts from 1955 onwards (Steinhaus, 1956; Lloyd, 1982; Ball and Hall, 1965; MacQueen, 1967). The k-means clustering method is by no means perfect, but no single clustering algorithm proposed up to now has been found to dominate over the other ones (Jain, 2010). Therefore, k-means clustering remains one of the most appealing options due to its simplicity and proven success for numerous applications.

The aim of this chapter is to explain the basics of k-means clustering. We start by discussing the method using a simple example and move on to the mathematical formulation. We then go

through some extensions of k-means clustering, with special attention for a regularisation approach of which variations are used in Chapters 3 and 4. Lastly, we discuss the relevance of a thorough understanding of the data when applying clustering methods. Throughout this chapter we primarily refer to clusters and cluster centres. In the chapters that follow these terms are used interchangeably with "regimes", where the regime pattern corresponds to the cluster centre and the cluster itself consists of a subset of the data assigned to the respective regime. For all results considered in this thesis the cluster centres represent realistic flow patterns (as based on visual inspection of numerous atmospheric flow fields, not shown), although this is not guaranteed as the average itself is not observed.

2.1 Method

To explain how k-means clustering works, we consider the example shown in Figure 2.1. The raw data is shown in the top left panel. On the basis of the visualisation one would divide this data into three clusters, as shown in the bottom left panel. The challenge of k-means clustering is how to obtain such a partition of the data algorithmically. This can be done iteratively as follows:

1. Select three (random) initial cluster centres, indicated by the pluses in panel 1.
2. Assign each datapoint to the nearest cluster centre (with respect to some spatial distance measure), as shown in panel 2.
3. Compute the new cluster centres as the average of all points assigned to each of the three clusters, shown by the stars in panel 3.
4. As in step 2 and 3, assign each datapoint to the nearest cluster centre (panel 4) and compute the updated cluster centres as the average of the data assigned to each of the clusters.
5. Repeat step 2 and 3 till convergence.

Following the procedure described above minimises the so-called k-means clustering functional $\mathbf{L}$, which is given by (Jain, 2010)

$$\mathbf{L} = \sum_{i=1}^{k} \sum_{x_n \in C_i} g(x_n, \theta_i). \tag{2.1}$$

Here $x_n \in \mathbb{R}^m$ is a datapoint, $\theta_i \in \mathbb{R}^m$ a cluster centre and $g : \mathbb{R}^m \to \mathbb{R}$ gives the distance between the two. The clustering functional is the sum of distances between each datapoint associated with cluster i and the centre of that cluster C_i over all k clusters. Minimisation of $\mathbf{L}$ yields a set of

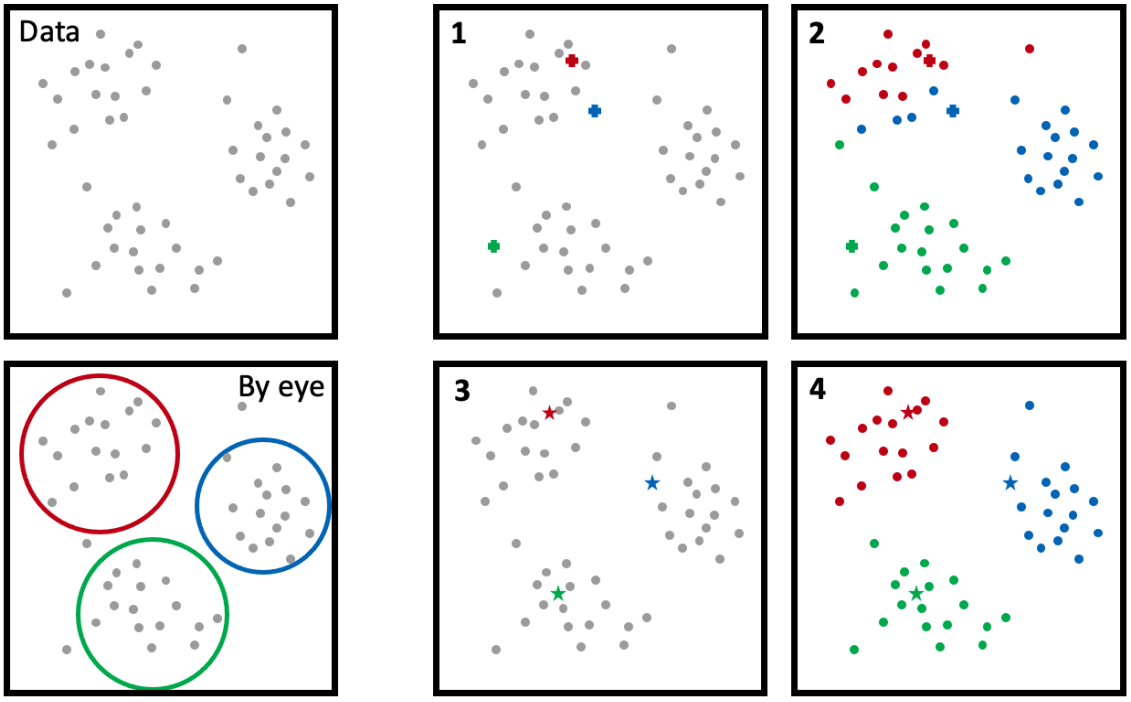


Figure 2.1: An example dataset to explain how k-means clustering works. The left two panels show the raw data (top) and a clustering based on visual inspection (bottom). The right four panels go through the steps of the k-means clustering algorithm.

clusters with low within-cluster variance and larger between-cluster variance, which is exactly what the iterative k-means method described above achieves. Ideally, minimisation of $\mathbf{L}$ occurs when all $x_n \in c_i$ are close to θ_i, but far from all other θ_j, $j \neq i$, which in practice is not always as straightforward.

There are a number of choices that have to be made when employing k-means clustering which have not been mentioned yet in this explanation. The first choice is on the number of clusters k that one wants to partition the dataset into. In the example in Figure 2.1 we decided to use $k = 3$ clusters based on visual inspection, but this is not feasible for all datasets one might be interested in. The common solution to this is to do the clustering multiple times for different values of k and determine which of these partitions is the best following some metric, e.g. the partition with the lowest value of the clustering functional $\mathbf{L}$. A second choice that has to be made is about the distance to use, where the standard Euclidian (or L^2) distance

$$g(x_n, \theta_i) = ||x_n - \theta_i||^2, \tag{2.2}$$

has been used in the example. This is a frequently made choice and yields good results in many

situations, but depending on the data other distance metrics g might be more suitable.

The last yet crucial choice is about selecting appropriate initial cluster centres, which can strongly impact the final clusters. In the shown example this may seem unlikely, but there are numerous cases where this can be an issue, as k-means clustering is only guaranteed to find a local, but not global minimum of $\mathbf{L}$. For that reason it is common practice to run the k-means clustering algorithm several times with randomly initialised cluster centres. Each of these runs has a corresponding minimal value of the clustering functional $\mathbf{L}$, which is a local minimum. The clusters corresponding to the smallest local minimum value of $\mathbf{L}$ over all runs with different initial cluster centres are then taken to be the optimal ones. That value of $\mathbf{L}$ is then considered as the global minimum, although this cannot be proven.

2.2 Extensions of k-means Clustering

Many extensions of k-means clustering have been discussed in literature. Some of these primarily focus on optimising the speed of convergence, for example by more efficiently assigning the data to the clusters (Pelleg and Moore, 1999) or by refining the initial conditions used (Fayyad et al., 1998). The latter approach has also been applied in some circulation regime studies with results comparable to those obtained using standard k-means clustering (e.g. Amini and Straus, 2018). Other approaches find the optimal number of clusters k automatically using a certain criterion (Pelleg and Moore, 2000) or consider the median of the data as the cluster representative instead of the mean (Kaufman and Rousseeuw, 2005). Another ansatz to extend k-means is to add a constraint of some form to the underlying optimisation procedure. This regularised version of the k-means clustering approach is introduced in some more detail, as it forms the basis of the methods used in Chapters 3 and 4.

2.2.1 Regularisation Approach

When clustering is employed to identify atmospheric circulation regimes, a comparison with methods for e.g. image segmentation as made at the start of this chapter might not be fair. In contrast to some of the common areas in which clustering methods are applied, the data considered for the regime identification has time as an extra dimension. The aspect of interest is the metastable dynamics of the data, for which time is an essential component. The regularised (finite element) approach proposed by Horenko (2010*a*) has been designed with exactly this time-aspect in mind. The approach is not specific to k-means clustering, but can be applied in this context. Here, we

introduce the concept of this method for the application of k-means clustering.

As discussed in Section 2.1, k-means clustering minimises the clustering functional $\mathbf{L}$ given in Equation (2.1). To allow for a discussion of the regularisation approach and include time as a dimension, here we consider

$$\mathbf{L}(\Theta, \Gamma) = \int_0^T \sum_{i=1}^{k} \gamma_i(t) g(x_t, \theta_i) dt. \tag{2.3}$$

As before θ_i is a cluster centre with $\Theta = \{\theta_i\}_{i=1,\dots,k}$ for k clusters and g is a distance functional. Instead of n, time t is used to identify the datapoints x_t changing the sum to an integral (although in practice it mostly still is considered as a sum, as measurements are taken at discrete times). The new term $\Gamma = \{\gamma_i(t)\}_{i=1,\dots,k,t\in[0,T]}$ with $\gamma_i(t) \in \mathbb{R}$ represents the cluster (or regime) assignment. Each $\gamma_i(t)$ lies between zero and one and for each time t their sum over all clusters i equals one. For k-means clustering $\gamma_i(t)$ is a indicator function, being one when the data at time t belongs to cluster i and zero otherwise. However, in general this does not need to be the case, thus extending the standard k-means approach.

The introduction of Γ is required for the regularisation of the clustering method. There are two ways in which this can be done. The first option is to add a regularisation term to the clustering functional itself, e.g. $\mathbf{L} + \int_0^T \sum_{i=1}^{k} \epsilon^2 (\partial_t \gamma_i(t))^2 dt$ (Horenko, 2010*a*). Secondly, one can impose a constraint on $\gamma_i(t)$, e.g. $\int_0^T \sum_{i=1}^{k} |\gamma_i(t+1) - \gamma_i(t)| \leq C$, which regularises the dynamics of the underlying system using the cluster assignment. Both approaches typically yield the same result when the regularisation parameters ϵ and C are chosen accordingly. The drawback of the first approach is that the fuzziness of the results makes it harder to interpret them (de Wiljes et al., 2013). In general it is more difficult to tune the ϵ parameter as there is no physical interpretation. On the other hand, the C parameter, used in the second approach, can directly be connected to shifts in the metastability of the system and might even correspond to a variable that one has prior information on. Therefore, in this thesis the second constrained regularisation is considered. Examples of such a regularisation approach are the enforcement of more persistent dynamics by imposing a constraint on the number of transitions between clusters (discussed in detail in Chapter 3, Section 3.3.1) or the regularisation over a model ensemble to identify a more pronounced non-stationary regime signal (Chapter 4, Section 4.2.2). Furthermore, a regularised approach has been applied to represent data using regression model clusters with Γ modelling the switches between them (Horenko, 2010*b*). A similar method has been applied to study the metastability of the southern hemisphere circulation (O'Kane et al., 2013) and identify a circulation regime model of the NAO (Quinn et al., 2020).

2.3 Importance of Understanding Data

With any clustering approach it is important to understand the data considered. Take for example the data shown on the left in Figure 2.2. Standard k-means clustering fails to identify the two clusters and when applied using $k = 2$ will simply split the data in half via some arbitrary middle line (depending on the initial cluster centres). However, when one considers this data not in the x-y-plane, but in radial coordinates, or projects the data onto the two leading eigenvectors, they are clearly separate and k-means clustering will succeed in identifying the correct clusters. Hence it is crucial to consider the nature of the data and to preprocess if necessary. This example also indicates that not all data is suitable for k-means clustering and you first need to understand your data before deciding on a suitable clustering method.

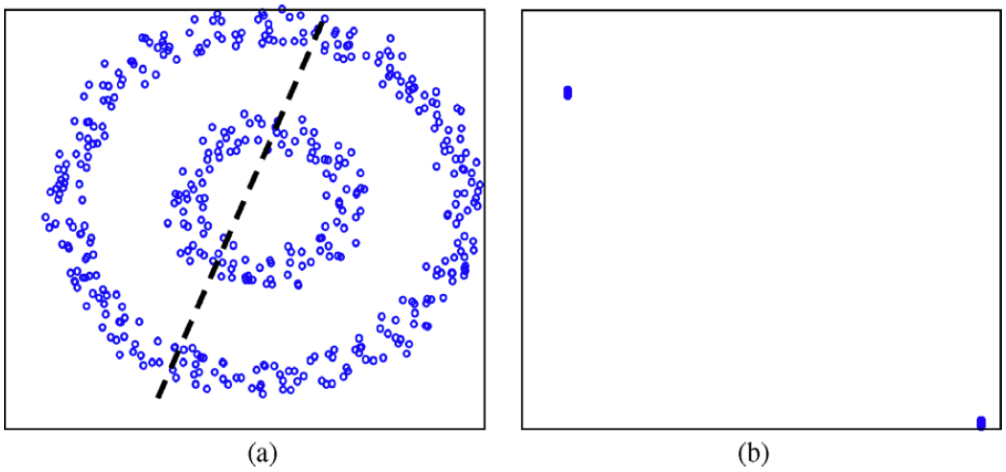


Figure 2.2: An example of a dataset for which k-means clustering fails to identify the correct clusters is shown in the left panel. By projecting onto the leading eigenvectors, shown in the right panel, k-means clustering can identify the two clusters. Figure by Jain (2010).

Another aspect for which it is important to first study the data is the number of clusters. The examples discussed so far had a clear correct number of clusters identifiable by visual inspection. However, it is not always as clear what the optimal number of clusters is. Consider the example in Figure 2.3 for which the data was generated with six clusters. If this information is not available, two or five clusters also appear reasonable values to consider for the k-means procedure. The optimal number of clusters can thus be ambiguous and depend on the features and information of interest. Several methods can be used to identify a suitable number of clusters to consider, but in cases such as the one shown in Figure 2.3 different methods may yield different results.

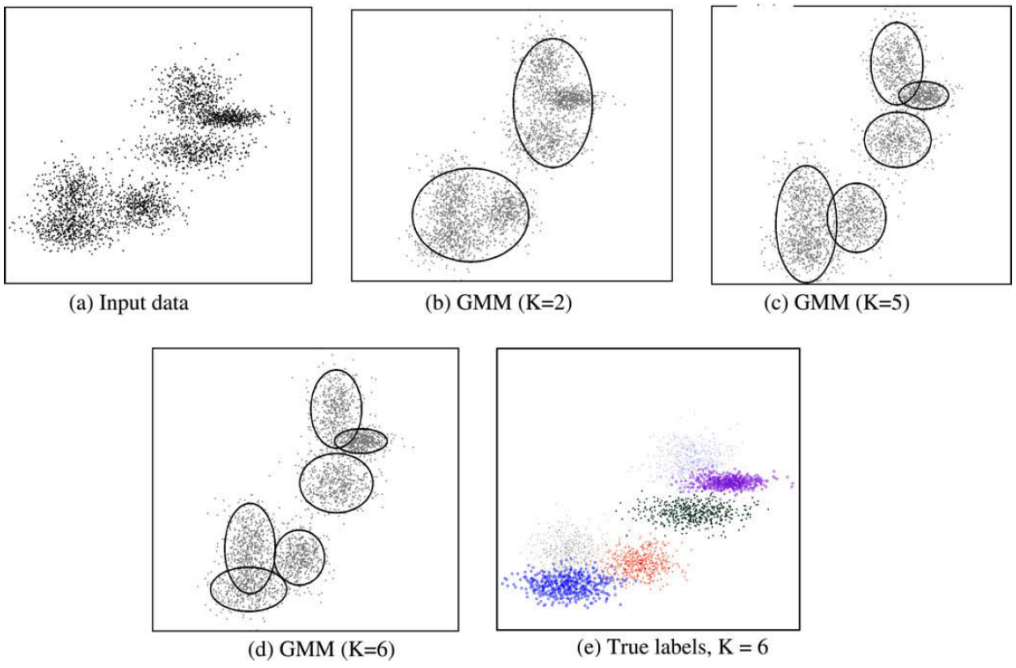


Figure 2.3: An example of a dataset, generated with six clusters (bottom right), where the optimal number of clusters is not straightforward to determine. Figure by Jain (2010).

For datasets which can straightforwardly be represented visually, it is relatively easy to understand what might be the suitable clusters (after some discussion when it is less clear). However, for high dimensional datasets, such as the ones we will use for the regime identification, this is more difficult. There are numerous options in which to visually represent the data and it is nearly impossible to find a way of representing the data in which the clusters are clearly visible without an understanding of the data. Considering the robustness of the clusters with respect to different initial seeds can help understanding whether the obtained clusters are a real feature. For example, the data shown in Figure 2.2 is likely to yield different clusters depending on the initial cluster centres, while the results for the data in Figure 2.3 are quite robust. This does not make it easier to define the best number of clusters to use, but it can help in determining whether the clusters for a given k are reasonable.

Chapter 3

Regime Identification Revisited

In this chapter the most commonly used approaches to circulation regime identification are revisited. We discuss the clustering method, as well as the filtering of the data that is often done before applying a clustering algorithm. Both can strongly affect the identified regimes. We start with a discussion of the standard methods in Section 3.1, focusing on regime identification in the wintertime Euro-Atlantic sector. Two aspects of this approach are then considered in the subsequent sections. Firstly, we compare the use of principal component (PC) data with that of gridpoint data and discuss the effect on the most suitable number of regimes and the regime robustness in Section 3.2. Secondly, we study the effect of applying a low-pass filter to remove high frequency oscillations on the regime dynamics and compare this with a novel time-regularised clustering method in Section 3.3. This novel algorithm, which is explained in Section 3.3.1, does not change the data itself, but instead enforces a level of persistence in the method for regime identification. Both comparisons give insight into the effect of filtering the data, either by employing PC analysis or by using a low-pass filter, before applying a k-means clustering algorithm. A summary and brief discussion are given in the final Section 3.4.

3.1 What is the Standard Regime Identification Approach?

The standard method for identifying circulation regimes is k-means clustering (e.g. Michelangeli et al., 1995; Straus et al., 2007), discussed in detail in Chapter 2. In addition to the clustering method, there are choices that have to be made concerning the data to be used for clustering. The raw data usually is in the form of gridpoint data, mostly on a longitude-latitude grid. Nearly always

this gridpoint data is projected onto the first several Empirical Orthogonal Functions (EOFs), after which the clustering algorithm is applied to the time series of the corresponding PCs (e.g. Vautard, 1990; Ferranti et al., 2015). EOFs represent the dominant modes of variability, with the PCs indicating their strength in time. They are a suitable way to reduce the dimensionality of the data, while retaining most of the atmospheric variability. In addition to PCs some studies apply a low-pass time filter to the data to remove high frequency, noisy oscillations and focus on the low-frequency behaviour (e.g. Straus et al., 2007; Grams et al., 2017). This latter approach enforces a higher persistence of the regimes compared to standard k-means clustering, which is independent of the time-ordering of the data. Since the data that is put into the clustering algorithm is key to the outcome (Section 2.3), it is likely that these decisions do affect the obtained regimes.

Clustering methods represent a projection of the data to a lower dimensional state space and thus applying clustering to the already filtered data of PCs means a projection of the data is done twice. Thirty years ago this approach was necessary, because computational limitations did not allow using the full gridpoint dataset. However this is no longer a constraint. Nevertheless, most studies continue to follow the original approach and use PCs. As EOFs give the modes associated with the most variability, while clusters give the recurrent patterns, the means of dimension reduction is quite different. The question thus arises of what the effect of this double filtering is on the resulting atmospheric circulation regimes. Similarly, applying a low-pass filter to remove the high-frequency behaviour before the cluster analysis also means the data is filtered twice. This is likely to not only affect the persistence, but also the occurrence of the found regimes and possibly the clusters themselves, thus raising the question of how strong this effect is.

Next to the initial projection and filtering of the data, also choices within the k-means clustering method have to be made. The number of clusters k is set a priori, making finding the optimal number of regimes part of the problem, as discussed in Section 2.1. The optimal number is one where the regimes accurately represent the data, without over-fitting and possibly mistaking noise to be part of the regime signal. Commonly-used methods to identify this optimum are the verification of significance using synthetic datasets (e.g. Straus et al., 2007; Dawson et al., 2012; Straus et al., 2017), using a classifiability index (e.g. Michelangeli et al., 1995; Plaut and Simonnet, 2001), and looking at the similarity of runs with different initial conditions (e.g. Jung et al., 2005). For the wintertime Euro-Atlantic sector most initial studies identified four as the optimal number of circulation regimes and many subsequent ones have simply taken this and computed their own regimes using $k = 4$.

3.1.1 Data, Preprocessing and k-means Clustering

To assess the effect of using PCs and applying a low-pass filter, we detail the data and standard method considered as a baseline in this chapter. We use 500 hPa geopotential height (z500) data from ERA Interim on a 2.5° by 2.5° longitude-latitude grid for a domain covering the Euro-Atlantic sector, 20° to 80°N and 90°W to 30°E (Dee et al., 2011). Daily data (00:00 UTC) is considered for the months December through March using 39 years of data (1979 - 2018). Deviations from a fixed background state are used throughout this period. The main argument for considering a fixed background state instead of a seasonally varying one is that when applying cluster analysis the data used is preferably as complete as possible to avoid any type of bias. This means that few to no assumptions, such as a seasonal cycle, are made in preparing the data to retain the information present in the data. Or, phrased differently, how can one compare two days if they are deviations with respect to a different background state? The risk of this approach is that seasonality affects the regimes that are found and introduces a bias in the occurrence and persistence. Thus there is a trade-off to be made between obtaining as large a sample size as possible, using data for more months, whilst minimizing such effects. The rationale for the choice of the period December-March is based on the difference in background state between the different months and discussed in detail in Falkena (2019) and the appendix to Falkena et al. (2020). Based on this analysis, we do not expect the found regimes to be sensitive to the removal of the seasonal cycle. Differences in the occurrence and persistence of the regimes cannot be ruled out and their dynamics throughout the winter season is discussed in Section 3.2.3.

The method used for the identification of circulation regimes is k-means clustering (Jain, 2010). That is, we minimise the clustering functional $\mathbf{L}$ given in Equation (2.1) using the standard Euclidian distance (L_2-norm) in Equation (2.2) weighted by the cosine of latitude (Chung and Nigam, 1999), following the procedure discussed in Section 2.1. This method is applied to both the gridpoint dataset, as well as the time series of the first 5, 10, 15 and 20 PCs. Furthermore, the method is applied to the gridpoint data after applying a 5- and a 10-day low-pass filter to remove high-frequency oscillations. Because k-means clustering can only identify local minima we run the k-means clustering algorithm 500 times with different random initial conditions. The initial condition at every location in space is drawn independently from a normal distribution around zero with the same standard deviation as the data. Note that this means there is no correlation in space, so as to not make any assumptions on the spatial patterns of the regimes. The tolerance used depends on k and is $0.0001/k^2$ for the gridpoint data and $0.001/k^2$ for the PC data. A comparison between the PC and gridpoint results is made in Section 3.2, while a discussion on the persistence and effect of

the low-pass filter is given in Section 3.3.

3.2 Number of Regimes

This section focuses on the question whether the optimal number of regimes is different between PC data and gridpoint data. We start with a discussion of information criteria, which are used to inform the selection of a suitable k as done in the section that follows. The identified optimal numbers of regimes differ between the two datasets, with $k = 4$ found to be most suitable for PC data, but $k = 6$ for gridpoint data. These two sets of regimes are compared in Section 3.2.3. We end this section with a discussion of an alternative clustering method in Section 3.2.4 to establish the robustness of the identified regimes.

3.2.1 Information Criteria

For k-means clustering the number of clusters k has to be set a priori and the question is how to determine the best value for k. The main methods, as mentioned in Section 3.1, consider consistency and robustness within the clustering algorithm results (e.g. Michelangeli et al., 1995) or compare results with those obtained for synthetic datasets (e.g. Straus et al., 2007). An alternative method is to use an information criterion (e.g. O'Kane et al., 2013), which is widely used in for example biological sciences (e.g. Volinsky and Raftery, 2000; Posada and Buckley, 2004; Arnold, 2010). An information criterion is a tool from model selection which is used to identify the optimal model (Burnham and Anderson, 2004); it strikes a balance between how well the model fits the data and the number of parameters needed, to prevent over-fitting. The optimal balance is where the information criterion is minimal. As the clusters are effectively a model representing the data, the concept can be applied here as well. Such a cluster model is described by the k cluster centres and the assignment of the data to them. Specifically, the parameters are k times those needed to describe the cluster centres and $k - 1$ (considering the cluster assignment as a k vector which sums to 1) times the length of the time series. The information criterion tool for identifying the optimal number of clusters has already been used in many applications and more theoretical studies (e.g Fraley and Raftery, 1998; Chen and Gopalakrishnan, 1998; Cobos et al., 2014).

The two information criteria that are used most widely are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Burnham and Anderson, 2004). The AIC is based in information theory and is an approximation of how different two probability distributions

(one for the data, one for the model) are (Akaike, 1973). It is given by

$$\mathrm{AIC} = -2\log(\mathcal{L}(\hat{\theta}|\mathrm{data})) + 2K, \tag{3.1}$$

where $\mathcal{L}(\hat{\theta}|\mathrm{data})$ is the likelihood of the optimal model $\hat{\theta}$ given the data, which measures how well the model fits the data, and K the number of parameters in the model. That is, $K = k \cdot m + (k-1) \cdot T_{nr}$ with m the dimension of the data, i.e. latitude times longitude or number of PCs, and T_{nr} the number of time steps. In contrast the BIC is based on the limiting behaviour of Bayes estimators, which minimizes the expectation value of the loss (e.g. error), and reads (Schwarz, 1978):

$$\mathrm{BIC} = -2\log(\mathcal{L}(\hat{\theta}|\mathrm{data})) + K\log(n), \tag{3.2}$$

where n is the sample size, here being the dimension of the data (number of gridpoints or number of PCs) times the number of days T_{nr}. Just as with the AIC, the BIC strikes a balance between the (log-)likelihood, i.e. how well the clusters fit the data, and the number of parameters in the model. For both the AIC and BIC we refer to the second term as the penalty term since it penalizes the use of many parameters in finding the optimal model to prevent over-fitting.

To compute the values of both information criteria the log-likelihood is needed. Assuming the errors of the model are independent and normally distributed the log-likelihood term can be written as (Burnham and Anderson, 2004)

$$-2\log(\mathcal{L}(\hat{\theta}|\mathrm{data})) = n\log(\hat{\sigma}^2), \tag{3.3}$$

where $\hat{\sigma}^2 = \sum \hat{\epsilon}_t^2/n$ is the error variance for residuals $\hat{\epsilon}_t$, the latter being the difference between the cluster centres and the data for every grid point or PC. This allows for a straightforward computation of both information criteria using the clustering functional $\mathbf{L}$. Note that the assumption of independent errors might not be strictly true for gridpoint data, however the 2.5° resolution is expected to be sufficiently large for the above to be a reasonable approximation. For higher resolution grids the dependence of errors can be an issue that would have to be taken into account.

The only difference between the AIC and the BIC is how they penalise the number of parameters in the model. The penalty term in the BIC takes into account the sample size, while the term in the AIC does not. This means the penalty term in the BIC is stronger with respect to the number of parameters, accounting for an assumed higher variability in a high-dimensional dataset, which increases the chances of over-fitting. To be physically plausible and useful the number of regimes ideally is larger than two, to represent the circulation in sufficient detail, but small enough to allow for an interpretable reduced description of the dynamics, i.e. roughly below twenty. Since the number of parameters needed to identify a cluster centre is set by the dimension of the data, it is different

Table 3.1: The values of the different terms in the AIC and BIC for both the PC and gridpoint data.

	20 PCs ($\cdot 10^4 + 26 \cdot 10^5$)			Gridpoint ($\cdot 10^4 + 17 \cdot 10^6$)		
k	$-2\log(\mathcal{L})$	$2K$	$K\log(n)$	$-2\log(\mathcal{L})$	$2K$	$K\log(n)$
2	14.04	0.95	5.46	153.59	1.44	11.2
3	12.50	1.90	10.9	132.75	2.63	20.5
4	11.33	2.85	16.3	117.22	3.82	29.7
5	10.43	3.80	21.8	105.11	5.01	39.0
6	9.68	4.75	27.2	95.29	6.20	48.3
7	9.16	5.70	32.7	88.41	7.39	57.5
8	8.70	6.65	38.1	82.30	8.58	66.8
9	8.27	7.60	43.6	76.68	9.77	76.1
10	7.89	8.55	49.0	71.68	11.0	85.4

between the PC and gridpoint data. This means that the suitability of the strength of the penalty term in either of the two information criteria may depend on the data considered. The dimension of PC data is significantly smaller than that of gridpoint data. As a consequence the penalty term of the BIC is stronger relative to the log-likelihood term for PC data compared to gridpoint data. This is illustrated in Table 3.1 using 20 PCs, which shows that the penalty term from the BIC is too strong for the PC data, while that of the AIC is not strong enough for the gridpoint data. For this reason the BIC is expected to not perform well for PC data, by which we mean that it will identify a very low number of clusters k to be optimal. On the other hand, the penalty term in the AIC likely is too weak to yield a realistic optimal k for the gridpoint data. A high number of clusters is expected to be found optimal, well beyond what is physically reasonable and suitable. When using either of these criteria one always has to judge whether the result is sensible for the purpose of the study.

3.2.2 What is the Optimal Number of Regimes?

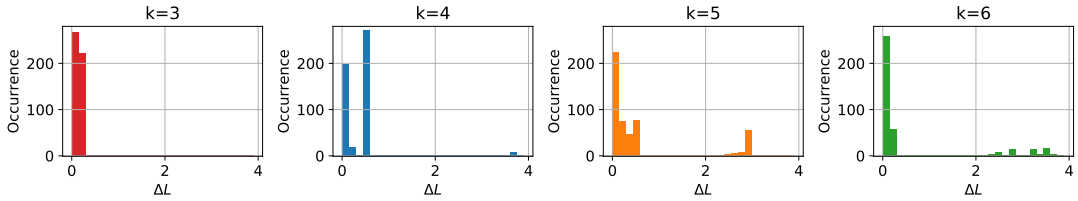
The standard number of wintertime regimes identified over the Euro-Atlantic sector in literature is four (e.g Vautard, 1990; Cassou, 2008; Dawson and Palmer, 2015) and few studies question this number (e.g. Fereday et al., 2008). This optimal number of four clusters has always been found in the context of PC data. The most-used argument for four regimes being optimal is based on how consistent, or similar, the results of the k-means algorithm are for different (random) initial

conditions. A way in which this is commonly assessed is the classifiability index introduced by Michelangeli et al. (1995), which uses pattern correlation to determine how similar two sets of regimes are. The value of the index is computed for both the data and a synthetic dataset with the same statistics, to see whether there exists a k for which the result is significantly different from a noise model. Another method to assess the similarity is to look at the assignment of the data to the different regimes (Fereday et al., 2008). This similarity measure considers the regimes for all data (instead of only their average) and thus possibly provides more information than the pattern correlation.

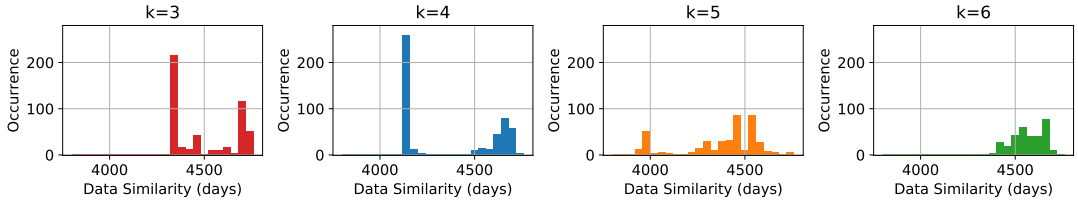
Here we briefly examine the data similarity, as well as the spread in the clustering functional $\mathbf{L}$ for the PC regimes with the main aim being to verify the reliability of the found regimes. Histograms for $\Delta\mathbf{L} = \mathbf{L}_{\text{run}} - \mathbf{L}_{\text{min}}$, being the difference of $\mathbf{L}$ for each run with the minimal value over all 500 runs of the clustering algorithm with different initial conditions, and the number of days assigned to the same regime (as the best result) are shown in Figure 3.1. The first aspect to note is that for some k not only a global minimum is found, but also a local one (e.g. for $k = 4$ in Figure 3.1a). In addition we see that the obtained regimes can be quite different in their assignment of data to certain regimes, indicating the found regime patterns are significantly different. Instead of only combining all accurate results (near the global minimum of $\mathbf{L}$) by computing the average data similarity (similar to the classifiability index Michelangeli et al. (1995)) we also look at its variance. The values for $k = 3, ..., 6$, given on the left side of Table 3.2, show that the variance is lowest for $k = 4$, which goes together with a high average. This is consistent with the results found in literature. Interestingly the mean for $k = 6$ is slightly higher than found for $k = 5$, while the opposite would be expected as more clusters allow for more variability.

Next we turn to the distribution of $\Delta\mathbf{L}$ and the data similarity for the gridpoint data, as shown in Figure 3.2. The first thing to note is that the distributions of $\Delta\mathbf{L}$ look similar to those for the PC data, indicating that using the high dimensional gridpoint data does not reduce the chance of finding the optimal regimes. Some differences with the PC result do occur for the data similarity, most notably the increased similarity for $k = 5$. Looking at the results for the global minimum we find that both $k = 5$ and $k = 6$ show a smaller variance than $k = 4$, in contrast to the PC results. Especially for $k = 5$ the difference is substantial. These differences indicate that by performing an EOF analysis some information is lost, resulting in a stronger consistency of k-means for $k = 5$ and $k = 6$ when using the gridpoint data.

In addition we look at the results of the clustering algorithm for a subset of the data, which is a standard approach to test the robustness of clustering methods (Jain, 2010). Here the results for the



(a) The difference of the clustering functional $\Delta\mathbf{L}$.

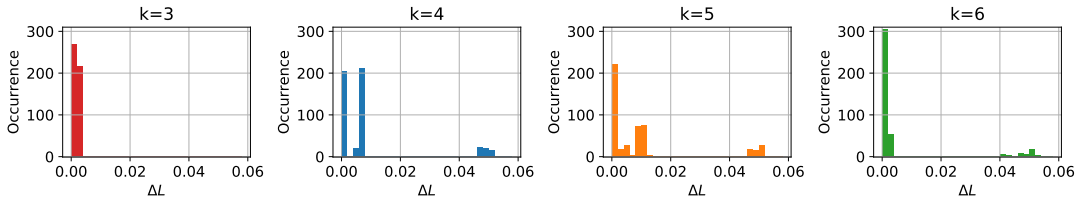


(b) The data similarity.

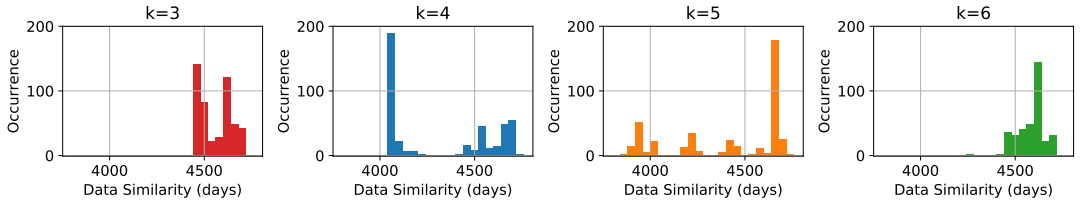
Figure 3.1: Histograms for the difference of the clustering functional with its minimum value $\Delta\mathbf{L} = \mathbf{L}_{\text{run}} - \mathbf{L}_{\text{min}}$ and the data similarity with respect to the optimal (minimal $\mathbf{L}$) result using the first 20 PCs for $k = 3, ..., 6$.

Table 3.2: The mean (μ) and variance per cluster (σ^2/k) of the data similarity for data with $\mathbf{DL} = \mathbf{L}_i - \mathbf{L}_{i+1}$, with i indicating a different initialisation of the clustering algorithm, where the $\mathbf{L}_i$ are sorted from small to large, below a set threshold for both the PC and gridpoint results. For the gridpoint results also the values for the odd and even years are given. The number of runs that are below the threshold is shown as well.

	20 PCs: $\mathbf{DL} < 0.01$			Gridpoint: $\mathbf{DL} < 0.0005$								
	All years			All years			Odd years			Even years		
k	μ	σ^2/k	#data	μ	σ^2/k	#data	μ	σ^2/k	#data	μ	σ^2/k	#data
3	4643	3649	254	4552	2109	485	2342	875	409	2246	140	156
4	4658	330	197	4607	1440	201	2359	452	248	2255	37	423
5	4509	978	265	4660	149	204	2187	3564	274	2243	132	137
6	4571	1103	315	4581	790	316	2296	322	60	1911	10478	210
7	-	-	-	3686	37440	417	1998	12145	165	2113	1470	170



(a) The difference of the clustering functional $\Delta\mathbf{L}$.

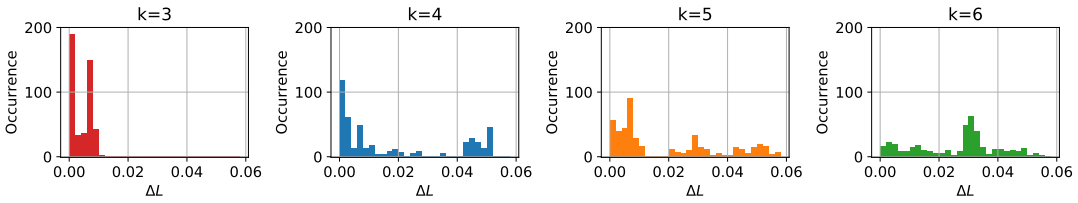


(b) The data similarity.

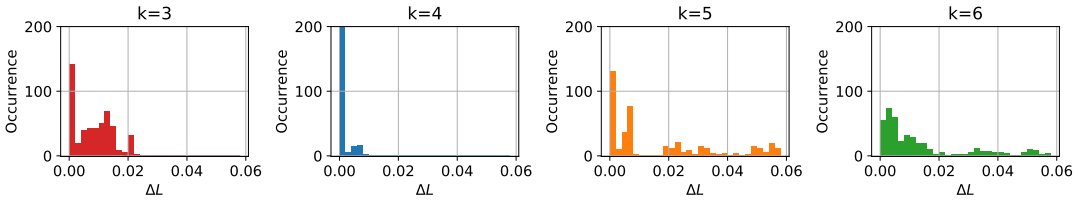
Figure 3.2: Histograms for the difference of the clustering functional with its minimum value $\Delta\mathbf{L} = \mathbf{L}_{\text{run}} - \mathbf{L}_{\text{min}}$ and the data similarity with respect to the optimal (minimal $\mathbf{L}$) result using the gridpoint data for $k = 3, ..., 6$.

datasets of odd and even years are studied, as stationarity of the dataset cannot be assumed. The data similarity and $\Delta\mathbf{L}$ for these two subsets of the data are shown in Figure 3.3, with the mean and variance of the data similarity given on the right side of Table 3.2. The differences between the results for odd and even years are found to be large, indicating either $k = 4$ or $k = 6$ having the smallest variance. Also for $k = 5$ differences between the two sets of years are large. These ambiguous results raise the question whether half the dataset is of sufficient length to draw reliable conclusions about the clustering results. This also means that non-stationarity of the regimes due to e.g. climate change is difficult to study accurately using clustering methods applied to reanalysis datasets.

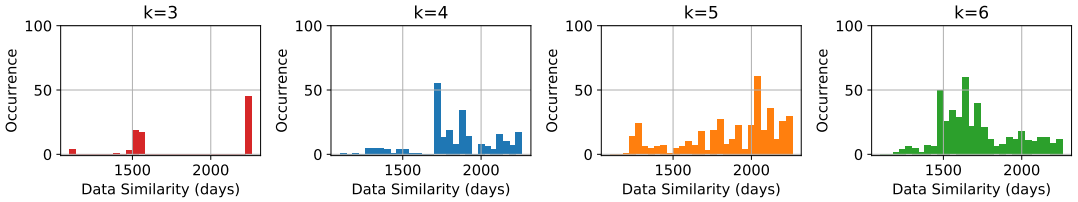
We refrain from drawing definite conclusions about the optimal number of clusters k from the above discussion on consistency, as there is some debate about its suitability for this purpose (Philipp et al., 2007). Instead, we use the AIC and BIC to identify the optimal number of regimes. The AIC is used when considering the PC results, as it is expected to give better results in that case, as discussed in Section 3.2.1. In Figure 3.4a the AIC is shown for using 5, 10, 15 and 20 PCs to identify the circulation regimes. A minimum at $k = 4$ is found when 20 PCs are used, although the AIC is also small for $k = 3$ and $k = 5$. For lower numbers of PCs the optimal number is found to be lower, while a higher number of PCs leads to a higher optimum for k. This is to be expected because the use of a limited number of PCs means that some variability of the original data is neglected. This



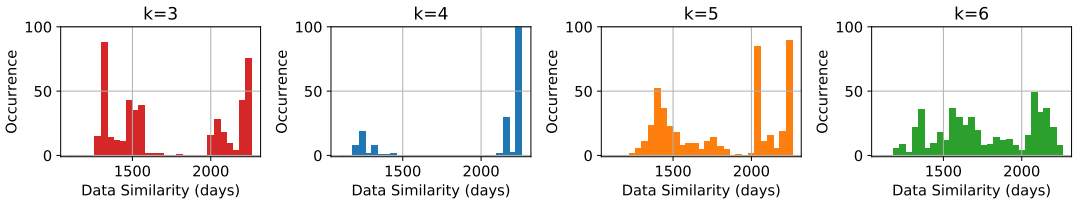
(a) The difference of the clustering functional $\Delta\mathbf{L}$ for the odd years.



(b) The difference of the clustering functional $\Delta\mathbf{L}$ for the even years.

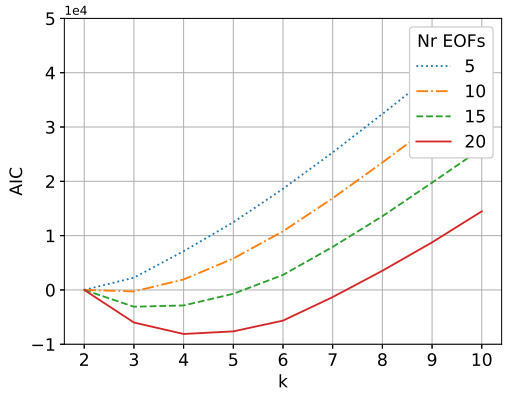


(c) The data similarity for the odd years.

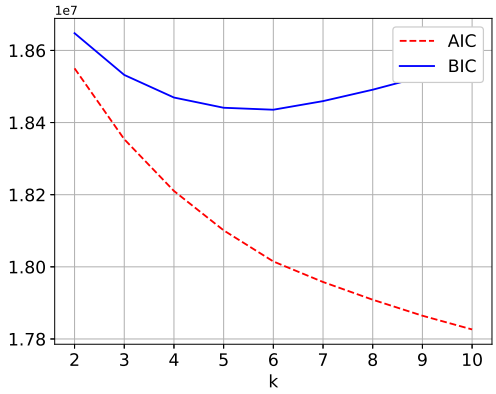


(d) The data similarity for the even years.

Figure 3.3: Histograms for the difference of the clustering functional with its minimum value $\Delta\mathbf{L} = \mathbf{L}_{\text{run}} - \mathbf{L}_{\text{min}}$ and the data similarity with respect to the optimal (minimal $\mathbf{L}$) result using either the odd (a,c) or even (b,d) years of the gridpoint data for $k = 3, ..., 6$.



(a) The AIC for different numbers of PCs.



(b) The AIC and BIC for the gridpoint data.

Figure 3.4: Information criteria for both the PC and gridpoint datasets for the range $k = 2, ..., 10$. Both the AIC and BIC, as given in Section 3.2.1, are shown for the gridpoint data. For the PC data only the AIC is shown.

loss of variability is larger when less PCs are used and as a consequence fewer clusters are needed to account for the variability of the PC data. The BIC has its minimum at $k = 2$ for every number of PCs considered, indicating the penalty term for the number of parameters is indeed too strong for the PC data (Section 3.2.1). Based on the AIC we conclude that $k = 4$ is indeed the optimal number when using 20 PCs, which corresponds with results from literature. However, for other numbers of PCs the optimal k according to the AIC is different, meaning this conclusion is not unambiguous.

The optimal number of regimes identified by an information criterion when using the gridpoint dataset is not the same. The BIC is more suitable for the gridpoint data than the AIC due to the dependence of the penalty term on the sample size, making it stronger than the penalty term of the AIC (Section 3.2.1). In Figure 3.4b both the AIC and BIC are shown for the gridpoint data. The BIC points towards an optimum of $k = 6$. The AIC does not show a minimum in the range considered as the penalty term is not strong enough for the high dimensional gridpoint data. Therefore, we base our decision on the optimal number of regimes on the BIC and find $k = 6$ to be optimal.

3.2.3 Circulation Regime Patterns and their Dynamics

The regimes that are obtained by using 10 or more PCs and the gridpoint data are essentially the same for the each k. Similarly, the occurrence rate and transition probabilities of the regimes do

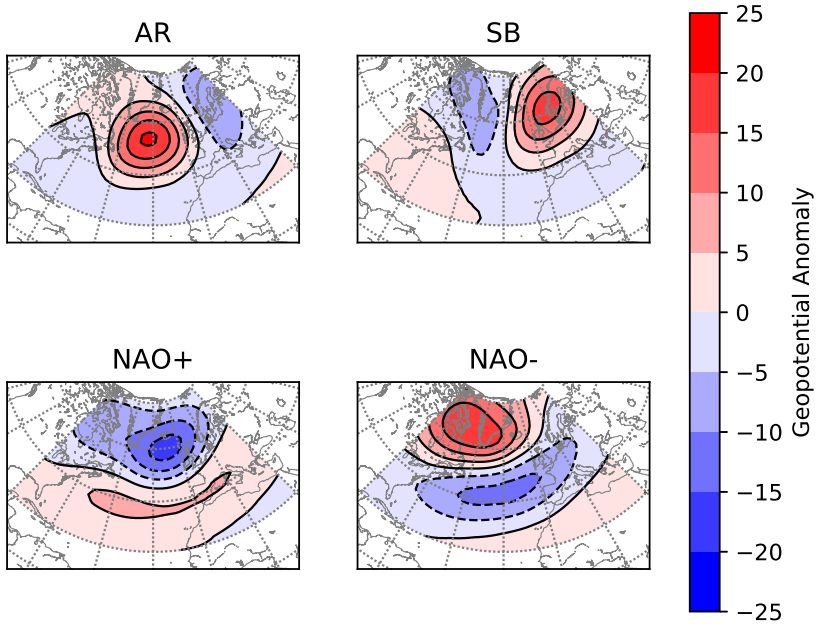


Figure 3.5: The clustering result of the standard k-means algorithm applied to the gridpoint data for $k = 4$.

not differ substantially. In Figure 3.5 the four regimes known from literature (e.g. Hannachi et al., 2012; Straus et al., 2007) are shown as obtained by applying k-means clustering to the gridpoint data for $k = 4$. They are the Atlantic Ridge (AR), Scandinavian Blocking (SB) and the two phases of the NAO. The transition probability of a regime to itself (or daily re-occurrence rate) and overall occurrence rate of these regimes can be found in Table 3.3. The positive phase of the NAO is the most frequently occurring regime, followed by SB. The high occurrence of the NAO+ regime may reflect the fact that it is the only regime associated with a northern low pressure area. Both phases of the NAO are found to be most persistent, i.e. transition to themselves with the highest probability, while the AR exhibits the least persistence. We note that the occurrence rates obtained are similar to those found in literature despite not using a seasonally varying background state.

In Figure 3.6 the regimes found using k-means clustering on the gridpoint data for $k = 6$ are shown. The first four regimes are in essence the same as those found for $k = 4$ (compare with Figure 3.5). Small differences occur in the location of the maximum high or low z500 area for the AR and NAO+. The two additional regimes have a low pressure area either in the central Atlantic or over Scandinavia. The first thus can be identified as the opposite phase of the AR with a pattern correlation of -0.57 and we refer to it as AR−, while the original regime is denoted by AR+. Similarly

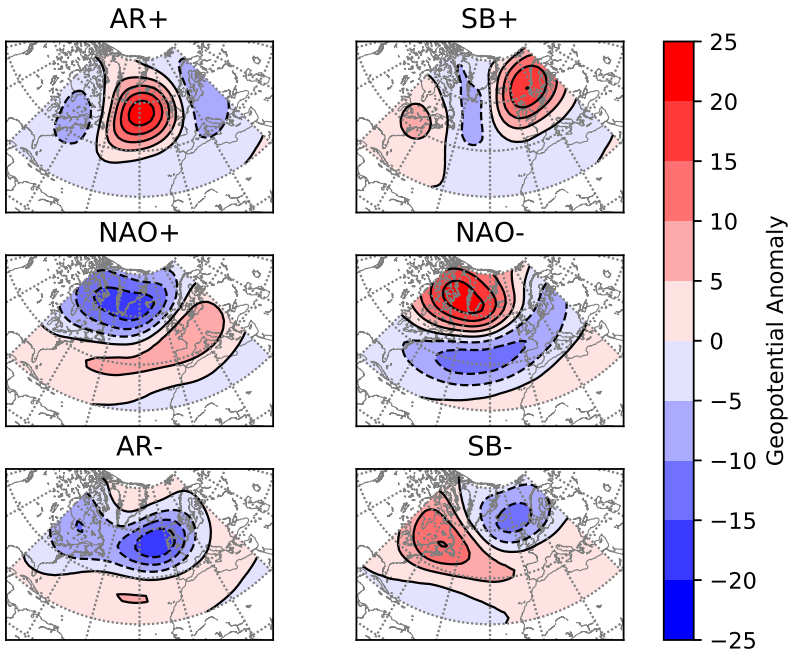


Figure 3.6: The clustering result of the standard k-means algorithm applied to the gridpoint data for $k = 6$.

Table 3.3: The values of the occurrence rate, self-transition probability, e-folding time and average regime duration for both $k = 4$ and $k = 6$.

$k = 4$	AR	SB	NAO+	NAO−		
Occurrence (%)	21.3	26.8	31.5	20.4		
Self-Transition Probability	0.756	0.792	0.850	0.849		
e-Folding Time (days)	3.6	4.3	6.2	6.1		
Average Duration (days)	4.1	4.8	6.7	6.6		
$k = 6$	AR+	SB+	NAO+	NAO−	AR−	SB−
Occurrence (%)	15.6	19.6	16.9	15.5	16.3	16.1
Self-Transition Probability	0.712	0.748	0.751	0.847	0.787	0.730
e-Folding Time (days)	2.9	3.4	3.5	6.0	4.2	3.2
Average Duration (days)	3.5	4.0	4.0	6.5	4.7	3.7

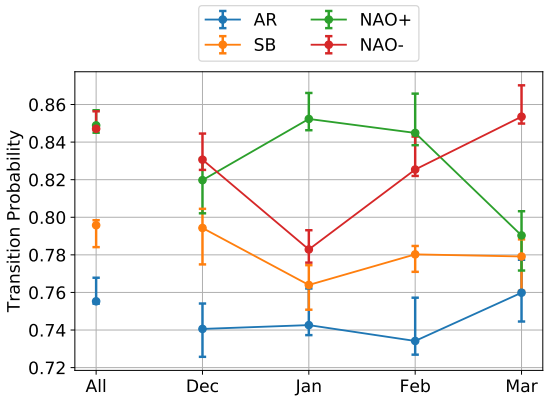
we refer to the second additional regime as SB−, as it represents the opposite phase of the SB regime (from now on denoted by SB+), where the pattern correlation -0.49 is slightly lower. Note that the pattern correlation of the two phases of the NAO is higher for $k = 6$, with the value being -0.93, versus -0.59 when using $k = 4$. The use of six clusters thus introduces a pleasing symmetry in the found regimes, with an equal number of regimes having a high and low z500 area in the north of the domain.

The occurrence rate and self-transition probability of the six regimes show different behaviour than found for $k = 4$, as can be seen in Table 3.3. Instead of the NAO+ regime, the SB+ regime is found to be the most frequently occurring regime. The NAO+ is the second ranked regime in occurrence, albeit with a small but significant difference relative to SB+. The other four regimes show similar occurrence rates. When looking at the self-transition probabilities, NAO− remains the most persistent, with exactly the same probability. The NAO+ regime however does lose some of its persistence, reducing its self-transition probability to a rate similar to that of SB+. AR− is found to be the second most persistent regime and the AR+ regime remains the least persistent. In Table 3.3 the e-folding time and average regime duration computed using the self-transition probability are also given for reference (see Appendix A for details).

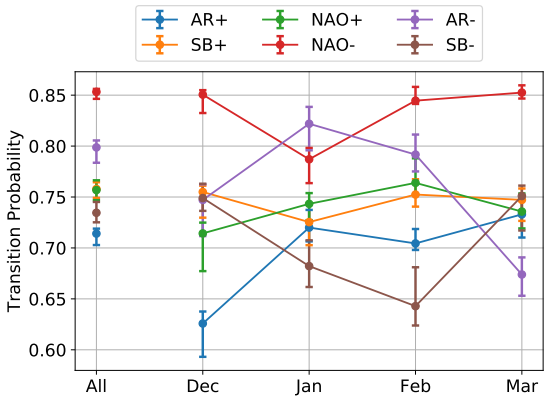
Since a constant background state (fixed climatology) is assumed instead of a seasonal varying one, there is variability in the occurrence and transition probabilities throughout the season. As discussed, the regimes found do not change significantly if a seasonal cycle is subtracted. However, there may be slight changes in the occurrence and persistence of some of the regimes. In Figure 3.7 the occurrence rates and self-transition probabilities throughout the winter months are shown for both $k = 4$ and $k = 6$. Some regimes, like AR for $k = 4$, show consistent behaviour throughout all months, while others, like AR− for $k = 6$, exhibit significant changes in occurrence and self-transition probability throughout winter. Note that for most regimes the occurrence and self-transition probability co-vary. This variation could reflect the effect of the seasonal cycle in the data, as well as intrinsic variability in the regime behaviour.

3.2.4 Robustness Verification using Spectral Clustering

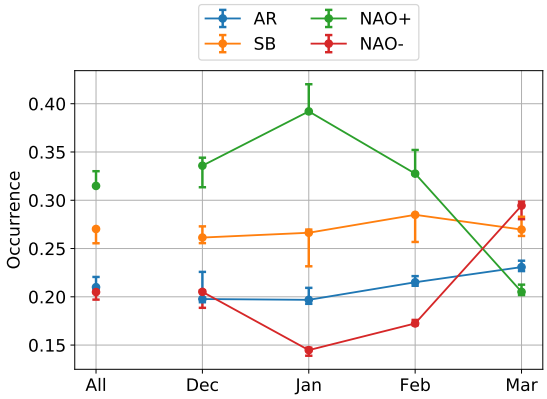
One of the drawbacks of k-means clustering is that it always converges to a set of clusters, independent of what the probability density function (pdf) of the system looks like. It converges to a local minimum and only in very simple cases it might be possible to tell whether this is a global minimum. Another method of clustering is spectral clustering. This is a graph based clustering



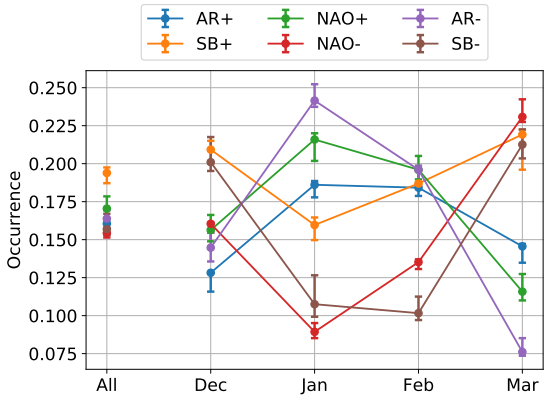
(a) The self-transition probabilities for $k = 4$.



(b) The self-transition probabilities for $k = 6$.



(c) The occurrence rates for $k = 4$.



(d) The occurrence rates for $k = 6$.

Figure 3.7: The occurrence rates and self-transition probabilities of the different regimes for $k = 4$ and $k = 6$ for the clustering results throughout the winter months. The error bars indicate the maximum and minimum value of occurrence/transition probability for clustering results with a slightly smaller $\mathbf{L}$ (the used bound on $\Delta\mathbf{L}$ is 0.002).

method where the associated function that needs to be optimized is convex. The advantage is that it always yields the same clusters (within accuracy bounds). Another drawback of k-means is that it makes (implicit) assumptions about the shape of the cluster by the choice of the norm. In Figure 2.2 (Section 2.3) an example of a dataset for which k-means clustering fails is shown. For this dataset spectral clustering does identify the correct clusters. Spectral clustering thus does not suffer from these two drawbacks of k-means clustering.

To verify the robustness of the regimes identified using k-means clustering we apply spectral clustering to the gridpoint data. This can help to indicate whether k-means clustering is indeed suitable for regime identification, i.e. whether the assumption of spherical clusters is realistic. First,

the basics of spectral clustering are explained, where we follow Von Luxburg (2007) and focus on the relevant aspects for the application to atmospheric circulation regimes. Next we turn to a brief discussion of the results when applied to atmospheric circulation data.

Where one considers individual datapoints in the context of k-means clustering, one focuses on a graph representing the batch of data for spectral clustering. The first step thus is to create this so-called similarity graph, which can be represented by a graph Laplacian matrix. The graph Laplacian then is used for the spectral clustering. The spectral clustering algorithm itself consists of two steps:

1. Compute the eigenvalues and eigenvectors of the graph Laplacian.
2. Apply a k-means clustering algorithm to the rows of the matrix with the first k eigenvectors as columns, where k is the number of clusters.

The graph Laplacian fully determines the outcome of the described spectral clustering approach. Therefore its definition needs to be considered thoroughly, as it can strongly impact the obtained clusters.

To obtain the graph Laplacian we start by considering the similarity graph. In such a graph each vertex represents a datapoint. Two vertices are connected if the similarity between the corresponding datapoints is positive or larger than a certain threshold. The edge between two connected vertices is weighted by the value of this similarity. There are different choices possible for the definition of the similarity between two datapoints, the clearest of which are related to the distance between the two points. The smaller the distance, the more similar two points are. For the spectral clustering of circulation regimes a fully connected graph is used. The similarity between datapoints x_i and x_j we consider is given by

$$s(x_i, x_j) = \exp\Big(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\Big). \tag{3.4}$$

Here σ affects how quickly the similarity falls off with increasing distance. It is a tuning parameter that can be used to optimise the obtained clusters. Alternative graphs are for example the ϵ-neighbourhood graph, where all points with distance smaller than ϵ are connected, or the k-nearest neighbour graph, connecting each vertex to its k-nearest neighbours. The choice for the level of connectedness is a direct result of the nature of the considered atmospheric data, having a strong similarity between data of subsequent timesteps. This means that a graph that is not fully connected will predominantly connect the data in time and thus limit the possibility of switching regimes.

The similarity graph can be represented by an adjacency matrix $W = (w_{ij})_{i,j=1,\dots,n}$, where $w_{ij} \neq 0$ gives the weight of the edge when two vertices are connected and $w_{ij} = 0$ if they are

not connected. For an undirected graph, where the edges do not carry a direction, this matrix is symmetric. In our case the weights are given by the similarity as defined in Equation (3.4), i.e. $w_{ij} = s(x_i, x_j)$. The other matrix of relevance to compute the graph Laplacian is the degree matrix D, which has the degrees of the vertices, given by

$$d_i = \sum_{j=1}^{n} w_{ij}, \tag{3.5}$$

on the diagonal. With the adjacency matrix W and degree matrix D, one can compute the graph Laplacian. There are several versions of the graph Laplacian. Here three, which all have been used in spectral clustering, are discussed. Firstly, there is the unnormalised version:

$$L = D - W. \tag{3.6}$$

Secondly, there are symmetric and non-symmetric normalised versions of the graph Laplacian:

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}, \qquad L_{rw} = D^{-1} L = I - D^{-1} W, \tag{3.7}$$

The non-symmetric normalised graph Laplacian L_{rw} is closely related to a random walk (rw). The decision on which of the three matrices to use for spectral clustering is predominantly determined by computational arguments, as they all yield the same result.

In the definition of the similarity between datapoints the choice of the tuning parameter σ is very important. As shown in Figure 3.8 for a toy example of two clusters, selecting the wrong σ can result in identifying incorrect clusters. When the data is strongly separated in different clusters, making a sensible choice of σ can be done by considering the eigenvalues. When they are all close to zero σ is too small (Figure 3.8d), when nearly none are close to zero σ is too large (Figure 3.8f). For data that is not as clearly separated as in this toy example the choice is not as clear. A possible solution is to make σ locally dependent, i.e. self-tuning (Zelnik-Manor and Perona, 2005). That is, for $s(x_i, x_j)$ set

$$2\sigma^2 = \|x_i - x_M\| \cdot \|x_j - x_M\|, \tag{3.8}$$

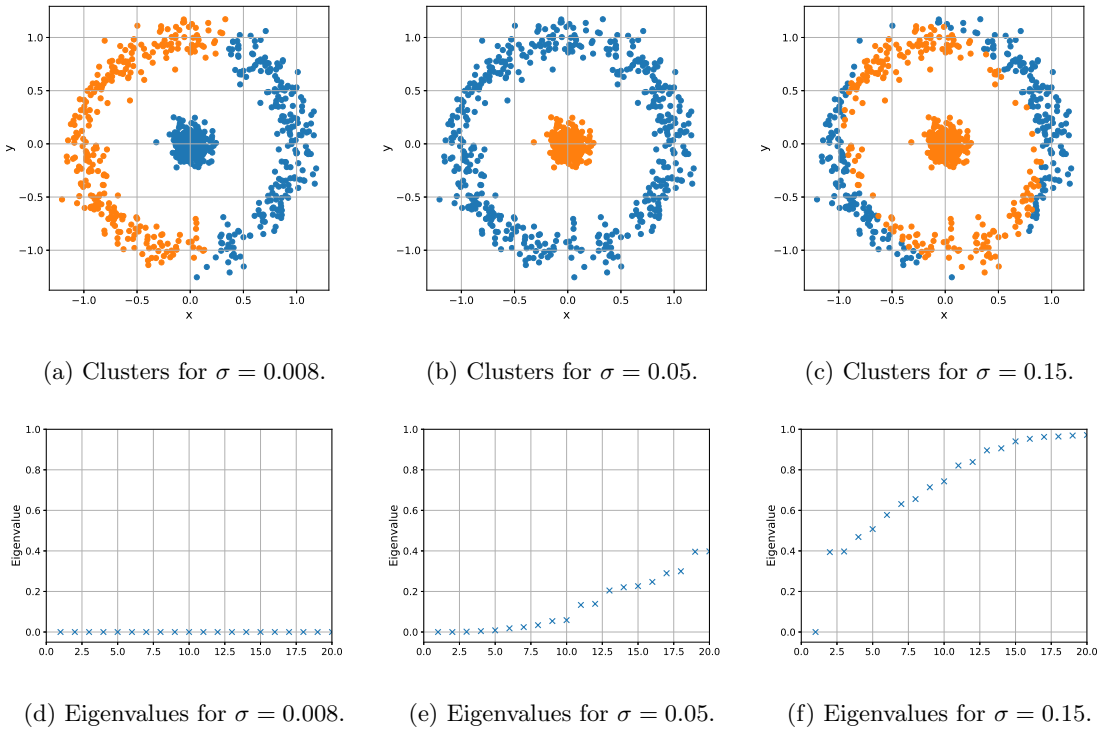
where x_M is the M-th neighbour (in distance) to $x_{i,j}$. This introduces a choice of M instead of σ. However, the eigenvalues are more consistent for different choices of M compared to σ and therefore this is the approach that is used here to study the identification of atmospheric circulation regimes using spectral clustering.

After considering the spectrum and the AIC and BIC, we apply the spectral clustering algorithm with $M = 7$ to the gridpoint data. The results are found to not be very sensitive to the exact value of M. As for the standard k-means approach, we consider the AIC and BIC to identify the optimal

number of regimes to be considered. Both are shown in Figure 3.9, with the BIC showing a clear minimum for $k = 5$. This is between the optimal values of $k = 4$ for PC data and $k = 6$ for gridpoint data identified in Section 3.2.2. The corresponding regimes, computed as the mean over all data assigned to each of the clusters, are shown in Figure 3.10. The AR+, SB+, NAO− and SB− are very similar to those identified for $k = 6$ (Figure 3.6), whereas the fifth regime (here referred to as NAO+) lies between the $k = 6$ NAO+ and AR− regimes.

However, we cannot interpret the regimes in the same way as for k-means clustering as the shape of the clusters is unknown. For example, the means of the data assigned to each of the two clusters in the example shown in Figure 3.8 are similar. This makes it difficult to interpret the cluster centres in the same way as in k-means clustering, as it is not clear how representative they are for the data in each cluster. The physical relevance of the regime representation thus is difficult to establish for spectral clustering and we refrain from considering these regimes for $k = 5$ further.

Considering the cluster centres of the spectral clustering results for $k = 4$ and $k = 6$, which are



(a) Clusters for $\sigma = 0.008$. (b) Clusters for $\sigma = 0.05$. (c) Clusters for $\sigma = 0.15$.

(d) Eigenvalues for $\sigma = 0.008$. (e) Eigenvalues for $\sigma = 0.05$. (f) Eigenvalues for $\sigma = 0.15$.

Figure 3.8: A toy example applying spectral clustering using different values of σ to define the similarity between datapoints. The obtained clusters are shown in (a-c) and the corresponding eigenvalues in (d-f).

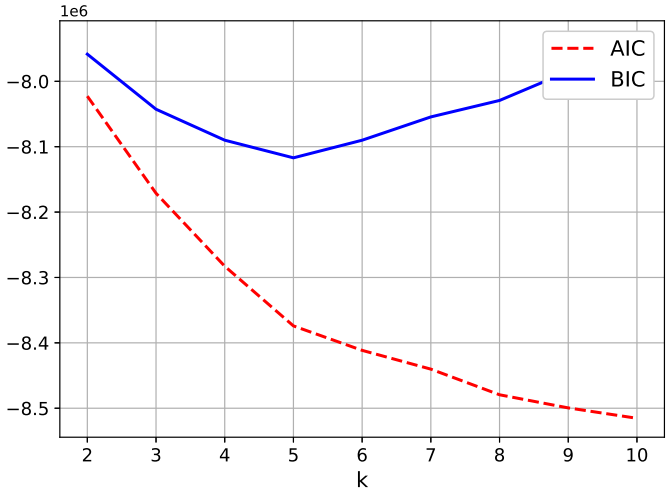


Figure 3.9: The AIC and BIC for spectral clustering of gridpoint data.

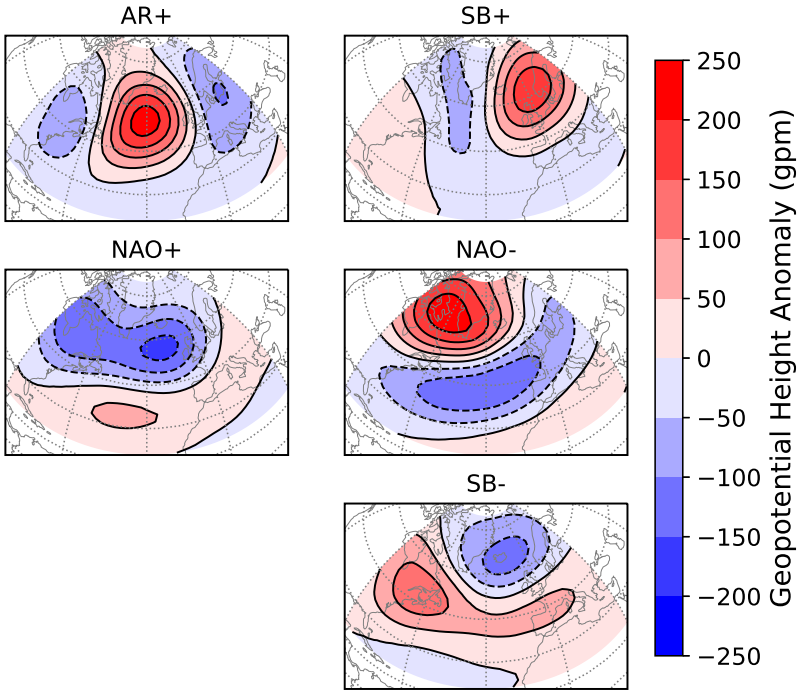


Figure 3.10: The regime centres as obtained using spectral clustering for $k = 5$.

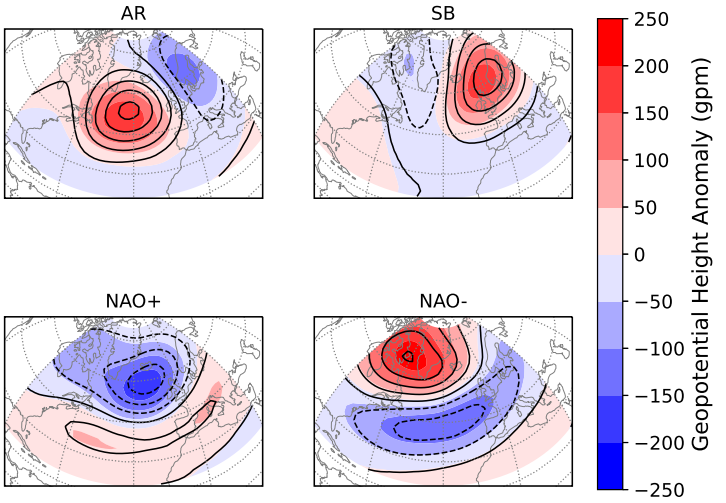
shown in Figure 3.11, we note that they do not differ strongly from the regimes identified using standard k-means clustering. This increases the confidence in the k-means clustering results, as similar regimes can be obtained using a different approach. It also indicates that the data belonging to the regimes is relatively spherically distributed, with spectral clustering mainly assigning data at the edge to a different regime. Both these points indicate that the regimes identified using k-means clustering are robust.

3.3 Persistent Regimes

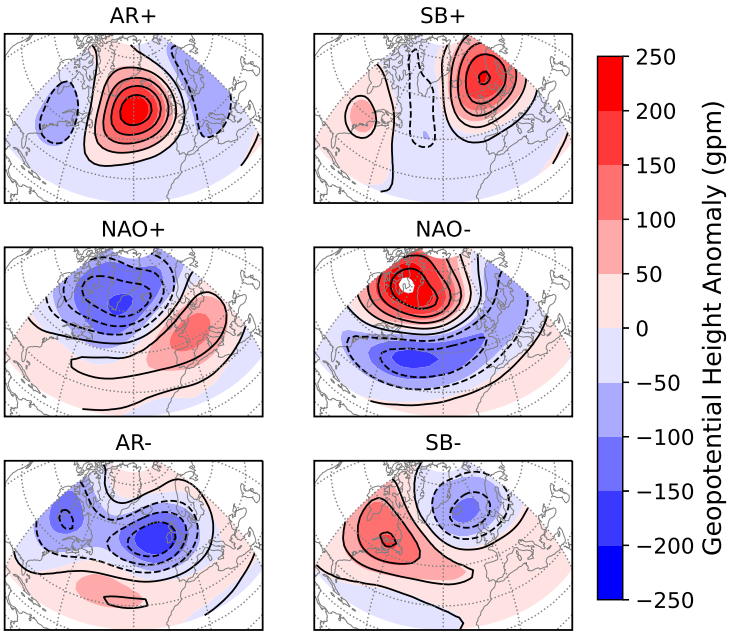
The previous section discussed the robustness of the circulation regimes, comparing the common used approach by projecting onto EOFs with the use of gridpoint data. The second aspect of the data filtering that is often applied prior to clustering the data is a low-pass filter. The aim of this is to filter out high-frequency dynamics and focus on the slower, low-frequency dynamics represented by the regimes. In this section we discuss an alternative approach that does not alter the data itself, but instead incorporates a persistence constraint in the clustering approach. This regularisation is based on the methods discussed in Section 2.2.1. We start with a detailed discussion of the regularised clustering method (Section 3.3.1), followed by an application to the Lorenz 63 system to show its effect (Section 3.3.2). Then the method is applied to identify the circulation regimes and their dynamics, comparing with the results obtained by first applying a 5- or 10-day low-pass filter to the data followed by a standard k-means clustering algorithm (Section 3.3.3).

3.3.1 Time-regularised k-means Clustering

The standard k-means clustering algorithm has been described in Chapter 2, including a brief discussion of regularisation approaches. Here, we introduce a time-regularised clustering approach in more detail, repeating some of the aspects mentioned in Chapter 2 for completeness. Given a dataset $\{x_t\}_{t\leq T} \in \mathbb{R}^m$, with t time, m the dimension of the data and T the span of the data in time, the aim of any clustering method is to find a set of k cluster centres that accurately describe the dataset based on some measure. Let $\Theta = (\theta_1, ..., \theta_k)$ be the set of parameters describing the k cluster centres. Here Θ represents the different circulation regimes for z500 anomaly data $\{x_t\}_{t\leq T}$. To assess how well the cluster centres represent the data, a model distance functional $g(x_t, \theta_i)$, giving the distance between a cluster centre and a datapoint, is required. We use the L_2-distance weighted by the cosine of latitude (Chung and Nigam, 1999) as for the standard k-means approach.



(a) $k = 4$



(b) $k = 6$

Figure 3.11: A comparison between the sets of regimes obtained using spectral clustering (colours) and standard k-means clustering (contours) for either four or six regimes.

In addition we consider the affiliation vector $\Gamma = (\gamma_1(t), ..., \gamma_k(t))$, which indicates the weight of a certain cluster at some point in time. In practice $\gamma_i(t)$ is nearly always either zero or one, indicating to which cluster that point belongs. This is because a linear optimization problem always has an optimal solution on the boundary of the admissable set (Cottle and Thapa, 2017). For this reason the affiliation vector is in general not considered when k-means clustering is discussed (Section 2.1). Here we do consider this vector because it allows for the incorporation of persistence in the clustering procedure, following the discussion on regularisation in Section 2.2.1.

The task of identifying the atmospheric circulation regimes best representing the data means one has to find the optimal parameters for the cluster centres Θ and the affiliations of the data Γ. This is done by minimizing the averaged clustering functional (Franzke et al., 2009)

$$\mathbf{L}(\Theta, \Gamma) = \sum_{t=0}^{T} \sum_{i=1}^{k} \gamma_i(t) g(x_t, \theta_i), \tag{3.9}$$

subject to

$$\sum_{i=1}^{k} \gamma_i(t) = 1, \quad \forall t \in [0, T], \qquad \gamma_i(t) \geq 0, \quad \forall t \in [0, T], i = 1, ..., k. \tag{3.10}$$

This is what the k-means procedure explained in Section 2.1 is doing implicitly, where $\gamma_i(t)$ is assumed to be zero or one. Finding the minimum of this functional minimizes the within-cluster variance, as $\mathbf{L}$ is a measure of the distance between the cluster centres and the datapoints assigned to it. Because the within-cluster variance is minimized simultaneously for all clusters, the distance between datapoints assigned to different clusters becomes large. In other words; the between-cluster variance is maximized.

This clustering functional does not yet incorporate any persistence; an arbitrary reshuffling of the data leads to exactly the same result. To include persistence in the clustering method we add a constraint on Γ that limits the number of transitions between regimes that is allowed (de Wiljes et al., 2014). This constraint on the number of transitions between regimes, or switches, that are permitted throughout the whole time-series is:

$$\sum_{i=1}^{k} \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq C, \quad i = 1, ..., k, \tag{3.11}$$

for some constant C. The value of C gives twice the number of switches allowed (for a transition $\theta_1 \to \theta_2$ both the switch out of θ_1 and the one into θ_2 are counted), so e.g. an average cluster length of five days corresponds to $C = 2 \times \#\text{days}/5 \approx 1900$. In Table 3.4 the average regime duration corresponding to several values of C is given. Note that for a first-order Markov process, which is a relatively accurate assumption for the unconstrained regime dynamics, the average regime duration T_{av} can be related to the e-folding time scale T_e according to $T_{av} = 1/(1-p)$ and $T_e = -1/\log(p)$,

Table 3.4: The value of C with corresponding average regime duration.

C	600	800	1000	1200	1400	1600	1800	2000	2200
Average Duration (days)	15.8	11.8	9.5	7.9	6.8	5.9	5.3	4.7	4.3

where p is the self-transition probability (a discussion of this is given in Appendix A). Hence T_e is slightly shorter than T_{av} (e.g. 4.8 days for an average duration of 5.3 days), and they approach each other as $p \to 1$. The rationale behind the introduced regularisation constraint is that in a chaotic atmospheric circulation not every datapoint can be straightforwardly assigned to a cluster. Some datapoints can be in-between clusters or outliers, e.g. transitioning between clusters or extreme events. K-means clustering assigns these points to the nearest cluster (by distance), while it can be more sensible to assign them to the same cluster as their neighbours if the distance to that cluster is also quite small. This is exactly what the constraint in Equation (3.11) is doing for reasonable values of C.

The minimization of the clustering functional $\mathbf{L}$ taking into account the persistence regularisation is done in two steps which are iterated until convergence:

1. For fixed Θ, minimize $\mathbf{L}$ over all possible values of Γ.

2. For fixed Γ, minimize $\mathbf{L}$ over all possible values of Θ.

The first part is done by linear programming using the Gurobi package for python (Gurobi Optimization LLC, 2019). The second part is done by k-means clustering. The computation is terminated when the difference between consecutive $\mathbf{L}$ becomes smaller than a set tolerance. As with standard k-means clustering, the regularised version is not convex and converges to a local, not global, minimum (Metzner et al., 2012). Therefore, the algorithm including the persistence constraint is run 100 times with different initial conditions; the reduced number of runs compared to standard k-means clustering is due to increased computation time by the incorporation of linear programming, which is similar for the PC and gridpoint data. The final result is chosen to be the one with the smallest clustering functional $\mathbf{L}$.

To make the effect of the incorporation of this constraint on the final clustering result more insightful a simple toy model is presented. Consider a system of three 2D clusters which are normally distributed around their respective centres $\mathbf{x}_i$, each with a different variance v_i; $\mathbf{x}_1 = (0.2, 0.4)$, $v_1 = 0.012$, $\mathbf{x}_2 = (0.8, 0.3)$, $v_2 = 0.018$ and $\mathbf{x}_3 = (0.7, 0.9)$, $v_3 = 0.01$. They transition into each

other according to a persistent transition matrix,

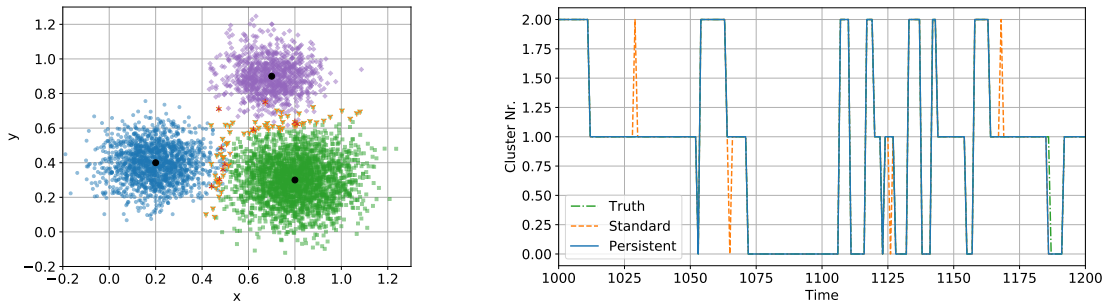
$$T = \begin{pmatrix} 0.92 & 0.02 & 0.06 \\ 0.04 & 0.95 & 0.01 \\ 0.02 & 0.1 & 0.88 \end{pmatrix}, \tag{3.12}$$

meaning there is a high probability of a cluster transitioning to itself. A time series of 5000 steps for this toy example is generated. As the focus in this section is not on identifying the number of regimes, we use the true number of $k = 3$. Both a standard and regularised k-means clustering algorithm are run, where a constraint of $C = 252$ is used (obtained by trial and error).

The clusters are shown in Figure 3.12a. When applying the standard (unconstrained) k-means algorithm the data that are assigned to the wrong cluster are located in-between the different clusters. In the assignment of the data to the clusters this leads to sudden jumps into one cluster and directly back to the original cluster, as can be seen in Figure 3.12b, which leads to the identification of a too short persistence. When the persistence constraint is incorporated many of the wrongly assigned datapoints are now assigned to the correct cluster. Furthermore we see that the short transitions into and directly out of a cluster are removed. Thus the persistence obtained using the regularised algorithm is closer to that of the real system. Note that despite the algorithm not enforcing a hard regime assignment, this is what is observed in the toy example. This is because the algorithm converges to solutions on the boundary of the admissible set (Metzner et al., 2012). The same is true when the more complex Lorenz 63 dynamics (Lorenz, 1963), discussed in the following section to identify some of the limitations of the method, and circulation regimes are considered.

3.3.2 A Simple Example: Lorenz 63

The Lorenz 63 system (Lorenz, 1963) is one of the most well-studied low-order systems derived from geophysical fluid dynamics. It exhibits regime behaviour, as well as chaos, making it a suitable system to test the performance of clustering methods (e.g. Hannachi and O'Neill, 2001), but also to use as an analogue for more complex systems in climate. For example Corti et al. (1999) used this regime feature to interpret the effect of climate change on atmospheric circulation regimes. Here we apply the two clustering approaches, standard and time-regularised k-means, to different realisations of the Lorenz 63 system. This allows for testing the accuracy and reliability of the methods used.



(a) The datapoints wrongly assigned by standard k-means clustering are shown in orange (Y-shape) and the points that are still wrongly assigned when the persistence constraint is incorporated are shown in red (inverted Y-shape).

(b) A part of the transition sequence which indicates in which cluster the data belongs. The truth is shown in green (dash-dotted, visible on the very right), in orange (dashed) is the result for standard k-means and in blue (solid) the result when the persistence constraint is incorporated. Note that the lines overlap most of the time.

Figure 3.12: The toy model (a) showing the clusters in blue, green and purple and (b) the assignment of the data to the different clusters in time.

The equations of the Lorenz 63 system are (Lorenz, 1963):

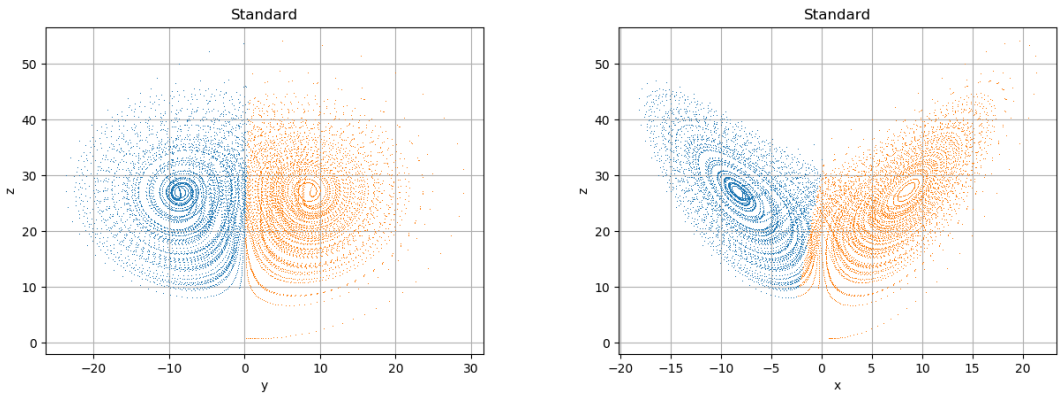
$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -\sigma x + \sigma y, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -xz + rx - y, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - bz. \end{aligned} \tag{3.13}$$

Here σ, r and b are parameters. The standard values used by Lorenz are $\sigma = 10$, $r = 28$ and $b = 8/3$ and give the well-known butterfly. We stick to these values here. The system is integrated using a standard Euler scheme with time steps of 10^{-2} for 10^4 steps.

We apply the standard k-means clustering algorithm to several realisations of the Lorenz 63 system (different initial conditions) projected onto either the y-z-plane or the x-z-plane. Mostly the clusters found in the y-z-plane correspond to the "correct" clusters, being the two wings of the butterfly separated by the line $y = 0$, as can be seen in Figure 3.13a. On the other hand, when the k-means algorithm is applied to the corresponding x-z-data the result is not as good, since it fails in identifying the two wings of the butterfly as the clusters, now separated by the $x = 0$ line (Figure 3.13b). The standard k-means algorithm thus fails to correctly identify the two wings of the butterfly in the Lorenz 63 system when only x and z data is taken into account. This could reflect in the results for atmospheric circulation regimes, as it is impossible to take into account all relevant dynamical factors.

To improve the result of the k-means algorithm for the x-z-plane we explore two methods that enforce consistency in time of the clustering result. The first method is to apply a low-pass filter to the data getting rid of high-frequency oscillations. The second method is to include a persistence constraint in the k-means algorithm itself as discussed in Section 3.3.1. For both approaches a parameter has to be chosen. For the time-filter this is the cut-off frequency and for the time-regularised algorithm this is the value of the persistence constraint. Here we set the cut-off frequency for the time-filtering to 150 time steps, being the equivalent of a 10-day filter used for atmospheric circulation data. The value of the persistence constraint is determined using an information criterion (see discussion in Section 3.2.1). Both the AIC and BIC are explored, where the BIC is found to yield better results. When the two criteria differ in the location of their minimum, the clusters for the BIC-minimum are closer to the best result as identified by visual inspection than those for the AIC-minimum. Note that the information criterion approach can also be used to determine the optimal cut-off frequency for the time-filtering.

For the same realisation as in Figure 3.13 the results for the time-filtered data and clustering using a time-regularisation are shown in Figure 3.14. Both results show a clear improvement with respect to the standard approach in Figure 3.13b. The result for the time-filtered data shows a slightly different assignment of data to clusters for the transition trajectories, but other than that the results of both methods are as desired. We note that for this realisation of the model the BIC shows a clearly identifiable minimum, which is not always the case.



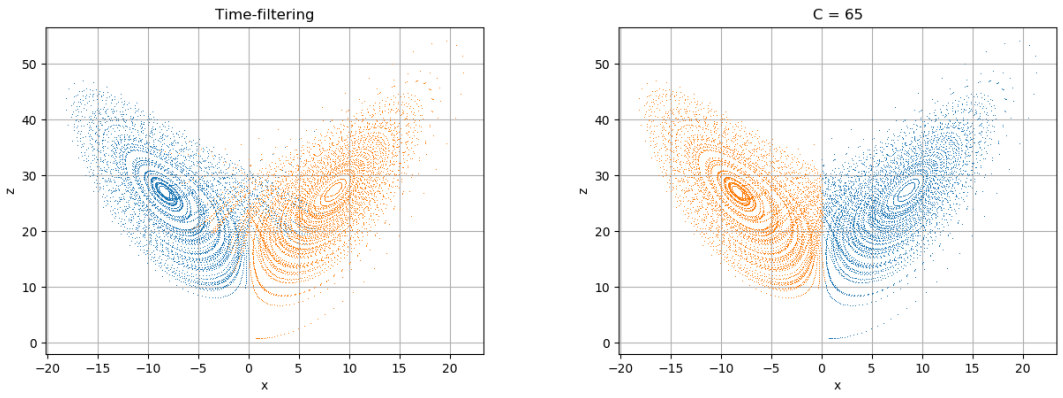
(a) y-z-plane.

(b) x-z-plane.

Figure 3.13: Clustering of the Lorenz 63 system in either the y-z- or x-z-plane. The dots show the datapoints of the simulation of the model and the colours indicate to which cluster these points are assigned.

The discussed results show a realisation of the model in which both methods work nicely. However, this is not the case for every realisation of the Lorenz 63 model. Already when applying k-means clustering to the y-z data the correct clusters are not always identified, as can be seen in Figure 3.15a. This result likely improves if more data is included, but as a limited amount of data is one of the difficulties of real world clustering it is important to note this limitation. Furthermore the BIC does not always point towards the correct result, which can be seen in Figure 3.15b. By looking at the clusters for different values of the constraint it is possible to identify a better value, but as this is impossible for the high-dimensional atmospheric data in which the circulation regimes are identified, this is not a desirable option. We note that also for the time-filtered data the result is not always as good as shown in Figure 3.14a, although in general it is slightly more robust than the results for the time-regularised approach.

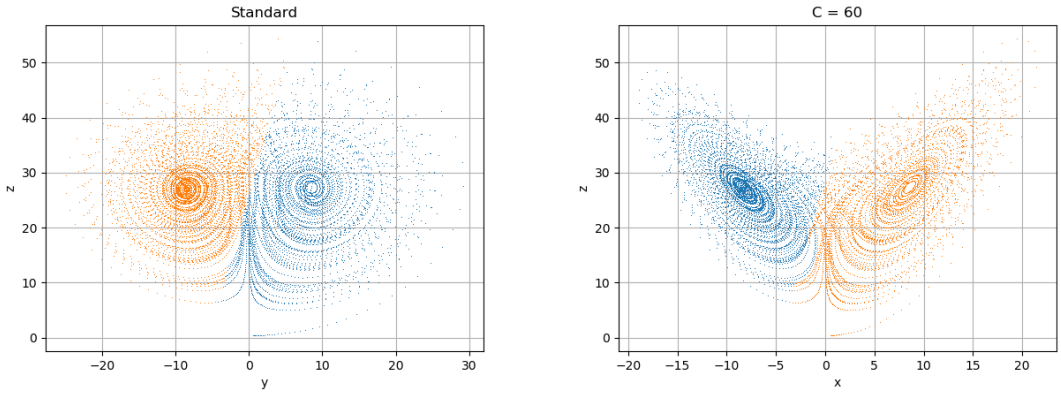
Applying the clustering methods used to identify circulation regimes in atmospheric data to the Lorenz 63 system teaches us to be careful in relying too much on the outcome of the algorithm. Even for the simple Lorenz 63 system the clustering algorithm does not always identify the correct clusters. For the even more complex atmospheric data in which the circulation regimes are identified, this is likely an even larger difficulty. This does not mean the result is not useful, but it is important to be aware of the limitations of the method. The data close to the cluster centres is always assigned correctly, but there is a substantial uncertainty in assigning the data further away from the cluster centres. Thus it is important to be careful when applying k-means clustering and not blindly trust



(a) Applying a low-pass filter to the data for a cut-off frequency of 150 time steps.

(b) Incorporating a persistence constraint in the clustering algorithm.

Figure 3.14: Clustering of the Lorenz 63 system in the x-z-plane using different methods to enforce persistence.



(a) Standard k-means applied to the y-z-data. (b) Incorporating a persistence constraint in the clustering algorithm.

Figure 3.15: The clustering approaches discussed do not always identify the correct clusters.

the result, especially as the method assigns every datapoint to a cluster even if it actually is in-between different clusters. Both the application of a low-pass filter, as well as the incorporation of a persistence constraint improve this aspect of the k-means result.

3.3.3 Persistent Regime Dynamics

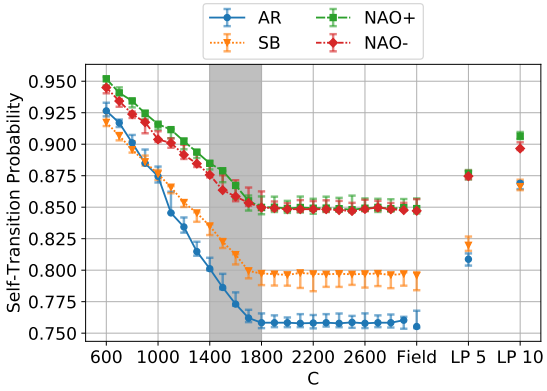
In Section 3.3.1 a time-regularised, or constrained, clustering method to enforce persistence of the atmospheric circulation regimes has been discussed. An alternative method to obtain more persistent regime dynamics is the standard approach of applying a time-filter to the data. The regime patterns found using these two methods do not differ substantially from those found and discussed in Section 3.2. When a low-pass filter is applied to the data the regimes are found to be slightly weaker, meaning the maximum and minimum geopotential height anomaly are smaller, but they do not show a visible difference in the configuration of high and low z500 anomalies. For the results using a persistence constraint differences in the regimes only emerge for very strong (unrealistic) constraint values in the form of slight shifts in the location of the centres of high and low z500 regions. For weak (realistic) constraints the regimes found are the same as for the unconstrained method and no weakening is found. By a 'realistic' constraint we mean one that does not force datapoints into regimes which are a large distance away, but only switches those datapoints that are in-between different regimes, as can be seen in the toy example in Section 3.3.1. In practice these are constraints corresponding to an average regime duration below circa 9 days (the corresponding C can be found in Table 3.4). Note that the average regime duration averaged over all unconstrained regimes following Table 3.3

is 5.5 days for $k = 4$ and 4.4 days for $k = 6$, which means the persistence constraint will only affect the result when it is below either $C \approx 1800$ or $C \approx 2200$ for $k = 4$ or $k = 6$ respectively (Table 3.4).

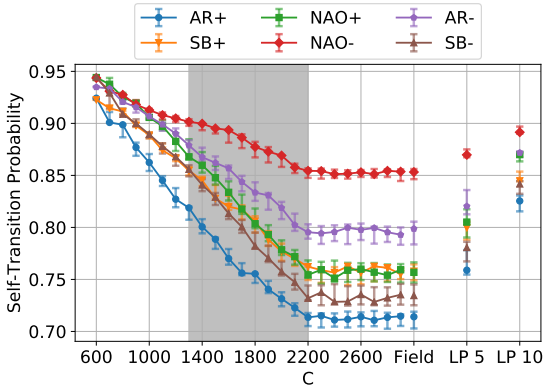
The effects of the time-filtering and time-regularised method on the self-transition probabilities are shown in Figures 3.16a ($k = 4$) and 3.16b ($k = 6$). On the left of each panel the results for the constrained algorithm are shown for various C and as expected the self-transition probability increases with decreasing C. The smaller the value of C, the less switches between regimes are allowed. The increase of the self-transition probability with decreasing C is approximately linear for all regimes, and starts at the 'raw' self-transition probability of the regimes. Consistently with the values computed using the average regime duration in Table 3.3, we find that the constraint starts to affect the self-transition probabilities around either $C = 2200$ for $k = 6$ (4.3 days) or $C = 1800$ for $k = 4$ (5.3 days). Note that when the constraint starts to affect the results the regime dynamics can no longer be described as a first-order Markov process and care must be taken interpreting the average regime duration and corresponding self-transition probabilities.

Comparing the results for time-filtered data with those of the constrained method in Figure 3.16 we see that using a 5-day low-pass filter corresponds to a constraint value of roughly 2000 for $k = 6$ and 1400 for $k = 4$. This difference is mainly due to the stronger effect of the constraint for a larger number of clusters. For the 10-day filter the corresponding values of C are approximately 1400 and 1100 for $k = 6$ and $k = 4$ respectively. Note that the self-transition probability of certain regimes differs slightly between the two methods. For example, the AR+ regime is found to increase its self-transition probability relatively stronger for the time-filtered data.

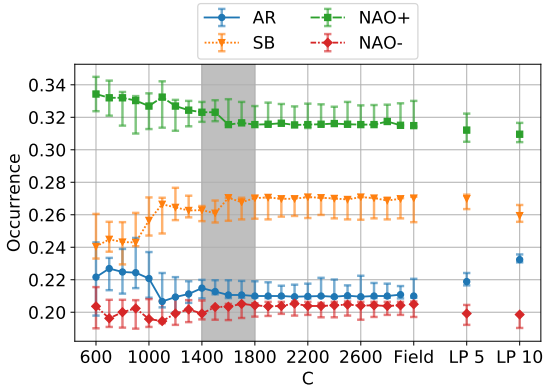
The occurrence rates of the different regimes are shown in Figures 3.16c and 3.16d for $k = 4$ and $k = 6$ respectively. We start by looking at the results of the time-regularised algorithm. The occurrence rate remains the same as for the unconstrained data, even for constraint values significantly stronger than the 'raw' persistence of the data. Only for very low C (strong constraints) do the occurrence rates start to differ. This indicates that the method indeed causes a switching of the 'in-between-cluster' points to the cluster of their neighbours instead of the cluster they are slightly closer to, as expected from the results of the toy model (Section 3.3.1). We regard the constraint as being 'weak' so long as the occurrence rates are not affected, and helping to identify true physical persistence. For these values of C we consider that the persistence constraint acts as a good filter, indicated by the gray bands in Figure 3.16. In contrast, the results for the time-filtered data show significant differences in the occurrence rates of the regimes. Especially for the 10-day filter the differences for e.g. the AR+ regime ($k = 4$) or the NAO+ regime ($k = 6$) are substantial. As this is the standard filter used in literature (e.g. Straus et al., 2017) it raises the question of how



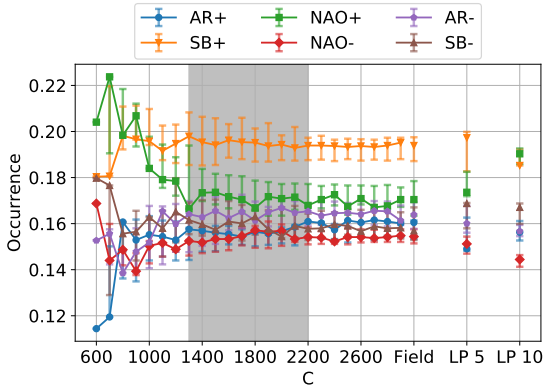
(a) The self-transition probabilities for $k = 4$.



(b) The self-transition probabilities for $k = 6$.



(c) The occurrence rates for $k = 4$.



(d) The occurrence rates for $k = 6$.

Figure 3.16: The occurrence and self-transition probabilities of the different regimes for $k = 4$ and $k = 6$ for the clustering results including the persistence constraint depending on the value of C (corresponding average regime durations can be found in Table 3.4). To the right the values for the unconstrained algorithm (field) and the 5- and 10-day low-pass filter are shown. The error bars indicate the maximum and minimum value of occurrence/self-transition probabilities for clustering results with a slightly smaller $\mathbf{L}$ (bounds for the difference are $\{0.00968, 0.00936, ..., 0.00232, 0.002, 0.002, 0.002\}$ decreasing with increasing C, which are chosen sufficiently small to give similar regimes according to the data correspondence. The gray bands indicate the region in which the persistence constraint is considered to act as a good filter.

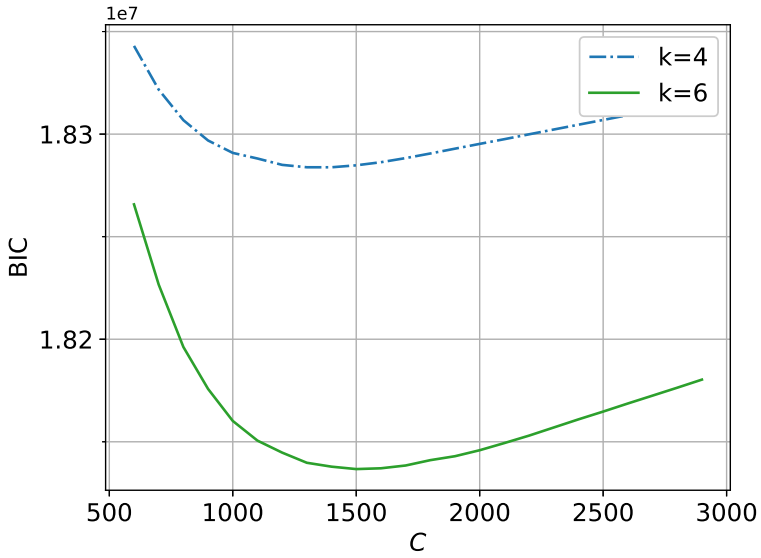


Figure 3.17: The BIC for the clustering with persistence constraint for $k = 4, 6$ and C in the range $\{600, ..., 2900\}$.

reliable the occurrence rates are and whether they are not solely a feature of the method used. In contrast, the inclusion of the constraint within the clustering procedure itself does not lead to such a bias and therefore provides a more robust way of finding persistent regimes, i.e. of isolating the signal from the noise.

When using the time-regularised algorithm one of the choices that needs to be made is which constraint value C is best to use. Here we base this choice on the BIC, as we did for finding the optimal number of clusters k. In Figure 3.17 the BIC is shown for both $k = 4$ and $k = 6$. For $k = 4$ the minimum is found for $C = 1400$ and for $k = 6$ it is found for $C = 1500$. These constraint values correspond to an average regime duration of 6.8 and 6.3 days respectively. The lower end of the region in which the BIC is close to its minimum coincides with the point beyond which smaller values of C start to affect the occurrence rates, giving a lower bound for the region where the persistence constraint is considered to act as a good filter as indicated by the gray bands in Figure 3.16. This increases the confidence of the optimal value of C being around these values. Interestingly, the optimal average regime duration for $k = 4$ and $k = 6$ differs by less than 10% ($\Delta C \approx 1500 - 1400 = 100$), whereas without the persistence constraint the average duration differs by 20% ($\Delta C \approx 2200 - 1800 = 400$). This confirms that the persistence constraint is helping identify a physical signal that is less dependent on the number of clusters chosen. The range where the BIC is very close to its minimum is between 6.3 ($C = 1500$) and 7.9 ($C = 1200$) days for $k = 4$. For

$k = 6$ this range is from 5.9 ($C = 1600$) to 6.8 ($C = 1400$) days. Twice this timescale, which is the minimum for recurrence of a regime (i.e. for regimes A and B we have $A - B - A$), thus corresponds roughly to a period of 12 to 14 days. This is somewhat longer than the timescale of synoptic weather systems (Blackmon et al., 1977; Boljka et al., 2018). As this is an average there are a substantial number of longer lasting regimes showing persistence well beyond the synoptic timescale.

3.4 Summary and Discussion

In this chapter we have shown, using an information criterion and further arguments based on the consistency of the clustering result, that the traditional number of four clusters is not optimal for representing wintertime Euro-Atlantic weather regimes when gridpoint data is considered. The traditional approach of applying clustering to the first few PCs involves a loss of information, which affects the number of regimes that is best to represent the data. The optimal number of regimes for the gridpoint data was identified using the Bayesian Information Criterion (BIC), which strikes a balance between how well the regimes fit the data and the number of parameters needed to describe them. This way we find that for the gridpoint data six regimes is the optimal choice. The two additional regimes are the opposite phases of the Atlantic Ridge and Scandinavian Blocking, introducing a pleasing symmetry in the found clusters. Furthermore, the dominant occurrence of the NAO+ when there are only four clusters, which likely is due to it being the only regime with a low pressure area in the north, is reduced by the addition of two regimes that also have this feature. Therefore, six regimes allow for more variability in their representation of the circulation and prevent all data with a more zonal flow from projecting onto the NAO+.

Next, we looked into ways to enforce persistence of the regimes. A common approach in literature is to apply a low-pass filter to remove high frequency oscillations and focus on the persistent behaviour (e.g. Bao and Wallace, 2015; Straus et al., 2017). This alters the data to which the clustering algorithm is applied, just as the use of PCs does. We have shown that this leads to a significant change in the occurrence rates of the circulation regimes. A new method, which incorporates a persistence constraint in the algorithm itself, does not change the data while still enforcing persistent regimes. The results for this time-regularised approach do not exhibit the change in occurrence rate found for the time-filtered data, as long as the constraint is not too strong, while still having an increased self-transition probability. Therefore, this method leads to a more robust and unbiased result compared to the time-filtering approach.

A choice that needs to be made in this adapted clustering method is the value of the constraint C. Using the BIC the optimal value of C is found to lie around an average regime duration of six to seven days. Interestingly, this matches the point beyond which smaller values of C start to affect the occurrence rates. Thus it can be viewed as a more accurate estimate of the physical persistence of the regimes than that provided by the raw data without the persistence constraint. Double this value, which is the minimum for recurrence of a regime, thus is slightly longer than the timescale of synoptic weather systems (Blackmon et al., 1977; Boljka et al., 2018). This shows that the atmospheric circulation indeed exhibits persistence beyond the synoptic timescale, suggesting the presence of predictable low-frequency modes.

Both results indicate that care must be taken when applying filtering methods (PCs, low-pass filter) to the data before a clustering algorithm is applied. Clustering itself provides a means of dimension reduction by projecting onto components representing recurrent patterns in the data. Since this is a method of filtering the data, it seems ill-advised to apply this to already filtered PC data, as it is not clear what the effect of this double filtering is on the result. A similar argument holds for applying a time-filter to the data before clustering. Information is lost in this procedure, introducing a bias in the resulting circulation regimes and their occurrence rates.

Chapter 4

Detection of Non-Stationary Regime Dynamics

The focus in this chapter is on whether the occurrence of the circulation regimes identified in Chapter 3 varies on (sub-)seasonal and interannual timescales in a predictable way. Such non-stationarity has a forced component, which yields the predictability, and an internal variability part, that is not predictable. The forced variability can be caused by links between the circulation regimes and other patterns of climate variability, such as sudden stratospheric warmings (SSWs) on sub-seasonal timescales (e.g. Charlton-Perez et al., 2018; Domeisen et al., 2020) or the El Niño Southern Oscillation (ENSO) on interannual timescales (e.g. Lee et al., 2019; Drouard and Cassou, 2019). A better understanding of the processes guiding the non-stationary regime dynamics can help improve predictions of the regimes themselves, as well as of the consequences for local and regional weather. To this end it is important to robustly identify predictable regime variations, given the inevitable presence of noise, which can conceal these possibly weak signals within the regime dynamics. A regularised clustering method is proposed in this chapter to do exactly that.

We start this chapter by motivating the use of this regularised clustering approach to identify non-stationary regime dynamics in Section 4.1. We discuss the problem setting of using clustering methods to identify circulation regimes in model ensembles exhibiting a wide spread in their regime representation, and show a motivational example for the regularisation method proposed to handle this spread and identify a more robust signal. This regularised method, and the data used, are then discussed in detail in Section 4.2, followed by a discussion on the choice of the regularisation

constraint using several criteria in Section 4.3. The obtained regime dynamics is presented in Section 4.4, looking into the effect of the regularisation and studying the non-stationary regime signal on both sub-seasonal and interannual timescales, as well as discussing the signal-to-noise problem (introduced in Section 1.4). We end with a summary and discussion of the results in Section 4.5.

4.1 Motivation for a Regularised Clustering Approach

The common approach for identifying atmospheric circulation regimes is to apply a k-means clustering algorithm to the 500 hPa geopotential height (e.g. Michelangeli et al., 1995; Cassou et al., 2005; Straus et al., 2007), as discussed in detail in Chapter 3. Applying this method to reanalysis data has shown consistent results between studies for e.g. the northern hemisphere or the Euro-Atlantic sector. However, reanalysis data only provides a single realisation (out of the many possible within the climate system) and thus is sensitive to sampling uncertainty in detecting the non-stationary signal. Reanalysis data mostly covers only the last 40 years (e.g. ERA-Interim), which is too short to reliably identify any non-stationary regime behaviour, especially on interannual timescales, but also on sub-seasonal timescales. For example, when one is interested in the effect of ENSO on the occurrence rate of the circulation regimes in winter, one only has a few years of data consisting of roughly 120 days each when using ERA-Interim. For six regimes, on average occurring 20 days each year, a few days more being assigned to one regime can significantly affect the regime frequencies. This makes it difficult to distinguish any signal from the noise.

One way to increase the sample size, and thus identify a more robust signal of the predictable component of the variability, is to use the UNSEEN method, in which model ensemble members with different lead times (moments of initialisation) are pooled to create a very large ensemble (e.g. Thompson et al., 2017; Kelder et al., 2020). Here, the lead times are beyond the deterministic predictability limit and there is no skill for predicting individual weather events. The ensemble members can thus be treated as plausible alternative realisations of reality, which are all equally affected by the sources of predictable variability. In line with this approach we use hindcast ensemble data of the ECMWF seasonal forecast system to study sub-seasonal and interannual regime dynamics. As the model has high levels of interannual ENSO forecast skill (Johnson et al., 2019), using seasonal forecasts allows for a more precise study of the interannual dynamics and effects of e.g. ENSO on the regimes. Here, we are not primarily concerned with the initial condition problem of weather forecasting, but rather require a high-resolution model with a small bias on the slightly longer timescales of interest for this study.

A difficulty here is that models are imperfect and exhibit a wide spread in their regime frequencies compared to reanalysis data (e.g. Fabiano et al., 2020). This behaviour may reflect the "signal-to-noise paradox" whereby models are more noisy than the real world (see further discussion in Section 4.4.5). It is exemplified by the domain dependence of the regimes, particularly the negative counterpart of the Atlantic Ridge (AR−), when considering the six regimes identified using the ECMWF SEAS5 hindcast ensemble shown in Figure 4.1. These six regimes for SEAS5 are similar to those identified for ERA-Interim in Chapter 3, but not the same. A discussion of the similarities and differences between the two datasets, both general and regime-specific, is given in Appendix B. On the other hand, for ERA-Interim data no domain dependence of the regimes is found, despite the smaller sample size. This domain dependence of the regimes within the model is undesirable from a physical and useability perspective.

When identifying circulation regimes the presence of noise can hide possible regime variability signals, both in model and reanalysis data. Specifically, small deviations in the distance of data

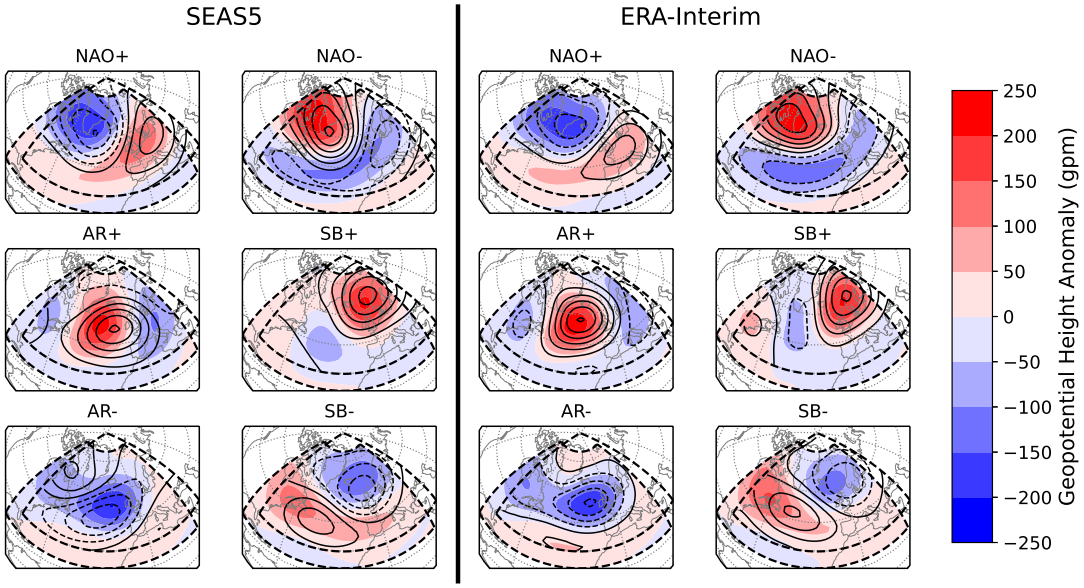
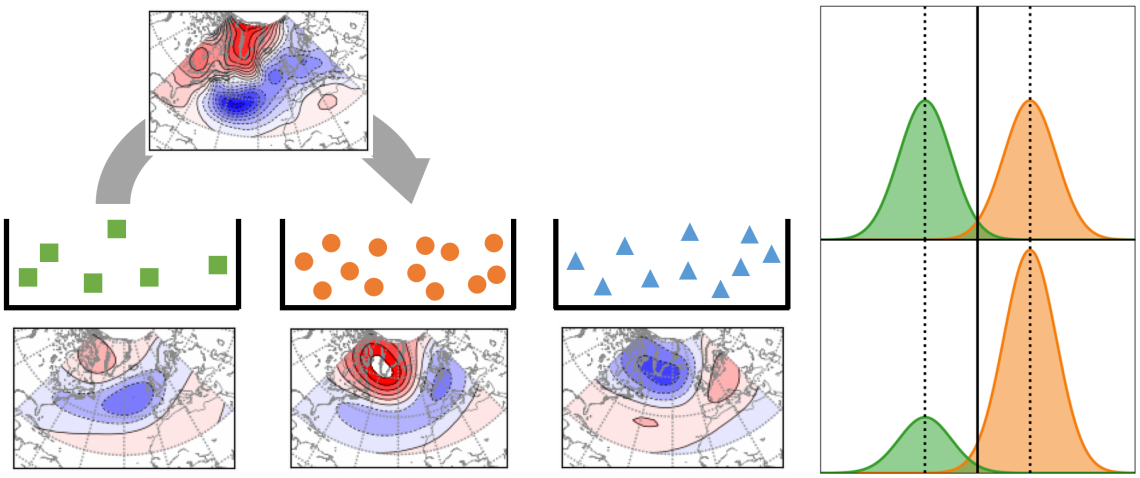


Figure 4.1: The regimes identified for the ECMWF SEAS5 hindcast ensemble members (left) and ERA-Interim (right) using standard k-means clustering for two slightly different domains (indicated by the dashed boxes). They are the positive and negative phase of the North Atlantic Oscillation (NAO), the Atlantic Ridge (AR+) and its negative counterpart (AR−), and the Scandinavian Blocking (SB+) and its negative counterpart (SB−). Regimes for domain A (20-80N, 90W-30E) are indicated by the colours and those for domain B (30-90N, 80W-40E) by the contours following the same 50 gpm difference between contour levels.

to the regimes can result in them being assigned to a different regime, obscuring the "true" signal. The method of k-means clustering deals poorly with data containing a lot of noise. The reason for this is that k-means clustering assigns every data point to the cluster centre, or regime, that it is closest to, even if only by a tiny margin. This makes it overly sensitive to noise, especially when the signal-to-noise ratio is small. A consequence is that the identification procedure lacks robustness and the informational gain is small. The following example visualises this issue by means of one possible scenario.

Consider Figure 4.2a which shows the distribution of ensemble members over three different clusters, and Figure 4.2b showing the (theoretical) distributions of data over two regimes (green, left and orange, right) when they are equally likely (top) and when the orange regime is more likely (bottom, note this is an exaggerated visualisation for illustration purposes). At a fixed time t it is possible that a data point, i.e. an ensemble member, falls in-between two (or more) regimes (e.g. the clusters associated with the green squares and orange circles). Here a standard k-means clustering algorithm assigns it to the regime it is closest to in distance, which is valid if the regimes are equally likely. However, due to the effect on the regimes of external forcing, such as ENSO, this is not always the case. If one regime is known to be more likely than the neighbouring one at that point in time, then it would be prudent to assign the ensemble member to the more likely regime. In this way the regime assignment of an ensemble member is not solely determined by its distance to the cluster centres, but also by a prior likelihood set by the distribution of the ensemble members over the regimes, which is picking up a non-stationary signal. Effectively, noise is being penalised. For example, in our visualisation the cluster comprised of the ensemble members indicated by the orange circles is more likely, i.e., occurs a lot more over all ensemble members, at a given time t. The shown ensemble member in distance falls between the green square and orange circle regimes, i.e. only slightly left of the solid line in the right figure. To assign it to the green square regime based on this small difference in distance places more weight on the noise than on the signal. For that reason it can be better to assign it to the orange circle regime, which has a higher probability as shown on the bottom of Figure 4.2b.

The aim is then to design a clustering method that penalises noise, to mitigate incorrect assignments of datapoints as exemplified above and avoid the misinterpretation of the regime signal. One possible way to achieve this is to regularise the k-means clustering method by implementing a constraint enforcing a level of similarity between the ensemble members at each moment in time (Section 2.2.1, see Bishop (2006) for a discussion of different types of regularisation). That is, we add information from the model ensemble to obtain a better informed regime identification method



(a) (b)

Figure 4.2: (a) An example of a possible reassignment of an ensemble member. The green squares, orange circles and blue triangles give the distribution of the ensemble members over the different regimes, that is, each marker indicates to which regime the corresponding ensemble member is assigned, but its location within the bin does not provide any further information. The arrow shows the desired reassignment of a data-point, which might plausibly be associated with either the green square or the orange circle regime. To better understand this reassignment, in (b) the distribution of data around the cluster centres (dotted lines) is shown for equally likely regimes (top), and a situation where the orange (right) regime is more likely, i.e. higher amplitude (bottom). When the two regimes are equally likely a point in the middle (solid line) has an equal probability of belonging to either of the regimes. However, when the orange (right) regime is more likely than the green one (left), data that lies half-way between has a larger probability of belonging to the orange regime than to the green one, and thus might better be reassigned to the orange regime. Such reassignments can help to reduce the effect of noise and identify a more robust signal.

that strengthens the non-stationary signal by penalising noise. This design has a physically meaningful basis as the preferred regimes should, on average, be represented by the overall ensemble, and if one regime is more populated than usual at a particular time, it makes it more likely that borderline cases belong to that regime (Figure 4.2). By introducing a constraint that prioritises similarities over small deviations it is possible to distinguish more pronounced regime behaviour, i.e. to discriminate better between the regimes. The underlying assumption here is that the distribution of the ensemble members over the regimes changes in time due to external factors such as ENSO and that the regularisation helps to better identify such weak non-stationary signals. Similar forms of regularisation, designed to increase persistence in time, have been successfully employed to de-

tect robust and meaningful regimes in the climate context (see Chapter 3 and e.g. Horenko, 2010*a*; Falkena et al., 2020).

Of course, such a regularised clustering algorithm could be over-confident. A curb on over-confidence is provided by accuracy. The overall goal is to keep the accuracy at a reasonably high level while significantly increasing the information gain (in the entropy sense) of the derived regime model. Specifically, we favour regimes with more informative dynamics over those that fit the data slightly better, since this can be advantageous in identifying weak signals. We thus monitor the effect of this empirical regularisation by quantifying the trade-off between accuracy and information, and assessing whether what it does is physically sensible. How to quantify this trade-off between accuracy and informativeness is discussed in Section 4.3. First we elaborate the regularised clustering method applied to ensemble data in the next section.

4.2 Ensemble-regularised k-means Clustering

When it comes to regime-analysis of model ensemble data there are two approaches one can take. The first is to assign the model data to the regimes obtained from reanalysis data (e.g. Ferranti et al., 2015; Grams et al., 2018). The second approach is to compute the regimes from the model data itself (e.g. Dawson and Palmer, 2015; Matsueda and Palmer, 2018). This latter approach means that the regimes identified in the model can differ from those of reanalysis data. On the other hand, it includes a natural bias correction of the model data in the regime representation, as possible differences in the location of the regime centres between the two datasets do not affect the regime dynamics of interest. Here we choose the latter approach for that reason. Before discussing how to implement the regularisation discussed in the previous section in the k-means clustering algorithm in Section 4.2.2, we first detail the model ensemble data used for the identification of the circulation regimes in Section 4.2.1.

4.2.1 Data

Daily 500 hPa geopotential height from the ECMWF hindcast ensemble of SEAS5 (Johnson et al., 2019) is used for the identification of the circulation regimes. The ensemble has 51 members for a November 1st start date and is available from 1981 to 2016. We use the 500hPa geopotential height (z500) on a 2.5° by 2.5° grid for a domain covering the Euro-Atlantic sector. To analyse the robustness of the obtained regimes we consider two slightly different domains; 20° to 80°N and

90°W to 30°E (domain A) and 30° to 90°N and 80°W to 40°E (domain B) (both are commonly used in literature, e.g. Cassou et al. (2005); Dawson et al. (2012)). The months December to March are considered using daily data (00:00 UTC), meaning forecast lead times of over a month are used for which there is no longer any discernible effect of the atmospheric initial conditions. Note that this loss of memory of the initial conditions does not imply there is no predictable variability, as other processes such as the month within the season or the phase of ENSO affect the circulation. We compute anomalies with respect to a DJFM average climatology to not make any assumptions on the (sub-)seasonal variability in the background climatology (for further reasoning on this point see Falkena (2019); Falkena et al. (2020) and Section 4.4.2).

To reduce the effect of weather noise, preprocessing methods are often used to focus on the larger-scale, predictable variability. In Fabiano et al. (2020) an Empirical Orthogonal Function (EOF) analysis was used to reduce the dimensionality of the data in a model ensemble, but still a large spread in the centroid distance and spatial correlation of the regimes was found. This indicates that this way of preprocessing is not sufficient to reduce the effect of noise on the identified regimes. Other methods of preprocessing the data, such as using a low-pass filter, could filter out some of the noise within the model as well. However, these methods can also lead to biases in the resulting regimes. For example, in Section 3.3 it was found that the use of a low-pass filter affects the regime frequencies. Therefore, we here focus on adapting the clustering method instead of preprocessing the data, to identify a more robust regime signal. This way we do not lose any information present in the data and avoid possibly introducing a bias by preprocessing. The regimes obtained for the SEAS5 hindcast data using the clustering method described in the following section are identified with the corresponding regimes in the ERA-Interim reanalysis (Dee et al., 2011) as obtained in Chapter 3 (for domain A). This means that the regimes for SEAS5 and for ERA-Interim are slightly different (see Figure 4.1), which allows for bias within the model. For consistency, the same period of 36 years for which the SEAS5 data is available is considered for ERA-Interim.

4.2.2 Method

The discussion of the ensemble-regularisation in this section follows the same steps as that on the time-regularisation in Section 3.3.1, repeating some aspects for completeness. The main difference, next to the regularisation itself, is that the data now also has an ensemble-member dimension. Let the considered data be of the form $x_{t,n} \in \mathbb{R}^{T\times N\times D}$, where T is the length of the time series, N the number of ensemble members and D the spatial dimension of the data (here latitude×longitude).

The aim of clustering the data is to find k cluster centres $\Theta = (\theta_1, ..., \theta_k) \in \mathbb{R}^{k \times D}$ (regimes) that best represent the data. The assignment of individual datapoints to the different clusters is given by the weights, or affiliation vector, $\Gamma = (\gamma_1(t,n), ..., \gamma_k(t,n)) \in \mathbb{R}^{k \times T \times N}$. This can be understood as the probability of a data point belonging to each of the different regimes.

Identifying the circulation regimes means that we need to find the optimal parameters for the cluster centres Θ and the data-affiliations Γ. To achieve this, a cost function, also referred to as the averaged clustering functional (in its discrete form), is minimized (Franzke et al., 2009) (compare Equation (3.9)):

$$\mathbf{L}(\Theta, \Gamma) = \sum_{t=0}^{T} \sum_{n=1}^{N} \sum_{i=1}^{k} \gamma_i(t,n) g(x_{t,n}, \theta_i). \tag{4.1}$$

Here $g(x_{t,n}, \theta_i)$ is the distance between the cluster centre and a datapoint, for which the L_2-norm $((\sum_{gridpoints}(x_{t,n} - \theta_i)^2)^{1/2})$ weighted by the cosine of latitude is used. The affiliations $\gamma_i(t,n) \geq 0$ are normalised following

$$\sum_{i=1}^{k} \gamma_i(t,n) = 1, \qquad \forall t \in [0,T], \quad \forall n \in [1,N]. \tag{4.2}$$

In practice, the $\gamma_i(t,n)$ values obtained via the optimisation are mostly equal to zero or one. In that case the datapoints are unambiguously assigned to one of the regimes. In traditional k-means clustering the assignment of Γ often does not exhibit persistence in time or with respect to the different ensemble members. This can be a sign of misinterpreting noise to be the signal. The aim is to mitigate this effect in order to identify a robust signal.

Previous studies have introduced a constraint within the clustering method to increase the temporal persistence of the regimes (Horenko, 2010*a*; de Wiljes et al., 2014, Chapter 3). Here we expand on that idea by implementing a constraint on the similarity between the ensemble members at every time-step, with the aim of identifying a more robust regime signal as discussed in Section 4.1. This constraint takes the form

$$\frac{1}{2} \sum_{i=1}^{k} \sum_{n_1, n_2 = 1}^{N} |\gamma_i(t, n_1) - \gamma_i(t, n_2)| \leq \phi \cdot C_{\text{eq}}, \qquad \forall t \in [0,T], \tag{4.3}$$

where the sum over n_1, n_2 is taken over all combinations of two ensemble members, that is $\sum_{n_1=1}^{N-1} \sum_{n_2=n_1+1}^{N}$. The division by two ensures that differences are not counted twice. C_{eq} is the maximum value that can be attained by the sum on the left-hand side and is given by

$$C_{\text{eq}} = \frac{N}{2}\big(N - \frac{N}{k}\big). \tag{4.4}$$

The maximum of C_{eq} is reached if the ensemble members are equally distributed over the k regimes. Thus ϕ represents the strength of the constraint relative to the maximum value C_{eq}. One can think

of ϕ as the proportion of the ensemble members that is not affected by the constraint. Note that by expressing the constraint in this way its strength, as given by ϕ, is independent of the ensemble size or the number of regimes. Instead, the ensemble size and number of regimes enter into C_{eq}. Since we implement the constraint separately for every time-step we do not make any assumptions on the form of the non-stationarity, but only ensure the algorithm can better discriminate between the regimes at a given time.

The regimes are obtained by minimising the clustering functional $\mathbf{L}$ in Equation (4.1) subject to the constraints in Equations (4.2) and (4.3). As for the time-regularised clustering, this minimisation is done in two steps:

1. For fixed Γ, minimise $\mathbf{L}$ over all possible values of Θ.
2. For fixed Θ, minimise $\mathbf{L}$ over all possible values of Γ.

The first step is realised via standard k-means clustering, while for the second we employ linear programming, i.e. optimisation of a linear function subject to constraints, using the Gurobi package for Python (Gurobi Optimization LLC, 2019). When the difference between subsequent $\mathbf{L}$ values comes below a set threshold the computation is terminated. Analogously to standard k-means clustering this presented algorithm only finds local minima. Therefore we run it at least 50 times starting from different initial seeds in an attempt to heuristically infer a global minimum (this approach is referred to as simulated annealing in the literature). The run with the lowest $\mathbf{L}$-value is then selected as the final result. We use $k = 6$ as the number of regimes representing the wintertime circulation over the Euro-Atlantic region, as this was identified to be optimal in Chapter 3. This result was obtained using ERA-Interim data. Using the same number of clusters k allows for a comparison between the two datasets and helps to better assess the predictable signals. A discussion of the differences between the SEAS5 hindcast ensemble and ERA-Interim data is given in Appendix B.

While $\gamma_i(t, n) \in \{0, 1\}$ within the standard k-means clustering, this is no longer the case for all time-steps or ensemble members when incorporating the constraint. Specifically $\gamma_i(t, n) \notin \{0, 1\}$ when it is not possible to numerically obtain a solution on the bounds of the admissible set of the optimisation problem; in that case $\gamma_i(t, n)$ is between zero and one. We use this as an indication that those data points cannot be unambiguously assigned to one of the regimes and employ it to define a 'no-regime' category. Note that this means that even if $\gamma_i(t, n)$ for some t, n is very close to one for a regime, that datapoint is still assigned to belong to the no-regime category. Using this

definition of a no-regime category the number of data points assigned to it increases approximately linearly with ϕ (not shown), i.e. the stronger the constraint (lower ϕ), the more data is assigned to the no-regime category.

4.3 Selection of a Regularised Regime Model

Using a constraint to regularise the outcome of the clustering algorithm, as introduced in the previous section, requires choosing a suitable constraint value. This value determines the regime-model that is used for the subsequent analysis of the regime dynamics and its non-stationarity. An appropriate value is highly dependent on the considered application and needs to be determined accordingly. Yet it is possible to employ several different selection criteria to aid the decision process, independently of the underlying physical processes. In this section we first introduce two main criteria that can be used for the constraint selection, as well as a check on the domain dependence of the regimes. We then evaluate these measures for the considered problem and discuss the arguments to arrive at our final choice of ϕ.

4.3.1 Selection Criteria

Numerous methods exist for deciding on the best model for the data at hand, i.e. to find an optimal value of the constraint. For example, for the choice of the optimal number of circulation regimes researchers have used verification by synthetic datasets (e.g. Straus et al., 2007), a classifiability index (e.g. Michelangeli et al., 1995), an information criterion (e.g. O'Kane et al., 2013) or cross-validation (Quinn et al., 2020). These methods can be used not only to determine the number of circulation regimes, but also the values of other hyper-parameters (e.g. Quinn et al., 2020, Chapter 3). Here we introduce three criteria, namely the Bayesian Information Criterion, Shannon entropy and a domain robustness measure, which can all be used to inform the choice of a suitable constraint value ϕ.

Bayesian Information Criterion

Information criteria are a popular tool for model selection and we have already seen the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Section 3.2.1. The aim is to find a balance between the accuracy and complexity of the model (in the spirit of Occam's

razor), that is, between how well the model fits the data and the number of parameters required (Burnham and Anderson, 2004). Since the ensemble gridpoint data considered in this chapter is high-dimensional the BIC is preferable over the AIC, as discussed in Section 3.2.1. For completeness we repeat the BIC, which is given by

$$\text{BIC} = -2\log(\mathcal{L}(\hat{\theta}|\text{data})) + K\log(m) = m\log(\hat{\sigma}^2) + K\log(m), \tag{4.5}$$

where m is the dimension of the data and K the number of parameters needed to describe the clusters. $\mathcal{L}(\hat{\theta}|\text{data})$ is the likelihood of the model parameters given the data, which can be expressed using the error variance $\hat{\sigma}^2$. Note that the error variance, being a measure of the accuracy of the regimes, is estimated using the clustering functional $\mathbf{L}$ in Equation (4.1). When computing the BIC we need to determine m and K. Each ensemble member has dimension $D \times T$, and the main question is in what way the number of ensemble members is incorporated. There are two choices that can be made for this. Firstly, one can simply use the number of ensemble members N. However, the regularisation constrains the number of combinations between any two ensemble members $\binom{N}{2}$ instead of the assignment of each of the N ensemble members individually. For this reason, we decide to use $\binom{N}{2}$ as the dimension of the ensemble, which yields $m = D \times \binom{N}{2} \times T$ and $K = k \times D + T \times (k-1) \times \phi \cdot \binom{N}{2}$.

Shannon Entropy

The second method considered to identify a suitable constraint value ϕ is to calculate the entropy, or information content. Entropy has already been used occasionally in evaluating model performance for circulation regimes (Fabiano et al., 2020), and some studies use it as a way of correcting information criteria for the distribution of the model residuals (Murari et al., 2019; Rossi et al., 2020). The goal of regularising the clustering algorithm is to identify a stronger non-stationary signal, i.e. to increase the amount of information in the resulting regime dynamics. The use of an information measure, such as entropy, thus follows naturally from the aim of implementing the constraint. The BIC validates model performance and does not capture this informativeness of the signal. Therefore, we also consider the Shannon entropy (Shannon, 1948; Shannon and Weaver, 1949), which is given by

$$H(p_i) = -\sum_{i=0}^{k} p_i \log_2 p_i, \tag{4.6}$$

where p_i, $i \neq 0$, is the occurrence rate of regime i and p_0 is the occurrence rate of no-regime. The Shannon entropy is low for an equal distribution of the data over the regimes. On the other hand it is larger for a more unequal distribution in which there is a stronger signal. Note that the information

gain by enforcing a less equal distribution of the data over the regimes comes at the cost of reduced accuracy.

Domain Robustness

Lastly, we discuss the domain robustness of the regimes to see whether a choice of constraint is suitable. By domain robustness we refer to the domain dependence of the regimes as obtained by the algorithm. Using standard k-means clustering the regimes of the SEAS5 hindcast data are found to be domain dependent, as discussed in Section 4.1 (Figure 4.1). From a physical perspective this is undesirable, as the conclusions on the regime dynamics can then depend on the chosen domain. To transform this domain dependence into a verification criterion for the constraint value we consider the clustering results for both domains A and B and compute the pattern correlation between the two sets of regimes over the overlapping section of these domains. The average pattern correlation is then computed as a measure of how well the regimes for the two domains match. Alternatively, one could take the lowest pattern correlation to indicate the quality of the match, i.e. what is the worst matching regime. This yields very similar results.

4.3.2 Selecting the Constraint Value

Now we evaluate the three discussed complementary criteria for the considered data. In Figure 4.3 these criteria are shown for a range of ϕ, where the BIC and Shannon entropy are shown for the two domains considered. For the BIC, which strikes a balance between accuracy and complexity, the optimal value is located at its minimum. In contrast, for the entropy, which compares information content and complexity, a higher value indicates a better result. The pattern correlation between the two domains ideally is as high as possible, indicating robustness of the regimes with respect to the choice of domain.

The BIC attains its minimum at $\phi = 0.96$ for both domains, indicating that for that constraint value the accuracy and complexity of the regime-representation of the data are in balance. When looking at the results we find that these regimes and the assignment of the data to them are very similar to those without the constraint, indicating that the constraint is too weak to have a strong impact. Furthermore, we find that this minimum depends strongly on the number of ensemble members considered, e.g. using 26 members the minimum of the BIC is found for $\phi = 0.92$. While the BIC is generally a good method to select certain hyper parameters, the goal of the regularisation is not just to identify the best statistical model and attain the highest accuracy, but also to obtain

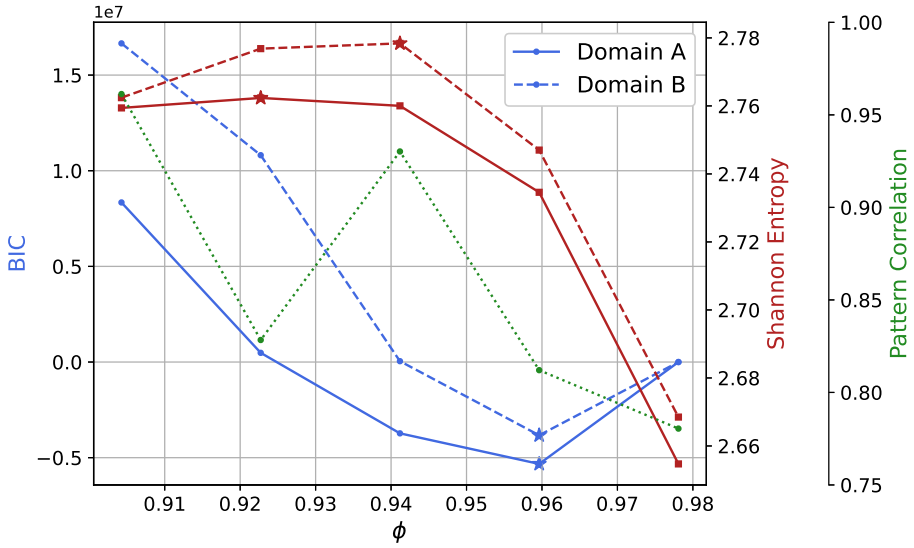


Figure 4.3: The BIC (blue, circles) and Shannon entropy (red, squares) for the two domains considered. The average pattern correlation between the regimes for the two domains (green, dotted) is shown as well. Stars indicate the lowest or highest value respectively, suggesting a suitable value for the constraint ϕ.

a more pronounced regime signal. To this end it can be desirable to lose some accuracy, by using a stronger constraint value, for gaining information.

The Shannon entropy indicates an optimal constraint value around $\phi = 0.92 - 0.94$, where it attains its maximum values. This is slightly stronger, i.e. a lower ϕ, than the optimum indicated by the BIC. For these values of ϕ most information, or signal strength, is gained by constraining the regime assignment of the data. Since the aim of implementing the constraint on the ensemble similarity is to identify a stronger signal, we decide to use the entropy results as the main guidance. To verify whether the model is still realistic for these parameter values we turn to the BIC as a complementary measure. Within the range of high Shannon entropy the lowest BIC is found for $\phi = 0.94$ and therefore this is the value we decide to work with. Also, for this constraint value the regimes are barely domain dependent, as indicated by the high average pattern correlation.

4.4 Non-Stationary Regime Dynamics

For this suitable value of $\phi = 0.94$ we study the resulting regime dynamics. We start by discussing the effect of the regularisation on the regimes themselves and their overall occurrence rates, after

which we turn to the non-stationary signals that can be identified. When discussing the non-stationarity we also consider the results obtained using standard k-means clustering, to further look into the effect of the regularisation. We look at variability on both the sub-seasonal and interannual timescales. For the interannual variability we discuss whether there is any predictable signal for the regime occurrence rates and compare this to the signal identified for an NAO-index, relating it to the signal-to-noise paradox.

4.4.1 Effect of the Regularisation on the Regimes

The regularisation affects the assignment of the data to the regimes, and thus their occurrence rates. In Figure 4.4 the average occurrence rates of the regimes are shown for standard and ensemble-regularised k-means clustering. For the unconstrained result, as well as ERA-Interim, the occurrence rates are close to an equal distribution of the data over all six regimes (see dotted line in Figure 4.4). In contrast, the occurrence rates differ significantly from an equal distribution when $\phi = 0.94$ is used as a constraint. Despite the relatively weak constraint several regimes, such as NAO+, have occurrence rates whose range barely overlaps with that of a uniform distribution (when corrected for the no-regime occurrence rate, dash-dotted line). This shows that the regularisation helps to discriminate better between the different regimes, identifying a stronger regime signal within the SEAS5 data. It follows that the uniformity of the occurrence rates found for the ERA-Interim regimes is potentially due to a lack of discrimination between the regimes, as there are not enough data available to control the noise.

The geographical regime structures are mostly unaffected by the constraint of $\phi = 0.94$, as can be seen in Figure 4.5. Only AR− changes noticeably, now having a weak positive z500 anomaly over Greenland. Overall these regimes correspond well to the regimes obtained for ERA-Interim, as can be seen in Table 4.1. The exception is AR−. However, the sample size for the regime identification in ERA-Interim is limited. Therefore, the poorer regime correspondence between ERA-Interim and the constrained results for AR− does not necessarily mean that the constrained SEAS5 AR− regime is incorrect. It may instead indicate that assigning SEAS5 data to the ERA-Interim regimes might not be the best approach for identifying a robust and statistically significant regime signal.

Overall the regularisation ensures that NAO+ occurs more often, while NAO−, AR+ and SB− occur less often compared to the unconstrained results. To study in detail how the assignment of the data changes between the unconstrained and constrained results, we look at the contingency table given in Table 4.2. For the cases where significant amounts of data are reassigned to a different

	NAO+	NAO−	AR+	SB+	AR−	SB−
Standard	0.92	0.98	0.97	0.91	0.95	0.96
Regularised	0.96	0.96	0.95	0.95	0.71	0.96

Table 4.1: The pattern correlation of the SEAS5 regimes (standard and regularised for $\phi = 0.94$) with the ERA-Interim regimes for domain A.

regime by using the constraint (over 5000, shaded red) we compute composites (Figure 4.6) to look into the z500 anomaly-structure of this data and interpret the changes.

A large proportion of the data that without constraint was assigned to AR− is assigned to NAO+, explaining the latter's increase in occurrence rate. In turn, AR− now contains a substantial part of the data that without constraint was assigned to NAO−. This reflects the change in the AR− regime with a higher positive anomaly in the north for the regularised results, which is exemplified by the composites shown in Figure 4.6. It also can be linked to the slight strengthening of the positive z500 anomaly for NAO−, as data with a relatively weak anomaly moves to the AR− regime. Interestingly,

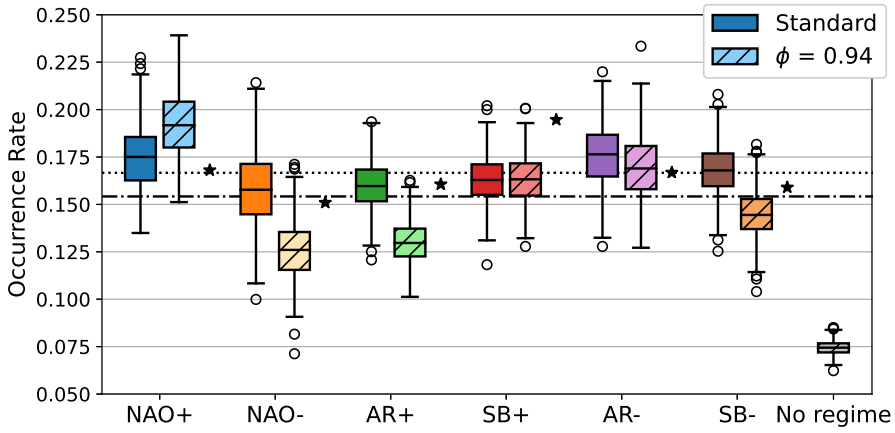


Figure 4.4: The overall occurrence rates of the different regimes for the results with and without a constraint. The boxes show the interquartile range (IQR) for bootstrapping with one (random) ensemble member per year, the whiskers extend 1.5 times the IQR on top of this (99.3% of the data falls within this range) and the circles are outlier points. Stars indicate the ERA-Interim values. The dotted line gives the $\frac{1}{6}$-line corresponding to an equal distribution of the data over the regimes, while the dash-dotted line corresponds to an equal distribution after correcting for the no-regime rate. Note that there is no ERA-Interim data assigned to no-regime by the way this category is defined using the outcome of the regularised algorithm.

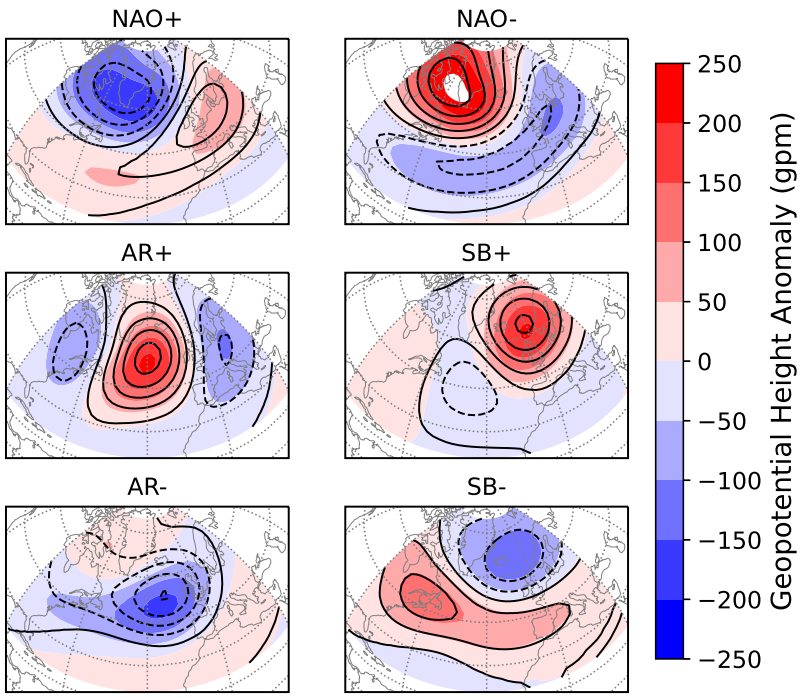


Figure 4.5: The regimes for domain A using a constraint value of $\phi = 0.94$ (colour) and without constraint (contours, with the same 50 gpm difference between contour levels).

		$\phi = 0.94$							
		NAO+	NAO−	AR+	SB+	AR−	SB−	No-regime	Total
Unconstrained	NAO+	27994	0	1443	5367	54	1297	2934	39089
	NAO−	0	23301	87	37	9432	310	2069	36227
	AR+	14	3409	25867	2349	480	426	2981	35526
	SB+	412	1132	208	28005	3881	50	2670	36358
	AR−	12512	0	640	71	22190	795	2757	38965
	SB−	1946	104	722	685	1474	29265	3254	37450
	Total	42878	27946	28967	36514	37511	32143	16656	222615

Table 4.2: A contingency table for the assignment of the SEAS5 data to the regimes for the results without constraint and for $\phi = 0.94$. Each column indicates the regime assignment following standard k-means clustering of the constrained regimes. Values over 5000 data points are coloured, blue for the same regime and red for a different regime.

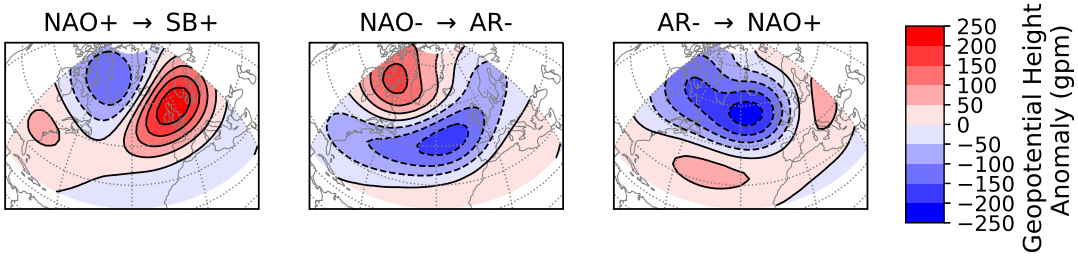


Figure 4.6: Composites of the most frequent regime reassignments between the results without constraint and for $\phi = 0.94$ (as shown by the red highlighting in Table 4.2). Shown are composites of data which without constraint are assigned to NAO+, NAO− and AR−, but which with constraint are assigned to SB+, AR− and NAO+ respectively.

NAO+ loses part of its data points to SB+ when the constraint is used. This concerns data with a north-western negative and north-eastern positive z500 anomaly (see Figure 4.6) where the balance of regime assignment is shifted by the regularisation. The unconstrained NAO+ has a relatively high positive z500 anomaly with its centre over the North Sea, which is lower when the constraint is used, corresponding to the positive z500 anomaly of SB+ being slightly weaker and located further south. The decrease in occurrence rate of AR+ is due to data having slightly off-centre positive anomaly areas now being assigned to NAO− or SB+. Data assigned to SB− when no constraint is used form the largest part of the no-regime set of the data, accounting for the majority of SB−'s decrease in occurrence rate.

Changes in transition probabilities between the regimes as a consequence of the implementation of the constraint are roughly in line with changes in the occurrence rates (not shown). That is, regimes that occur more often become more persistent and less likely to transition to another regime, and the other way around for regimes that occur less often. One notable change is the increase in the number of transitions from NAO− to AR−, which is due to the change in the AR− regime ensuring both regimes have an area of positive z500 anomalies over Greenland (Figure 4.5). This change is one-way, as there is no increase in the transition probability from AR− into NAO−.

The above discussion of the effect of regularising the clustering algorithm shows that the regularisation works as expected. That is, the occurrence rates of the regimes become more distinct indicating that a more pronounced regime signal is identified. The changes in the regime patterns are in line with changes in the assignment of the data to them and no unexpected changes in transition probabilities are found. In the next sections we turn to discussing the non-stationary behaviour of the regimes. We start with a brief discussion of the sub-seasonal signal, followed by a more detailed study of the interannual signal, including a discussion on predictability and the signal-to-noise issue.

4.4.2 Sub-Seasonal Regime Variability

Since we consider anomaly data with respect to a constant background climatological state it is expected that there is a seasonal signal in the occurrence rates of the regimes. For SEAS5 this is shown in Figure 4.7 by the dash-dotted black line for the ensemble-regularised results. The sub-seasonal variability obtained using standard k-means clustering is shown as well (dotted black line), exhibiting similar behaviour in time as found for the regularised results. The sub-seasonal variability for ERA-Interim falls within the ensemble spread of the SEAS5 results. This variability is not shown since the sample size is too small to draw reliable conclusions.

NAO+ exhibits the largest variability throughout the season with a maximum occurrence rate close to 0.3 in mid-January and a minimum below 0.1 in March, with the identified variability being amplified by the regularisation. All other regimes exhibit a seasonal cycle as well, where the amplitude of the variations differs between the regimes. We see that AR+ and AR− have a peak in occurrence rate in February, whereas NAO−, SB+ and SB− have a minimum in January and/or February. Most of this variability exceeds the sampling uncertainty as shown by the shaded area bounded by the grey dotted lines. Looking at the variability on sub-seasonal timescales found in other studies, comparison is difficult because the number of regimes considered is different (e.g. Cortesi et al., 2021). The seven year-round regimes of Grams et al. (2017) come closest and show comparable dynamics for AR+, AR− (Atlantic trough) and SB− (Scandinavian trough), while for NAO+ (zonal regime) an opposite signal emerges with lowest occurrence rates in January (when looking at DJFM). This may be linked to the different way in which a no-regime state is determined. A similar story holds for NAO− (Greenland blocking) although the difference is less robust.

To study whether this sub-seasonal variability is solely due to the changing background state within winter we correct for this effect. This is done by computing an anomaly data set with respect to a sub-seasonal climatology, instead of a fixed one, and assigning the obtained z500 fields to the closest of the regimes shown in Figure 4.5. The sub-seasonal climatology is computed by fitting a fourth order polynomial to the daily averaged fields. The assignment of the sub-seasonal anomaly data is done by first computing the distance of the data to the regimes and then minimising the clustering functional $\mathbf{L}$ over all values of Γ subject to the constraint (for $\phi = 0.94$), i.e. we apply the second step of the constrained clustering algorithm (Section 4.2.2). This ensures that the corrected occurrence rates are comparable with the standard constrained results. Note that this will result in more data being assigned to no-regime, as the minima of $\mathbf{L}$ for the sub-seasonal anomaly data are likely to differ slightly from those of the anomalies with respect to a constant climatology.

An alternative approach would be to cluster the data again after having removed the sub-seasonal climatology, however this would require rerunning the clustering for each choice of the sub-seasonal climatology one wishes to consider, whereas the approach taken here does not require re-running the clustering.

The sub-seasonal occurrence rates corrected for the seasonal cycle of the mean climatology are shown in colour (solid line) in Figure 4.7. As expected the occurrence rate of no-regime has increased, with an approximate doubling of the number of data points that are difficult to assign. This leads

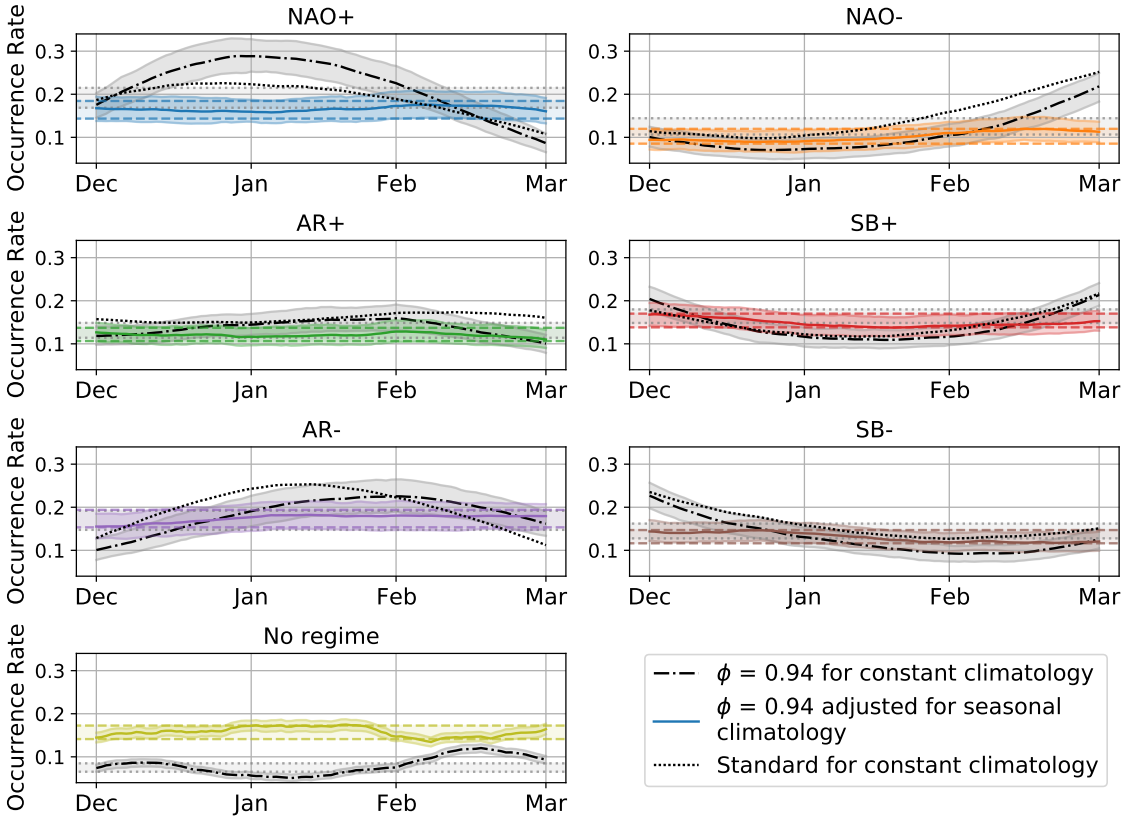


Figure 4.7: The sub-seasonal variation in occurrence rates (30-day running mean) of the different regimes for the constrained SEAS5 results with respect to a constant climatology (grey, dash-dotted black line) and corrected for a seasonally varying background state (colour, solid colour line) with the shaded area corresponding to the 2-standard deviation range. The shaded areas bounded by the dotted and dashed lines give the 2-standard deviation noise level for the constant and seasonal climatology results respectively. Error bounds are determined using bootstrapping with 3 members for every 30-day period in each year. In addition the sub-seasonal variation obtained using standard k-means clustering for SEAS5 is shown by the black dotted line.

to an overall decrease in the occurrence rate of the regimes, which is largest for NAO+. The sub-seasonal variability is significantly reduced after correcting for the background climatology, and for all regimes the average falls within the sampling uncertainty. There still is some variability, e.g. SB− appears to be more likely in early winter, but this is not statistically significant. Thus, we do not find any significant sub-seasonal variability in the regime occurrence rates when using a sub-seasonal climatology. We conclude that the seasonal cycle in the occurrence rates primarily reflects the seasonal cycle in the mean climatology, rather than any seasonal cycle in the variability itself.

Note that any attempt to correct for a varying climatology will be dependent on the choice of sub-seasonal climatology. For example, when one uses a 90-day running mean as reference climatology (e.g. Grams et al., 2017) it is possible that part of the sub-seasonal signal in occurrence rates seen in Figure 4.7 remains. Furthermore, often meteorological data is grouped according to the season (e.g. DJF) and sub-seasonal variations are not considered to first order. Thus, we deem it better to use a fixed climatology before clustering and identify the sub-seasonal signal afterwards. This way there is no assumption on the form of the sub-seasonal climatology, which could affect the regimes themselves and the attribution of the data to them.

4.4.3 Interannual Regime Variability

In Figure 4.8 the wintertime interannual variability in regime occurrence rates is plotted for the ensemble-regularised results (colour), as well as for standard k-means clustering for comparison (black dashed). These results are with respect to a constant climatology; the interannual variability when correcting for a seasonal background climatology is comparable (not shown). The signal identified in the interannual variability, as indicated by the SEAS5 ensemble mean, is slightly stronger for the regularised results compared to that obtained without constraint. The average (over all regimes) mean standard deviation of the interannual occurrence rate for the bootstrapped results (using 25 members) is 0.029 with a 95% confidence interval of [0.027, 0.031] for the regularised results, while it is 0.026 [0.024, 0.028] without the constraint. Thus the variability on interannual timescales is slightly amplified by using the regularisation. In this and the next paragraph the discussion is focused on the regularised results. Note that the standard deviation of single ensemble members is of the same order as that of ERA-Interim, albeit slightly smaller on average (0.087 versus 0.104). For NAO+, AR− and SB− we find strong interannual variability in the ensemble mean occurrence rates, whereas for AR+ and SB+ no significant signal is found. Interestingly, the interannual variability in the occurrence rate of NAO− is weaker than that of NAO+, suggesting a

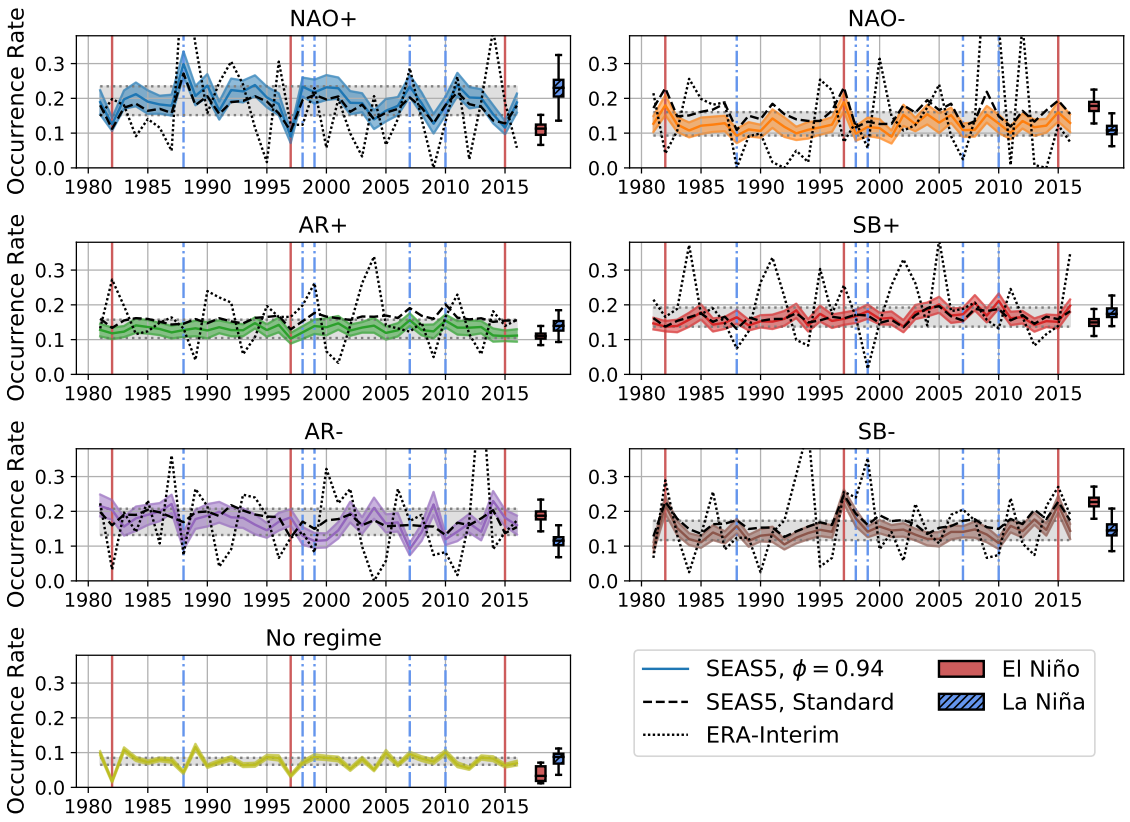


Figure 4.8: The yearly winter occurrence rates of the different regimes for the constrained results (colour). The grey bands give a noise level, the dashed black line shows the unconstrained SEAS5 results and the dotted black line the ERA-Interim occurrence rates. The bandwidth is given by the 2-standard-deviation range when using bootstrapping with 25 ensemble members, where for the noise level ensemble members for different (random) years are considered. On the right the average occurrence rates of very strong El Niño years (indicated by the solid red lines) and strong La Niña years (indicated by the dash-dotted blue lines) are shown.

smaller predictable signal.

The majority of the signal coincides with El Niño or La Niña years, as shown in the boxplots on the right-hand side of Figure 4.8. Here we refrain from separating early (Nov & Dec) and late winter (Jan & Feb) as is sometimes done (e.g. Moron and Plaut, 2003; Ayarzaguena et al., 2018), because we would not want to include November (due to the initialisation on November 1st). During winters in which there was a very strong El Niño, indicated by the solid red lines, we find an increase in occurrence of SB− and NAO−, and a decrease of NAO+. In those years there also is less data that cannot be attributed to one of the regimes, indicating that the ensemble members are more similar

in their dynamics. On the other hand we see an increase of NAO+ and decrease of AR− occurrence during strong La Niña winters (indicated by the dash-dotted blue lines), with the highest NAO+ value in 1988-89 for both SEAS5 and ERA-Interim. This is in line with previous studies looking into links between ENSO and the NAO (Li and Lau, 2012; Toniazzo and Scaife, 2006; Ayarzaguena et al., 2018, see also Figure 4.10), although here we capture this relation in more detail by using the regime variability. Note that data assigned to both NAO+ and SB− would tend to be assigned to the positive phase of the NAO when considering only four regimes. However, their response to a strong El Niño is opposite in sign, i.e. SB− becomes more frequent while NAO+ occurs less often. The distinction between these two regimes thus allows for better understanding of the details of the response of the circulation to ENSO.

4.4.4 Predictability of the Regime Signal

To see whether the SEAS5 ensemble provides a predictive signal for the ERA-Interim occurrence rates we regress the ERA-Interim annual occurrence rates against those of SEAS5. The rationale behind this regression approach is as follows. Assume there is a true signal given by $c(t)$. For an observational time series $y(t)$ we can then write

$$y(t) = ac(t) + e_y(t), \tag{4.7}$$

where $e_y(t)$ represents noise. Note that we explicitly allow for the possibility that the index (e.g. the regime frequency) we consider is only a projection of the 'true' signal, which probably is not perfect, hence $a \leq 1$. In a similar way we have a statistical model for the time series of an ensemble member $x_i(t)$ given by

$$x_i(t) = bc(t) + e_{x_i}(t), \tag{4.8}$$

now with a coefficient b as the model likely is imperfect. For the ensemble mean $\bar{x}(t)$ we then obtain

$$\bar{x}(t) = bc(t) + e_{\bar{x}}(t). \tag{4.9}$$

We regard the ensemble mean as the best estimate of the signal, and ask how well it can predict the observations. Thus we regress $y(t)$ onto $\bar{x}(t)$, i.e. estimating $y(t) = A\bar{x}(t) + E_y(t)$, which yields $A = a/b$ as the regression coefficient. This is the ratio of signal strengths, with the model prediction being well calibrated if $a = b$.

The regression coefficient thus provides information on the signal strength without having to explicitly address the noise of the observations, nor of the model. Since estimates of the noise in the observations (i.e. e_y) are especially uncertain, it is beneficial to avoid having to quantify them when

	Regime	NAO+	NAO−	AR+	SB+	AR−	SB−	NAO-index	MLR	NAO−
$\phi = 0.94$	Regression coefficient	0.99	0.98	-0.69	0.53	-0.32	1.14	1.872	NAO+ SB−	-1.35 -1.76
	p-value	0.04	0.35	0.50	0.49	0.58	0.03	0.01		0.03
	Bayes Factor	8.81	1.60	1.28	1.29	1.18	13.20	53.82		54.10
Standard	Regression coefficient	1.23	1.11	-0.22	0.65	1.29	1.02		NAO+ SB−	-1.44 -1.72
	p-value	0.05	0.15	0.81	0.51	0.14	0.10			0.06
	Bayes Factor	7.39	3.04	1.03	1.26	3.15	4.42			16.10

Table 4.3: The results for linear regression of the regime occurrence rates, linear regression of the NAO-index and MLR of the ERA-Interim NAO− against the SEAS5 ensemble mean NAO+ and AR− occurrence rates for both the regularised and standard results.

estimating the signal strength. The regression coefficient a/b can be linked to more conventional measures of signal strength such as the Anomaly Correlation Coefficient (ACC) or the Ratio of Predictable Components (RPC) (Eade et al., 2014). One can derive that

$$\text{ACC} = \frac{a}{b}\frac{\sigma_{\bar{x}}}{\sigma_y}, \qquad \text{RPC} = \frac{a}{b}\frac{\sigma_{x_i}}{\sigma_y}, \tag{4.10}$$

where $\sigma_{y,\bar{x},x_i}$ are the standard deviations of the residuals for the respective variables. So long as $\sigma_{x_i} \approx \sigma_y$, the RPC provides the same information as the regression coefficient. However the regression coefficient is a more robust estimate, as it does not require estimating the noise.

The scatter plots for the regression of ERA-Interim against SEAS5 occurrence rates are shown in Figure 4.9. The slopes and p-values are given in Table 4.3 for each of the regimes for both the constrained and unconstrained results. In addition, the Bayes factor is given. As discussed in Section 1.6 the Bayes factor is the ratio of the probability of the data given a hypothesis for two different hypotheses H_1 and H_2, i.e. $P(D|H_1)/P(D|H_2)$ (Kass and Raftery, 1995) and has been recently used in climate studies (Kretschmer et al., 2020). Here, the first hypothesis H_1 is the linear regression model and the second hypothesis H_2 is a constant occurrence rate following the overall value. A value above one indicates that H_1 is more likely than H_2, while the converse is true for a value below one. To have strong evidence towards the hypothesis of linear regression the Bayes factor would have to be much larger than one. The Bayes factor allows for the comparison of different hypotheses, whereas the p-value only indicates whether the null hypothesis can be rejected without providing an alternative (Wagenmakers, 2007; Shepherd, 2021).

This linear regression analysis indicates that there is a predictive signal for NAO+ and SB− with p-values below 0.05 using $\phi = 0.94$, while without the constraint only the NAO+ signal is found to be significant at the 95% level (Table 4.3). The Bayes factor for both constrained regimes is substantially larger than one, albeit not very large. This constitutes positive, but not yet particularly strong, evidence that the signal seen in the model is reflected in the observations (Kass and Raftery, 1995, in which values of 3-20 are said to constitute positive evidence, while values over 20 yield strong evidence). Note that these two regimes are characterised by a zonal flow pattern. Comparing the regularised result with the standard approach, the constraint adds a significant predictable signal for SB− which would not otherwise have been found. The regression coefficient is around 1 for both the NAO+ and SB− regimes, indicating just as strong a signal in SEAS5 as in ERA-Interim. For NAO− the regression coefficient is around 1 as well, but this is not significant as indicated by the high p-value and Bayes factor close to one. No predictable signal is found for the other three regimes

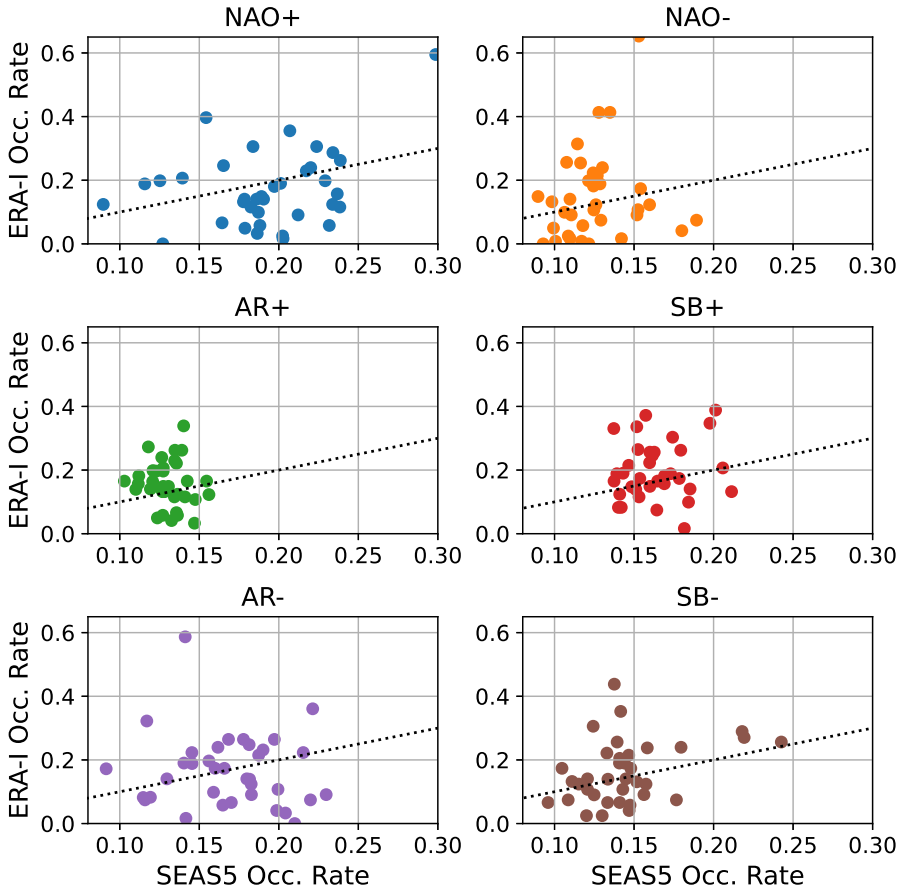


Figure 4.9: Scatter plots of the annual winter occurrence rates of ERA-Interim against those of the SEAS5 ensemble mean for each of the six regimes. The dotted lines show a one-to-one relation.

either.

The absence of a significant signal for NAO− is intriguing, as we do obtain a signal for NAO+. Interestingly, we obtain a strong predictable signal for the NAO− regime by applying multiple linear regression (MLR) using the NAO+ and SB− occurrence rates (i.e. regressing the observed NAO− onto the SEAS5 NAO+ and SB−, last column of Table 4.3). For these regimes the response of the occurrence rate of SB− to a strong El Niño is similar to that of NAO−, but that of NAO+ is opposite (Figure 4.8). The Bayes factor here is very large and constitutes strong evidence towards this being a real signal. Hence, the NAO− regime is predictable from SEAS5, just not from the SEAS5 NAO− regime signal itself. Again regularisation significantly improves the predictability, as that of the standard k-means results has a considerably larger p-value and smaller Bayes factor. Both the NAO+ and SB− regime patterns project well onto the positive phase of the NAO-index, which could in part explain the strong signal obtained from these two regimes for the predictability of the NAO− regime, which projects on its negative phase (in line with the negative regression coefficients).

4.4.5 Signal-to-noise Problem

The discussion of the signal-to-noise problem of the North Atlantic is often focused on the NAO-index. Here, we compare the regression results of the regimes with those of the NAO-index. We compute the NAO-index as the first principal component of the daily 500 hPa geopotential height fields for December till March (Weisheimer et al., 2017). The yearly NAO-index is then computed as the average index over all days in that winter and shown as the dashed black line in Figure 4.10 for SEAS5. The regression for this NAO-index is shown in Figure 4.11 with the coefficient and statistics given in Table 4.3. The signal for this NAO-index is strong with a Bayes factor of over 50 (comparable to that attained for the NAO− regime using NAO+ and SB− as predictors). The regression coefficient is roughly 2, indicating that the SEAS5 model is underpredicting the signal in the observed NAO-index by about a factor of 2. Assuming that the variance of the error is comparable between observations and model ensemble members, this result is in line with previous studies on the signal-to-noise paradox for the NAO where RPC ≈ 2 has been found as a lower bound (Eade et al., 2014; Scaife and Smith, 2018). (We also computed the RPC directly and found a value of around 2.)

Thus, there is a significant difference in the model representation of the regimes compared to that of the NAO-index. While for the NAO-index we see an underestimation of the signal in the

model compared to observations, in line with the signal-to-noise paradox, this is not the case for the signal in the occurrence rates of the two zonal regimes NAO+ and SB−, which are the regimes with interannual predictability. In order to analyse whether this discrepancy is due to only considering the *occurrence rates* of the regimes, we need to address whether there is a possible signal-to-noise problem in the *amplitude*, i.e. strength, of the regimes. To this end we compute the average NAO-index for each regime, i.e. averaging the NAO-index over all days assigned to a regime, for both SEAS5 and ERA-Interim. The results are shown in Table 4.4. As expected the NAO+ and NAO− regimes contribute most to the respective phases of the NAO-index. Approximating the NAO-index in SEAS5 using the annual occurrence rates and average NAO-indices for these two regimes, that is multiplying the NAO− and NAO+ regime NAO-indices from Table 4.4 by their annual occurrence as shown in Figure 4.8 and adding the two, provides a good estimate of the NAO-index variability for both regularised and standard results as can be seen in Figure 4.10. In addition we compute the average annual winter NAO-indices for each regime in SEAS5, which are found to be uncorrelated with their respective regime frequencies (not shown). This indicates that the regime occurrence and the regime strength (in terms of its projection on the NAO-index) are independent. Hence we do

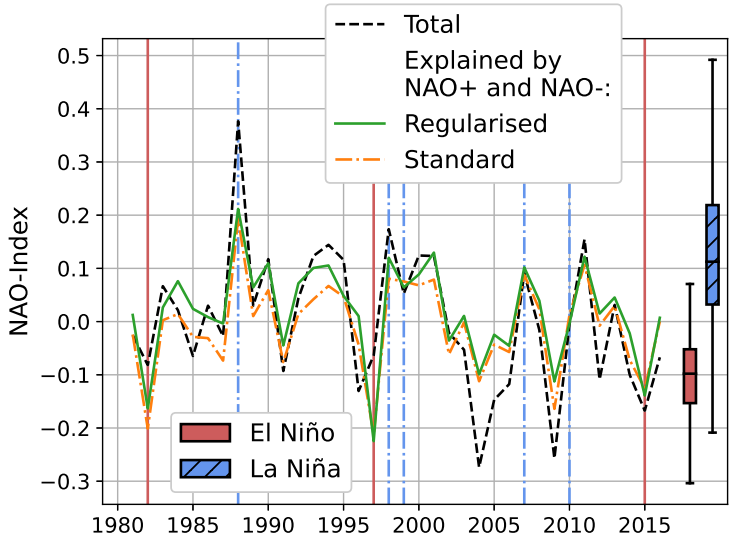


Figure 4.10: The SEAS5 NAO-index (black, dashed) and an approximation using the NAO+ and NAO− occurrence rates and projected NAO-indices for the regularised (green, solid) and standard (orange, dash-dotted) approach. On the right the average NAO-indices of very strong El Niño years (indicated by the solid red lines) and strong La Niña years (indicated by the dash-dotted blue lines) are shown.

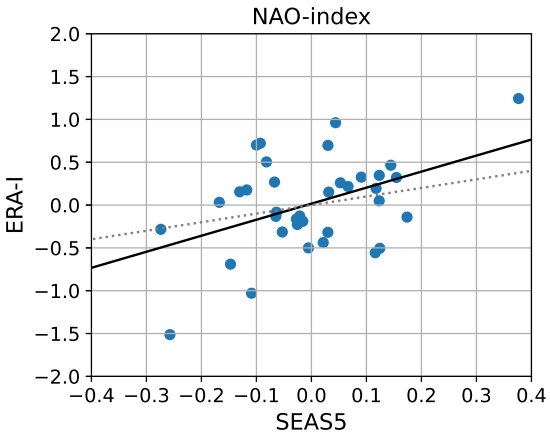


Figure 4.11: Linear regression of the NAO-index (black solid line), being the winter-average of the first PC of the DJFM daily z500. The dotted line shows a one-to-one relation.

Regime	NAO+	NAO−	AR+	SB+	AR−	SB−
SEAS5 Regularised	1.26	-1.79	-0.31	-0.16	-0.33	0.69
SEAS5 Standard	1.33	-1.53	-0.53	-0.47	0.30	0.69
ERA-Interim	1.38	-1.69	-0.31	0.06	0.06	0.31

Table 4.4: The average NAO-index for each of the six regimes in both SEAS5 (regularised and standard) and ERA-Interim.

not find evidence of a signal-to-noise problem in relation to the regime strengths, e.g. a regime is not weak when its occurrence rate is high.

This leaves us with the discrepancy between the signal strength for the regime frequencies and for the NAO-index. Using the regularisation, we found that SEAS5 has a predictable signal for the two zonal regimes with a regression coefficient around one. On the other hand, no signal was found for the non-zonal regimes and the NAO− signal was not manifest directly, though could be detected from NAO+ and SB−. Thus, the signal-to-noise paradox for the NAO-index might be linked to certain regimes being poorly represented within the model, i.e. the NAO-index cannot provide all the relevant information of the atmospheric flow structure for predictability in the Euro-Atlantic sector. It is not necessarily the case that the amplitude of the predictable signal in response to remote forcings such as ENSO is too weak (Scaife and Smith, 2018), but rather that the signal is only present in part of the dynamics, while other aspects are incompletely represented. The first regime to consider in this regard is NAO−, which represents a blocking over Greenland, as it is unsuccessfully predicted from the SEAS5 NAO− regime, even though a strong signal has been

identified using the NAO+ and SB− regimes. This also points to the negative phase of the NAO being at the heart of the signal-to-noise problem.

4.5 Summary and Discussion

To identify a regime variability signal in model hindcast ensemble data a constraint on the similarity between ensemble members has been implemented. In this way a stronger and more informative regime signal is identified by considering the trade-off between accuracy and entropy. Different criteria are used to identify the optimal settings for the ensemble-regularisation method, yielding an optimal constraint value. This optimal value is sufficiently strong to increase the information gain (as indicated by the entropy), but not so strong to lose a lot of accuracy (as indicated by the BIC). The constraint helps better discriminate between the different regimes, which is reflected in the overall occurrence rates of the regimes being more distinct. The regime patterns themselves are not strongly affected, increasing confidence in this approach.

When considering the non-stationary regime dynamics, we find that the average sub-seasonal variability is primarily determined by variability in the average background climatology. When one looks at a seasonal climatology, such as the DJF average, a large part of the found variability will remain. A question when removing a background climatology based on daily averages is whether one is removing part of the signal, for the differences in regime occurrence throughout the season do reflect the changes in the background climatology. Therefore, we regard it as cleaner to consider a constant climatology within the season and account for the background variability in the interpretation.

On interannual timescales the NAO+, NAO−, AR− and SB− regimes show significant variability, which is enhanced by the ensemble-regularisation compared to standard k-means clustering results. In large part this is related to ENSO, with El Niño leading to SB− and NAO− being more frequent, while La Niña corresponds to increased occurrence rates of NAO+ and decreased frequencies of AR−. When considering only four regimes most data now assigned to either NAO+ or SB− would be allocated to the NAO+ regime. Thus the use of six regimes (instead of four) allows for a more nuanced view of the regime response to ENSO. Note that the limited length of the timeseries available, with only a hand-full of ENSO events, means that the robustness of these responses cannot be fully established.

We have used linear regression to identify the signal on interannual timescales, as it allows a direct estimation of the ratio of signal strengths between observations and model, without requiring

estimation of the noise levels. The SEAS5 ensemble has a predictable signal for the occurrence rates of the two zonal regimes (NAO+ and SB−), but not for the other regimes. The Bayes factors show a substantial improvement of predictability, especially for SB−, with the regularisation compared to the standard results. Interestingly, a strong predictive signal for NAO− is obtained by considering multiple linear regression using the model NAO+ and SB−, whereas no such signal is found using the model NAO− frequencies. The regression coefficients that come out of the linear regression are around one for both NAO+ and SB−, indicating that the SEAS5 signal is of the same magnitude as that found in ERA-Interim. This implies there is no signal-to-noise paradox for these two flow regimes. Note that also for NAO− the regression coefficient is around one, but is not statistically significant.

In contrast we find that for an NAO-index the regression analysis results in a regression coefficient of 2, in line with previous studies on the signal-to-noise paradox for the North-Atlantic sector that show that the model underpredicts the observed NAO by a similar factor (e.g. Eade et al., 2014). Our regime analysis suggests that the NAO signal-to-noise paradox largely manifests itself in the non-zonal phase of the NAO (and related regimes), i.e. in its negative phase rather than its positive phase. Improving the regime representation of NAO− (and AR−) within the SEAS5 model could improve not only the regime dynamics, but also help shed light on the signal-to-noise paradox over the Euro-Atlantic domain.

Chapter 5

A Bayesian Approach to Regime Assignment

In the previous two chapters we discussed how to identify robust and pronounced regime signals by applying regularised k-means clustering methods to gridpoint data. In Chapter 3 the regularisation enforced a level of persistence, helping to identify the true regime persistence in atmospheric circulation data. In Chapter 4 the regularisation was on the model ensemble, enforcing a level of similarity between ensemble members to identify the non-stationary variability signal. Both these approaches constrain the clustering method in a way that allows to identify regime signals within data that is affected by noise. A drawback of these regularisation approaches is that there is a parameter that needs to be selected. For this selection we have used different criteria, such as the Bayesian Information Criterion (BIC). However, it would be desirable to have an approach that does not rely on such parameter selection. Another limitation of (regularised) k-means clustering is that it yields a hard, categorical, assignment of the data to the regimes. As a consequence it is difficult to quantify the uncertainty of the regime assignment, since data close to the regime centre is treated the same as data that is only just (by distance) assigned to that regime.

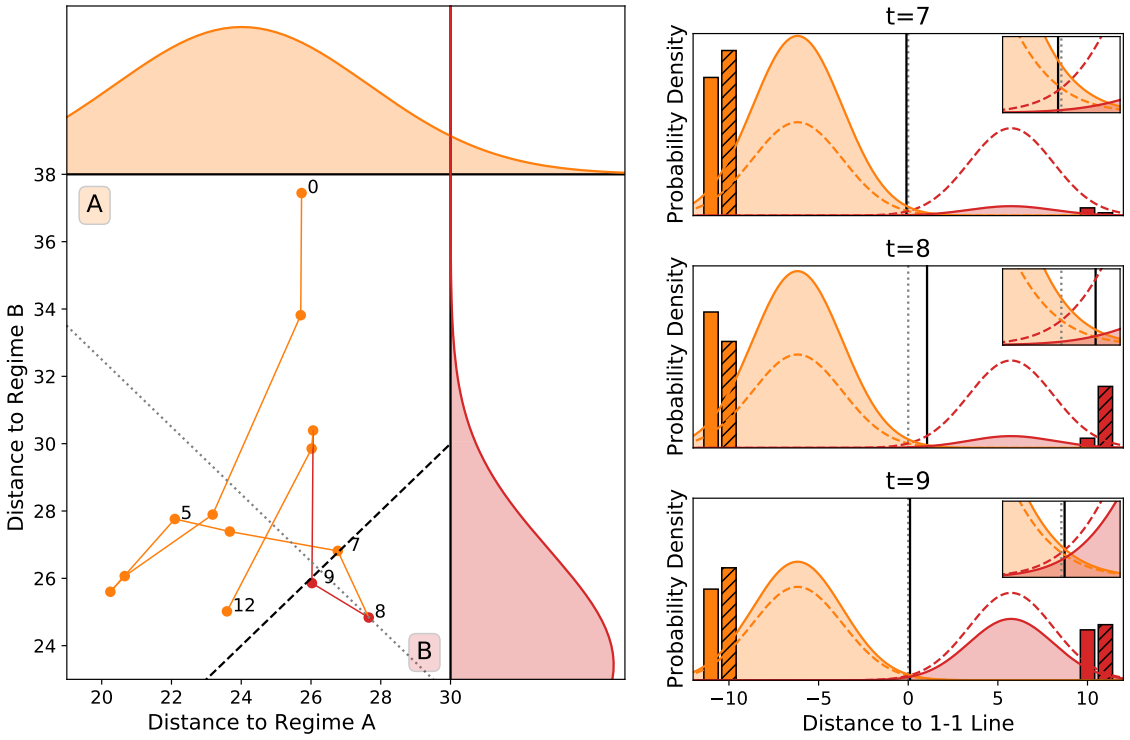
In this chapter a Bayesian approach to obtain a probabilistic regime assignment is proposed. Such a probabilistic regime assignment can overcome some of the drawbacks of a hard regime assignment, while not requiring a regularisation parameter to be selected. In the first section the rationale behind this Bayesian approach is explained. Next, after a brief discussion of the data and the standard clustering methods used for comparison in Section 5.2, two methods of Bayesian regime

assignment are discussed. The first sequential approach makes use of knowledge on the persistence of the regime dynamics and is discussed in Section 5.3. The second ensemble approach employs the information within the ensemble to focus on the non-stationary signal, similar to the regularisation in Chapter 4, and is explained in Section 5.4. A summary and brief discussion are given at the end.

5.1 Motivation Towards a Probabilistic Regime Assignment

As mentioned, the hard regime assignment of k-means clustering means that the result is susceptible to noise. Consider Figure 5.1a which shows the distance of the data to two regimes in time for a real case (discussed later in detail), over a period of 12 days. Initially, the data clearly is categorised to belong to regime A, being significantly closer in distance to regime A than to regime B. However, from day 7 to 9 the data makes a brief excursion into a part of the phase diagram that is closer to regime B, after which it moves back to being closest to regime A. The question is whether this is a real signal or simply the effect of noise in the observed distance. Since the regime dynamics is quite persistent in time it is likely to be the latter, but this possibility is not picked up by the hard assignment of a standard k-means clustering approach. Often a low-pass filter is applied to remove this high-frequency variability (e.g. Straus et al., 2007; Grams et al., 2017), but in Chapter 3 it was shown that low-pass filtering can lead to a bias in the observed regime frequencies. Another solution is to use a regularised clustering algorithm which constrains, or bounds, the number of transitions between the regimes so that is in line with the natural metastability of the underlying dynamics (Chapter 3, Falkena et al., 2020). Such a method allows to better identify the signal within the noise, but does require selecting a constraint parameter. This introduces a parameter selection, where e.g. an information criterion is used to decide on a suitable constraint value.

An alternative approach is to make the regime assignment probabilistic rather than deterministic, allowing for a more nuanced and informative regime assignment in the presence of noise. Methods such as mixture modelling provide such a probabilistic regime assignment (e.g. Hannachi and O'Neill, 2001; Smyth et al., 1999), but are not widely used. In studies that look into forecasting of regimes on sub-seasonal timescales, the probability of being in a regime is often considered by looking at the empirical distribution of the (hard) regime assignment across an ensemble (e.g. Vigaud et al., 2018; Cortesi et al., 2021; Büeler et al., 2021, and Chapter 4). A limitation of this method is that it requires availability of ensemble data, where typically the ensemble size is small, and verification is done against a hard regime assignment from reanalysis. A probabilistic regime assignment that does not require this availability of ensemble data would help in better assessing the skill in predicting



(a) (b)

Figure 5.1: A conceptual example of the difficulty k-means clustering has when noise affects the data, showing what a probabilistic approach can bring. (a) An example trajectory of the data as a function of the distances to two regimes A (orange) and B (red). The 1-1 line is shown black dashed, meaning the region above is closer to regime A and the region below to regime B. Numbers indicate the day corresponding to that point in the trajectory. The likelihood functions shown along the top and right give the climatological probability of those distances given hard assignment to regime A (orange, top) or B (red, right). The dotted grey line indicates a slice through the probability space along which the pdfs in panel (b) are considered. (b) A slice of the likelihood functions, weighted by the prior probabilities following Bayes Theorem, for each of the regimes (solid lines, A: orange, B: red) along the grey dotted line in (a), perpendicular to the 1-1 line, for the 7th, 8th, and 9th day. The location of the data on each day is indicated by the vertical black lines, and the bars at the edge of the plots show the prior (left) and posterior (right, hatched) probabilities for each of the regimes (A: orange, left edge, B: red, right edge). The climatological likelihood functions are shown dashed in all panels and the vertical grey dotted line indicates the location of the 1-1 line. The insets in each panel show an enlargement of the region around the 1-1 line.

regimes, as it could be applied to reanalysis data, which is also subject to noise, as well. Such a regime assignment would allow to identify the instances in which the observations cannot be clearly assigned to one regime or in which a wrong hard assignment is potentially due to noise. This approach allows for a fairer verification of the model by taking some degree of observational uncertainty into account.

For a probabilistic regime assignment it is the probability of being in a regime given the data, i.e. $P(\text{Regime}|\text{Data})$, that is of interest, where the data considered here is the location in phase space at a given time. Following Bayes Theorem this is given by

$$P(\text{Regime}|\text{Data}) = \frac{P(\text{Data}|\text{Regime})P(\text{Regime})}{P(\text{Data})}, \tag{5.1}$$

combining prior knowledge of the probability of being in a regime $P(\text{Regime})$ with an observed likelihood given a regime $P(\text{Data}|\text{Regime})$. The latter can be computed from the climatological data, which can be obtained from e.g. a standard k-means clustering result. In Figure 5.1a the observed (climatological) likelihood functions for both regimes are shown next to the trajectory. The working of Bayes Theorem for such a trajectory is shown in Figure 5.1b. The climatological likelihood functions of the two regimes A and B, as indicated by the dashed lines, are re-weighted using the prior regime probabilities, shown by the non-hatched bars at the edge of the panels. This yields two new likelihood functions (solid lines) with respect to which the posterior probabilities are computed as the values of the likelihood functions at the datapoint (vertical black line). The obtained Bayesian probabilities are indicated by the hatched bars and used to inform the prior probabilities for the next time-step, employing climatological information about transition probabilities.

Figure 5.1b shows how the inclusion of prior information $P(\text{Regime})$ following Bayes Theorem (5.1) affects the posterior $P(\text{Regime}|\text{Data})$ for the trajectory at days 7, 8 and 9, following a section along the dotted line in Figure 5.1a. At day 7 the prior information indicates a very high probability of being in regime A as all previous days belonged clearly to that regime. This increases the probability of $t = 7$ belonging to regime A and decreases that of belonging to regime B with respect to the climatological likelihood, which would otherwise be evenly balanced between the two regimes as the data lies roughly in the middle. Thus, there is a high probability that the data at day 7 belongs to regime A. Given the known persistence of regimes, the prior information for day 8 again then indicates a high probability of being in this regime, albeit slightly smaller than at $t = 7$, which weights the likelihood functions accordingly. Although the data is closer to regime B, the prior information gives that there is an approximately equal probability of being in either of the two regimes. The prior for $t = 9$ does not weight the likelihood functions as much as for $t = 7$, and thus the data at day 9 being equally close to both regimes means that again the probability of being in

either of the regimes is close to a half. This discussion shows how the inclusion of prior information can be used to compute the probability of a regime given the data, and thereby soften the effects of noise, following the fundamental principles of probability as encoded in Bayes Theorem (5.1).

Other aspects than the persistence discussed here can affect the prior regime likelihood as well. It is likely that non-stationary external factors, such as the El Niño Southern Oscillation (ENSO) or Sudden Stratospheric Warmings (SSWs), have an influence on the prior regime probabilities (e.g. Toniazzo and Scaife, 2006; Ayarzaguena et al., 2018; Domeisen et al., 2020). The Bayesian approach allows to incorporate such information in the prior regime probabilities, either by looking at e.g. an ENSO index or by making use of the availability of ensemble data. In Chapter 4 a regularised clustering method helped to identify a more pronounced interannual regime signal by making use of the information available in an ensemble. Similarly, having a more informative prior for Bayes Theorem (5.1), incorporating information from external processes, can help in identifying a stronger non-stationary regime signal.

The next step is to formalise the intuition of Figure 5.1 and study how to use Bayes Theorem to obtain a probabilistic regime assignment. Before elaborating on this method we discuss the data that is used and the application of standard k-means clustering to obtain the circulation regimes that we consider in this chapter in Section 5.2. The two sections that follow explain two different ways in which Bayes Theorem can be used for the regime assignment, translating the intuition acquired in this section into an empirical approach.

5.2 Data and k-means Clustering

For the identification of the circulation regimes the 500 hPa geopotential height fields (z500) from two datasets are used: the ECMWF SEAS5 hindcast ensemble dataset (Johnson et al., 2019) and the ERA-Interim reanalysis dataset (Dee et al., 2011). For both datasets, daily (00:00 UTC) gridpoint z500 data over the Euro-Atlantic sector (20° to 80°N, 90°W to 30°E) is considered for all winters (DJFM) for which the SEAS5 ensemble data is available (1981-2016). The regimes are computed using anomaly data, where the anomalies are computed with respect to the average DJFM climatology (see Falkena (2019) and the appendix to Falkena et al. (2020) for the rationale behind this choice). The ERA-Interim and SEAS5 anomalies are computed with respect to their own respective climatologies. The SEAS5 hindcast ensemble has 51 members and is initialised on November 1st, which means that by considering data from December onwards the effect of the atmospheric initial

conditions has been effectively lost. This allows us to treat the ensemble members as alternative realisations of the atmospheric circulation, subject to the non-stationary influences for that year (notably ENSO).

A standard k-means clustering algorithm, as discussed in Chapter 2, with a Euclidian norm to compute the distance between the data and regimes, is used to identify six circulation regimes over the Euro-Atlantic sector for both ERA-Interim and the SEAS5 hindcast ensemble (Jain, 2010). Six was identified as a suitable number of regimes for such unfiltered data in Section 3.2. The regimes for the SEAS5 hindcast ensemble are shown in Figure 5.2 and are the two phases of the North Atlantic Oscillation (NAO), the Atlantic Ridge (AR), Scandinavian Blocking (SB) and both their counterparts (as in Chapters 3 and 4). Note that these regimes are slightly different in their patterns from those of ERA-Interim (see Section 4.4.1 and Appendix B for details on this), thereby providing an inherent bias correction between the model and reanalysis. These hard, categorical, regime assign-

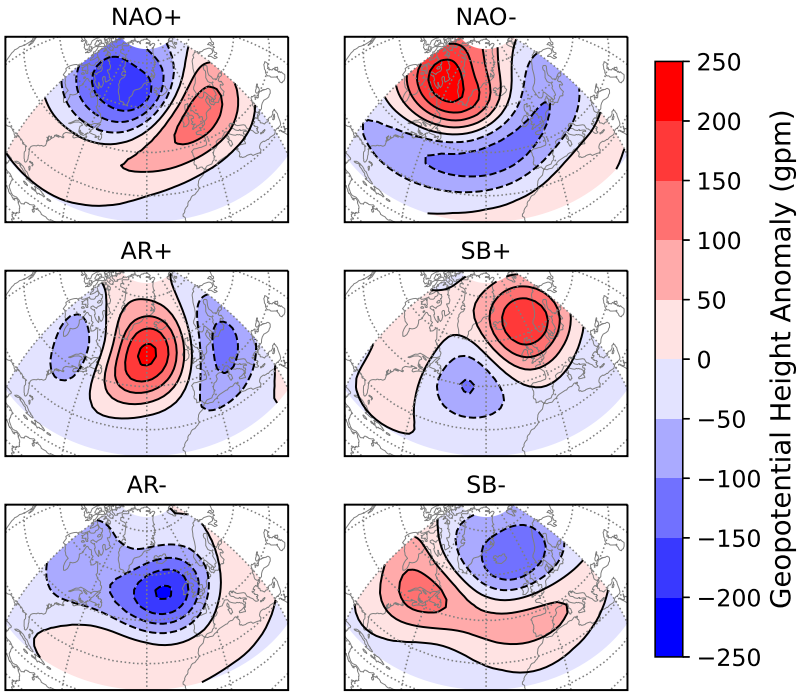


Figure 5.2: The six circulation regimes obtained for the SEAS5 ensemble using k-means clustering. From top-left to bottom-right: NAO+, NAO-, Atlantic Ridge (AR+), Scandinavian Blocking (SB+), AR-, SB-.

ments are used to compute the likelihood functions that are considered in the Bayesian approach (details in Section 5.3.1). In addition we consider the (hard) regime assignments obtained using the time-regularised clustering algorithm discussed in Chapter 3.3. This allows for a comparison of different approaches to identify the persistent regime signal.

5.3 Sequential Bayesian Regime Assignment

In this section a sequential Bayesian approach to regime assignment, which can be applied to ERA-Interim data as well as single ensemble realisations, is discussed. We start with the details of the method itself in Section 5.3.1, followed by an evaluation and comparison with the results of both a standard and time-regularised k-means clustering method in Section 5.3.2.

5.3.1 Bayes Theorem for Regime Assignment

The starting point for sequential Bayesian regime assignment is the six regimes obtained using k-means clustering discussed in Section 5.2 and shown in Figure 5.2 (and earlier Sections 3.2.3 and 4.4.1). The likelihood functions in Bayes Theorem (5.1) are computed based on the distance to these regimes and remain fixed throughout the sequential Bayesian regime assignment. The discussion of the method as phrased below is general, and can be applied to all types of regime dynamics as long as the regimes themselves and the likelihood functions are specified a priori.

Let r be a discrete random variable indicating a regime, i.e. taking values in $\{1, ..., k\}$ for k regimes, and let $\mathbf{d}_t \in \mathbb{R}^k$ be a vector containing the distances to each of the regimes at time t (here the Euclidian distance is used, which is also the standard cost function in the k-means setting (Chapter 2)). Specifically, $\mathbf{d}_t$ is the data we consider in our Bayesian approach. At a given time t we are interested in the probability to be in a regime r given the data, i.e. $P(r|\mathbf{d}_t)$. Bayes Theorem tells us that

$$P(r|\mathbf{d}_t) = \frac{P(\mathbf{d}_t|r)P(r)}{P(\mathbf{d}_t)}. \tag{5.2}$$

Here, $P(r)$ is the prior probability of regime r, which may or may not depend on time t, and $P(\mathbf{d}_t)$ is the likelihood of the data at time t. Since we only consider a discrete number of regimes which are mutually exclusive and exhaustive, the latter can be computed by

$$P(\mathbf{d}_t) = \sum_{r=1}^{k} P(\mathbf{d}_t|r)P(r), \tag{5.3}$$

making it a normalisation factor.

Lastly, $P(\mathbf{d}_t|r)$ is the likelihood of the data at time t given a regime r. The likelihood of the data can be determined from the distance to each of the regimes by considering how the data falls within the conditional distance distributions, i.e. the distributions conditioned on data belonging to one of the regimes. For each datapoint in either the SEAS5 or ERA-Interim timeseries we have this distance to each of the k regimes, which has been computed in the k-means clustering procedure to determine the hard regime assignment (Section 5.2). This gives the distributions of the distances to each of the regimes conditional on regime r, which for SEAS5 are shown in Figure 5.3. There are a few things to note concerning these distributions. Firstly, the distance to the regime the data is assigned to is generally the smallest, but can still be larger than the distance to other regimes for a different data point belonging to that regime. Secondly, for data assigned to AR+, SB+, AR− and SB− the distances to the other regimes are roughly equally distributed with the means being relatively close to each other. However, for data assigned to either NAO+ or NAO− the distance to the other phase is larger than that to the other four regimes. Thus these two regimes are further away from each other than the rest of the regimes, and information on the proximity to one regime is providing information on the proximity to the other.

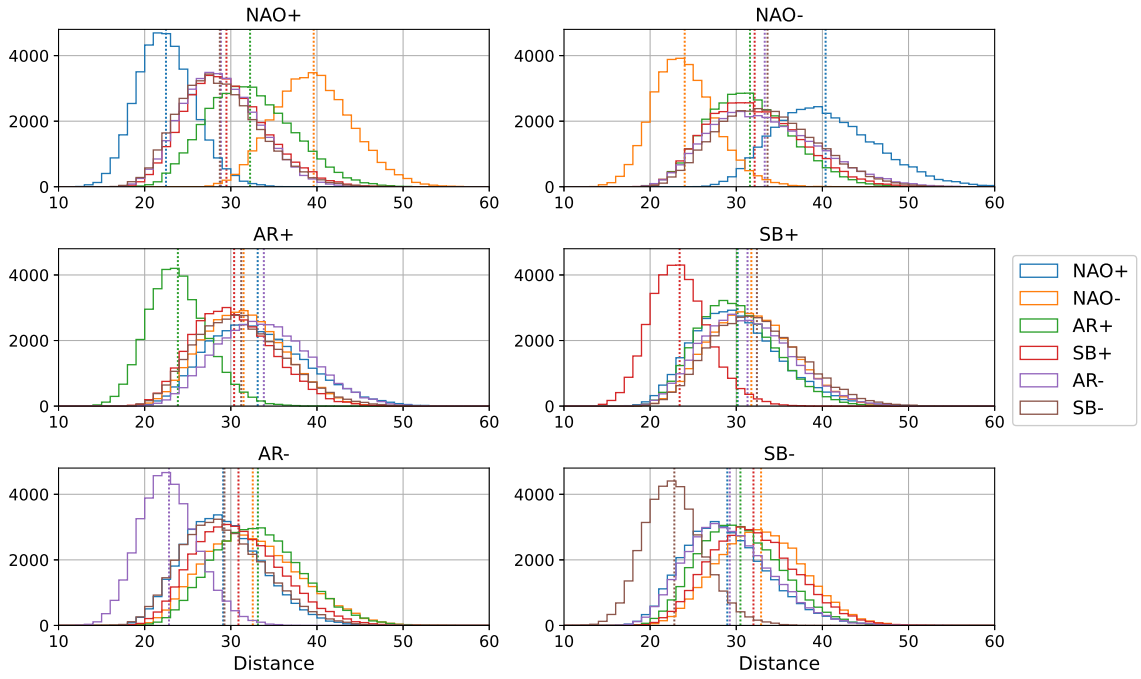


Figure 5.3: The distributions of the distances to each of the regimes (color) conditional on the SEAS5 hindcast data being assigned to the regime given in the title. The means of each distribution are indicated by the vertical dotted lines.

Also, we see that these distributions are approximately normal, justifying us to approximate the corresponding k-dimensional conditional probability density functions (pdf) by a multivariate normal. The likelihood $P(\mathbf{d}_t|r)$ is then given by the value of the conditional pdf, that is

$$P(\mathbf{d}_t|r) = \frac{\exp(-\frac{1}{2}(\mathbf{d}_t - \mu_r)^T \Sigma_r^{-1}(\mathbf{d}_t - \mu_r))}{\sqrt{(2\pi)^k |\Sigma_r|}}, \tag{5.4}$$

where $|\cdot|$ represents the determinant. The mean μ_r and covariance Σ_r, representing the variability around the cluster centre, are estimated from the conditional distance distributions obtained from the k-means clustering results for each regime. These estimates are done separately for ERA-Interim and SEAS5 to avoid biases due to the regimes being slightly different. The estimates of the mean and covariance are surprisingly similar between both datasets, indicating that, apart from the slight difference in regimes, the model does a reasonable job in representing the variability of the regime dynamics. A further discussion on this, including a robustness analysis of the distance distributions is given in Appendix C.

To obtain the prior probability $P(r)$ there is a natural choice from propagating the probabilities of the previous time-step forward, which makes the prior time dependent $P(r) = P_t(r)$. From k-means clustering an estimate of the regime dynamics is known, which is characterised by the climatological regime frequencies P^c and transition probabilities T^c_{ij} between the regimes. For SEAS5 these are given by

$$P^c = \begin{pmatrix} 0.176 \\ 0.158 \\ 0.160 \\ 0.163 \\ 0.175 \\ 0.168 \end{pmatrix}, \qquad T^c = \begin{pmatrix} 0.728 & 0.000 & 0.039 & 0.062 & 0.060 & 0.112 \\ 0.000 & 0.822 & 0.050 & 0.046 & 0.053 & 0.029 \\ 0.079 & 0.054 & 0.702 & 0.075 & 0.021 & 0.069 \\ 0.069 & 0.058 & 0.065 & 0.739 & 0.037 & 0.031 \\ 0.072 & 0.032 & 0.035 & 0.045 & 0.771 & 0.045 \\ 0.065 & 0.033 & 0.095 & 0.029 & 0.070 & 0.708 \end{pmatrix}, \tag{5.5}$$

as obtained in Chapter 4. Starting from the regime probabilities at time $t-1$, a best estimate of the prior probabilities for the next time step is

$$P_t(\mathbf{r}) = T^c P(\mathbf{r}|\mathbf{d}_{t-1}), \tag{5.6}$$

where $P_t(\mathbf{r})$ is the vector of prior probabilities $\{P_t(r)\}_{r=1,\ldots,k}$ at time t and $P(\mathbf{r}|\mathbf{d}_{t-1})$ the vector of posterior probabilities $\{P(r|\mathbf{d}_{t-1})\}_{r=1,\ldots,k}$ at time $t-1$. Note that in the transition matrix T^c the diagonal elements — corresponding to persistence of the current regime — dominate. At the start of each winter, on December 1st, there is no previous regime probability to use, and thus little prior information on the likelihood of each of the regimes. For that reason the climatological regime

frequencies P^c are used as a prior in these cases. Note that this is nearly as uninformative as using a uniform distribution.

Using the prior probabilities $P_t(r)$ and likelihood of the data $P(\mathbf{d}_t|r)$ following the conditional distance distributions we can compute the posterior probability of a regime given the data $P(r|\mathbf{d}_t)$ using Bayes Theorem (5.2) in every time-step. This can be regarded as a direct implementation of the optimal filter, which is feasible since we consider a finite state Markov chain. Following this approach, we obtain a sequential probabilistic regime assignment, where the regime probabilities of one day are used to obtain a prior for the next day. Applying this method to ERA-Interim data and the ensemble members of the SEAS5 ensemble yields a probability of being in each of the six regimes at every day in winter. From here on we refer to this posterior Bayesian probability simply as the Bayesian probability.

The above described sequential Bayesian regime assignment is simple and allows for a straightforward comparison with the commonly used hard regime assignment, as well as with the regularised clustering results (without the need of selecting a constraint parameter). However, there are other options to model the uncertainty and to update the corresponding model parameters sequentially. For instance one can model each regime individually and associate its center estimates with the mean of a Gaussian. The updating procedure for such a model is called the Kalman filter (Kalman, 1960) and the corresponding Monte Carlo approximation the Ensemble Kalman Filter (Evensen and van Leeuwen, 2000). Of course various other methods for more general distributions as well as iterative assimilation of incoming information exist (e.g. Kantas et al., 2014; Hu and van Leeuwen, 2021; Acevedo et al., 2017). The method used here is closer to a particle filter (Del Moral, 1997; Doucet et al., 2001) as our ensemble members are weighted with importance weights stemming from the likelihood rather than using an analytic formula such as being used in the Kalman filter. In this thesis the aim is specifically to stay close to the methods discussed in previous Chapters 3 and 4, and model the process of hard regime assignments as random variables in each time step. This allows for a straightforward implementation which can be readily applied in an operational setting. Furthermore, using this method we can investigate whether the results are comparable to those found using regularised clustering methods, without the need to select a constraint parameter.

5.3.2 Effect on the Regime Dynamics

The first question to answer is what the effect is of this Bayesian approach in practice, and whether this matches the intuition behind the method. How does the prior affect the Bayesian probabilities?

A next step is to compare the probabilistic approach with results obtained using a hard regime assignment, as given by k-means clustering. Is the average regime frequency affected? What is the effect on the regime persistence? In this section we start by discussing the first question by looking at some examples to get an idea of how the method is working in practice. Next we look at the statistics of the results compared to a k-means clustering approach to answer the other questions.

To start, consider the example given in Figure 5.4 for the 23rd ensemble member which shows the prior and Bayesian regime probabilities together with the climatological likelihood corresponding to the observed datapoint. A first aspect to note is that most of the time the regime likelihood $P(\mathbf{d}|r)$ gives a clear indication of the regime the data belongs to. Secondly, we see that the prior quite closely follows the Bayesian probabilities with a delay of one day, corresponding to the high persistence in the transition matrix in Equation (5.5). The initial prior, given by the climatological values, is uninformative and in that case the regime likelihood nearly fully determines the Bayesian probabilities. Subsequently, the prior is much more informative but in most cases the regime likelihood still strongly determines the final probability. However, when the likelihood does not clearly point towards one regime, e.g. around days 8-12, the prior information shifts the probabilities towards stronger persistence, in this case of the AR+ regime. This can also be seen around days 99-101, corresponding to days 7-9 in the example shown in Figure 5.1 in Section 5.1, where the inclusion of prior information increases the importance of persistence over a short excursion away from the most likely regime. In this way the Bayesian regime assignment allows for identifying stronger persistence, i.e. high probability of the dominant regime, without losing the signal of other regimes entering the dynamics as they still have some non-zero probability. The effect of this approach for ERA-Interim data is similar.

The Bayesian probabilistic regime assignment allows to understand some of the subtleties of the regime dynamics, e.g. regime transitions occur in the form of a decrease/increase of the regime probabilities. How does such an approach compare to the commonly used hard regime assignment obtained using k-means clustering? The bar at the bottom of Figure 5.4 shows the hard regime assignment corresponding to this time series. The Bayesian regime probabilities vary more smoothly, and do show less short back-and-forth transitions between regimes, which occur several times for the hard regime assignment, e.g. around day 9 and 20. In Section 3.3 a constraint on the number of transitions between regimes was introduced to reduce the number of these short back-and-forth regime transitions (Falkena et al., 2020). This was shown to increase the regime persistence without affecting the regime occurrence rates, provided the constraint parameter was chosen appropriately. In Figure 5.5 a comparison between the regime likelihood, Bayesian regime probabilities and a

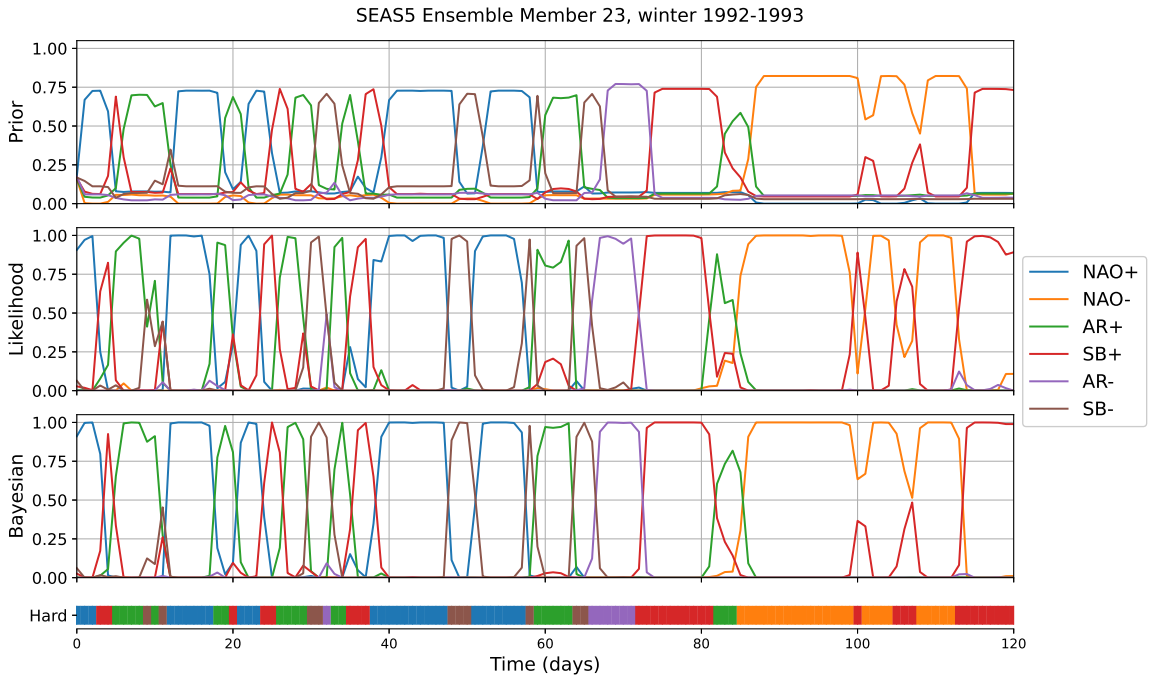


Figure 5.4: The prior probability, conditional regime likelihood and Bayesian regime probability for the 23rd ensemble member in the sequential Bayesian regime assignment procedure for the winter of 1992-1993. The bar at the bottom indicates the hard regime assignment following k-means clustering.

hard regime assignment obtained using either a standard or this time-regularised k-means clustering approach is shown for ERA-Interim for the winter of 1993-1994. The regularisation does reduce the number of regime transitions, by e.g. removing the NAO+ regime between two occurrence of SB− around day 18. At the same time the Bayesian probabilities show a small increase in the NAO+ likelihood, with SB− still having the highest probability. Here the regularisation and Bayesian approach thus yield similar results. On the other hand, around e.g. day 84 and 107 the regularisation eliminates some regime transitions where the Bayesian probabilities still show some signal of the corresponding regimes. The probabilistic approach thus allows to identify the data where the observations are less clear, showing an increase in probability instead of a hard regime change. It also retains some regime transitions that the time-regularised clustering eliminates due to it being difficult to select the "correct" constraint, which in the probabilistic approach show as increases in the corresponding regime probability. This analysis confirms that the Bayesian approach seems to be doing something sensible, without having to tune any parameters.

The impact of the sequential Bayesian approach on the regime frequencies, computed as the average Bayesian regime probability for this method, and (1-day) autocorrelation is shown in Figure

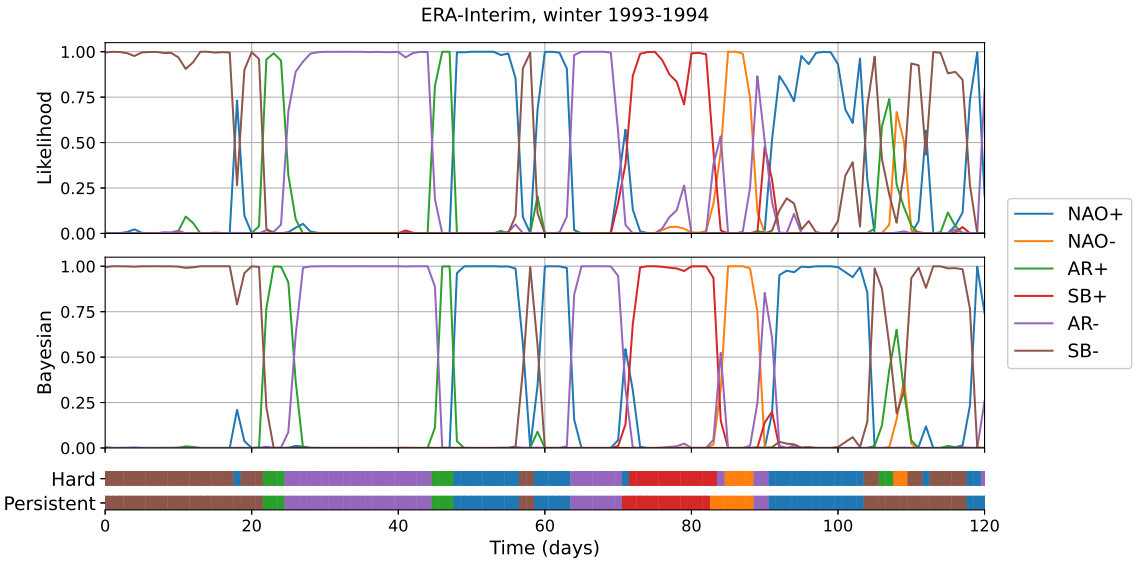


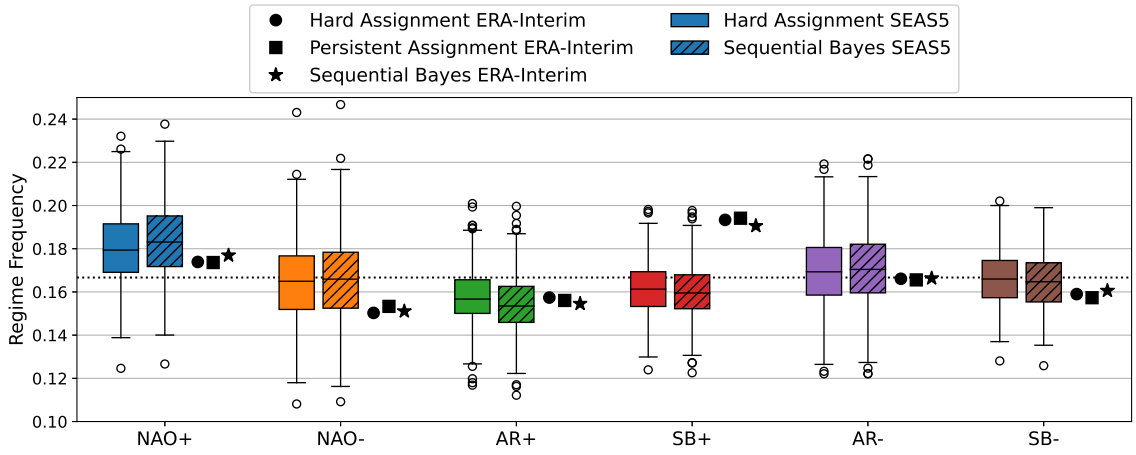
Figure 5.5: The observed regime likelihood and Bayesian regime probability for ERA-Interim, with the hard assignment using a standard or time-regularised k-means algorithm shown by the bars for the winter of 1993-1994.

5.6. The average frequencies of the regimes do not change when using the Bayesian regime assignment, as can be seen in Figure 5.6a. This holds both for the SEAS5 hindcast ensemble data and for ERA-Interim, where also the results of the time-regularised k-means clustering algorithm are shown for comparison. On the other hand the autocorrelation, being an indication of the persistence of the regimes, is strongly affected (Figure 5.6b). For ERA-Interim we see that the sequential Bayesian approach increases the autocorrelation even beyond that obtained using a regularised clustering algorithm that contains a persistence constraint. Also for SEAS5 a strong increase in autocorrelation is found using the sequential Bayesian regime assignment compared to a standard categorical assignment. For most regimes the ERA-Interim values lie at the top of the SEAS5 autocorrelation range, both for the standard and Bayesian approach. We find that the Bayesian approach does not alter the regime frequencies, but does lead to more persistent regime dynamics, as we might hope.

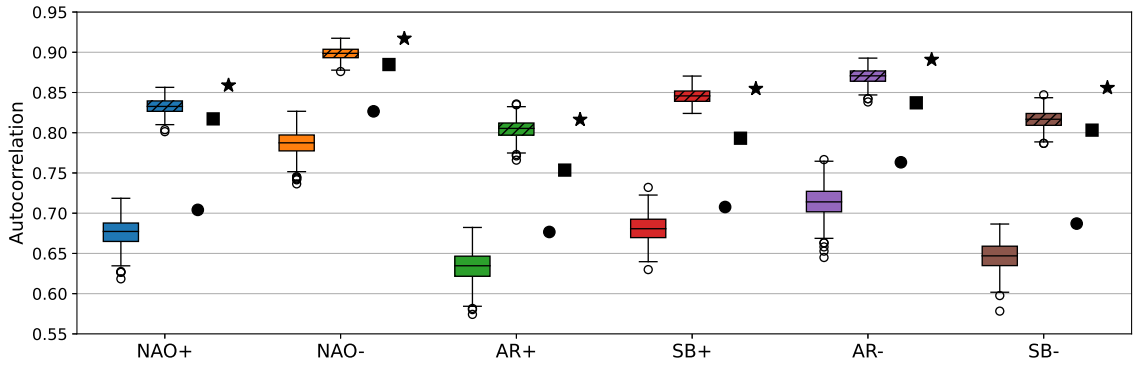
5.4 Ensemble Bayesian Regime Assignment

The implicit assumption made in the sequential Bayesian approach as discussed in the previous section is that the regime dynamics is statistically stationary in time. That is, the climatological likelihood functions and transition probabilities do not change in time. This is a reasonable and

minimal first assumption yielding good results, but it is likely that external factors such as ENSO affect some aspects of the regime dynamics, as discussed in Section 5.1 and seen in Section 4.4.3. These external factors act on timescales of several weeks to years and are unlikely to be picked up by the sequential approach. Nevertheless, the sequential regime assignment does impact the statistics on shorter timescales and could be relevant for sub-seasonal to seasonal (S2S) applications. There are two obvious ways in which to include the effect of external forcing in the Bayesian approach. The first is to update the regime likelihood functions in time. The second is to update the prior



(a)



(b)

Figure 5.6: The regime frequencies and 1-day autocorrelation as obtained using either standard k-means clustering or a sequential Bayesian regime assignment for the SEAS5 hindcast ensemble (boxes) and ERA-Interim (circles and stars), for which also the values obtained with the time-regularised k-means clustering method are shown (squares). Error bounds are determined using bootstrapping, where the boxes indicate the interquartile range with the whiskers extending 1.5 times on top of that and the circles being outlier points.

probabilities.

In the following analysis we focus on the latter approach and not the first. The main reason for this is the lack of data availability. Even though the SEAS5 hindcast ensemble has 51 members for each year, this still is insufficient to allow for e.g. weekly updating of the likelihood functions. An option for which sufficient data is available would be to compute the likelihood function during e.g. strong El Niño years, and use those to change the likelihood functions each year. However, this relies on the hypothesis that the regions in phase space belonging to each of the regimes shift as a consequence of ENSO forcing, while it may simply be the case that some regions are visited more often than others. As there are only 36 years of data available it is impossible to test this hypothesis and thus we refrain from pursuing this approach further. On the other hand, there is sufficient data to update the prior probabilities in time. There are several ways in which this can be done. For example, one can use information on ENSO to shift the prior probabilities, or one can make use of the ensemble information by allowing the transition probabilities to change in time. We pursue the latter approach, as it makes use of the information within the SEAS5 ensemble and does not require any external information. This ensemble Bayesian approach is explained and evaluated in the next two sections, followed by an analysis of the resulting interannual variability in Section 5.5.

5.4.1 Updating the Transition Probabilities

To obtain more informative prior regime probabilities, we focus on the transition matrix T. The transition probabilities T_{ij} from regime i to j are updated following the ensemble behaviour. This allows not only for (fixed) persistence to inform the prior, but also non-stationary external factors such as ENSO. Although there is not sufficient data to robustly estimate the transition probabilities directly, they can be inferred from the occurrence rates. The main assumption we make when updating the transition matrix T in time is that the regime probabilities are approximately stationary with respect to the current best estimate of the transition matrix. That is, we look for a transition matrix $T(t)$ for which the regime probabilities averaged over the ensemble at time t, $\bar{P}(t)$, are approximately stationary:

$$T(t)\bar{P}(t) = \bar{P}(t) + \epsilon^t. \tag{5.7}$$

Here ϵ^t is a noise term. Note that the climatological transition probabilities P^c are (nearly) stationary with respect to the transition matrix T^c. The aim thus is to find a transition matrix $T(t)$ for which Equation (5.7) holds. In addition we have that a transition matrix is normalised, meaning its

columns each sum to unity:

$$\sum_{i=1}^{k} T_{ij}(t) = 1, \qquad \forall j \in 1, ..., k, \forall t. \tag{5.8}$$

This gives two equations which can be used to update $T(t)$ at each time-step t. The problem of finding the values of the transition matrix $T(t)$ is ill-posed as there are not sufficient constraints, i.e. there are many transition matrices $T(t)$ which have $\bar{P}(t)$ as their invariant distribution. This means some choices need to be made in order to obtain a unique solution. The approach we propose in the following paragraph is one that follows the regime dynamics closely and is least biased in the sense that the deviations from T^c are equally distributed over all six regimes. With this approach the updating of the transition matrix differs from common methods used to identify hidden Markov models (e.g. Rabiner, 1989; Ghahramani, 2001; Majda et al., 2006) and no assumptions on the form of the noise are made.

The regime dynamics is dominated by persistence, i.e. the probability of a regime to transition to itself corresponding to the diagonal elements of the transition matrix, as can be seen in Equation (5.5). Therefore we focus on these diagonal elements $T_{ii}(t)$ for updating the matrix $T(t)$ in time. Writing out Equation (5.7) elementwise while separating the diagonal and off-diagonal elements yields

$$T_{ii}(t)\bar{P}_i(t) + \sum_{j \neq i}^{k} T_{ij}(t)\bar{P}_j(t) = \bar{P}_i(t) + \epsilon_i^t, \qquad \forall i \in 1, ..., k. \tag{5.9}$$

There are several ways to progress from here. As the diagonal terms dominate, we assume the off-diagonal elements do not differ much from the climatological values, that is $T_{ij}(t) \approx T_{ij}^c$ for all $i \neq j$. This yields an approximate equation for the diagonal elements of $T(t)$:

$$T_{ii}(t)\bar{P}_i(t) \approx \bar{P}_i(t) - \sum_{j \neq i}^{k} T_{ij}^c \bar{P}_j(t). \tag{5.10}$$

When a particular regime is less populated than it is in climatology, the other regimes will conversely be more populated, implying a larger negative term on the right-hand side of (5.10) and thus a smaller value of the self-transition probability, which makes physical sense. Note that this approximation breaks down when $\bar{P}_i(t)$ is very small compared to the other $\bar{P}_j(t)$, in which case we set $T_{ii}(t) = 0$ to prevent negative values. Starting from the updated diagonal elements, the off-diagonal elements are computed using Equation (5.8) with an equal distribution of the perturbation from the climatological value over the off-diagonal terms, that is

$$T_{ij}(t) = T_{ij}^c - \frac{1}{k-1}(T_{jj}(t) - T_{jj}^c). \tag{5.11}$$

The above method is equivalent to considering $T(t)$ as the climatological transition matrix plus a perturbation, i.e. $T(t) = T^c + T'(t)$, and subsequently assuming that the perturbations to the

off-diagonal terms are small. To see this substitute the above perturbation into Equation (5.7) and neglect T'_{ij} for $i \neq j$. An alternative way of looking at this is by considering it as a Markov regression model (Hamilton, 1989; Krolzig, 1997). That is, we write the transition matrix T as

$$T(t) = T^c + \sum_m \alpha_m(t) T_m. \tag{5.12}$$

Here T_m are matrices that set the shape of the perturbations to the climatological transition matrix, where the sum over each of the columns is zero for every m, and $\alpha_m(t)$ gives the strength of that term at time t. For a choice of

$$T_m = \begin{pmatrix} 0 & \dots & -\frac{1}{k-1} & \dots & 0 \\ & & \vdots & & \\ \vdots & & 1 & & \vdots \\ & & \vdots & & \\ 0 & & -\frac{1}{k-1} & & 0 \end{pmatrix}, \tag{5.13}$$

where the m-th column is non-zero this is exactly equivalent to the approach mentioned before. Here the α_m can be computed using the same assumptions as discussed before (compare with the perturbation approach and Equation (5.11)). This shows that there are several ways of looking at the problem that yield the same outcome, increasing the confidence in this approach.

5.4.2 Effect on the Regime Dynamics

To get an idea of how this approach can inform the prior probabilities consider Figure 5.7, which shows both the sequential and ensemble Bayesian regime assignments for the 42nd ensemble member during the winter of 1992-93. This is the same winter for which the 23rd ensemble member is shown in Figure 5.4. As an example, consider the probability of AR−. Around days 5-10 the ensemble indicates this regime is less likely, as shown by a lower self-transition probability, lowering the prior probability of the regime. On the other hand, from day 25 onward AR− is more likely according to the ensemble, increasing its prior probability compared to the sequential approach. In most cases changes to the final Bayesian probabilities are small. The only exceptions occur when a regime is deemed very unlikely, i.e. does not occur in any of the other ensemble members, as happens twice for the SB+ regime between days 60 and 90. In these two cases a high observed likelihood for SB+ is reduced substantially in the Bayesian probabilities in favour of the second most-likely regime according to the likelihood, e.g. a 90% likelihood is reduced to a 35% Bayesian probability. Yet importantly, the Bayesian probability of this regime is still non-zero, so it can quickly respond to

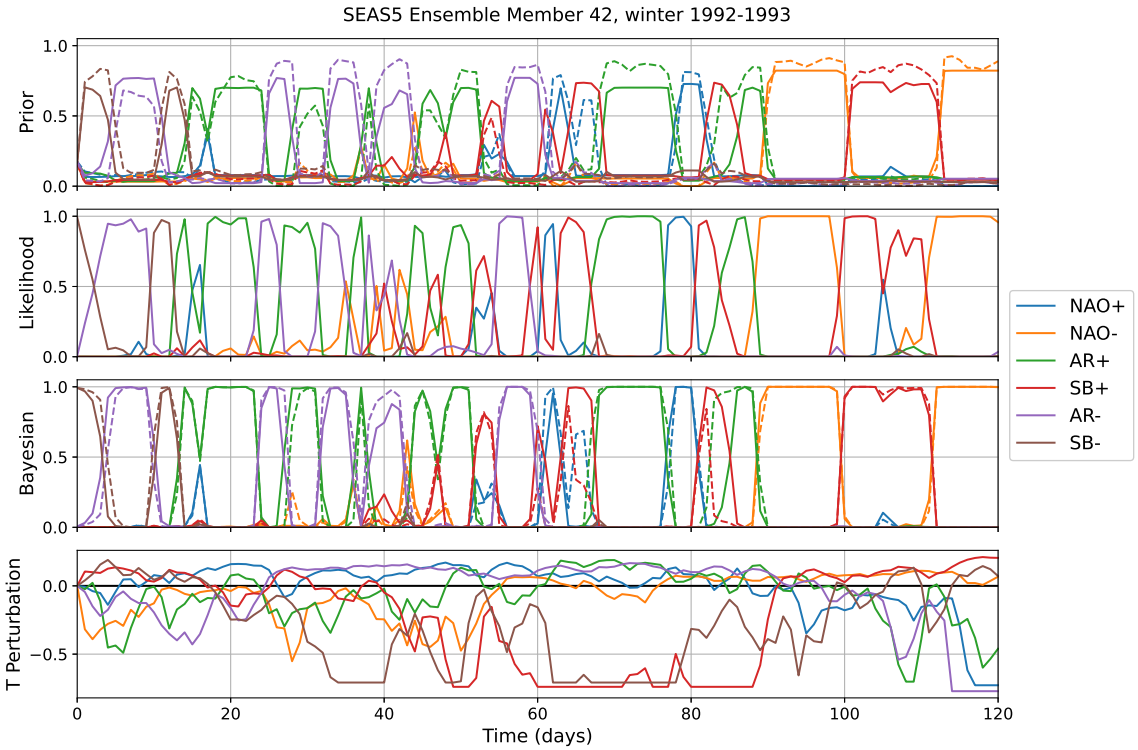


Figure 5.7: In the top three panels the prior probability, conditional regime likelihood and Bayesian regime probability for the 42nd ensemble member in the Bayesian regime assignment procedure for the winter of 1992-1993 are shown. The solid line shows the sequential Bayesian approach and the dashed line the ensemble approach discussed in this section. The bottom panel shows the difference between the updated self-transition probabilities in the ensemble approach and the climatological values.

new information. The overall regime frequencies and autocorrelation are not affected and remain as shown in Figure 5.6 for the sequential approach.

5.5 Interannual Regime Variability

The interannual variability as obtained using the ensemble Bayesian regime assignment is shown in Figure 5.8, with the result of the sequential Bayesian approach shown for reference (the interannual variability of the sequential Bayesian approach is nearly identical to that obtained using a k-means clustering assignment). The primary signal in the variability is found during very strong El Niño years (red vertical lines) with SB− and NAO− showing an increase in frequency, while AR+, AR− and NAO+ show a decrease in frequency. The signal during La Niña years is less clear, with an

increase in NAO+ frequency but no clear signal for the other regimes. These results reflect the well-known nonlinearity in the response to ENSO (Straus and Molteni, 2004; Toniazzo and Scaife, 2006) and are in line with the results obtained in Chapter 4 using a regularisation on the ensemble members, even though the regimes themselves are slightly different. The boxes on the right of each panel show the average regime frequencies during very strong El Niño and La Niña years for both the sequential and ensemble Bayesian approach. Some enhancement of the signal is found using the ensemble Bayesian regime assignment, which is most clear for the AR− and SB− regimes. The ERA-Interim variability obtained using the sequential Bayesian approach is shown as well.

To further consider the effect the updating of the transition matrix in the ensemble approach has on the interannual variability we look at Figure 5.9, which shows the difference between the sequential and ensemble Bayesian regime assignment as well as the yearly average change to the

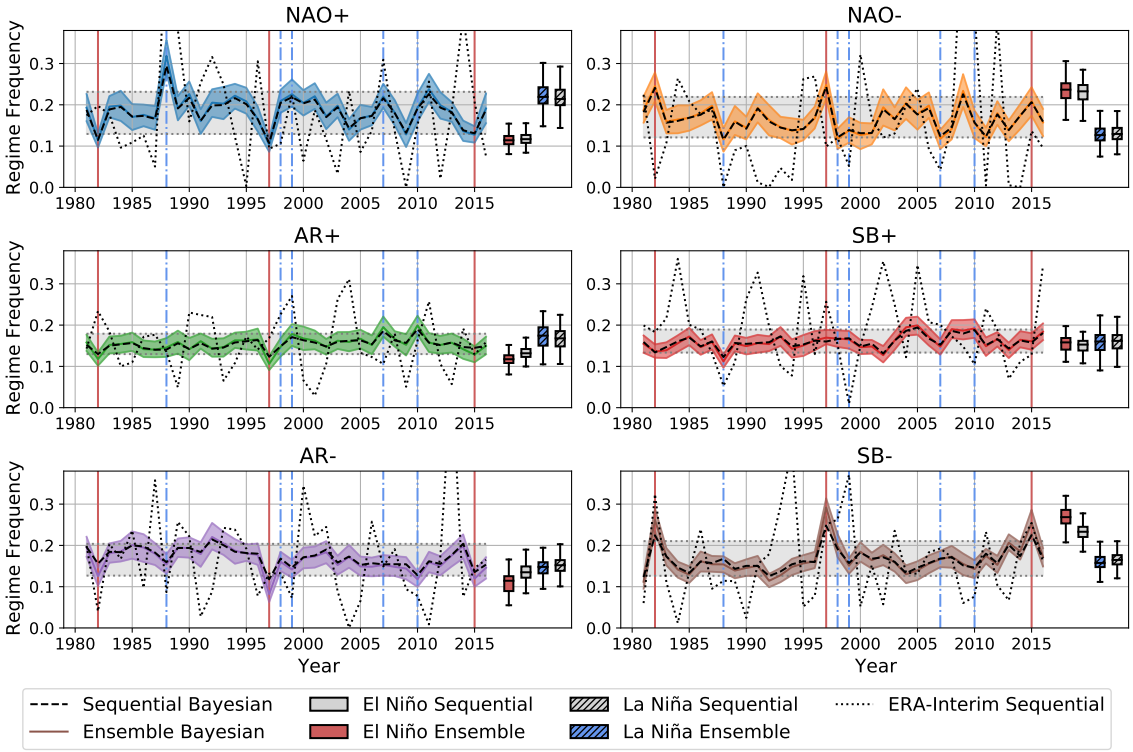


Figure 5.8: The interannual variability of the occurrence rates for the ensemble Bayesian regime assignment for SEAS5 (colour, with 95% confidence interval shaded), with the standard approach indicated by the black dashed lines. The grey shaded areas bounded by the grey dotted lines indicate the 10th and 90th percentile of the ensemble Bayesian assignment for each regime over all years. The black dotted line shows the ERA-Interim variability and the boxes on the right show the average occurrence rate during strong El Niño and La Niña years.

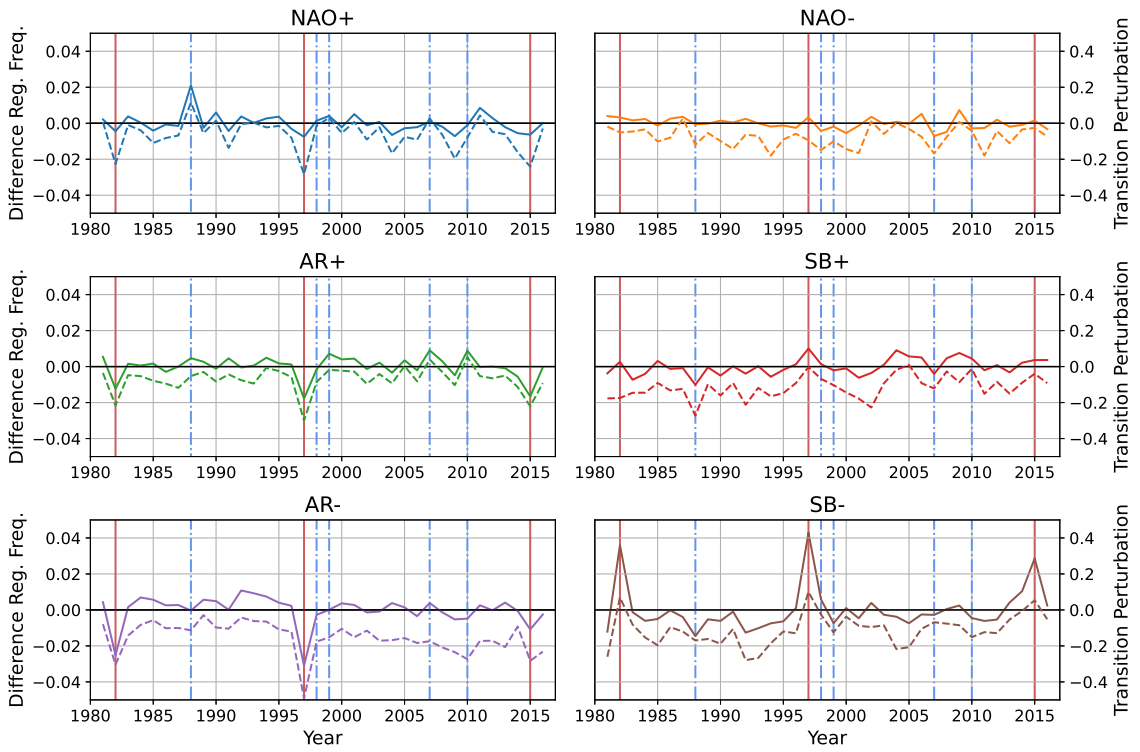


Figure 5.9: The difference in interannual variability of the occurrence rates between the standard and ensemble Bayesian regime probabilities (solid), as well as the change in the self-transition probability for the regimes following the ensemble (dashed).

self-transition probabilities, or persistence, of the regimes following the ensemble approach. Note that on average the perturbation to the self-transition probabilities is negative. The effect of the ensemble Bayesian approach on the regime frequencies is clearly visible for AR+, AR− and SB−, where the signal in response to El Niño is enhanced. For NAO+ a strong increase in regime frequency is found for the 1988-1989 La Niña, together with a weak change during El Niño years. NAO− and SB+ do not show much difference in interannual variability between the two methods, although in the latter case there is little signal to enhance. The changes in the self-transition probabilities in general match those found in the regime frequencies, as expected. One aspect to note is that for NAO+ the changes in the self-transition probability are relatively larger than those in the regime frequencies, especially when comparing to SB−.

The response of the changes in regime frequency to El Niño events found using the ensemble Bayesian approach appears to show a true signal and is very unlikely to have arisen by chance. To understand this, consider the change in regime frequency for SB−. The marginal probability of a very strong El Niño event is 3/36 (3 events in 36 years), so the chance of the first increase

in SB− frequency aligning with El Niño is 3/36. Then, given the first El Niño event has already happened, the probability of the second spike aligning is 2/35 and for the third 1/34. This gives a p-value of $3/36 \cdot 2/35 \cdot 1/34 \approx 10^{-4}$ for the alignment occurring by chance. The alignment of the increase/decrease in frequency for the other regimes only further decreases the probability of this being by chance. Also note that the response of both AR+ and AR− is a decrease in regime frequency during El Niño years, indicating another aspect of nonlinearity in the circulation response to ENSO.

Some of these signals in response to ENSO can already be picked up using 10-member ensembles. In Figure 5.10 the interannual variability of the regime frequency is shown for 50 random 10-member ensembles obtained from the full SEAS5 ensemble. For the full ensemble the strongest signal is found for SB− during very strong El Niño years, and this is the signal that jumps out most strongly again. To quantify this the Probability of Detection (POD) and False Alarm Ratio (FAR) for the 10-member ensembles are considered for peaks or troughs in regime frequency aligning with El Niño (Figure

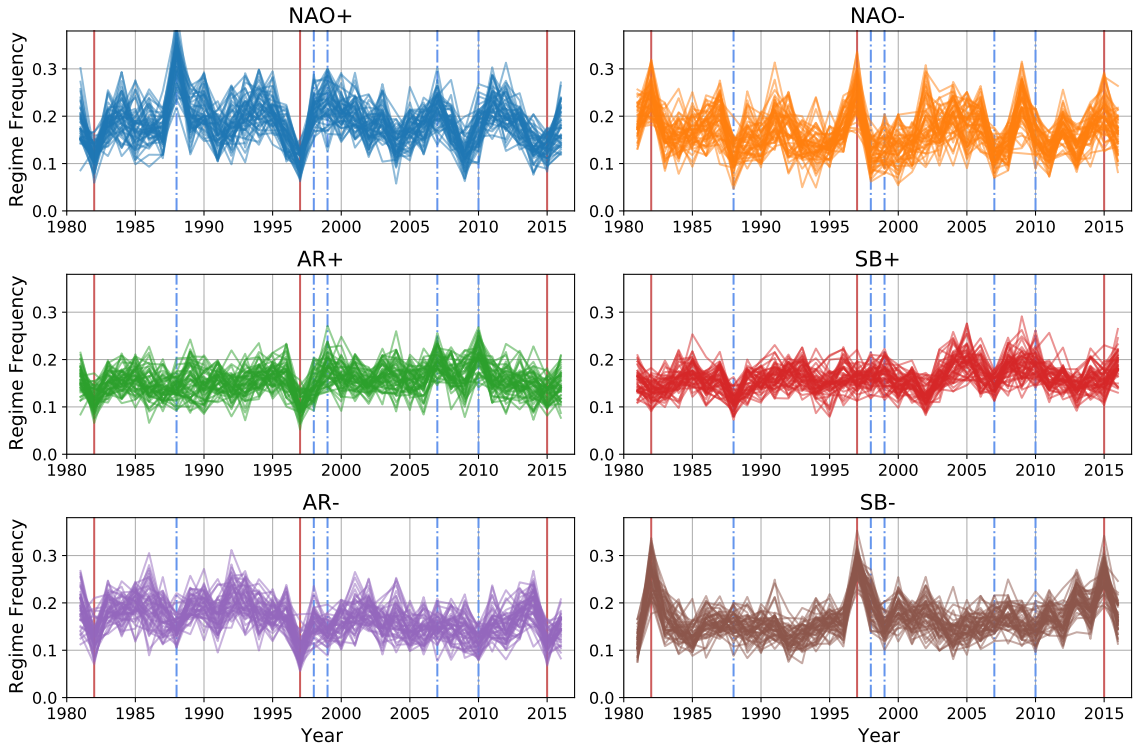
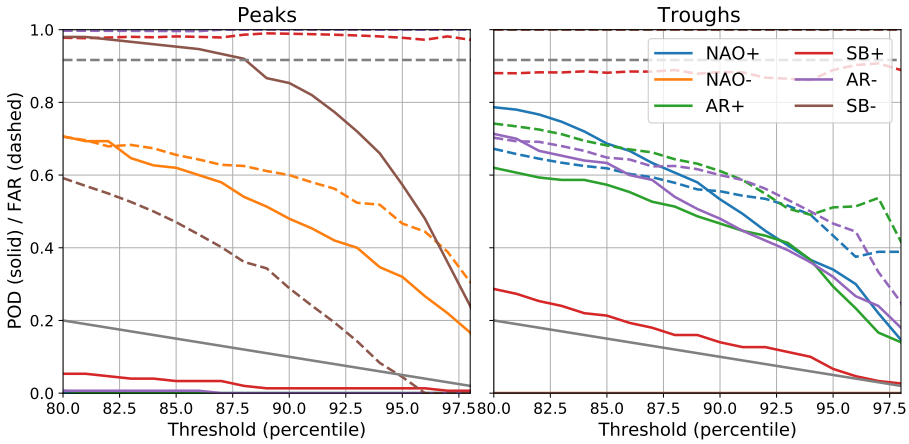


Figure 5.10: The interannual variability of the regime frequency for the ensemble Bayesian approach when applied to (random) ensembles of 10 members. In total 50 random ensembles are shown. The solid red and dash-dotted blue lines indicate very strong El Niño and strong La Niña years respectively.

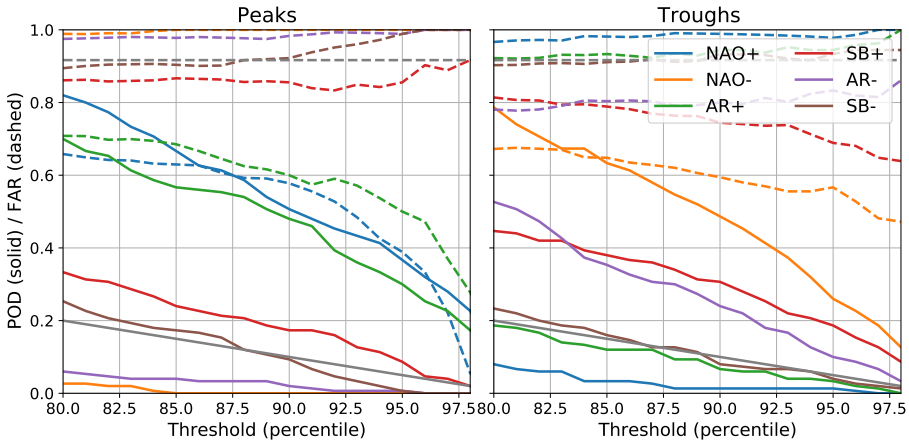
5.11a). Here, peaks and troughs are considered as exceedances with respect to the nth percentile. The POD is computed as the number of peaks/troughs aligning with El Niño years over the total number of El Niño years, and the FAR is computed as the number of peaks/troughs outside those El Niño years divided by the total number of peaks/troughs. As expected, there is a high POD for peaks in the SB− regime frequency with a relatively low FAR. Also for NAO− (peaks), NAO+, AR+ and AR− (troughs) there is some signal, with the FAR being comparable to the POD. For La Niña years there is some signal for NAO+, AR+ (peaks) and NAO− (troughs), but it is not as strong as for SB− in El Niño years (Figure 5.11b). This is to be expected as it is unlikely we can identify strong signals using a smaller ensemble if they are not clear in the full ensemble. Nevertheless, the relatively high PODs for these three regimes are encouraging.

To see whether the found response to ENSO for some regimes also reflects a predictable signal in the observations we regress the ERA-Interim interannual variability onto the SEAS5 one, as done in Chapter 4.4.4. The results for this, looking at the sequential and ensemble Bayesian approach, are shown in Table 5.1. In addition to the p-value, we also compute the Bayes factor which is the fraction of the probabilities of the data given two different hypotheses H_1 and H_2, i.e. $P(D|H_1)/P(D|H_2)$ (Kass and Raftery, 1995, and Section 1.6). Here the first hypothesis H_1 is that of a linear regression model, whereas the second hypothesis H_2 assumes a constant, climatological, regime frequency. Values of the Bayes factor above one indicate H_1 is more likely, with values between 3 and 20 constituting positive evidence and values over 20 yielding strong evidence towards it (Kass and Raftery, 1995).

Using the sequential Bayesian approach we already find some predictable signal for the NAO+ and SB− regimes, with Bayes factors of 7.6 and 5.1 respectively (Table 5.1). The Bayes factor for NAO− is also above 3, but here the p-value is larger reducing the confidence in this being a true signal. These results are comparable with those found in Chapter 4.4.4, with the regression coefficients being close to one for NAO+, NAO− and SB−. These regression coefficients around one indicate the signal in SEAS5 is of similar magnitude to that in ERA-Interim, showing no evidence of a signal-to-noise problem for the regime frequencies, in contrast to the NAO-index (as discussed in Section 4.4.5). Using the ensemble information to update the transition probabilities increases the predictable signal for NAO+ and SB−, with smaller p-values and higher Bayes factors. Also the AR− signal is enhanced with a Bayes factor over 3, although the p-value is still relatively large. The enhancement of the NAO+ signal is comparable to that found using an ensemble-regularised clustering approach, whereas the change for SB− is weaker (a Bayes factor of 13.2 compared to 5.5, see Section 4.4.4 for comparison). On the other hand, the decrease in Bayes factors for NAO− and



(a) El Niño years.



(b) La Niña years.

Figure 5.11: The probability of detection (solid) and false alarm ratio (dashed) for a peak or trough in regime frequency in 10-member subsamples of the SEAS5 ensembles occurring in the same year as a very strong El Niño or La Niña, as a function of the percentile used for the definition of the peaks and troughs. The colored lines indicate the regime values, and the grey lines the values for peaks and troughs occurring in random years, i.e. no signal.

AR− using a regularised approach is not found using the ensemble Bayesian method, which shows small increases of the Bayes factors. In Chapter 4.4.4 a significant signal was found using multiple linear regression (MLR) of ERA-Interim NAO− onto the SEAS5 NAO+ and SB−. Here we find this signal as well with a Bayes factor of 21.1 for the sequential method increasing to 26.6 using the ensemble approach. Comparing the two methods, we find that the ensemble Bayesian regime assignment allows to identify more pronounced interannual variability signals for some regimes, while still accounting for the signal of the other regimes.

Regime	NAO+	NAO−	AR+	SB+	AR−	SB−	MLR	NAO−
Sequential Bayes								
Regression Coefficient	1.170	1.094	-0.504	0.258	1.207	1.083	NAO+	-1.369
							SB−	-1.838
p-value	0.052	0.139	0.592	0.795	0.174	0.082		0.047
Bayes Factor	7.579	3.251	1.167	1.037	2.696	5.054		21.108
Ensemble Bayes								
Regression Coefficient	1.066	1.035	-0.435	0.225	1.037	0.785	NAO+	-1.429
							SB−	-1.412
p-value	0.044	0.133	0.527	0.782	0.136	0.075		0.041
Bayes Factor	8.910	3.365	1.240	1.042	3.306	5.487		26.641

Table 5.1: The regression coefficient, p-value and Bayes factor for linear regression of the interannual variability in regime frequency (ERA-Interim onto SEAS5) for all six regimes. In addition, the result of MLR of the ERA-Interim NAO− frequency against the SEAS5 ensemble mean NAO+ and SB− regime frequencies is shown. Values for both the sequential as well as the ensemble approach are shown.

5.6 Summary and Discussion

A new approach exploiting Bayes Theorem (5.1) is proposed to obtain a probabilistic regime assignment. The approach combines climatological likelihood functions with prior information from the previous day, using the known regime persistence from prior studies, to obtain a Bayesian regime probability. This sequential Bayesian regime assignment allows for smoother transitions between the regimes and indicates when data does not clearly belong to one regime. In contrast to the methods discussed in Chapters 3 and 4 that use a regularised k-means clustering algorithm, there is no parameter, other than the number of regimes k, that has to be selected. Applying the approach to six wintertime circulation regimes over the Euro-Atlantic sector yields an increase in persistence, without affecting the average regime frequencies for both SEAS5 and ERA-Interim (Figure 5.6). In addition, for ERA-Interim the 1-day autocorrelation was found to be higher than that obtained using a time-regularised k-means clustering approach containing a persistence constraint (as introduced in Section 3.3).

A yet more informative prior for the Bayesian approach can be obtained by continuously updating the prior probabilities by taking information from the full SEAS5 ensemble into account. Starting

from the assumption of approximate stationarity of the ensemble mean regime frequencies, the regime transition matrix is updated. This in turn affects the prior probabilities, leading to more pronounced interannual variability for some regimes. When considering the interannual variability, the response to three very strong El Niño events in recent decades clearly stands out (Figure 5.8). During these three winters SB− and NAO− increase in frequency, while NAO+, AR+ and AR− decrease. The signals for AR+, AR− and SB− are enhanced by the ensemble Bayesian approach compared to the sequential method. The signal during La Niña winters is less pronounced, with the increase in NAO+ frequency during 1988-89 standing out most clearly.

This response to ENSO in the SEAS5 ensemble can already be identified using only a 10-member ensemble. The increase in SB− occurrence during El Niño years is a particularly strong signal and is found in nearly all 10-member ensembles considered (Figure 5.10). Also for NAO+, NAO−, AR+ and AR− significant probabilities of detection for peaks or troughs coinciding with El Niño events are found. However, here there also is a substantial false alarm ratio indicating that many peaks or troughs in the ensemble occur in non-El Niño years. These results suggest that one may not need a very large ensemble to identify regime signals in response to ENSO.

We also use a linear regression analysis to identify predictable signals in the observations on interannual timescales. Here, as in Chapter 4.4.4, NAO+ and SB− were found to be predictable from the SEAS5 ensemble with regression coefficients around one (Table 5.1), suggesting no signal-to-noise deficit for these regimes. The ensemble approach leads to an increase in Bayes factor compared to the sequential method for all regimes, with the largest improvement for NAO+. The signal for NAO+ is comparable to that obtained using the ensemble-regularised clustering approach discussed in Section 4.4.4, whereas the signal for SB− is slightly weaker. This difference likely is related to the small differences between the regime patterns used for the two methods. On the other hand, none of the signals for the other regimes are reduced, which is the case for the regularised clustering approach.

Chapter 6

Discussion

A thorough understanding of regional circulation dynamics is essential to increase our knowledge of the effects of climate change on different parts of the world. Non-stationary factors within the climate system, varying on (sub-)seasonal to interannual timescales, are key to this understanding, as they are driving aspects of the atmospheric circulation around the world via teleconnection patterns. When embedding these remote drivers in a storyline approach to better understand the circulation response, it helps to discretise the atmospheric circulation into a discrete number of states: regimes. This thesis addresses the very first step towards developing such a storyline approach, discussing the identification of robust circulation regime signals using different approaches. The focus here is on the wintertime Euro-Atlantic sector, where circulation regimes have been well established and widely studied (e.g. Michelangeli et al., 1995; Kageyama et al., 1999; Yiou and Nogaj, 2004).

6.1 Summary

In Chapter 3 the standard approach to identifying atmospheric circulation regimes is revisited. Most studies apply both Principal Component (PC) analysis (a spatial filter), using only the first couple of PCs, and a low-pass time filter before applying a clustering algorithm. The use of PCs is found to yield a different optimal number of regimes compared to the use of gridpoint data. Information criteria indicate four regimes (two phases of the North Atlantic Oscillation (NAO), Atlantic Ridge (AR), Scandinavian Blocking (SB)) as being optimal for PC data, in line with literature (e.g Michelangeli et al., 1995; Yiou and Nogaj, 2004). On the other hand six regimes is found to be optimal for gridpoint data, adding the opposite of the Atlantic Ridge and Scandinavian

Blocking to the standard four regimes. Furthermore, the use of a low-pass filter is found to introduce a bias in the regime frequencies, whereas the use of a time-regularised clustering algorithm including a persistence constraint does not lead to such a bias. Both these results show that it is important to be careful with filtering the data, whether through PC analysis or a low-pass filter, as it can remove some of the important nuances of the signal. Changing the clustering method itself, instead of the data, allows for identifying equally persistent regime signals without introducing a bias. These results are in line with the mathematical understanding that it is better to treat the signal and noise at the same time, rather than attempting to first remove the noise.

In Chapter 4 we turn to the non-stationary regime signal, using the six circulation regimes identified in Chapter 3. A novel ensemble-regularised clustering approach is introduced, enforcing a level of similarity between different members of a model ensemble, to identify a more pronounced non-stationary regime variability signal. The idea behind this regularisation is that the majority of ensemble members respond to e.g. an El Niño event, with some regimes being preferred over others. Noise may lead to some members not showing as strong a response to such forcing and the constraint on the ensemble similarity corrects this. It is found that there is no robust sub-seasonal regime variability in the climatology, other than that introduced by changes in the background circulation. On interannual timescales a clear response to the El Niño Southern Oscillation (ENSO) is found, with an increase in SB− and NAO− frequency and decrease of NAO+ during very strong El Niño events. Predictable signals are found for both NAO+ and SB−, with an NAO− signal only emerging by using multiple linear regression with NAO+ and SB− as predictors. In contrast to an NAO-index, no signal-to-noise problem is found for the regime occurrence rates, with regression coefficients around one.

To further study the persistent and non-stationary regime dynamics another method to identify the regime signal is discussed in Chapter 5. This Bayesian approach to regime assignment does not require any parameter selection, as is required for the regularised clustering methods in Chapters 3 and 4, but instead uses prior information to obtain a better informed regime probability. This probabilistic approach also captures the subtleties of the regime dynamics better, as it is not enforcing a hard, categorical assignment of data to a regime. A sequential Bayesian regime assignment method leads to more persistent dynamics, while an ensemble Bayesian approach in addition enhances the non-stationary regime signal. The identified interannual regime variability is comparable to that obtained using the ensemble-regularised method, with predictable signals for both NAO+ and SB−. Small differences are found when considering the response to ENSO, with the Bayesian method also identifying an ENSO signal for AR+, in addition to the NAO+, NAO−, AR− and SB− response

signals that are obtained using the ensemble-regularised approach. Some of the regime variability signals in response to ENSO can already be identified in a 10-member ensemble, showing the robustness of the signal within the model.

6.2 Outlook

6.2.1 Regime Variability Signals

This thesis only adds one small part to the extensive literature on atmospheric circulation regimes and many questions remain. The interannual regime variability signal is affected by many factors, where ENSO is the driver that can be identified most clearly regardless of the method used and has been discussed in most detail in this thesis. In response to very strong El Niño events SB− and NAO− become more frequent, while NAO+, AR+ and AR− decrease in occurrence rate. The response to La Niña events is less strong, exhibiting a larger spread, with increased occurrence rates of NAO+ and decreased frequencies of AR−. In this respect, it would be interesting to look at the response for early and late winter separately, as those responses might differ (e.g. Moron and Plaut, 2003; Ayarzaguena et al., 2018). One can imagine that such a difference in response could be related to the Stratospheric Polar Vortex (SPV) dynamics and the occurrence of Sudden Stratospheric Warmings (SSWs). As the SPV is known to affect the North Atlantic atmospheric circulation (e.g. Charlton-Perez et al., 2018; Domeisen et al., 2020), it would be interesting to investigate the relation between the SPV, ENSO and the regime response to them. Also the response of the regime dynamics to other remote drivers, such as the Madden-Julian Oscillation (MJO) (e.g. Cassou, 2008; Straus et al., 2015; Lee et al., 2019, 2020), would be valuable to consider in future studies. Here, especially the probabilistic regime assignment obtained using the Bayesian approach can be valuable in understanding the links between the regime dynamics and remote drivers.

The use of six instead of four regimes can aid in better understanding the nuances of the regime response to external drivers and maybe identify some aspects where models lack skill. When considering only four regimes, which is standard in many studies, most data now assigned to either NAO+ or SB− would be allocated to the NAO+ regime. The separation between NAO+ and SB− in this work could help better understand the relation between ENSO, the MJO and the North Atlantic circulation. Previous studies found a more frequent NAO+ after MJO phases 1-5 in El Niño years using four regimes (Lee et al., 2019), whereas the overall signal is a decrease of the NAO-index during El Niño (Figure 4.10). El Niño was found to enhance the teleconnections between the MJO

and the regimes, which is also apparent in the increase of the NAO− frequency after MJO phases 7 and 8. On the other hand La Niña on average inhibits such teleconnections between the MJO and the four regimes studied. It would be interesting to examine the teleconnection signal for the additional regimes (AR− and SB−) discussed here, to see whether these relations can be captured in more detail.

On interannual timescales predictable signals for NAO+ and SB− are found for both the regularised and Bayesian approach using a regression method. It is interesting that no clear signals are found for the other regimes, especially for NAO− as this is part of the NAO teleconnection pattern. As ENSO is important to obtain this predictable signal, the weaker NAO− response compared to that of NAO+ and SB− can probably be linked to this lack of signal. It indicates the model might not represent this NAO− regime accurately, which could be related to it resembling a blocking pattern over Greenland, as models are known to have difficulty representing blocking patterns (e.g. Tibaldi and Molteni, 1990; Pelly and Hoskins, 2003). A potential reason for the lack of a predictable signal in NAO− could lie in the role of SSWs, which are known to induce negative NAO states (Baldwin and Dunkerton, 2001; Hitchcock and Simpson, 2014; Domeisen et al., 2020). Portal et al. (2022) have shown that seasonal forecast models, including the SEAS5 model studied here, tend to overpredict the SSW response to ENSO. Consistent with that, there is a strong negative NAO response to ENSO, which is not seen in observations. However, if SSWs are playing an important role in seasonal predictability, then this role will be difficult to assess from the limited sample size provided by the reanalysis record.

The lack of predictable signal for NAO− likely ties in with the signal-to-noise problem found for an NAO-index, discussed in Chapter 4. NAO− shows a weaker response to ENSO than NAO+. Linked to this weak NAO− response, Hardiman et al. (2022) found a weak teleconnection between ENSO and the Arctic Oscillation (AO), which is highly correlated with the NAO, to be linked to a lack of eddy feedback within models. The eddy feedback within a model in turn was found to be correlated to the Ratio of Predictable Components (RPC), which gives the ratio of the predictable component in observations over that in the model, with lower RPC values for stronger eddy feedback. One pathway between ENSO and the North Atlantic atmospheric circulation runs via the stratosphere (e.g. Butler et al., 2014), which links this to the previous discussion on the relation with SSWs. It indicates that the low signal-to-noise ratios for an NAO-index and lack of NAO− predictability both could (in part) be caused by weak eddy feedback in the model. It would be interesting to quantify this effect further using the regime approach.

The regime variability signals obtained can feed into the development of storylines of regional

climate change. One can imagine that the regime occurrence rates change depending on the level of warming. These changes in the regime frequencies describe the change in the atmospheric circulation and in turn can inform the effect on e.g. the surface temperature. For example, suppose there is one regime that is known to increase the probability of cold spells. When in a model increasing temperatures lead to this regime occurring less often, we can infer from this that cold spells are becoming less likely. This type of reasoning can also be applied to possible changes in the regime response to e.g. ENSO under climate change and the effect that would have on regional weather. In this study we have not dealt with the surface impact of the different circulation regimes, which is a crucial factor when developing storylines. Thus assessing the effect of the six circulation regimes on surface variables would be valuable next step for the identified regime signals to aid in better understanding the regional effects of climate change.

6.2.2 Regime Identification Methods

A significant part of the work described in this thesis has been devoted to discussing different methods for regime signal identification. In Chapters 3 and 4 two different regularisation approaches were applied to a standard k-means clustering algorithm to focus on either the persistent or non-stationary regime dynamics. One can imagine other types of regularisation, enforcing for example a level of similarity in the regime assignment of ensemble members over a longer period of time. The regularisation can be tailored to the question of interest and used as an alternative to filtering the data before applying a clustering method. Here one has to be careful not to enforce unphysical regime dynamics. Information criteria can aid in selecting suitable constraint values for the regularisation, preventing overfitting which could indicate such unphysical results.

The Bayesian regime assignment approach discussed in Chapter 5 does not require such a selection of a constraint value. The inclusion of prior information allows to identify both more persistent and non-stationary regime dynamics. The two methods of identifying the prior regime probabilities, using the regime transition matrix and assuming approximate stationarity in time to update this transition matrix, are by no means the only ways in which the prior can be obtained. For example, one could link the computation of the prior regime probabilities to external drivers such as ENSO or SSWs to obtain more informed posterior regime probabilities. Such better informed prior regime probabilities could be relevant for predictions on sub-seasonal to seasonal (S2S) timescales. The clear improvement in persistence obtained from the sequential Bayesian method and the probabilistic nature of the regime assignment might be useful in this regard, even if the seasonal averages are

not much affected. Priors that are informed by a persistent transition matrix and external climatic processes could be used for both model ensembles as well as reanalysis datasets, and aid in better distinguishing the signal from the noise in these S2S predictions.

Another aspect of the Bayesian regime assignment that one can consider to obtain a better informed posterior regime probability is the climatological likelihood function. In Chapter 5 this is assumed to be constant, but one can imagine situations where remote drivers impact these functions. For example, it could be the case that there is a difference in the regime likelihood functions between El Niño and La Niña years. The difficulty here is how to assess such a hypothesis, as the reanalysis time series is not of sufficient length to obtain a reasonable sample size. It could be an option when using ensemble data, but testing its value remains a challenge due to the lack of reanalysis results. Another direction one could imagine is to have seasonally varying likelihood functions, capturing the variability in the background climatology.

The use of the Bayesian regime assignment approach is not limited to atmospheric circulation regimes, but can be applied to any situation in which the data can be separated into two or more regimes. For example, one can think of the two phases of the NAO or the jet latitude (Woollings et al., 2010), but one can also imagine applications in topics outside of climate such as finance and healthcare. For the application one needs some information on the regime likelihood function and a way to obtain an informative prior. In most cases the latter will be the most challenging and requires a thorough understanding of the processes involved. For circulation regimes a prior based on climatological transition probabilities, which automatically builds in persistence, was shown to be a suitable and natural choice, and incorporating information from a full ensemble enhanced the interannual signal. Depending on the regime process considered other choices for the prior may be more suitable.

There are many directions that can be taken to advance the presented methods for regime identification and assignment in this thesis. One can apply the methods to other clustering problems. One can adapt the regularisation within the k-means clustering algorithm focused on different research questions. One can use external drivers to inform the prior in the Bayesian approach to regime assignment. One can dive into the results to study the relation between regimes and remote drivers. One can research these and many more aspects building on the methods and results of this thesis. Studying any of these aspects will aid in better understanding the atmospheric circulation signal, have valuable applications in e.g. developing storylines of regional climate change, and hopefully lead to model improvement for even better predictability in the future.

Bibliography

Acevedo, W., de Wiljes, J. and Reich, S. (2017), 'Second-order accurate ensemble transform particle filters', *SIAM Journal on Scientific Computing* **39**(5), A1834–A1850.

Akaike, H. (1973), Information Theory as an Extension of the Maximum Likelihood Principle, *in* B. Petrov and F. Csaki, eds, 'Second International Symposium on Information Theory', Budapest, pp. 267–281.

Ambaum, M. H. (2010), 'Significance Tests in Climate Science', *Journal of Climate* **23**(22), 5927–5932.

Ambaum, M. H., Hoskins, B. J. and Stephenson, D. B. (2001), 'Arctic Oscillation or North Atlantic Oscillation?', *Journal of Climate* **14**(16), 3495–3507.

Amini, S. and Straus, D. M. (2018), 'Control of Storminess over the Pacific and North America by Circulation Regimes', *Climate Dynamics* **52**(7), 4749–4770.

Arnold, T. W. (2010), 'Uninformative Parameters and Model Selection Using Akaike's Information Criterion', *Journal of Wildlife Management* **74**(6), 1175–1178.

Ayarzaguena, B., Ineson, S., Dunstone, N. J., Baldwin, M. P. and Scaife, A. A. (2018), 'Intraseasonal Effects of El Niño–Southern Oscillation on North Atlantic Climate', *Journal of Climate* **31**(21), 8861–8873.

Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A. and Scaife, A. A. (2018), 'An Intercomparison of Skill and Overconfidence/Underconfidence of the Wintertime North Atlantic Oscillation in Multimodel Seasonal Forecasts', *Geophysical Research Letters* **45**(15), 7808–7817.

Baldi, P. and Hatfield, G. (2002), *DNA Microarrays and Gene Expression: from experiments to data analysis and modelling*, Cambridge University Press.

Baldwin, M. P. and Dunkerton, T. J. (2001), 'Stratospheric Harbingers of Anomalous Weather Regimes', *Science* **294**(5542), 581–584.

Ball, G. H. and Hall, D. J. (1965), ISODATA, a novel method of data analysis and pattern classification, Technical report, Stanford Research Institute, Stanford, CA.

Bao, M. and Wallace, J. M. (2015), 'Cluster Analysis of Northern Hemisphere Wintertime 500-hPa Flow Regimes during 1920–2014', *Journal of the Atmospheric Sciences* **72**(9), 3597–3608.

Barnes, E. A., Samarasinghe, S. M., Ebert-Uphoff, I. and Furtado, J. C. (2019), 'Tropospheric and Stratospheric Causal Pathways Between the MJO and NAO', *Journal of Geophysical Research: Atmospheres* **124**(16), 9356–9371.

Barnston, A. G. and Livezey, R. E. (1987), 'Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns', *Monthly Weather Review* **115**, 1083–1126.

Bauer, P., Thorpe, A. and Brunet, G. (2015), 'The quiet revolution of numerical weather prediction', *Nature* **525**, 47–55.

Beerli, R. and Grams, C. M. (2019), 'Stratospheric modulation of the large-scale circulation in the Atlantic–European region and its implications for surface weather events', *Quarterly Journal of the Royal Meteorological Society* **145**(725), 3732–3750.

Bell, C. J., Gray, L. J., Charlton-Perez, A. J., Joshi, M. M. and Scaife, A. A. (2009), 'Stratospheric communication of El Niño teleconnections to European winter', *Journal of Climate* **22**(15), 4083–4096.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.

Blackmon, M. L., Wallace, J. M., Lau, N.-C. and Mullen, S. L. (1977), 'An Observational Study of the Northern Hemisphere Wintertime Circulation', *Journal of the Atmospheric Sciences* **34**(7), 1040–1053.

Bloomfield, H. C., Brayshaw, D. J. and Charlton-Perez, A. J. (2020), 'Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types', *Meteorological Applications* **17**(1), e1858.

Boljka, L., Shepherd, T. G. and Blackburn, M. (2018), 'On the Coupling between Barotropic and Baroclinic Modes of Extratropical Atmospheric Variability', *Journal of Atmospheric Sciences* **75**(6), 1853–1871.

Brönnimann, S., Luterbacher, J., Staehelin, J., Svendby, T., Hansen, G. and Svenøe, T. (2004), 'Extreme climate of the global troposphere and stratosphere in 1940–42 related to El Niño', *Nature* **431**, 971–974.

Brönnimann, S., Xoplaki, E., Casty, C., Pauling, A. and Luterbacher, J. (2007), 'ENSO influence on Europe during the last centuries', *Climate Dynamics* **28**, 181–197.

Büeler, D., Ferranti, L., Magnusson, L., Quinting, J. F. and Grams, C. M. (2021), 'Year-round sub-seasonal forecast skill for Atlantic–European weather regimes', *Quarterly Journal of the Royal Meteorological Society* **147**(741), 4283–4309.

Burnham, K. P. and Anderson, D. R. (2004), 'Multimodel inference: Understanding AIC and BIC in model selection', *Sociological Methods and Research* **33**(2), 261–304.

Butler, A. H., Polvani, L. M. and Deser, C. (2014), 'Separating the stratospheric and tropospheric pathways of El Niño–Southern Oscillation teleconnections', *Environmental Research Letters* **9**(024014).

Casola, J. H. and Wallace, J. M. (2007), 'Identifying weather regimes in the wintertime 500-hPa geopotential height field for the Pacific-North American sector using a limited-contour clustering technique', *Journal of Applied Meteorology and Climatology* **46**(10), 1619–1630.

Cassou, C. (2008), 'Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation', *Nature* **455**(7212), 523–527.

Cassou, C., Laurent, T. and Phillips, A. S. (2005), 'Tropical Atlantic Influence on European Heat Waves', *Journal of Climate* **18**(15), 2805–2811.

Charlton-Perez, A. J., Bröcker, J., Stockdale, T. N. and Johnson, S. (2019), 'When and where do ECMWF seasonal forecast systems exhibit anomalously low signal-to-noise ratio?', *Quarterly Journal of the Royal Meteorological Society* **145**(725), 3466–3478.

Charlton-Perez, A. J., Ferranti, L. and Lee, R. W. (2018), 'The influence of the stratospheric state on North Atlantic weather regimes', *Quarterly Journal of the Royal Meteorological Society* **144**(713), 1140–1151.

Charney, J. G. and DeVore, J. G. (1979), 'Multiple Flow Equilibria in the Atmosphere and Blocking', *Journal of Atmospheric Sciences* **36**, 1205–1216.

Chen, S. S. and Gopalakrishnan, P. S. (1998), 'Clustering via the Bayesian information criterion with applications in speech recognition', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2**, 645–648.

Cheng, X. and Wallace, J. M. (1993), 'Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns', *Journal of Atmospheric Sciences* **50**(16), 2674–2696.

Christiansen, B. (2007), 'Atmospheric circulation regimes: Can cluster analysis provide the number?', *Journal of Climate* **20**(10), 2229–2250.

Chung, C. and Nigam, S. (1999), 'Weighting of geophysical data in Principal Component Analysis', *Journal of Geophysical Research* **104**(D14), 16925–16928.

Cobos, C., Muñoz-Collazos, H., Urbano-Muñoz, R., Mendoza, M., León, E. and Herrera-Viedma, E. (2014), 'Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion', *Information Sciences* **281**, 248–264.

Cortesi, N., Torralba, V., Lledó, L., Manrique, A., Nube, S. and Reviriego, G. (2021), 'Yearly evolution of Euro-Atlantic weather regimes and of their sub-seasonal predictability', *Climate Dynamics* **56**, 3933–3964.

Corti, S., Molteni, F. and Palmer, T. N. (1999), 'Signature of recent climate change in frequencies of natural atmospheric circulation regimes', *Nature* **398**(6730), 799–802.

Cottle, R. W. and Thapa, M. N. (2017), *Linear and Nonlinear Optimization*, Springer.

Dawson, A. and Palmer, T. N. (2015), 'Simulating weather regimes: impact of model resolution and stochastic parameterization', *Climate Dynamics* **44**, 2177–2193.

Dawson, A., Palmer, T. N. and Corti, S. (2012), 'Simulating regime structures in weather and climate prediction models', *Geophysical Research Letters* **39**(21), L21805.

de Wiljes, J., Majda, A. and Horenko, I. (2013), 'An adaptive markov chain monte carlo approach to time series clustering of processes with regime transition behavior', *Multiscale Modeling & Simulation* **11**(2), 415–441.

de Wiljes, J., Putzig, L. and Horenko, I. (2014), 'Discrete nonhomogeneous and nonstationary logistic and markov regression models for spatiotemporal data with unresolved external influences', *Communications in Applied Mathematics and Computational Science* **9**(1).

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N. and Vitart, F. (2011), 'The ERA-Interim reanalysis: Configuration and performance of the data assimilation system', *Quarterly Journal of the Royal Meteorological Society* **137**(656), 553–597.

Del Moral, P. (1997), 'Nonlinear filtering: Interacting particle resolution', *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* **325**(6), 653–658.

Deutscher Wetterdienst (2019), 'Grosswetterlagen Forecast (GWL)'.
URL: *https://www.dwd.de/EN/research/weatherforecasting/met_applications/nwp_applications/grosswetterlagen_forecast.html*

Doblas-Reyes, F., Sörensson, A., Almazroui, M., Dosio, A., Gutowski, W., Haarsma, R., Hamdi, R., Hewitson, B., Kwon, W.-T., Lamptey, B., Maraun, D., Stephenson, T., Takayabu, I., Terray, L., Turner, A. and Zuo, Z. (2021), Linking Global to Regional Climate Change, *in* 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, eds, 'Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1363–1512.

Dole, R. M. and Gorden, N. D. (1983), 'Persistent Anomalies of the Extratropical Northern Hemisphere Wintertime Circulation: Geographical Distribution and Regional Persistence Characteristics', *Monthly Weather Review* **111**(8), 1567–1586.

Domeisen, D. I. V., Garfinkel, C. I. and Butler, A. H. (2019), 'The Teleconnection of El Niño Southern Oscillation to the Stratosphere', *Reviews of Geophysics* **57**, 5–47.

Domeisen, D. I. V., Grams, C. M. and Papritz, L. (2020), 'The role of North Atlantic-European weather regimes in the surface impact of sudden stratospheric warming events', *Weather and Climate Dynamics* **1**, 373–388.

Doob, J. L. (1953), *Stochastic processes*, Wiley New York.

Dorrington, J. and Strommen, K. (2020), 'Jet Speed Variability Obscures Euro-Atlantic Regime Structure', *Geophysical Research Letters* **47**(15), e2020GL087907.

Doucet, A., De Freitas, N., Gordon, N. J. et al. (2001), *Sequential Monte Carlo methods in practice*, Vol. 1, Springer.

Drouard, M. and Cassou, C. (2019), 'A Modeling- and Process-Oriented Study to Investigate the Projected Change of ENSO-Forced Wintertime Teleconnectivity in a Warmer World', *Journal of Climate* **32**(23), 8047–8068.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L. and Robinson, N. (2014), 'Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?', *Geophysical Research Letters* **41**(15), 5620–5628.

Evensen, G. and van Leeuwen, P. J. (2000), 'An ensemble kalman smoother for nonlinear dynamics', *Monthly Weather Review* **128**(6), 1852–1867.

Fabiano, F., Christensen, H. M., Strommen, K., Athanasiadis, P., Baker, A., Schiemann, R. and Corti, S. (2020), 'Euro-Atlantic weather Regimes in the PRIMAVERA coupled climate simulations: impact of resolution and mean state biases on model performance', *Climate Dynamics* **54**, 5031–5048.

Fabiano, F., Meccia, V., Davini, P., Ghinassi, P. and Corti, S. (2020), 'A regime view of future circulation changes in Northern mid-latitudes', *Weather and Climate Dynamics* **2**, 163–180.

Falkena, S. K. J. (2019), Atmospheric Circulation Regimes: Identification, Number and Statistics, Technical report, University of Reading.

Falkena, S. K. J., de Wiljes, J., Weisheimer, A. and Shepherd, T. G. (2020), 'Revisiting the Identification of Wintertime Atmospheric Circulation Regimes in the Euro-Atlantic Sector', *Quarterly Journal of the Royal Meteorological Society* **146**(731), 2801–2814.

Falkena, S. K. J., de Wiljes, J., Weisheimer, A. and Shepherd, T. G. (2022), 'Detection of interannual ensemble forecast signals over the North Atlantic and Europe using atmospheric circulation regimes', *Quarterly Journal of the Royal Meteorological Society* **148**(742), 434–453.

Fayyad, U., Reina, C. and Bradley, P. S. (1998), Initialization of Iterative Refinement Clustering Algorithms, *in* 'Knowledge Discovery in Databases (KDD)', AAAI.

Feldstein, S. B. and Franzke, C. L. (2017), Atmospheric Teleconnection Patterns, *in* C. L. Franzke and T. J. O'Kane, eds, 'Nonlinear and Stochastic Climate Dynamics', Cambridge University Press, Cambridge, pp. 54–104.

Fereday, D. R., Knight, J. R., Scaife, A. A., Folland, C. K. and Philipp, A. (2008), 'Cluster Analysis of North Atlantic–European Circulation Types and Links with Tropical Pacific Sea Surface Temperatures', *Journal of Climate* **21**(15), 3687–3703.

Ferranti, L., Corti, S. and Janousek, M. (2015), 'Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector', *Quarterly Journal of the Royal Meteorological Society* **141**(688), 916–924.

Fraley, C. and Raftery, A. E. (1998), 'How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis', *The Computer Journal* **41**(8), 578–588.

Franzke, C., Horenko, I., Majda, A. J. and Klein, R. (2009), 'Systematic Metastable Atmospheric Regime Identification in an AGCM', *Journal of the Atmospheric Sciences* **66**(7), 1997–2012.

Franzke, C. L. E., Woollings, T. and Martius, O. (2011), 'Persistent Circulation Regimes and Preferred Regime Transitions in the North Atlantic', *Journal of the Atmospheric Sciences* **68**(12), 2809–2825.

Garfinkel, C. I., Benedict, J. J. and Maloney, E. D. (2014), 'Impact of the MJO on the boreal winter extratropical circulation', *Geophysical Research Letters* **41**(16), 6055–6062.

Garfinkel, C. I. and Hartmann, D. L. (2008), 'Different ENSO teleconnections and their effects on the stratospheric polar vortex', *Journal of Geophysical Research* **113**(D18114).

Ghahramani, Z. (2001), An introduction to hidden markov models and bayesian networks, *in* 'Hidden Markov models: applications in computer vision', World Scientific, pp. 9–41.

Gigerenzer, G. (2004), 'Mindless statistics', *The Journal of Socio-Economics* **33**, 587–606.

Giuntoli, I., Fabiano, F. and Corti, S. (2022), 'Seasonal predictability of Mediterranean weather regimes in the Copernicus C3S systems', *Climate Dynamics* **58**, 2131–2147.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I. and Wernli, H. (2017), 'Balancing Europe's wind-power output through spatial deployment informed by weather regimes', *Nature Climate Change* **7**(8), 557–562.

Grams, C. M., Magnusson, L. and Madonna, E. (2018), 'An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study', *Quarterly Journal of the Royal Meteorological Society* **144**(717), 2577–2591.

Gurobi Optimization LLC (2019), 'Gurobi Optimizer Reference Manual'.
URL: *http://www.gurobi.com*

Hamilton, J. D. (1989), 'A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle', *Econometrica* **57**(2), 357–384.

Hannachi, A. and O'Neill, A. (2001), 'Atmospheric multiple equilibria and non-Gaussian behaviour in model simulations', *Quarterly Journal of the Royal Meteorological Society* **127**(573), 939–958.

Hannachi, A., Straus, D. M., Franzke, C. L. E., Corti, S. and Woollings, T. (2017), 'Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere', *Reviews of Geophysics* **55**(1), 199–234.

Hannachi, A., Woollings, T. and Fraedrich, K. (2012), 'The North Atlantic jet stream: A look at preferred positions, paths and transitions', *Quarterly Journal of the Royal Meteorological Society* **138**(665), 862–877.

Hansen, A. R. and Sutera, A. (1986), 'On the Probability Density Distribution of Planetary-Scale Atmospheric Wave Amplitude', *Journal of Atmospheric Sciences* **43**(24), 3250–3265.

Hardiman, S. C., Dunstone, N. J., Scaife, A. A., Smith, D. M., Comer, R., Nie, Y. and Ren, H.-l. (2022), 'Missing eddy feedback may explain weak signal-to-noise ratios in climate predictions', *npj Climate and Atmospheric Science* **5**(57).

Henderson, S. A., Maloney, E. D. and Barnes, E. A. (2016), 'The Influence of the Madden–Julian Oscillation on Northern Hemisphere Winter Blocking', *Journal of Climate* **29**(12), 4597–4616.

Hitchcock, P. and Simpson, I. R. (2014), 'The Downward Influence of Stratospheric Sudden Warmings', *Journal of the Atmospheric Sciences* **71**(10), 3856–3876.

Hochman, A., Messori, G., Quinting, J. F., Pinto, J. G. and Grams, C. M. (2021), 'Do Atlantic-European weather regimes physically exist?', *Geophysical Research Letters* **48**(e2021GL095574).

Holton, J. R. and Tan, H.-C. (1980), 'The Influence of the Equatorial Quasi-Biennial Oscillation on the Global Circulation at 50 mb', *Journal of the Atmospheric Sciences* **37**(10), 2200–2208.

Horel, J. D. and Wallace, J. M. (1981), 'Planetary-Scale Atmospheric Phenomena Associated with the Southern Oscillation', *Monthly Weather Review* **109**(4), 813–829.

Horenko, I. (2010*a*), 'Finite element approach to clustering of multidimensional time series', *SIAM Journal on Scientific Computing* **30**(3), 62–83.

Horenko, I. (2010*b*), 'On clustering of non-stationary meteorological time series', *Dynamics of Atmospheres and Oceans* **49**(2-3), 164–187.

Hoskins, B. J. and Karoly, D. J. (1981), 'The Steady Linear Response of a Spherical Atmosphere to Thermal and Orographic Forcing', *Journal of the Atmospheric Sciences* **38**(6), 1179–1196.

Hu, C.-C. and van Leeuwen, P. J. (2021), 'A particle flow filter for high-dimensional system applications', *Quarterly Journal of the Royal Meteorological Society* **147**(737), 2352–2374.

Hurrell, J. W. (1995), 'Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation', *Science* **269**(5224), 676–679.

Ineson, S. and Scaife, A. A. (2009), 'The role of the stratosphere in the European climate response to El Niño', *Nature Geoscience* **2**, 32–36.

IPCC (2021), Summary for Policymakers, *in* 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, eds, 'Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 3–32.

Jain, A. K. (2010), 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters* **31**(8), 651–666.

Jain, A. K. and Flynn, P. (1996), 'Image segmentation using clustering', *Advances in Image Understanding* pp. 65–83.

Johnson, N. C. and Feldstein, S. B. (2010), 'The continuum of North Pacific sea level pressure patterns: Intraseasonal, interannual, and interdecadal variability', *Journal of Climate* **23**(4), 851–867.

Johnson, S. C. (1967), 'Hierarchical Clustering Schemes', *Psychometrika* **32**, 241–254.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P., Mogensen, K., Zuo, H. and Monge-Sanz, B. M. (2019), 'SEAS5: The new ECMWF seasonal forecast system', *Geoscientific Model Development* **12**(3), 1087–1117.

Jung, T., Palmer, T. N. and Shutts, G. J. (2005), 'Influence of a stochastic parameterization on the frequency of occurrence of North Pacific weather regimes in the ECMWF model', *Geophysical Research Letters* **32**(L23811).

Kageyama, M., D'Andrea, F., Ramstein, G., Valdes, P. J. and Vautard, R. (1999), 'Weather regimes in past climate atmospheric general circulation model simulations', *Climate Dynamics* **15**(10), 773–793.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Transaction of the ASME Journal of Basic Engineering* pp. 35–45.

Kantas, N., Beskos, A. and Jasra, A. (2014), 'Sequential monte carlo methods for high-dimensional inverse problems: A case study for the navier–stokes equations', *SIAM/ASA Journal on Uncertainty Quantification* **2**(1), 464–489.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kaufman, L. and Rousseeuw, P. J. (2005), *Finding groups in data: An introduction to cluster analysis*, Wiley series in Probability and Statistics.

Kelder, T., Müller, M., Slater, L. J., Marjoribanks, T. I., Wilby, R. L., Prudhomme, C., Bohlinger, P., Ferranti, L. and Nipen, T. (2020), 'Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes', *npj Climate and Atmospheric Science* **3**(47).

Kidson, J. W. (1988), 'Interannual Variations in the Southern Hemisphere Circulation', *Journal of Climate* **1**(12), 1177–1198.

Kimoto, M. and Ghil, M. (1993), 'Multiple Flow Regimes in the Northern Hemisphere Winter. Part I: Methodology and Hemispheric Regimes', *Journal of Atmospheric Sciences* **50**(16), 2625–2643.

Kretschmer, M., Zappa, G. and Shepherd, T. G. (2020), 'The role of Barents–Kara sea ice loss in projected polar vortex changes', *Weather and Climate Dynamics* **1**, 715–730.

Krolzig, H.-M. (1997), *Markov-Switching Vector Autoregressions*, Springer-Verlag.

Kumar, A. (2009), 'Finite samples and uncertainty estimates for skill measures for seasonal prediction', *Monthly Weather Review* **137**(8), 2622–2631.

Lee, J. C. K., Lee, R. W., Woolnough, S. J. and Boxall, L. J. (2020), 'The links between the Madden-Julian Oscillation and European weather regimes', *Theoretical and Applied Climatology* **141**, 567–586.

Lee, R. W., Woolnough, S. J., Charlton-Perez, A. J. and Vitart, F. (2019), 'ENSO Modulation of MJO Teleconnections to the North Atlantic and Europe', *Geophysical Research Letters* **46**, 13,535–13,545.

Li, Y. and Lau, N.-C. (2012), 'Impact of ENSO on the Atmospheric Variability over the North Atlantic in Late Winter—Role of Transient Eddies', *Journal of Climate* **25**(1), 320–342.

Liu, H., Tosi, E. and Tibaldi, S. (1994), 'On the relationship between northern hemispheric weather regimes in wintertime and spring precipitation over China', *Quarterly Journal of the Royal Meteorological Society* **120**(515), 185–194.

Lloyd, S. (1982), 'Least squares quantization in PCM', *IEEE Transactions on Information Theory* **28**(2), 129–137.

Lorenz, E. N. (1963), 'Deterministic Nonperiodic Flow', *Journal of Atmospheric Sciences* **20**(2), 130–141.

Lorenz, E. N. (1969), 'The predictability of a flow which possesses many scales of motion', *Tellus* **21**(3), 289–307.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', pp. 281–297.

Madden, R. A. and Julian, P. R. (1971), 'Detection of a 40-50 Day Oscillation in the Zonal Wind in the Tropical Pacific', *Journal of the Atmospheric Sciences* **28**(5), 702–708.

Madonna, E., Li, C., Grams, C. M. and Woollings, T. (2017), 'The link between eddy-driven jet variability and weather regimes in the North Atlantic-European sector', *Quarterly Journal of the Royal Meteorological Society* **143**(708), 2960–2972.

Majda, A. J., Franzke, C. L., Fischer, A. and Crommelin, D. T. (2006), 'Distinct metastable atmospheric regimes despite nearly Gaussian statistics: A paradigm model', *Proceedings of the National Academy of Sciences* **103**(22), 8309–8314.

Mathieu, P.-P., Sutton, R. T., Dong, B. and Collins, M. (2004), 'Predictability of Winter Climate over the North Atlantic European Region during ENSO Events', *Journal of Climate* **17**(10), 1953–1974.

Matsueda, M. and Palmer, T. N. (2018), 'Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts', *Quarterly Journal of the Royal Meteorological Society* **144**(713), 1012–1027.

Metzner, P., Putzig, L. and Horenko, I. (2012), 'Analysis of Persistent Nonstationary Time Series and Applications', *Communications in Applied Mathematics and Computational Science* **7**(2), 175–229.

Michel, C. and Rivière, G. (2011), 'The Link between Rossby Wave Breakings and Weather Regime Transitions', *Journal of the Atmospheric Sciences* **68**(8), 1730–1748.

Michelangeli, P.-A., Vautard, R. and Legras, B. (1995), 'Weather Regimes: Recurrence and Quasi Stationarity', *Journal of Atmospheric Sciences* **52**(8), 1237–1256.

Mo, K. C. and Ghil, M. (1987), 'Statistics and Dynamics of Persistent Anomalies', *Journal of Atmospheric Sciences* **44**(5), 877–901.

Mo, K. and Ghil, M. (1988), 'Cluster analysis of multiple planetary flow regimes', *Journal of Geophysical Research* **93**(D9), 10,927–10,952.

Molteni, F., Tibaldi, S. and Palmer, T. N. (1990), 'Regimes in the wintertime circulation over northern extratropics. I: Observational Evidence', *Quarterly Journal of the Royal Meteorological Society* **116**, 31–67.

Moron, V. and Plaut, G. (2003), 'The Impact of El Nino-Southern Oscillation Upon Weather Regimes over Europe and the North Atlantic During Boreal Winter', *International Journal of Climatology* **23**, 363–379.

Murari, A., Peluso, E., Cianfrani, F., Gaudio, P. and Lungaroni, M. (2019), 'On the Use of Entropy to Improve Model Selection Criteria', *Entropy* **21**(4), 394.

Nicholls, N. (2001), 'The Insignificance of Significance Testing', *Bulletin of the American Meteorological Society* **82**(5), 981–986.

O'Kane, T. J., Risbey, J. S., Franzke, C., Horenko, I. and Monselesan, D. P. (2013), 'Changes in the Metastability of the Midlatitude Southern Hemisphere Circulation and the Utility of Nonstationary Cluster Analysis and Split-Flow Blocking Indices as Diagnostic Tools', *Journal of the Atmospheric Sciences* **70**(3), 824–842.

O'Reilly, C. H., Weisheimer, A., Woollings, T., Gray, L. J. and Macleod, D. (2019), 'The importance of stratospheric initial conditions for winter North Atlantic Oscillation predictability and impli-

cations for the signal-to-noise paradox', *Quarterly Journal of the Royal Meteorological Society* **145**(718), 131–146.

Ortizbeviá, M. J., Sánchezgómez, E. and Alvarez-García, F. J. (2011), 'North Atlantic atmospheric regimes and winter extremes in the Iberian peninsula', *Natural Hazards and Earth System Science* **11**(3), 971–980.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A., Orfila, B., Rogel, P., Terres, J.-M. and Thomson, M. (2004), 'Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER)', *Bulletin of the American Meteorological Society* **85**(6), 853–872.

Pelleg, D. and Moore, A. (1999), Accelerating Exact k-means Algorithms with Geometric Reasoning, *in* 'Knowledge Discovery in Databases (KDD)', AAAI, pp. 277–281.

Pelleg, D. and Moore, A. (2000), X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *in* '17th International Conference on Machine Learning', pp. 727–734.

Pelly, J. L. and Hoskins, B. J. (2003), 'How well does the ECMWF Ensemble Prediction System predict blocking?', *Quarterly Journal of the Royal Meteorological Society* **129**, 1683–1702.

Philander, S. G. H. (1983), 'El Nino Southern Oscillation phenomena', *Nature* **302**, 295–301.

Philipp, A., Della-Marta, P. M., Jacobeit, J., Fereday, D. R., Jones, P. D., Moberg, A. and Wanner, H. (2007), 'Long-term variability of daily North Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering', *Journal of Climate* **20**(16), 4065–4095.

Plaut, G. and Simonnet, E. (2001), 'Large-scale circulation classification, weather regimes, and local climate over France, the Alps and western Europe', *Climate Research* **17**(3), 303–324.

Polvani, L. M., Sun, L., Butler, A. H., Richter, J. H. and Deser, C. (2017), 'Distinguishing Stratospheric Sudden Warmings from ENSO as Key Drivers of Wintertime Climate Variability over the North Atlantic and Eurasia', *Journal of Climate* **30**(6), 1959–1969.

Polvani, L. M. and Waugh, D. W. (2004), 'Upward Wave Activity Flux as a Precursor to Extreme Stratospheric Events and Subsequent Anomalous Surface Weather Regimes', *Journal of Climate* **17**(18), 3548–3554.

Portal, A., Ruggieri, P., Palmeiro, F. M., García-Serrano, J., Domeisen, D. I. V. and Gualdi, S. (2022), 'Seasonal prediction of the boreal winter stratosphere', *Climate Dynamics* **58**, 2109–2130.

Posada, D. and Buckley, T. R. (2004), 'Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests', *Systematic Biology* **53**(5), 793–808.

Quinn, C., Harries, D. and O'Kane, T. J. (2020), 'Dynamical analysis of a reduced model for the North Atlantic Oscillation', *Journal of the Atmospheric Sciences* **78**(5), 1647–1671.

Rabiner, L. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Rasmusson, E. M. and Wallace, J. M. (1983), 'Meteorological Aspects of the El Nino/Southern Oscillation', *Science* **222**(4629), 1195–1202.

Rex, D. F. (1950), 'Blocking Action in the Middle Troposphere and its Effect upon Regional Climate. II: The climatology of Blocking Action', *Tellus* **2**(4), 275–301.

Riddle, E. E., Stoner, M. B., Johnson, N. C., L'Heureux, M. L., Collins, D. C. and Feldstein, S. B. (2013), 'The impact of the MJO on clusters of wintertime circulation anomalies over the North American region', *Climate Dynamics* **40**, 1749–1766.

Robertson, A. W. and Ghil, M. (1999), 'Large-scale weather regimes and local climate over the western United States', *Journal of Climate* **12**(6), 1796–1813.

Rossi, R., Murari, A., Gaudio, P. and Gelfusa, M. (2020), 'Upgrading model selection criteria with goodness of fit tests for practical applications', *Entropy* **22**(4), 447.

Ruggieri, P., Alvarez-Castro, M. C., Athanasiadis, P., Bellucci, A., Materia, S. and Gualdi, S. (2020), 'North Atlantic Circulation Regimes and Heat Transport by Synoptic Eddies', *Journal of Climate* **33**(11), 4769–4785.

Sardeshmukh, P. D. and Hoskins, B. J. (1988), 'The Generation of Global Rotational Flow by Steady Idealized Tropical Divergence', *Journal of the Atmospheric Sciences* **45**(7), 1228–1251.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J. and Williams, A. (2014), 'Skillful long-range prediction of European and North American winters', *Geophysical Research Letters* **41**(9), 3307–3314.

Scaife, A. A. and Smith, D. (2018), 'A signal-to-noise paradox in climate science', *npj Climate and Atmospheric Science* **1**, 28.

Schwarz, G. (1978), 'Estimating the Dimension of a Model', *The Annals of Statistics* **6**(2), 461–464.

Shannon, C. E. (1948), 'A Mathematical Theory of Communication', *Bell System Technical Journal* **27**(3), 379–423.

Shannon, C. and Weaver, W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

Shepherd, T. G. (2019), 'Storyline approach to the construction of regional climate change information', *Proceedings of the Royal Society A* **475**(20190013).

Shepherd, T. G. (2021), 'Bringing physical reasoning into statistical practice in climate-change science', *Climatic Change* **169**(2).

Shukla, J. (1998), 'Predictability in the Midst of Chaos: A Scientific Basis for Climate Forecasting', *Science* **282**(5389), 728–731.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R. and Arribas, A. (2016), 'A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability?', *Journal of Climate* **29**(3), 995–1012.

Smyth, P., Ide, K. and Ghil, M. (1999), 'Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering', *Journal of the Atmospheric Sciences* **56**(21), 3704–3723.

Steinhaus, H. (1956), 'Sur la division de corp materiels en parties', *Bulletin L'Académie Polonaise des Science* **IV**(C1.III), 801–804.

Stephenson, D., Hannachi, A. and O'Neill, A. (2004), 'On the existence of multiple climate regimes', *Quarterly Journal of the Royal Meteorological Society* **130**(597), 583–605.

Straus, D. M., Corti, S. and Molteni, F. (2007), 'Circulation regimes: Chaotic variability versus SST-forced predictability', *Journal of Climate* **20**(10), 2251–2272.

Straus, D. M. and Molteni, F. (2004), 'Circulation Regimes and SST Forcing: Results from Large GCM Ensembles', *Journal of Climate* **17**(8), 1641–1656.

Straus, D. M., Molteni, F. and Corti, S. (2017), *Atmospheric regimes: The link between weather and the large-scale circulation*, Cambridge University Press.

Straus, D. M., Swenson, E. and Lappen, C.-L. (2015), 'The MJO Cycle Forcing of the North Atlantic Circulation: Intervention Experiments with the Community Earth System Model', *Journal of the Atmospheric Sciences* **72**(2), 660–681.

Strommen, K., Mavilia, I., Corti, S., Matsueda, M., Davini, P., Hardenberg, J., Vidale, P. and Mizuta, R. (2019), 'The Sensitivity of Euro-Atlantic Regimes to Model Horizontal Resolution', *Geophysical Research Letters* **46**, 7810–7818.

Strommen, K. and Palmer, T. N. (2019), 'Signal and noise in regime systems : A hypothesis on the predictability of the North Atlantic Oscillation', *Quarterly Journal of the Royal Meteorological Society* **145**, 147–163.

Swenson, E. T. and Straus, D. M. (2017), 'Rossby wave breaking and transient eddy forcing during Euro-Atlantic circulation regimes', *Journal of the Atmospheric Sciences* **74**(6), 1735–1755.

Thompson, V., Dunstone, N. J., Scaife, A. A., Smith, D. M., Slingo, J. M., Brown, S. and Belcher, S. E. (2017), 'High risk of unprecedented UK rainfall in the current climate', *Nature Communications* **8**(107).

Tibaldi, S. and Molteni, F. (1990), 'On the operational predictability of blocking', *Tellus* **42A**, 343–365.

Toniazzo, T. and Scaife, A. A. (2006), 'The influence of ENSO on winter North Atlantic climate', *Geophysical Research Letters* **33**(L24704).

Toth, Z. (1993), 'Preferred and Unpreferred Circulation Types in the Northern Hemisphere Wintertime Phase Space', *Journal of Atmospheric Sciences* **50**(17), 2868–2888.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C. and Ropelewski, C. (1998), 'Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures', *Journal of Geophysical Research* **103**(C7), 14,291–14,324.

van den Dool, H. M. and Chervin, R. M. (1986), 'A Comparison of Month-to-Month Persistence of Anomalies in a General Circulation Model and in the Earth's Atmosphere', *Journal of the Atmospheric Sciences* **43**(14), 1454–1466.

van der Wiel, K., Bloomfield, H. C., Lee, R. W., Stoop, L. P., Blackport, R., Screen, J. A. and Selten, F. M. (2019), 'The influence of weather regimes on European renewable energy production and demand', *Environmental Research Letters* **14**(094010).

Vautard, R. (1990), 'Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors', *Monthly Weather Review* **118**, 2056–2081.

Vigaud, N., Robertson, A. W. and Tippett, M. (2018), 'Predictability of Recurrent Weather Regimes over North America during Winter from Submonthly Reforecasts', *Monthly Weather Review* **146**, 2559–2577.

Volinsky, C. T. and Raftery, A. E. (2000), 'Bayesian information criterion for censored survival models', *Biometrics* **56**, 256–262.

Von Luxburg, U. (2007), 'A Tutorial on Spectral Clustering', *Statistics and Computing* **17**(4), 395–416.

Wagenmakers, E.-J. (2007), 'A practical solution to the pervasive problems of p values', *Psychonomic Bulletin & Review* **14**(5), 779–804.

Walker, G. T. and Bliss, E. (1932), 'World Weather V', *Memoirs of the Royal Meteorological Society* **4**(36), 53–84.

Wallace, J. M. and Gutzler, David, S. (1981), 'Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter', *Monthly Weather Review* **109**(4), 784–812.

Wang, B., Lee, J.-Y., Kang, I.-S., Shukla, J., Park, C.-K., Kumar, A., Schemm, J., Cocke, S., Kug, J.-S., Luo, J.-J., Zhou, T., Wang, B., Fu, X., Yun, W.-T., Alves, O., Jin, E., Kinter, J., Kirtman, B., Krishnamurti, T., Lau, N., Lau, W., Liu, P., Pegion, P., Rosati, T., Schuibert, S., Stern, W., Suarez, M. and Yamagata, T. (2009), 'Advance and prospectus of seasonal prediction: assessment of the APCC / CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004)', *Climate Dynamics* **33**, 93–117.

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T. N., Johnson, S. and Palmer, T. N. (2019), 'How confident are predictability estimates of the winter North Atlantic Oscillation?', *Quarterly Journal of the Royal Meteorological Society* **145**(S1), 140–159.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A. and Palmer, T. (2017), 'Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution', *Quarterly Journal of the Royal Meteorological Society* **143**(703), 917–926.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., Vitart, F., Perez, C. D., Ray, A. J., Murray, V., Bharwani, S., Macleod, D., Fleming, L., Morse, A. P.,

Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., Mcevoy, D., Depledge, M., Perkins-kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-johnston, I., Buontempo, C., Lamb, R., Meinke, H. and Zebiak, S. E. (2017), 'Potential applications of subseasonal-to-seasonal (S2S) predictions', *Meteorological Applications* **24**, 315–325.

Woollings, T., Hannachi, A. and Hoskins, B. (2010), 'Variability of the North Atlantic eddy-driven jet stream', *Quarterly Journal of the Royal Meteorological Society* **136**(649), 856–868.

Yadav, P., Straus, D. M. and Swenson, E. T. (2019), 'The Euro-Atlantic Circulation Response to the Madden-Julian Oscillation Cycle of Tropical Heating: Coupled GCM Intervention Experiments', *Atmosphere - Ocean* **57**(3), 161–181.

Yeh, S.-W., Cai, W., Min, S.-K., Mcphaden, M. J., Dommenget, D., Dewitte, B., Collins, M., Ashok, K., An, S.-I., Yim, B.-Y. and Kug, J.-S. (2018), 'ENSO Atmospheric Teleconnections and Their Response to Greenhouse Gas Forcing', *Reviews of Geophysics* **56**(1), 185–206.

Yiou, P., Goubanova, K., Li, Z. X. and Nogaj, M. (2008), 'Weather regime dependence of extreme value statistics for summer temperature and precipitation', *Nonlinear Processes in Geophysics* **15**(3), 365–378.

Yiou, P. and Nogaj, M. (2004), 'Extreme climatic events and weather regimes over the North Atlantic: When and where?', *Geophysical Research Letters* **31**(7).

Zappa, G. and Shepherd, T. G. (2017), 'Storylines of atmospheric circulation change for European regional climate impact assessment', *Journal of Climate* **30**(16), 6561–6577.

Zelnik-Manor, L. and Perona, P. (2005), 'Self-Tuning Spectral Clustering', *Advances in neural information processing systems* **17**, 1601–1608.

Zhang, W. and Kirtman, B. (2019), 'Understanding the Signal-to-Noise Paradox with a Simple Markov Model', *Geophysical Research Letters* **46**(13), 308–317.

Appendix A

Relating Average Regime Duration, Self-Transition Probability and e-Folding Time Scale

In this appendix we discuss the derivation of relationships between the self-transition probabilities and the corresponding e-folding timescale and average regime duration. The main assumption for deriving these relations is that the regime dynamics is can be approximated by a first-order Markov chain, meaning the current state fully determines the probability of the state the next day.

Let p be the probability of a regime transitioning into itself. Consider the exponential e^{-t/T_e}, where t is time and T_e the e-folding time (both in days), describing the decay in likelihood of the atmosphere still being in the same regime after time t. For $t = 1$ day we can relate the transition probability to the e-folding time scale by

$$\begin{aligned} \mathrm{e}^{-1/T_e} &= p, \\ -1/T_e &= \log(p), \\ T_e &= -\frac{1}{\log(p)}, \end{aligned} \tag{A.1}$$

meaning that if we know one, we can compute the other.

Next we relate the self-transition probability to the expected (or average) regime duration. Starting from a regime with self-transition probability p (day 0), the expected time it takes to transfer out of that regime (O) is

$$\begin{aligned}
\mathbb{E}(\text{days to } O) &= \sum_{n=1}^{\infty} nP(O \text{ at day } n), \\
&= \sum_{n=1}^{\infty} np^{n-1}(1-p), \\
&= (1-p)\sum_{n=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}p} p^n, \\
&= (1-p)\frac{\mathrm{d}}{\mathrm{d}p}\sum_{n=0}^{\infty} p^n, \\
&= (1-p)\frac{\mathrm{d}}{\mathrm{d}p}\Big(\frac{1}{1-p}\Big), \\
&= (1-p)\frac{1}{(1-p)^2} = \frac{1}{1-p}.
\end{aligned} \tag{A.2}$$

We denote the found average regime duration by $\mathbb{E}(\text{days to } O) = T_{av}$. Note that the Taylor series of $-\log(p)$ around 1 is

$$-\log(p) = (1-p) + \mathcal{O}(p^2), \tag{A.3}$$

meaning that in the limit $p \to 1$ we have that T_e and T_{av} become equal (by comparing Equation (A.1) and (A.2)).

The found relations between p and either T_e or T_{av} allow us to also express the e-folding time T_e as a function of T_{av} and the other way around. This yields

$$T_e = -\frac{1}{\log(T_{av}-1) - \log(T_{av})}, \qquad T_{av} = \frac{1}{1-\exp(-1/T_e)}. \tag{A.4}$$

Appendix B

A Comparison Between Regimes for ERA-Interim and SEAS5 Data

In this thesis both ERA-Interim and ECMWF SEAS5 model hindcast data are used for the regime identification. When comparing the regime results between both these datasets it is important to be aware of the differences between them and the biases present in the model that could affect the regimes. In this appendix we discuss these correspondences and differences, starting with some notes on the general differences in Section B.1, followed by a regime-tailored discussion in Section B.2.

B.1 General Differences

The main interest in this thesis is on the circulation regimes and their dynamics, these can be impacted by different aspects of the data. In this section of Appendix B we focus on differences in the climatology, which is used as a background state with respect to which anomalies are computed, in the variance and in the autocorrelation. These three aspects are studied for two domains; 20° to 80°N and 90°W to 30°E (domain A, dashed box) and 30° to 90°N and 80°W to 40°E (domain B, dash-dotted box). These are the two domains considered in Chapter 4.

We start by looking at the difference in the background climatology as shown in Figure B.1. We find that SEAS5 has a lower geopotential height over the central Atlantic extending north-eastwards over Europe and a higher geopotential height over Greenland and the northern parts of both domains. The regimes are computed using anomaly data with respect to these respective background

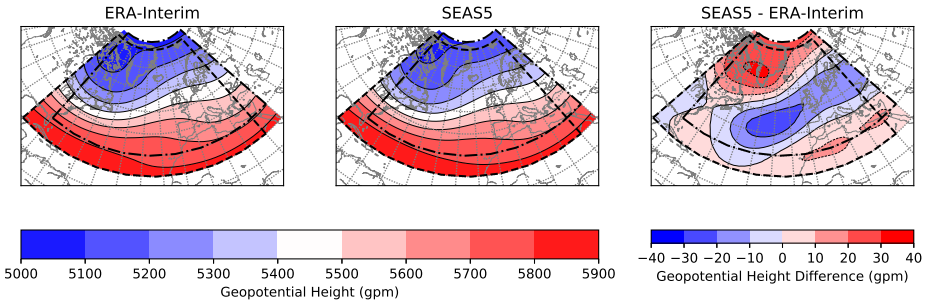


Figure B.1: The average climatology of ERA-Interim and SEAS5 and their difference. The boxes indicate the two domains considered.

climatologies, meaning these do not provide direct information on the anomalies. Nevertheless, it is possible that the differences in climatology do affect the identified regime patterns.

Next we turn to the variance around the climatology, where we not only consider deviations with respect to the fixed climatologies shown in Figure B.1, but also with respect to a seasonally varying climatology. The variance over the domain for both datasets and climatologies is shown in Figure B.2. Firstly, we note that the variance is smaller when using a seasonal climatology than when considering a fixed one. In general this difference is small, except over Canada/North America, which is the case for both ERA-Interim and SEAS5. Interestingly this effect is larger for ERA-Interim than for SEAS5. The difference in the effect of the used climatology between SEAS5 and ERA-Interim shows a larger variance difference for SEAS5 in the central and northern Atlantic, while over the United States and Canada the variance difference between the two climatologies is a lot smaller. The differences between ERA-Interim and SEAS5 are very similar for both climatologies. On average the variance of SEAS5 is slightly higher, with largest differences off the coast of Spain/France followed by Eastern Europe. There is a small area near Iceland where the variance of SEAS5 is slightly lower than that of ERA-Interim.

Lastly, we discuss the autocorrelation for a lag of one day shown in Figure B.3. As for the variance we see that using a seasonal climatology reduces the autocorrelation compared to a fixed climatology. Again, this effect is stronger for ERA-Interim than for SEAS5, with the differences occurring in roughly the same regions, albeit with different strengths. We also see that the difference increases with increasing lag, for both the differences between the two climatologies and between ERA-Interim and SEAS5. The differences in autocorrelation between ERA-Interim and SEAS5 show a more varied pattern than those for the variance. We find a lower autocorrelation for SEAS5 in the

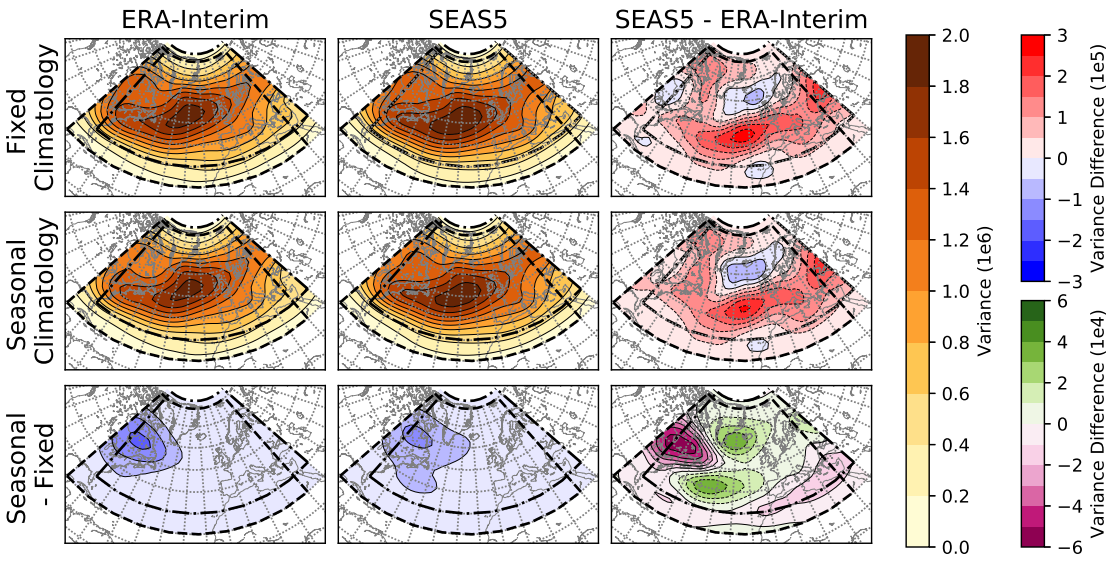


Figure B.2: The variance of ERA-Interim and SEAS5 and their difference using either a fixed or seasonal climatology and their differences.

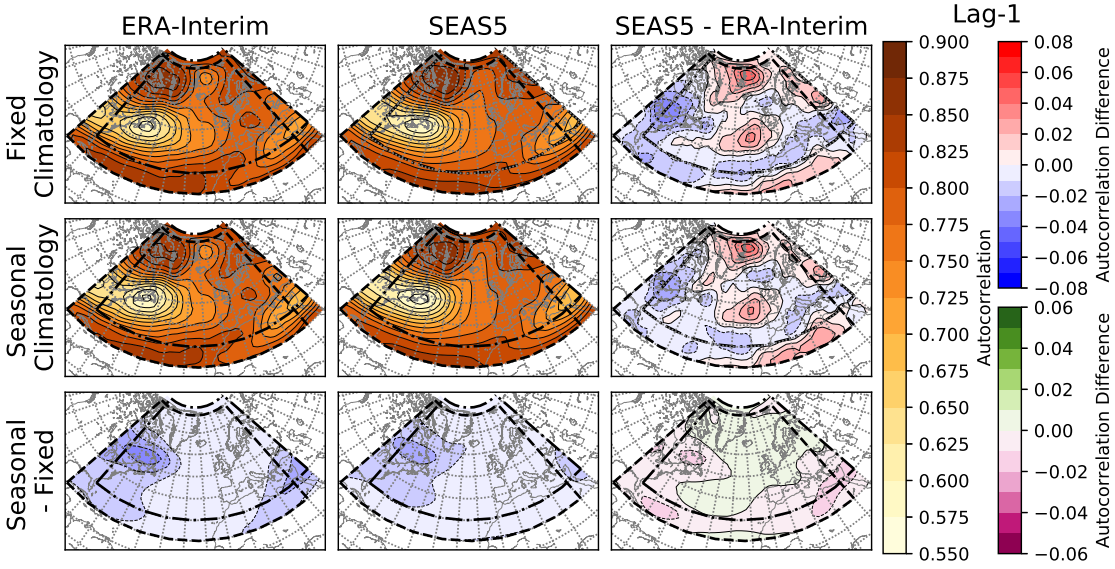


Figure B.3: The autocorrelation of ERA-Interim and SEAS5 using either a fixed or seasonal climatology and their differences at a 1-day lag

Table B.1: The pattern correlation between the SEAS5 ensemble regimes and the ERA-Interim regimes for $k = 4$ and $k = 6$.

k	AR+	SB+	NAO+	NAO-	AR-	SB-	Mean
4	0.97	0.98	0.97	0.99			0.98
6	0.97	0.89	0.91	0.98	0.95	0.96	0.94

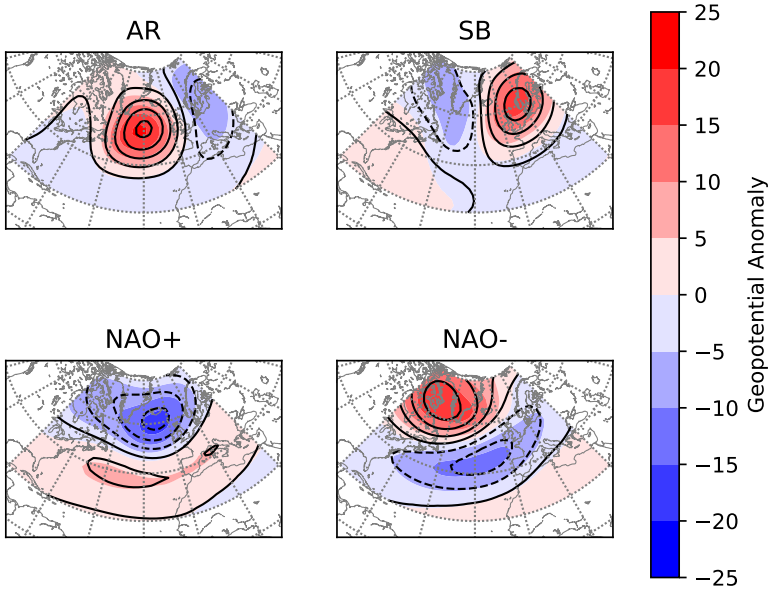
south-western part of the domain, as well around Italy and north of the UK. A higher autocorrelation occurs over the Baltic states and in a band throughout the centre of the domain. These difference patterns are roughly the same for longer lags, becoming stronger the longer the lag.

B.2 Differences in the Regimes

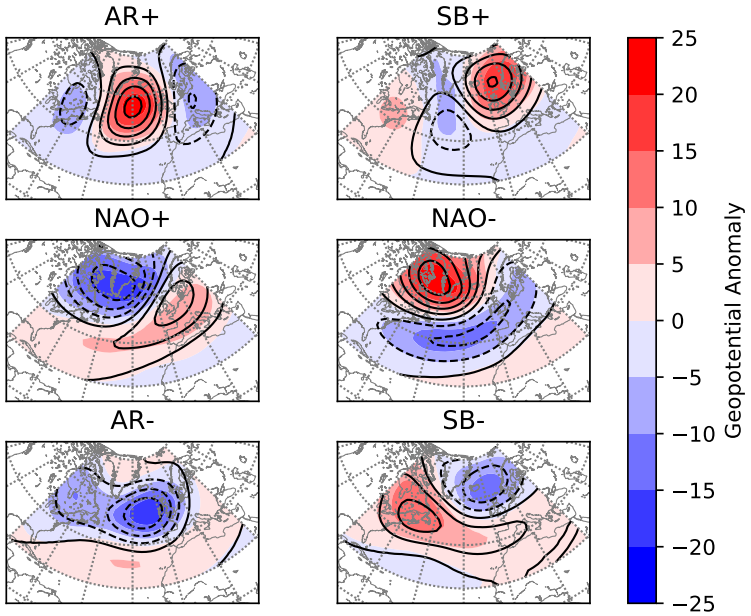
The discussed differences in climatology, variance and autocorrelation can affect the regimes identified, as well as their dynamics. In this section we study the difference between the regimes of the two datasets, both for $k = 4$ and $k = 6$, where also the effects of the background climatology and domain are considered. All of this is done using a standard k-means clustering algorithm applied to gridpoint data. It is not expected that the SEAS5 and ERA-Interim results are the same, but it is desirable for the ERA-Interim results to fall within the ensemble spread of SEAS5. We start with some notes on the regime patterns themselves, after which the optimal number of regimes and the regime dynamics are considered. In the end some brief comments are made on the incorporation of a persistence regularisation within the k-means clustering algorithm applied to the SEAS5 model ensemble data (method discussed in Chapter 3).

B.2.1 Regime Patterns

The regime patterns differ between ERA-Interim and SEAS5, but for both $k = 4$ and $k = 6$ they can straightforwardly be linked one-to-one as shown in Figure B.4. The regimes overall look quite similar, but changes in the location of high- and low geopotential height centres are found. To assess the correspondence of the regime patterns, we compute the pattern correlation between the regimes. These are given in Table B.1 and show a very high correspondence for all regimes for $k = 4$. When using six regimes the correspondence is slightly lower but still good, with the lowest pattern correlation for SB+ and NAO+.



(a) Regimes for $k = 4$.



(b) Regimes for $k = 6$.

Figure B.4: The regimes for ERA-Interim (colour) and those for the SEAS5 ensemble (contours at same intervals, dashed for negative values).

These results are obtained using a constant climatology for the primary domain A considered in this thesis. However, the domain and background climatology do affect the regimes for SEAS5, as also discussed in Chapter 4, while for ERA-Interim the regime patterns are robust to changes in these two aspects. When using four regimes the results are still quite robust, but for $k = 6$ this is no longer the case. The use of a seasonal climatology does substantially alter the regime patterns, with two NAO−-like regimes and the loss of AR−. This likely is due to a different seasonal cycle in the model. The effect of a different background climatology, also means the good correspondence with the ERA-Interim regimes is lost. The same is true when using domain B for SEAS5, for which the regime patterns change as well. These differences could be related to the more northern domain (B) having on average a higher (positive) difference in the variance between ERA-Interim and SEAS5 (Figure B.2). Thus, for domain A with a fixed climatology, being the focus in this thesis we find the highest correspondence between the ERA-Interim and SEAS5 regimes.

It is relevant to know what the ensemble spread in regime patterns is, as the amount of data available is a lot lower for ERA-Interim compared to using all 51 SEAS5 ensemble members. When clustering single ensemble members we find that the obtained regimes show a large spread. This indicates that for the SEAS5 data, the 36 years available are not sufficient to robustly identify the circulation regimes. This holds for both four and six regimes. The higher variability in SEAS5 likely contributes to this less pronounced regime behaviour. The regimes obtained for ERA-interim do fall within the ensemble spread, but are more identifiable as they are quite robust to subsampling within the ERA-Interim dataset. The more ensemble members are considered, the more robust the regimes become, although there still are cases where the regimes are substantially different when considering the pattern correlation. Overall, this indicates that the SEAS5 model shows less robust regime dynamics than ERA-Interim, but due to the ensemble size still yields reliable regime dynamics.

B.2.2 Regime Dynamics

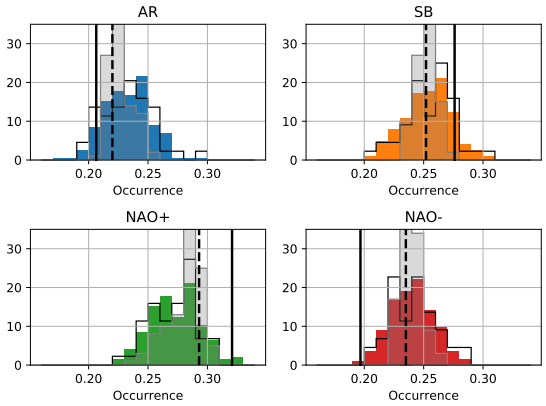
Next to the regime patterns it is important to check that the regime dynamics of ERA-Interim falls within the ensemble spread of SEAS5. Here, the average regime frequency and transition probabilities are considered. We consider the spread in both these aspects using one, five or ten ensemble members a year to investigate whether the ERA-Interim values fall within the ensemble spread.

In Figure B.5 the distributions of occurrence rates for the ensemble are shown for $k = 4$ and $k = 6$. The first thing to note is that the value of ERA-Interim always falls within the ensemble

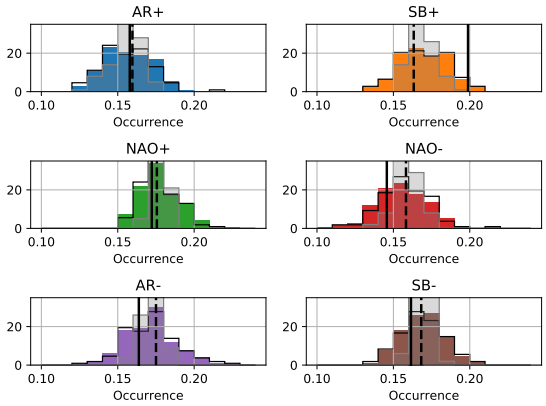
spread, albeit at edge for a significant number of regimes. Using four regimes this is more true than using six regimes. Furthermore, we see that by increasing the number of ensemble members a year the distributions move closer to the full ensemble value, as is expected, leading to the ERA-Interim value falling outside of the ensemble spread in some cases. Most notable are the underestimation of NAO+ and overestimation of NAO− occurrence rates for $k = 4$. For six regimes primarily the occurrence of SB+ is underestimated.

Next we look at the transition probabilities, i.e. the frequency of regime A transitioning to regime B throughout the time series. The distributions for SEAS5 are shown in Figure B.6. As for the regime frequencies the ERA-Interim values fall within the ensemble spread. On average the regime persistence, i.e. the transition probability of a regime to itself, of ERA-Interim lies on the higher end of the ensemble spread. This lower SEAS5 persistence could be related to the lower autocorrelation (on average) over domain A, as found in Figure B.3. For the regime transition probabilities ERA-Interim values lie more on the edge of the SEAS5 spread for transitions out of SB when considering four regimes. For $k = 6$ the picture is more mixed, with a number of cases where ERA-Interim values lie at the edge of the ensemble spread.

Overall we find that for both the regime frequencies and transition probabilities the results for the ERA-Interim regimes fall within the ensemble spread of SEAS5. This indicates that, after a bias correction by using the respective regimes obtained for the ERA-Interim and SEAS5 datasets, the

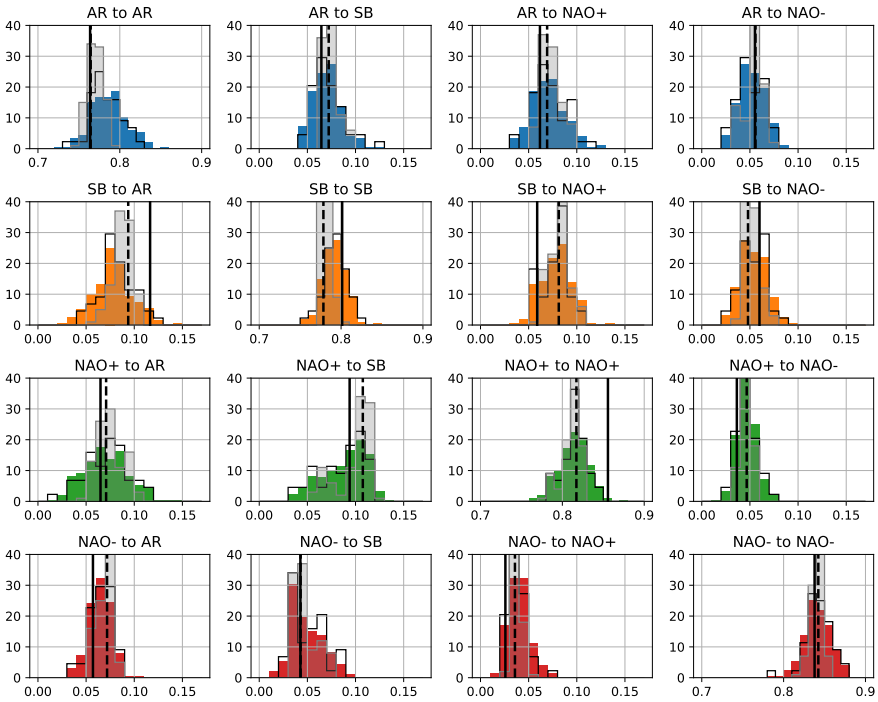


(a) Regime frequency for $k = 4$.

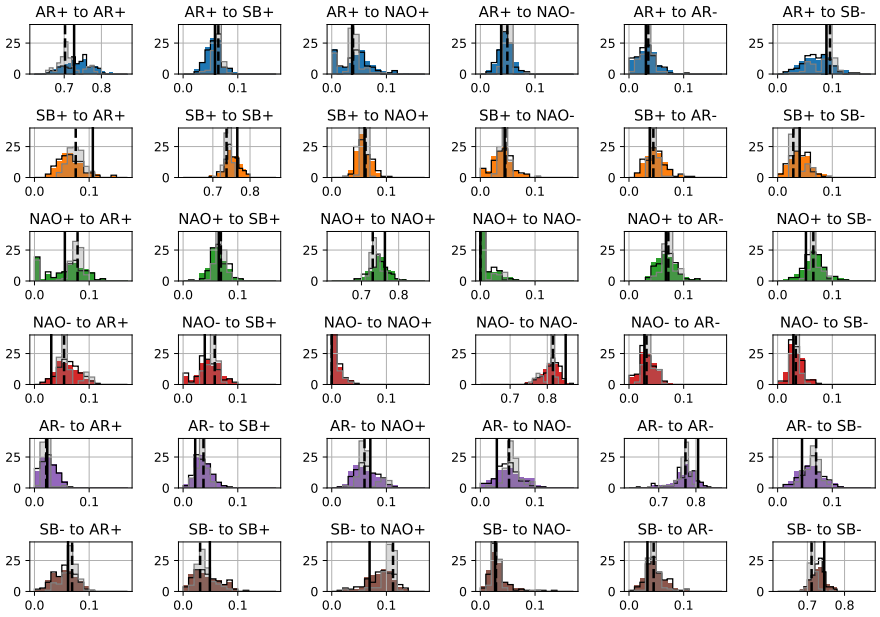


(b) Regime frequency for $k = 6$.

Figure B.5: The distribution of regime frequencies for the SEAS5 ensemble members (color), constrained on a one-to-one match with ERA-Interim and using 5 (black, blank) or 10 (grey, shaded) members a year. The vertical lines give the values for ERA-Interim (solid) and the full SEAS5 ensemble (dashed).



(a) Transition probabilities for $k = 4$.



(b) Transition probabilities for $k = 6$.

Figure B.6: The distribution of transition probabilities for the SEAS5 ensemble members (color), constraint on a one-to-one match with ERA-Interim (black line) and using 5 (black, blank) or 10 (grey, shaded) members a year. The vertical lines give the values for ERA-Interim (solid) and the SEAS5 ensemble (dashed).

ERA-interim regime dynamics is sufficiently represented within the SEAS5 model.

B.2.3 Persistence in SEAS5

Due to the large size of the SEAS5 ensemble it is not possible to use the time-regularised k-means clustering algorithm discussed in Section 3.3.1 to identify persistent regime dynamics in the full SEAS5 ensemble. However, it is possible to apply the time-regularised algorithm to individual ensemble members to get an idea of the persistent regime dynamics in the ensemble data. The first aspect to look at in this regard is what a suitable constraint value would be. Here we consider the constraint value in terms of the average regime duration (ARD) corresponding to a C for consistency between ERA-Interim and SEAS5 data, as the time series are of different length. The BIC is considered to identify a suitable value, as shown in Figure B.7 for $k = 4$ and $k = 6$. For ERA-Interim an average regime duration of approximately 7 days is found to be optimal for both four and six regimes. For $k = 4$ the majority of ensemble members identifies this to be the optimal regime duration as well, while for $k = 6$ the majority finds 6 days to be optimal. A shift towards shorter regime duration is expected as the use of more regimes allows for more variability and thus more regime transitions. This also corresponds to the results obtained in Section 3.3.

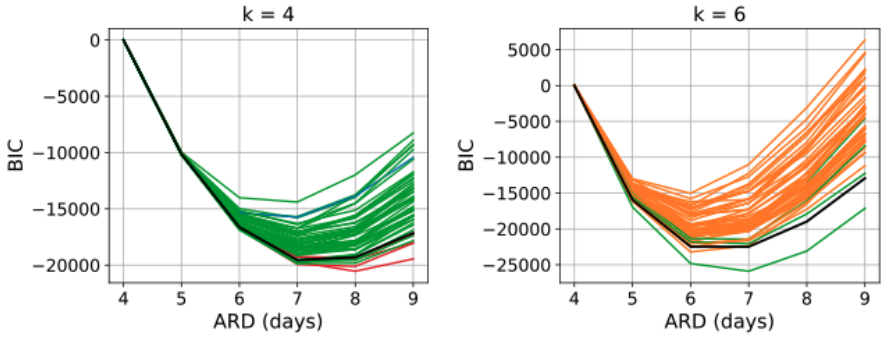
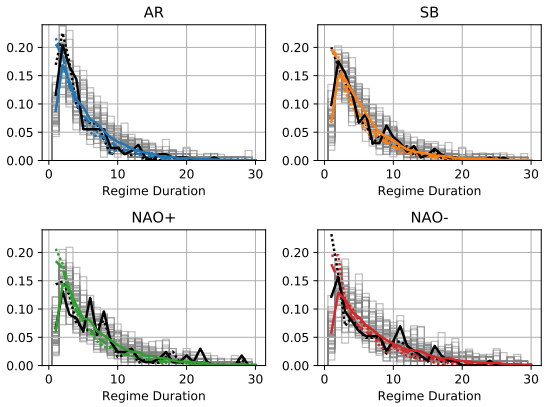
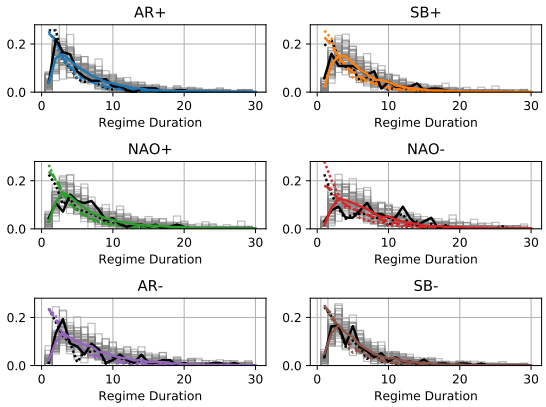


Figure B.7: The BIC for ERA-Interim (black) and single SEAS5 ensemble members where the colour indicates the location of the minimum; red for an average regime duration of 8 days, green for 7 days and orange for 6 days.

Since also for $k = 6$ the BIC for a 7 day average regime duration is close to its minimum, we decide to consider this constraint value for further analysis. The regime patterns, occurrence rates and transition probabilities are not strongly affected when the persistence constraint is used. Therefore, we focus on the effect on the distribution of the regime duration, which is shown in Figure B.8. The constraint reduces the number of one-day regimes that occurs and has a similar effect on



(a) Regime duration for $k = 4$.



(b) Regime duration for $k = 6$.

Figure B.8: The distributions of regime duration using a persistence constraint enforcing an average regime duration of 7 days. The grey bars give the histograms of the different ensemble members, the coloured solid lines give the ensemble with constraint and in black solid the result for ERA-I with constraint is shown. Furthermore the results for ERA-Interim (black dotted), full SEAS5 ensemble (colour dotted) and SEAS5 ensemble mean (colour, dashed) without constraint are shown.

the SEAS5 ensemble members as on ERA-Interim. As before, the ERA-Interim distributions fall within the ensemble spread. The noise within the ERA-Interim results is substantial, especially for a long regime duration, making it difficult to say anything more on the correspondence with SEAS5. Overall we have no reason to believe the persistent regime dynamics differs significantly between ERA-Interim and SEAS5.

Appendix C

Robustness of Distance Distributions

This last appendix relates to the robustness of the Bayesian regime assignment discussed in Chapter 5. The likelihood $P(\mathbf{d}|r)$, i.e. the likelihood of the data given a regime, is a key element in the Bayesian regime assignment approach. This likelihood is determined by fitting a multivariate normal to the conditional distance distributions for ERA-Interim and SEAS5. Thus, it is important to determine how robust those estimates are. We use a bootstrap analysis to obtain an error estimate of the mean and covariance of the fitted multivariate normal distributions. Here we use 20 years for ERA-Interim and single ensemble members for SEAS5 in the bootstrap analysis. Results for the mean are shown in Figure C.1 and for the covariances in Figures C.2, C.3, C.4, C.5, C.6, C.7.

Firstly note that the spread in the mean distance is small and that the ERA-Interim and SEAS5 values are comparable. The only exception to the latter is the distance to NAO+ and AR− conditional on being assigned to NAO+, where the ERA-Interim and SEAS5 values differ. This is related to the regime patterns being slightly different. Also for the covariances mostly the spread is small and comparable between the two datasets, indicating that also the spread around the mean is relatively robust. The largest spread in covariance values is found for data assigned to NAO−, indicating this is the least stable regime.

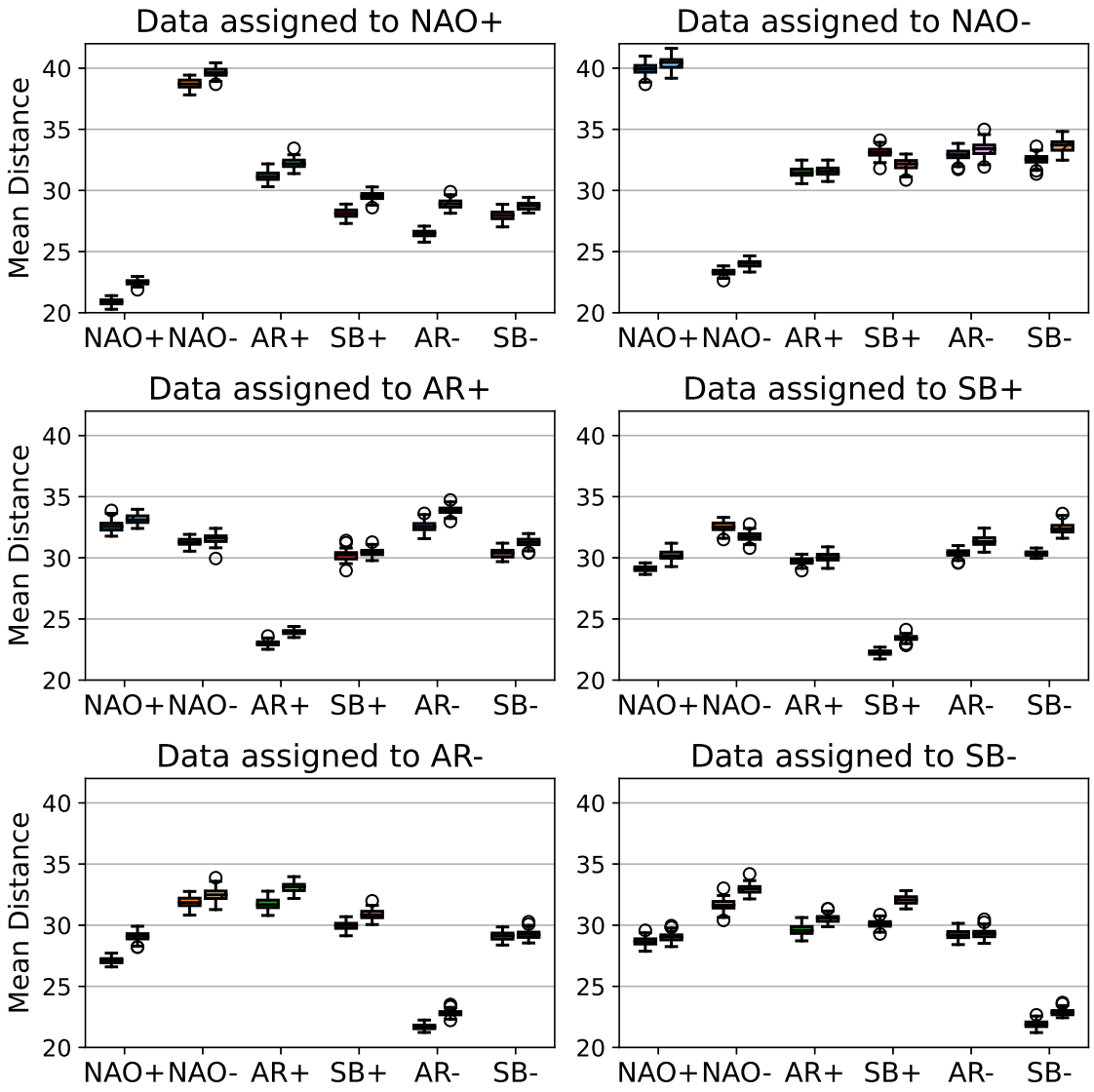


Figure C.1: The mean distance to each of the regimes conditional on the data being assigned to the regime in the caption. Results for ERA-Interim are shown left, for SEAS5 right.

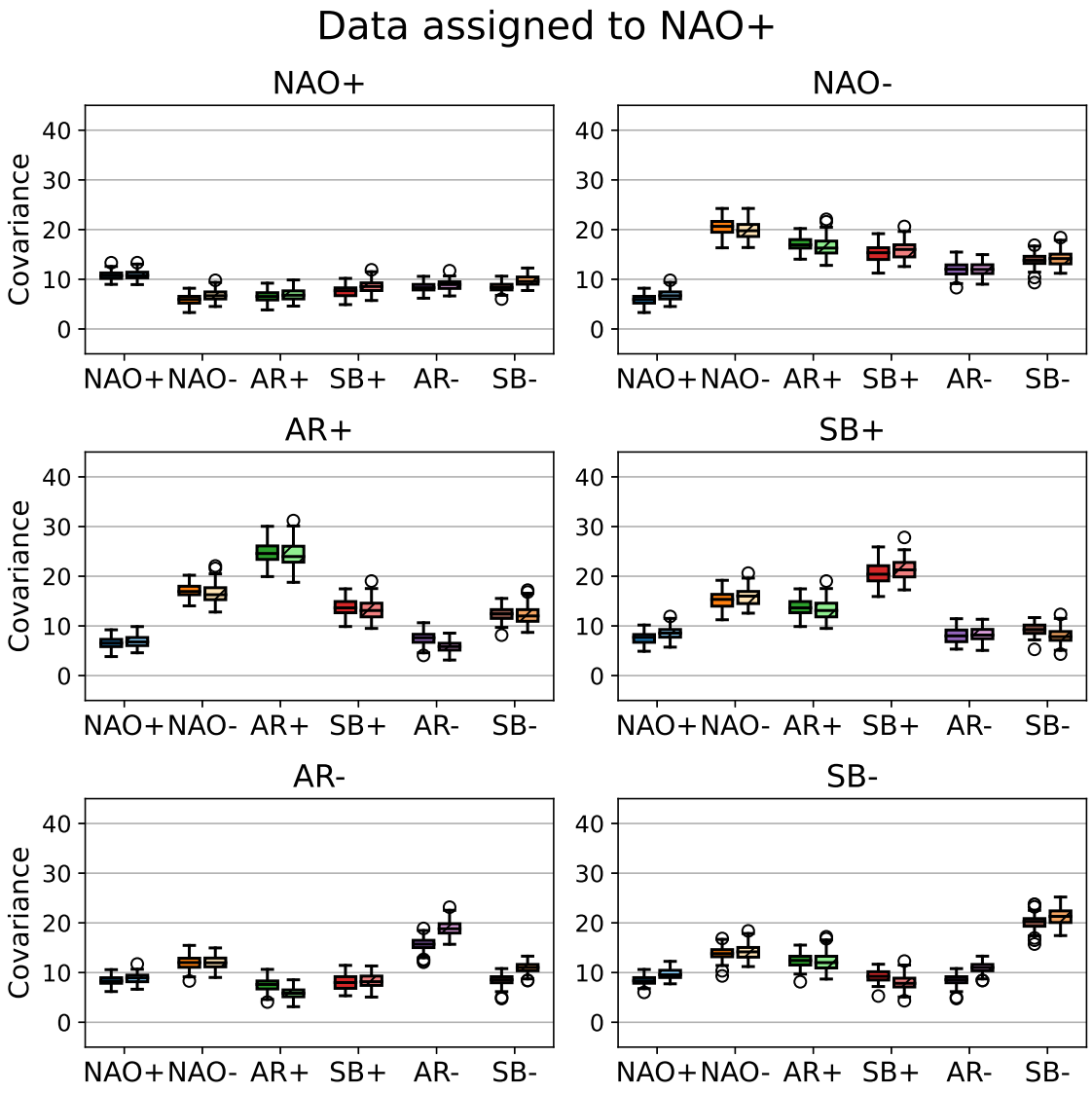


Figure C.2: The covariance of the distance conditional on the data being assigned to NAO+. Top left to bottom right shows the covariance matrix row by row. Results for ERA-Interim are shown left, for SEAS5 right.

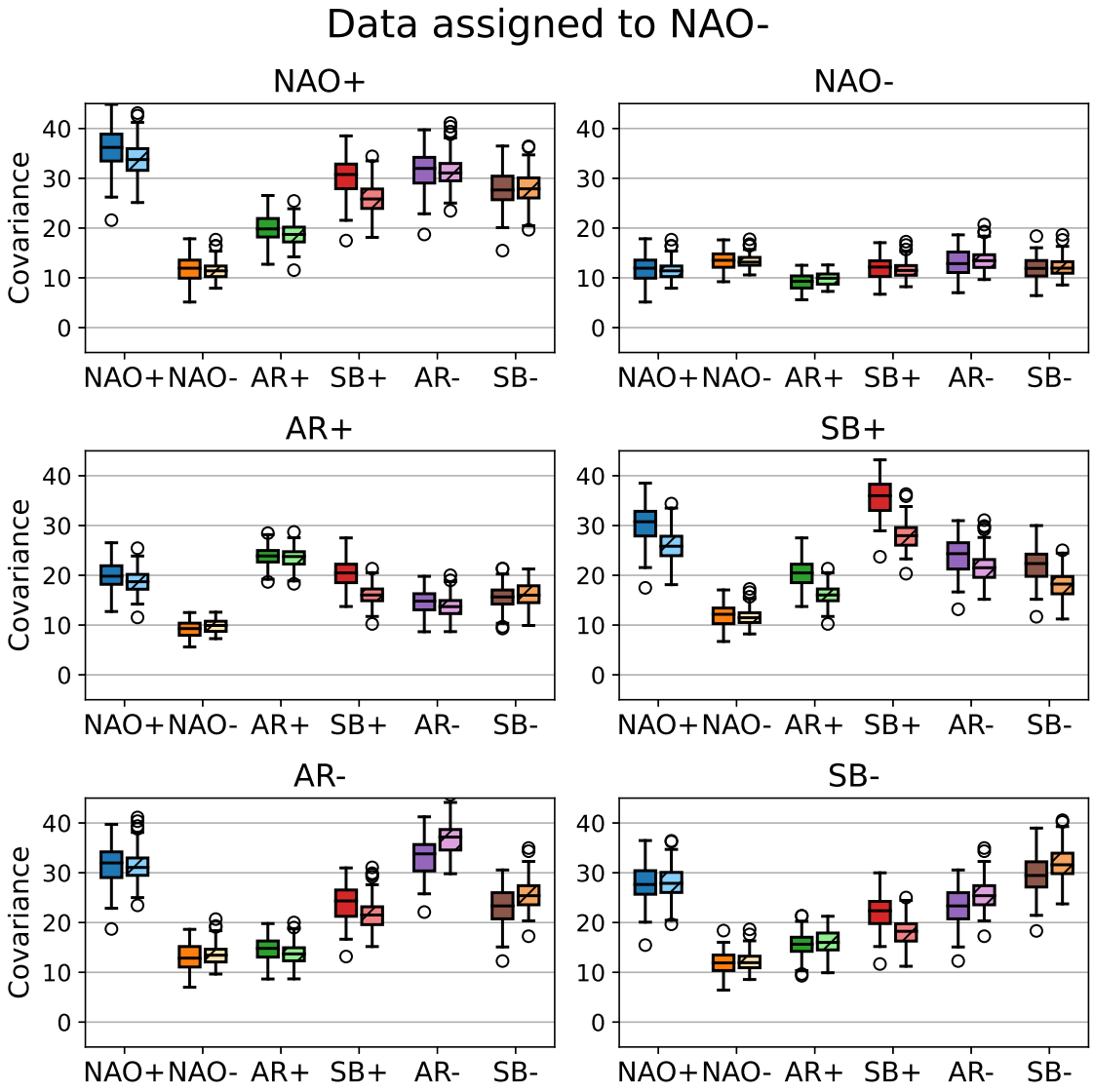


Figure C.3: As C.2 for data assigned to NAO−.

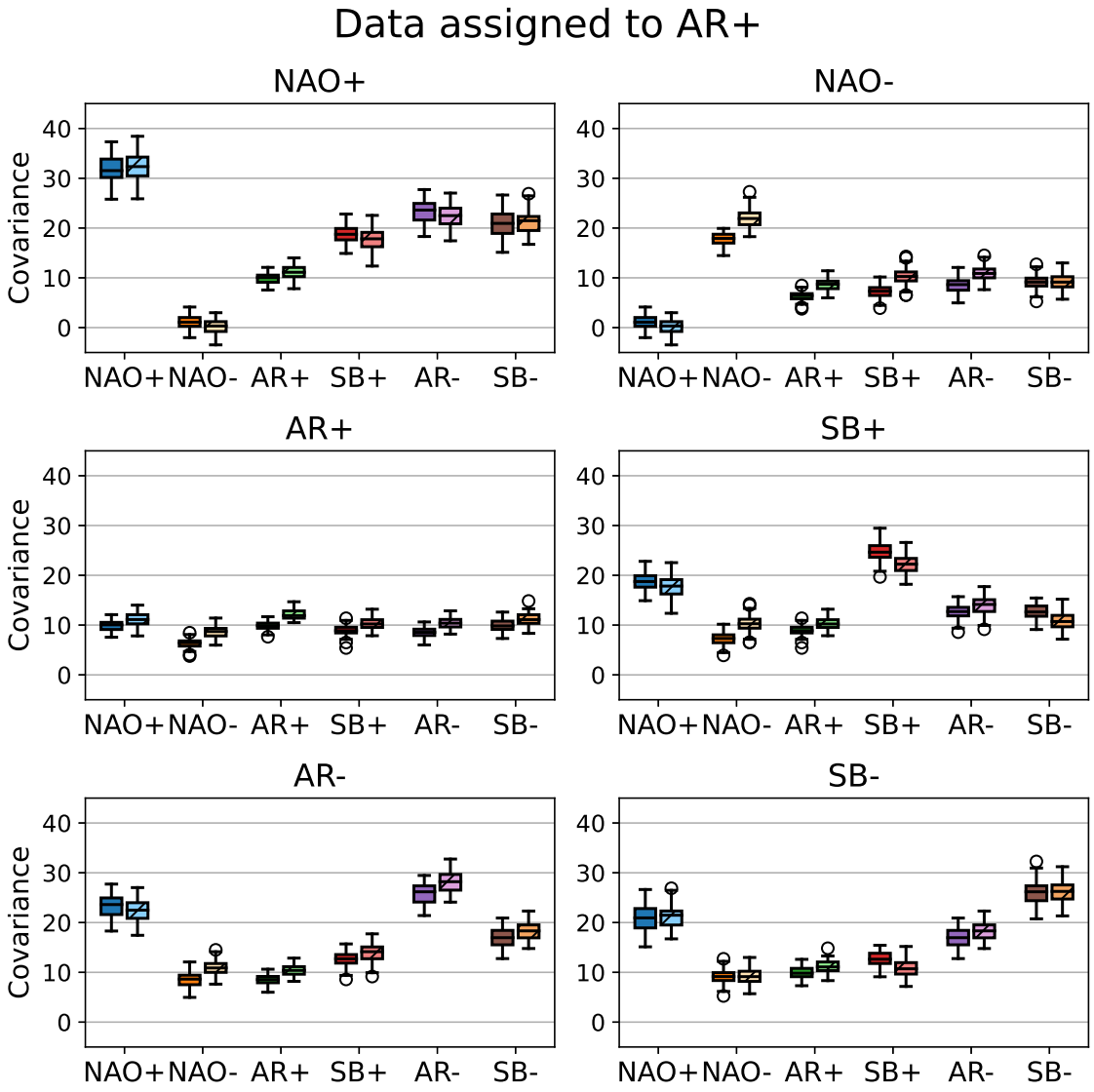


Figure C.4: As C.2 for data assigned to AR+.

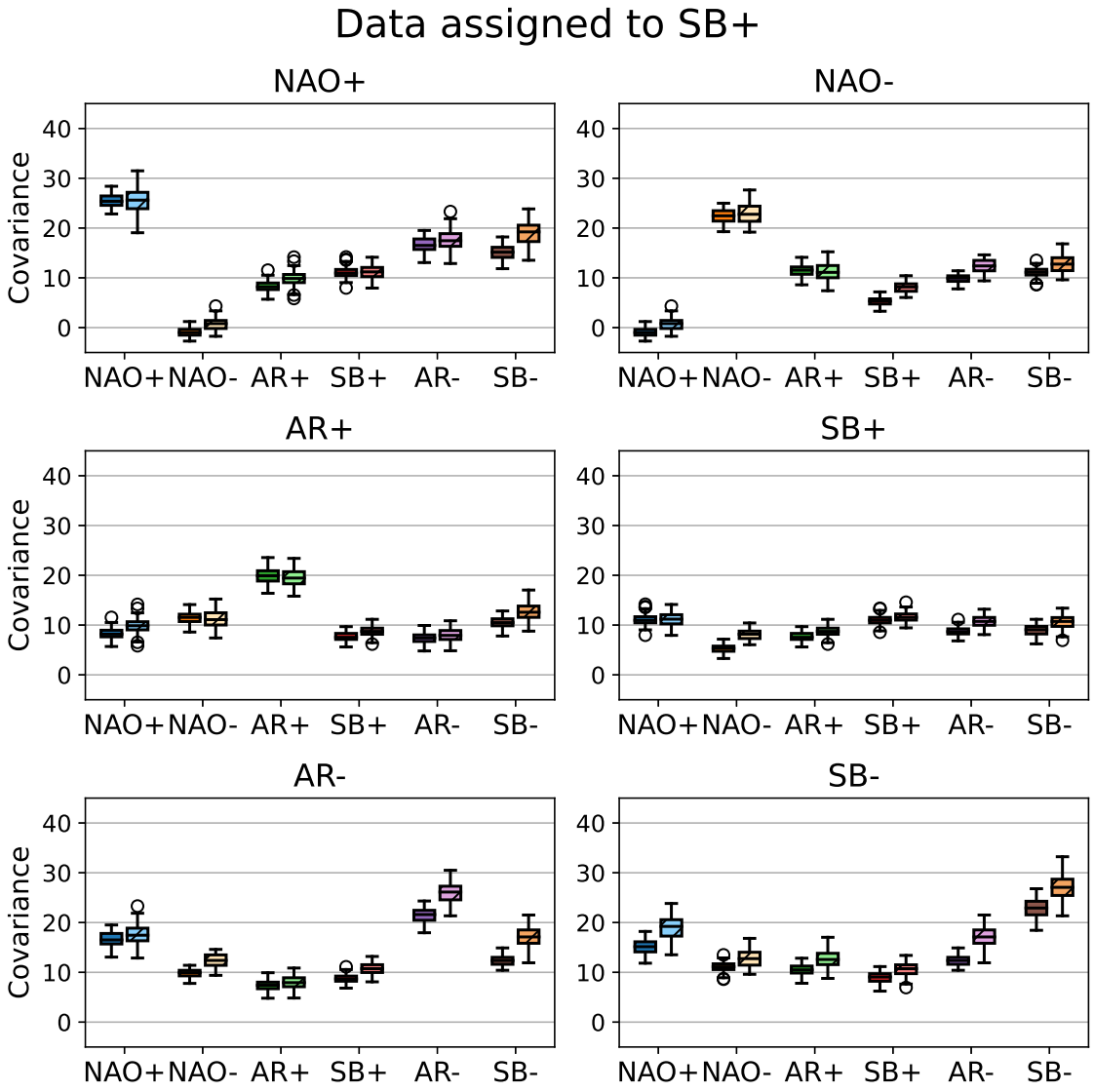


Figure C.5: As C.2 for data assigned to SB+.

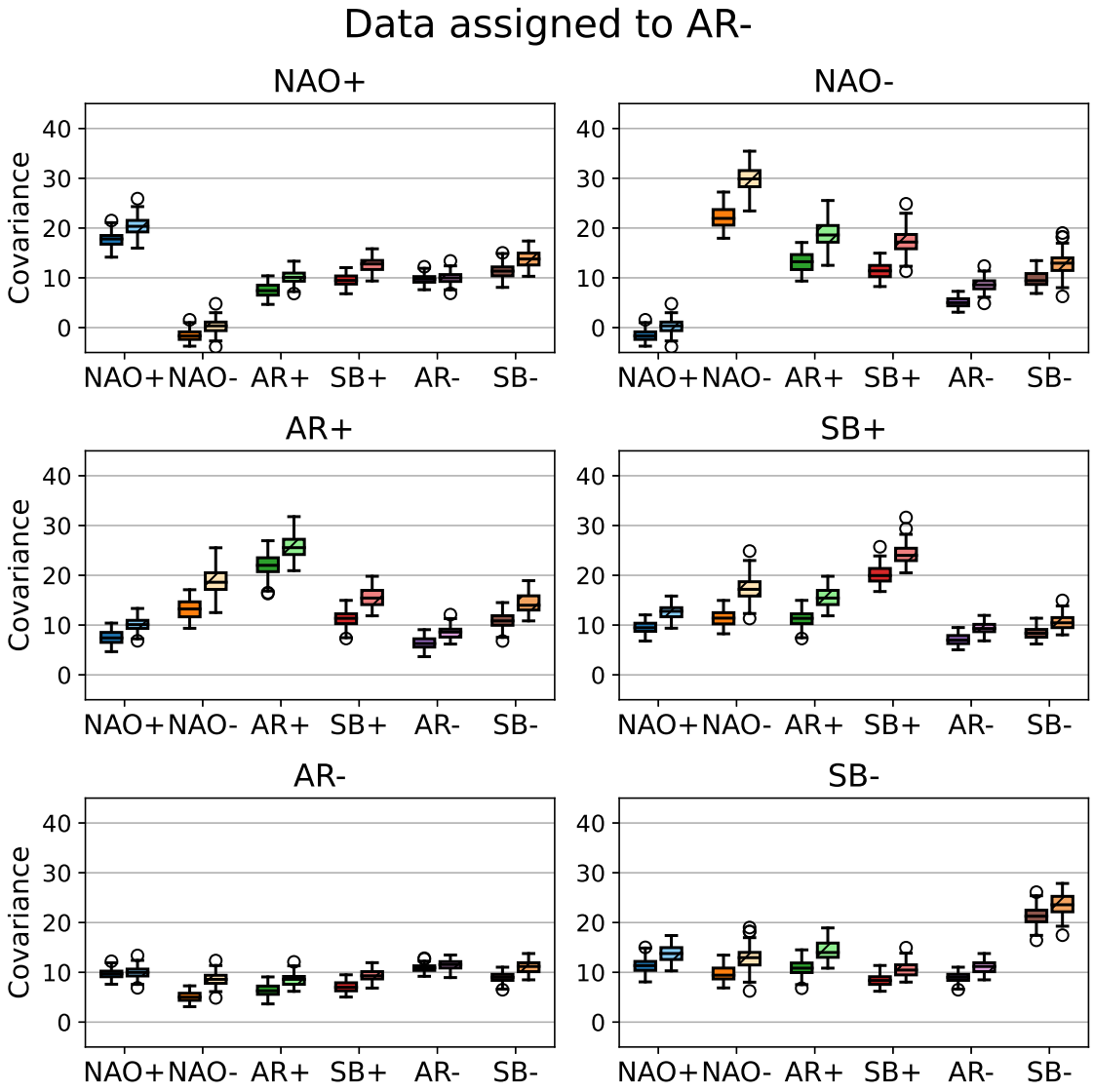


Figure C.6: As C.2 for data assigned to AR−.

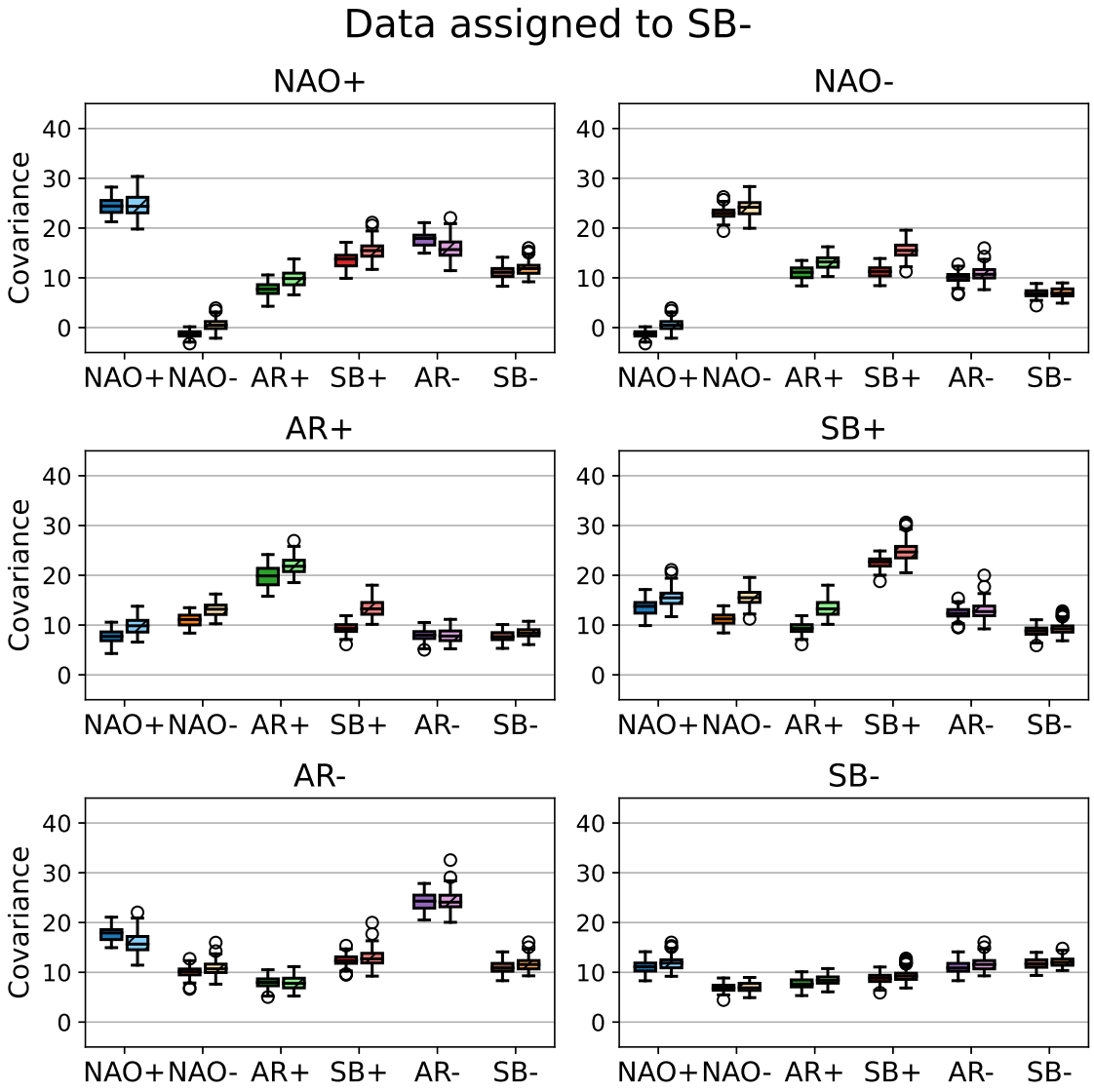


Figure C.7: As C.2 for data assigned to SB−.