Statistical learning for climate models

Nemo Malhomme

université PARIS-SACLAY

Statistical learning for climate models

Apprentissage statistique pour les modèles climatiques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 579, Sciences mécaniques et énergétiques, matériaux et géosciences (SMEMAG)
Spécialité de doctorat : Mécanique des fluides
Graduate School : Sciences de l'ingénierie et des systèmes. Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **Laboratoire Interdisciplinaire des Sciences du Numérique (Université Paris-Saclay, CNRS)** et **Laboratoire des Sciences du Climat et de l'Environnement (Université Paris-Saclay, CNRS, CEA, UVSQ)**, sous la direction de **Bérengère PODVIN**, directrice de recherche, et sous le co-encadrement de **Davide FARANDA**, directeur de recherche, et **Lionel MATHELIN**, chargé de recherche.

Thèse soutenue à Paris-Saclay, le 04 Décembre 2024, par

Nemo MALHOMME

Composition du jury

Membres du jury avec voix délibérative

François YVON Directeur de recherche, CNRS, ISIR, Sorbonne Université	Président
Marc BOCQUET Professeur, CEREA, École des ponts, IP Paris	Rapporteur & Examinateur
Corentin HERBERT Chargé de recherche HDR, CNRS, ENS de Lyon	Rapporteur & Examinateur
Mathieu VRAC Directeur de recherche, CNRS, LSCE, IPSL	Examinateur

THESE DE DOCTORAT
NNT : 2024UPAST165

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences mécaniques et énergétiques, matériaux et géosciences (SMEMAG)

Titre : Apprentissage statistique pour les modèles climatiques
Mots clés : Allocation de Dirichlet, Climat, Apprentissage automatique, Modèles synoptiques

Résumé :

Les modèles climatiques peinent à représenter précisément les structures de circulation atmosphérique liées aux événements extrêmes, et notamment leurs variations régionales. Cette thèse explore comment l'Allocation Latente de Dirichlet (LDA), une méthode d'apprentissage statistique issue du traitement du langage naturel, peut être utilisée pour évaluer la représentation par les modèles climatiques de données telles que la pression au niveau de la mer (SLP). La LDA identifie un jeu de structures locales (ou motifs) à l'échelle synoptique, interprétables physiquement comme des cyclones et des anticyclones. La même base de motifs peut servir à décrire les données issues des modèles et des réanalyses, permettant de représenter toute carte SLP par une combinaison parcimonieuse de ces motifs. Les coefficients, ou poids, de ces combinaisons fournissent une information locale sur la configuration synoptique de la circulation. Les analyser permet de caractériser la structure de la circulation dans les réanalyses et les modèles, et ainsi d'identifier localement des biais globaux ou spécifiques aux événements extrêmes. Une erreur dynamique globale peut être définie à partir des différences de poids des données modèles avec les réanalyses. Cette méthodologie a été appliquée à quatre modèles de CMIP6. Bien que les modèles représentent correctement en général la circulation à grande échelle, leurs erreurs sont plus élevées pour les vagues de froid et de chaleur. Une source d'erreur dans tous les modèles est liée aux motifs méditerranéens. Des critères d'évaluation supplémentaires ont été proposés. L'un s'appuie sur la fréquence d'apparition des motifs dans la représentation des cartes de pression. L'autre consiste à combiner l'erreur dynamique globale avec l'erreur de température, ce qui permet de différentier entre les modèles. Ces résultats démontrent le potentiel de la LDA pour l'évaluation et la préselection des modèles.

Title : Statistical learning for climate models
Keywords : Dirichlet allocation, Climate, Machine learning, Synoptic models

Abstract : Climate models face challenges in accurately representing atmospheric circulation patterns related to extreme weather events, especially regarding regional variability. This thesis explores how Latent Dirichlet Allocation (LDA), a statistical learning method originating from natural language processing, can be adapted to evaluate the ability of climate models to represent data such as Sea-Level Pressure (SLP). LDA identifies a set of local synoptic-scale structures, physically interpretable as cyclones and anticyclones, referred to as motifs. A common basis of motifs can be used to describe reanalysis and model data so that any SLP map can be represented as a sparse combination of these motifs. The motif weights provide local information on the synoptic configuration of circulation. By analyzing the weights, we can characterize circulation patterns in both reanalysis data and models, allowing us to identify local biases, both in general data and during extreme events. A global dynamic error can be defined for each model run based on the differences between the average weights of the run and reanalysis data. This methodology was applied to four CMIP6 models. While large-scale circulation is well predicted by all models on average, higher errors are found for heatwaves and cold spells. In general, a major source of error is found to be associated with Mediterranean motifs, for all models. Additional evaluation criteria were considered : one was based on the frequency of motifs in the sparse map representation. Another one involved combining the global dynamic error with the temperature error, thus making it possible to discriminate between models. These results show the potential of LDA for model evaluation and preselection.

Résumé en français

Le terme "météo" désigne les conditions atmosphériques à un moment et un lieu précis, caractérisées par des variables telles que l'humidité, la température, le vent, la pression ou la couverture nuageuse. Une bonne compréhension de la météo offre de nombreuses applications pratiques. Cela permet notamment des prévisions précises, utiles au quotidien, mais c'est également un outil essentiel pour des secteurs tels que l'agriculture. Le climat, quant à lui, décrit les structures météorologiques et leurs arrangements dans le temps sur des échelles temporelles plus élevées. Comprendre le climat permet de prévoir la quantité de pluie qui tombe sur un an, les températures attendues pour la prochaine saison ou l'intensité potentielle des tempêtes. De plus, la connaissance du climat est essentielle pour projeter les changements climatiques futurs en fonction des comportements sociétaux. Ces projections permettent la lutte contre le changement climatique, soit par l'atténuation (en modifiant les activités humaines pour prévenir les changements les plus graves), soit par l'adaptation (en ajustant les infrastructures et les pratiques pour minimiser l'impact des changements inévitables). Une meilleure compréhension du climat représente donc un défi théorique fondamental avec un impact direct sur la vie humaine.

L'état de l'art actuel de la connaissance sur le climat et la météo permet la construction de modèles, des simulateurs qui répliquent le système Terre. Bien que ces modèles soient imparfaits et maintiennent un certain degré d'incertitude, ils sont capables de prouesses tant en météorologie qu'en climatologie. Par exemple, la météo peut être prédite plusieurs jours à l'avance. Les climats de nombreux scénarios passés, présents ou futurs peuvent être explorés. Et, en particulier, les modèles permettent de projeter l'évolution du réchauffement climatique, et de prédire les fréquences et intensités attendues des évènements extrêmes, tels que les vagues de chaleur, de froid, ou encore les ouragans. Ces évènement sont d'une importance capitale en raison de leurs impacts : ils peuvent causer de nombreux problèmes de santé, des morts, des baisses de rendement agricole ainsi que des dommages aux infrastructures. En prédisant le passé et le futur proche, les modèles offrent un cadre d'expérimentation pour tester et améliorer notre connaissance des processus impliqués. En retour, cette compréhension du climat permet de développer des modèles plus fidèles à la réalité, créant ainsi un cercle vertueux.

Malgré ces progrès, cependant, de nombreux aspects du comportement physique de l'atmosphère ne sont pas compris, ou seulement partiellement. Entre autres, c'est notamment le cas des statistiques des évènements extrêmes et

de leur variabilité d'une région du monde à une autre. En particulier, pour les évènements d'extrême chaud ou froid, ces difficultés demeurent même dans les régions bien documentées telles que l'Europe. Les modèles ne sont pas encore capables de reproduire ces aspects de la réalité. Une cause probable de cette lacune est la représentation incomplète de la circulation atmosphérique dans les modèles du climat. La circulation atmosphérique est le moyen principal par lequel les variables météorologiques, telles que la température et l'humidité, sont transportées spatialement. Par conséquent, la structure de ces flux de circulation, appelée la configuration synoptique, joue un rôle important dans l'élaboration des conditions météorologiques, à la fois au quotidien et pour les cas extrêmes. Cependant, la circulation atmosphérique est un système complexe, composé de cyclones et d'anticyclones transportant de l'air dont la température est influencée par divers processus thermodynamiques. Identifier les sources exactes des erreurs de représentation est une tâche difficile. Les modèles climatiques sont à la fois le produit de la compréhension actuelle du système Terre, et un outil pour évaluer cette compréhension. Il est crucial de continuer à développer de nouvelles façons de tester la représentation de la circulation atmosphérique dans les modèles, afin de pouvoir quantifier leurs incertitudes, identifier leurs défauts, et s'assurer qu'ils continuent de s'améliorer.

L'objectif de ce travail est de **construire un protocole d'évaluation de la représentation de la circulation atmosphérique par les modèles.** En particulier, il s'agit de développer une méthode capable **d'estimer localement les erreurs de représentation dans la configuration synoptique, avec une interprétation physique claire et directe.** Dans ce but, sera fait l'usage d'une méthode d'apprentissage statistique originaire du domaine du traitement du langage naturel : l'**Allocation Latente de Dirichlet** (en anglais Latent Dirichlet Allocation, LDA).

Les récents progrès en puissance de calcul, ainsi que les développements dans l'instrumentation météorologique, ont considérablement augmenté le volume de données disponibles pour la science du climat. En conséquence, il y a un besoin croissant de méthodes permettant d'interpréter efficacement ces données et d'en extraire des informations pertinentes. Pour cette raison, des techniques d'apprentissage automatique sont d'ores et déjà appliquées dans divers sous-domaines des sciences climatiques, notamment l'étude de la circulation atmosphérique et l'évaluation des performances des modèles climatiques. En général, les outils d'apprentissage automatique sont conçus pour des tâches bien spécifiques, ce qui limite leur domaine d'application. Cependant, des questions fondamentales dans différents domaines scientifiques peuvent partager des caractéristiques communes, suggérant que des solu-

tions issues d'un domaine peuvent être adaptées à un autre. Dans notre cas, des éléments communs ont été identifiés entre le traitement du langage naturel et la classification de la circulation atmosphérique. Les deux visent à modéliser la structure sous-jacente des données en identifiant un petit nombre de motifs latents permettant de décrire les configurations observées. Ce travail de thèse explore comment la LDA peut être utilisée pour extraire de l'information de données climatiques. Ces informations seront ensuite utilisées pour évaluer la représentation de la circulation atmosphérique dans les modèles, dans le but ultime d'identifier et de caractériser d'éventuels défauts dans leurs représentations.

Le chapitre 1 de ce manuscrit introduit les concepts et outils de climatologie et de météorologie sur lesquels ce travail s'appuie. Y sont décris les mécanismes physiques à l'origine de la circulation atmosphérique, l'état des connaissances sur le changement climatique et ses effets, ainsi que les variables les plus adaptées à décrire les phénomènes étudiés. La notion d'évènement extrême y est discutée, et en particulier les vagues de chaleur et de froid et les différentes façons de les définir. Ce chapitre contient également une introduction générale à la modélisation du climat, et en particulier aux General Circulation Models (GCM), qui sont parmi les modèle les plus avancés disponible, et aux méthodes existantes pour leur évaluation. Dans le cadre de ce travail, quatre GCM seront comparés à des données de référence, les réanalyses.

Le chapitre 2 fait le tour des méthodes existantes permettant la classification des configurations synoptiques. Ce chapitre introduit notamment la Principal Component Analysis (PCA), une méthode de réduction dimensionnelle communément utilisée dans le pré-traitement des données, avant la classification. Plusieurs méthodes de classification automatique sont présentées. Ces méthodes sont catégoriques, c'est-à-dire que chaque configuration synoptique est associée à une et une seule catégorie. Par la suite, sont présentées des méthodes de classification non-catégoriques, dont les modèles statistiques (dont fait partie la LDA).

Dans le chapitre 3, la méthode LDA est introduite dans le contexte de son domaine d'application originel, le traitement du langage naturel. Plus spécifiquement, il s'agit du domaine du "topic modelling", qui cherche à représenter des documents textuels par les sujets dont ils traitent. Différents modèles statistiques permettant de représenter la composition en sujets d'un texte sont présentés. Le chapitre détaille les avantages de la LDA par rapport à d'autres modèles, le fonctionnement de l'apprentissage automatique des sujets dans un corpus de textes, ainsi que comment cette procédure est implémentée en pratique.

Les résultats de ce travail de thèse sont présentés dans les chapitres 4, 5 et 6. Tout d'abord, le chapitre 4 présente la méthodologie d'application de la LDA aux jeux de données climatiques, ainsi que les résultats de l'application à la pression au niveau de la mer. Dans ce cas, les sujets appris automatiquement par la LDA prennent la forme de motifs spatiaux. Ces motifs sont analogues à des cyclones et des anticyclones, et permettent de décrire les configurations synoptiques avec parcimonie et interprétabilité physique. La pertinence des motifs ainsi obtenus est testée par une application à des données d'origines différentes, qui donnent des jeux de motifs similaires. Enfin, la décomposition en motif proposée par la LDA est comparée à la décomposition donnée par la PCA, afin de déterminer les avantages et les inconvénients de chacune.

Dans le chapitre 5, l'outil LDA est utilisé pour caractériser la circulation atmosphérique dans les modèles et les réanalyses. La caractérisation des statistiques des motifs sert de base pour l'évaluation des modèles. Ce procédé est utilisé pour évaluer les configurations synoptiques produites par les modèles, en général ainsi que dans les cas de vagues de chaleur et de froid localisées dans divers pays d'Europe. Les erreurs de représentation des modèles sont quantifiées de manière locale. On observe que les modèles représentent satisfaisamment la circulation générale, bien qu'ils sur-représentent les cyclones et anticyclones méditerranéens. Dans les cas extrêmes, les modèles montrent de nombreux biais d'amplitude supérieure.

Le chapitre 6 présente des critères d'évaluation additionnels pour l'évaluation des modèles. Premièrement, les propriétés statistiques des motifs sont évaluées en plus grand détail afin de définir des notions de fréquence d'apparition et d'intensité moyenne. Deuxièmement, les erreurs mesurées antérieurement sur les configurations synoptiques sont comparées avec une erreur sur la représentation par les modèles de la température, dans le but de déterminer leur origine. Ces deux systèmes sont utilisés pour extraire des informations supplémentaires des données, pour évaluer et comparer les modèles en plus grand détail.

Remerciements

Avant toute autre chose, je voudrais commencer par remercier tous ceux sans qui la réalisation de ce travail de thèse n'aurait pu être possible.

Je souhaite remercier les membres de mon jury, pour le travail conséquent que représente l'évaluation d'une thèse. Merci à François Yvon en tant qu'examinateur et président du jury, à Marc Bocquet et Corentin Herbert en tant que rapporteurs et examinateurs, et à Mathieu Vrac en tant qu'examinateur. Je remercie également Mathieu Vrac et Alexandre Allauzen en tant que membres de mon comité annuel de suivi de thèse.

Je tiens également à présenter ma gratitude envers chaque membre de ma direction de thèse : Bérengère Podvin, Davide Faranda, et Lionel Mathelin. D'abord, pour m'avoir accepté comme candidat pour ce fascinant projet de recherche, et ensuite pour leur soutient continu et leur conseil au cours de ces trois années. Je suis heureux d'avoir eu la chance de travailler avec vous.

Merci également à tous mes collègues, pour l'environnement de travail chaleureux et encourageant, et pour les interactions productives et enrichissantes que nous avons pu avoir. En particulier, merci à mes collègues du LSCE, au sein de l'équipe ESTIMR et plus généralement du Thème 3, en tant que permanents, post-doc, doctorants ou stagiaires : Romeo Hatchi, Pascal Yiou, Soulivanh Thao, Jean-Yves Peterschmitt, Pradeebane Vaittinada Ayar, Meriem Krouma, Stella Bourdin, Lucas Fery, Mireia Ginesta-Fernandez, Paula Gonzalez, Robin Noyelle, et Camille Cadiou.

Enfin, je voudrais remercier mes proches, en particulier mes parents et mes deux adelphes, ainsi que mes amis, Cédric, Louis, Charlotte, Marie, Louis, ainsi Nathan, Blaise, et Marvin pour leur soutient et pour les moments heureux passés en leur compagnie.

Contents

Acronyms and notations

Acronyms and abbreviations

CMIP6	-	6th Coupled Model Intercomparison Projet
EOF	-	Empirical Orthogonal Function
FCM	-	Fuzzy C-Means
GCM	-	General Circulation Model
IPCC	-	Intergovernmental Panel on Climate Change
KL	-	Kullback-Leibler
LDA	-	Latent Dirichlet Allocation
ML	-	Machine Learning
PC	-	Principal Component
PCA	-	Principal Component Analysis
pLSI	-	Probabilistic Latent Semantic Indexing
POD	-	Proper Orthogonal Decomposition
SLP	-	Sea-Level Pressure
SOM	-	Self-Organizing Maps
UK	-	United Kingdom
z500	-	Geopotential height at 500 hPa
z700	-	Geopotential height at 700 hPa

Notations for matrix and vectors

Matrices and vectors are represented in bold. Elements of a matrix or vector are represented by specifying the positions of the element as index on the matrix name. The letters chosen for indexes are generally the lowercase of the letter indicating the length of the indexed dimension. Example: Let $\boldsymbol{A}$ be a $I \times J$ matrix. Its elements are written $\boldsymbol{A}_{i,j}$, $\forall i \in \{1, ..., I\}$, $\forall j \in \{1, ..., J\}$.

Introduction

Weather refers to the atmospheric conditions at a specific time and place, characterized by parameters such as humidity, temperature, wind, pressure, and cloud cover. Though often taken for granted, weather plays a crucial role in both society and everyday life. A good understanding of weather has many practical applications, including accurate forecasting, which is of everyday use for event planning, but also a critical tool for sectors such as agriculture.

Climate, on the other hand, describes long-term patterns of weather across extended time scales. Understanding climate allows for predictions of how much rain falls in a year, what temperatures can be expected for the next season, or the potential intensity of storms. In addition, knowledge of climate is essential for projecting future climatic changes, based on societal behaviors. Such projections are key to addressing climate change, either through mitigation — modifying human activities to prevent the most severe changes — or adaptation — adjusting infrastructure and practices to minimize the impact of unavoidable changes. A better understanding of climate therefore represents a fundamental theoretical challenge with a direct global impact for human life.

After more than a century of steady progress, the current state of knowledge on climate has reached quite an impressive level. The dynamics of weather variables are well understood enough that we can construct models — simulators designed to replicate the Earth system. While models remain imperfect and are associated with a degree of uncertainty, they are able to achieve feats in both meteorology and climatology. Weather can be forecast several days in advance. The behavior of climate in various scenarios past, present or future can be explored. In particular, models allow us to project the trajectories of global warming, and predict the frequencies and intensities to expect of extreme weather events, such as heatwaves, cold spells, or hurricanes. Extreme weather events are of particular importance because of their impacts ; they can cause a variety of health issues, deaths, losses of agricultural yields, as well as damages to infrastructures. By predicting the past and the close future, models enable reliable experimentation to test our understanding, improving our knowledge of climate processes. In turn, this understanding of climate allows for the development of models that more faithfully reproduce the Earth system, creating a virtuous circle of knowledge.

However, despite significant progresses, gaps remain in our understanding of climate. Many aspects of the physical behavior of the atmosphere are not, or only partially, understood. Key areas of uncertainty include the effects

of clouds on global warming, the influence of turbulence in air flows, and a limited understanding of extreme weather event statistics, and their regional variability. In particular, for extreme heat and cold events, these difficulties remain even in well-documented regions such as Europe. Models are not yet able to reproduce these aspects of reality.

The probable cause of this limitation is the incomplete representation of atmospheric circulation in climate models. Atmospheric circulation is the primary process by which weather variables, such as temperature and humidity, are transported across different regions. It dictates where and how these variables move. Therefore, the patterns of atmospheric circulation play a significant role in shaping both everyday weather, and extreme weather events. However, atmospheric circulation is a complex system of patterns and structures, composed of cyclones and anticyclones transporting air whose temperature is determined by thermodynamic processes. Identifying the exact sources of representation error is a difficult task.

Climate models are both the product of our current understanding of the Earth system, and a tool for evaluating that understanding. It is crucial to continuously test the ability of models to represent atmospheric circulation patterns and evaluate their performances in new ways. Doing so allows us to quantify uncertainties, to identify model flaws, and to ensure that they can continue to improve.

The goal of this work is to **construct a protocol to evaluate model representation of atmospheric circulation.** Specifically, we aim to develop a method capable of **estimating local errors in the representation of circulation patterns, with a clear and direct physical interpretation.**

To achieve this, we use a statistical learning method originating from the field of natural language processing: **Latent Dirichlet Allocation (LDA)**.

Recent advances in computational power, as well as developments in meteorological instrumentation, have significantly increased the volume of data available to climate science. As a result, there is a growing need for methods to efficiently interpret this data and extract relevant information. Machine learning (ML), a field of research that focuses on solving this type of problem, has therefore become increasingly valuable to climate science. Machine learning techniques are now applied across various sub-fields of climate sciences, including the study of atmospheric circulation, and the evaluation of climate model performance.

Generally speaking, ML tools are tailored to well-identified tasks. As such, their popularity only propagates within the limits of their scientific discipline of application. However, fundamental questions in different scientific fields may share common features, suggesting that solutions from one domain can be adapted to another. In our case, we have identified common elements between natural language processing and atmospheric circulation classification. Both aim to model the underlying structure of data, and do so by identifying a small number of latent patterns to describe datapoints with. This thesis work explores how LDA can be used to extract information from climate data. This information will then be used to evaluate model representation of atmospheric circulation, with the ultimate goal of identifying and characterizing any potential flaws in their representations.

In chapter 1 of this manuscript, we introduce the concepts and tools of climatology and meteorology on which our work is based. We focus in particular on atmospheric circulation, extreme weather events, as well as climate models, their performances and their limitations. In chapter 2, we provide an overview of the existing methods tackling the issue of classifying patterns of atmospheric circulation. Then, in chapter 3, we introduce our method of choice, Latent Dirichlet Allocation, detailing its origins, the way it functions, as well as the rationale for choosing it. Chapter 4 details our methodology of application of LDA, including how it was converted from studying corpora of text documents to our set of climate data. This chapter also presents our preliminary results. In chapter 5, the LDA tool is used to characterize atmospheric circulation in reanalysis data, and use it as a basis of evaluation of the models. Model representation of the circulation is evaluated, both in general and for various European extreme weather events. Chapter 6 relates additional evaluation criteria for the evaluation of models, that allowed us to extract additional information from the data. Then, we provide a conclusion, as well as perspectives detailing how this work might be pursued in the future.

1 - Climate and climate extremes

1.1 . Introduction

Meteorology is the study of weather. It considers the atmosphere, specifically on the lower layers, where weather takes place, and studies the variables connected to weather, such as temperature, precipitations, pressure, humidity, or wind speed and direction. Meteorology works with a time scale of a few days. On the other hand, according to the IPCC glossary (IPCC, 2024), climatology is the study of climate, defined as:

> "average weather," or more rigorously, the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands of years.

While considering the same variables as meteorology, climatology works with statistical descriptions rather than values. It considers the atmosphere, and the other systems with which it interacts, on a timescale of decades. This thesis work focuses on climatology. However, discussing statistical description of weather variables also requires knowledge on the processes driving those variables.

This chapter presents the key concepts and tools of meteorology and climatology relevant to this thesis. We start by introducing basic mechanisms of weather and the main climate variables used in their study. Our work focuses specifically on atmospheric circulation, which describes the large-scale movements of air in the atmosphere. We then introduce the notion of extreme weather events, the challenges associated with defining them, and how they are expected to evolve with climate change. Finally, we introduce climate models, and their importance to climate sciences. We present a state of the art of model performances, recent progresses, and evaluation methods, as well as our choices regarding the data we use in our work.

1.2 . Climate study and variables

1.2.1 . Basic weather mechanisms

To facilitate later discussions, we present here basic meteorological processes (Wallace and Hobbs, 2006). The sun is the main driver of weather-related phenomena. A part of solar radiation is absorbed by the surface of the Earth, causing it to heat up. Because of its temperature, the Earth emits infrared radiation. This radiation is in large part absorbed and diffused by gases and

clouds in the atmosphere, heating up the surface and the air close to it. This is referred to as the greenhouse effect. These processes are summarized in figure 1.1.

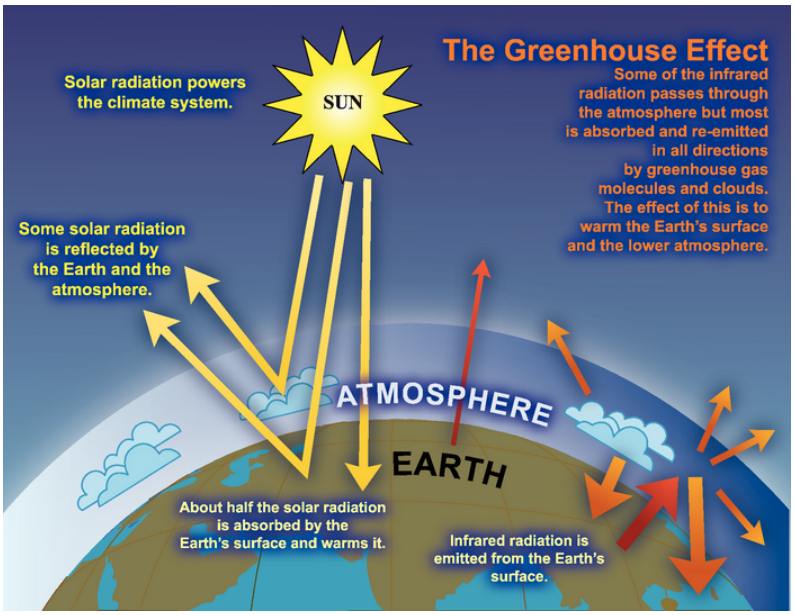


Figure 1.1 – A simplified model of radiative energy transfers in the Earth system. This figure is from (Solomon et al., 2007).

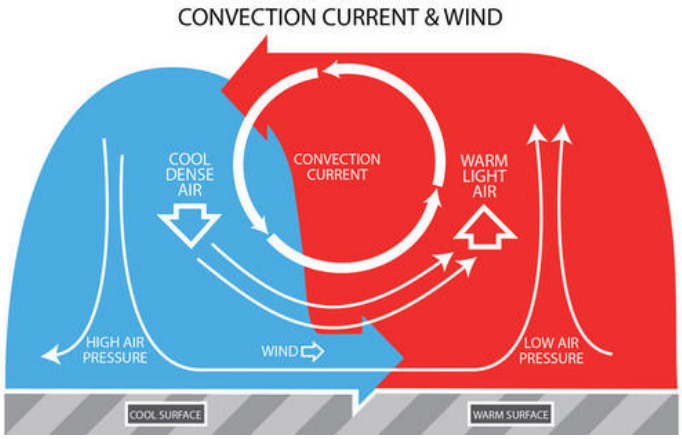


Figure 1.2 – Schematic representation of the circulation of air induced by surface temperature gradients. Cold air is represented in blue, and warm air in red. This figure is from (Foundation and Christopher AuYeung, 2015).

Hot air is able to contain more humidity, and raises, as it is lighter than cold

air. This causes regions of lower pressure to exist, while cold air descending causes regions of higher pressure, as shown in 1.2. Wind flows from high pressures to low pressures, but due to the rotation of the Earth, movement twists to the right in the northern hemisphere, and to the left in the southern, through a process known as baroclinic instability (Grotjahn, 2003). This is why cyclones, which are region of low pressure, induce inward spiralling movement (counterclockwise in the northern hemisphere), and anticyclones, which are region of high pressure, outward spiralling movement (clockwise in the northern hemisphere). Earth rotation also causes high-altitude, almost permanent currents of wind that goes from west to east, that we call jet stream, or jet (Lamb, 1948; Rossby and Willett, 1948). When hot air reaches higher altitudes, it cools down. No longer able to contain as much humidity, the water condenses and forms clouds. The processes associated with cyclones and anticyclones are summarized in figure 1.3.

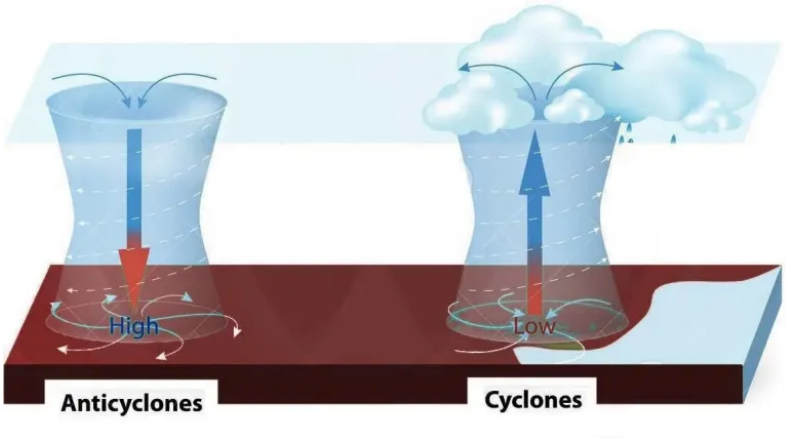


Figure 1.3 – Schematic representation of the movements of air close to the surface and in the upper atmosphere associated with anticyclones (left) and cyclones (right) in the northern hemisphere. Air spiral outwards, clockwise near an anticyclone, and spirals inwards, counterclockwise near a cyclone. Air warms as it gets closer to the surface, and grows colder as it raises in altitude. Colder air liberates moisture, which can lead to clouds and precipitations. This figure is from (Kumbhar, 2023).

In this work, we focus on atmospheric circulation, i.e. on the structure of the movement of air in the atmosphere. On the global scale, phenomena describe in the previous section induce a 3-cell atmospheric circulation structure, as shown in figure 1.4. The cells are separated by belts of alternatively high and low pressure.

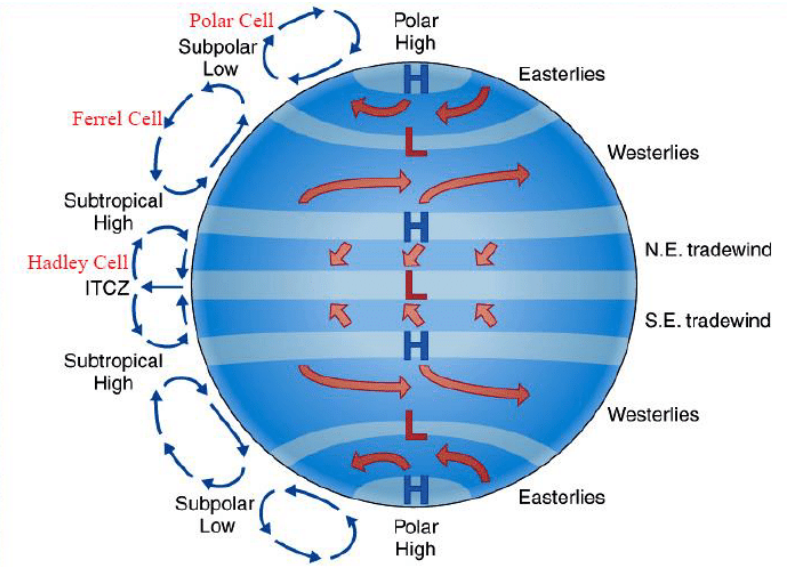


Figure 1.4 – Global structure of atmospheric circulation. This figure is from (Aguado and Burt, 2014).

The dynamics of atmospheric circulation are described by the intensity and localization of high and low pressures. This means that pressure can be a good descriptor of atmospheric circulation. Cyclones and anticyclones typically exist at the scale of ≈ 1000 km, which is called the synoptic scale. These objects are therefore referred to as synoptic objects. Since synoptic objects characterize the structure of air movement, i.e. atmospheric circulation, it is possible to study and predict changes in weather based on them (Bluestein, 1992). This is called synoptic meteorology (Bergeron, 1980). These structure have a typical persistence time of ≈ 5 to 10 days (Bluestein, 1992). However, the jet is sometimes subject to undulations called Rossby waves, or planetary waves (Madden, 1979; Rossby, 1949), capable of forming big meanders. Those meanders can be stable for longer periods of time (Krishnamurti, 1961), and lock synoptic objects in place. This leads to what is known as blockings, where a cyclone or anticyclone can remain for extended periods of times over the same region (Faranda et al., 2016; Lupo, 2021). This can cause extreme weather events, i.e. weather events of high intensity, able to cause damages to human infrastructure and health (see section 1.3).

1.2.2 . Variables of study

Atmospheric circulation can be studied through synoptic meteorology using pressure-related variables, which best describe synoptic objects. While circulation and pressure are 3-dimensional fields, for easier visualization and

processing, we will only consider horizontal 2-dimensional maps. Horizontal pressure maps contain the information on horizontal circulation, but also some information on vertical circulation: a high pressure means air is entering the current height level, and a low that it is exiting it. The most natural variable choice is Sea-Level Pressure (SLP). It describes circulation at the bottom of the atmosphere, where the weather is relevant to human activities.

However, if we want to study circulation patterns at higher levels, we need another variable. Geopotential is defined as the gravitational potential energy of one unit of mass. When divided by the gravitational pull of the Earth ($g \approx 9.8$ m/s^2), assumed constant everywhere, it becomes a vertical coordinate known as geopotential height. When studying circulation patterns, it is common to look at geopoptential height along a surface of fixed pressure. The National Weather Service of the United States of America (US Department of Commerce, 2024) defines geopotential height as:

> The height above sea level of a pressure level. For example, if a station reports that the 500 mb height at its location is 5600 m, it means that the level of the atmosphere over that station at which the atmospheric pressure is 500 mb is 5600 meters above sea level. This is an estimated height based on temperature and pressure data.

A common value of pressure to examine geopotential height at is 500 hPa, shortened to z500. 500 hPa corresponds to an altitude of about 5.5 km, which corresponds roughly to the middle of the atmosphere (Wallace and Hobbs, 2006). While still considering circulation relevant to weather, this altitude allows for less influence from the topography of the surface.

Most climate variables follow an annual cycle of evolution. When considering variations irrespective of their location in the annual cycle, it can be desirable to eliminate this cycle from the data. To that end, are often computed what is referred to as anomalies. Anomalies are obtained by removing from each data point the average of all maps corresponding to the same day of the year. For more detail, see chapter 4.

Examples of z500 anomalies graphs are shown figure 1.5. The lines represent isohypses (lines of equal geopotential height). This can be used to locate pressure highs and lows. Furthermore, dense lines indicate regions of high geopotential height gradient, which corresponds to high pressure gradient, and therefore to strong winds.

1.2.3 . Climate change

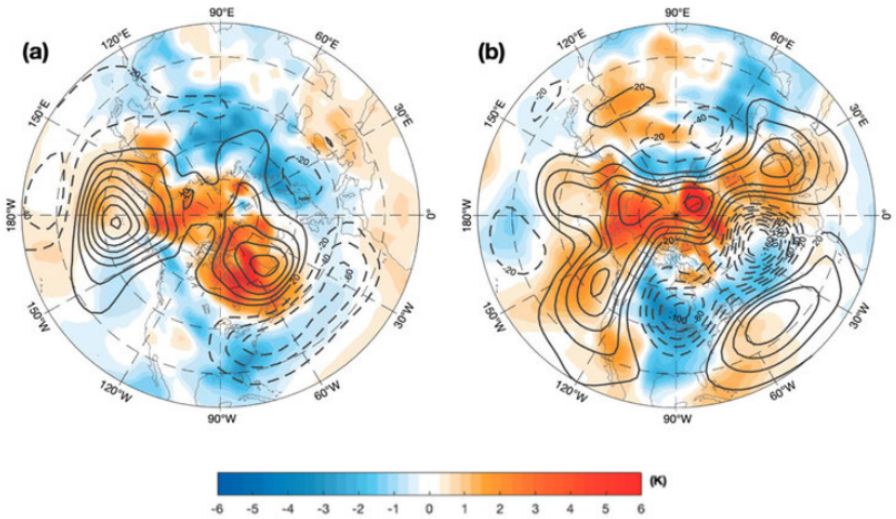


Figure 1.5 – Example of depiction of z500 (contours) and temperature (color), representing average anomalies (a) during the 2010-2011 winter (b) during the 2013-2014 winter. z500 is represented through contour lines of equal geopotential height, with a contour interval of 20 m. Dashed lines represent negative anomalies. This figure is from (Luo et al., 2020).

The term "climate change" can refer to any change in climate over time (IPCC, 2024). However, it is known that the climate has undergone, and is still undergoing, changes since the pre-industrial period, and it is know that these changes are due to human influence (anthropogenic). The pre-industrial period is the period of time that precedes human influence due to industrialization. While lacking a general consensus on its definition, references for pre-industrial climate are generally taken from some point between 1750 and 1850 (Carslaw et al., 2017). The Intergovernmental Panel on Climate Change (IPCC, (programme, 2024)) was created by the United Nations to assess the science on climate change, and especially anthropogenic climate change. Its work consists in reviewing what is currently known on the subject, and compiling it into reports. The following discussion is based on the report of the most recent iteration to date, Assessment Report 6 (IPCC, 2023e).

A main consequence of anthropogenic climate change is global warming. At time of writing, the current world average surface temperature increased by between 1.3°C and 1.6°C since pre-industrial times (Ipcc, 2022). However, temperature does not increase at the same rate everywhere. In particular, temperature has increased more inland than at sea, at the poles than in the tropics. See figure 1.6, from (Intergovernmental Panel on Climate Change (IPCC),

2023), for observed and simulated local temperature increases for various world average temperature increases.

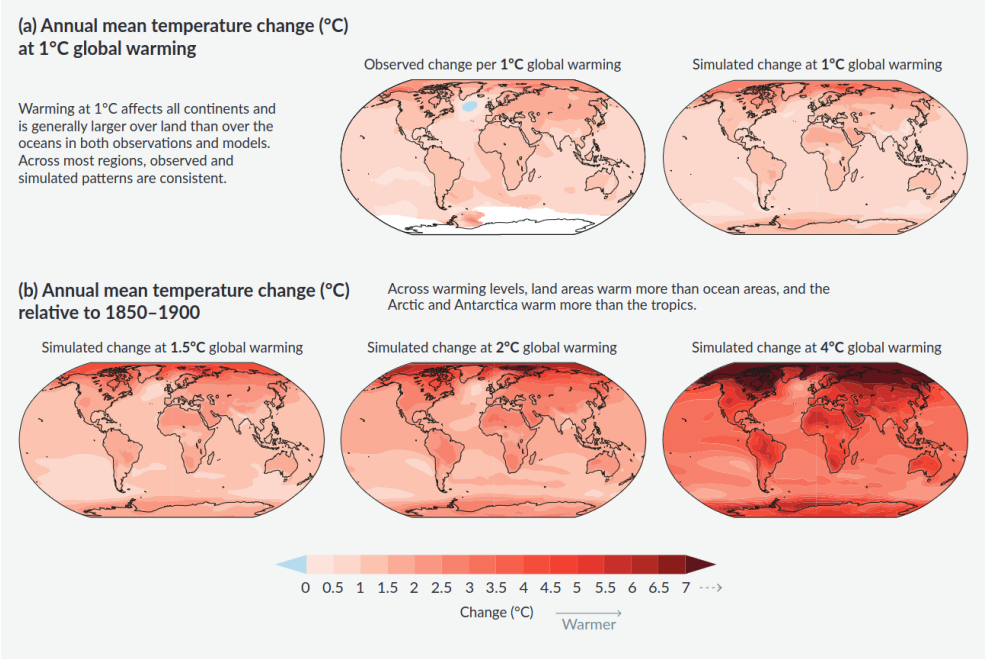


Figure 1.6 – Observed and simulated local changes in annual mean surface temperature. In (a) is shown a comparison of observed and simulated temperature changes with global warming levels as observed between 1850 and 2020. In (b) are shown simulated changes in temperature for higher levels of global warming. This figure is from the IPCC summary for policymakers of the 2021 report on the physical science basis for climate change (Intergovernmental Panel on Climate Change (IPCC), 2023).

Variations in the temperature gradients across hemispheres cause changes in the atmospheric circulation those gradients generate, which makes it difficult to predict changes in other weather variables, such as precipitations. Climate change-induced variations in temperatures can therefore be due to two distinct but interconnected drivers. *Dynamic* effects are changes in the patterns of atmospheric circulations. For example, the increased average temperature caused a higher frequency of circulation patterns that bring in hotter air from equatorial regions, is a change driven by dynamic factors. *Thermodynamic* effects are changes in air temperature and mechanisms of temperature exchange with atmosphere. For example, the increased average temperature due to enhanced greenhouse effect, irrespective of circulation patterns, is a change driven by thermodynamic factors.

1.3 . Extreme weather events

1.3.1 . Motivation

In its report, the IPCC defines extreme weather events (sometimes just called extreme events) as "an event that is rare at a particular place and time of year" (IPCC, 2023b). The definition of extreme events can depend on the locality and the time of year, but is based on the concept of rarity. The rarity of extreme events is due to low probability in the relevant statistical distributions. Often enough, extreme events are improbable because they require climate variables to be far enough from average, during a long enough period. This is notably the case for heatwaves and cold spells, which are usually defined as a period of at least several days where the temperature is significantly higher than (respectively, below) average. Despite their rarity, extreme weather events have significant impacts on human society. They can cause severe damages to human health (Weilnhammer et al., 2021) and infrastructure (Añel et al., 2017), motivating their study.

Heatwaves, in particular, are already responsible for significant societal impacts. We can cite, as examples, the European heatwave of 2003 (Fink et al., 2004; Stott et al., 2004), or that of 2018 (McCarthy et al., 2019), both responsible for tens of thousands of deaths. Cold spells also represent a significant hazard (López-Bueno et al., 2021). For instance, we can cite the cold spell of 2017 over the Balkans, which had consequent socio-economic impacts (Anagnostopoulou et al., 2017). In particular, a cold spell occurring in spring, during the budding period of plants, can break off the developing cycle of plants and heavily reduce agricultural yields (Papagiannaki et al., 2014). One such example is the cold spell of April 2021 described in (Vautard et al., 2023). Climate change will not prevent such occurrences. Furthermore, their properties, including intensity and frequency, are projected to change with climate (IPCC, 2023b). Specifically, according to the IPCC report, Heatwaves were observed to have become more frequent. These trends are projected to continue in the future (IPCC, 2023b).

Extreme weather events are difficult to study owning to their rarity: data concerning them is scarse. The limited amount of observed events leads to incomplete knowledge about their precursors and statistical properties. Furthermore, this incomplete understanding propagates to their representation by models. This leads to difficulties in forecasting, and in simulation.

1.3.2 . Definition of extreme events

There is no general consensus about the definitions of extreme weather events (McPhillips et al., 2018). Here, we list the three main approaches used to de-

fine them.

Impacts-based definition: The first approach is to define an extreme weather event by its impacts. In this case, an event is categorized as such if it meets certain criteria regarding numbers of dead or injured, economical losses, or amount of affected people. One commonly used open access database recording events using this method is the Emergency Events Database (EM-DAT) (Delforge et al., 2023). Following EM-DAT methodology, an event is recorded if it meets at least one of the following criteria: 10 or more deaths, 100 or more affected people, a call for international assistance or a declaration of a state of emergency occurred. In addition to these metrics, economic impact is also recorded by EM-DAT. Costs associated with extreme events are estimated in reinsurance reports, providing generally accessible data. However, estimating the economic impacts of extreme events is a complex problem. Among other reasons, this is because of the problem of estimating intangible, long-term effects on society and the time and efforts required for recovery. For a detailed review of the methodologies used to the estimation of economic impacts of extreme events, see (Doktycz and Abkowitz, 2019). A problem with this approach is that it makes extreme event detection dependent on whether the affected location is populated, and if so, on whether it has adaptation procedures in place to reduce impacts.

Physics-based definition: A second approach is to use physical thresholds to discriminate extreme events. Extreme wind gusts may be defined using a threshold in wind speed. Extreme temperature events may be defined using thresholds in temperature and duration. However, the choice of the variables to apply the threshold to is not obvious and can have an impact on the results. In the case of heatwaves, for example, temperature thresholds can be applied to daily maximum temperature, daily minimum temperature, daily average temperature, or more complex indexes taking more information into account (Perkins and Alexander, 2013), such as Excess Heat Factor (EHF) (Nairn et al., 2009). Similarly, thresholds can be chosen to be relative to the data (such as a given percentile), or absolute. And the choice of a threshold necessarily involves some arbitrariness. Because of these reasons, definitions can vary significantly on a study-to-study basis (Grotjahn et al., 2016). Furthermore, for more structurally complex extremes, such as tropical cyclones, automating the detection in climate datasets poses its own difficulties. There is no consensus on detection algorithms to use, and the choice of methods influences the resulting data (Bourdin et al., 2022).

Machine Learning-based definition: The third approach to extreme event definition is to use machine learning (or deep learning) for automatic de-

tection and classification in climate datasets (McGovern et al., 2017). These methodologies require an input of extreme events labelled by humans to train the model on. This is particularly common for more structurally complex extreme events such as cyclones, due to the difficulty of the task of devising an algorithm for automatic detection. The techniques involved are methodologies of image classification, as is done in (Liu et al., 2016), or image segmentation, as is done in (Kumler-Bonfanti et al., 2020).

In this thesis work, when studying extreme weather events, we focus specifically on heatwaves and cold spells. We chose the physical approach, so as to ensure our definition of extreme event does not depend on population vulnerability. Our definition of choice is the following:

- A **Heatwave** over a specific region designates any period of at least three consecutive days where the spatial average of daily mean temperature over the region is beyond the 97th percentile.
- A **Cold Spell** over a specific region designates any period of at least three consecutive days where the spatial average of daily mean temperature over the region is below the 3rd percentile.

The regions considered will be European countries. Daily mean temperature was chosen for its widespread availability in many datasets. The threshold is chosen to be relative so that the definition is adaptable to different regions. Arbitrariness is unavoidable in the choice of the threshold value. The value we picked was chosen to ensure a number of extreme events per year similar to estimations from organisms of weather study (in particular, Météo-France). (See chapter 5.)

1.3.3 . Time evolution

Cold spells are not projected to increase in frequency nor intensity with climate change (IPCC, 2023b). Heatwaves, however, are projected to increase in both frequency and intensity (Alexander et al., 2006; Frich et al., 2002; IPCC, 2023b). See figure 1.7, from (Intergovernmental Panel on Climate Change (IPCC), 2023), for a summary of the currently observed trends in heatwaves, and the knowledge regarding human contribution. An increase in heatwaves can be due to either thermodynamic drivers or dynamic drivers, or a mix of the two. Thermodynamic drivers can make heatwaves more likely or intense, by way of a general increase in temperature, and increased ability of circulation patterns to cause intense heat (Chan et al., 2022). Dynamic drivers, on the other hand, can make heatwaves more likely or intense through changes in the statistics of atmospheric circulation patterns associated with heatwaves. In particular, the article (Rousi et al., 2022) demonstrates a link between increasing heatwaves trends over Europe, and an increasing tendency for the

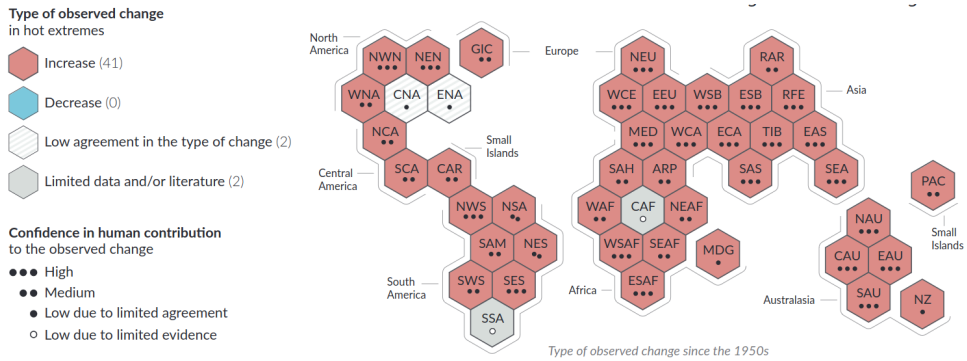


Figure 1.7 – Observed regional changes in hot extremes between pre-industrial and current periods, and level of confidence in the observed changes. This figure is from the IPCC summary for policymakers of the 2021 report on the physical science basis for climate change (Intergovernmental Panel on Climate Change (IPCC), 2023).

jet to split in two, trapping an anticyclone over Europe, causing a blocking event.

Extreme events are generally described in terms of "return time". The return time of an event is the average time it would take for an event of similar or higher intensity to happen again, assuming a fixed statistical distribution. It is inversely proportional to the probability of the event. As an example, a heatwave with a $\frac{1}{50}$ chance to occur any given year (based on current climate statistics) currently has a return time of 50 years. The probabilistic representations of extreme events are based on Extreme Value Theory (Coles, 2001; Gumbel, 1942, 2019; Wilks, 2011), which will not be detailed here. The return times of events can change when the climate changes, making them less or more likely.

In general, the return time of an extreme event grows with its intensity, which means that heatwaves of higher temperatures are less likely to happen. When it is said that heatwaves increase in both frequency and intensity, it is meant that the statistical distribution is shifting towards the extremes. Events of fixed intensity have their return time decrease. In parallel, events of fixed return times become associated with higher intensities. These two effects come together, as consequences of the same changes in weather statistics. See figure 1.8, from (Intergovernmental Panel on Climate Change (IPCC), 2023), for a summary of the effects of climate change on heatwaves, seen either as changes in frequency of events at fixed intensity, or changes in intensity of events at fixed frequency.

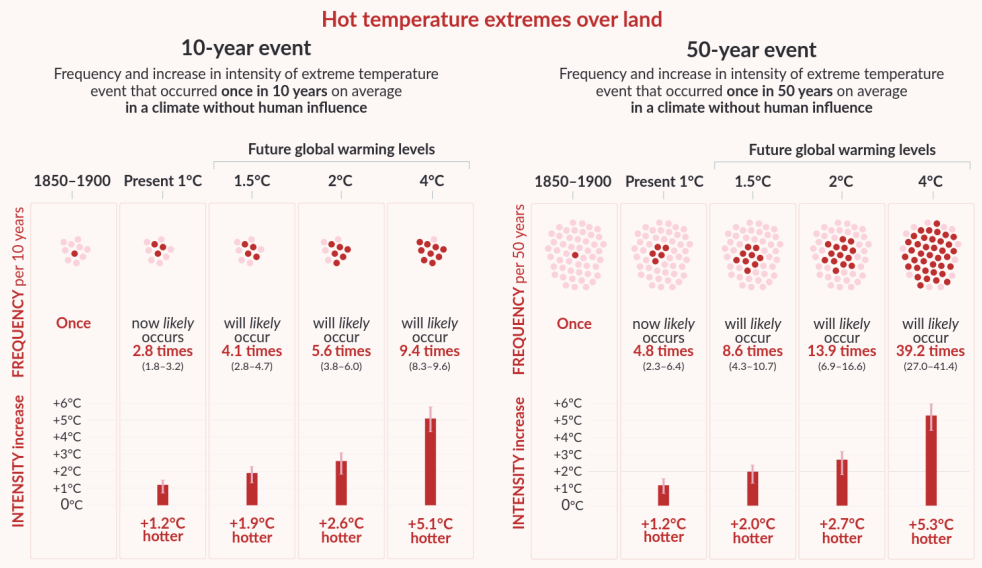


Figure 1.8 – Projected changes in the intensity and frequency of 10-year and 50-year return times hot temperature extremes over land, for different levels of global warming. Values shown are relative to the 1850-1900 period, representing a climate without human influence. This figure is from the IPCC summary for policymakers of the 2021 report on the physical science basis for climate change (Intergovernmental Panel on Climate Change (IPCC), 2023).

1.4 . Climate models

1.4.1 . Introduction to climate modelling

Data used in climate science comes either from observations, or models. Observations refers to data measured from the Earth system using instruments. When studying weather, this mainly refers to weather stations and satellites. There are weather observation stations in many places in the world, equipped with instruments such as thermometers, barometers, hygrometers or anemometers. They providing regular measurements of climate observables such as, respectively, temperature, pressure, humidity, or wind speed and direction. Satellites provide information on many weather-related variables, including cloud cover as well as temperature (Foelsche et al., 2008). Observation datasets are compiled from weather station observations (using between 10,000 and 40,000 stations, depending on the dataset), satellite data, and older records (Rohde and Hausfather, 2020).

On the other hand, climate models are simulators, build to generate climate data similar to reality. Given an initial situation and external forcings (such as solar radiation, and greenhouse gas emissions), they compute the future of the Earth system based on a set of assumptions that vary among models. Because climate is chaotic, it is common practice to run the models several times with slight random differences in initial conditions, in order to explore the space of possible outcomes. These are called model runs.

There are several types of climate models, with varying levels of complexity (Edwards, 2011; Saunois, 2022). On the simpler side, Energy Balance Models compute the radiative balance of the Earth and its atmosphere, without taking any spatial effect into consideration (North et al., 1981). On the opposite end, General Circulation Models (GCM) solve the fluid equations in the atmosphere and ocean on 3D grids. These models are the highest performing and are used in many projects that aim at producing realistic data. GCM are comprised of several sub-models each taking care of one component of the Earth system. This includes the atmosphere, the ocean, sea ice, but also, in some models, vegetation, land use, or ocean biochemistry. These elements are connected through the use of a coupler. As an example, the structure of the ACCESS-ESM1.5 model (Ziehn et al., 2020) is shown in figure 1.9.

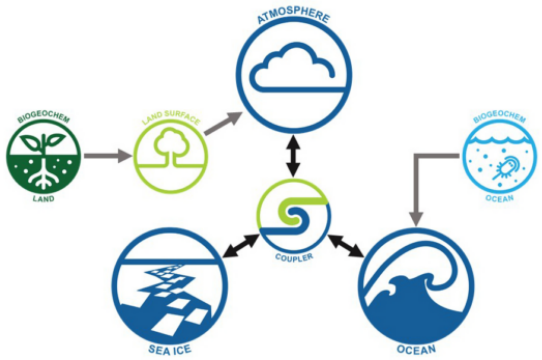


Figure 1.9 – Example of the structure of a General Circulation Model, with several modules dealing with different parts of the Earth system, linked together by a coupler. This example is of the ACCESS-ESM1.5 model, taken from (Infrastructure, 2024).

Due to domain size, which is the entire Earth, GCM are limited in their grid resolutions by computation times and data volume. In practice, GCM with finer grid can only reach cell size of one degree of latitude and longitude, or

a fraction of a degree, which still represents hundreds of kilometers. Therefore, there are several phenomenons that GCM are incapable of resolving, including, for example, turbulence, and cloud microphysics. Because of this, approximations have to be formulated to include the influence of sub-cell scale phenomena. These are called parametrizations. Several parametrization schemes exist, and there is no general consensus on which are the best (Loehle, 2018). Some regional circulation models can afford to have grid cells on the scale of ≈ 1 km. Such models can resolve some smaller-scale phenomena such as convection rain (Lucas-Picher et al., 2021), but still require parametrizations for even smaller-scale phenomena, such as cloud formation, or turbulence.

Models play a key role in climate science, due to the limited amount of data that can feasibly be measured, and the fact that there is only a single Earth system trajectory available. They fulfill several roles:

- *The completion of missing observation data.* Observations are spatially sparse, due to only being available locally, at weather stations. Models can be used to interpolate this data using physical knowledge, resulting in complete datasets known as "reanalysis" (Dee et al., 2014). Reanalyses are the closest complete datasets to reality available.
- *Generation of additional data.* By applying to the models forcings similar to that of the present climate, but different (often randomized) initial conditions, models can be used to generate alternative trajectories for the Earth system. This augments the amount of data available, allowing for statistical analyses that would be impossible otherwise.
- *Test of our understanding of the processes involved in climate.* Models have many parameters and parametrizations. By studying the similarity between generated data and observations, we can asses how well the underlying assumptions of models represent reality.
- *Prediction and forecast of the future.* By taking as initial conditions our current knowledge about the present, climate models can be used to simulate the future. On short enough timeframes, this can be done for a goal of weather prediction (Lynch, 2008). On larger timescales, this can be done for a goal of climate projection, which is prediction based on assumptions about future forcings (IPCC, 2023d).
- *Simulation of alternative climates.* By applying different forcings to the climate models, that correspond to what we know of different conditions, we can reproduce and study the climate of different periods, such as the pre-industrial period, or ancient times ("paleoclimates") (Overpeck, 1995).

1.4.2 . Model evaluation

The formulation and parametrization of a climate model involves a high number of parameters. Values for these parameters must be determined in a way that ensures the convergence of computations, the verisimilitude of the underlying physics (for example, the conservation of energy), as well as similarity with observations of some climate phenomena. The process of selecting values for the parameters of a model is called tuning, and the climate variables chosen for the model to mimic observations are referred to as "tuning targets" (Schmidt et al., 2017). There exist many different tuning strategies that aim at different tuning targets. The choice of tuning strategy is crucial, because it can significantly influence the behavior of the model (Hourdin et al., 2017). As an example, the tuning process of the IPSL-CM6A-LR model is described in (Mignot et al., 2021). Therefore, in order to know how much the output of models can be trusted, it is necessary to evaluate their performance.

To begin with, the most obvious way to evaluate a model is on the similarity of its output with known aspects of reality. However, a model cannot be validated as a whole. Models must be evaluated independently on each specific purpose they are used for (Parker, 2009, 2020; Winsberg, 2018). As an example, the adequacy of a model in representing current precipitations over northern Europe does not guarantee its adequacy for predicting future changes in precipitation in the region. Furthermore, model evaluation must take into account the uncertainties on model outputs, as well as their robustness with respect to changes in parameters, parametrizations, or tuning (Notz, 2015). The relevant evaluation metrics and processes vary depending on the purpose in question. When a model is found to not be fit for an intended purpose, a modeller must either increase the resolution to better represent smaller scales, improve the parametrization scheme of relevant phenomenon, or find a better suited tuning strategy. However, one must note that making a model more fit for one purpose may make it worse for a different task.

Every few years, the Coupled Model Intercomparison Project (CMIP) provides a framework of comparison of climate modelling (Meehl et al., 2000). Its goal is the production of standardized model datasets, to facilitate the evaluation and comparison of currently existing GCM, and the assessments of what can be learned from them about the Earth system. This is achieved by having each participating model run the same set of standardized experiments. At time of writing, the most recent iteration of this project is the 6th. The different experiments run in CMIP6 are detailed in (Eyring et al., 2016).

Thanks to CMIP6, large quantities of model data, with varied parameters, forcings, or initial conditions, are widely available. Ensemble methods are methodologies designed to take advantage of modern processing capabili-

ties and the abundance of data to evaluate the performance of models. By running several models with the same initial conditions and forcings, it is possible to evaluate the uncertainty among models. Using many models reduces the influence of individual choices in parameters, parametrization and tuning. The aggregation of runs of different models can lead to well-quantified uncertainty ranges and therefore higher confidence projections (Knutti et al., 2010). However, this approach has limits. Many models share structures and components (Masson and Knutti, 2011), which can lead to common sources of biases that are systematic across different models (Abramowitz et al., 2019; Boé, 2018). Alternatively, running the same model with a large number of slightly or significantly different initial conditions allows for estimating the robustness of the model output and distinguishing the model's internal variability from its errors (Hawkins et al., 2016; Kay et al., 2015; Maher et al., 2019). This approach can also be used to evaluate parameterizations (Phillips et al., 2004), as done on a large scale in projects like (Williams et al., 2013), and to estimate the uncertainty and robustness of a model with respect to its parameters (Lee et al., 2011; Murphy et al., 2004; Shiogama et al., 2014).

1.4.3 . Recent progress in models

There is a large amount of studies comparing GCMs with observations or reanalyses. In chapter 3 of the 2021 report on the science basis behind climate changes (IPCC, 2023c), the IPCC provides an overview assessment of the ability of models to reproduce the properties of reanalysis and observed climate. Based on this report, we provide here a short state of the art on the performance and progress of GCM climate models.

In the 2013 IPCC report on the physical science basis behing climate change (Intergovernmental Panel on Climate Change (IPCC), 2014), it was found that models that participated in CMIP5 are generally able to represent the properties and evolutions of climate variables, such as temperature, as well as atmospheric circulation (Rodrigues et al., 2018a). Similar conclusions are reached for models participating in CMIP6. Nevertheless, several biases are fond in model outputs relative to observed behavior of the climate. Such biases include underestimation of global warming trends (Oldenborgh et al., 2009), mis-representations of clouds and winds (Lauer et al., 2018), mis-representation of the arctic sea temperatures (Găinuşă-Bogdan et al., 2018; Kuhlbrodt et al., 2018), as well as underestimation of atmopheric blocking events in the north hemisphere (Crueger et al., 2018; Davini and D'Andrea, 2020; Scaife et al., 2010). These biases were found in CMIP5 participants, and were reduced, but not eliminated, in CMIP6. Details regarding recent progresses of climate models are included below.

Despite these difficulties, models have made continued progress on various aspects, as detailed in (IPCC, 2023a). To begin with, modellers were able to significantly increase resolutions between the two last iterations of CMIP. CMIP6 included HighResMIP, a selection of high-resolution simulation experiments, destined to explore the effects of increased resolution on model performance. In HighResMIP, models reached up to 10 km resolutions, which is an order of magnitude above typical GCM resolutions. It was shown that higher resolution allows models to better represent atmospheric mechanisms (Bock et al., 2020). This has resulted in reduced biases, notably on cyclogenesis as well as blocking events (Jiaxiang et al., 2020; Schiemann et al., 2020). However, resolution increase is not by itself sufficient to eliminate biases, several of which were shown to be at least partially caused by model formulation and parametrization. Such biases include the misrepresentation of the persistence of blocking events (Davini and D'Andrea, 2020), and the duplication of the intertropical belt of precipitation (Tian and Dong, 2020).

Parametrizations have also seen improvements between CMIP5 and CMIP6. One consequences of the increase in resolution is that some phenomenon that used not be resolved now are, allowing for parametrization to be simplified. This notably the case for ocean eddies and narrow currents, no longer needing to be parameterized (Hewitt et al., 2017). A commonly used mean to compare and evaluate models is the Climate sensitivity, which designates the increase of temperature that would follow from doubling the CO_2 levels relative to pre-industrial times. The largest source of model spread in the estimation of climate sensitivity is the representation of the effects of short-lived aerosols in the atmosphere. While it remains the main source of spread among models, progress was made in their parametrizations (Meehl et al., 2020). Furthermore, our understanding and representation of within-ice processes have progressed (Faria et al., 2014; Hanna et al., 2020), leading to better ice-ocean interactions representation (Asay-Davis et al., 2017). Similarly, a better understanding of various biogeochemistry processes have helped models reduce uncertainties in carbon cycle representation (Jones and Friedlingstein, 2020). Finally, the inclusion of stochasticity in parameterization schemes was shown to be able to reduce model biases (Berner et al., 2017; Palmer, 2019).

Building models that capture the properties of extreme weather events is difficult, owning to their rarity. Similarly, the limited amount of observation data makes it difficult to evaluate the ability of a model to reproduce the properties of real extreme events. Despite the challenges, extreme events are reasonably well represented by CMIP6 models (Kharin et al., 2013; Li et al., 2021). For temperature extremes in particular, the statistics of simulated extremes rea-

sonably match observations. However, model representation of temperature extremes still contain biases. While CMIP6 models were shown to better represent underlying physical causes than CMIP5, temperatures biases remain similar in model outputs (Di Luca et al., 2020a,b). Additional biases also affect the estimation of trends, with heatwave trends being often underestimated, and cold spell trends underestimated (Fischer and Knutti, 2014; Ringard et al., 2016; Sillmann et al., 2014). Furthermore, there is high model-to-model variability on the sign and amplitudes of many local changes (Borodina et al., 2017). Strong regional differences in extreme events statistics are typically not that well represented by models (Donat et al., 2017).

1.4.4 . Selected datasets

One of the main goals of this work is to evaluate models by assessing their similarity with reality in their representation of atmospheric circulations and extreme events. As several methods (including our own, see chapter 4) require data to be complete (no holes), it is standard practice to use reanalysis as ground truth. Relevant reanalysis datasets include NCEP (Kalnay et al., 1996; Kistler et al., 2001; Saha et al., 2010), and ERA5 (Hersbach et al., 2020), with ERA5 being the general go-to at time of writing due to a combination of availability, higher resolution, and verisimilitude. Regarding the model data to evaluate, models participating to CMIP generate data that is systematically made widely available. It is used in many studies, and is therefore well documented. Furthermore, it represents the latest advances in climate models. In this work, we take our model data from CMIP6, the most recent CMIP at time of writing. Specifically, we take GCM data, and we focus on GCMs that have a large amount of available runs. Based on these criteria, our choice of models is:

- **IPSL-CM6A-LR**, from the Institut Pierre-Simon Laplace, in France (Boucher et al., 2020).
- **MIROC6**, from the Atmosphere and Ocean Research Institute, in Japan (Tatebe et al., 2019).
- **ACCESS-ESM1.5**, from the ACCESS-National Research Infrastructure, in Australia (Ziehn et al., 2020).
- **CanESM5**, from the Canadian Centre for Climate Modelling and Analysis, in Canada (Swart et al., 2019).

1.5 . Summary

In this thesis work, we mainly focus on atmospheric circulation, i.e. the movements of air in the atmosphere, carrying with it properties of the air masses

such as temperature or humidity. Atmospheric circulation is mostly determined by the localization and intensity of cyclones and anticyclones. Some patterns of atmospheric circulation can cause extreme weather events, such as heatwaves or cold spells. Extreme weather events are a relevant subject of study due to their impacts on society. Furthermore, circulation patterns and their effects on temperature are projected to change with anthropogenic climate change.

Climate models are tools of simulation that are used for many purposes in climate sciences, from enriching observation data to making projections about the future of the climate. They must be very complex, in order to be able to faithfully reproduce many of the phenomena and behaviors observed in reality. However, despite their continuous progress, models still struggle with systematic biases in their representations. In particular, there is a lot of progress to be made on their representation of extreme events. In order for models to be able to progress, it is crucial to keep developing evaluation methods, so that model flaws can be identified, and the models can be improved.

In this thesis work, we want to evaluate how models represent atmospheric circulation and extreme events. To that end, pressure-related variable maps, such as Sea-Level Pressure, are commonly used, as they contain the information on cyclone and anticyclones. It is standard practice to use reanalysis data, which is incomplete observations filled in using models, as ground truth to compare the models to. The models that we will study are selected from the CMIP6 project, which consists in a standardized set of experiments meant to evaluate models and produce large quantities of available data.

2 - Classification of atmospheric circulation patterns

2.1 . Introduction

Weather is in large part determined by the circulation of air masses in the atmosphere. Atmospheric circulation is itself described by highly dimensional data, generally pressure maps. One way to overcome the high dimensionality hurdle is to classify daily maps into a finite set of weather patterns. Weather patterns are configurations of the circulation of air masses in the atmosphere, described through the intermediary of pressure variable maps. Here, we are concerned with methodologies for the identification of atmospheric circulation patterns, and the classification of maps along them. These methods aim to separate individual maps into a finite set of classes which have, ideally, relevant and distinct effects on weather.

In this chapter, we go over the different existing classification methods, with applications to the problem of atmospheric circulation classification. We present Principal Component Analysis (PCA), a dimensionality reduction technique commonly used as preprocessing before classification. We show how PCA can be applied to determine weather regimes, which are classifications of circulation patterns of large spatial and temporal scales.

Then, we focus on classification methods that apply to individual daily maps of pressure-related data. We first present hard-clustering methods, which attribute to each map a unique class. Among them are hierarchical clustering, k-means, the rotated T-mode PCA, as well as self-organizing maps. We continue with soft-clustering methods, which attribute to each map a distribution over classes instead. Among them, we present several methods of fuzzy clustering, as well as a type of statistical model called mixture model. Part of the discussion this chapter is based on (Huth et al., 2008).

2.2 . Generalities

While classification is inherently an approximation, as atmospheric circulation is continuous, and not naturally separated into discrete configurations, a classification scheme provides a proxy for studying the dynamics of atmospheric circulation dynamics. Insight on circulation dynamics can be obtained by studying the properties of each class, such as frequencies of appearance,

persistence, and preferred transitions to other classes. We present several methodologies used to that effect, and reference results they have produced, based on a review article from Huth et al. (Huth et al., 2008). Atmospheric circulation pattern classification uses as input data 2-dimensional fields of a pressure-related variables, generally either SLP or geopotential height at a pressure level of 500 or 700 hPa. Some studies have demonstrated that there is little to be gained in using more than one pressure level, due the high amount of dependence among them (Esteban et al., 2005b; Kidson, 1997). Therefore, all methodologies presented here will be made to work with individual 2-dimensional pressure level maps as input and as objects to classify. According to (Huth et al., 2008), here is a list of five properties to evaluate a classification scheme by:

- *Consistency*: The classification does not vary excessively with small variations of its hyperparameters.
- *Separability*: Maps within a class are similar to each other, while maps of different classes are dissimilar.
- *Stability*: The classification does not vary excessively with small alterations in the data, like re-sampling, or altering its grid resolution.
- *Structure*: Do the classes contain similar amounts of maps, or is there one class containing most of the data.
- *Reproduction of predefined types*: Expected results, gained from physical knowledge of the system, are reproduced.

Classification of atmospheric circulation patterns is a long-standing practice historically used in the context of weather forecast. This was known as synoptic classification, and would include synoptic catalogues of circulation patterns. In the Hess-Brezowsky catalogue (Hess and Brezowsky, 2010), the patterns are chosen using expert knowledge of their relevance and influence on weather. In the Lamb catalogue (Lamb, 1972), they are based on geometric physical information such as the direction of associated wind. Such catalogues can allow for high-performing and pertinent classification, but require pressure variable maps to be classified by hand and eye, making them difficult to generalize to big datasets. Furthermore, while physical and expert knowledge is included in the construction of the classes, synoptic catalogues have unavoidable arbitrariness in the class definitions and map attribution in boundary, limit cases. Recent works have adapted these classification system for computer automation. See (James, 2007) for the Hess-Brezowsky catalogue and (Jones et al., 1993) for the Lamb catalogue. However, that does not remove the problem of arbitrariness in the class definition. Furthermore, the automatic classification is meant only to facilitate application to large datasets, and not to replace the need for classification by a human expert These automated methods are not meant to constitute a self-sufficient

classification procedure by themselves.

2.3 . Principal Component Analysis

2.3.1 . Dimensional reduction

Attempting to apply any classification method to pressure data raises two problems. Firstly, the high number of dimensions of the datapoints causes difficulties. A high dimensionality complicates the problem, increasing computation times, memory requirements, and the amount of local minima to avoid during optimization. Secondly, a bias emerges from the fact that there are highly correlated variables within the input data maps. Some classification methods, such as clustering (see below) are based on a measure of distance between datapoints, that attributes an equal weight to each data variable. Variables containing redundant information are a problem in these classification methods, because the redundant information would be given additional weight by the amount of variables it appears in (Davis and Kalkstein, 1990; Fovell and Fovell, 1993). As an example, if the pressure in six neighbouring grid points is always identical, then the pressure in this region would be given six times the weight in the estimation of which circulation maps should be considered similar. Furthermore, this increase in weight would be dependant on the choice of grid resolution. For these two reasons, it is preferable to preprocess the data before classification. Specifically, the data must be converted into a new basis, one that has a lower amount of dimensions, and that avoids or reduces correlations between the variables.

Principal Component Analysis (PCA) (Kherif and Latypova, 2020; Preisendorfer, 1988; Wold et al., 1987) is an unsupervised dimensional reduction method. It is known in some domains of research as Proper Orthogonal Decomposition, or POD, and in the field of atmospheric sciences, it is often known instead as Empirical Orthogonal Function analysis (EOF analysis) (Hannachi et al., 2007). The different names all refer to the same methodology. The goal of PCA is to find modes (linear combinations of the data variables), all orthogonal to one another, that maximally explain the variance within the data. Furthermore, the modes it generates are ordered, with the first mode being optimized to explain as much of the data variance as possible, and each subsequent mode explaining as much variance as possible, among what is not yet explained by previous modes. Those modes can then be substituted to the data variables.

Let $(\boldsymbol{P}_{d,i})_{d\in\{1,\dots,D\},i\in\{1,\dots,I\}}$ designate a set of D pressure maps, each containing I variables. Each datapoint $\boldsymbol{P}_d$ is a vector of length I. The input data can

be written as a matrix, $\boldsymbol{P}$, of shape $D \times I$, D being the amount of datapoints (or realizations, or time steps). Then, we are searching for modes, themselves vectors of length I. Let them be noted $\boldsymbol{m}$. If there are K of them, then they can be written as a matrix $\boldsymbol{M}$ of shape $I \times K$, such that:

$$\boldsymbol{T} = \boldsymbol{P}\boldsymbol{M} \tag{2.1}$$

Where $\boldsymbol{T}$ is the matrix of shape $D \times K$ representing the projection of the data in the PCA mode basis.

PCA modes are called principal components. Let us start by searching for the first one. It is defined as the linear combination of the I variables that lead to the highest explained variance. The explained variance of a mode is defined as the variance of the projection of the data in that mode. Therefore, we search for $\boldsymbol{m}_1$ that minimizes:

$$\begin{aligned}\sum_d \mathsf{Var}(\boldsymbol{T}_{d,1}) &= \sum_d \mathsf{Var}\left(\sum_i \boldsymbol{P}_{d,i}\boldsymbol{M}_{i,1}\right) \\ &= \sum_d \mathsf{Var}\left(\boldsymbol{P}_d \cdot \boldsymbol{m}_1\right) \\ &= \sum_d \left(\boldsymbol{P}_d \cdot \boldsymbol{m}_1 - \frac{1}{D}\sum_d \boldsymbol{P}_d \cdot \boldsymbol{m}_1\right)^2\end{aligned} \tag{2.2}$$

Where $\cdot$ designates the scalar product. The solution can be shown to be an eigenvector of the correlation matrix of the data, $\boldsymbol{P}^T\boldsymbol{P}$. Specifically, it is the eigenvector with the highest associated eigenvalue. Furthermore, the proportion of explained variance explained by this mode is equal to this eigenvalue divided by the sum of all eigenvalues of the data matrix, i.e. the trace of $\boldsymbol{P}^T\boldsymbol{P}$.

$$\frac{\sum_d \mathsf{Var}(\boldsymbol{T}_{d,1})}{\mathsf{Var}(\boldsymbol{P})} = \frac{\lambda_{\mathsf{max}}}{\mathsf{Tr}\left(\boldsymbol{P}^T\boldsymbol{P}\right)} \tag{2.3}$$

To compute further modes, one must subtract from the data its projection into all previous modes, expressed into the basis of the original I variables. It can be proven that modes constructed this way, maximizing at each step the explained variance, correspond to the eigenvectors of the data correlation matrix, sorted in descending order of the eigenvalues.

This process can be continued until I modes have been found. At that point, all the variance will be explained, and one has obtained an alternative basis for the data, of the same dimension. As the variables of this new basis are the eigenvector of the data correlation matrix, which is Hermitian, they are all orthogonal. This ensures that they are not linearly related to each other. By construction, this basis avoids highly-correlated variables, and therefore

the bias due to redundant information. Furthermore, the contribution to explained variance decreases with each new mode, and in high dimension, it will typically become negligible before I modes are found. PCA functions as a dimensional reduction method by truncating the process, and keeping only the first few modes. The amount of modes to keep is either directly chosen as a hyperparameter of the method, or by keeping only modes higher than a predefined threshold of explained variance.

Therefore, PCA outputs a series of modes, in decreasing order of relevance. Those modes form a set of linearly independent dimensions and can serve as a basis to project the data into. As this process ensure both that the dimension of the data is reduced, and no variables contain redundant information (at least in the form of linear dependencies), it solves the two problems of clustering data that were previously identified. This is why PCA is often used as preprocessing before the application of classification methods.

The PCA methodology, under the name of EOF analysis (Fukuoka, 1951), was introduced in atmospheric sciences by (Lorenz, Edward N, 1956). Like in many other domains, it has seen and still sees a lot of application (Richman, 1990). For its application in atmospheric circulation patterns classification, see the rest of this chapter. Most of the classification methodologies we discuss begin by an application of PCA for data preprocessing, and it has been a standard for a significant time (Christensen and Bryson, 1966).

2.3.2 . Application: Weather Regimes

Weather regimes are a commonly used type of atmospheric circulation classification, introduced in (Rex, 1950). Here, we present them as both an introduction to the concept of atmospheric circulation classification, and as an example application of PCA to this domain. Unlike other methods we look at in this chapter, weather regimes are a classification focusing on low-frequency dynamics. Weather regimes typically exist on temporal scales of at least 10 days, and consider spatial fluctuations at synoptic scale or higher. As stated in (Michelangeli et al., 1995), weather regimes are a small set of structures that are defined by either:

- *Persistence*: Weather regimes are the patterns that last the longest amount of time on average.
- *Recurrence*: Weather regimes are the patterns that occur the most often. In other words, they are the maxima of the probability distribution function in phase space.
- *Quasi-stationarity*: Weather regimes are patterns that are stationary, i.e. for which derivatives cancel out and vanish.

Baroclinic disturbances in the atmosphere, which have a life cycle on the temporal scale of about a week, and can cause fluctuations at smaller spatial scale, are only considered as modulations of the regimes. As such, most individual daily maps are transitory, not coinciding with any regime directly. Weather regimes are applied to classify agglomerated data over several consecutive days, rather than daily maps.

The computation of weather regimes, whether based on *Recurrence* or *Quasi-stationarity*, typically makes use of PCA as a dimensionality reduction tool. Because these definitions of weather regimes require the exploration of phase space, it is useful to reduce its dimension to a more manageable amount. In addition to this, PCA is also used as a form of filtering. PCA modes are generally associated with a specific spatial frequency. Therefore, by carefully selecting which and how many modes to take into account, it is possible to filter out frequencies that are not of interest (in the case of weather regimes, higher frequencies). This allows the reduced phase space to incorporate only properties relevant to weather regimes.

PCA has notably been applied with these two goals to study circulation dynamics in the North-Atlantic region in (Michelangeli et al., 1995) as well as in (Vautard, 1990). We show in figure 2.1 the 4 regimes obtained for the winter in the north-Atlantic region in (Michelangeli et al., 1995) with a quasi-stationarity based method.

However, in this thesis work, we chose to focus on methodologies adapted to the classification of maps on a daily scale, and take into account fluctuations due to baroclinic disturbances.

2.4 . Hard clustering

Clustering, or cluster analysis, designates the unsupervised process of automatically finding arrangements of unlabelled datapoints into classes (or clusters) in such a way that all elements within a class are as similar as possible, while elements from different classes are as different as possible (Han et al., 2022). This corresponds to the *separability* property defined earlier. Depending on the method, the number of clusters K can either be user-chosen, or automatically optimized. There exist two main types of clustering methods: hard clustering, which attributes to each map a unique cluster, and soft clustering, which attributed to each maps a distribution over clusters.

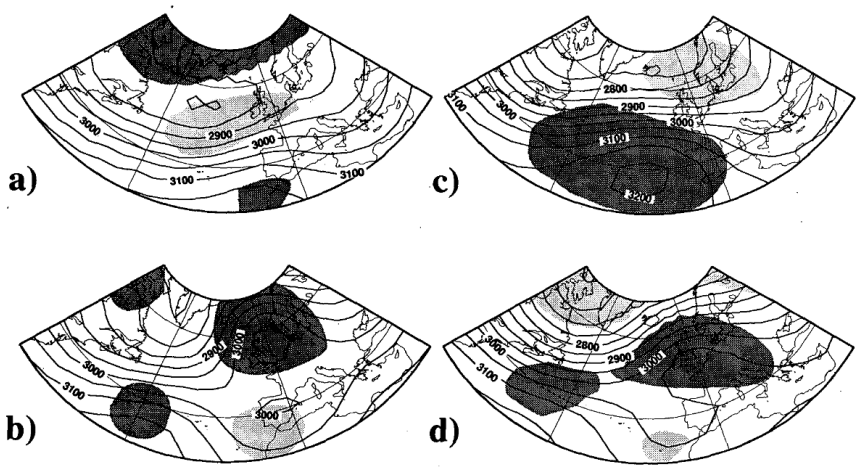


Figure 2.1 – 4 weather regimes of 700 hPa geopotential height obtained for winter in north-Atlantic region using a quasi-stationarity based method. Contour interval is 50m. Dark shaded areas are where the anomaly is more than 50m higher than the average. Light shaded areas are where the anomaly is less than -50m higher than the average. This figure is from (Michelangeli et al., 1995).

2.4.1 . Hierarchical clustering

Hard clustering methods can be split into two categories: hierarchical, and non-hierarchical methods. Hierarchical methods consists in building a hierarchy of nested clusters. It begins with attributing each datapoint to its own cluster, then iteratively merging the two closest clusters until only one cluster is left. (To be exhaustive, one should mention this only the *agglomerative* type of hierarchical models. *Divisive* types, which start with one cluster and then iteratively divide clusters until each datapoint is its own cluster, also exist.) The metric of cluster similarity, based on which the choice of clusters to merge is based, greatly influences the result of the clustering. There are two layers to this choice. First, a measure of distance between datapoints is picked ; more often than not it is the mean square distance. Then, based on this distance metric, one must choose a way to determine the distance between two clusters. This choice of generalization of the distance metric to clusters is called the linkage. Simple examples include *minimum linkage*, or *maximum linkage*, in which the distance between two clusters is the minimum (respectively, maximum) of their element-wise distances. These and many other linkage algorithms are detailed in Murtagh, 1985 (Chambers et al., 1985). (Fovell and Fovell, 1993) also discusses in detail the methodological biases of several previously discussed methods, such as the different linkage choices of hierarchical clustering.

Hierarchical clustering has seen application atmospheric circulation classification. Examples of such applications of hierarchical cluster analysis use the following linkages.

- *Average linkage*: The distance between two clusters is the average of their element-wise distances. In (Mote, 1998), average linkage was applied to find correlations between synoptic patterns and Greenland ice melting.
- *Ward linkage*: The two clusters to be merged are those that would result in the lowest total within-cluster variance. In (Cheng and Wallace, 1993), the Ward method is used to classify large scale circulation in the northern hemisphere. In (Fernau and Samson, 1990) and (Vrac et al., 2007), it is used to identify circulation patterns in North-American atmospheric circulation.

Among these two linkages, (Kalkstein et al., 1987) finds that while the Ward method reliably leads to clusters of similar sizes, average linkage is the best suited to the goal of clustering extreme events separately from regular days. The COST733 Action (Huth, 2010; Philipp et al., 2016) compared the performance for atmospheric circulation classification of many different classification methods, including several hierarchical clustering linkage.

As an example, we show in figure 2.2 the centroids of 3 clusters obtained in anomalies of z500 height on the norther hemisphere in (Cheng and Wallace, 1993), using Ward linkage hierarchical clustering.

Figure 2.2 – Average 500 hPa geopotential height anomaly maps of 3 clusters obtained by applying Ward linkage hierarchical clustering to northern hemisphere winter data. Contour interval is 50 m. Dashed lines indicate negative anomalies. This figure is from (Cheng and Wallace, 1993).

2.4.2 . k-means

Non-hierarchical clustering takes as an input parameter the number of clusters K, and uses centroids to represents them, associating each datapoint to

one centroid. Here, we will detail one of the most common and representative non-hierarchical clustering methods, k-means (Ahmed et al., 2020; Hartigan, 1975). K-means is a simple model that takes as input a set of D datapoints $\boldsymbol{P}_d$, in the shape of vectors of length I, and a cluster amount K. The clusters S_k, for $k \in \{1, ..., K\}$, are defined by centroids, themselves vectors of length I. The goal of k-means is to attribute each datapoints to one of the clusters in a way that minimizes the within-cluster variance, written in equation 2.4.

$$\sum_{k=1}^{K} \sum_{\boldsymbol{P}_d \in S_k} \left|\left| \boldsymbol{P}_d - \frac{1}{N_{S_k}} \sum_{\boldsymbol{P}_d \in S_k} \boldsymbol{P}_d \right|\right|_2 \tag{2.4}$$

where $||\cdot||_2$ designates the ℓ_2 norm, and N_{S_k} the amount of elements in cluster S_k.

We present in algorithm 1 a standard algorithm for k-means. It is known as the "naive" algorithm due to more efficient options existing. To summarize, each

Algorithm 1 Naive k-means learning process

Initialize cluster centroids $(S_k)_{k \in \{1,...,K\}}$
while hasn't converged **do**
  for d in $\{1, ..., D\}$ **do** ▷ Assignment step
    Assign $\boldsymbol{P}_d$ to S_k such that $k = \text{argmin}_k ||\boldsymbol{P}_d - S_k||_2$
  end for
  for k in $\{1, ..., K\}$ **do** ▷ Update step
    Update centroids to $S_k = \frac{1}{N_{S_k}} \sum_{\boldsymbol{P}_d \in S_k} \boldsymbol{P}_d$
  end for
end while

datapoint is associated to the closest centroid, then each centroid moves to the average of all its assigned datapoint. The process is then repeated until convergence.

We now present some examples of application of this method on atmospheric classification. Note that these applications use more efficient optimization algorithms than the one presented before, as well as refined initialization schemes. In (Brinkmann, 1999), k-means is applied to z700 data over eastern North America data to explore the link between temperature anomalies and circulation patterns. In (Corte-Real et al., 1998), using SLP, it is applied to explore the circulation patterns associated with dry weather or precipitation over Portugal, and explore their variations over years. These associations are then tested for in a GCM model (HADCM2) in (Corte-Real et al., 1999). In (Esteban et al., 2005a), the authors determine circulation patterns responsible for

heavy snowfall in Andorra. And in (Esteban et al., 2006), the same methodology is applied on NCEP reanalysis data to construct a catalogue of circulation patterns over western Europe.

However, k-means has downsides. In particular, its lacks *Consistency*: the clusters are highly dependant on the choice of initialization. K-means optimization is likely to lead to local minima, failing to find the global minimum. Simulated annealing is a mean of ensuring *Consistency* when solving the clustering problem (Selim and Alsultan, 1991). Rather than a clustering method, simulated annealing is an optimisation process. It can be applied to the problem of choosing clusters that minimize internal variance to avoid local maxima and find the global minimum (Laarhoven and Aarts, 1987).

In consists in the following. Let us call "loss" any function that we want to minimize. In this cluster analysis case, it is the total within-cluster variance. At each step, rather that systematically making a change that would reduce loss, like gradient descent would do, simulated annealing picks a random move in parameter space. Then, the random move is accepted at a probability that exponentially decreases with the increase in loss. By analogy with physical processes, this loss increase is treated like an increase in energy ΔE. More specifically, the probability of accepting a move associated with ΔE is:

$$1 - \exp \frac{\Delta E}{kT} \tag{2.5}$$

Continuing the analogy, T is called the temperature. At higher temperature, simulated annealing will have an easier time escaping from local minima, but a harder time converging, and vice-versa. During the simulated annealing process, T starts high and then is progressively decreased, following a predefined "annealing schedule". k is a constant factor that relates temperature to energy, analogous to Boltzmann's constant. It exists to relate values of temperature with values of the loss function, and ensure they are of comparable scales. While technically a hyperparameter of the method, changing k is equivalent to a multiplicative change of variable in the annealing schedule defining the values of T.

A example of application to circulation patterns classification would be (Philipp et al., 2007). In this article, the authors use simulated annealing to describe atmospheric circulation patterns over the North-Atlantic-European region, and study their changes since pre-industrial times.

2.4.3 . PCA as a classification method

T-mode PCA

This subsection is mainly based on the work of Michael B. Richman, in (Richman, 1986). While PCA is useful as a tool for dimensional reduction, notably due to the linear independence of its modes, it can also be applied as a classification method itself. However, that requires using it with a different methodology from what was previously presented, as PCA modes do not typically resemble common atmospheric circulation patterns. In the domain of circulation pattern classification, when PCA is used as preprocessing, the data matrix has datapoints as rows, and data variables as columns. This, which has been presented in an earlier section, is referred to a S-mode PCA, and leads to the PCA modes (Principal Components, PC) being linear combinations of the data variables. When used as a classification tool, the data matrix is transposed. Now of shape $I \times D$ instead of $D \times I$, the rows are data variables and the columns are datapoints. This is referred to as T-mode PCA, and leads to PC being linear combinations of the datapoint indexes. Figure 2.3 summarizes the differences between these two methods. However, this is not enough yet to provide a classification scheme.

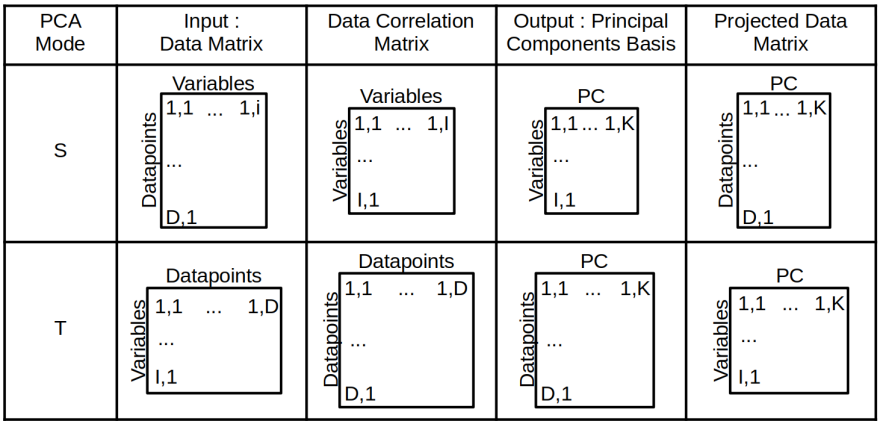


Figure 2.3 – Difference between S-mode and T-mode of PCA. In T-mode, the data matrix is transposed before PCA application. In S-mode, the principal components (PC) are defined over data variables. In T-mode, PC are defined over datapoints. This figure is adapted from (Richman, 1986).

Rotated PCA

The PC modes obtained from PCA, either in S-mode or T-mode, have a set of characteristics that can be undesirable:

- *Dependence to domain shape*: The first few PC obtained by PCA are generally low-frequency modes covering the entire spatial domain. Generally, these modes depend only on the shape of the spatial domain, and not on the input data.
- *Subdomain instability*: Because of the previous property, applying PCA to a subdomain of space usually does not yield subdomains of the PC, but rather compressed versions of them.
- *Sampling error*: When eigenvalues are close, the order of the modes may change depending on the sampling of the data.
- *Lack of physical interpretability*: In meteorological contexts, PC generally fail to represent meaningful or interpretable states on their own, and only have physical meaning when combined (Lamb, 1985).

A solution to these four problems is to apply a rotation to the PCA basis in phase space. Applying a rotation to the PCA basis allows one to modify it along new degrees of freedom, while preserving the orthogonality of the PC. The specific rotation is chosen to optimize for desired properties. This is done at the sacrifice of the variance maximization property of the PC (Richman, 1986).

There are many different ways to choose a rotation to apply. Historically, the optimization criterion was motivated by the desire for the PC to be associated with human intuition-compatible physical interpretations. The first objectively defined criterion for finding useful rotations of the PC was described by Thurstone in (Thurstone, 1933). However, more recent applications, especially in the study of weather variables, use the varimax criterion defined in (Kaiser, 1958). The varimax criterion optimizes for a balanced spread of the explained variance among the rotated PC. It also has the effect of making the PC *sparse* - only a limited amount of variables are non-negligibly represented in each given PC. Thus, application of such a rotation to the PC basis results in a set of data-informed, interpretable modes that are generally local and sparse. This can be done to S-mode PCA, resulting in localized, physically interpretable modes. But when done to T-mode PCA, where the resulting PC are defined over datapoint indexes, the sparsity property means that in each PC, only a limited amount of datapoint are represented non-negligibly. By interpreting each PC as a cluster, and considering a datapoint index being represented in a PC, as it being included in the corresponding cluster, we have applied PCA as a clustering method.

Rotated T-mode PCA classification has been shown to be capable of capturing previously known circulation patterns classes in artificial test data (Compagnucci and Richman, 2008), and in European midlatitudes (Huth, 1993). It has also been applied to study circulation in the northern hemisphere (Bartzokas and Metaxas, 1996). In (Huth, 1996), the author explores the optimal ways to

apply T-mode rotated PCA on geopotential height maps to classify circulation over Europe. We show in figure 2.4 example circulation patterns obtained by (Huth, 1996) with a rotated T-mode PCA method in the European and north-Atlantic region, using the z500 variable. This methodology is then adapted for the purposes of model evaluation in (Huth, 2000; Huth, 1997). SLP has also seen use under this methodology for the classfication and study of the temporal variations of circulation patterns in (Jacobeit et al., 2003), and to explore links between circulation patterns and extreme cold in the wet Pampas, Argentina, in (Müller et al., 2003).

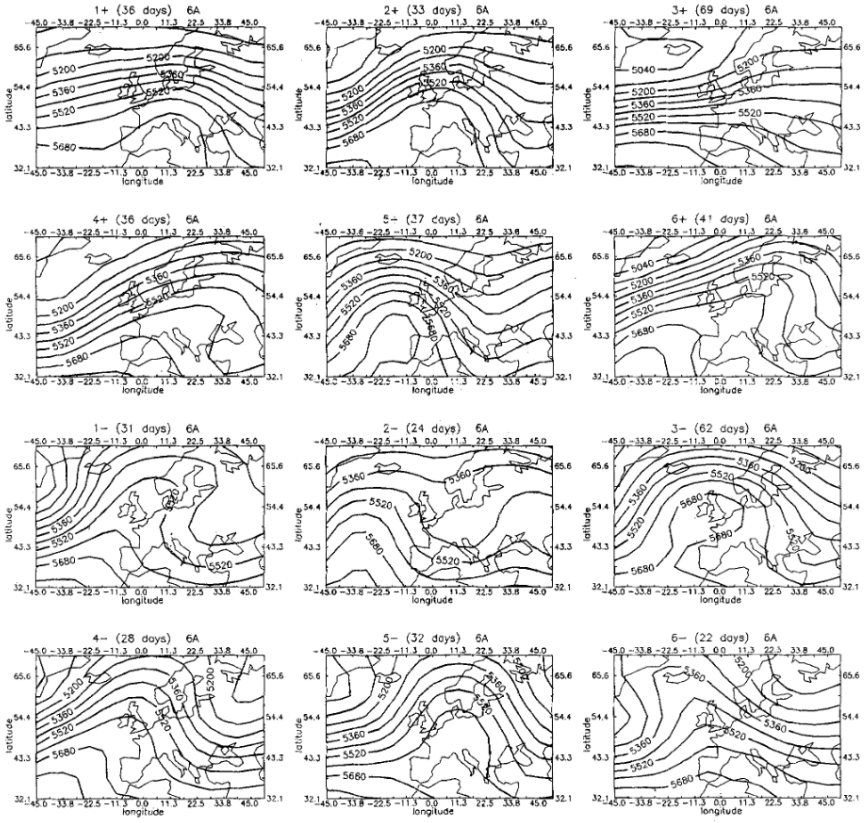


Figure 2.4 – Mean 500 hPa geopotential height (in meters) for example classes obtained by rotated T-mode PCA. This figure is from (Huth, 1996).

2.4.4 . Self-Organizing Maps

A neural network is a machine learning model using an architecture based on units called neurons, arranged in a set of interconnected layers. Each neuron takes values as its input (either the output of the previous layer, or the input data if it is the first), applies a non-linear function depending on some param-

eters to it, and produces a value as its output. As long as there is non-linearity in the neuron, a sufficiently deep (number of layers) and wide (amount of neurons per layer) network can in theory model any function. However, as is often the case in machine learning, deep and/or wide networks have many parameters to optimize and therefore require a lot of data to train. Here we will focus specifically on Self-Organizing maps (SOM), a popular neural network methodology in the domain of atmospheric circulation study.

SOM are an architecture for unsupervised learning initially developed by Kohonen in (Kohonen, 1990; Kohonen, 1995). The goal of SOM is twofold. First, SOM are a clustering method. It aims to sort individual datapoints among a user-chosen number of clusters. The clusters are defined by centroids, i.e. points in the phase space of the data, which are optimized in the learning process. Secondly, the main particularity of SOM is that the clusters also have positions relative to each other in an abstracted cluster space. Some clusters are close and others are further away, giving additional meaning to the cluster attributions of the datapoints. As such, SOM also serves as a sort of dimensional reduction method, with discrete output. The distances between the clusters are determined by the positions of nodes in cluster space associated with each cluster. The distribution of the cluster-nodes in cluster space, as well as its dimension, are user-chosen. In this discussion, we will only consider 2-dimensional cluster-space and regular arrangements of nodes.

SOM have one layer of neurons, which are identified with the nodes in cluster-space. Each node has I parameters that define the centroid of the associated cluster. This is referred to as the "reference vector". In order to determine the cluster attributed to a datapoint by an already optimized SOM model:

- The datapoint is compared in parallel to the reference vector of each node, using a distance metric (in general, either the Euclidean distance, or the scalar product).
- The node with lowest distance becomes the attributed cluster.

Non-linearity is necessary to any neural network, in order to be able to represent arbitrary functions. But the distance metric chosen to compare the datapoint with the reference vectors needs not be non-linear in the data variables, because the non-linearity of SOM networks comes in the "best fit" selection process. The architecture of SOM summarized in figure 2.5.

Most neural network methodologies use backpropagation to optimize the parameters. Backpropagation is a methodology of gradient estimation in neural networks commonly used for parameter optimization. It was originally named in (Orbach, 1962), however the first implementation was described in (KELLEY,

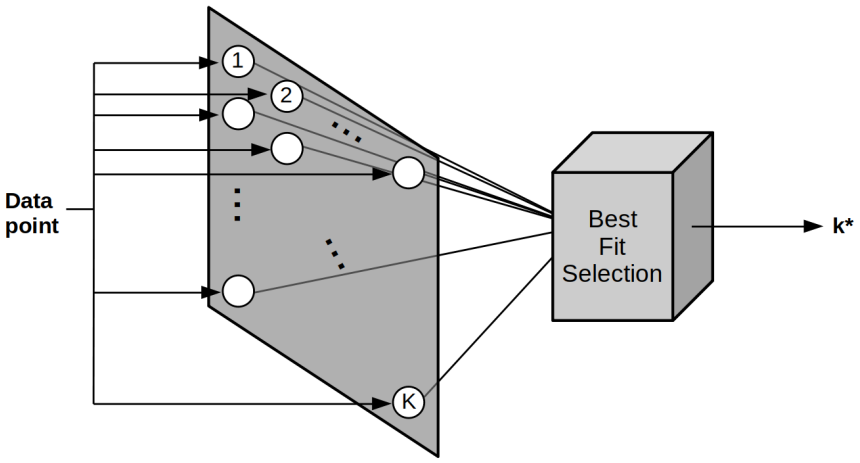


Figure 2.5 – Schematic representation of Self-Organizing Maps. An input map is compared to the reference vectors of K nodes, and a best fit $k*$ is selected as output. In this example, the nodes are arranged in a 2-dimensional cluster-space.

1960). However, SOM utilize competitive learning, detailed in algorithm 2. The symbol $\leftarrow$ means: "Is set to the value of". Let K be the amount of clusters. $\forall k \in \{1, ..., K\}$, let $\boldsymbol{m}_k$ designates the reference vector associated with cluster k. And finally, let $||\cdot|\cdot||$ be the distance metric chosen to compare datapoints with reference vectors, and $||\cdot||$ the distance metric in cluster space. In each step t of the learning process, a datapoint $\boldsymbol{P}_t$ is classified by being compared to every reference vector and picking a best fit $\boldsymbol{m}_{k^*}$. The best fit node, as well as its neighbours in cluster space are then updated to have their reference vector become closer to the classified datapoint. How much a given node k is to be adjusted is determined by the neighborhood function, noted $h_{k^*k}(t)$. It is a function of distance between the positions of node k, r_k and that of the best fit node k^*, r_{k^*}, and time t (i.e. learning steps): $h_{k^*k}(t) = h(||r_k - r_{k^*}||, t)$. The neighbourhood function determines how close nodes must be to be considered neighbours. It goes to zero when $||r_k - r_{k^*}||$ increases to ensure locality of the learning, and goes to zero when t increases to ensure convergence. This local (in cluster space) update process is where the spatial structure of the nodes comes into play, and the source of the global ordering in the resulting clustering scheme. As a result, the *structure* of the clustering provided by SOM is different from other methods, such as k-means. Where these tend to generate few, populated clusters where data is dense, and many where data is sparse, SOM on the other hand generates many clusters where data is dense, and few where it is sparse.

Algorithm 2 Self-Organizing Maps learning process

Randomly initialize the m_k, $\forall k \in \{1, ..., K\}$
for t in $\{1, ..., D\}$, repeated until convergence **do**
  $k^* \leftarrow \text{argmin}_k \, ||\boldsymbol{P}_d | \boldsymbol{m}_k||$
  for k in $\{1, ..., K\}$ **do**
    $\boldsymbol{m}_k(t+1) \leftarrow \boldsymbol{m}_k(t) + h_{k^*k}(t) \times (\boldsymbol{P}_t - \boldsymbol{m}_k(t))$
  end for
end for

A visual example of the learning process, from (Kohonen, 1995), is displayed in figure 2.6. In this simple case, both input data space and cluster space are 2-dimensional. The probability distribution of the data fed to the learning model is assumed to be uniform on the square region. We see the reference vectors, initialised randomly inside a small centered circle, progressively adjust to represent an approximation of the probability distribution of the input data.

To the author's knowledge, this method was first applied to weather data by Cavazos in (Cavazos, 1999), then in (Cavazos, 2000), to the classification of weather patterns associated with extreme events. It was demonstrated that SOM were capable of revealing underlying structure in its input data in (Hewitson and Crane, 1994). They have been successfully used to extract patterns associated with extreme weather events. It has notably permitted studies of the temporal evolution, in frequency and aspect, of these patterns (Hewitson and Crane, 2002). Since then, the SOM method has seen a lot of use in the domain, owning to its ability to provide detailed classifications with high number of classes ; while retaining approachable physical interpretability thanks to the visualization of the underlying structure of classes (Sheridan and Lee, 2011). This method continues to see new applications in the field of atmospheric circulation classification (Mwasiagi, 2011).

2.5 . Soft clustering

Up to now, the methods that were presented are for hard clustering, where the attribution of a datapoint to a cluster is:

- *Boolean*: A datapoint is either included or not included in a cluster, without third option such as partial inclusion.
- *Unique*: A datapoint is only included in exactly one cluster.

The only exception would be T-mode PCA clustering, which is not *unique*, as it may attribute datapoints to numbers of clusters other than one. However

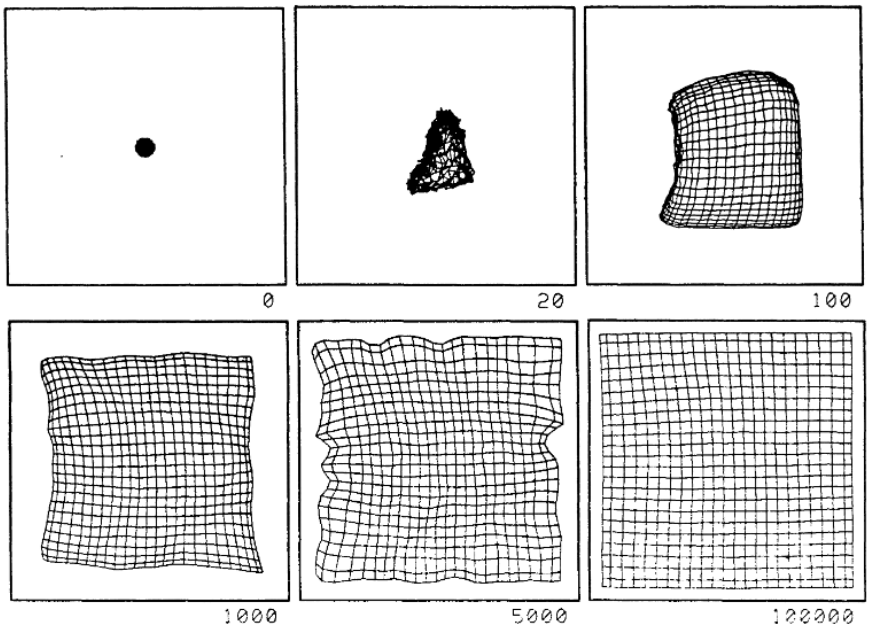


Figure 2.6 – Example representation of the SOM learning process. The background square represents the 2-dimensional space of possible input vectors, which we assume to be of uniform probability distribution. The reference vectors of the nodes after varying amount of learning steps are represented by a grid, with the links representing the nearest neighbours of each node, revealing the 2-dimensional topological structure of the cluster space. This figure is from (Kohonen, 1995).

there exist clustering methodologies, referred to as "soft clustering", that aim to provide more subtle information on datapoint categorization by attributing them cluster attributions in the form of numbers between 0 and 1 for each cluster. Those numbers may symbolize either a degree of fulfillment of conditions that describe a cluster, or a probability of belonging to it, rather than categorical attributions. Because it provides more detailed information on the datapoints, soft clustering will be our approach of choice for the circulation classification performed in this thesis work.

2.5.1 . Fuzzy relation-based methods

Among soft clustering methods, there exists a set of methods referred to as fuzzy clustering, which uses values between 0 and 1 instead of Boolean to deal with imprecise and subjective statements, including the degree of membership of a datapoint to a cluster (Yang, 1993). Fuzzy clustering designates a wide array of methods. First we look at a method based on fuzzy relations, or

rules. This methods consists in defining clusters using imprecise statements such as "there is very low pressure at this position". For each datapoint, the truth value of these statements are represented by values in $[0, 1]$, using the concept of fuzzy logic as proposed by Zadeh in 1965 (Zadeh, 1965). From these statements are then computed the degrees of membership of datapoints to clusters (see below for more detail).

This method was introduced to atmospheric circulation classification in (Bardossy et al., 1995; Bárdossy et al., 2002). Its performance was compared to other classification methods in (Stehlík and Bárdossy, 2003). It remains at time of writing an uncommon method, with few articles applying it to the domain of atmospheric circulation classification.

In order to apply fuzzy rules-based classification to data, it must first be brought to the $[0, 1]$ range. This is to ensure a regular scale shared for each datapoint, and facilitates the definition of terms such as "low" or "very high". Using equation 2.6, input data maps are normalized to set each daily map minimum at 0 and maximum at 1.

$$\forall d \in \{1, ..., D\}, i \in \{1, ..., I\}, \boldsymbol{X}_{d,i} = \frac{\boldsymbol{P}_{d,i} - \min_j (\boldsymbol{P}_{d,j})}{\max_j (\boldsymbol{P}_{d,j}) - \min_j (\boldsymbol{P}_{d,j})} \tag{2.6}$$

The classes are defined by fuzzy rules consisting of a list of I statements, one for each data variable. Those statements are either VERY LOW, MODERATELY LOW, MODERATELY HIGH, VERY HIGH or INDIFFERENT. In (Bardossy et al., 1995), those were chosen manually based on the Hess-Brezowsky catalogue (Hess and Brezowsky, 2010), with the aim of representing these patterns as the classes. From these Hess-Brezowsky catalogue maps were selected a set of representative locations of very low, medium low, medium high and very high pressures ; the others were set to INDIFFERENT. In (Bárdossy et al., 2002), the fuzzy rules were instead a parameter to optimize. The evaluation of the statement associated with class $k, \forall k \in \{1, ..., K\}$ on variable $i, \forall i \in \{1, ..., I\}$, is done by the way of membership functions $f_{k,i} : [0, 1] \longrightarrow [0, 1]$. Membership functions take as argument the normalized map variable of index i, $\boldsymbol{X}_{d,i}$, and return a value reflecting how true the statement associated with class k is for that variable. In (Bardossy et al., 1995; Bárdossy et al., 2002), there are 5 possible membership functions, one for each of the 5 possible statements attributed. They are triangular fuzzy numbers, i.e. functions of the form:

$$f(x) = \begin{cases} \frac{x-a}{b-a} \text{ if } a \leq x \leq b \\ \frac{c-x}{c-b} \text{ if } b \leq x \leq c \\ 0 \text{ otherwise} \end{cases} \tag{2.7}$$

with three parameters (a, b, c). The choices of parameters made for each statement in (Bardossy et al., 1995; Bárdossy et al., 2002) are shown in table 2.1.

Statement	Fuzzy triangular number parameters
VERY LOW	$(0, 0, 0.4)$
MEDIUM LOW	$(-0.2, 0.2, 0.5)$
MEDIUM HIGH	$(0.5, 0.8, 1.2)$
VERY HIGH	$(0.6, 1, 1)$
INDIFFERENT	$\mu_{k,i}$ is the constant 1

Table 2.1 – Triangular fuzzy number parameters chosen to mathematically represent each statement in (Bardossy et al., 1995; Bárdossy et al., 2002).

Arbitrariness in the choices of thresholds for the triangular functions is unavoidable.

To classify an input field $\boldsymbol{X}_d$, one must compute its "degree of fulfillment" of the properties of class k, $\mu_k(\boldsymbol{X}_d)$, for each class. This degree of fulfillment serves as soft cluster attributions of the datapoints. For each variable $i \in \{1, ..., I\}$, the associated statement are evaluated by the corresponding membership function $f_{k,i}$. Then, the $f_{k,i}(\boldsymbol{X}_{d,i})$ are agglomerated for all i to compute the degree of fulfillment. The exact equation used to obtain the degree of fulfillment can vary depending on user choice. Generally, the membership functions are first agglomerated by statement type (except INDIFFERENT which can be safely ignored), then the 4 remaining values are agglomerated into the degree of fulfillment. That is done by the use of AND: $a, b \longmapsto a \times b$, OR: $a, b \longmapsto a + b - ab$, or weighted linear combinations of the two. Thus, a degree of fulfillment is obtained for every class k, functioning as a soft cluster attribution.

2.5.2 . Fuzzy C-means

In the previous subsection, we have seen a soft clustering method based on the truth value of fuzzy rules. There also exist soft clustering methods that are based on functions to optimize, as we have seen in hard clustering, but do so using fuzzy (soft) cluster attributions (Yang, 1993). Such methods have seen limited application to the field of atmospheric circulation pattern classification, but are presented here for the sake of exhaustivity.

A representative method of function optimization-based soft clustering is fuzzy C-means (FCM) (Suganya and Shanthi, 2012). FCM is a soft clustering counterpart to k-means, based on the idea of fuzzy logic (Zadeh, 1965). Compared to k-means, it adds the possibility of representing datapoints that are intermediary between clusters. Cluster attributions can be partial and datapoints spread between clusters. FCM was initially described in (Dunn, 1973), and an algorithm was proposed in (Bezdek, 2013). Similarly to k-means, the goal of FCM is to optimize for low within-cluster variance, and maximal between-cluster variance, but with a soft clustering formulation of variance.

Let $\mu_{d,k}, \forall k \in \{1, ..., K\}$ denote the cluster attribution of datapoint d to cluster k, also called the degree of membership. It is a number between 0 and 1, such that $\forall d \in \{1, ..., D\}, \sum_{k=1}^{K} \mu_{d,k} = 1$. FCM seeks to optimise the following:

$$\sum_{d=1}^{D} \sum_{k=1}^{K} \mu_{d,k}^{m} ||\boldsymbol{P}_d - S_k||_2 \tag{2.8}$$

with $m \in [1, \inf)$, a metaparameter, representing a degree of fuzziness. We present in algorithm 3 a standard algorithm for FCM, adapted from (Yang, 1993). To summarize, at each step, the cluster centroids move to the average

Algorithm 3 Fuzzy c-means learning process

Initialize degrees of memberships $\mu_{d,k}$
while hasn't converged **do**
  for k in $\{1, ..., K\}$ **do** ▷ Update step
    Update centroids to $S_k = \frac{\sum_{d=1}^{D} \mu_{d,k}^{m} \boldsymbol{X}_d}{\sum_{d=1}^{D} \mu_{d,k}^{m}}$
  end for
  for d in $\{1, ..., D\}$ **do** ▷ Assignment step
    Assign degrees of membership $\mu_{d,k} = \left(\sum_{k'=1}^{K} \left(\frac{||\boldsymbol{X}_d - S_k||_2}{||\boldsymbol{X}_d - S_{k'}||_2} \right)^{\frac{2}{m-1}} \right)^{-1}$
  end for
end while

of all its assigned datapoint, weighted by their degrees of membership. Then, the degrees of membership are computed, based on their relative distance to each centroid, and the value of m. As was the case for k-means, many variations of this algorithm exist, some more efficient than it.

Several generalizations and modifications of the FCM objective and algorithm exist (Yang, 1993). One particular generalization of FCM consists in adding an entropy-based regularization term to the objective function (Honda et al.,

2014). This particular FCM generalization can be demonstrated to be identical to a clustering method based on a probabilistic description of the process of data creation, where each datapoint is assumed to come from one of K distinct Gaussian distributions with different parameters, corresponding to each cluster (Ichihashi et al., 2001). This type of clustering methods, based on a probabilistic description of the data generation process, with each datapoint coming from one of K separate probability distribution each modelled by one of the clusters, are called mixture models and are detailed in the next section.

2.5.3 . Mixture models

We now look at a different kind of soft clustering methods: mixture models (Lindsay, 1995). This is the type of method that this thesis research work is based on. Mixture models are statistical models (McCullagh, 2002). This means that their aim is to represent, in the form of probability distributions, the process of datapoint generation (Cox, 2006). Mixture models, in particular, assume that there exists a number K of distinct, independent processes of datapoint generation that are each associated to one of the K classes. They are described by the probability distributions $p_k, \forall k \in \{1, ..., K\}$. The general probability distribution of the data is therefore expressed as a weighted average, i.e. mixture, of those other probability distribution, with the weights corresponding to the probabilities of a datapoint being generated by each of the K processes. For a datapoint $\boldsymbol{P}_d$:

$$p(\boldsymbol{P}_d) = \sum_{k=1}^{K} m_k p_k(\boldsymbol{P}_d) \tag{2.9}$$

with m_k denoting the prior probability of a datapoint being generated by the process k. To ensure that $p(\boldsymbol{P}_d)$ is a probability function, we necessarily have $\sum_{k=1}^{K} m_k = 1$. Each of the probability distributions p_k is then written as a function of a set of parameters $\boldsymbol{\theta}_k$: $p_k(\boldsymbol{P}_d, \boldsymbol{\theta}_k)$.

In order to classify datapoints, the model must be fitted to the dataset. The parameters, $(\boldsymbol{\theta}_k)_{k\in\{1,...,K\}}$ and $(m_k)_{k\in\{1,...,K\}}$, are optimized to maximize the probability of the data (or a lower bound estimate if it is intractable) using any optimization process such as gradient descent. Then, a datapoint can be attributed a probability of belonging to each class by using Bayesian inference.[1] If one needs categorical class attributions, they can take the values of k that maximize the probability of belonging. The most commonly used mixture model is the Gaussian mixture model, where all of the p_k are assumed Gaussian. This is the type of mixture model proven to result in an identical optimization process than regularized FCM, see (Ichihashi et al., 2001). But there

1. See chapter 3 for more on statistical models and details on Bayesian inference.

exists a great variety of possible mixture models ; using distinct types of probability distributions as class descriptors. The advantage of the mixture model classification method is that class attributions provide information on the statistical properties of the data. Oftentimes, in contexts where mixture models are used, the values of the fitted parameters, and the shapes of the distributions associated with each class, are of more interest than the attributions themselves. The downside of this method is that it requires assumptions on the shape of the distributions associated with each class.

There are relatively few applications of mixture models to atmospheric circulation classification. They were used in (Vrac et al., 2007) to identify seasonal circulation patterns over north America. This methodology is included in the latest versions of the methodological intercomparison project COST733 (Philipp et al., 2016), where it was proven to be able to lead to useful results.

2.6 . Summary

Atmospheric circulation can be described and studied through high-dimensional data. Attempting to classify daily maps into a finite set of circulation patterns, or classes, is one way to deal with the high dimensionality. It makes it possible to measure variations in time of frequencies of appearances of the patterns, or to measure their correlations with other observables or events. We have given in this chapter an overview of the many different existing methods of classifying atmospheric circulation patterns.

Most of these methods require preprocessing of the data to avoid linear dependencies in the variables (which can induce biases) and reduce the number of dimensions. This preprocessing is usually done by the application of Principal Component Analysis (PCA, alternatively referred to as POD or EOF analysis), which extracts from a dataset a set of linearly independent components that maximize explained variance.

One common way to categorize atmospheric circulation is to use weather regimes, defined as states of maximal persistence, frequency, or stationarity. Weather regimes can be a very efficient way to describe the low-frequency, large scale aspects of circulation. However, in what follows, we focus on methodologies that are adapted to the classification of individual maps on a daily timescale.

Several different classification methods have been conceived, based on different criteria to optimize. Such methods include catalogues made by experts

for manual classification, designed specifically to facilitate weather forecast. Many others are automated methodologies that optimize for good categorizations, along user-defined metrics. We first look at hard clustering, a set of methods whose goal is to provide unique class attributions to any pressure map. More specifically, it seeks to organise data into clusters such that the datapoints sharing a cluster are as similar as possible, while being as dissimilar as possible to datapoints in other clusters. Those include hierarchical clustering, which constructs nested structures of clusters, or k-means, which groups datapoints according to cluster centroids they are the closest to. By applying it to a transposed data matrix and rotating the obtained components, PCA itself can also be used as a clustering algorithm itself. Another popular method, Self-Organizing Maps, is a neural network that produces clusters endowed with an underlying ordered structure, facilitating physical interpretation.

However, there also exists a different kind of classification methodology that does not categorically attributes clusters to maps, but instead provides cluster attributions in the form of numbers between 0 and 1. They are called soft clustering methods, and are able to represent more information on the properties of datapoints. For example, fuzzy rules-based clustering handles imprecise conditions for cluster belonging, and measures degrees of fulfillment of these conditions. Other methodologies of fuzzy clustering utilize partial levels of cluster membership to optimize some criteria. This includes fuzzy c-means, a fuzzy counterpart to k-means that allows for each datapoint to be spread over several clusters. On the other hand, mixture models are statistical models, using a probabilistic representation of the process of data generation. They assume the existence of underlying classes with distinct statistical properties, and estimate probabilities of a map being a part of each cluster. However, in their applications in the domain of circulation patterns classifications, these methods are generally used as categorical classifications by selecting for each datapoint the cluster with highest degree of fulfillment or probability. In this thesis work, we will use statistical models for soft clustering as a way to extract more information from pressure-related data.

3 - Introduction of LDA through topic modelling

3.1 . Introduction

The goal of our work is to study atmospheric circulation through an approach based on clustering. As seen in the previous chapter, most clustering methods aim to attribute clusters to maps *categorically*, i.e. with cluster attributions that are both *Boolean* and *unique*. However, there exist "soft clustering" methodologies, that provide cluster attributions in the form of a number between 0 and 1 associated with each cluster, representing a form of *partial* attribution. In particular, statistical models are models that represent the process of data generation in a probabilistic way. In mixture models, the clusters are defined as distinct probabilistic processes of data generation. These models provide soft cluster attributions to datapoints in the form of a posteriori probabilities of them having been generated through the processes associated with each cluster.

In this chapter, our goal is to find a statistical model method to provide soft clustering of the data, in a way that provides meaningful and physically interpretable information. That method is Latent Dirichlet Allocation. In order to clarify why it is the model we have chosen to use, we will introduce it progressively by presenting several intermediary statistical models in increasing order of complexity. In order to facilitate their introduction, we start here by introducing the graphical model formalism. The models are introduced in the original domain of application of LDA - natural language processing - and more specifically, topic modelling. We also go over the notion of Bayesian inference, i.e. how to optimize model parameters from the data, and detail two possible algorithms for optimization of LDA: Gibbs sampling, and variational inference. The information contained in this chapter is mainly based on (Jordan, 2004), (Cowell et al., 2007), (Blei et al., 2003), and (Yvon, 2022).

3.2 . Graphical models

3.2.1 . Networks and probability distributions

The relationships of interdependencies among random variables can be modelled using graphical model representation (also called Bayesian networks). Therefore, we will begin by introducing notions on the study of graphs. A network, or graph, G is defined by a set of nodes, called $\mathbb{V}$, and a set of edges $(u, v) \in \mathbb{V} \times \mathbb{V}$, called $\mathbb{E}$. Such a network is directed if each edge is associated

with a direction. $(u, v) \in \mathbb{E} \not\Leftrightarrow (v, u) \in \mathbb{E}$

A path of length l in G is a sequence of edges $(u_1, v_1), ..., (u_l, v_l)$ such that $\forall 0 < i < l, (u_i, v_i) \in \mathbb{E}$ and $u_{i+1} = v_i$. A network is acyclic, if there exists no path of nonzero length ending on the same node that it starts. G acyclic $\Leftrightarrow \forall$ path in $G, l = 0$ or $u_1 \neq v_l$.

Let π_u denotate the set of parents of a node $u \in \mathbb{V}$. It is defined as the set of all nodes $v \in \mathbb{V}$ such that $(v, u) \in \mathbb{E}$, i.e. with an edge going from v to u.

Graphical models are acyclic directed graphs, where each node is associated to a random variable. Let variables $(x_i)_{i\in\{1,...,I\}}$ be a set of random variables associated with the nodes of $G = (\mathbb{V}, \mathbb{E})$. Then, the joint probability distribution of the $(x_i)_{i\in\{1,...,I\}}$ is of the form:

$$p((x_i)_{i\in\{1,...,I\}}) = \prod_{x_i}^{\mathbb{V}} p(x_i|\pi_{x_i}) \tag{3.1}$$

Thus, a graph corresponds to a family of joint probability distributions over the random variables. From now on, the nodes of a graph and the associated random variables will be identified: $(x_i)_{i\in\{1,...,I\}} = \mathbb{V}$.

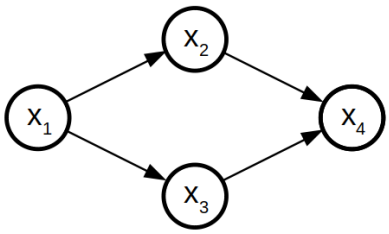


Figure 3.1 – Example of a directed, acyclic graph. Each node corresponds to a random variable.

As an example, an acyclic, directed graph is represented in figure 3.1. According to the graphical model formalism, this graph implies that the joint probability distribution of the variables $x_1, ..., x_4$ is of the form described in equation3.2.

$$p(x_1, x_2, x_3, x_4) = p(x_4|x_3, x_2)p(x_3|x_1)p(x_2|x_1)p(x_1) \tag{3.2}$$

In simpler terms, what a graph tells you is that the probability distribution of a random variable is only and fully defined by the value of the variables corresponding to its parent nodes.

A graphical model also represents a generative process, i.e. a process according to which new values of the random variables can be generated. First, draw every variable associated to a node without parents. Then, draw every variable whose parent nodes have all been drawn. Taking again the example of figure 3.1, we would first draw the value of x_1 from its probability distribution $p(x_1)$, which does not depend on any of the other variables. Then, we can draw x_2 and x_3 from $p(x_2|x_1)$ and $p(x_3|x_1)$ respectively, using the value obtained from x_1. Finally, we can draw x_4 from $p(x_4|x_3, x_3)$, using the values of x_2 and x_3.

To simplify the notations of graphs with high number of nodes, we introduce the plate notation, see figure 3.2. Sets of nodes playing similar roles in the structure of the graph can be represented within a plate, functioning similarly to a "for" loop.

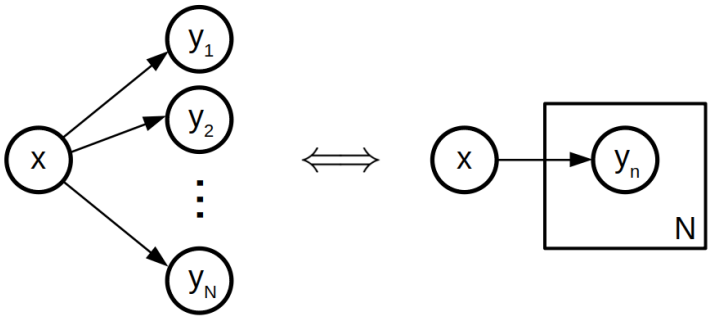


Figure 3.2 – Example of a graph illustrating the plate notation.

When studying systems using graphical models, it can be useful to differentiate between hidden variables, and observed variables, the values of which are considered known. To that end, observed variables are shown with a grayed node. See figure 3.3 for an example.

3.2.2 . Inference in graphical models

Let $G = (\mathbb{V}, \mathbb{E})$ a Bayesian network. Let $\mathbb{X} \subset \mathbb{V}$ and $\mathbb{Y} \subset \mathbb{V}$ be two disjoint sets of variables, with $\mathbb{X}$ non-empty. The problem of inference consists in finding the conditional probability distribution $p(\mathbb{X}|\mathbb{Y})$, i.e. finding the distribution over $\mathbb{X}$ when the variables in $\mathbb{Y}$ are observed. Solving the inference problem is key to fitting statistical models to observed data.

By summing over all possible realizations of a random variable, we can elim-

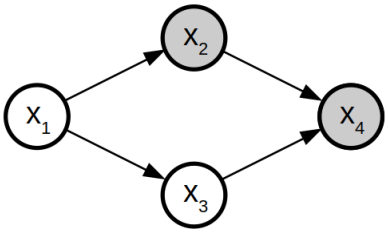


Figure 3.3 – Example of a graph displaying two observed variables (x_2 and x_4, in gray), and two unobserved variables (x_1 and x_3, in white).

inate this variable from the equations. This is called marginalisation of the joint probability distribution.

$$p(x) = \sum_{y_0} p(x, y = y_0) \tag{3.3}$$

To shorten notations, summing over all possible realizations of a random variable y will be written by putting the random variable as index of the sum: $p(x) = \sum_y p(x, y)$. If the set of possible realizations of y is continuous, and not discrete, then we write $p(x) = \int_y p(x, y)\, dy$ instead.

By definition of conditional probability, $p(x|y) = \frac{p(x,y)}{p(y)}$. Using marginalisation, we can write:

$$p(\mathbb{X}|\mathbb{Y}) = \frac{p(\mathbb{X}, \mathbb{Y})}{\sum_{\mathbb{X}} p(\mathbb{X}, \mathbb{Y})} \tag{3.4}$$

Both the numerator and the denominator can be rewritten as marginalisations of the joint probability distribution $p(\mathbb{V})$:

$$p(\mathbb{X}|\mathbb{Y}) = \frac{\sum_{\mathbb{V}\setminus\mathbb{X}\setminus\mathbb{Y}} p(\mathbb{V})}{\sum_{\mathbb{X}} \sum_{\mathbb{V}\setminus\mathbb{X}\setminus\mathbb{Y}} p(\mathbb{V})} \tag{3.5}$$

Where $\sum_{\mathbb{X}}$ designates summing over all possible realizations of all variables in $\mathbb{X}$, and $\setminus$ designates set subtraction. The solution to the inference problem is obtaining a expression for this formula. In some cases this can easily be done, is others it can be analytically intractable.

In Bayesian networks, it is possible to simplify this computation process thanks to the constraints imposed on the form of $p(\mathbb{V})$. Let $u, v \in \mathbb{V}$. u and v are independent $\iff p(u, v) = p(u)p(v)$. In such a case, knowing the realization of one of the independent variables has no effect on the distribution of the other. It can be shown that the form taken by the probability distribution of variables in a Bayesian network (equation 3.1) implies the following relationships of independence. Let $u, v \in \mathbb{V}$ such that $u \notin \pi_v, v \notin \pi_u$. If the variables

π_u are observed, then u and v are independent. In many cases, these relationships can be used to simply the inference computations.

3.3 . Topic modelling

3.3.1 . Introduction

Topic modelling consists in building representations of the underlying word patterns in collections of text documents, which are sequences of words. Typically, this is done with statistical models by using latent variables, referred to as "topics" (Blei et al., 2010). Each topic represents a different probability distribution over words, generally containing words which commonly appear in the same documents. For example, a topic related to sports would attribute high probabilities to words like "athlete" or "training", and a lower probability to words such as "stratosphere" or "wiring". This section is based on the works of David Blei in (Blei et al., 2003) and François Yvon in (Yvon, 2022).

Let $(X_d)_{d\in\{1,...,D\}}$ be the corpus of documents. Each document X_d is defined by a sequence of N_d words: $(w_{d,n})_{d\in\{1,...,D\},n\in\{1,...,N_d\}}$. The set of all possible words is referred to as the vocabulary, and its size is noted V. In order to simplify the problem, we consider the order of the words to be irrelevant. This is referred to as the "bag of words" hypothesis. A document is only defined by the number of times each word appears in it. In practice, it means we are only interested in the patterns of occurrences of words in documents, and not in the patterns of sequences of words.

3.3.2 . Unigram model

One of the simplest models we can make is to assume each word of each document is drawn from the same categorical probability distribution defined over the vocabulary. This is called the unigram model, and would be described by the graph shown in figure 3.4, and equation 3.6.

$$p((w_{d,n})_{d\in\{1,...,D\},n\in\{1,...,N_d\}}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{d,n}) \tag{3.6}$$

The unigram model has one parameter: that of the probability distribution over the vocabulary $p(w = v), \forall v \in \{1, ..., V\}$. Since the vocabulary $v \in \{1, ..., V\}$ is a finite discrete set, this distribution is a categorical distribution. Its parameter is a normalized vector, each element of it being the probability associated to the corresponding element in the set. For the unigram

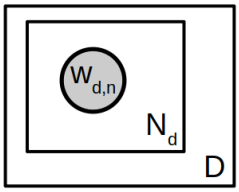


Figure 3.4 – Graphical model representation of the unigram model. The D-plate represents iteration over all documents, and the N_d-plate iteration over words.

model, this parameter is noted $\boldsymbol{\beta}$, and is a vector of length V, such that $\forall v \in \{1, ..., V\}, p(w = v) = \boldsymbol{\beta}_v$.

Every statistical model is also a generative model, able to be used to generate new documents. For the generation process, the model parameters are assumed known. In the unigram model, the generation process is as shown in algorithm 4. The symbol $\sim$ means: "Is drawn from the probability distribution parameterized by".

Algorithm 4 Unigram model document generation

Pick a document length N
for n in $\{1, ..., N\}$ **do**
  $w_n \sim \boldsymbol{\beta}$
end for

Note that the model does not innately provide a way to pick a document length N. This will be the case for all the models presented in this section. One way of doing it is to draw N from a Poisson law fitted on the lengths of documents in the corpus $(N_d)_{d \in \{1,...,D\}}$. While an integral part of document generation, modelling document length is a distinct problem from topic modelling, and will not be explored further in this work.

3.3.3 . Mixture model

The unigram model captures the frequencies of each word, but it does not represent the 2-variable statistics, such as correlations between words. In order to increase the complexity of the model, we introduce a document-level latent variable, called "topic attribution". Topic attributions are noted $\boldsymbol{z}$, and are a vector of length D. The topic attribution of document d, $\boldsymbol{z}_d$, is a number in $\{1, ..., K\}$. Qualitatively, $\boldsymbol{z}_d$ represents which, among a set of K different

topics, best describes the distribution of words observed in the document. The number of topics K is a hyperparameter of the method, equivalent to a number of clusters. This model is a discrete mixture model (Nigam et al., 2000) (see chapter 2). It is described by the graph shown in figure 3.5, and the corresponding equation 3.7.

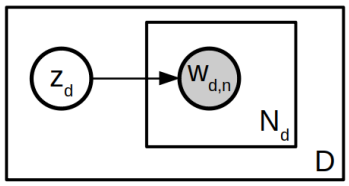


Figure 3.5 – Graphical model representation of the mixture model. The D-plate represents iteration over all documents, and the N_d-plate iteration over words.

$$p((w_{d,n})_{d\in\{1,...,D\},n\in\{1,...,N\}}, \boldsymbol{z}) = \prod_{d=1}^{D} p(\boldsymbol{z}_d) \prod_{n=1}^{N_d} p(w_{d,n}|\boldsymbol{z}_d) \tag{3.7}$$

Instead of assuming that each word is drawn from the same probability distribution, the mixture model assumes the existence of K distinct topics, each associated to a different probability distribution on the vocabulary. Each document is then assumed to have been generated following the probability distribution associated with one of the topics. The parameters of the mixture model are noted $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. In the mixture model, the parameter $\boldsymbol{\beta}$ is a matrix of shape $K \times V$, with each row $\boldsymbol{\beta}_k, \forall k \in \{1, ..., K\}$, the parameter of the categorical distribution over the vocabulary associated with the k-th topic. $\forall k \in \{1, ..., K\}, \forall v \in \{1, ..., V\}, p(w = v|z = k) = \boldsymbol{\beta}_{k,v}$. $\boldsymbol{\alpha}$ is a vector of length K, parameter of the probability distribution over topic attributions, which is also a categorical distribution. $\forall k \in \{1, ..., K\}, p(z = k) = \boldsymbol{\alpha}_k$. In the mixture model, the generation process is as shown in algorithm 5.

Algorithm 5 Mixture model document generation

Pick a document length N
$z \sim \boldsymbol{\alpha}$
for n in $\{1, ..., N\}$ **do**
    $w_n \sim \boldsymbol{\beta}_z$
end for

3.3.4 . pLSI

The mixture model is capable of representing more detail than the unigram model, but it still has limitations. It assumes that a document is always defined by a single topic, and does not allow for documents sharing a topic to have differing statistical properties. We want the ability to represent documents that contain a mix of different topics. To that end, topic attribution $\boldsymbol{z}_d$ is changed to be a word-level variable $\boldsymbol{z}_{d,n}$, and each document is associated a composition in topics.

Probabilistic Latent Semantic Indexing (pLSI), is a model that functions this way. It considers each document index d to be associated with its own probability distribution over topics. pLSI is described by the graph shown in figure 3.6, and the corresponding equation 3.7.

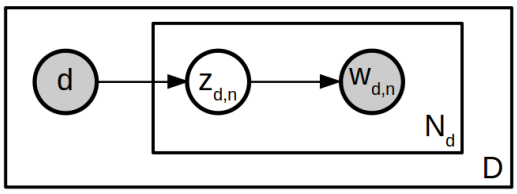


Figure 3.6 – Graphical model representation of pLSI. The D-plate represents iteration over all documents, and the N_d-plate iteration over words.

$$p((w_{d,n})_{d\in\{1,...,D\},n\in\{1,...,N\}}, \boldsymbol{z}|d) = \prod_{d=1}^{D}\prod_{n=1}^{N_d} p(\boldsymbol{z}_{d,n}|d)p(w_{d,n}|\boldsymbol{z}_{d,n}) \tag{3.8}$$

pLSI has some specificities compared to the other models seen thus far. It uses d, the index of documents in the corpus, as one of the variables, without modelling a $p(d)$. In effect, when fitted, it associates to each document d in the learning corpus its own probability distribution over topic attributions. The parameters of pLSI are noted $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. $\boldsymbol{\theta}$ is a matrix of shape $D \times K$, parameterizing the document-topic distributions. Each row $\boldsymbol{\theta}_d$ is the parameter of the categorical distribution over topic attributions associated with document d, such that $\forall d \in \{1, ..., D\}, \forall k \in \{1, ..., K\}, p(z = k|d) = \boldsymbol{\theta}_{d,k}$. Similarly to the mixture model case, $\boldsymbol{\beta}$ is a matrix of shape $K \times V$, such that $\forall k \in \{1, ..., K\}, \forall v \in \{1, ..., V\}, p(w = v|z = k) = \boldsymbol{\beta}_{k,v}$.

pLSI does not exactly generate new documents, but it describes the process of generating a new instance of a document at a given index d. This process

Algorithm 6 pLSI document generation

Pick a document index $d \in \{1, ..., D\}$
Pick a document length N
for n in $\{1, ..., N\}$ **do**
  $z_n \sim \boldsymbol{\theta}_d$
  $w_n \sim z_n$
end for

is described in algorithm 6.

Due to its particularity, pLSI comes with the drawback of being unable to handle documents that are not part of the learning corpus. All other models described here can, given new documents, determine values for their latent variables using Bayesian inference, without needing to re-fit the model. But adding a new document to pLSI would imply a new document index $d = D+1$, and therefore new model parameters $\boldsymbol{\theta}_{D+1}$ that would need to be fitted. Furthermore, the amount of values contained in the model parameter $\boldsymbol{\theta}$ grows linearly with the amount of documents in the corpus, which can become impractical for large datasets.

For these reasons, we introduce a new model that shares the structure of a word-level latent variable describing topic attribution, and a document-level latent variable that describes topic composition, but without involving the document index. This model is Latent Dirichlet Allocation.

3.4 . Latent Dirichlet Allocation

3.4.1 . Introduction

Latent Dirichlet Allocation (LDA) is an unsupervised statistical learning method introduced by David Blei in (Blei et al., 2003). Like the models presented above, its purpose is to model the statistical relationships between words in a corpus of text documents, using latent variables known as "topics", characterized by a probability distribution over the vocabulary. The number of topics K is a hyperparameter of the method, equivalent to a number of clusters. LDA is a soft clustering technique: each of the D documents in the corpus is associated with a distribution over the topics. LDA treats documents under the bag of words hypothesis, which assumes that the ordering of words in documents is irrelevant.

The graph representing the structure of LDA is shown figure 3.7, and the as-

sociated equation is equation 3.9. From now on, for the sake of simpler notations, we will abuse of the vector notation by referring to $(w_{d,n})_{d\in\{1,...,D\},n\in\{1,...,N_d\}}$ as $\boldsymbol{w}$ as if it were a matrix, despite the lengths of each row, N_d, not being equal in general.

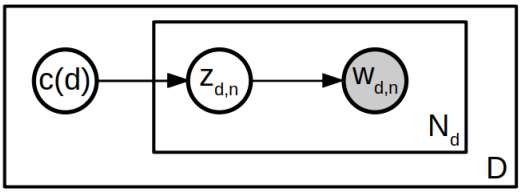


Figure 3.7 – Graphical model representation of LDA. The D-plate represents iteration over all documents, and the N_d-plate iteration over words.

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}) = \prod_{d=1}^{D} p(\boldsymbol{c}(d)) \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{c}(d)) p(w_{d,n}|z_{d,n}) \tag{3.9}$$

LDA introduces a prior probability distribution on the topic composition associated to a document d. This topic composition $\boldsymbol{c}(d)$ is the parameter of the document-topic categorical distribution. Therefore, it is a vector of length K summing to 1. Since the value of $\boldsymbol{c}$ can be inferred for any document once the model is fitted using Bayesian inference, including documents outside training data, $\boldsymbol{c}$ will be treated as a function defined on documents. This is why it is denoted $\boldsymbol{c}(d)$ and not $\boldsymbol{c}_d$. The set of possible topic composition describes a simplex of dimension $K-1$. A simple example, with only $K=3$ topics, is shown figure 3.8.

In LDA, the prior probability distribution on the topic composition associated to a document is assumed to be a Dirichlet distribution, of parameter $\boldsymbol{\alpha}$. Thus, LDA has two parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. $\boldsymbol{\beta}$ is a matrix of shape $K \times V$ governing the topic-word categorical distributions associated with each topic. Each row $\boldsymbol{\beta}_k, \forall k \in \{1, ..., K\}$, is the parameter of the categorical distribution over the vocabulary associated with topic k. $\boldsymbol{\alpha}$ is a vector of length K, parameter of the Dirichlet distribution over the topic composition of documents. There is no additional parameter governing the categorical distribution over topic attribution $p(z_{d,n}|\boldsymbol{c}(d))$, because the parameter of this law is $\boldsymbol{c}(d)$.

LDA is also a generative model, capable of generating new documents from the fitted model parameters. The generation process of LDA is as shown in

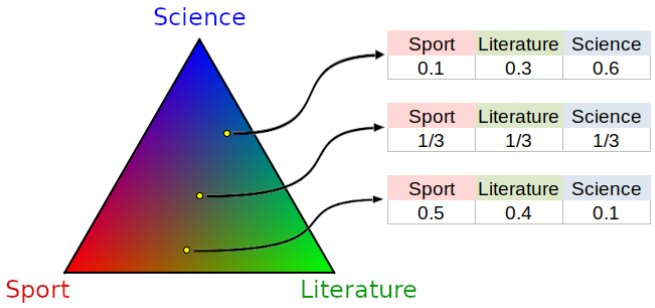


Figure 3.8 – Example representation of the space of possible topic compositions, on which we define the Dirichlet probability distribution parameterized by α.

Algorithm 7 LDA document generation

Pick a document length N
$\boldsymbol{c} \sim \boldsymbol{\alpha}$
for n in $\{1, ..., N\}$ **do**
  $\boldsymbol{z}_n \sim c$
  $\boldsymbol{w}_n \sim z_n$
end for

algorithm 7.

While LDA sees use in natural language processing, it is also applied in several other fields of science (Jelodar et al., 2019). Notably, text mining techniques have direct applications to fields such as social sciences, as well as political sciences, which are also concerned with the study of large quantities of natural language. See (Gross and Murthy, 2014) for a social science example, regarding the search of organizations in topics in social network exchanges, and (Greene and Cross, 2015) for a political science example, regarding the time evolution of political terms and topics in European parliament debates. However, LDA can also be applied to data other than text corpora. For instance, it has been applied in biogeography, to study the spatial distribution of the presence of species of birds (Valle et al., 2018). Finally, closer to what is done in this thesis work (see chapter 4), there are applications to fluid mechanics, where it is used to find structures within turbulent flows (Frihat et al., 2021; Podvin et al., 2024).

3.4.2 . Joint distribution

To facilitate a more detailed examination of LDA, we present a graphical model representation including the model parameters in figure 3.9.

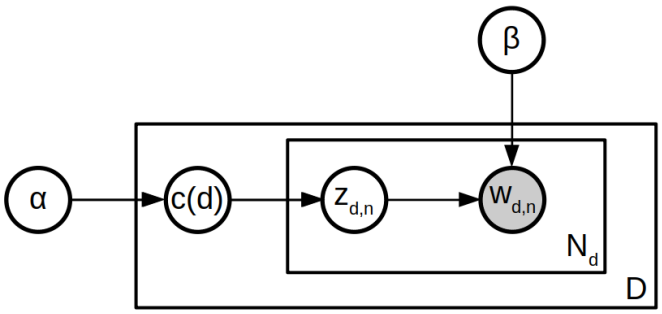


Figure 3.9 – Graphical model representation of LDA, including model parameters. The D-plate represents iteration over all documents, and the N_d-plate iteration over words.

The joint distribution of all variables in LDA, knowing the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, is equation 3.10

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p(\boldsymbol{c}(d)|\boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{c}(d)) p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}) \tag{3.10}$$

where

- $\boldsymbol{c}(d)$ is drawn from the Dirichlet distribution of parameter $\boldsymbol{\alpha}$:

$$p(\boldsymbol{c}(d)|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} (\boldsymbol{c}_k(d))^{\boldsymbol{\alpha}_k - 1}, \qquad B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\boldsymbol{\alpha}_k)}{\Gamma(\sum_{k=1}^{K} \boldsymbol{\alpha}_k)} \tag{3.11}$$

- $z_{d,n}$ is drawn from the categorical distribution parameterized by $\boldsymbol{c}(d)$:

$$p(z_{d,n} = k|\boldsymbol{c}(d)) = \boldsymbol{c}_k(d) \tag{3.12}$$

- $w_{d,n}$ is drawn from the categorical distribution parameterized by $\boldsymbol{\beta}_{z_{d,n}}$:

$$p(w_{d,n} = v|z_{d,n}, \boldsymbol{\beta}) = \boldsymbol{\beta}_{z_{d,n},v}. \tag{3.13}$$

3.4.3 . Smooth LDA

Ideally, when the model is fit to the data, the $\boldsymbol{\beta}_k$ should reflect the frequencies of apparition of words within in each topic. This can be a problem because

of the strong sparsity of natural language data: it is very common for documents to contain words that appear in no other. Words that do not appear in the learning corpus, and by extension documents containing them, would therefore be assigned a probability of 0. That would render meaningful inference of latent parameters for such documents impossible. To solve this problem, according to (Jelinek, 1998), we must "smooth" the parameters $\boldsymbol{\beta}$ to ensure that every word of the vocabulary is assigned a nonzero probability, even if it does not appear in the learning corpus. In this case, the smoothing of $\boldsymbol{\beta}$ is achieved by assuming the distribution parameter of each topic, the $\boldsymbol{\beta}_k$, is obtained from a Dirichlet prior of parameter $\boldsymbol{\eta}$.

$$\forall k \in \{1, ..., K\}, p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) = \frac{1}{B(\boldsymbol{\eta})} \prod_{i=1}^{V} (\boldsymbol{\beta}_{k,i})^{\boldsymbol{\eta}_i - 1} \tag{3.14}$$

The full graphical model representation of smoothed LDA is represented in figure 3.10. This is the version of LDA we use in the rest of this work.

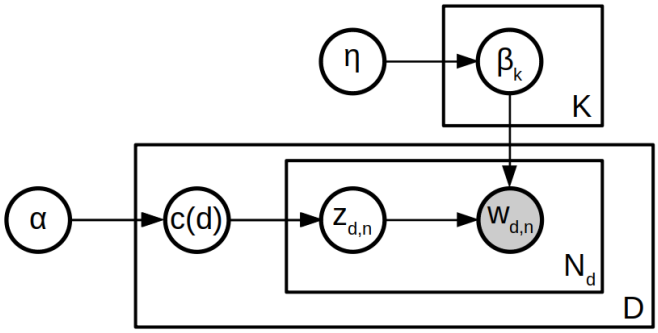


Figure 3.10 – Graphical model representation of smoothed LDA, including model parameters. The D-plate represents iteration over all documents, the N_d-plate iteration over words, and the K-plate iteration over topics.

3.5 . Learning for LDA

3.5.1 . Inference for model parameters

Analysis of a corpus of documents with LDA consists in examining the posterior distribution of the topics $\boldsymbol{\beta}$, topic proportions $\boldsymbol{c}$, and topic assignments $\boldsymbol{z}$. To apply LDA in practice, we use the Python module Gensim (Řehůřek and Sojka, 2010), whose implementation is based on the works of M. Hoffman, D. Blei and F. Bach. in (Hoffman et al., 2010), itself based on (Blei et al., 2003). The methodology presented here is from these articles.

Using conditional probabilities, we can write:

$$p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})}$$

Let us obtain an expression for the denominator, $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})$. To that end, we write the joint probability law and then marginalize over $\boldsymbol{\beta}$, $\boldsymbol{c}$, and $\boldsymbol{z}$, by integrating over all possible values of $\boldsymbol{\beta}$, $\boldsymbol{c}$, and summing over all possible values of $\boldsymbol{z}$.

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\boldsymbol{\beta}} \int_{\boldsymbol{c}} \sum_{\boldsymbol{z}} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\eta})\, d\boldsymbol{\beta}\, d\boldsymbol{c}$$

Based on 3.10, this takes the form:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\boldsymbol{\beta}} \int_{\boldsymbol{c}} \sum_{\boldsymbol{z}} \left(\left(\prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) \right) \prod_{d=1}^{D} p(\boldsymbol{c}(d)|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(\boldsymbol{z}_{d,n}|\boldsymbol{c}(d), \boldsymbol{\beta}) p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}) \right) d\boldsymbol{c}\, d\boldsymbol{\beta}$$

Which, when developed, yields:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{\Gamma(\sum_{k=1}^{K} \boldsymbol{\alpha}_k)}{\prod_{k=1}^{K} \Gamma(\boldsymbol{\alpha}_k)} \times \int_{\boldsymbol{\beta}} \left(\prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) \right) \int_{\boldsymbol{c}} \sum_{\boldsymbol{z}} \prod_{d=1}^{D} \left(\prod_{k=1}^{K} \boldsymbol{c}_k(d)^{\boldsymbol{\alpha}_k - 1} \right) \left(\prod_{n=1}^{N} \boldsymbol{c}_{\boldsymbol{z}_{d,n}}(d) \boldsymbol{\beta}_{\boldsymbol{z}_{d,n}, \boldsymbol{w}_{d,n}} \right) d\boldsymbol{c}\, d\boldsymbol{\beta} \quad (3.15)$$

The problem here is the coupling of $\boldsymbol{\beta}$ and $\boldsymbol{c}$ in the marginalisation process. Because of this, the likelihood of the data, $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})$, is not tractable in practice. Therefore, $p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta})$ can not be derived through purely analytical means.

3.5.2 . Gibbs sampling

This section presents one method used to overcome this analytical intractability: Gibbs sampling (Casella and George, 1992). While not the method used in our implementation of LDA learning, we detail it here for exhaustivity. The Gibbs sampling method is based on the idea that it is easier to sample from conditional distributions than to marginalize the joint distribution. It functions by generating a sequence of samples, iteratively sampling each variable from its conditional distribution knowing all others. In the end, the intractable joint distribution can be approximated using the set of samples.

The application of Gibbs sampling to generate J samples on a set of I random variables $\mathbb{V} = (x_i)_{i \in \{1,\dots,I\}}$ described by a joint probability $p(\mathbb{V})$ is described by algorithm 8.

Algorithm 8 Gibbs sampling

Initialize variables $\left(x_i^{(0)}\right)_{i\in\{1,...,I\}}$
for $j = 1$ to J **do**
  for $i = 1$ to I **do**
    $x_i^{(j)} \sim p(x_i|x_0^{(j-1)}, ..., x_{i-1}^{(j-1)}, x_{i+1}^{(j)}, ..., x_I^{(j)})$
  end for
end for

To summarize, the process starts by choosing, as random initialisation, a starting value for each random variable of $\mathbb{V}$. Gibbs sampling consists in iteratively sampling each of the random variables, from their conditional distribution knowing all the others variables, using the last sampled value. By repeating this process a large amount of times, we obtain a set of samples for each of the random variables from which the joint probability distribution can be approximated.

Gibbs sampling can be applied to fit a LDA model. The following discussion and algorithm are adapted from (Pochon and Favre, 2022), and concerns a non-smoothed LDA model. Since this is a learning process, words $\boldsymbol{w}$ are considered known and are not sampled. Furthermore, the only variables needing to be sampled are the topic attributions $\boldsymbol{z}$. The variables $\boldsymbol{c}(d), \forall d \in \{1, ..., D\}$, and $\boldsymbol{\beta}$ are directly measured at each given step by looking at the topic attributions of the words $\boldsymbol{z}_{d,n}$. Additionally, if the words $\boldsymbol{w}$ are considered known, it can be demonstrated that:

$$\begin{aligned}\forall k \in \{1, ..., K\},\\ p(\boldsymbol{z}_{d,n} = k|\boldsymbol{w}_{d,n}, \boldsymbol{c}(d), \boldsymbol{\beta}) &\propto p(\boldsymbol{z}_{d,n} = k|\boldsymbol{c}(d)) \times p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}, \boldsymbol{\beta})\\ &= \boldsymbol{c}_k(d) \times \boldsymbol{\beta}_{\boldsymbol{z}_{d,n},\boldsymbol{w}_{d,n}}\end{aligned} \tag{3.16}$$

The process of applying Gibbs sampling to fit a (non-smoothed) LDA model is described in algorithm 9.

Specifically, the estimation of $\boldsymbol{c}(d)^{(j)}$, at document d, at word n, is done based on the frequencies of each topic index in the sampled topic attributions in the document:

$$\forall k \in \{1, ..., K\},$$
$$\boldsymbol{c}_k(d)^{(j)} = \frac{1}{N_d - 1}\left(\sum_{n'=1}^{n-1}\left[\boldsymbol{z}_{d,n'}^{(j)} = k\right] + \sum_{n'=n+1}^{N_d}\left[\boldsymbol{z}_{d,n'}^{(j-1)} = k\right]\right) \tag{3.17}$$

Algorithm 9 Gibbs sampling for LDA

Initialize topic attributions $\boldsymbol{z}^{(0)} = \left(\boldsymbol{z}_{d,n}^{(0)}\right)_{d\in\{1,...,D\},n\in\{1,...,N_d\}}$
for $j = 1$ to J **do**
  for $d = 1$ to D **do**
    for $n = 1$ to N_d **do**
      Estimate $\boldsymbol{c}(d)^{(j)}$ based on the topic attributions of other words in document d.
      Estimate $\boldsymbol{\beta}^{(j)}$ based on the topic attributions of other words in the corpus.
      $\boldsymbol{z}_{d,n}^{(j)} \sim p(z|\boldsymbol{w}_{d,n}, \boldsymbol{c}(d), \boldsymbol{\beta}) \propto \boldsymbol{c}_z(d) \times \boldsymbol{\beta}_{\boldsymbol{z}_{d,n},\boldsymbol{w}_{d,n}}$
    end for
  end for
end for

With $[\cdot]$ designating the Iverson Bracket:

$$[P] = \begin{cases} 1 \text{ if P is true} \\ 0 \text{ otherwise} \end{cases} \tag{3.18}$$

And the estimation of $\boldsymbol{\beta}^{(j)}$, at document d, at word n, is done based on the frequencies of each word with each topic attribution, in the sampled topic attributions in the entire corpus:

$$\begin{aligned} &\forall k \in \{1, ..., K\}, \forall v \in \{1, ..., V\}, \\ &\boldsymbol{\beta}_{k,v}^{(j)} = \frac{1}{\left(\sum_{d'=1}^{D} N_{d'}\right) - 1} \Big(\sum_{d'=1}^{d-1} \sum_{n'=1}^{n_{d'}} \left[\boldsymbol{z}_{d',n'}^{(j)} = k\right] \left[\boldsymbol{w}_{d',n'} = v\right] \\ &\quad + \sum_{n'=1}^{n-1} \left[\boldsymbol{z}_{d',n'}^{(j)} = k\right] \left[\boldsymbol{w}_{d',n'} = v\right] + \sum_{n'=n+1}^{N_d} \left[\boldsymbol{z}_{d',n'}^{(j-1)} = k\right] \left[\boldsymbol{w}_{d',n'} = v\right] \\ &\quad + \sum_{d'=d+1}^{D} \sum_{n'=1}^{n_{d'}} \left[\boldsymbol{z}_{d',n'}^{(j-1)} = k\right] \left[\boldsymbol{w}_{d',n'} = v\right] \Big) \end{aligned} \tag{3.19}$$

3.5.3 . Variational approach

Gibbs sampling is a way around the intractability of the joint distribution. However, it can take a very high amount of samples for Gibbs sampling to provide a satisfying approximation of the joint probability distribution, especially in cases with high amount of variables to samples. We resort instead

to a variational approach (Attias, 1999; Jordan et al., 1999). The basic idea of the variational approach is to estimate an intractable function to optimize by constructing another, strictly inferior, simpler function that depends on additional variables called variational parameters. This function is called the lower bound. By maximizing the lower bound with respect to the variational parameters, we obtain an estimate of the original function to optimize, that can then be optimized with respect to its original variables.

Let q designate a set of probability distributions distinct from the ones described by p. With $q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z})$ an arbitrary probability distribution over $\boldsymbol{c}$, $\boldsymbol{z}$, and $\boldsymbol{\beta}$, we can rewrite the log likelihood of the data, and use Jensen's inequality:

$$\begin{aligned}
\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) &= \log \int_{\boldsymbol{c}} \int_{\boldsymbol{\beta}} \sum_{\boldsymbol{z}} p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) d\boldsymbol{c} d\boldsymbol{\beta} \\
&= \log \int_{\boldsymbol{c}} \int_{\boldsymbol{\beta}} \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z})} d\boldsymbol{c} d\boldsymbol{\beta} \\
&\geq \int_{\boldsymbol{c}} \int_{\boldsymbol{\beta}} \sum_{\boldsymbol{z}} q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}) \log p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}) d\boldsymbol{c} d\boldsymbol{\beta} \\
&\quad - \int_{\boldsymbol{c}} \int_{\boldsymbol{\beta}} \sum_{\boldsymbol{z}} q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}) \log q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}) d\boldsymbol{c} d\boldsymbol{\beta} \\
&= E_q[\log p(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})] - E_q[\log q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z})]
\end{aligned} \tag{3.20}$$

With $E_q[\cdot]$ designating the expectation over the distributions described by q. We have obtained a lower bound for the log likelihood $\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})$. Furthermore, it can be proven that the difference between the two sides of the inequality corresponds to the Kullback-Leibler (KL) divergence between the variational and real posterior distributions. Therefore, we can ensure the tightness of the lower bound by minimizing the KL divergence.

To simplify computations, we choose a fully factorized distribution of the form shown in equation 3.21.

$$q(\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{z}) = \prod_{k=1}^{K} q(\boldsymbol{\beta}_k) \times \prod_{d=1}^{D} q(\boldsymbol{c}(d)) \prod_{n=1}^{N_d} q(\boldsymbol{z}_{d,n}) \tag{3.21}$$

Tractability is ensured by removing all couplings, and choosing either categorical of Dirichlet distributions for the variables. The graph corresponding to the equation, including the variational parameters, is represented in figure 3.11.

- $\boldsymbol{\lambda}$ is a matrix of shape $K \times V$. Each row $\boldsymbol{\lambda}_k, \forall k \in \{1, ..., K\}$ corresponds to the parameter of the Dirichlet distribution $q(\boldsymbol{\beta}_k)$.
- $\boldsymbol{\gamma}$ is a matrix of shape $D \times K$. Each row $\boldsymbol{\gamma}_d, \forall d \in \{1, ..., D\}$ corresponds to the parameter of the Dirichlet distribution $q(\boldsymbol{c}(d))$.

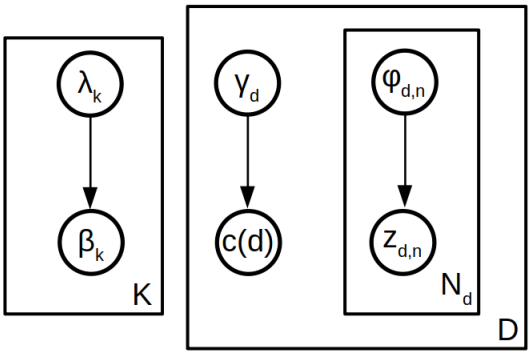


Figure 3.11 – Graphical model representation of the variational parameter scheme for LDA inference. The D-plate represents iteration over all documents, the N_d-plate iteration over words, and the K-plate iteration over topics.

- $\boldsymbol{\phi}$ is a 3D tensor of shape $D \times V \times K$. Each row $\boldsymbol{\phi}_{d,v}, \forall d, v \in \{1, ..., D\}, \{1, ..., V\}$ corresponds to the parameter of the categorical distribution over $\boldsymbol{z}_{d,n}$, knowing $\boldsymbol{w}_{d,n} = v$: $q(\boldsymbol{z}_{d,n} = k) = \boldsymbol{\phi}_{d,\boldsymbol{w}_{d,n},k}$.

From this, by computing the derivatives of the KL divergence with respect to the variational parameters, and setting them equal to zero, update equations are obtained for the parameters. Using the bag of words hypothesis, we reformulate the information of the sequences of words contained in each documents, $\boldsymbol{w}_d$, as word counts $\boldsymbol{X}_d$. $\boldsymbol{X}$ is a $D \times V$ matrix such that $\forall d \in \{1, ..., D\}, \forall v \in \{1, ..., V\}, \boldsymbol{X}_{d,v}$ is the number of times the word v appears in document d. With a corpus of documents described by $\boldsymbol{X}$ fixed and known, we can iterate the application of the update equations on each parameter until convergence. This process, including the update equations, is summarized in algorithm 10, taken directly from (Hoffman et al., 2010). Then, we have a satisfying approximation on the posterior distributions on $\boldsymbol{\beta}$, $\boldsymbol{c}$, and $\boldsymbol{z}$. The variables $\boldsymbol{\beta}$, $\boldsymbol{c}$, and $\boldsymbol{z}$ are then selected to maximize this probability, completing the analysis of the corpus by LDA.

3.5.4 . Online learning

In practice, however, we use a slightly different version of this algorithm. Instead of going over every document at each update step, we only cover each document once.[1] This allows it converge faster for large datasets, and function in contexts where data arrives progressively in time. For this reason, it

1. The number of times each document is covered by online learning can be increased, as a hyperparameter called `passes`, in the gensim implementation.

Algorithm 10 Variational Bayes learning for LDA

Initialize $\boldsymbol{\lambda}$ randomly
while relative improvement to lower bound > 0.00001 **do**
  for $d = 1$ to D **do**
    Initialize $\boldsymbol{\gamma}_{d,k} = 1$. (The constant 1 is arbitrary.)
    while $\frac{1}{K}\sum_k |\text{change in}\boldsymbol{\gamma}_{d,k}| < 0.00001$ **do**
      Set $\boldsymbol{\phi}_{d,v,k} \propto exp\left(E_q[log(\boldsymbol{c}_k(d))] + E_q[log(\boldsymbol{\beta}_{k,v})]\right)$
      Set $\boldsymbol{\gamma}_{d,k} = \boldsymbol{\alpha} + \sum_v \boldsymbol{\phi}_{d,v,k}\boldsymbol{X}_{d,v}$
    end while
  end for
  Set $\boldsymbol{\lambda}_{k,v} = \boldsymbol{\eta} + \sum_d \boldsymbol{X}_{d,v}\boldsymbol{\phi}_{d,v,k}$
end while

referred to as "online" learning for LDA. In this algorithm 11 (also taken from (Hoffman et al., 2010)), documents are indexed by timestep t, and the update process on $\boldsymbol{\lambda}$ is modified in the following way. During the consideration of each new document t, $\tilde{\boldsymbol{\lambda}}$ is estimated as the optimal value for $\boldsymbol{\lambda}$ if all currently observed document were copies of the current one. $\boldsymbol{\lambda}$ is then updated using a weighted average of its previous value and $\tilde{\boldsymbol{\lambda}}$. The respective weights are determined by a new variable $\rho_t = (\tau_0 + t)^{-\kappa}$, with $\kappa \in]0.5, 1]$, meant to progressively decrease the influence of $\tilde{\boldsymbol{\lambda}}$ as more documents are seen. κ and τ_0 are hyperparameters of the method.

Algorithm 11 Online variational Bayes learning for LDA

Set $\rho_t = (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly
for $t = 0$ to $\inf$ **do**
  Initialize $\boldsymbol{\gamma}_{t,k} = 1$. (The constant 1 is arbitrary.)
  while $\frac{1}{K}\sum_k |\text{change in}\boldsymbol{\gamma}_{t,k}| < 0.00001$ **do**
    Set $\boldsymbol{\phi}_{t,v,k} \propto \exp\left(E_q[log(\boldsymbol{c}_k(t))] + E_q[log(\boldsymbol{\beta}_{k,v})]\right)$
    Set $\boldsymbol{\gamma}_{t,k} = \boldsymbol{\alpha} + \sum_v \boldsymbol{\phi}_{t,v,k}\boldsymbol{X}_{t,v}$
  end while
  Compute $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\eta} + t\boldsymbol{X}_{t,v}\boldsymbol{\phi}_{t,v,k}$
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t\tilde{\boldsymbol{\lambda}}$
end for

With this algorithm, $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ can be considered hyperparameters. However, when fitting the model on the learning corpus, we would like to also obtain estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$. This is be done based on the same variational approach as for model parameter inference. The lower bound estimate of the log likelihood can be optimized with respect to either $\boldsymbol{\alpha}$ or $\boldsymbol{\eta}$, yielding two additional

update equations. This process can then be adapted to the online learning algorithm in a way similar to $\boldsymbol{\lambda}$: The optimal values of $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ if all known documents were identical to the current one, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\eta}}$, are computed, and weighted with the previous value using ρ_t.

Once the model parameters have been estimated by the learning process, a dimension-reduced approximation of the documents is obtained as the combination of their topic composition $\boldsymbol{c}$, and the basis of topics $\boldsymbol{\beta}$. New documents can be studied under LDA through their representation in this already learned basis. In order to do that, the same algorithm is used, without updating the model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ that are considered known. This way, a topic composition $\boldsymbol{c}$ can be obtain from any document. For this reason, once a LDA model is learned, $\boldsymbol{c}$ is considered to be a function over documents.

3.6 . Summary

To summarize, we have defined an unsupervised statistical learning method called Latent Dirichlet Allocation. This method aims at modeling underlying patterns in the occurrences of words within text documents through the use of latent variables referred to as topics. The topics are defined by probability distributions over the set of possible words. Each document is attributed a distribution over topic indexes, noted $\boldsymbol{c}$, which can be interpreted as its composition in topics, providing a soft clustering of the input data. This methods was chosen over simpler models for its ability to capture 2-variables statistics and to represent documents with more than a single topic.

Direct parameter optimization is impossible due to analytical intractability of the posterior probability over model parameters. However, we can still approximate optimal parameters by using variational Bayes inference. This method is accessible in the Python module Gensim, which we use in the following applications.

4 - Application of LDA to climate data

4.1 . Introduction

In the previous chapter, we introduced the Latent Dirichlet Allocation (LDA) method, which is traditionally used to extract a set of topics from a corpus of text documents, and associate each document with its own topic composition. However, as was demonstrated in (Griffiths and Steyvers, 2004), topic modelling methods can be applied to images, and by extension, to physical variable maps. In this chapter, we will go over how LDA is applied to extract a set of patterns called motifs from a dataset of bidimensional climate variable maps, and associate each map with it own motif composition. Our methodology closely follows that of (Fery et al., 2022).

We present a set of motifs learned from Sea-Level Pressure (SLP) data in the NCEP reanalysis dataset (Kalnay et al., 1996; Kistler et al., 2001). The motifs are found to be similar to localized cyclonic and anticyclonic structures, providing them with direct physical interpretations. To verify whether these motifs represent general features of atmospheric circulation rather than artifacts specific to NCEP data, we apply LDA to a different reanalysis dataset, ERA5 (Hersbach et al., 2020), of higher spatial resolution. The results show a similar set of motifs, confirming the generalizability of the features captured by LDA to different datasets.

To further test the robustness of these motifs, meaning their stability across different parameter values and datasets, we compute additional LDA bases with different value of the parameters, and on model data. We find that all tested cases result in similar bases, therefore confirming the robustness of our set of motifs.

LDA serves as a dimensionality reduction tool, representing data on a small set of relevant modes. As such, despite its application for soft clustering, it fulfills a function similar to that of Principal Component Analysis (PCA). In the final section, we compare the advantages and drawbacks of LDA and PCA in reducing the dimensionality of SLP data. We develop a method to evaluate how efficiently each technique captures the information contained in the data. We find that PCA generally performs better when allowed to use all learned modes. However, with its sparse representation of maps, LDA captures the information of the maps, especially the extrema, using only a limited amount of motifs. The physically interpretable nature of the LDA motifs, along with their locality and the sparsity of the decomposition, make LDA a relevant

tool for studying atmospheric circulation.

4.2 . Methodology of application to climate data

4.2.1 . From text documents to climate maps

Latent Dirichlet Allocation is applied to datasets of bidimensional climate variables maps by using the equivalences summarized in table 4.1. Each spatial map is reinterpreted as a document. Grid points, or cells, are reinterpreted as the words, with the list of cells taking the role of the vocabulary. Field values at each cell are reinterpreted as word counts. In this case, the distribution over the vocabulary associated with the motifs, parameterized by $\boldsymbol{\beta}_k, \forall k \in \{1, ..., K\}$, are defined over space and are called "motifs" instead of topics. An illustration of these correspondences is shown in figure 4.1.

Natural Language Processing	Climate data application
Corpus	Dataset of 1-variable 2-D maps
Document	1-variable 2-D map
Word	Gridpoint
Wordcount	Map field value
Topic	Motif

Table 4.1 – Equivalences between application of LDA in the context Natural Language Processing and in our context. The maps used are 1-variable and 2-D, but contain 2 channels, one for positive values, and one for negative values.

Since the climate variable values are interpreted by LDA as word counts, they have to be digitized and made non-negative. In order to both satisfy this need, and still be able to process negative field values, the real variable maps are separated into two channels, one for positive and one for negative values. This process is equivalent to doubling the grid size over which the maps are defined, and results in a vocabulary size V of twice the amount of grid cells I (see also (Fery et al., 2022)). This process of separating the data into two channels is schematically represented in figure 4.2. Moreover, LDA computation times increase with the total amount of words in the documents, which corresponds in our case to the total sum of absolute field values in the maps. In order to manage computation times, a rescaling factor A is applied to the data before digitization. To sum up, let $\boldsymbol{x}_i$ designate the value of a map $\boldsymbol{x}$ at

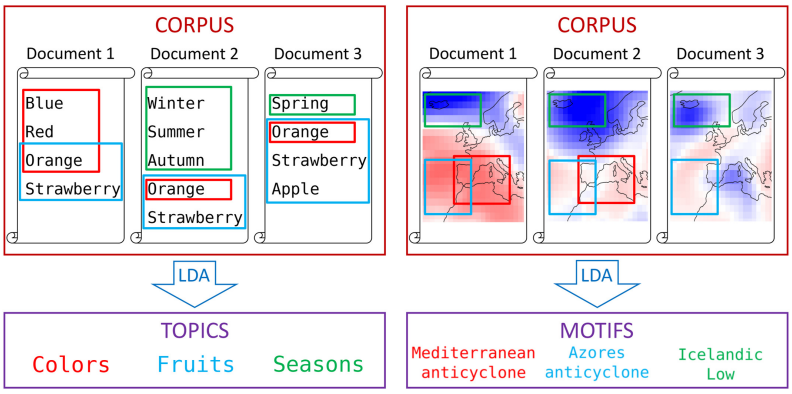


Figure 4.1 – Schema illustrating the objects that LDA takes as input, and seeks as topics/motifs, in the applications to text documents (left) and climate data maps (right). This figure is from (Fery et al., 2022).

cell i, out of a total number of I cells. The LDA formatted version of the map, $\boldsymbol{X}$, is defined as follows:

$$\forall i \in \{1, ..., I\}, \begin{cases} \boldsymbol{X}_i = max(int(A \times \boldsymbol{x}_i), 0) \\ \boldsymbol{X}_{i+I} = max(-int(A \times \boldsymbol{x}_i), 0) \end{cases} \tag{4.1}$$

where $int()$ designates the truncate to integer function. $\boldsymbol{X}$ is therefore a non-negative, integer-valued map defined on $V = 2I$ cells. The motifs $\boldsymbol{\beta}_k$ are therefore defined over the $V = 2I$ cells of the two channels.

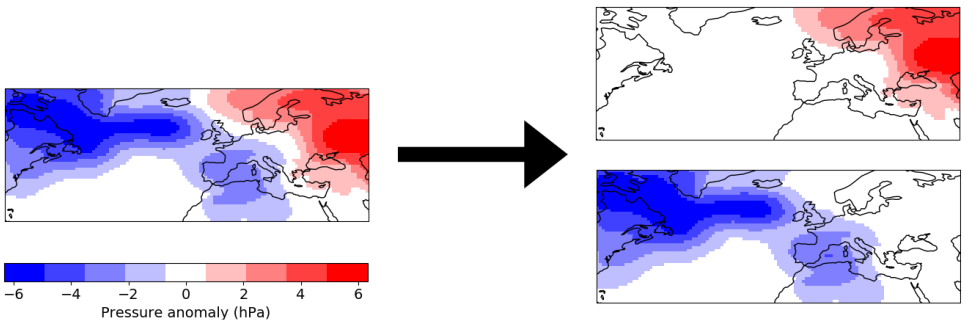


Figure 4.2 – Example representation of the process of separating the positive and negative values into two distinct channels. As LDA only takes positive values as input, both channels can be set to contain only positive or zero values, without loss of information.

Analysis of a corpus of documents with LDA consists in examining the posterior distribution of the topics $\boldsymbol{\beta}$, topic proportions $\boldsymbol{c}$, and topic assignments

z. These are determined via a variational Bayes approach (for more detail on the optimisation process, see section 3.5). This process generally outputs values for the Dirichlet parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ that respectively ensure the sparsity of the document-topic and of the topic-word distributions: there are generally only a few topics with non-negligible weight in each document, and each topic is characterized by high occurrences of a few vocabulary words. In the climate data case, this manifests as each data map only containing a few motifs, and each motif only involving a limited amount of grid points. This sparsity property makes LDA particularly suited to provide models and decomposition bases that can be interpreted easily.

For a given set of D maps, LDA returns motif distributions over grid cells $\boldsymbol{\beta} = (\boldsymbol{\beta}_k)_{k\in\{1,...,K\}}$, as well as the map compositions $(\boldsymbol{c}_k(d))_{d\in\{1,...,D\},k\in\{1,...,K\}}$, where $\boldsymbol{c}_k(d)$ denotes the weight of motif k in map d. The sum of all motif weights in a map is 1, such that $\forall d \in \{1, ..., D\} \sum_{k=1}^{K} \boldsymbol{c}_k(d) = 1$. The motif weights $\boldsymbol{c}(d)$ are always positive, unlike in other methods such as PCA. The set of distributions $\boldsymbol{\beta}$ can be considered as a basis of motifs. Any map $\boldsymbol{X}$ defined on the grid (but not necessarily part of the learning set) can be approximated in this basis by its $K-$dimensional motif composition, noted $\boldsymbol{c}(\boldsymbol{X})$. Different sets of maps can thus be compared efficiently through examination of their motif compositions.

4.2.2 . Application to sea-level pressure

We use the reanalysis dataset NCEP (Kalnay et al., 1996; Kistler et al., 2001) as our source of data to train LDA on. Our variable of study is the SLP, which contains the synoptic information relevant atmospheric circulation.[1] Our input dataset is NCEP SLP data from the north-Atlantic region between 22.5° and 70° latitude and 80° and 50° longitude, from 1948 to 2018. The temporal correlation time of synoptic circulation patterns is of a few days. Daily data, which is widely available, is therefore well suited to hold all the necessary information regarding atmospheric circulation patterns. We use the highest available spatial resolution for NCEP on this region, which is 20 by 53. We want to eliminate the seasonal cycle from the data, so as to look only at anomalous atmospheric circulation patterns, and ignore the average changes in pressure due to seasonal effects. Before applying equation 4.1 to format our data for LDA, we first compute SLP anomalies. Let $(\boldsymbol{P}_d)_{d\in\{1,...,D\}}$ denote the set of raw variable maps in the input dataset. As the data maps are daily, the d index can be interpreted as indicating the date of the corresponding map. Therefore, each map can be re-indexed by a combination of the year, noted y, and

1. See 1.2 for more information on pressure variables and the information they contain.

the day of the year, noted d'. In order to eliminate the seasonal cycle from the data, we compute anomalies, $\boldsymbol{x}$, from the input dataset $\boldsymbol{P}$ using equation 4.2.

$$\boldsymbol{x}_{d',y,i} = \boldsymbol{P}_{d',y,i} - \left\langle \boldsymbol{P}_{d',y,i} \right\rangle_y \tag{4.2}$$

With $\langle \cdot \rangle_y$ designating the average over all values of the variable y.

Our choice for the number of motif is based on the average local dimension of data, a value obtained using a dynamical systems methodology (Rodrigues et al., 2018b). Local dimension represents the number of degrees of freedom accessible to the system at any given point in phase space. In (Fery et al., 2022), it was found that NCEP SLP anomalies over our spatial domain have an average local dimension of 28. Therefore, we chose to use $K = 28$ motifs. The rescaling factor is set to $A = 0.5$ to alleviate computation time. Some arbitrariness exists in the choice of the factor. The effects of changing this parameter in the resulting motif basis are explored in section 4.3. The motif basis resulting from this choice of parameters is shown in figure 4.3. The motifs $\boldsymbol{\beta}_k$ are defined over the $V = 2I$ cells of the two channels. In order to represent them on the original grid size I, we show them as the difference of the two channels: at grid point i, $\boldsymbol{\beta}_{k,i} - \boldsymbol{\beta}_{k,i+I}$. We found that no motif attributes non-negligible probability to both channels at any grid point, which reflects the anticorrelation of the two channels. The motifs are shown in decreasing order of their average weights in the learning dataset. These average weights are show in figure 4.4.

Motifs take the form of localized objects at synoptic scale (of the order of 1000 km). In most cases, motifs are exclusively of one sign, either positive or negative. As such, the motifs can be seen as similar to cyclonic or anticyclonic structures. These properties of the motifs correspond to what is observed in previous applications of the method, in (Fery et al., 2022). Based on this physical interpretation, and to make discussion easier, names based on their signs and geographical locations were assigned to the motifs. Therefore, the motif weights inferred from an SLP anomaly map directly measure the contribution of the associated synoptic objects to the synoptic configuration of the map.

In the learning process, there is no constraint that imposes that the LDA motifs be localized. However, the motifs typically contain grid points with highly correlated field values. Therefore, the locality of the motifs is likely due to the local correlations caused by fluid interactions in the atmosphere. We note that several motifs are approximate opposites of one another, such as the Mediterranean anticyclone (motif 2) and the Genoa low (motif 8), or the Greenland high (motif 10) and Greenland low (motif 13). This allows motif compositions to represent both cyclonic and anticyclonic structures at the same location. In practice, we observe that no map is assigned non-negligible weights

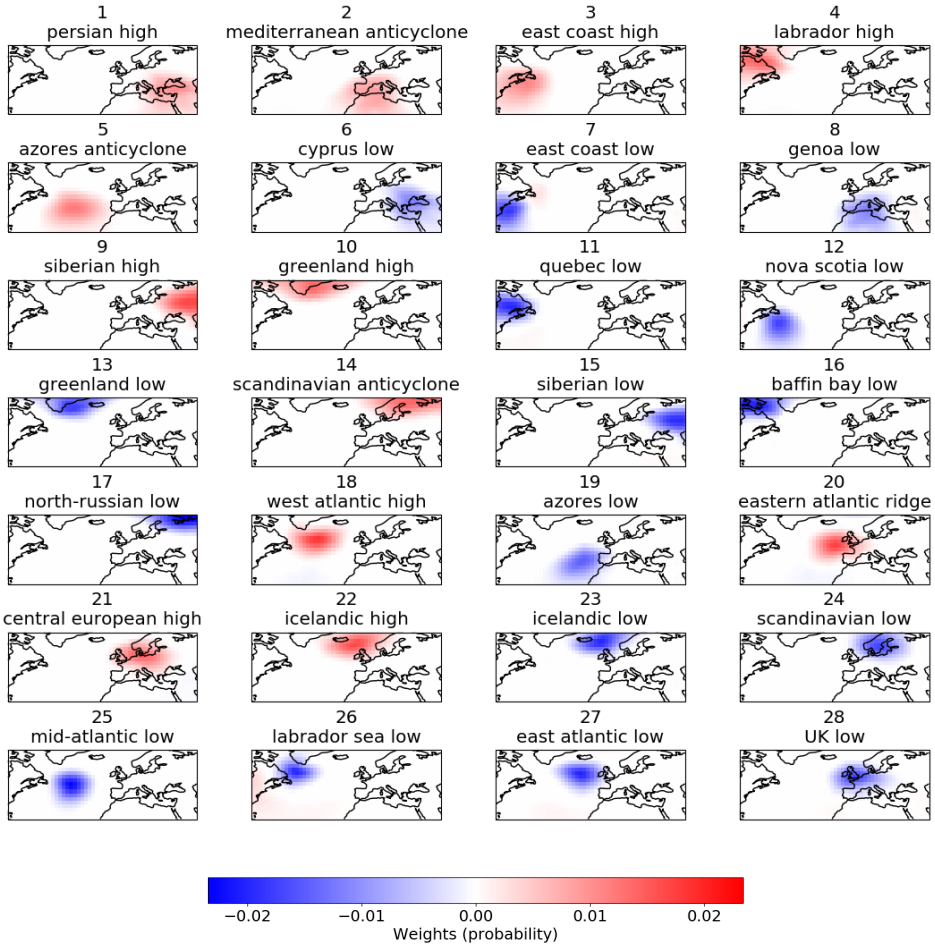


Figure 4.3 – The basis of 28 motifs learned by LDA from NCEP SLP anomaly fields. Each motif is defined as a probability distribution over space, with positive and negative channels. The names were given based on sign and geographical location.

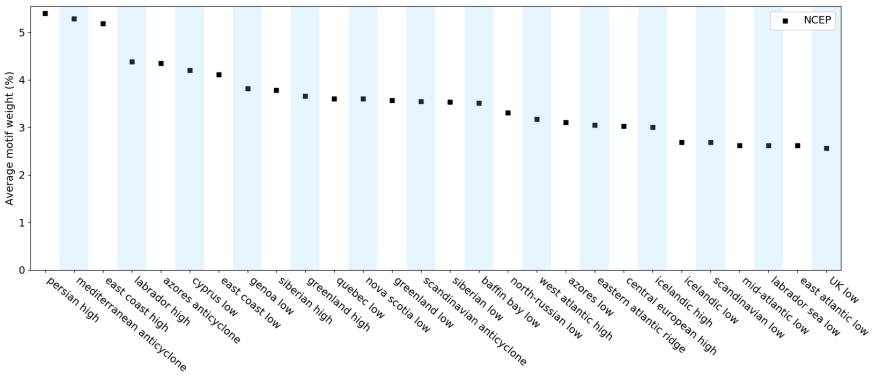


Figure 4.4 – Average weight of the 28 motifs learned by LDA in NCEP SLP data.

to two opposite motifs at the same time.

To provide a rough estimate, the computation of the LDA learning process, on this dataset of 25574 maps, with 2000 iterations of 20 passes through the training data, organized in chunks of 2048 maps, takes from 2 to 5 hours to complete on a computation cluster, with no particular parallelization of the code.

4.2.3 . Map reconstruction

LDA offers the possibility of reconstructing maps from a motif composition. The reconstruction of map $\boldsymbol{X}$, noted $\boldsymbol{X}^*$, is obtained based on equation 4.3.

$$\boldsymbol{X}^* = ||\boldsymbol{X}||_1 \sum_{k=1}^{K} \boldsymbol{c}_k(\boldsymbol{X}) \boldsymbol{\beta}_k \tag{4.3}$$

where:

- $\boldsymbol{\beta}_k$ is the spatial distribution associated with motif k.
- $\boldsymbol{c}_k(\boldsymbol{X})$ is the weight of the k-th motif in the weight vector associated with the pressure map $\boldsymbol{X}$.
- $||\boldsymbol{X}||_1 = \sum_{v=1}^{V} |\boldsymbol{X}_i|$, v iterating over the V grid points in both channels, is the ℓ_1 norm of map $\boldsymbol{X}$. This term is a renormalization factor, allowing for direct comparison with physical fields.

Reconstruction has no common equivalent in natural language applications of LDA. It is connected to, but different from, the generative process described in section 3.4. In the LDA generative process, a document length N is chosen, and motif weights are drawn from a Dirichlet distribution. N times, a motif is randomly chosen based on the weights, and a gridpoint is drawn from the

probability distribution described by the chosen motif. The field value at that gridpoint is then incremented. The reconstruction equation is obtained by applying the following changes to this process.

- Reconstruction is done from known motif weights: $\boldsymbol{c}(\boldsymbol{X})$.
- In order to limit map values dependancy to N, which has no physical meaning, the output map is normalized. The total sum of the map absolute values (i.e. its ℓ_1 norm) is set to 1.
- We want to eliminate all probabilistic aspects. To that end, we take the $N \longrightarrow \infty$ limit, so that the random sampling of motifs and gridpoints returns the exact probability distributions. Therefore, for each motif, gridpoints are attributed field values in direct proportion to the associated probability, and the field values associated with each motif are summed, weighted according to the motif weights.

Note that, while LDA takes as input text documents in the shape of word count matrices, it has no model for the amount of words in a document (see chapter 3). In the case of this climate application, this means that the LDA model does not represent the ℓ_1 norm of input maps. The output motifs $\boldsymbol{\beta}$ are all probability distributions over the two-channels grid, and the map weights $\boldsymbol{c}(\boldsymbol{X})$ are a probability distribution over motif indices. Therefore, without renormalization, a reconstructed field will have a ℓ_1 norm of 1. Patterns are preserved, but information regarding their intensity is lost. In order to restitute this information to reconstructed maps, they must be given an ℓ_1 norm with physical meaning. The value of the renormalization factor can be arbitrary in several cases. However, for reconstructed maps with available real field equivalents, the renomalization is taken directly from the original map.

Maps can be reconstructed from their associated motif weight vectors. We also can reconstruct maps from arbitrary weight vectors, even if they were not directly inferred from a real map. For example, we can reconstruct the average weights of cold spells and heatwaves maps in a given model. In this case, $\boldsymbol{c}_k(\boldsymbol{X})$ is replaced with $\langle\langle \boldsymbol{c}_k(\boldsymbol{X})\rangle\rangle$, where $\langle\langle . \rangle\rangle$ designates the average over all time steps associated with a given type of extreme event, and the renormalization factor $||\boldsymbol{X}||_1$ is replaced with $||\langle\langle \boldsymbol{X}\rangle\rangle||_1$.

4.3 . Robustness of the motif basis

4.3.1 . Robustness for different datasets

In this section, we explore whether the motifs identified by LDA capture general features of atmospheric circulation. To that end, we compare the basis of motif obtained in section 4.2 with a basis obtained on a different reanal-

ysis dataset, ERA5 (Hersbach et al., 2020). If the two bases exhibit sufficient similarity, which will be measured through spatial correlation, it will demonstrate that the features captured by LDA motifs represent general features of atmospheric circulation, rather than being specific to the NCEP data. In this case, we say that the resulting basis is robust with respect to changes in input dataset.

Our ERA5 learning corpus consists in daily maps from 1957 to 2021. We maintain the same spatial domain as for NCEP: between 22.5° and 70° latitude and 80° and 50° longitude. Although higher resolutions are available, we used a spatial resolution of 1°. It was found to be sufficient to contain all relevant information about circulation patterns on the synoptic scale, while maintaining manageable computation times. Our resolution is 48 points in latitude, 130 points in longitude, and we have two channels for positive and negative values. Therefore, the total number of values per map, associated with the vocabulary size V, is $2 \times 48 \times 130 = 12480$, an increase from the $2 \times 20 \times 53 = 2120$ of the NCEP case. We maintain the number of motifs, $K = 28$.

The 28 motifs learned from this ERA5 dataset are shown in figure 4.5. They are sorted by their average weights in decreasing order, as shown in figure 4.6. As was done previously, names are assigned to the motifs based on sign and localisation.

We evaluate the impact of this change in input data on the learned basis. Since motifs are defined by a spatial distribution, we use spatial correlation to measure their similarity. A $K \times K$ correlation matrix for the two bases is obtained as follows. All motifs are set to the same 1° resolution by linear interpolation. For each matrix entry, the Pearson correlation coefficient ρ_{kl} between motif k of basis $\boldsymbol{\beta}$ and motif l of basis $\boldsymbol{\beta}'$ is computed as shown in equation (4.4).

$$\rho_{kl} = \frac{\overline{\left(\boldsymbol{\beta}_k - \overline{\boldsymbol{\beta}_k}\right)\left(\boldsymbol{\beta}'_l - \overline{\boldsymbol{\beta}'_l}\right)}}{\sqrt{\overline{\left(\boldsymbol{\beta}_k - \overline{\boldsymbol{\beta}_k}\right)^2}}\sqrt{\overline{\left(\boldsymbol{\beta}'_l - \overline{\boldsymbol{\beta}'_l}\right)^2}}} \tag{4.4}$$

We show in figure 4.7 the correlation between the bases learned respectively on NCEP and ERA5, which are respectively associated with cell-motif distributions $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Motifs were reordered in order to give the same rank to motifs with the highest correlation. We see that 25 out of 28 motifs have a correlation of at least 0.7 with one in the other basis, defining a clear, unique equivalent. Motif differences are usually due to slight spatial translations, or the split of larger motifs into several smaller ones. The bases learned by LDA

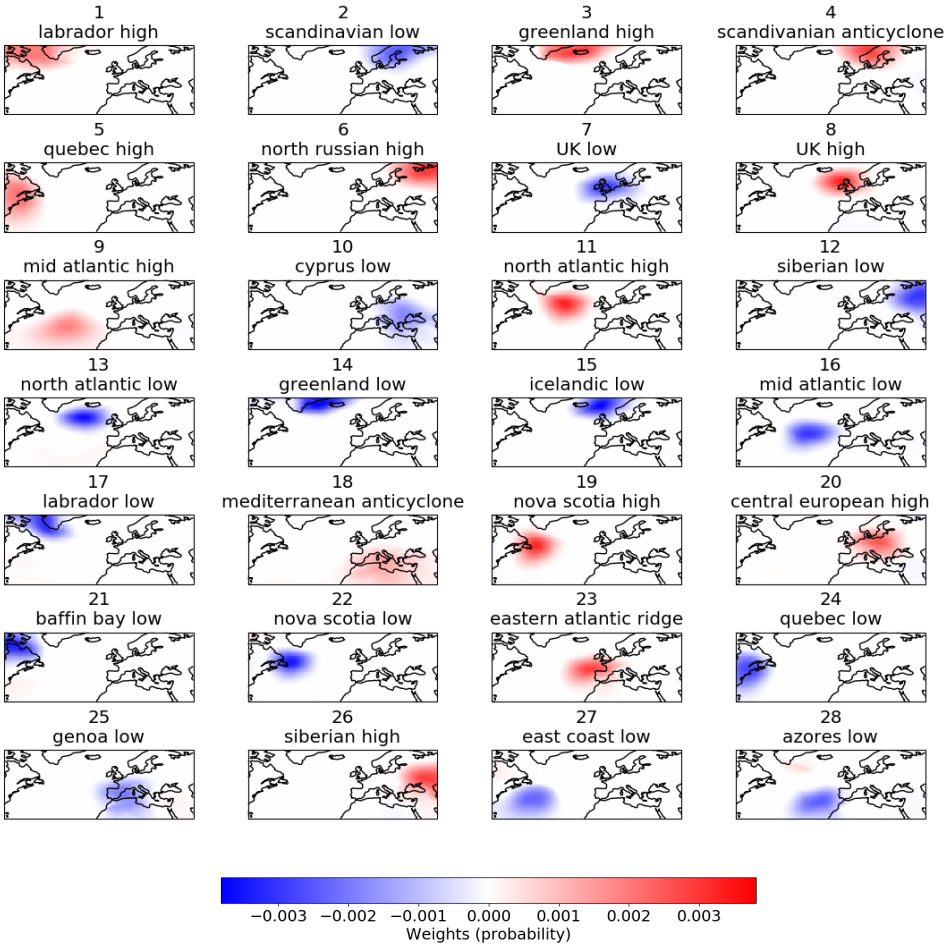


Figure 4.5 – The basis of 28 motifs learned by LDA from ERA5 SLP anomaly fields. Each motif is defined as a probability distribution over space, with positive and negative channels. The names were given based on sign and geographical location.

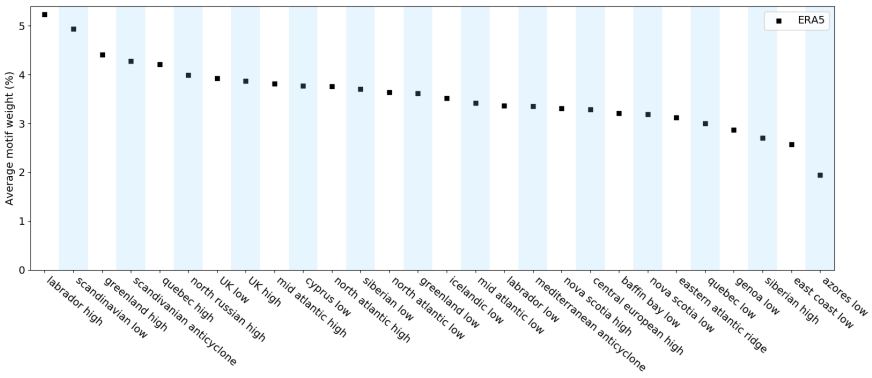


Figure 4.6 – Average weight of the 28 motifs learned by LDA in ERA5 SLP data.

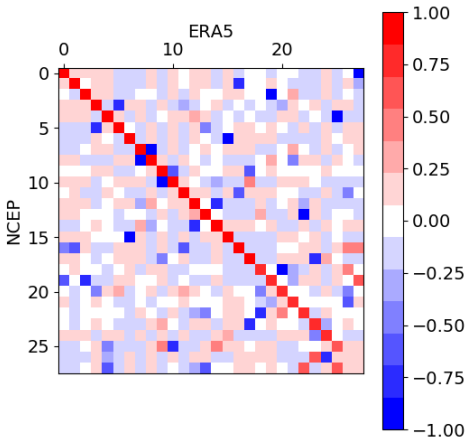


Figure 4.7 – Spatial correlation between the motifs of the bases obtained by applying LDA on NCEP (vertical) and on ERA5 (horizontal). The order of the motifs has been adjusted to put the highest correlations on the diagonal.

from NCEP and ERA5 SLP datasets are found to be similar despite the difference in origin and resolution. This demonstrates that the motifs do in fact represent features that are generally relevant to atmospheric circulation, and not just artifacts due to specific datasets.

4.3.2 . Applicability to model data

Our goal is to use motif weights to characterize atmospheric circulation in model and reanalysis data, and to use this characterization as a basis for

model evaluation. However, we must first demonstrate that the LDA basis, which was extracted from reanalysis data, is also suitable for the study of model data. If the underlying patterns of model SLP anomalies are too different to be efficiently represented with the same set of motifs, computing motif weights of model data in the ERA5 basis would result in loss of information, and induce biases in our results.

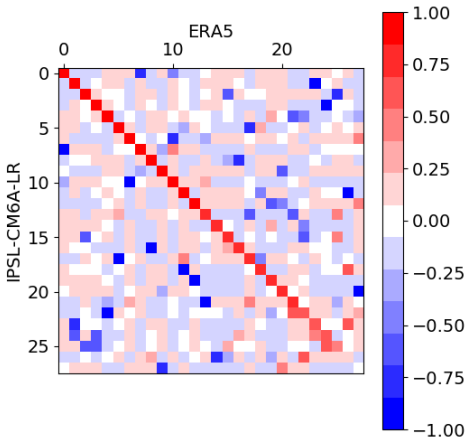


Figure 4.8 – Spatial correlation between the motifs of the bases obtained by applying LDA on ERA5 (horizontal) and on **IPSL-CM6A-LR** run 1 (vertical). The order of the motifs has been adjusted to put the highest correlations on the diagonal.

In order to explore how different a basis learned from model data would be, we compute a LDA basis from run 1 of the model **IPSL-CM6A-LR** over the historical period, from 1950 to 2014, with the same parameters as for the ERA5 application. figure 4.8 shows the correlation matrix between this basis and the ERA5 one. For the case considered, 22 out of 28 motifs have a clear equivalent in the other basis with correlation of 0.7 or higher. Based on these results, we consider that the features of atmospheric circulation captured by reanalysis motifs are also relevant to model data. The motif basis learned from ERA5 can therefore be considered suitable to represent model data.

4.3.3 . Robustness for different parameter values

To explore further the impact of data resolution on LDA, the correlation matrix between bases learned on ERA5 with respectively 20×53 and 48×130

resolution is shown in figure 4.9.

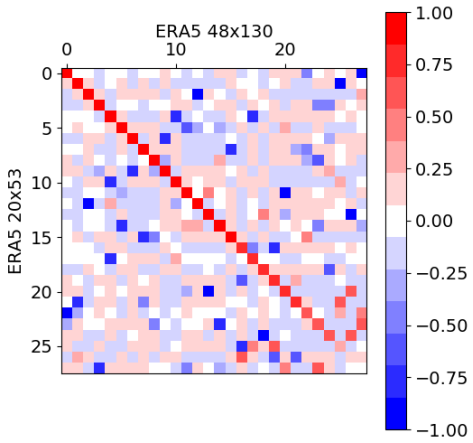


Figure 4.9 – Spatial correlation between the motifs of the bases obtained by applying LDA on ERA5 with 20×53 resolution (vertical), and 48×130 resolution (horizontal). The order of the motifs has been adjusted to put the highest correlations on the diagonal.

The bases are similar, with a majority of the motifs having an equivalent in the other basis with high correlation. However, there are 6 motifs that do not have a correlation of at least 0.7 with any from the other basis. Specifically, this is due to regions of high or low pressure being cut into a different amount of motifs in the two bases, as shown in figure 4.10. While the two bases are mostly similar, resolution does have an impact on LDA basis.

Finally, we examine the influence of the choice of rescaling factor on the LDA basis. The correlation between the bases learned on ERA5 respectively with rescaling factors $A = 1$ and $A = 0.5$ is shown in figure 4.11. All motifs have a unique equivalent in the other basis with a correlation of at least 0.85. Rescaling factor seems to have little impact on the basis learned by LDA.

4.4 . Comparison with PCA

4.4.1 . Methodology of comparison

LDA is used here as a dimensionality reduction technique, to provide new avenues of statistical analysis of SLP data. As such, it faces competition from

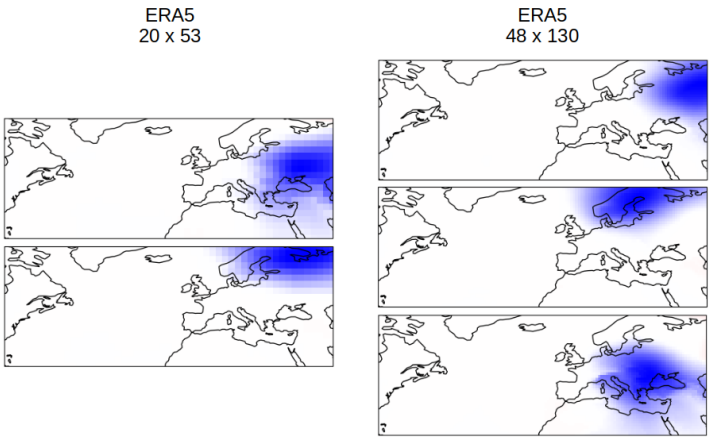


Figure 4.10 – Example of a region of low pressure being cut differently into motifs in bases learned from the same ERA5 SLP dataset, but with different resolution.

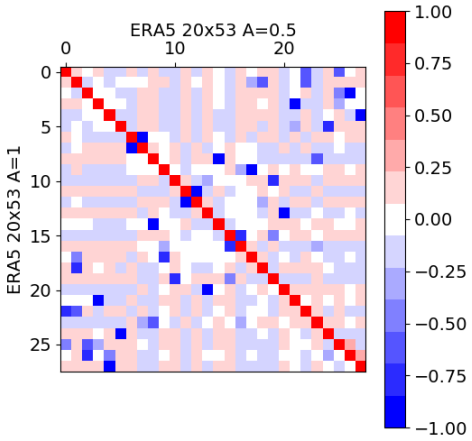


Figure 4.11 – Spatial correlation between the motifs of the bases obtained by applying LDA on ERA5 with, 1° resolution, with a rescaling factor of 1 (vertical), and 0.5 (horizontal). The order of the motifs has been adjusted to put the highest correlations on the diagonal.

more commonly used dimensionality reduction tools. In order to highlight the properties, advantages and drawbacks, of using LDA for this purpose, we compare its performance with PCA (Wold et al., 1987) using various metrics detailed below.

The performance of the dimensionality reduction methods will be compared through their ability to capture the information contained in the data. We compute LDA and PCA bases on the same dataset: daily NCEP SLP anomalies, with the same spatial and temporal domains and resolutions as detailed in section 4.2. In order to facilitate comparison with LDA, the number of modes chosen for PCA is the same $K = 28$. For both methods, the number of iterations is independently chosen to ensure convergence of the modes. Additional iterations produced only negligible effects in the modes. We compute the weights of SLP maps from the NCEP dataset in the reduced spaces of modes as learned by LDA and PCA. Then, the maps are reconstructed using equation 4.3, and compared with the original using one of the following six metrics.

- ℓ_2 norm, defined in equation 4.5.

$$\frac{1}{V}\sqrt{\sum_{v=1}^{V}\left(\boldsymbol{X}_v^* - \boldsymbol{X}_v\right)^2} \tag{4.5}$$

- Extremum value, defined in equation 4.6. A similar definition is used for the minimum.

$$\left|max_v\left(\boldsymbol{X}_v^*\right) - max_v\left(\boldsymbol{X}_v\right)\right| \tag{4.6}$$

- Extremum position, defined as the haversine distance between the coordinates of the grid point where the maximum (res. minimum) is reached in the reconstructed map, versus the original, in kilometers.
- Correlation, defined in equation 4.7.

$$\frac{\overline{\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}^* - \overline{\boldsymbol{X}^*}\right)}}{\sqrt{\overline{\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^2}}\sqrt{\overline{\left(\boldsymbol{X}^* - \overline{\boldsymbol{X}^*}\right)^2}}} \tag{4.7}$$

Where $\bar{\cdot}$ designates the spatial average: $\overline{\boldsymbol{X}} = \sum_{v=1}^{V} \boldsymbol{X}_v$. A low reconstruction error (or high correlation) indicates that the dimensionality reduction method has efficiently captured the information from the data into its modes and weights.

When reconstructing the maps, instead of using all 28 modes, we only use a limited number of modes, noted K_{lim}. This incomplete reconstruction using only K_{lim} modes is done following equation 4.8, assuming $\boldsymbol{c}_k(\boldsymbol{X})$ designates the weight associated to mode k in map $\boldsymbol{X}$ for both methods.

$$\boldsymbol{X}^* = ||\boldsymbol{X}||_1 \sum_{k=1}^{K_{lim}} \boldsymbol{c}_k(X)\boldsymbol{\beta}_k \tag{4.8}$$

With

$$\sum_{d}^{D}|\boldsymbol{c}_0(d)| \geq \sum_{d}^{D}|\boldsymbol{c}_1(d)| \geq ... \geq \sum_{d}^{D}|\boldsymbol{c}_K(d)| \tag{4.9}$$

The incomplete reconstruction process assumes that the motifs are ordered from highest to lowest total absolute weight in the considered dataset. This order is computed independently for the general, cold spell, and heatwave cases. In the extreme cases, the total absolute weights are computed using only maps associated with the corresponding type of extreme event. Here, K_{lim} is treated as a variable of study, and we will explore reconstruction errors as a function of K_{lim}.

4.4.2 . Properties of LDA and PCA

In fig 4.12 we represent the median and quartiles reconstruction errors as a function of K_{lim}, for LDA and PCA, using each of the previously defined error metrics, in the general, heatwave, and cold wave cases.

When all 28 motifs are used, PCA achieves lower reconstruction error than LDA for every metric, in every case. PCA also scores better than LDA when using the ℓ_2 norm and correlation error metrics, for all valus of K_{lim}. However, LDA reaches its optimum score with notably fewer motifs than PCA. On metrics based on extrema, the convergence of the LDA reconstruction error requires even fewer motifs. Furthermore, on lower motif amounts, LDA can reach lower reconstruction errors than PCA on median. This is especially visible with the minimum value error metric (g, h and i). The fast convergence of LDA allows it to reach low error values, performing better than PCA when using only few motifs, even if it is overtaken when more motifs are available.

The behavior of LDA is likely due to the sparsity of the map representation, and the sparsity of the motif themselves. In the LDA basis, each map is only attributed non-negligible weights to a limited number of motifs. This likely explains the fast convergence of LDA reconstruction error with K_{lim}. Additionally, the motifs are localized, only having non-negligible values in a limited spatial region. LDA approximates maps by cutting them down into localized objects whose weights are proportional to the relative local intensity of the field. Therefore, extrema are typically represented by LDA using a single motif with high weight, while PCA requires the combination of several motifs to reproduce its shape and intensity. These properties make LDA well-suited to the representation of extrema in a map, within few motifs. Since the dynamics of atmospheric circulation are described by the intensity and localization of high and low pressures, and the fluctuations around the middle values are less relevant, we believe that LDA is adapted to the task of representing it. Furthermore, the locality of the motifs, as well as the two levels of sparsity, contribute to the ease of associating a physical interpretation to individual LDA weights.

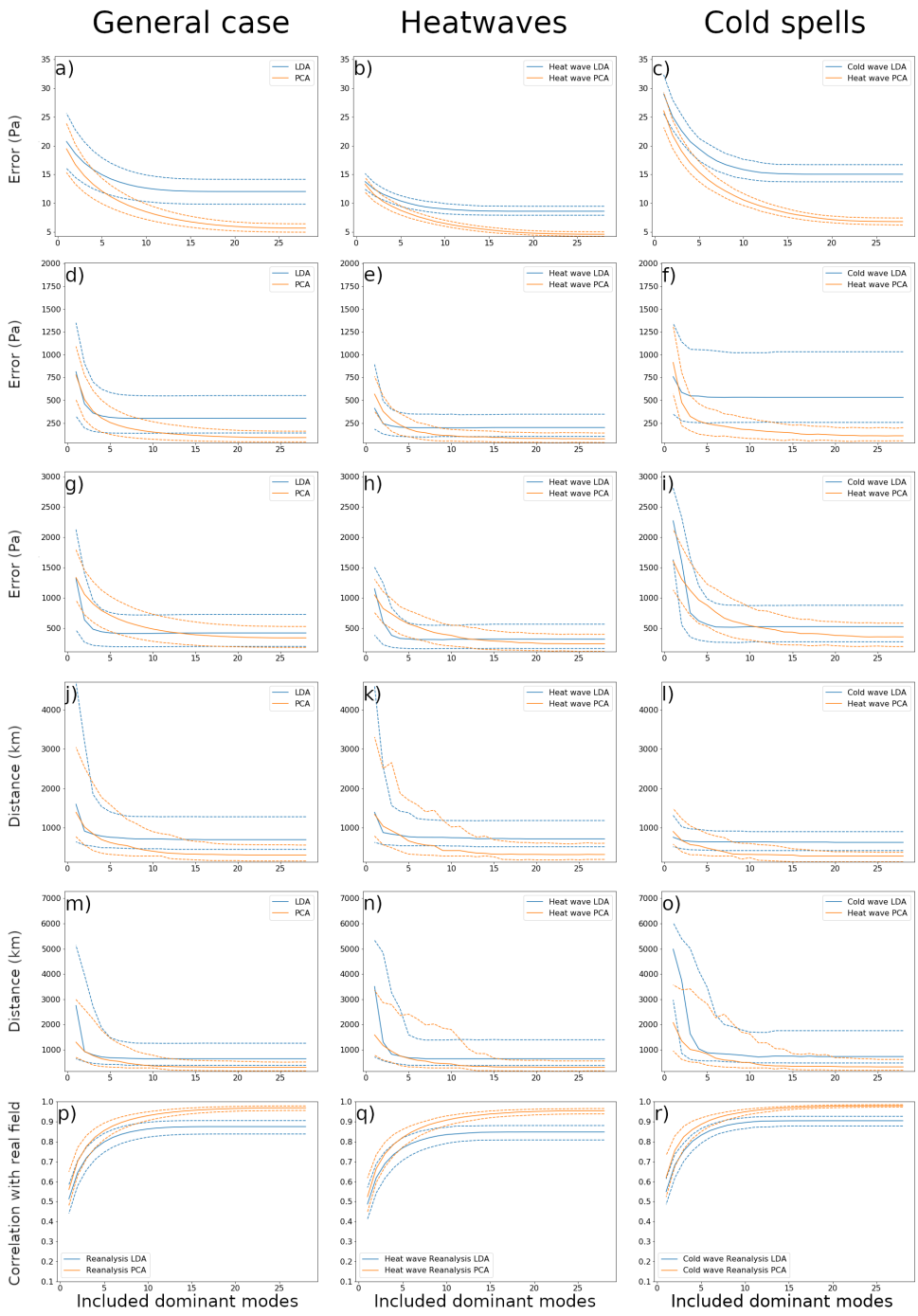


Figure 4.12 – Median (solid line) and quartile (dashed lines) reconstruction error of LDA and PCA, as a function of the number of modes (out of 28) used for reconstruction. Left to right: general case, heatwave, and cold spell case. Top to bottom: ℓ_2 norm error (a-c), maximum value error (d-f), minimum value error (g-i), maximum location error (j-l), minimum location error (m-o), correlation (p-r).

Note, however, that there are limits to the relevance of this comparison. While PCA modes can be associated with a positive or negative weight for approximating a map, LDA motifs can only have positive weights. This means that unlike PCA, to capture the possibility of a structure being either positive or negative, LDA requires the use of two modes instead of one. Therefore, using the same number of modes does not imply that PCA and LDA will be able to capture a straightforwardly comparable amount of information.

In conclusion, PCA outperforms LDA in capturing the information present in the dataset maps. However, LDA possesses the advantages of sparsity of the map representation, and sparsity and locality of the motifs. Such properties are appropriate to a study of atmospheric circulation using pressure data.

4.5 . Summary

In this chapter, we applied LDA to daily SLP anomaly data from reanalysis datasets. This was done by reinterpreting physical maps as text documents, with grid points corresponding to a vocabulary of words, and field values to word counts. The latent variables of LDA, topics, became spatial distribution that we call motifs. The motif basis learned from the data is comprised of synoptic-scale localized objects that can be physically interpreted as cyclones and anticyclones. We tested the robustness of the basis by comparing it to others learned from different datasets, including model data, with different resolutions, and/or different parameters. We find that the set of motifs is robust and represents features generally relevant to atmospheric circulation.

Finally, the ability of LDA to capture information from the data was compared to that of PCA. We show that each method has its advantages, and that that of LDA lie in the sparsity of the motifs, the sparsity of the map representation, and the ease of access of a physical interpretation for motif compositions.

SLP maps, even outside of the training dataset, can be associated with a composition in motifs. This composition characterizes the synoptic configuration of the map. Therefore, we now have a tool that allows us to compute the synoptic configuration of any SLP map or dataset. This characterization can then be used as a basis for comparing the synoptic configurations of several datasets, which can be applied for model evaluation.

5 - Analysis of synoptic composition

5.1 . Introduction

In the previous chapter, it was established that applying LDA to ERA5 SLP anomaly data yields a basis that decomposes any SLP map into its synoptic components. In this part, we use this decomposition to characterize average atmospheric circulation in ERA5 by the average motif weights. We obtain this characterization in the general case, as well as for extreme temperature events (**Cold Spells** and **Heatwaves**) in several European countries: France, Germany, Poland, Italy, Spain, and the United Kingdom (UK). This allows us to discuss the atmospheric circulation patterns associated with extreme cold and heat, and explore their variations across European countries.

We introduce four GCM from the CMIP6 project: **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**. By applying the same characterization to the models, we can evaluate their ability to capture the properties observed in ERA5 reanalysis data. We find that models generally replicate the atmospheric circulation associated with ERA5, but still display significant biases. Extreme event circulation, however, is less well represented by the models. We observe significant errors that vary on a model-to-model basis as well as per region of interest.

Since we measure representation error on individual motifs, our methodology allows us to detect and measure model biases in a localized manner. In this chapter we provide several examples of model biases, either shared or model-specific, in the general case, as well as for extreme temperature events in different countries.

5.2 . Synoptic configurations in ERA5

5.2.1 . Average atmospheric circulation

In the following, we use the basis of 28 motifs learned on ERA5 with 48×130 (1°) resolution, with a rescaling factor of $A = 0.5$ (see figure 4.5). This data will be treated as ground truth to compare the models to.

In figure 5.1, we show the average motif weight of anomaly SLP fields in the ERA5 dataset, between 1957 and 2014. Uncertainties are estimated by a resampling method called bootstrapping. Many alternative sets of maps, of the

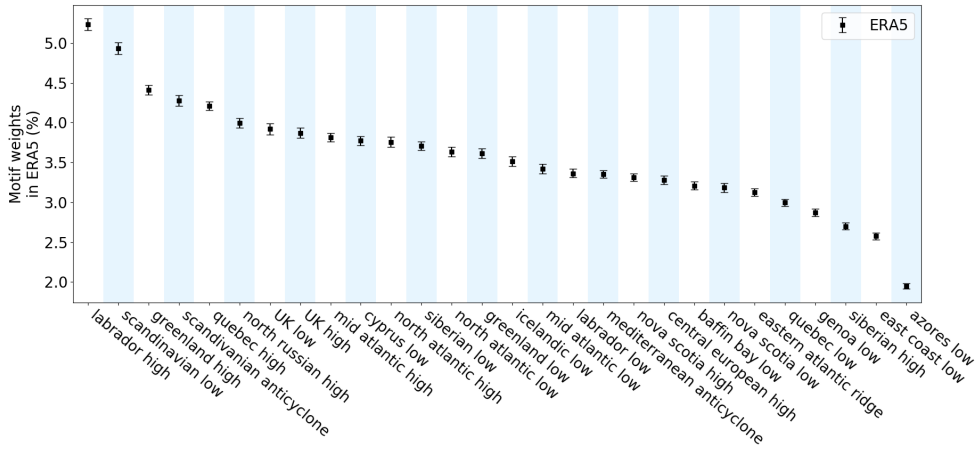


Figure 5.1 – Average motif weights in ERA5 SLP anomaly maps. The motifs have been ordered by decreasing average weight. 90% confidence uncertainties are determined by bootstrapping.

same size as the original, are generated by randomly sampling with replacements from the original data. The average motif weights in these datasets are computed, and the 5th and 95th percentile weights for each motif are used as lower and upper errors, provide a 90% confidence intervals. The choice of 90% is customary for the computation of confidence intervals. The number of resampled datasets used for this is 500, as it was found to be sufficient to reach statistical convergence.

To each motif is associated an average weight, representing the percentage of the data it explains. As motifs are associated with synoptic objects, motif weights also indicate the prevalence of the associated synoptic object in the data. Therefore, we refer to the set of average motifs weights of a dataset as its synoptic configuration. We find that some motifs explain more of the data than others, with weights varying between 2% and more than 5%. Most of the dominant motifs are northern anticyclones, such as the Labrador high, or the Greenland anticyclone, with the exception of the Scandinavian low. The motif with the smallest overall contribution in ERA5 data is the Azores low. This characterization of the synoptic configuration of ERA5 reanalysis data will serve as a basis of comparison when exploring the synoptic configurations of other datasets. For this reason, the order the motifs will be presented in, in all subsequent figures, will be the decreasing ERA5 average weight order shown here.

5.2.2 . Synoptic configuration of extreme events

Synoptic configuration can also be used to characterize extreme weather events. We want to identify the average atmospheric circulation patterns associated with extreme cold or heat events, and compare them with average circulation in the general case. This is accomplished by selecting only maps corresponding to days associated with a given type of extreme event, then computing average weights over those maps. We use the following definitions for extreme temperature events:

- A **Cold Spell** over a specific region designates any period of at least three consecutive days where the spatial average of daily mean temperature over the region is below the 3rd percentile.
- A **Heatwave** over a specific region designates any period of at least three consecutive days where the spatial average of daily mean temperature over the region is beyond the 97th percentile.

The patterns associated with extreme temperatures events in one region are expected to differ from those associated to such events in another. This is the reason why our definitions are region-based. The considered regions are: France, Germany, Italy, Poland, Spain, and the United Kingdom (UK). In order to filter the data by country, we use masks based on the NUTS region system (Commission Européenne, 2005). These masks are shown in figure 5.2.

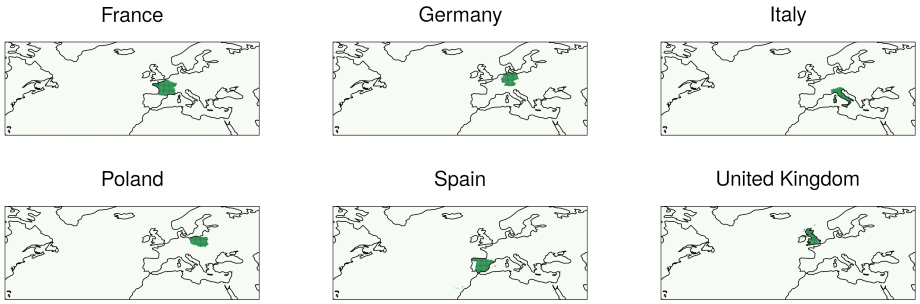


Figure 5.2 – Masks used to define the different regions. They are based on the NUTS region system (see Commission Européenne, 2005).

We first focus our study on extreme temperature events occurring in France. The synoptic configurations corresponding to **Cold Spells** and **Heatwaves** in France are shown and compared to the general synoptic configuration of reanalysis data in figure 5.3. Uncertainties for **Cold Spells** and **Heatwaves** were obtained through the same method of bootstrapping.

We find that the synoptic configuration of extreme events is significantly different from the average configuration of the general data. This is an expected

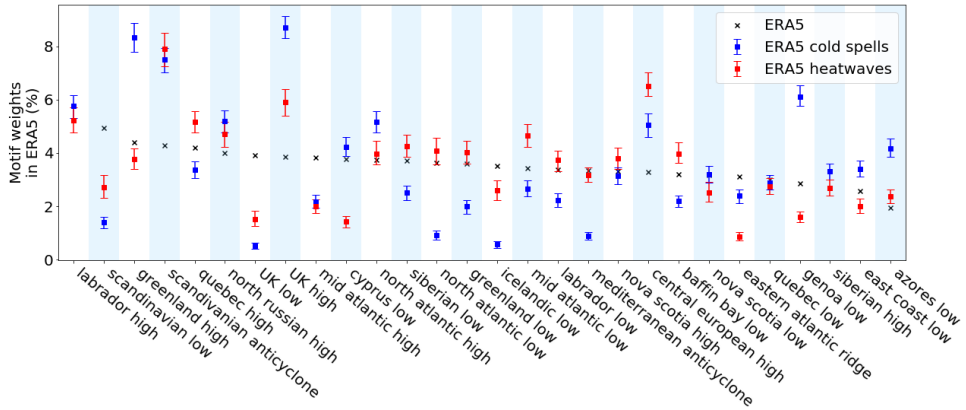


Figure 5.3 – Average motif weights in ERA5 SLP anomaly maps, in the general case (black), in the case of **Cold Spells** in France (blue), and in the case of **Heatwaves** in France (red). 90% confidence uncertainties are determined by bootstrapping.

result, as specific atmospheric circulation patterns are known to be drivers of extreme temperature events. In particular, in both the **Cold Spell** and **Heatwave** cases, the synoptic configuration displays a higher amplitude of variation between the motifs with highest and lowest weights. **Cold Spell** circulation is dominated by northern anticyclones such a the Greenland high, the Scandinavian anticyclone and the UK high, with more than 6% weights each. Correspondingly, the low pressure objects over those regions have less than half of their general case weights. The Genoa low is also a key motif in French **Cold Spells**, with the fourth highest weight. Its opposite, the Mediterranean anticyclone, has half of its general case weight. **Heatwave** circulation is dominated by a smaller set of high-weights motifs, mainly consisting of the Scandinavian anticyclone, and the central European high. Both types of extremes are associated with above-average weights of the Scandinavian anticyclone and UK high.

However, our definitions of extreme temperature events induces a bias regarding the period of the year these events are from. **Cold Spells**, with our definition, are extremely likely to occur during winter, and **Heatwaves**, similarly, in summer. Therefore, it is important to determine whether the significant differences observed in synoptic configuration are due extreme circulation, or seasonal circulation. To that end, we look at the synoptic configuration of ERA5 maps corresponding to winter and summer, and compare it to the previously shown extreme event synoptic configuration. This is shown in figure 5.4. Uncertainties for seasonal data were obtained through bootstrap-

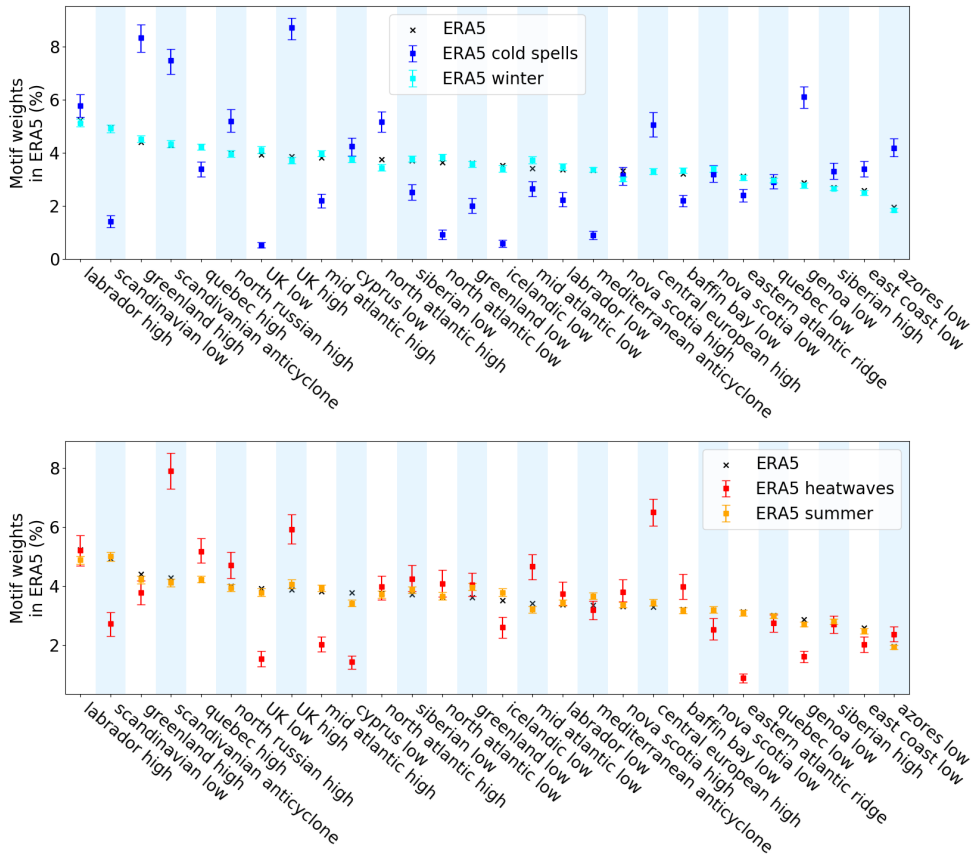


Figure 5.4 – Top: Average motif weights in ERA5 SLP anomaly maps, in the general case (black), in the case of winter (cyan), and in the case of **Cold Spells** in France (blue). Bottom: Average motif weights in ERA5 SLP anomaly maps, in the general case (black), in the case of summer (orange), and in the case of **Heatwaves** in France (red). 90% confidence uncertainties are determined by bootstrapping.

ping, as in the previous figures. We find that the amplitude of differences in average motif weights between seasonal data and general data is significantly smaller than with the extreme temperature cases. Therefore, we conclude that the different motif weights observed in extreme cases cannot be explained by seasonal circulation, and are instead the product of circulation patterns specifically associated with extreme temperature events.

5.2.3 . Extreme events across European countries

We have established that our method is able to characterize circulation patterns associated with extreme temperature events. In order to explore the

possible geographical dependencies of the synoptic configurations associated with extreme events, we compare average motif weights during extreme events across several western European countries. The list of countries studied, in no particular order, is: France, Germany, Poland, Italy, Spain, and the United Kingdom (UK) (see figure 5.2 for the definition of the regions).

In figure 5.5 are shown the average synoptic configurations of **Cold Spells** days for each of the six considered countries. (The data for France is the same as in figure 5.3.)

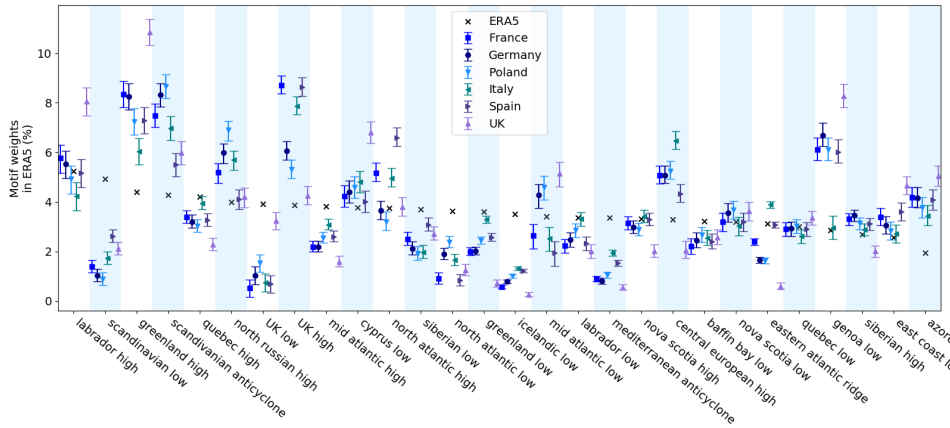


Figure 5.5 – Average motif weights in ERA5 SLP anomaly maps, in the general case (black), and in the case of **Cold Spells** in various European countries. 90% confidence uncertainties are determined by bootstrapping.

There is a shared structure among the synoptic configurations of **Heatwaves** in the six regions, which generally involves higher weights for highs in the north and low in the south, and correspondingly, lower weights for low in the north and highs in the south. The five most prevalent motifs during **Cold Spells** in France are UK high, Greenland high, Scandinavian anticyclone, Genoa low, and Central European high. However, we observe that there are differences in the synoptic configurations associated with **Cold Spells** in other countries. In particular, the UK differs from the other countries. In the UK, **Cold Spells** are associated with higher average weights for Labrador high, Greenland high, Cyprus low, Genoa low and east coast low, and lower weights for Quebec high, UK high, and Siberian high. The larger differences in synoptic configuration of UK **Cold Spells**, relative to other countries, may be due to its position being further north than other countries, as well as being the only insular region considered. Comparatively, France, Germany and Poland display **Cold Spell** synoptic configuration within similar ranges. Notable dif-

ferences include largely lower weights for UK high and north Atlantic high for Germany and Poland, compared to France, as well as higher weights for north Russian high, north Atlantic high, and mid Atlantic low. This similarity can be explained by geographical proximity. In Italy, unlike all other considered countries, **Cold Spells** are not associated with higher Genoa low weights compared to the general case. Similarly, the weights for Quebec high, Greenland low, and Labrador low do not significantly deviate from the general case, while they do for all other considered countries. Italy **Cold Spell** synoptic configuration shows lower weight for Greenland high, and higher weights for central European high, and eastern Atlantic ridge. Finally, **Cold Spells** in Spain are associated with relatively lower weight differences with the general case for Scandinavian high and low motifs, indicating a lower influence of the Scandinavian synoptic objects over **Cold Spells** in Spain, relative to other countries. This is also the case for the eastern Atlantic ridge. The Spain **Cold Spell** synoptic configuration also displays higher weights for north Atlantic high and Siberian low, as well as lower weight for central European high.

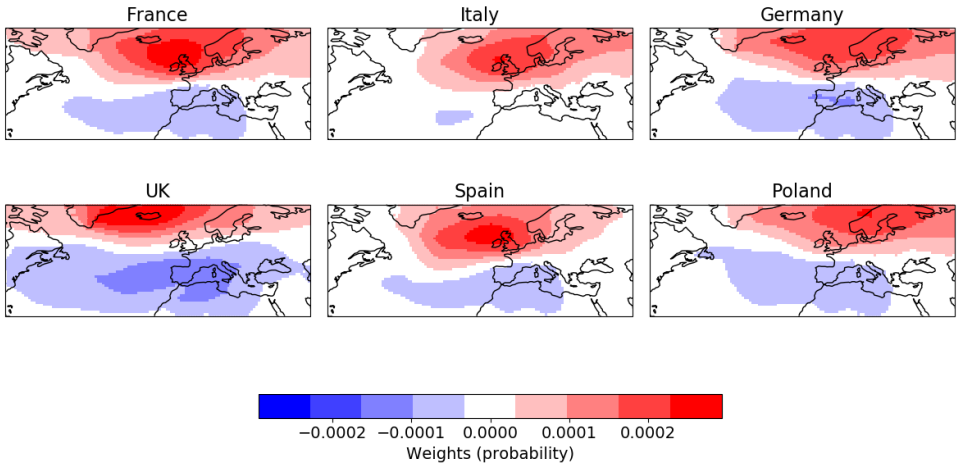


Figure 5.6 – Reconstructions from the average **Cold Spell** weights in six European countries (normalized).

In order to visualize the different structures associated with these differences in weights, we provide normalized reconstructed maps in figure 5.6. The reconstructions displays the properties we have highlighted from the differences in **Cold Spell** weights. In particular, the lack of a strong Genoa low in the case of Italy sets it apart relative to the other countries. We also observe the anticyclonic structure extending further over the north Atlantic for France than for Germany and Poland. In the case of the UK, we see a stronger cy-

clonic structure in the south, which extends further west, as well as a lack of UK high, compared to the other countries.

Figure 5.7 shows the average synoptic configurations associated with **Heatwaves** for each of the six considered countries. (The data for France is the same as in figure 5.3.)

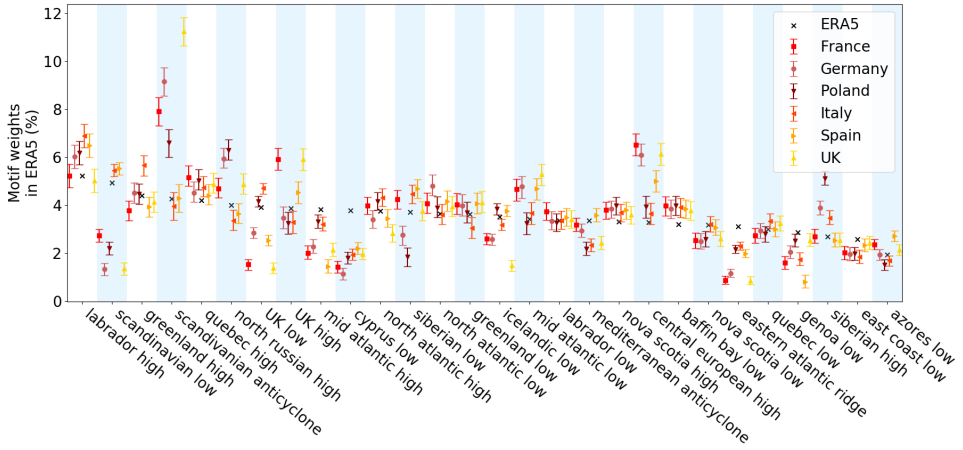


Figure 5.7 – Average motif weights in ERA5 SLP anomaly maps, in the general case (black), and in the case of **Heatwaves** in various European countries. 90% confidence uncertainties are determined by bootstrapping.

Heatwave results differs from **Cold Spells** by having larger variations in motif weights among countries. However, simultaneously, there are more motifs without significant differences among countries. Those are Quebec high and low, Labrador low, Nova Scotia high and low, Baffin bay low, and East Coast low. Unlike all other regions, Italy and Spain show little to not weight difference between **Heatwaves** and the general case for Scandinavian anticyclone, Scandinavian low, and north Russian high. For Italy and Poland, this also includes mid-Atlantic high, mid-Atlantic low, UK high, UK low and central European high. **Heatwaves** in Spain are also associated with a lower Genoa low weight. France, Germany and Poland have similar variations in average synoptic configuration relative to he general case for most motifs. Exceptions include Scandinavian anticyclone and low, UK low, and Siberian high and low. Finally, UK **Heatwaves** are associated with weights for Scandinavian anticyclone much higher that in any other case, for any country. Furthermore, they are associated with lower weights for Icelandic low. They are also associated with higher weights for UK high and lower weights for UK low, which is con-

sistent with known drivers of **Heatwaves**, as detailed in chapter 1.

All of these differences in motif weights can be visualized using normalized reconstructed maps, as shown in figure 5.8. We recognize in figure 5.8 several of

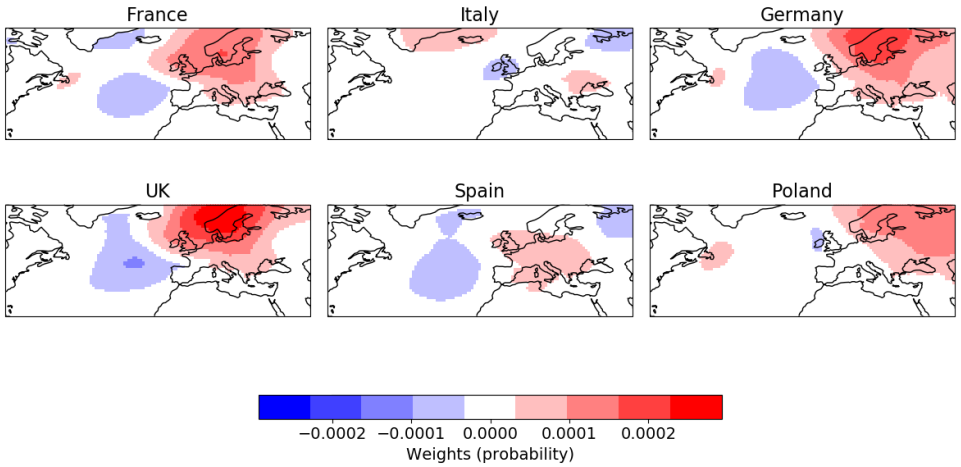


Figure 5.8 – Reconstructions from the average **Heatwave** weights in six European countries (normalized).

the properties that were seen from the weights. Notably, we can see that Italy is closer to the average than other countries, or that for the UK, **Heatwave** synoptic configuration is dominated by a strong Scandinavian anticyclone.

To conclude, we have observed that extreme temperature events in each region, both hot and cold, are associated with distinct average synoptic configurations. This provides a detailed characterization of the atmospheric circulation associated with extremes in reanalysis data, on which models can then be evaluated.

5.3 . Synoptic configurations in models

5.3.1 . General case

Using this characterization of atmospheric circulation, we will now evaluate the ability of climate models to reproduce the synoptic configurations observed in ERA5 reanalysis data. The results presented in this part were published in (Malhomme et al., 2025). We consider 4 GCM models from CMIP6: **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**. The spatial resolution and number of available run of each model are detailed in table 5.1

Model	Resolution	SLP runs	Reference
IPSL-CM6A-LR	38×53	33	Boucher et al., 2020
MIROC6	34×92	50	Tatebe et al., 2019
ACCESS-ESM1.5	39×69	30	Ziehn et al., 2020
CanESM5	17×46	25	Swart et al., 2019

Table 5.1 – Spatial resolution and amount of available runs for each of the 4 considered models.

In our methodology of model evaluation, ERA5 reanalysis data is treated as ground truth. Note, however, that reanalysis datasets combine observational data with model outputs, through interpolation and other techniques. As a result, reanalysis data can potentially inherit some of the biases from the models used in their construction. As our methodology relies on comparing of the two, it is blind to any shared biases. This is a limitation of our method. Nonetheless, because reanalyses incorporate actual observations, they remain the closest data available to reality, and are likely to be less biased than the models alone. Therefore, if the two are found to disagree, we have reason to believe that reanalyses are closer to the truth, and our approach can still detect biases in the models.

For each historical run of the four models, the ERA5 motif basis weights are computed by Bayesian inference. Then, the average synoptic configuration is computed. At first, we consider the general data (no filtering by extreme event). For each run, the relative difference between the average motif weights in the model and in the reanalysis is computed for all K motifs following equation 5.1.

$$\forall k \in \{1, ..., K\}, E_k = \frac{\langle \boldsymbol{c}_k(\boldsymbol{X}^{m,r}) \rangle - \langle \boldsymbol{c}_k(\boldsymbol{X}) \rangle}{\langle \boldsymbol{c}_k(\boldsymbol{X}) \rangle} \tag{5.1}$$

where $\boldsymbol{X}$ corresponds to reanalysis maps, $\boldsymbol{X}^{m,r}$ corresponds to maps from run r of model m, and $\langle \cdot \rangle$ designates the average over all maps in the dataset (model run or reanalysis).

The statistics of this relative error, for each of the four considered models, are shown in figure 5.9, using box plots. The mean motifs weights in the reanalysis data are also shown, for reference. Box plots represent data in the following way: The black horizontal line inside the box designates the median of the values. The edges of the box represent the first and third quartiles of

the values. Whiskers extend to the maximum and minimum values, but only up to 1.5 times the difference between the first and third quartiles. The remaining values are represented as colorless circles.

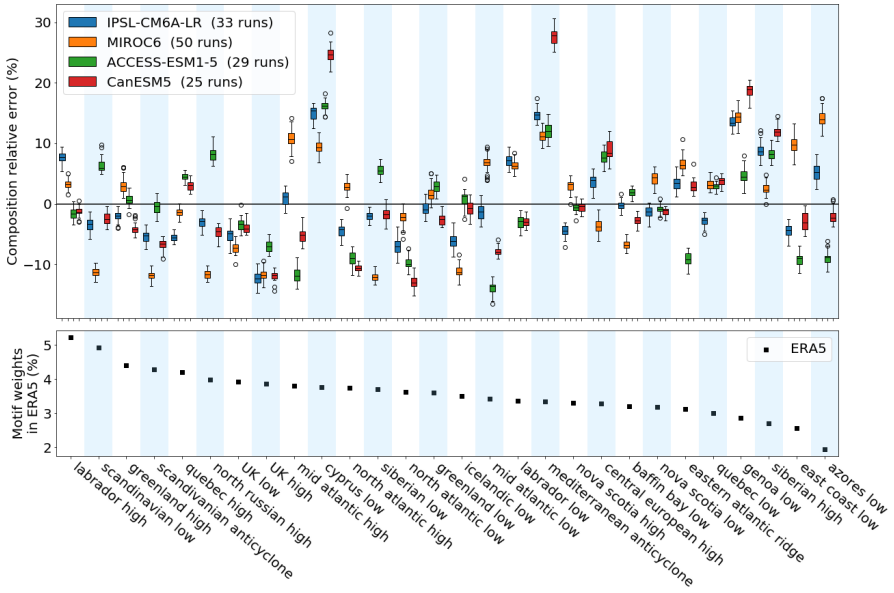


Figure 5.9 – Top: Relative error on average motif weight between models and ERA5 reanalysis. The box edges correspond to 1st and 3rd quartiles. The black line is the median. The whiskers extend to the furthest datapoint, up to 1.5 times the difference between the 1st and 3rd quartiles. Datapoints beyond the whiskers are represented as colorless circles. Bottom: average motif weight in the synoptic configuration of ERA5 fields.

Here, we see that median relative errors are mostly contained within $\pm 20\%$, which is relatively small. In particular, all model relative errors are less than 15% for the eight most prevalent motifs in the reanalysis. Therefore, models reproduce well the synoptic configuration observed in reanalysis. Relative errors made by **IPSL-CM6A-LR**, **MIROC6** and **ACCESS-ESM1.5** are all below 20%. The largest errors, with a median reaching 25%, are made by **CanESM5** on the Mediterranean anticyclone and Cyprus low motifs. It is possible that these larger errors are due to the lower resolution of **CanESM5**, compared to the other models (see table 5.1).

Models can be evaluated and compared based on the errors they make relative to reanalysis. However, another key factor to model evaluation is to compare the errors with model inner variability, represented here by the height of the boxes. If the inner variability is smaller than the errors, it shows that

all runs of the model make similar predictions, and indicates the presence of a bias inherent to the model. Here, the inner variability of the models is significantly smaller than the error in 96 out of 112 cases, so 87.5% of the time. Therefore, the models do have biases, deviating from the motif weights that characterize atmospheric circulation in ERA5 by up to around 20%.

Additionally, we observe that the motifs associated with largest relative errors tend to be the same from one model to another. A multimodel ensemble method would therefore not eliminate these biases. The largest errors are made on motifs located on the Mediterranean region. Specifically, the Mediterranean anticyclone and Cyprus low motifs weights are systematically over-estimated in every run of every model. Every model run also over-represents Genoa low and under-represents UK high and low. Finally, all models but **ACCESS-ESM1.5** systematically under-represent the Scandinavian anticyclone, which is the fourth most prevalent motif in the reanalysis, with an average weight of more than 4%. These similarities in model errors suggest that the origin of the errors could be common to all models.

In the end, our results indicate that average atmospheric circulation is generally well represented by the models, according to the angle of study of LDA-based synoptic configurations. Models still contain biases, however, of relatively low, but not insignificant, amplitude. Some biases are different from model to model, and can therefore be eliminated, or at least diminished, by ensemble methods. Some biases, however, are common to all the considered models.

5.3.2 . Cold spells in France

In order to study how well models represent the circulation associated with extreme events, we now apply this method to extreme temperature events data, starting with **Cold Spells** occurring in France. To begin with, we evaluate the ability of models and the LDA basis weights to reproduce the overall structure of SLP anomalies associated with extremes. Real and reconstructed averages of **Cold Spell** daily maps, for ERA5 and each model, are shown in figure 5.10. The real average is obtained by averaging over daily maps associated with a **Cold Spell**. The reconstructed average is obtained from the average motif weights of daily maps associated with a **Cold Spell**, using equation 4.3. To identify the most significant motifs associated with each model, the two most prevalent cyclonic and the two most prevalent anticyclonic motifs in each case are annotated on the figure. We observe that the overall synoptic structure associated with French **Cold Spells** consists of an anticyclonic structure in the north, and a cyclonic structure in the south. The corridor between

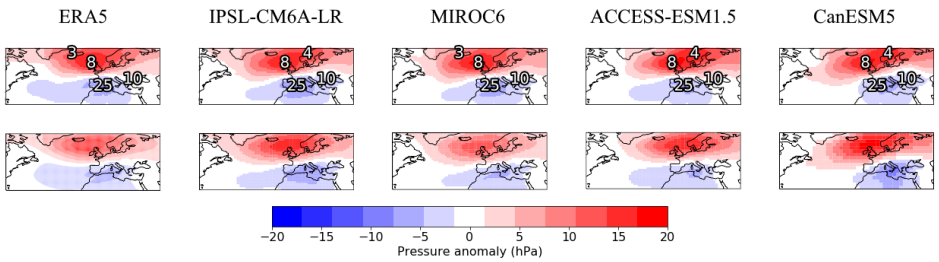


Figure 5.10 – Top line: Reconstruction of the average motif composition of **Cold Spells** in France according to different models (columns). The two cyclones and the two anticyclones with highest average weights in each case are annotated. Bottom line: Average SLP field for **Cold Spells** in France according to different models (columns).

the two is slanted northeast-southwest, passing through the middle of France.

First, we compare real average anomaly maps with the corresponding LDA reconstructions. We find that, for all models, the reconstructed average reproduces the overall structure, as well as most of the detail of the shapes of the high and low pressures. This confirms that LDA captures the synoptic information contained in the real maps. Then, we compare models with reanalysis. Model average maps are in good agreement with ERA5. They have the same two most prevalent cyclones as ERA5, Cyprus low and Genoa low, and reproduce motif 8, UK high, as a dominant motif. However, some discrepancies are present. All models underestimate the westward extent of the anticyclonic structure over the Atlantic. Only **MIROC6** reproduces the fact that Greenland high (motif 3) is more prevalent than Scandinavian anticyclone (motif 4), though as seen in section 5.2, Greenland high and Scandinavian anticyclone are both relevant for French **Cold Spells** (near 8% weights).

Model relative errors on synoptic configurations in the case of **Cold Spells** occurring in France are shown in figure 5.11, using box plots.

The amplitudes of model errors are significantly higher in the **Cold Spell** case than for the general case. The variability among the runs of each models is also higher. The five most prevalent reanalysis motifs during French **Cold Spells** are UK high, Greenland high, Scandinavian anticyclone, Genoa low, and Central European high. The significantly higher weights of UK high and Genoa low during **Cold Spells** are well reproduced by all models with an error within the internal variability of all four models. The central European high weight is also well represented by all models except by **ACCESS-ESM1.5** which overestimates it by at least 25%. The weight of Scandinavian antic-

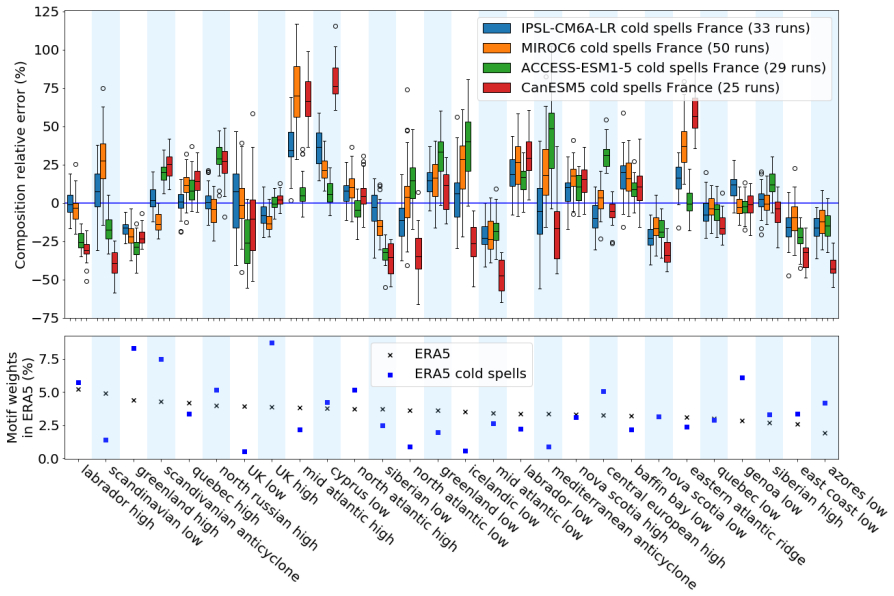


Figure 5.11 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in France. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

clone is well reproduced by **IPSL-CM6A-LR** and **MIROC6**. However, it is overestimated by around 25% by the two other models. All four models underrepresent the Greenland high motif by around 25%. Larger relative errors are made on less relevant motifs where reanalysis weights are lower. The most over-represented motifs are Cyprus low and Mid-Atlantic high for all models except **ACCESS-ESM1.5**. We note that larger relative errors are generally observed for the lower resolution model **CanESM5**.

In the end, our study shows that models have significantly larger biases in their representation of the average atmospheric circulation associated with **Cold Spells** in France. Many of these biases are shared among the models, and therefore cannot be eliminated by ensemble methods. One possible conclusion would be that models do not use, on average, the same circulation patterns to cause **Cold Spells** as reanalysis.

5.3.3 . Heatwaves in France

We apply the same methodology to **Heatwaves** occurring in France. Similarly to the previous case, we first evaluate the ability of models and projection onto LDA basis to reproduce the overall structure of SLP anomalies as-

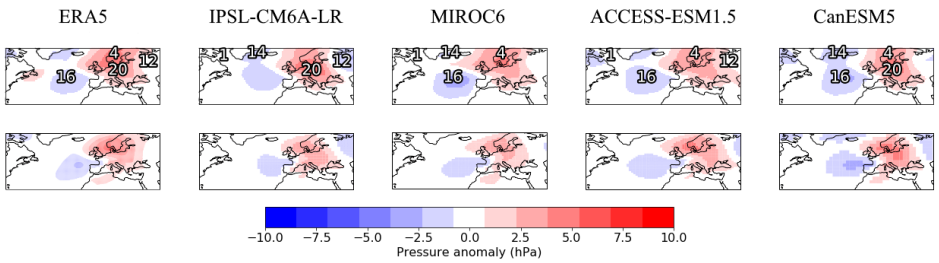


Figure 5.12 – Top line: Reconstruction of the average motif composition of **Heatwaves** in France according to different models (columns). The two cyclones and the two anticyclones with highest average weights in each case are annotated. Bottom line: Average SLP field for **Heatwaves** in France according to different models (columns).

sociated with extremes. Real and reconstructed averages of **Heatwave** daily maps, for ERA5 and each model, are shown in figure 5.12. To identify the most significant motifs in each case, the two most prevalent cyclonic and the two most prevalent anticyclonic motifs are annotated on the figure. SLP anomaly values are weaker than in the **Cold Spell** case. This is because **Heatwaves** are more varied in configuration, leading to average anomaly values closer to zero.

In the **Heatwave** case, there are differences between the real and reconstructed maps. In every dataset, the anticyclonic structure over Europe has a crescent-like shape around the Atlantic cyclone. However, it appears in reconstruction with an arrow-like shape, without an extension over north Africa. Despite this, the overall structure consisting of anticyclones over northern and central Europe, with a depression over the Atlantic is preserved by LDA reconstruction. Regarding model performance, we find that similarly to the **Cold Spell** case, model reproduce the overall structure of ERA5 average circulation, with anticyclonic conditions on northern and central Europe and cyclones over the Atlantic. However, models disagree with ERA5 and each other on the detail of the shapes of those synoptic structures. The most prevalent anticyclones in reanalysis data are the Scandinavian anticyclone (motif 4) and the Central European high (motif 20). Only **CanESM5** reproduces this property. For the other models, this leads to an anticyclonic structure that is weaker in the north for **IPSL-CM6A-LR**, and in the south for **MIROC6** and **ACCESS-ESM1.5**. The most prevalent cyclones in reanalysis data are Siberian low (motif 12) and Mid-Atlantic low (motif 16). Only **ACCESS-ESM1.5** reproduces this property, which is shown by deviations from ERA5 in the other models regarding the shape and intensities of the synoptic structures.

Relative errors in motif weights between reanalyses and models for **Heatwaves** occurring in France are computed and shown in figure 5.13.

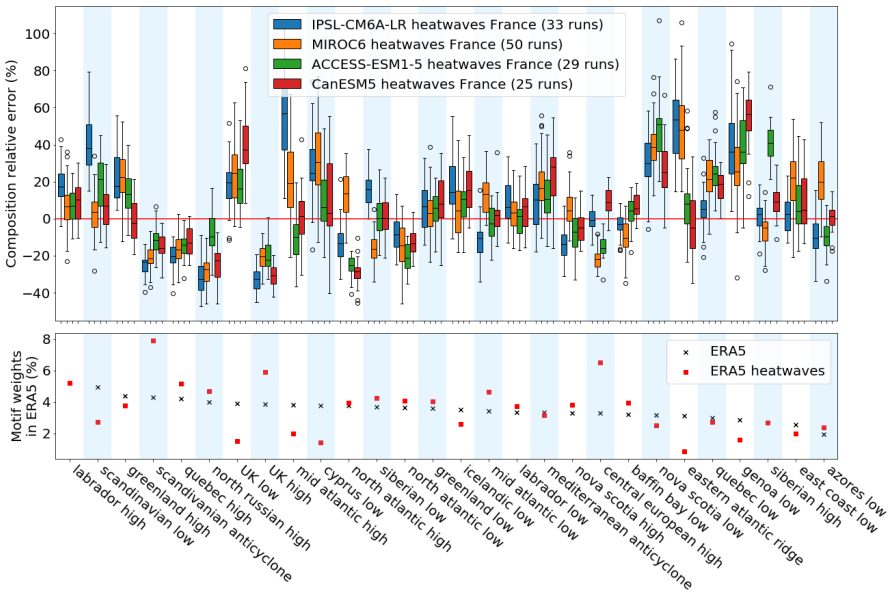


Figure 5.13 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in France. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

Similarly to the **Cold Spell** case, both relative errors and internal variabilities are higher in the **Heatwave** case than in the general case. Which motifs are or are not relevant to **Heatwaves** in France is generally well reproduced by the models. However, the most relevant motifs tend to be under-represented. Every model systematically under-represents the contribution of the most prevalent motif, which is the Scandinavian anticyclone, except **ACCESS-ESM1.5**, which still underestimate its average weight by more than 10% on median. The weight of the second most prevalent motif, central European high, is well represented by **IPSL-CM6A-LR** and **CanESM5**, but underestimated by about 20% by **MIROC6** and **ACCESS-ESM1.5**. UK high, the third most prevalent motif, is under-represented by 20% or more by almost all runs of all models. In general, motifs that have higher weights than in the general case tend to be under-represented (including Quebec high and north Russian high), while motifs that have lower weights (UK low, Nova Scotia low, and Genoa low) are over-represented. This shows that models underestimate the changes in SLP patterns associated with **Heatwaves**.

In the end, similarly to the **Cold Spell** case, we find that models display significant and shared biases in their representation of the average circulation associated with **Heatwaves** in France. As was said for **Cold Spells**, a possible conclusion is that, on average, models do not use the same circulation patterns to cause extreme events.

5.3.4 . Cold spells across European countries

We have established that our method can evaluate the ability of models to reproduce atmospheric circulation patterns associated with extreme events. In order to explore the possible geographical dependencies of model errors, we now apply the same methods to other western European countries. The list of countries studied is the same as previously, France, Germany, Poland, Italy, Spain, and the UK. We first look at **Cold Spells**.

Model relative errors on synoptic configurations in the case of **Cold Spells** occurring in Germany are shown in figure 5.14, using box plots. Model errors

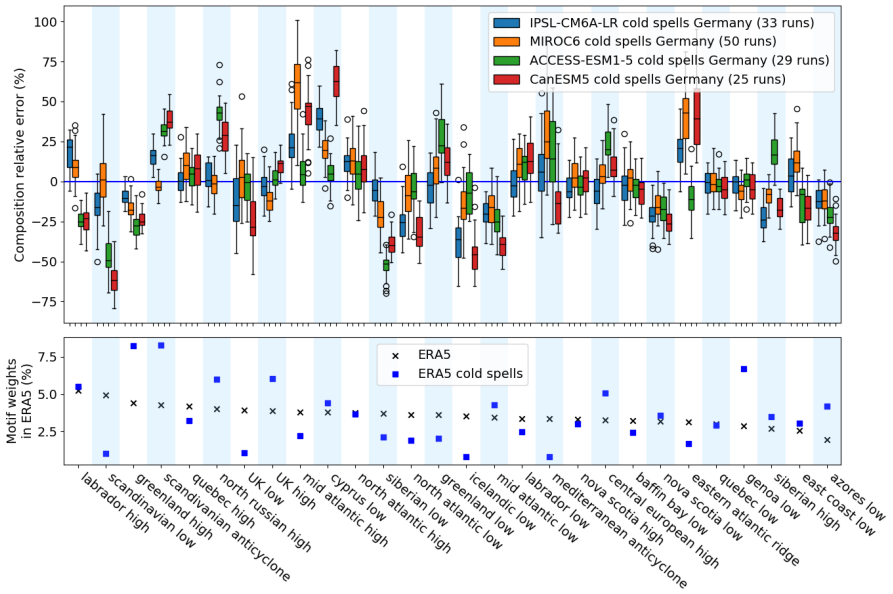


Figure 5.14 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Germany. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

are higher than in the general case, on a scale similar to what was observed for France. Furthermore, model errors are mostly the same as in the case of France, with models generally over-representing and under-representing

the same motifs, by similar amounts. This is an expected result, considering the geographical proximity of the two countries. As some of the countries we consider display similar behavior from the models, we chose not to go over all of them in this part, so as to avoid repetition. We will only look at Italy and UK, ignoring Poland, which is similar to Germany and France, and Spain, which is similar to Italy. Figures representing the data for the other countries are available in Appendix A.

Figure 5.15 shows model relative errors on synoptic configurations in the case of **Cold Spells** occurring in Italy. We find that in the case of **Cold Spell** occur-

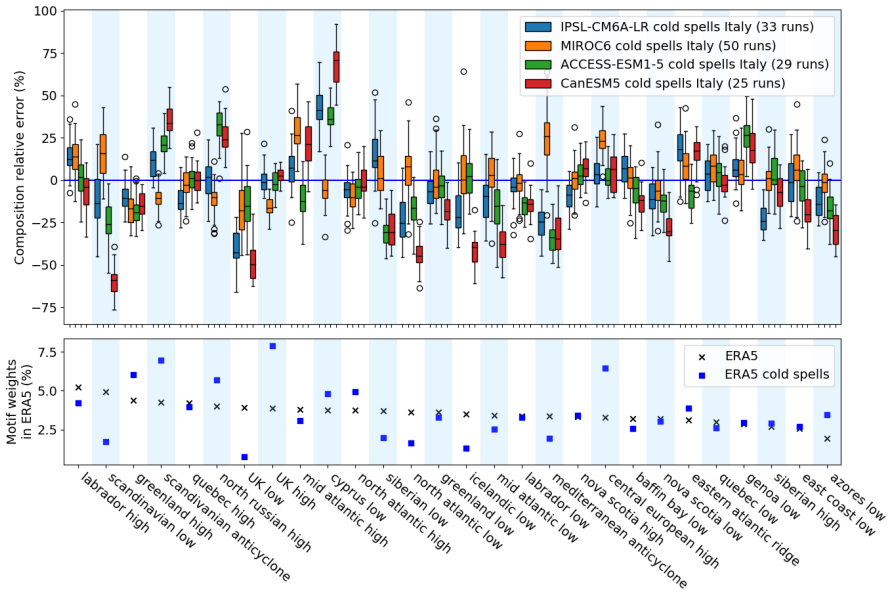


Figure 5.15 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Italy. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

ring in Italy, model relative errors differ from the France and Germany cases. **MIROC6** is the only model that does not over-represent Cyprus low, while it was **ACCESS-ESM1.5** for France and Germany. Inversely, **MIROC6** is the only model to systematically over-represent central European high, which is a key motif in this case, while **ACCESS-ESM1.5** was the only model to do this for France and Germany. The relative errors are overall lower on Genoa low, and models over-represent significantly less some motifs such as mid Atlantic high and eastern Atlantic ridge.

Figure 5.16 shows model relative errors on synoptic configurations in the case

of **Cold Spells** occurring in the UK. This case is mainly characterized by high

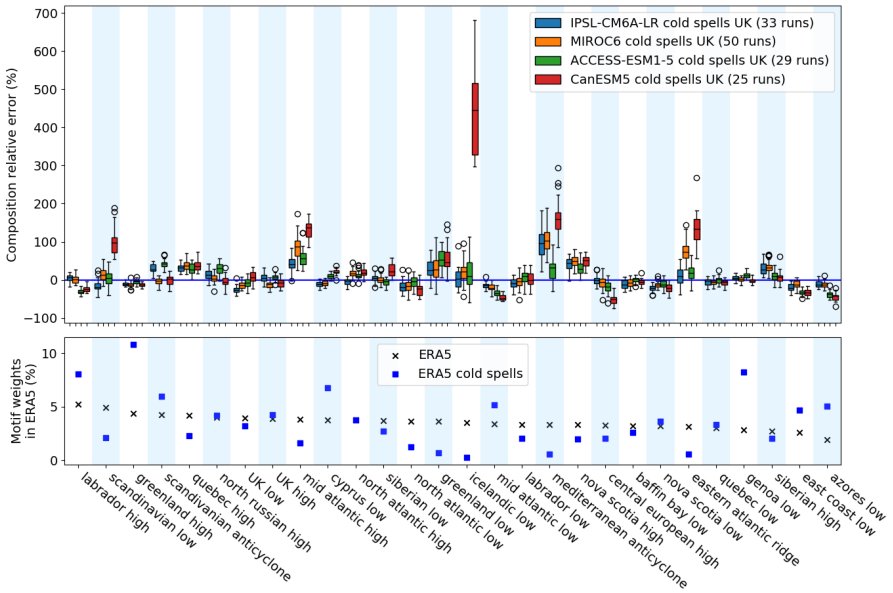


Figure 5.16 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in the UK. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

overestimation of some motif weights, especially by the model **CanESM5**. This concerns mainly Icelandic low, a model with low average weight in this case which is overestimated by a factor between 3 and 6 by this model. **CanESM5** also severely over-represents the Scandinavian low, mid Atlantic high, Mediterranean anticyclone and eastern Atlantic ridges by up to 200%. While the reanalysis value of the Icelandic low weight is within model variability for the three other models, they also over-represent, although less severely, the mid Atlantic high. **MIROC6** over-represents the eastern Atlantic ridge and Mediterranean anticyclone as well. **IPSL-CM6A-LR** over-represents the Mediterranean anticyclone, and under-represents the Scandinavian low. All four models also overestimate the weights of the Quebec high, as well as the Nova Scotia high.

5.3.5 . Heatwaves across European countries

We now apply the same method to **Heatwaves**. figure 5.17 shows model relative errors on synoptic configurations in the case of **Heatwaves** occurring in Germany. We find that models generally reproduce the significantly higher weight of the Scandinavian anticyclone motifs compared to the general case. However, it is systematically underestimated by all four models by around

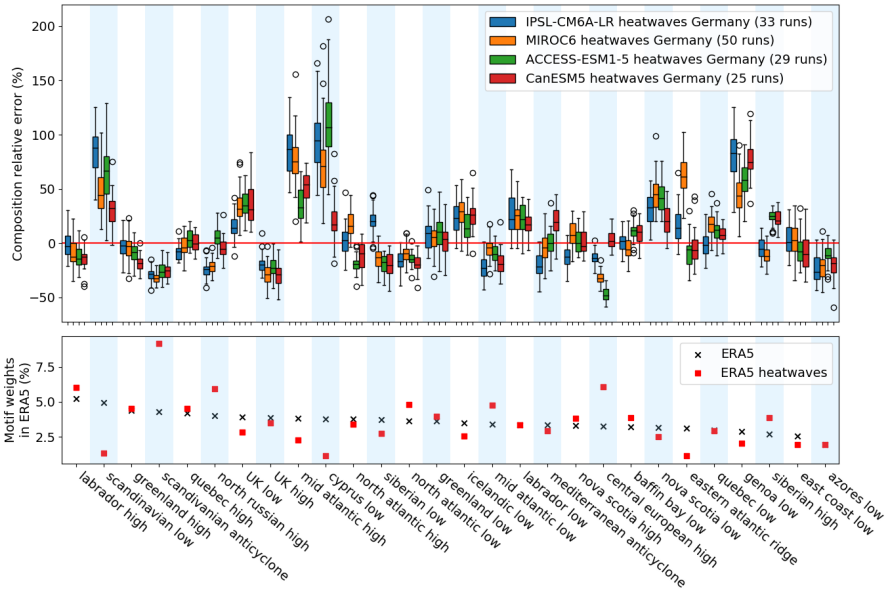


Figure 5.17 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Germany. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

25%. The highest relative errors are committed on motifs that are not relevant to German **Heatwaves**, that models tend to over-represent. This includes Scandinavian low, mid-Atlantic high, Cyprus low, or Genoa low. Relative errors on these motifs can reach 100%. Similarly to the case of France, it appears that models underestimate the changes in atmospheric circulation associated with **Heatwaves**. However, some differences exist. Overestimation of Scandinavian low weight is more severe in this case than for France. **IPSL-CM6A-LR** under-represents the central European high, while **MIROC6** and **ACCESS-ESM1.5** do so more severely. However, the errors on most motifs are of similar sign and amplitude than in the case of **Heatwaves** occurring in France. For similar reasons as in the **Cold Spell** case, all six countries will not be examined in detail here. Instead, we will examine Germany, Italy, and the UK. Figures relating to the other countries are included in Appendix A.

Figure 5.18 shows model relative errors on synoptic configurations in the case of **Heatwaves** occurring in Italy. The synoptic configuration associated with **Heatwaves** in Italy is closer to the general case than in the other countries. As one of the main differences, Cyprus low weight is significantly lower during extreme heat. This difference is not well represented by the models, who overestimate this weight in almost every case by a factor that reaches 100%

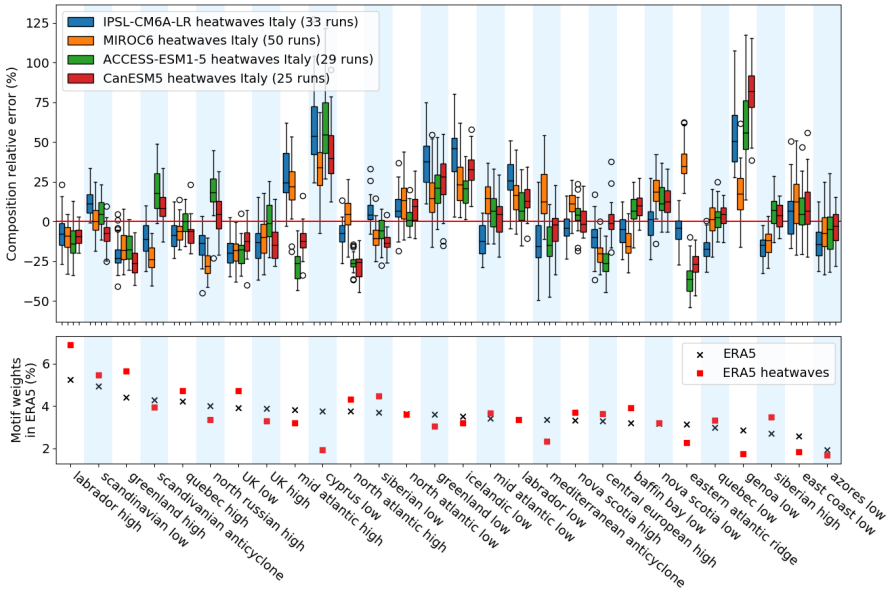


Figure 5.18 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Italy. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

for **IPSL-CM6A-LR** and **ACCESS-ESM1.5**. Models disagree on the weight associated to the eastern Atlantic ridge, with it being systematically overestimated by **MIROC6**, and systematically underestimated by **ACCESS-ESM1.5** and **CanESM5**. Unlike in Germany and France, the weight associated with the UK low is underestimated by the models, and the weight associated with the Icelandic low is systematically overestimated.

Figure 5.19 shows model relative errors on synoptic configurations in the case of **Heatwaves** occurring in the UK. The synoptic configuration associated with extreme heat in the UK is dominated by the Scandinavian anticyclone, by a wider margin than in the other countries. Its average weight is underestimated by **IPSL-CM6A-LR**, **MIROC6** and **ACCESS-ESM1.5**. However, **CanESM5** reproduces the value observed in reanalysis. Models disagree on the average weight of Scandinavian low. It is underestimated by **CanESM5**, and overestimated by **IPSL-CM6A-LR** and **ACCESS-ESM1.5**.

5.3.6 . Discussion

To sum up, when considering variations across several European countries, we reach similar conclusions for both **Heatwaves** and **Cold Spells**. Models

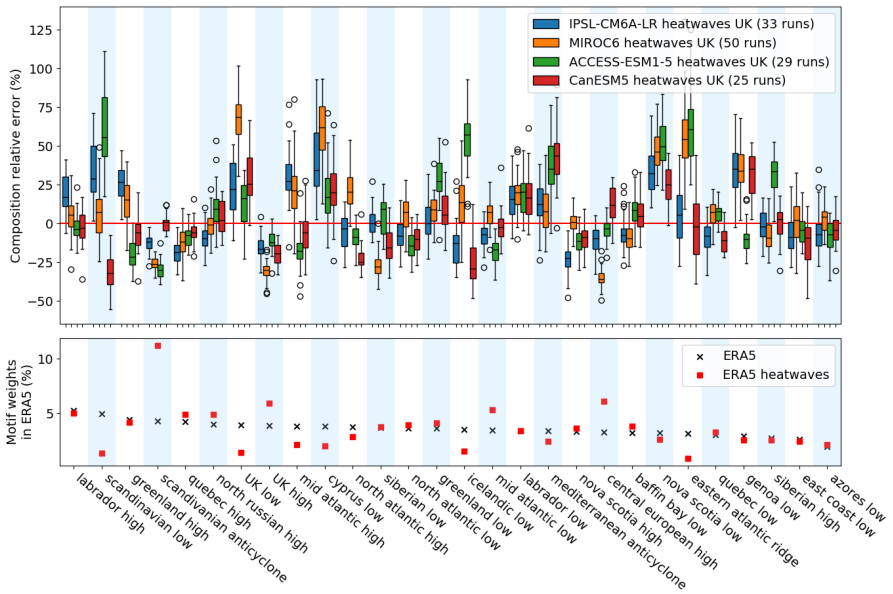


Figure 5.19 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in the UK. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

make relative errors on extreme synoptic configurations that are of similar scales for every considered country. This implies that, in every country, models associate extreme temperature events with different synoptic configurations than reanalysis. Furthermore, the errors made by models tend to have the same sign on most countries, for most motifs. This especially the case for neighbouring countries, such as France and Germany, where models misrepresentations are similar.

However, we observe variations in the localizations and amplitudes of model errors, depending on the region of interest. For examples on **Cold Spells**, we can cite **MIROC6** and **ACCESS-ESM1.5** switch their errors on the central European high and Cyprus low when looking at Italy and France, or the fact that extreme cold in the UK is heavily mis-represented by **CanESM5** especially, which over-represents motifs of little relevance. For examples on **Heatwaves**, models fail to capture the much lower weight of the Cyprus low during extreme heat in Italy, nor the much higher weight of the Scandinavian anticyclone during extreme heat in the UK.

By providing these detailed and local estimation of model representation errors, LDA allows informed preselection of a model best suited for a specific

task. As examples, among the four models we considered, **CanESM5** would be best suited to represent circulation in the north during **Heatwaves** in the UK, but **IPSL-CM6A-LR** would be preferable for the representation of key circulation patterns during **Cold Spells** in Italy.

5.4 . Summary

In this chapter, we have demonstrated that LDA, through its ability to characterize the atmospheric circulation of datasets, can be used as a tool for evaluating model representation of synoptic configurations observed in reanalysis data. Due to the locality of the LDA motifs, we were able to measure model representation error locally. This was done for four CMIP6 models, **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**, in the general case, and for **Cold Spells** and **Heatwaves** in several European countries.

We have been able to detect several biases, either shared or model-specific, including a general over-representation of synoptic objects located on the Mediterranean sea. We have shown that model representation of circulation associated with extreme events is associated with significant errors, higher than in the general case. The errors are of a similar order of magnitude for extremes located in every examined country.

However, each extreme event is associated with errors that vary model-to-model. As such, given a specific type of event, and a location of interest, the LDA study of synoptic configurations is able to rank the representations of atmospheric circulation by the models. This provides a tool for preselecting models best suited to provide useful data on specific extreme events. Furthermore, this method can be generalized to any kind of extreme event that can be automatically detected in the data.

6 - Higher-dimensional evaluation criteria

6.1 . Introduction

We have shown in chapter 5 that the synoptic configurations, defined as the sets of average weights associated with a given basis of motifs, provide a characterization of the atmospheric circulation. This average can be either unconditional or conditional on specific events. Comparison of synoptic configurations provides a local measure of the differences in circulation patterns, in particular those associated with extreme events, and can be used to evaluate the accuracy of a model. In this chapter, we discuss two possible extensions of our evaluation protocol, in order to obtain a fuller characterization of the model relevance.

Firstly, we examine in more detail the characteristics of the weights. The motif representation is sparse: in any individual map, many of the motifs will have negligible weights. This intermittency effect is averaged out in our measure of synoptic configuration. A lower (resp. higher) average weight of a motif could be due to one or both of two possible effects:

- The motif has a negligible weight (i.e. is absent) in a higher (resp. lower) proportion of the maps.
- The weight values of the motif, when not negligible (i.e. present), are lower (resp. higher).

Our previous analysis does not allow us to differentiate between these two causes for a change in the average value of the weights. In this part, we will use this intermittency information to produce a more detailed characterization of atmospheric circulation, and apply it for model evaluation.

Secondly, while differences between synoptic configurations in reanalysis and models can be used to define an error on model representation of dynamics, another key aspect of evaluating climate models is the representation of temperature. We therefore introduce another component to our model evaluation protocol: the representation error of the average temperature. We explore how the conjunction of these two evaluation metrics can, or cannot, be used to discriminate between the four models: **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**, in the general case, **Cold Spells**, and **Heatwaves**. Moreover, we consider the correlation between temperature-based and dynamics-based model representation errors. Model errors on temperature can be caused either by a misrepresentation of circulation, which advects air from different regions, or by a misrepresentation of thermodynamics, which governs the exchanges of heat between air masses and their en-

vironment (see section 1.2). Evidence of correlation could therefore indicate a possible causal link between model dynamic error and model temperature error.

6.2 . Motifs frequencies and intensities

6.2.1 . Defining motif presence

The logarithmic histograms of all motifs weights in ERA5 as well as in run 1 of each model is shown in figure 6.1. All motifs, in each dataset, show a bi-

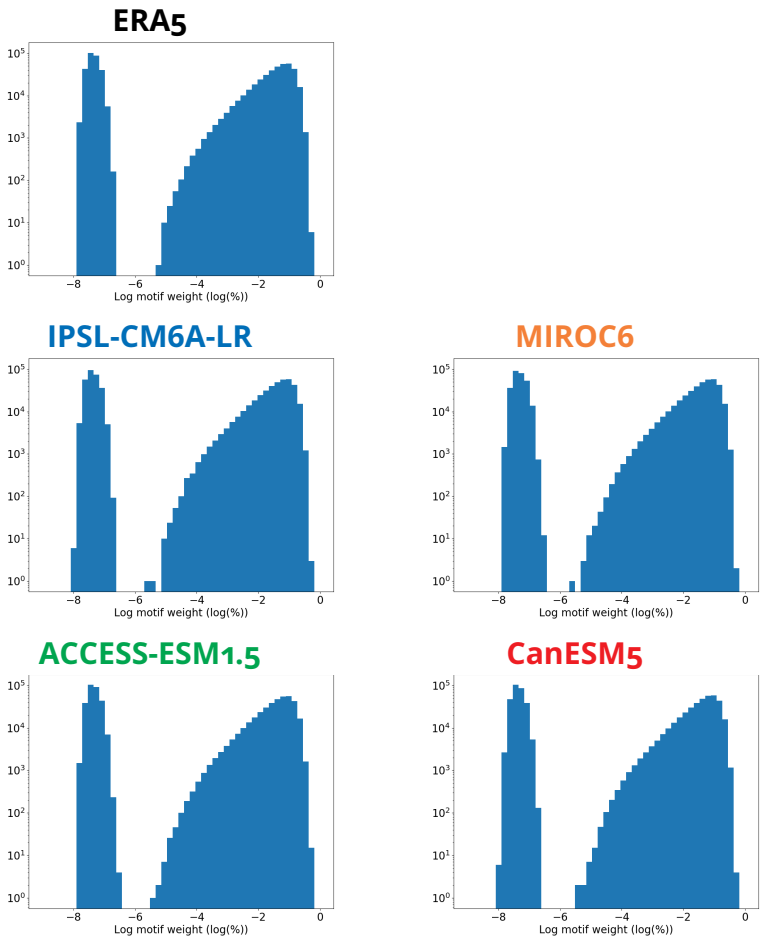


Figure 6.1 – Logarithmic histogram of motif weights, all 28 motifs included. Left to right, top to bottom: ERA5, run 1 of **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**.

modal distribution, separated by a gap located around 10^{-6}. This shows that motifs are intermittent, and can be used to define the notion of motif *pres-*

ence or *absence*. A motif is *present* in a map if it has a weight higher than $\epsilon = \exp -14 \approx 8.3 \times 10^{-7}$, and *absent* otherwise. This likely overestimates the presence of motifs in the representation, as probabilities in the lower end of the high probability lobe are much lower than those in the higher end, and a motif weight can still be negligible despite being present. However, it has the advantage of being a robust way of defining intermittency, as the results do not change with small variations of ϵ. Mathematically, the presence of motif k in map d is noted $[\boldsymbol{c}_k(d) > \epsilon]$, where $[\cdot]$ is the Iverson bracket:

$$[P] = \begin{cases} 1 \text{ if P is true} \\ 0 \text{ otherwise} \end{cases} \tag{6.1}$$

This idea can be further extended into the notion of motif *frequency*, i.e. the proportion of maps a motif is present in. The frequency of motif k in a dataset is noted $\boldsymbol{f}_k$ and is defined in equation 6.2.

$$\boldsymbol{f}_k = \langle [\boldsymbol{c}_k(d) > \epsilon] \rangle \tag{6.2}$$

Where $\langle \cdot \rangle$ designates the average over $d \in \{1, ..., D\}$.

From now on, we will consider all values in the left lobe to be 0, which makes no difference within round-off errors. The frequency of a motif in a dataset contains a portion of the information that makes up the average weight. Specifically, it contains the information of: "*how many maps the motif plays a role in*". The part of the motif average weight that is not contained in frequencies is "*the value of the weight when it is non-zero*". We call "intensity" of a motif in a dataset the average of these values. For a motif k, let it be noted $\boldsymbol{i}_k$. Its definition is provided in equation 6.3.

$$\boldsymbol{i}_k = \langle \boldsymbol{c}_k(d) \times [\boldsymbol{c}_k(d) > \epsilon] \rangle \tag{6.3}$$

Within round-off error, the average weight of a motif is the product of its frequency and intensity. We note that these are themselves values aggregated using average.

$$\forall k \in \{1, ..., K\}, \boldsymbol{f}_k \times \boldsymbol{i}_k = \langle \boldsymbol{c}_k(\boldsymbol{X}) \rangle \tag{6.4}$$

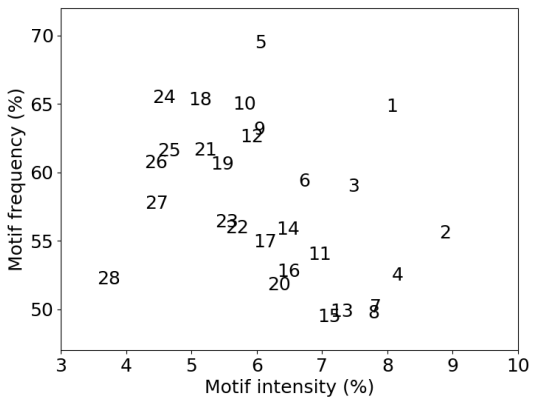
Thus, differences in motif weights between two datasets can be attributed in proportion to these two factors.

However, while motif frequencies are physically interpretable, the $\boldsymbol{i}_k$ do not have a direct physical interpretation. This is due to two reasons. To begin with,

LDA has no model of the norm of a map - the total sum of all absolute field values. As a result, an SLP anomaly map which is twice as intense but with the same distribution as another would be treated as identical by LDA. Therefore, what we would interpret as the physical intensity of a motif in an SLP map - say, how strong an anticyclone is at a given location - is in fact not an information represented in the motif weight. Furthermore, maps are associated with a composition in motifs, which must always sum to 1. Therefore, a motif may have a lower weight simply because more motifs are present, rather than because it appears with lower intensity, even relatively to the other motifs. As a result, while we will refer to the $\boldsymbol{i}_k$ as "motif intensities" in the following, they are not to be understood as the physical intensity of the corresponding motif in the SLP anomaly map. Intensities and frequencies are the two elements that make up the average weight, and the two sources that can participate to its changes between datasets. However, since only the frequencies have a clear physical interpretation, we will mainly focus on them in the rest of this study.

6.2.2 . Motif frequencies and intensities in ERA5

We now show how the computation of motif frequencies and intensities can improve the characterization of atmospheric circulation provided by average weights. This can be applied to every case - models, extreme events, regions of interest - studied in chapter 5. In this part, we will apply this enriched characterization on some of these cases, to explore what can be learned from it.



Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.2 – Motif frequencies versus motif intensities, in ERA5. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

We plot motif intensities versus frequencies in ERA5 in figure 6.2. Each motif is annotated with its motif index. We recall that motifs are ordered by decreasing average weight. Unsurprisingly, the most prevalent motif (1, Labrador high) has both a high intensity and high frequency, while the least prevalent motif (28, Azores low) has a low intensity and frequency. However, the highest frequency is associated with motif 5 (Quebec high), while the lowest corresponds to motif 15 (Icelandic low). Considering that some motifs are opposites of each other, it may seem counter-intuitive that almost every motif has a frequency of at least 50%. This is due to the fact that the gap with which we define the presence of motifs is very low, and there are therefore values of weight that are considered present despite still being negligible. The frequency measured with this indicator can be seen as overestimated compared to the proportion of maps a motif plays a significant role in. However, it remains relevant for comparison purposes.

We show the motif intensities and frequencies in ERA5 in 6.3, using the usual order of decreasing average weights. We find that neither intensities nor frequencies reproduce the prevalence order obtained for average weights. As we saw in figure 6.2, we find examples of motifs that have high weight due to a high frequency, despite a low intensity, such as the Quebec high (motif 5), as well as motifs that have high average weight despite a relatively low frequency, such as the Scandinavian anticyclone (motif 2). While both intensities and frequencies can be relevant, intensities are closer to reproducing the order of prevalence obtained from average weights. Furthermore, motif intensities are spread from 4% up to 9%, over a factor of 2.5. Motif frequencies, on the other hand, are spread over a factor of 1.5. Therefore, more of the variance observed for average weights is explained by intensities than by frequencies.

We also learn the following about synoptic configurations in ERA5. Motifs located on Quebec (Quebec high, motif 5, and Quebec low, motif 24) are particularly common. That is also the case of Mediterranean motifs, such as the Mediterranean anticyclone (motif 18), and the Genoa low (motif 10). On the opposite side, motifs located on the UK (motif 7 and 8) only appear in a lower percentage of maps. The features identified by the frequency/intensity splitting can then be compared with other datasets, such as models, or extreme events.

We now look at extreme temperature events. Our goal is to determine if the differences in synoptic configuration observed with the general case are due, or not, to motif frequency. We focus on extreme events occurring in France. In figure 6.4, we show motif intensities and frequencies during **Cold Spells**,

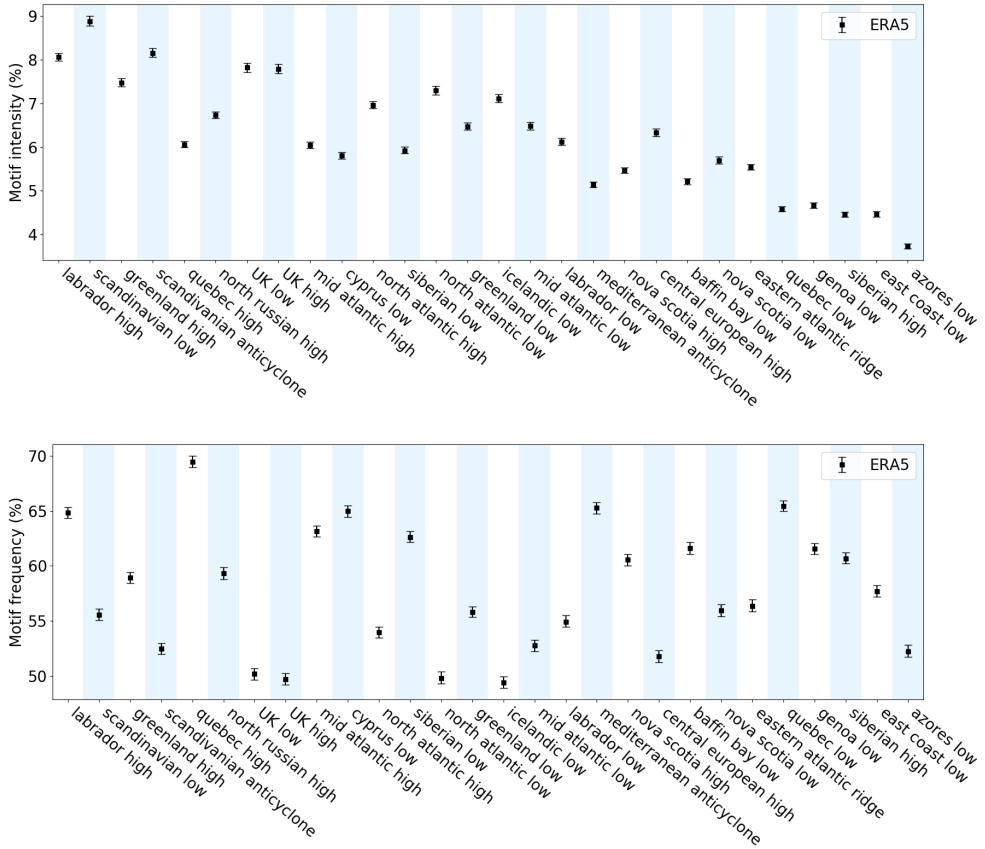


Figure 6.3 – Top: Average motif intensities in ERA5 SLP anomaly maps. Bottom: Average motif frequencies in ERA5 SLP anomaly maps. 90% confidence intervals are determined by bootstrapping.

according to ERA5.

We find that both motif intensities and motif frequencies display changes bigger that the 90% confidence interval when compared to the general case. Variations in frequencies across motifs are of much higher amplitude than in the general case. We also observe that the spread of frequency values is higher than for the general case, with frequencies reaching as low as 20% for the UK low. Most changes are shared between the two metrics: when one increases, the other does too, and vice versa. This stronger correlation between the two can be better observed in figure 6.5, where motif intensities are plotted versus frequencies in the case of ERA5 **Cold Spells**. Each motif is annotated with its motif index. This strong correlation, as well as the spread between the motifs, show that **Cold Spells** in France are associated with a specific synoptic configuration.

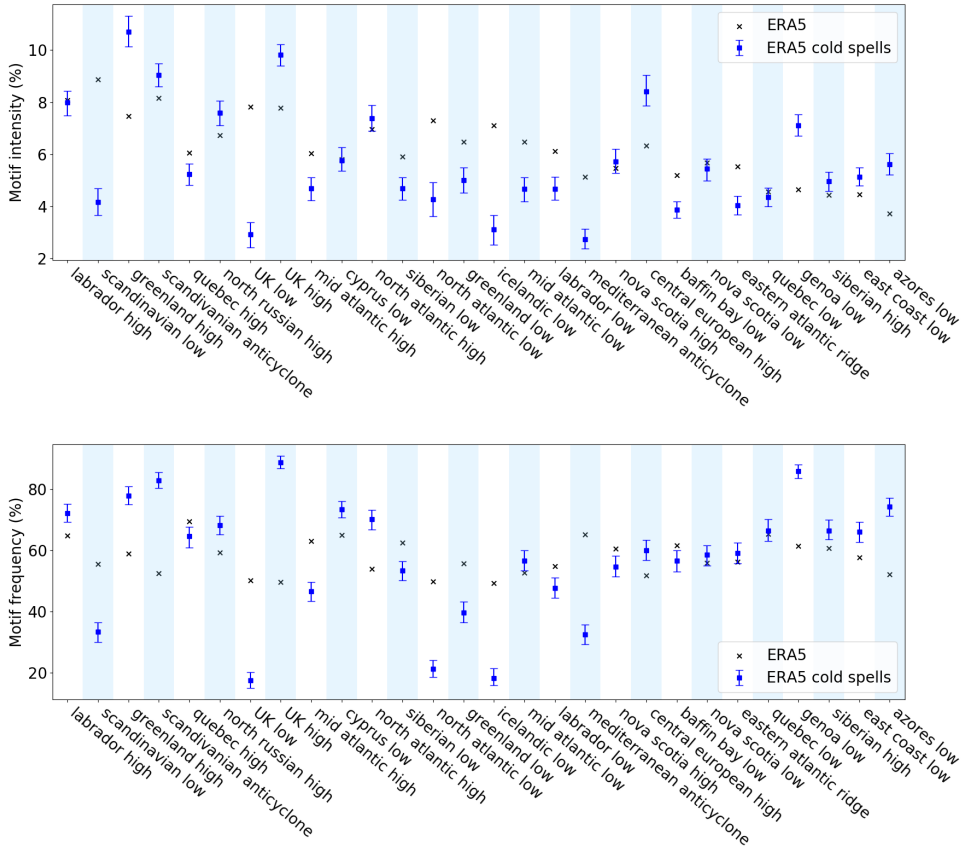
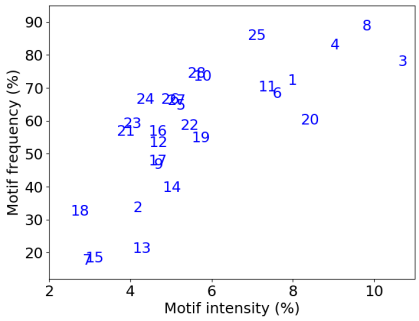


Figure 6.4 – Top: Motif intensities in ERA5 SLP anomaly maps, in the case of **Cold Spells** in France (blue), and in the general case (black). Bottom: Motif frequencies in ERA5 SLP anomaly maps, in the case of **Cold Spells** in France (blue), and in the general case (black). 90% confidence intervals are determined by bootstrapping.

In figure 6.6, we show motif intensities and frequencies during **Heatwaves**, according to ERA5. We observe variations in motif intensities and frequencies during **Heatwaves**, compared to the general case. In most cases, these variations are of higher amplitude than the 90% confidence interval.

For a more detailed analysis, intensities and frequencies are plotted against each other in 6.7, for the case of ERA5 **Heatwaves**. Each motif is annotated with its motif index. The spread in motif frequencies in reduced compared to the **Cold Spell** case. However, the spread in intensities is higher, with the Scandinavian anticyclone (motif 4) having more than 11% intensity. We also observe less correlation between frequencies and intensities than for the ex-



Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.5 – Motif frequencies versus motif intensities, in ERA5 **Cold Spells**. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

treme cold case.

In the end, the frequency/intensity split has allowed us to characterize atmospheric circulation in the ERA5 dataset further than average motif weights. We have obtained an indicator of motif intermittency, which can be used to compare the synoptic configuration of extreme events with the general case.

6.2.3 . Motif frequencies and intensities in models

We will now use the features identified by the frequency/intensity split as a model evaluation criterion. We plot motif intensities versus frequencies for general data in each model in figure 6.8. Each motif is annotated with its motif index, and the values are averaged over all model runs. We find that models generally reproduce the overall structure observed in ERA5. The Azores low (motif 28) has both low frequency and low intensity in every model, while the Labrador High (motif 1) has high frequency and intensity. Models also reproduce the occurrence of motifs with high frequency but low intensity, such as Quebec high (motif 5), or low frequency but high intensity, such as the Scandinavian low (motif 2). Furthermore, most motifs occupy a similar region of the intensity-frequency space in models and reanalysis.

We now look at **Cold Spells** events occurring in France. We plot motif intensities versus frequencies in **Cold Spells** for each model in figure 6.9. Each motif is annotated with its motif index, and the values are averaged over all model runs. All four models reproduce the larger variability and stronger correlation between frequencies and intensities compared to the general case as observed in ERA5. Motifs also seem to occupy the same overall region of

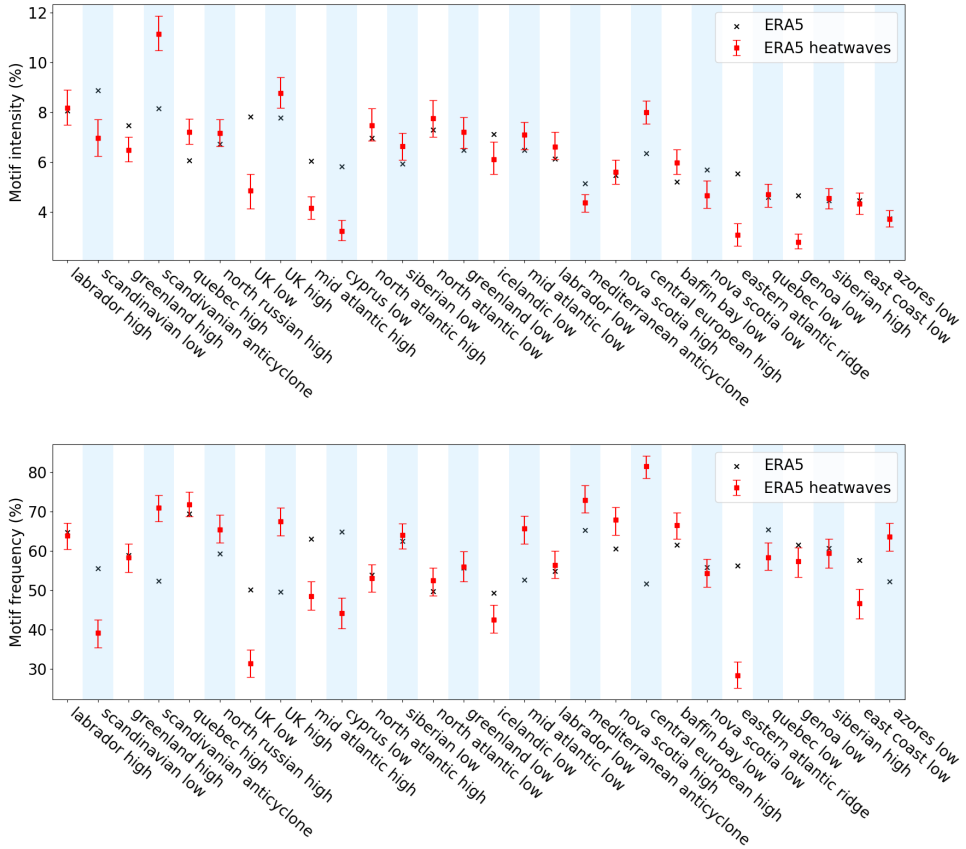
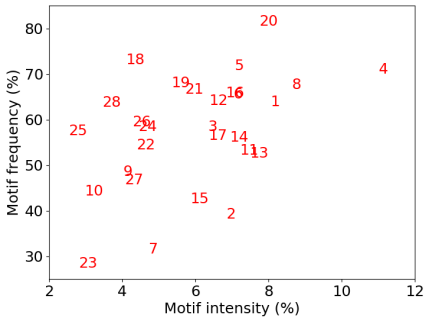


Figure 6.6 – Top: Motif intensities in ERA5 SLP anomaly maps, in the case of **Heatwaves** in France (red), and in the general case (black). Bottom: Motif frequencies in ERA5 SLP anomaly maps, in the case of **Heatwaves** in France (red), and in the general case (black). 90% confidence intervals are determined by bootstrapping.

intensity-frequency space than in ERA5, but there are bigger differences than for the general case. All models except **MIROC6** overestimate the intensity of Scandinavian low (motif 4) compared to Greenland high (motif 3). Furthermore, the central European high (motif 20) is mis-represented in different ways by each model. **IPSL-CM6A-LR** underestimates ts intensity, **MIROC6** overestimates its frequency, **ACCESS-ESM1.5** overestimates both its frequency and intensity, and **CanESM5** satisfyingly represents the ERA5 values.

We now look at **Heatwaves** events occurring in France. We plot motif intensities versus frequencies in **Heatwaves** for each model in figure 6.10. Each motif is annotated with its motif index, and the values are averaged over all model runs. In the extreme heat case, the spread among models is higher



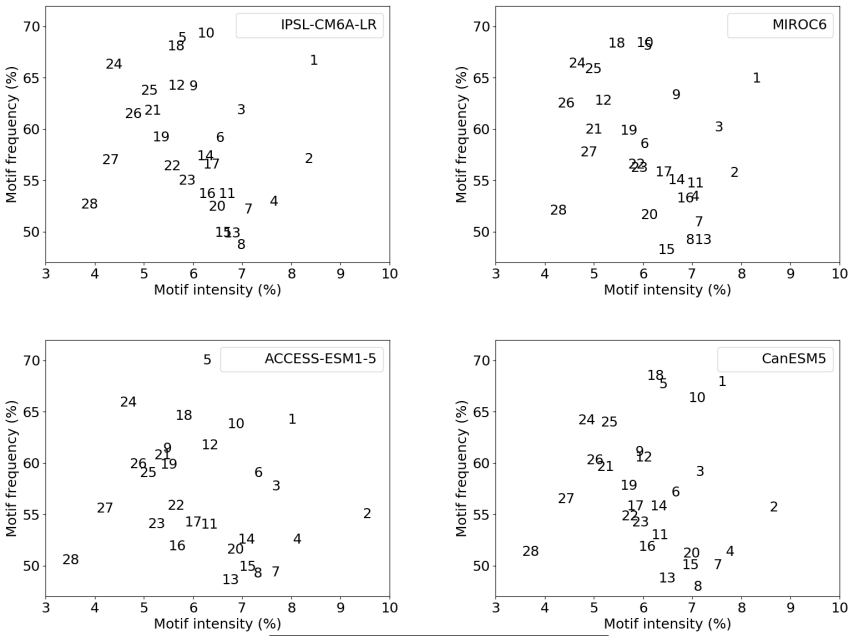
Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.7 – Motif frequencies versus motif intensities, in ERA5 **Heatwaves**. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

for both frequencies and intensities. No model reproduces the high intensity of Scandinavian high (motif 4) observed in ERA5. The high frequency of the central European high (motif 20) in ERA5 is reproduced by **IPSL-CM6A-LR** and **CanESM5**, but not **MIROC6** nor **ACCESS-ESM1.5**. On the other side, the low frequency of the eastern Atlantic ridge (motif 23) is reproduced by **ACCESS-ESM1.5** and **CanESM5**, but not **IPSL-CM6A-LR** nor **MIROC6**. The models also tend to overestimate, to various degrees, the frequency of the Scandinavian anticyclone (motif 2)

Similarly to what was done with average motif weight in chapter 5, we now examine the model relative error on the representation of motif intensities and frequencies, using boxplots. Relative errors in motif intensities and frequencies between reanalyses and models are shown in figure 6.11. Model represent relatively well the motif intensities and frequencies observed in ERA5. We find that model relative errors on intensities are of similar scale than for average motif weights. In fact, the main errors on intensities are the same that were observed for average weights. The Mediterranean motifs, the Cyprus low and the Mediterranean anticyclone, are overestimated by every run of every model. Relative errors in frequencies, on the other hand, behave differently. They reach lower values, only up to $\sim 8\%$, unlike the $\sim 25\%$ reached for intensities. **ACCESS-ESM1.5** underestimates the frequencies of both Mediterranean motifs. However, despite the lower errors on frequencies, inner model variability is still smaller than the error in a majority of cases.

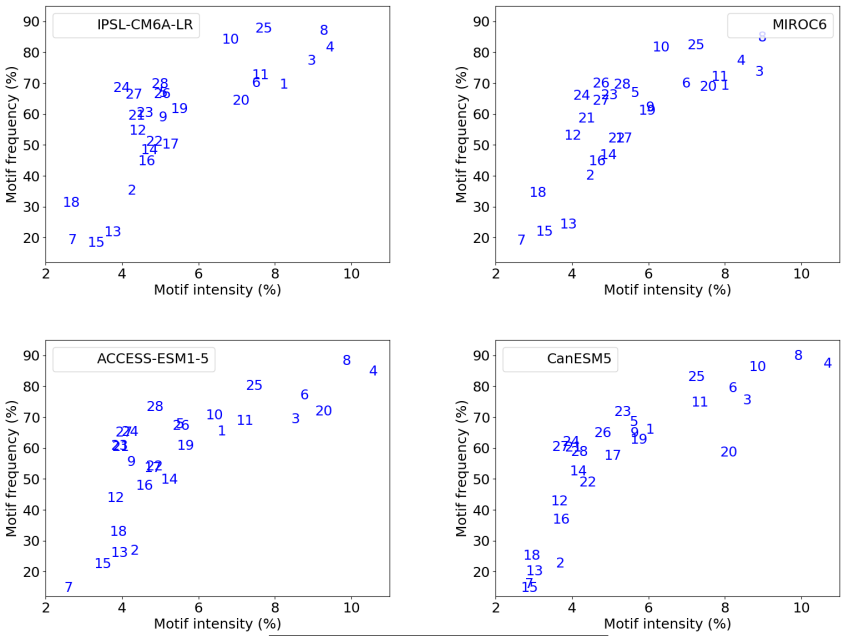
Relative errors in motif intensities and frequencies between reanalyses and models for **Cold Spells** occurring in France are shown in figure 6.12. Unlike the general case, the amplitude of relative errors on intensity and frequencies are



Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.8 – Motif frequencies versus motif intensities, averaged over all model runs, for each model. Top-left: **IPSL-CM6A-LR**. Top-right: **MIROC6**. Bottom-left: **ACCESS-ESM1.5**. Bottom-right: **CanESM5**. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

similar. However, the errors models make on intensities and frequencies are different. Models systematically underestimate the intensity of the Greenland high, but generally reproduce the frequency observed in ERA5. They systematically overestimate overestimate the frequency of the mid Atlantic high, and underestimate mid Atlantic low. Models also generally overestimate the frequencies of the Greenland low, as well as the Nova Scotia high, and underestimate that of the Nova Scotia low. The relative errors on intensities and frequencies both show differences with the errors on average weights seen in chapter 5. Therefore, the intensity/frequency split provides additional information to the characterization of the way models mis-represent atmospheric

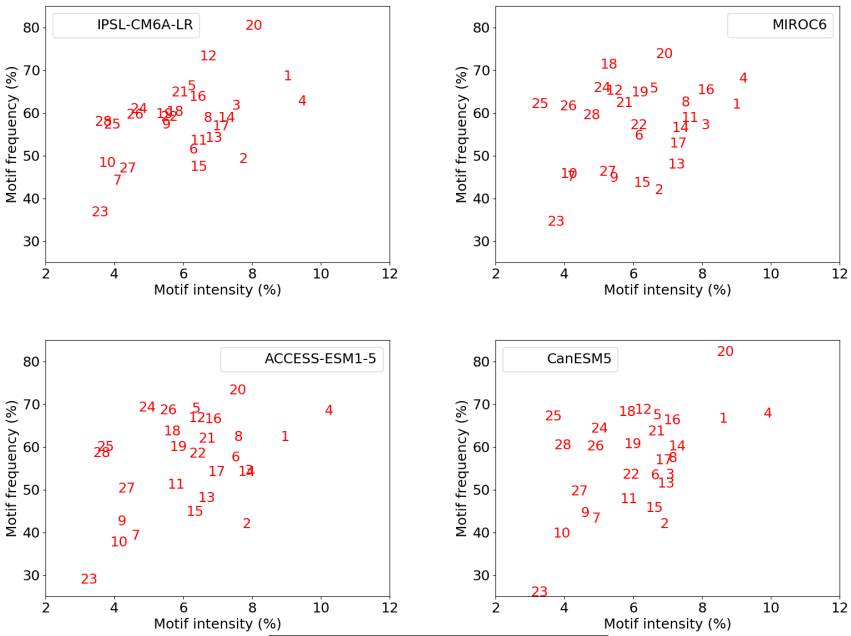


Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.9 – Motif frequencies versus motif intensities, averaged over all model runs, for model **Cold Spells**. Top-left: **IPSL-CM6A-LR**. Top-right: **MIROC6**. Bottom-left: **ACCESS-ESM1.5**. Bottom-right: **CanESM5**. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

circulation associated with **Cold Spells**.

Relative errors in motif intensities and frequencies between reanalyses and models for **Heatwaves** occurring in France are shown in figure 6.13. Similarly to the extreme cold case, models make different errors for intensities and frequencies. Models tend to overestimate the intensities of the Greenland high, the Cyprus low, the Mediterranean anticyclone and the Nova Scotia low. In



Motif	Motif
1 : Labrador H	15 : Icelandic L
2 : Scandinav. L	16 : mid Atlantic L
3 : Greenland H	17 : Labrador L
4 : Scandinav. H	18 : Mediterr. H
5 : Quebec H	19 : Nova Scotia H
6 : north Russian H	20 : C. European H
7 : UK L	21 : Baffin bay L
8 : UK H	22 : Nova Scotia L
9 : mid Atlantic H	23 : E. Atlantic ridge
10 : Cyprus L	24 : Quebec L
11 : N. Atlantic H	25 : Genoa L
12 : Siberian L	26 : Siberian H
13 : N. Atlantic L	27 : east coast L
14 : Greenland L	28 : Azores L

Figure 6.10 – Motif frequencies versus motif intensities, averaged over all model runs, for model **Heatwaves**. Top-left: **IPSL-CM6A-LR**. Top-right: **MIROC6**. Bottom-left: **ACCESS-ESM1.5**. Bottom-right: **CanESM5**. Each motif is annotated with its index. Indices and names of the 28 LDA motifs are provided in the table. They are sorted in decreasing order of average weight in ERA5. L: low. H: high.

ERA5, the frequency of the UK low motif during **Heatwaves** is among the lowest. This is not reproduced by the models, who systematically overestimate its frequency. The eastern Atlantic ridge also has among the lowest frequencies in ERA5. **ACCESS-ESM1.5** reproduces the ERA5 value, but it is overestimated by **IPSL-CM6A-LR** and **MIROC6**, and underestimated by **CanESM5**. All models also underestimate the frequency associated with the north Russian high, and the UK high. As for the extreme cold case, the intensity/frequency split

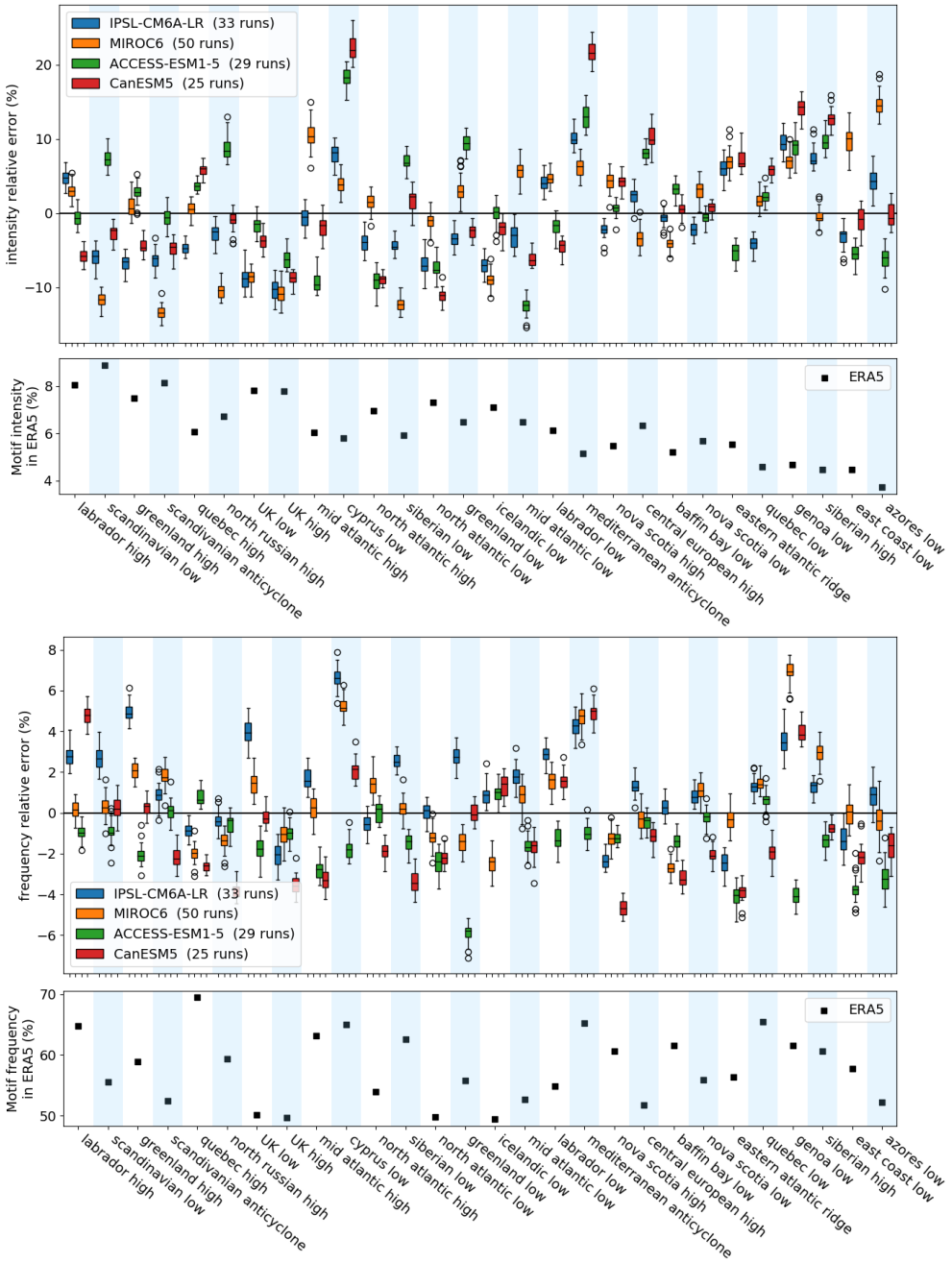


Figure 6.11 – From top to bottom: Relative error on motif intensities between models and ERA5 reanalysis. Motif intensities in ERA5 data. Relative error on motif frequencies between models and ERA5 reanalysis. Motif frequencies in ERA5 data.

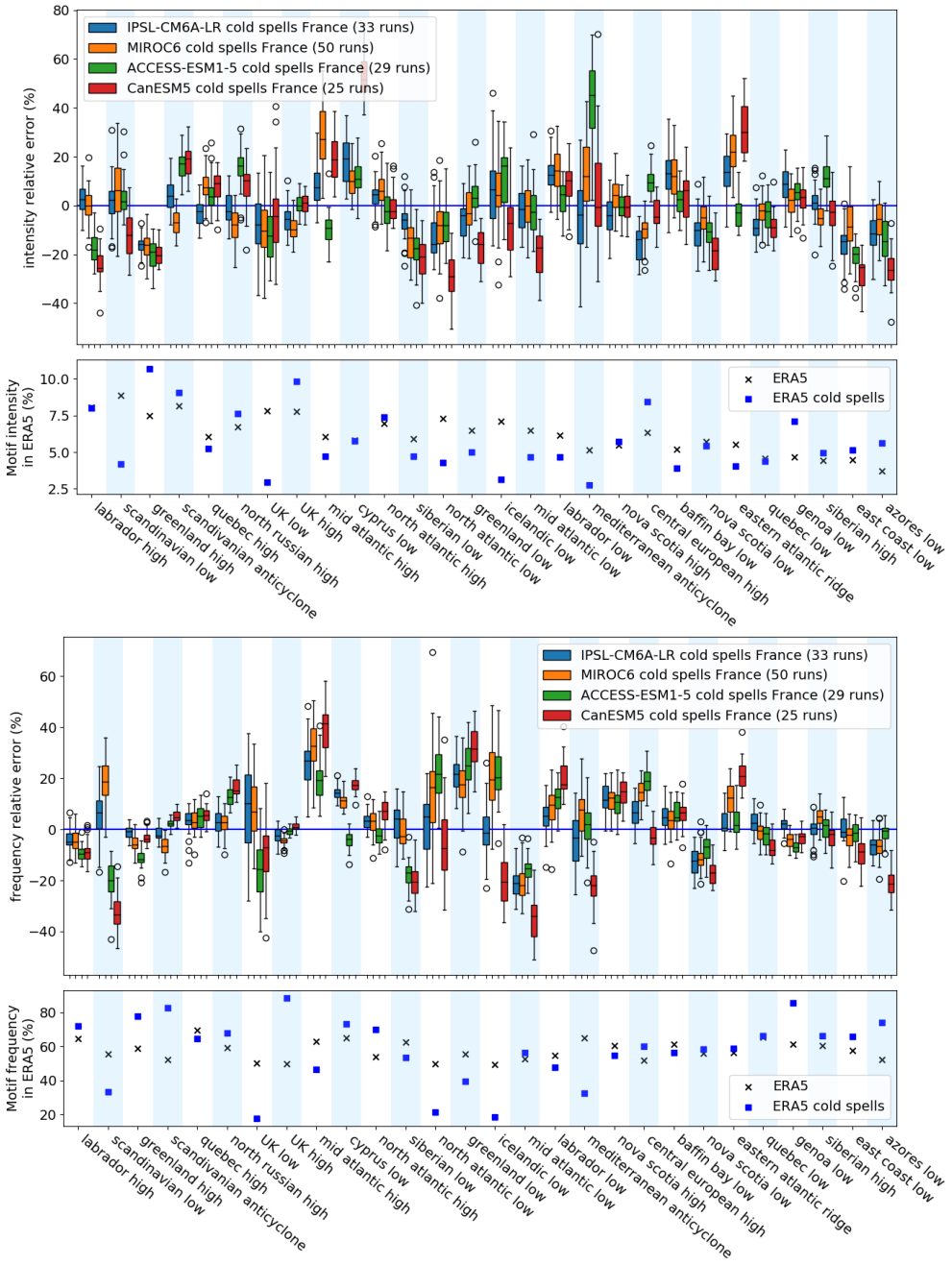


Figure 6.12 – From top to bottom: Relative error on motif intensities between models and ERA5 reanalysis in the case of **Cold Spells** occurring in France. Motif intensities in ERA5 data, for **Cold Spells** and in the general case. Relative error on motif frequencies between models and ERA5 reanalysis in the case of **Cold Spells** occurring in France. Motif frequencies in ERA5 data, for **Cold Spells** and in the general case.

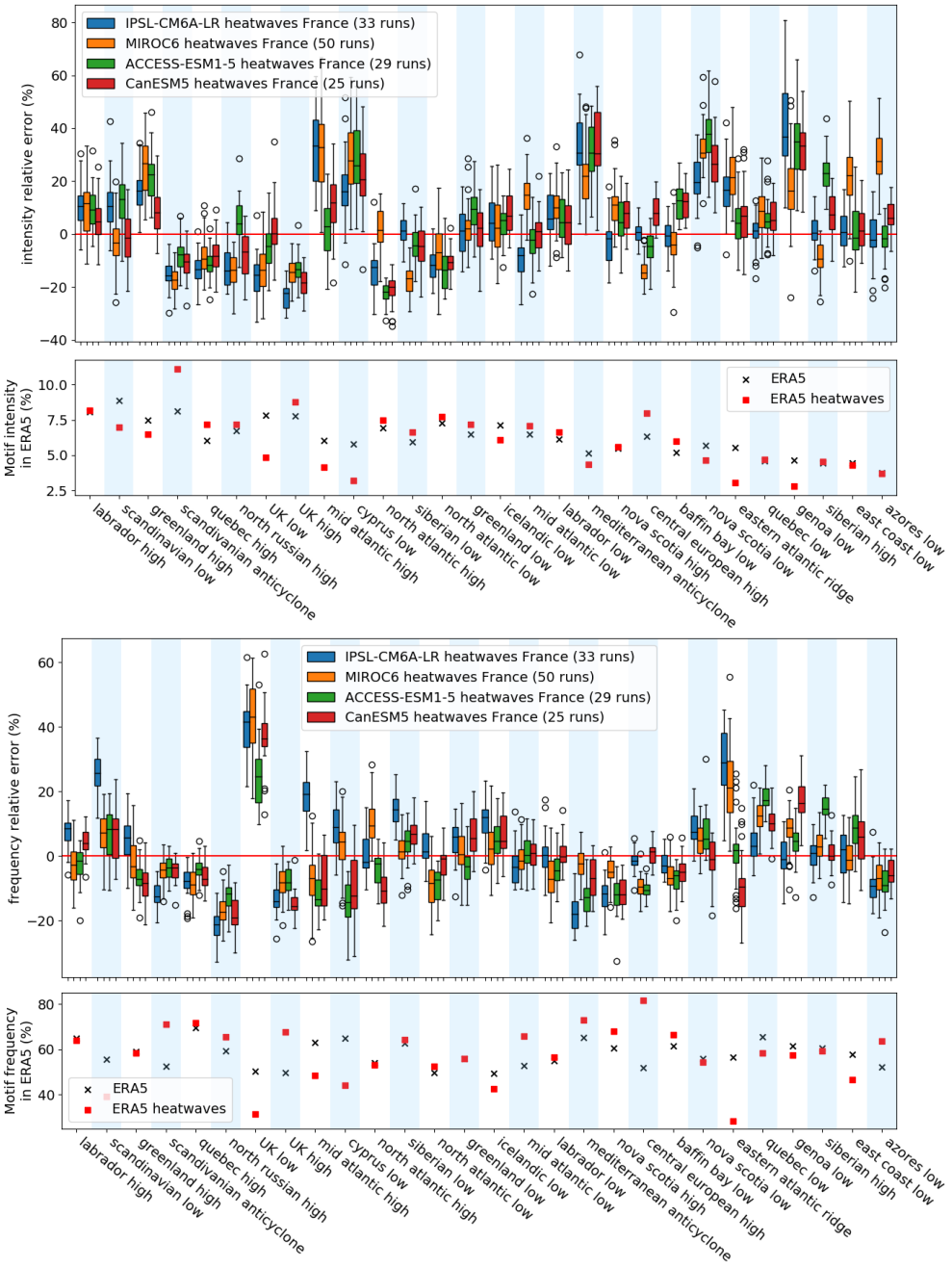


Figure 6.13 – From top to bottom: Relative error on motif intensities between models and ERA5 reanalysis in the case of **Heatwaves** occurring in France. Motif intensities in ERA5 data, for **Heatwaves** and in the general case. Relative error on motif frequencies between models and ERA5 reanalysis in the case of **Heatwaves** occurring in France. Motif frequencies in ERA5 data, for **Heatwaves** and in the general case.

has provided additional characterization of the way models mis-represent atmospheric circulation associated with **Heatwaves**.

Lastly, we wish to quantify whether models errors on intensities and frequencies are good indicators of model errors on motif weights. To that end, we compute the Pearson correlation coefficient (see equation 4.4) between model relative error on motif intensities (respectively: motif frequencies) and model relative error on average motif weights. The results are shown in table 6.1.

Intensities	**General case**	**Cold Spells**	**Heatwaves**
IPSL-CM6A-LR	0.95	0.82	0.82
MIROC6	0.96	0.84	0.85
ACCESS-ESM1.5	0.98	0.83	0.82
CanESM5	0.97	0.81	0.89
Frequencies	**General case**	**Cold Spells**	**Heatwaves**
IPSL-CM6A-LR	0.55	0.72	0.74
MIROC6	0.49	0.61	0.81
ACCESS-ESM1.5	0.23	0.64	0.76
CanESM5	0.56	0.72	0.84

Table 6.1 – Top: Correlations between model relative errors on the representation of motif average weights, and motif intensities. Bottom: Correlations between model relative errors on the representation of motif average weights, and motif frequencies. For each model, the correlation is computed over all motifs and runs.

We find that, in the general case, model relative errors on intensities are a good indicator of model relative errors on average weights, with correlations of at least 0.95, while model relative errors on frequencies are a poor indicator. Intensity errors are always better correlated with average weight errors than frequency errors. However, in the extreme event cases, the correlation between model relative errors on frequencies and average weights are higher, especially for **Heatwaves**, where frequency errors and intensity errors are comparable.

To sum up, we have been able to enrich our characterization of the average atmospheric circulation in ERA5 and models. This was done through a robust measure of motif intermittency, on the basis of which we were able to separate average motif weights into motif frequencies, and motif intensities,

which represent the average weight of a motif when it is nonzero. Model errors are split into a frequency and intensity contribution, which provides additional characterization of the model accuracy. We find that extreme events are characterized by a strong disparity between motifs, both in frequency and in intensity, which is relatively well captured by the models.

6.3 . Dynamic and temperature representation errors

6.3.1 . Definition of the errors

We have shown in chapter 5 that the synoptic configurations obtained by projecting SLP anomalies into the LDA basis provides a characterization of atmospheric circulation. Differences in synoptic configurations between models and reanalysis can be used as a quantitative measurement of model representation error. This error specifically assesses how well atmospheric circulation dynamics are represented, and is therefore referred to as the dynamic error. The dynamic error of run r of model m, $E_{\boldsymbol{X}}^{m,r}$, is defined as the sum of individual motif errors, where the error associated with each motif is defined as the absolute difference between the average weights in the model and in reanalysis data. It is computed according to equation 6.5.

$$E_{\boldsymbol{X}}^{m,r} = \sum_{k=1}^{K} |\langle \boldsymbol{c}_k(\boldsymbol{X}^{m,r}) \rangle - \langle \boldsymbol{c}_k(\boldsymbol{X}) \rangle| \tag{6.5}$$

This dynamic error can be used to evaluate models comparatively, and produce rankings.

Alternatively, models can be comparatively evaluated based on their representation of average temperature, i.e. the average temperature difference between models and reanalysis data. For run r of model m, the temperature error is computed as shown in equation 6.6.

$$E_{\boldsymbol{T}}^{m,r} = \left\langle \overline{\boldsymbol{T}^{m,r}} \right\rangle - \left\langle \overline{\boldsymbol{T}} \right\rangle , \tag{6.6}$$

where $\boldsymbol{T}$ refers to reanalysis temperature maps, and $\boldsymbol{T}^{m,r}$ to temperature maps from run r of model m. Two sources can contribute to the temperature representation error: errors in the representation of thermodynamic processes (regulating the exchanges of heat), and errors in the representation of circulation dynamics (responsible for the advection of warm or cold air from different longitudes/latitudes) (Wehrli et al., 2018). The dynamic error, as previously defined, may propagate to the temperature error, but it is a priori unclear to what extent the latter is determined by the former.

This section will attempt to answer two questions. Both questions relate to whether dynamic and temperature errors contain distinct or redundant information. First, we will determine whether the two measures of error can be combined to provide a comparative evaluation of model performance. We will explore how different the rankings resulting from each error, and whether they suffice to discriminate between models. Then, we will explore whether the two errors are linked. Knowing that temperature error can be caused either by misrepresentation of dynamics or thermodynamics, the correlation between the two errors will inform us on the cause of temperature errors in the models.

6.3.2 . Model comparative evaluation

We represent each model run as a point in the error plane $(E_{\boldsymbol{X}}^{m,r}, E_{\boldsymbol{T}}^{m,r})$, shown in figure 6.14. In addition, we annotate for each run the index of the motif which contributes the most to the dynamic error: $min_k \boldsymbol{c}_k(\boldsymbol{X}^{m,r}) - \boldsymbol{c}_k(\boldsymbol{X})$. For each model, we show on the right of the figure the two motifs that appear most frequently as the largest contributors to the error of a run (the proportion of runs each motif corresponds to is indicated between parentheses) - except for **ACCESS-ESM1.5**, where the largest contributor is always Cyprus low.

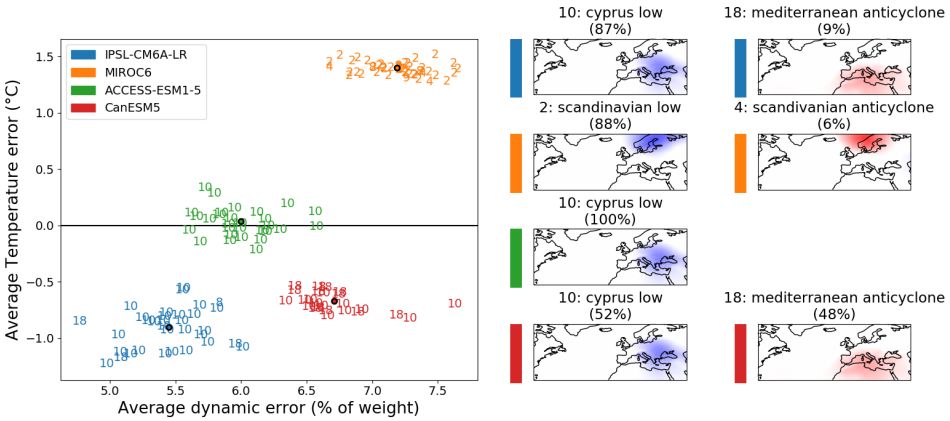


Figure 6.14 – Run-average temperature model error (average temperature difference with reanalysis), versus run-average dynamic model error (average motif weights difference with reanalysis). The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

Each model is associated with a well-identified cluster in the 2-dimensional error plane. If only one of the error was considered, there would be some overlap between the model clusters. **MIROC6** is the model with the highest temperature error, but with the lowest temperature variability. Unlike other models, it over-predicts the temperature by almost 1.5°C. In contrast, the **IPSL-CM6A-LR** model has the highest temperature variability, and under-predicts the temperature by around 1°C. **IPSL-CM6A-LR** also corresponds to the lowest dynamic error. **CanESM5** has a slightly lower temperature error that **IPSL-CM6A-LR**, and has the second highest average dynamic error, behind **MIROC6**. **ACCESS-ESM1.5** has the lowest temperature error, and a relatively low dynamic error.

As mentioned earlier, each run is annotated with the index of the motif with the highest contribution to dynamic error. This makes it possible to attribute errors to specific motifs and regions in space. The Cyprus low (motif 10) is the least well represented motif for all or almost all runs of **ACCESS-ESM1.5** and **IPSL-CM6A-LR**, as well as most runs of **CanESM5**. Mediterranean anticyclone (motif 18), the opposite of Cyprus low, is occasionally the least well represented in runs of **CanESM5** and **IPSL-CM6A-LR**. Both are eastern Mediterranean motifs.

We note that these motifs, which contribute the most to the error in many model runs, are not the most prevalent motifs. If they were, as we are using absolute errors, the large model error on their average weight could be explained by small relative fluctuations. The relative error associated to these Mediterranean motifs is therefore necessarily large. This confirms that the representation of atmospheric circulation over the eastern Mediterranean is a significant issue for all models, particularly for **IPSL-CM6A-LR**, **ACCESS-ESM1.5**, and **CanESM5**. **MIROC6** differs from the other models. Its error on the mean temperature is larger, it is the only model that overestimates average temperature, and its dynamic error is attributed to different motifs than other models, namely the Scandinavian low and Scandinavian anticyclone (motifs 2 and 4). This points to there being different sources of error between **MIROC6** and the other models.

In the end, comparatively evaluating models based on both dynamic and temperature errors provides more detailed characterizations of model misrepresentations than any error on its own. This can be used to select the model best suited to a specific representation task.

6.3.3 . Case of extreme temperature events in France

We now consider extreme temperature events and compute the dynamic and temperature errors associated with **Heatwaves** and **Cold Spells**. When considering extreme events, we are only interested in the components of error specific to the extreme events. Indeed, the biases of the general case were already examined previously. We want to examine the biases that are specific to extremes. Therefore, we eliminate the average bias, as shown in figure 6.14. This error specific to extreme temperature events is called the anomalous error. We define the anomalous dynamic error $E^{m,r}_{\boldsymbol{X},ex}$ similarly for **Heatwaves** and **Cold Spells** following equation 6.7.

$$E^{m,r}_{\boldsymbol{X},ex} = \sum_{k=1}^{K} |\langle\langle \boldsymbol{c}_k(P^{m,r})\rangle\rangle - \langle\langle \boldsymbol{c}_k(P)\rangle\rangle| - E^{m,r}_{\boldsymbol{X}} \tag{6.7}$$

The anomalous temperature error $E^{m,r}_{\boldsymbol{T},ex}$ is defined for **Heatwaves** and **Cold Spells**, for run r of model m following equation 6.8.

$$E^{m,r}_{\boldsymbol{T},ex} = \left\langle\left\langle \overline{\boldsymbol{T}^{m,r}}\right\rangle\right\rangle - \left\langle\left\langle \overline{\boldsymbol{T}}\right\rangle\right\rangle - E^{m,r}_{\boldsymbol{T}} \tag{6.8}$$

In each remaining figure in this chapter, the dynamic and temperature errors represented will be the anomalous errors defined above. The average errors studied in figure 6.14 are eliminated. However, we note that the general conclusions reported below did not change when these errors were taken into account.

Figure 6.15 shows model anomalous temperature error against model anomalous dynamic error in the case of **Cold Spells** occurring in France.

Model internal variability is higher in the cold extreme case than in the general case, both for dynamic and temperature error. There are three distinct clusters in the error plane, separated by gaps larger than internal model variabilities. One cluster corresponds to **CanESM5**. It is the model with the highest dynamic error, and has a high temperature error. A second cluster corresponds to **ACCESS-ESM1.5**. **ACCESS-ESM1.5** has the highest temperature error. Its overestimates **Cold Spell** temperatures by more than 1.5°C on average. **ACCESS-ESM1.5** dynamic error is comparable to that of **MIROC6**, and both are made in majority on the same motif (Greenland high). The last cluster consists of two models, **IPSL-CM6A-LR** and **MIROC6**. With the general bias removed, the temperature value from reanalysis is within the internal variability of both these models. They are also associated with the lowest dynamic error. This cluster appears to be closest to reanalysis. On average, **IPSL-CM6A-LR**

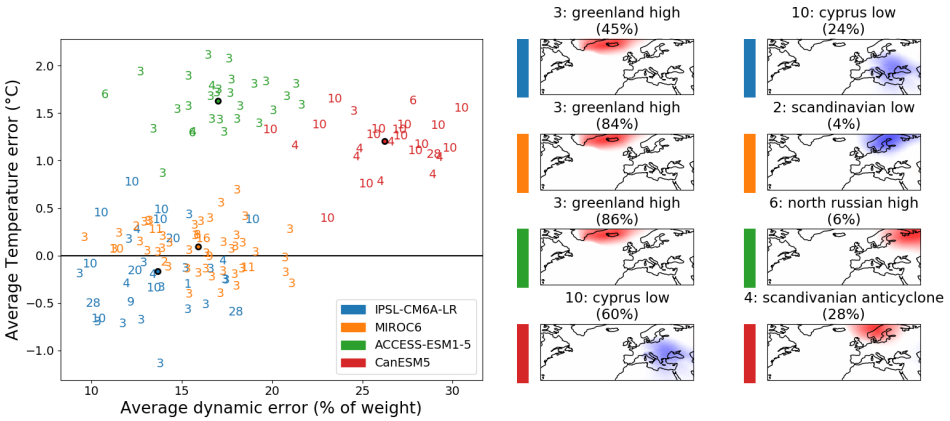


Figure 6.15 – Run-average temperature model error (average temperature difference with reanalysis) on **Cold Spells** in France, versus run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

has a slightly lower dynamic error than **MIROC6**, but the difference is lower than internal variability.

The Greenland high (motif 3) is the least well represented motif on more than 80% of **MIROC6** and **ACCESS-ESM1.5** runs, as well as 45% of **IPSL-CM6A-LR** runs. However this does not necessarily indicate a major model error in the local atmospheric circulation during **Cold Spells**. The relative error on Greenland high weight is small (see figure 5.11). The significant contribution may simply reflect the predominance of the motif in the composition of **Cold Spells**. In contrast, for a majority of **CanESM5** runs, as well as 24% of **IPSL-CM6A-LR** runs, the largest contribution to dynamic error is due to the Cyprus low (motif 10). The Cyprus low is not a particularly dominant motif, but one on which models make a significant relative error (75% on median, see figure 5.11). This suggests a major flaw in the model representation of local circulation over the Mediterranean, specific to **Cold Spells**.

In figure 6.16, we plot model anomalous temperature error against model anomalous dynamic error in the case of **Heatwaves** occurring in France.

Internal model variability for **Heatwaves** is similar to the **Cold Spell** case. However, both temperature and dynamic errors associated with the models

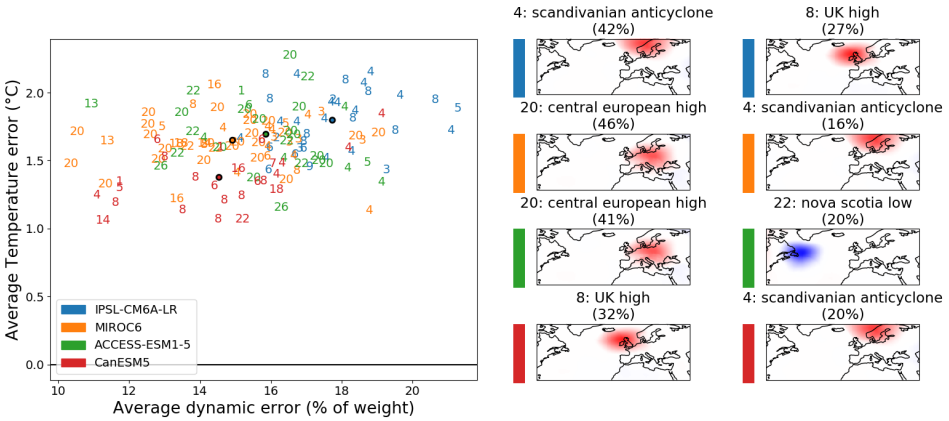


Figure 6.16 – Run-average temperature model error (average temperature difference with reanalysis) on **Heatwaves** in France, versus run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

are closer, so that in error plane, the regions occupied by each model are overlapping. The models are not separated into distinct clusters in this case. All four models are associated with temperature errors between $+1.0$ and $+2.5$°C - as these biases are all positive, they cannot be removed by the use of ensemble methods. Still, some differences can be made between the models. **CanESM5** has the lowest of both types of error on average, and **IPSL-CM6A-LR** the highest, but the differences are lower than model internal variabilities.

In addition, motifs that contribute the most to dynamic error vary significantly more from run to run than for both the general and **Cold Spells** cases. In particular, no motif is the highest contributor to the error in more than 50% of runs of any model. The Central European high (motif 20) appears frequently as the highest contributor to the error in runs of both **MIROC6** and **ACCESS-ESM1.5**. The Scandinavian anticyclone (motif 4) is the largest error contributions in multiple runs of **IPSL-CM6A-LR**, **MIROC6** and **CanESM5**. However, both the Central European high and the Scandinavian anticyclone are dominant motifs in **Heatwaves**, so their presence does not necessarily reflect a significant bias in the models. To sum up, all models appear to perform comparably for the representation of heat waves.

6.3.4 . Case of extreme temperature events in Italy

We now look at model errors on extreme events occurring in a different country. Italy was chosen as its extreme temperature events are known to behave differently than that of France (see chapter 5). Figure 6.17 shows model anomalous temperature error against model anomalous dynamic error in the case of **Cold Spells** occurring in Italy.

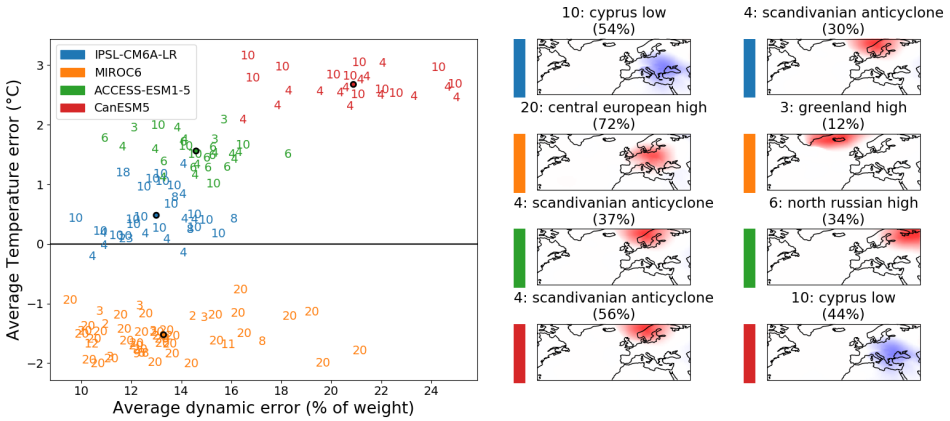


Figure 6.17 – Run-average temperature model error (average temperature difference with reanalysis) on **Cold Spells** in Italy, versus run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

Temperature error reaches higher values than in the case of France, up to $+3$°C for **CanESM5**, and -2°C for **MIROC6**. Each model is associated with its own cluster in the error plane in the case of Italian **Cold Spells**. **IPSL-CM6A-LR** is the cluster associated with lowest average temperature error. It is tied for lowest dynamic error with **MIROC6**, although the internal variability of this second one reaches higher. **MIROC6** is the only model to underestimate the **Cold Spell** temperature, with a temperature error between -1°C and -2°C. **ACCESS-ESM1.5** has higher average dynamic error than **IPSL-CM6A-LR**, however, their internal variabilities are overlapping for this error. **ACCESS-ESM1.5** is associated with a temperature error 1° higher than **IPSL-CM6A-LR**. **CanESM5** is associated with both highest temperature error, between $+2$°C and $+3$°C, and highest dynamic error.

The Scandinavian anticyclone (motif 4) concentrates the highest dynamic error on 30% of **IPSL-CM6A-LR** runs, 37% of **ACCESS-ESM1.5** runs, and 56% of **CanESM5** runs. The Cyprus low (motif 10) also plays an important role, being the worst offender in 54% of **IPSL-CM6A-LR** runs and 44% of **CanESM5** runs, despite its major influence on error in the general case being ignored. Once again, the motifs contributing the most to **MIROC6** dynamic errors are different motifs than the other models. In 72% of runs, it is the central European high. The Scandinavian anticyclone and central European high are dominant motifs in **Cold Spells** occurring in Italy. These error do not demonstrate a bias in the model. The Cyprus low, however, has lower average weight. Therefore, the fact that it is the largest contributor to dynamic error in some runs demonstrates a bias in the representation of the atmospheric circulation specific to extreme cold in Italy. With the exception of **MIROC6**, temperature and dynamic error are in general agreement regarding the rankings of model performance.

Figure6.18 shows model anomalous temperature error against model anomalous dynamic error in the case of **Heatwaves** occurring in Italy.

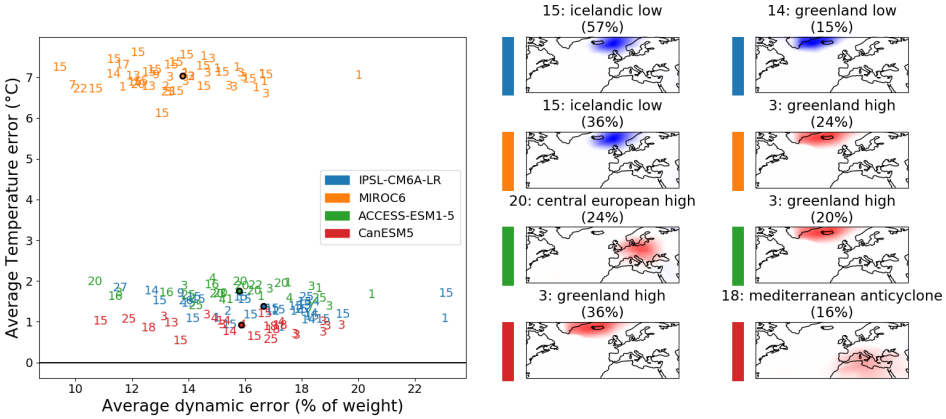


Figure 6.18 – Run-average temperature model error (average temperature difference with reanalysis) on **Heatwaves** in Italy, versus run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

In the case of **Heatwaves** occurring in Italy, we observe two well-separated cluster: one containing only **MIROC6**, and another cluster regrouping the

three other models. Internal model variability on dynamical error is higher in this case than in all the other extreme events considered thus far. The **MIROC6** cluster is characterized by high temperature error, around $+7$°C on average. It is the model with the lowest average dynamic error. **IPSL-CM6A-LR**, **ACCESS-ESM1.5** and **CanESM5** are characterized by similar, overlapping ranges of dynamic error, with **IPSL-CM6A-LR** being slightly higher on average. While there is also overlap in the ranges of temperature errors, the ranking can unequivocally be established. In increasing temperature error order, it is **CanESM5**, **IPSL-CM6A-LR**, and **ACCESS-ESM1.5**.

The Icelandic low (motif 15) contributes the most to dynamic error in most of **IPSL-CM6A-LR** runs and 36% of **MIROC6** runs. The Greenland high (motif 3) concentrates the most error in a fourth of **MIROC6** runs, a fifth of **ACCESS-ESM1.5** runs, and a third of **CanESM5** runs. High representation errors on the Greenland high do not necessarily signify a major bias in the models, as the error could simply be due to the high values of average weight. However, the representation error on the Icelandic low indicates a model bias on the representation of circulation at the north of Europe, specifically during **Heatwaves** occurring in Italy. Preselection of the best suited model in this case is not obvious. It can depend on whether accurate event intensity, or accurate circulation patterns are more valued. However, dynamic representation plays a much weaker role than temperature representation in the discrimination of models for this type and location of extremes.

6.3.5 . Link between the two errors

As was established in the previous subsections, there are cases where dynamic and temperature errors are in agreement regarding model representation performance rankings, and cases where they are not. We will now explore the correlations between the two types of model errors to determine their relationship in different extreme events. Since temperature error arises from two sources — dynamic and temperature misrepresentation — a correlation between these errors could suggest a causal link from dynamic misrepresentation to temperature misrepresentation.

We compute the Pearson correlation coefficient (see equation 4.4) between the absolute value of temperature error and dynamic error, using each run of each model as a realization of the two variables. This is done in the general case. For extreme cold and heat event in each of the 6 considered countries, we compute the correlation between the absolute value of the anomalous temperature error, and the anomalous dynamic error. The results are shown in table 6.2

General case	0.52

	Cold Spells	Heatwaves
France	0.41	0.29
Germany	0.48	0.30
Poland	0.29	0.03
Italy	0.61	-0.45
Spain	0.60	0.22
UK	0.50	0.05

Table 6.2 – Correlations between temperature error and dynamic error, in absolute value, in the general case and in extreme temperature events in six European countries. When considering extreme events, only the anomalous error, excluding the average error observed in the general case, is considered.

In most cases, we observe a positive, but partial correlation between the two types of error. Here, we are particularly interested in the correlations between the misrepresentations of the differential of temperature and circulation specific to extreme, as measured by the anomalous error metrics in the extreme event cases. To begin with, we observe that correlation is systematically lower for **Heatwaves** than for **Cold Spells**. This means that dynamic aspects are less relevant to the error models make on extreme temperature values, or at least the *surface level* circulation dynamics described by SLP anomalies. This means that surface-level circulation plays less of a role in extreme heat temperature events that in extreme cold temperatures.

We find a correlation close to zero for **Heatwaves** in Poland and the UK. This indicates that errors in SLP anomaly have little to no bearing on the misrepresentation of extreme heat temperature in those countries. This can indicate that errors in temperature are mostly due to thermodynamic misrepresentation, or at the very least that atmospheric circulation at surface level is not relevant to these kinds of extremes. **Cold Spells** in Italy and Spain have the highest level of correlation between anomalous temperature error and anomalous dynamic error. This implies that misrepresentation of atmospheric circulation at the surface level is an important reason why models make errors on the temperature differences with the general case. The case of **Heatwaves** in Italy is the only one where the correlation coefficient is negative. The position of **MIROC6**, with very large temperature error and lower dynamic error that other models, increases this effect - whithout **MIROC6**, which is an outlier in this case, the correlation would be of -0.15. Nevertheless, this too would indicate that misrepresentation of surface level circulation patterns is not the cause of the temperature errors.

6.4 . Summary

In this chapter, we define additional criteria to represent model errors in a higher-dimensional space. This is used to extract discriminating features from the data, with the goal of improving the evaluation of the models, in particular with respect to the representation of extreme events. Two analysis protocols are explored. One corresponds to a more detailed analysis of the average motif weight, which is factorized into a frequency and an intensity contribution using a robust measure of motif intermittency. The large spread in motif frequencies makes it possible to identify specific motifs contributing to the extreme events, such as the UK high and the Scandinavian anticyclones for **Cold Spells**, or the central European high for **Heatwaves** in the case of France. Model errors can be evaluated separately on intensities and frequencies, providing additional insight on model mis-representation of circulation. In the general data case, a clear correlation is observed between the average weight of the motif and its intensity, but this is not the case for the frequency. However, for extreme events, the intensities and frequencies of events are both well correlated with the average weight, particularly for cold spells.

The second protocol consists in taking into account both dynamic and temperature errors. Differences between model synoptic configuration in reanalyses and models can be used to define a global dynamic error. But models can also be evaluated based on the temperature error, which corresponds to the difference between the average temperature observed in reanalysis and that predicted by the models. By extracting which motif contributes the most to the total dynamic error, and by comparing the dynamic error with the temperature error, it is possible to characterize model errors. We show that the conjunction of these two model evaluation metrics can discriminate between the performances of the models, especially in the general and **Cold Spell** cases. This can be used as a mean to pre-select models that are best suited to faithfully represent specific types of extreme events.

Moreover, model temperature error can have two possible sources, misrepresentation of dynamic processes, or misrepresentation of thermodynamic processes. By studying the correlation between model temperature error and model dynamic error, we can estimate in each case how much of the temperature error is due to the dynamic error as we measure it. In particular, we find that the two error metrics are correlated in the case of **Cold Spells** occurring in Italy, which would point toward a causality link. On the other hand, there is little correlation observed in the case of **Heatwaves** in Italy, indicating that

the model misrepresentation of temperature is not due to the misrepresentation of low-altitude circulation. However, we note that our definition of the dynamic error is based on the surface level atmospheric circulation described by SLP anomalies. Therefore, this assessment is only valid for low-altitude circulation.

Conclusion

Summary and main results

The stated goal of this thesis research project was to devise a protocol for locally evaluating model representation of atmospheric circulation. **In this work, we have developed methodologies to evaluate climate models by comparing learned representations of the model atmospheric circulation with those of reanalysis data.** We have chosen to work with Sea Level Pressure (SLP), which specifically represents atmospheric circulation close to the surface. Latent Dirichlet Allocation (LDA), a machine learning tool from natural language processing, was used to extract local latent information from climate data related to atmospheric circulation. We explored various ways in which this latent information could be used to characterize atmospheric circulation within a dataset and develop methods for model evaluation. These evaluation methods were applied to four General Circulation Models (GCM) from the CMIP6 project - **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**- to evaluate their representation of atmospheric circulation, both generally and specifically for extreme temperature events (**Heatwaves** and **Cold Spells**) in several European countries. We were thus able to identify biases in the models, in both the general case, as well as for extreme events. Some of the biases were shared by all models, while others were model-specific. Below, we summarize the main findings of our work.

The LDA technique could be adapted to identify a common basis of localized synoptic objects to represent SLP maps from any dataset. The data representation provided by LDA is low-dimensional, local and physically interpretable. LDA is an unsupervised learning method developed for topic modelling in natural language processing. In its original version, it takes as input a corpus of text documents, all defined only by their word counts (and ignoring their order), and extracts from them a set of latent variables called topics, which are distributions over the set of possible words. Each document is then associated with a topic composition, thus providing soft clustering. The method was adapted for SLP maps, which represents atmospheric circulation close to the surface. We focused on the European and North-Atlantic region and have considered daily maps from the period of $\sim$ 1950-2010. When applied to SLP from any dataset, LDA yields a basis of motifs which are localized objects at synoptic scale. These motifs are physically interpretable, as they can be recognized as synoptic objects such as cyclones and anticyclones. Latent variables, or weights, describing the motif composition can be obtained from any SLP anomaly map. Furthermore, the motif compositions of maps

are sparse, so that each map is characterized by a few motifs. Therefore, in addition to being a dimensionality reduction method, LDA also provides an accessible physical interpretation to the motif weights.

Extreme events can be associated with characteristic synoptic configurations. LDA motif weights can be used to characterize the atmospheric circulation in a dataset. In particular, we use the average weight of each motif in a dataset as a characterization of its synoptic configuration. This allows us to measure differences between the atmospheric circulation patterns associated with extreme temperature events, and the general average circulation. As an example, we have found that cold extremes in a country are generally associated with a wide anticyclonic structure at the north of the country, and a cyclonic structure at the south. The synoptic patterns vary per country, as shown in the case of Italy, which show little to no southern anticyclonic structure, and the UK, which shows an intense cyclonic structure. These observations can be directly obtained from the motif weights and given a physical interpretation.

Models can be evaluated and compared with each other, based on synoptic configurations. By comparing synoptic configurations in a model with reanalysis data, we can evaluate the model representation of atmospheric circulation. This evaluation method can be applied to general data over the whole domain, but also to the subset of maps associated with a **Cold Spell** or **Heatwave** occurring in a specific region. The motif representation provides a local assessment of the error, and makes it possible to identify common or model-specific sources of errors in the model. Among the results, we note that:

- **All models misrepresent circulation around the Mediterranean region.** Using the previously defined methodology, we were able to identify biases in the four models we have studied: **IPSL-CM6A-LR**, **MIROC6**, **ACCESS-ESM1.5** and **CanESM5**. Some of these biases were model-specific, and therefore can be eliminated using multimodel ensemble methods. Other biases, however, were shared by all four models. In particular, we have found that every model systematically overestimates the presence of cyclones and anticyclones located over the Mediterranean.

- **Models use different recipes than reanalyses for extreme temperature events.** Model representation errors are several times higher when studying specifically extreme temperature events. While that varies on a model-to-model basis, synoptic configurations associated with extreme temperature events are substantially different from results in

ERA5. This implies that models cause extreme temperatures using different patterns of circulation, on average.

Further characterization can be obtained by higher-dimensional representations, at both a local and global level.

- **Locally, each average weight can be decomposed as the product of a characteristic frequency and an intensity.** Each motif is then represented by an intensity and a frequency, each of which can be compared independently. Representation errors can then be locally measured for each model, specifically in intensities or frequencies, the latter being physically interpretable. We find that errors in average weights are mostly due to intensities in the general case. But during extreme temperature events, distinct relevant errors are observed on both indicators, allowing for the characterisation of model error in higher detail. The inter-model variability of those errors is generally higher than the inner model variability.

- **The combination of global dynamic error and global temperature error in each model run can discriminate model performance in the general case, and for some extreme events.** The global dynamic error in each model run can be compared with the global temperature error, and associated with the dominant motif contributing the most to the error. Studied in conjunction, the two model errors allow for discriminating the performances of the models in the general and **Cold Spell** cases. Depending on the country of interest, model generally perform similarly on **Heatwaves**.

- **We can estimate the influence of model misrepresentation of circulation on model temperature errors.** Comparing the dynamic and temperature errors makes it possible to estimate the influence of the model misrepresentation of circulation on the model temperature error. Temperature errors in models arise from inaccuracies in either the dynamic or thermodynamic components. By examining the correlation between dynamic errors and temperature errors, we can estimate how much of the temperature error is attributable to dynamic factors. This can be applied to either type of extreme temperature events in any region. In particular, we have found a strong link between the two in **Cold Spells** in Italy, while there is close to no link in the **Heatwaves** of this country.

These methodologies can be used for model pre-selection. Each method of model evaluation we have presented can be used to compare them with

each other. By producing rankings based on either global or local criteria, we can determine which model best represents the specific aspects of atmospheric circulation that each method examines. Therefore, we can determine which model is best suited to realistically reproduce atmospheric circulation in a given region, in general or for a specific type of extreme event. This can be helpful for future studies requiring large amounts of data, as they will be able to select the models best suited to generate it.

Perspectives

This thesis work concerns the application of a relatively new methodology to atmospheric sciences. As such, it is still largely exploratory and merely scratches the surface of its potential uses. Below, we propose several directions for future research to build upon this work. We divide these potential developments into two categories: extracting more information from the motif weights, or exploring application of LDA to different datasets.

Extracting more information from motif weights

Considering higher moments of the weight distributions. When using LDA motif weights to characterize atmospheric circulation in a dataset, we have so far only looked at averages. More information could be obtained by looking at higher orders of the weight distributions, such as variances. This would provide a more detailed characterization, and may allow the detection of subtler biases in model representation of circulation, in general, or during extreme events.

Studying motif weights as time series. The motifs weights of a dataset of daily maps form a set of time series. Exploring the properties of these times series may result in new insight regarding atmospheric circulation in a dataset. For instance, one could explore whether the motif weights show any trends, and if so, whether they can be attributed to climate change.

Exploring dynamics of the motif weights. Because of the accessible physical interpretation, the dynamics of SLP in the reduced space of motif compositions could provide interesting insight. In particular, the motif to motif transfer rates could be explored to determine which motifs tend to follow after which. With a choice of time step, for example, the 5 days of synoptic correlation time, a matrix of motif-to-motif transfer rates can be obtained under a stationarity assumption. This would provide synoptic-level information on meteorology and would be applicable to short-term qualitative predictions by humans, thanks to the sparsity of LDA decomposition. For a less rudimen-

tary model, the transfer rates could be made to vary by season, or with day of the year.

Extending the field of application of LDA

Producing rankings of every CMIP6 model to identify internal sources of error. Without changes to the method, data from additional GCM could be included in a further study. This would provide a more complete picture of the performances of CMIP6 models, and therefore give further insight regarding which biases are shared and which are model-specific. Furthermore, exploring in more detail the links between model biases and their resolutions, parametrization schemes, or tuning strategies would help locating the sources of these biases and finding solutions. Considering additional models would also provide more complete rankings of their performance, and thus facilitate the task of model preselection.

Considering other climate variables. This method could also be applied to variables other than SLP, and thus evaluate model performances on other aspects of atmospheric circulation, or climate in general, with little to no changes to the methodology. As an example, 500 hPa geopotential height (z500) also contains relevant information on atmospheric circulation, at higher altitude. Alternatively, LDA could also be applied to temperature data, to extract information about the thermodynamics aspects of climate.

Building an LDA motif basis on multiple variables, possibly including temporal aspects. In our methodology of application for LDA, an input document was one daily map of SLP anomalies. But it is possible to give LDA daily maps of several variables simultaneously, as a single document. This would result in multi-variable motifs. Similarly, by considering several consecutive daily maps as a single document, LDA could learn motifs that include a short temporal evolution. However, this would require learning a higher amount of motifs. A limitation of this approach is that it would require a high amount data to converge. 60 years of daily maps may therefore not be sufficient, and combining several datasets may be mandatory.

Exploring other spatial domains, and other definitions of extreme events. We have only applied our model evaluation methods to the north-Atlantic region, looked only at extremes defined in western European countries, and considered only temperature-related extremes. Without changes to the method, more could be learned by applying our methodology to other regions, and exploring any definitions of extreme events that can be automatically detected in the data.

Comparing motif bases from different periods. Finally, the time evolution of the LDA motif basis can be explored. By dividing the input data into a set of overlapping sub-periods, and computing the LDA motif basis separately on each one, a time dependency can be introduced into the motifs themselves. Corresponding motifs would be recognizable from successive similar bases, and the progressive evolution - or discrete transformations - of the motifs that best describe the data could be measured. Then, the question of whether such changes are attributable to climate change could be explored.

References

Abramowitz, G., N. Herger, E. Gutmann, D. Hammerling, R. Knutti, et al. (2019). « ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing ». English. In: *Earth System Dynamics* 10.1, pp. 91–105. doi: 10.5194/esd-10-91-2019 (see page 34).

Aguado, E. and J. Burt (2014). *Understanding Weather and Climate*. English. 7th edition. Boston: Pearson (see page 22).

Ahmad, N. (2008). « Characterization of convective rain in Klang Valley, Malaysia ». In.

Ahmed, M., R. Seraj, and S. M. S. Islam (2020). « The k-means Algorithm: A Comprehensive Survey and Performance Evaluation ». en. In: *Electronics* 9.8, p. 1295. doi: 10.3390/electronics9081295 (see page 47).

Alexander, L. V., X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, et al. (2006). « Global observed changes in daily climate extremes of temperature and precipitation ». en. In: *Journal of Geophysical Research: Atmospheres* 111.D5. doi: 10.1029/2005JD006290 (see page 28).

Anagnostopoulou, C., K. Tolika, G. Lazoglou, and P. Maheras (2017). « The Exceptionally Cold January of 2017 over the Balkan Peninsula: A Climatological and Synoptic Analysis ». en. In: *Atmosphere* 8.12, p. 252. doi: 10.3390/atmos8120252 (see page 26).

Añel, J. A., M. Fernández-González, X. Labandeira, X. López-Otero, and L. De la Torre (2017). « Impact of Cold Waves and Heat Waves on the Energy Production Sector ». en. In: *Atmosphere* 8.11, p. 209. doi: 10.3390/atmos8110209 (see page 26).

Asay-Davis, X. S., N. C. Jourdain, and Y. Nakayama (2017). « Developments in Simulating and Parameterizing Interactions Between the Southern Ocean and the Antarctic Ice Sheet ». en. In: *Current Climate Change Reports* 3.4, pp. 316–329. doi: 10.1007/s40641-017-0071-0 (see page 35).

Attias, H. (1999). « A Variational Baysian Framework for Graphical Models ». In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press (see page 79).

Bardossy, A., L. Duckstein, and I. Bogardi (1995). « Fuzzy rule-based classification of atmospheric circulation patterns ». en. In: *International Journal of Climatology* 15.10, pp. 1087–1097. doi: 10.1002/joc.3370151003 (see pages 56, 57).

Bárdossy, A., J. Stehlík, and H.-J. Caspary (2002). « Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules ». en. In: *Cli-*

mate Research 23.1, pp. 11–22. doi: 10.3354/cr023011 (see pages 56, 57).
Bartzokas, A. and D. A. Metaxas (1996). « Northern Hemisphere gross circulation types. Climatic change and temperature distribution ». en. In: *Meteorologische Zeitschrift*, pp. 99–109. doi: 10.1127/metz/5/1996/99 (see page 50).
Baumberger, C., R. Knutti, and G. Hirsch Hadorn (2017). « Building confidence in climate model projections: an analysis of inferences from fit ». en. In: *WIREs Climate Change* 8.3, e454. doi: 10.1002/wcc.454.
Bergeron, T. (1980). « Synoptic meteorology: An historical review ». en. In: *pure and applied geophysics* 119.3, pp. 443–473. doi: 10.1007/BF00878152 (see page 22).
Berner, J., U. Achatz, L. Batté, L. Bengtsson, A. d. l. Cámara, et al. (2017). « Stochastic Parameterization: Toward a New View of Weather and Climate Models ». EN. In: *Bulletin of the American Meteorological Society* 98.3, pp. 565–588. doi: 10.1175/BAMS-D-15-00268.1 (see page 35).
Bezdek, J. C. (2013). *Pattern Recognition with Fuzzy Objective Function Algorithms*. en. Springer Science & Business Media (see page 58).
Blei, D., L. Carin, and D. Dunson (2010). « Probabilistic Topic Models ». In: *IEEE Signal Processing Magazine* 27.6, pp. 55–65. doi: 10.1109/MSP.2010.938079 (see page 67).
Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). « Latent dirichlet allocation ». In: *The Journal of Machine Learning Research* 3.null, pp. 993–1022 (see pages 63, 67, 71, 75).
Bluestein, H. B. (1992). *Synoptic-dynamic Meteorology in Midlatitudes: Observations and theory of weather systems*. en. Taylor & Francis (see page 22).
Bock, L., A. Lauer, M. Schlund, M. Barreiro, N. Bellouin, et al. (2020). « Quantifying Progress Across Different CMIP Phases With the ESMValTool ». en. In: *Journal of Geophysical Research: Atmospheres* 125.21, e2019JD032321. doi: 10.1029/2019JD032321 (see page 35).
Boé, J. (2018). « Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity ». en. In: *Geophysical Research Letters* 45.6, pp. 2771–2779. doi: 10.1002/2017GL076829 (see page 34).
Borodina, A., E. M. Fischer, and R. Knutti (2017). « Potential to Constrain Projections of Hot Temperature Extremes ». EN. In: *Journal of Climate* 30.24, pp. 9949–9964. doi: 10.1175/JCLI-D-16-0848.1 (see page 36).
Boucher, O., J. Servonnat, A. L. Albright, O. Aumont, Y. Balkanski, et al. (2020). « Presentation and Evaluation of the IPSL-CM6A-LR Climate

Model ». en. In: *Journal of Advances in Modeling Earth Systems* 12.7, e2019MS002010. doi: 10.1029/2019MS002010 (see pages 36, 110).
Bourdin, S., S. Fromang, W. Dulac, J. Cattiaux, and F. Chauvin (2022). « Intercomparison of four algorithms for detecting tropical cyclones using ERA5 ». English. In: *Geoscientific Model Development* 15.17, pp. 6759–6786. doi: 10.5194/gmd-15-6759-2022 (see page 27).
Brinkmann, W. A. R. (1999). « Application of Non-hierarchically Clustered Circulation Components to Surface Weather Conditions: Lake Superior Basin Winter Temperatures ». en. In: *Theoretical and Applied Climatology* 63.1, pp. 41–56. doi: 10.1007/s007040050090 (see page 47).
Carslaw, K. S., H. Gordon, D. S. Hamilton, J. S. Johnson, L. A. Regayre, et al. (2017). « Aerosols in the Pre-industrial Atmosphere ». en. In: *Current Climate Change Reports* 3.1, pp. 1–15. doi: 10.1007/s40641-017-0061-2 (see page 24).
Casella, G. and E. I. George (1992). « Explaining the Gibbs Sampler ». In: *The American Statistician* 46.3, pp. 167–174. doi: 10.1080/00031305.1992.10475878 (see page 76).
Cavazos, T. (1999). « Large-Scale Circulation Anomalies Conducive to Extreme Precipitation Events and Derivation of Daily Rainfall in Northeastern Mexico and Southeastern Texas ». EN. In: *Journal of Climate* 12.5, pp. 1506–1523. doi: 10.1175/1520-0442(1999)012<1506:LSCACT>2.0.CO;2 (see page 54).
— (2000). « Using Self-Organizing Maps to Investigate Extreme Climate Events: An Application to Wintertime Precipitation in the Balkans ». EN. In: *Journal of Climate* 13.10, pp. 1718–1732. doi: 10.1175/1520-0442(2000)013<1718:USOMTI>2.0.CO;2 (see page 54).
Chambers, J. M., J. Gordesch, A. Klas, L. Lebart, and P. P. Sint (1985). *Lectures in Computational Statistics: Multidimensional Clustering Algorithms*. English. Springer Verlag (see page 45).
Chan, P. W., J. L. Catto, and M. Collins (2022). « Heatwave–blocking relation change likely dominates over decrease in blocking frequency under global warming ». en. In: *npj Climate and Atmospheric Science* 5.1, pp. 1–8. doi: 10.1038/s41612-022-00290-2 (see page 28).
Cheng, X. and J. M. Wallace (1993). « Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns ». EN. In: *Journal of the Atmospheric Sciences* 50.16, pp. 2674–2696. doi: 10.1175/1520-0469(1993)050<2674:CAOTNH>2.0.CO;2 (see page 46).
Christensen, W. I. and R. A. Bryson (1966). « AN INVESTIGATION OF THE POTENTIAL OF COMPONENT ANALYSIS FOR WEATHER CLASSIFICATION ». EN. In: *Monthly Weather Review* 94.12, pp. 697–709. doi: 10.1175/1520-0493(1966)094<0697:AIOTPO>2.3.CO;2 (see page 43).

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. London: Springer. doi: 10.1007/978-1-4471-3675-0 (see page 29).
Commission Européenne (2005). *Communication from the Commission to the European Parliament and the Council on the appropriateness of establishing rules on a Europe-wide basis for more detailed levels in the NUTS Classification*. en (see page 103).
Compagnucci, R. H. and M. B. Richman (2008). « Can principal component analysis provide atmospheric circulation or teleconnection patterns? » en. In: *International Journal of Climatology* 28.6, pp. 703–726. doi: 10.1002/joc.1574 (see page 50).
Corte-Real, J., B. Quian, and H. Xu (1999). « Circulation patterns, daily precipitation in Portugal and implications for climate change simulated by the second Hadley Centre GCM ». en. In: *Climate Dynamics* 15.12, pp. 921–935. doi: 10.1007/s003820050322 (see page 47).
Corte-Real, J., B. Qian, and H. Xu (1998). « Regional climate change in Portugal: precipitation variability associated with large-scale atmospheric circulation ». en. In: *International Journal of Climatology* 18.6, pp. 619–635. doi: 10.1002/(SICI)1097-0088(199805)18:6<619::AID-JOC271>3.0.CO;2-T (see page 47).
Cowell, R. G., P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. en. Springer Science & Business Media (see page 63).
Cox, D. R. (2006). *Principles of Statistical Inference*. en. Cambridge University Press (see page 59).
Crueger, T., M. A. Giorgetta, R. Brokopf, M. Esch, S. Fiedler, et al. (2018). « ICON-A, The Atmosphere Component of the ICON Earth System Model: II. Model Evaluation ». en. In: *Journal of Advances in Modeling Earth Systems* 10.7, pp. 1638–1662. doi: 10.1029/2017MS001233 (see page 34).
Davini, P. and F. D'Andrea (2020). « From CMIP3 to CMIP6: Northern Hemisphere Atmospheric Blocking Simulation in Present and Future Climate ». EN. In: *Journal of Climate* 33.23, pp. 10021–10038. doi: 10.1175/JCLI-D-19-0862.1 (see pages 34, 35).
Davis, R. E. and L. S. Kalkstein (1990). « Development of an automated spatial synoptic climatological classification ». en. In: *International Journal of Climatology* 10.8, pp. 769–794. doi: 10.1002/joc.3370100802 (see page 41).
Dee, D. P., M. Balmaseda, G. Balsamo, R. Engelen, A. J. Simmons, et al. (2014). « Toward a Consistent Reanalysis of the Climate System ».

EN. In: *Bulletin of the American Meteorological Society* 95.8, pp. 1235–1248. doi: 10.1175/BAMS-D-13-00043.1 (see page 32).
Delforge, D., V. Wathelet, R. Below, C. L. Sofia, M. Tonnelier, et al. (2023). *EM-DAT: the Emergency Events Database*. doi: 10.21203/rs.3.rs-3807553/v1 (see page 27).
Di Luca, A., R. de Elía, M. Bador, and D. Argüeso (2020a). « Contribution of mean climate to hot temperature extremes for present and future climates ». In: *Weather and Climate Extremes* 28, p. 100255. doi: 10.1016/j.wace.2020.100255 (see page 36).
Di Luca, A., A. J. Pitman, and R. de Elía (2020b). « Decomposing Temperature Extremes Errors in CMIP5 and CMIP6 Models ». en. In: *Geophysical Research Letters* 47.14, e2020GL088031. doi: 10.1029/2020GL088031 (see page 36).
Doktycz, C. and M. Abkowitz (2019). « Loss and Damage Estimation for Extreme Weather Events: State of the Practice ». en. In: *Sustainability* 11.15, p. 4243. doi: 10.3390/su11154243 (see page 27).
Donat, M. G., A. J. Pitman, and S. I. Seneviratne (2017). « Regional warming of hot extremes accelerated by surface energy fluxes ». en. In: *Geophysical Research Letters* 44.13, pp. 7011–7019. doi: 10.1002/2017GL073733 (see page 36).
Dunn, J. C. (1973). « A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters ». In: *Journal of Cybernetics* 3.3, pp. 32–57. doi: 10.1080/01969727308546046 (see page 58).
Edwards, P. N. (2011). « History of climate modeling ». en. In: *WIREs Climate Change* 2.1, pp. 128–139. doi: 10.1002/wcc.95 (see page 31).
Esteban, P., P. D. Jones, J. Martín-Vide, and M. Mases (2005a). « Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees ». en. In: *International Journal of Climatology* 25.3, pp. 319–329. doi: 10.1002/joc.1103 (see page 47).
— (2005b). « Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees ». en. In: *International Journal of Climatology* 25.3, pp. 319–329. doi: 10.1002/joc.1103 (see page 40).
Esteban, P., J. Martin-Vide, and M. Mases (2006). « Daily atmospheric circulation catalogue for western Europe using multivariate techniques ». en. In: *International Journal of Climatology* 26.11, pp. 1501–1515. doi: 10.1002/joc.1391 (see page 48).
Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, et al. (2016). « Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization ». English. In: *Geoscientific Model Development* 9.5, pp. 1937–1958. doi: 10.5194/gmd-9-1937-2016 (see page 33).

Faranda, D., G. Masato, N. Moloney, Y. Sato, F. Daviaud, et al. (2016). « The switching between zonal and blocked mid-latitude atmospheric circulation: a dynamical system perspective ». en. In: *Climate Dynamics* 47.5, pp. 1587–1599. doi: 10.1007/s00382-015-2921-6 (see page 22).

Faria, S. H., I. Weikusat, and N. Azuma (2014). « The microstructure of polar ice. Part II: State of the art ». In: *Journal of Structural Geology*. Microdynamics of Ice 61, pp. 21–49. doi: 10.1016/j.jsg.2013.11.003 (see page 35).

Fernau, M. E. and P. J. Samson (1990). « Use of Cluster Analysis to Define Periods of Similar Meteorology and Precipitation Chemistry in Eastern North America. Part I: Transport Patterns ». EN. In: *Journal of Applied Meteorology and Climatology* 29.8, pp. 735–750. doi: 10.1175/1520-0450(1990)029<0735:UOCATD>2.0.CO;2 (see page 46).

Fery, L., B. Dubrulle, B. Podvin, F. Pons, and D. Faranda (2022). « Learning a Weather Dictionary of Atmospheric Patterns Using Latent Dirichlet Allocation ». en. In: *Geophysical Research Letters* 49.9, e2021GL096184. doi: 10.1029/2021GL096184 (see pages 83–85, 87).

Fink, A. H., T. Brücher, A. Krüger, G. C. Leckebusch, J. G. Pinto, et al. (2004). « The 2003 European summer heatwaves and drought - synoptic diagnosis and impacts ». en. In: *Weather* 59.8, pp. 209–216. doi: 10.1256/wea.73.04 (see page 26).

Fischer, E. M. and R. Knutti (2014). « Detection of spatially aggregated changes in temperature and precipitation extremes ». en. In: *Geophysical Research Letters* 41.2, pp. 547–554. doi: 10.1002/2013GL058499 (see page 36).

Foelsche, U., G. Kirchengast, A. K. Steiner, L. Kornblueh, E. Manzini, et al. (2008). « An observing system simulation experiment for climate monitoring with GNSS radio occultation data: Setup and test bed study ». en. In: *Journal of Geophysical Research: Atmospheres* 113.D11. doi: 10.1029/2007JD009231 (see page 30).

Foundation, C.-1. and Christopher AuYeung (2015). *Air Movement | CK-12 Foundation*. en (see page 20).

Fovell, R. G. and M.-Y. C. Fovell (1993). « Climate Zones of the Conterminous United States Defined Using Cluster Analysis ». EN. In: *Journal of Climate* 6.11, pp. 2103–2135. doi: 10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2 (see pages 41, 45).

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, et al. (2002). « Observed coherent changes in climatic extremes during the second half of the twentieth century ». en. In: *Climate Research* 19.3, pp. 193–212. doi: 10.3354/cr019193 (see page 28).

Frihat, M., B. Podvin, L. Mathelin, Y. Fraigneau, and F. Yvon (2021). « Coherent structure identification in turbulent channel flow using Latent Dirichlet Allocation ». In: *Journal of Fluid Mechanics* 920, A27. doi: 10.1017/jfm.2021.444 (see page 73).

Fukuoka, A. (1951). « The Central Meteorological Observatory,A study on 10-day forecast(A synthetic report) ». en. In: *Geophysical Magazine* 22.3, pp. 177–208 (see page 43).

Găinuşă-Bogdan, A., F. Hourdin, A. K. Traore, and P. Braconnot (2018). « Omens of coupled model biases in the CMIP5 AMIP simulations ». en. In: *Climate Dynamics* 51.7, pp. 2927–2941. doi: 10.1007/s00382-017-4057-3 (see page 34).

Greene, D. and J. P. Cross (2015). « Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis ». In: *Proceedings of the ACM Web Science Conference*. WebSci '15. New York, NY, USA: Association for Computing Machinery, pp. 1–10. doi: 10.1145/2786451.2786464 (see page 73).

Griffiths, T. L. and M. Steyvers (2004). « Finding scientific topics ». en. In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235. doi: 10.1073/pnas.0307752101 (see page 83).

Gross, A. and D. Murthy (2014). « Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing ». In: *Neural Networks*. Special Issue on "Affective Neural Networks and Cognitive Learning Systems for Big Data Analysis" 58, pp. 38–49. doi: 10.1016/j.neunet.2014.05.008 (see page 73).

Grotjahn, R. (2003). « BAROCLINIC INSTABILITY ». In: *Encyclopedia of Atmospheric Sciences*. Ed. by J. R. Holton. Oxford: Academic Press, pp. 179–188. doi: 10.1016/B0-12-227090-8/00076-2 (see page 21).

Grotjahn, R., R. Black, R. Leung, M. F. Wehner, M. Barlow, et al. (2016). « North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends ». en. In: *Climate Dynamics* 46.3, pp. 1151–1184. doi: 10.1007/s00382-015-2638-6 (see page 27).

Gumbel, E. J. (1942). « On the Frequency Distribution of Extreme Values in Meteorological Data ». In: *Bulletin of the American Meteorological Society* 23.3, pp. 95–105 (see page 29).

— (2019). *Statistics of Extremes*. en. Columbia University Press. doi: 10.7312/gumb92958 (see page 29).

Han, J., J. Pei, and H. Tong (2022). *Data Mining: Concepts and Techniques*. en. Morgan Kaufmann (see page 44).

Hanna, E., F. Pattyn, F. Navarro, V. Favier, H. Goelzer, et al. (2020). « Mass balance of the ice sheets and glaciers – Progress since AR5 and chal-

lenges ». In: *Earth-Science Reviews* 201, p. 102976. doi: 10.1016/j.earscirev.2019.102976 (see page 35).
Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007). « Empirical orthogonal functions and related techniques in atmospheric science: A review ». en. In: *International Journal of Climatology* 27.9, pp. 1119–1152. doi: 10.1002/joc.1499 (see page 41).
Hartigan, J. A. (1975). *Clustering Algorithms*. 99th. USA: John Wiley & Sons, Inc. (see page 47).
Hawkins, E., R. S. Smith, J. M. Gregory, and D. A. Stainforth (2016). « Irreducible uncertainty in near-term climate projections ». en. In: *Climate Dynamics* 46.11, pp. 3807–3819. doi: 10.1007/s00382-015-2806-8 (see page 34).
Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, et al. (2020). « The ERA5 global reanalysis ». en. In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. doi: 10.1002/qj.3803 (see pages 36, 83, 91).
Hess, N. P. and H. Brezowsky (2010). « Katalog der Großwetterlagen Europas ». de. In: (see pages 40, 56).
Hewitson, B. C. and R. G. Crane (2002). « Self-organizing maps: applications to synoptic climatology ». en. In: *Climate Research* 22.1, pp. 13–26. doi: 10.3354/cr022013 (see page 54).
Hewitson, B. C. and R. G. Crane (1994). *Neural Nets: Applications in Geography: Applications for Geography*. en. Springer Science & Business Media (see page 54).
Hewitt, H. T., M. J. Bell, E. P. Chassignet, A. Czaja, D. Ferreira, et al. (2017). « Will high-resolution global ocean models benefit coupled predictions on short-range to climate timescales? » In: *Ocean Modelling* 120, pp. 120–136. doi: 10.1016/j.ocemod.2017.11.002 (see page 35).
Hoffman, M., F. Bach, and D. Blei (2010). « Online Learning for Latent Dirichlet Allocation ». In: *Advances in Neural Information Processing Systems*. Vol. 23. Curran Associates, Inc. (see pages 75, 80, 81).
Honda, K., S. Oshio, and A. Notsu (2014). « FCM-type fuzzy co-clustering by K-L information regularization ». In: *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2505–2510. doi: 10.1109/FUZZ-IEEE.2014.6891747 (see page 58).
Hourdin, F., T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, et al. (2017). « The Art and Science of Climate Model Tuning ». EN. In: *Bulletin of the American Meteorological Society* 98.3, pp. 589–602. doi: 10.1175/BAMS-D-15-00135.1 (see page 33).
Huth, R. (1993). « An example of using obliquely rotated principal components to detect circulation types over Europe ». English. In: *Meteorologische Zeitschrift. Neue Folge; (Germany)* 2:6 (see page 50).

— (1996). « Properties of the circulation classification scheme based on the rotated principal component analysis ». en. In: *Meteorology and Atmospheric Physics* 59.3, pp. 217–233. doi: 10.1007/BF01030145 (see pages 50, 51).
— (2000). « A circulation classification scheme applicable in GCM studies ». en. In: *Theoretical and Applied Climatology* 67.1, pp. 1–18. doi: 10.1007/s007040070012 (see page 51).
Huth, R. (1997). « Continental-Scale Circulation in the UKHI GCM ». EN. In: *Journal of Climate* 10.7, pp. 1545–1561. doi: 10.1175/1520-0442(1997)010<1545:CSCITU>2.0.CO;2 (see page 51).
— (2010). « Synoptic-climatological applicability of circulation classifications from the COST733 collection: First results ». In: *Physics and Chemistry of the Earth, Parts A/B/C*. Classifications of Atmospheric Circulation Patterns – Theory and Applications 35.9, pp. 388–394. doi: 10.1016/j.pce.2009.11.013 (see page 46).
Huth, R., C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, et al. (2008). « Classifications of Atmospheric Circulation Patterns ». en. In: *Annals of the New York Academy of Sciences* 1146.1, pp. 105–152. doi: 10.1196/annals.1446.019 (see pages 39, 40).
Ichihashi, H., K. Miyagishi, and K. Honda (2001). « Fuzzy c-means clustering with regularization by K-L information ». In: *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*. Vol. 2, 924–927 vol.3. doi: 10.1109/FUZZ.2001.1009107 (see page 59).
Infrastructure, A. E.-S. S. .-. N. R. (2024). *Earth System Model 1.5 (ESM1.5)*. en-US (see page 31).
Intergovernmental Panel On Climate Change (2023). « Chapter 2: Changing State of the Climate System ». en. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. doi: 10.1017/9781009157896.
« Chapter 9: Evaluation of Climate Models » (2014). In: *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Intergovernmental Panel on Climate Change (IPCC). Cambridge: Cambridge University Press, pp. 741–866. doi: 10.1017/CBO9781107415324.020 (see page 34).
« Summary for Policymakers » (2023). In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by Intergovernmental Panel on Climate Change (IPCC). Cambridge: Cambridge University Press, pp. 3–32. doi: 10.1017/9781009157896.001 (see pages 24, 25, 28–30).

IPCC (2023a). « Chapter 1: Framing, Context, and Methods ». en. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. doi: 10.1017/9781009157896 (see page 35).
— (2023b). « Chapter 11: Weather and Climate Extreme Events in a Changing Climate ». en. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. doi: 10.1017/9781009157896 (see pages 26, 28).
— (2023c). « Chapter 3: Human influence on the climate system ». en. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. doi: 10.1017/9781009157896 (see page 34).
— (2023d). « Chapter 4: Future Global Climate: Scenario-based Projections and Near-term Information ». en. In: *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. doi: 10.1017/9781009157896 (see page 32).
— (2023e). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press. doi: 10.1017/9781009157896 (see page 24).
Ipcc (2022). *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. en. 1st ed. Cambridge University Press. doi: 10.1017/9781009157940 (see page 24).
IPCC, 2. (2024). *Glossary — IPCC* (see pages 19, 24).
Jacobeit, J., H. Wanner, J. Luterbacher, C. Beck, A. Philipp, et al. (2003). « Atmospheric circulation variability in the North-Atlantic-European area since the mid-seventeenth century ». en. In: *Climate Dynamics* 20.4, pp. 341–352. doi: 10.1007/s00382-002-0278-0 (see page 51).
James, P. M. (2007). « An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe ». en. In: *Theoretical and Applied Climatology* 88.1, pp. 17–42. doi: 10.1007/s00704-006-0239-3 (see page 40).
Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. en. MIT Press (see page 75).

Jelodar, H., Y. Wang, C. Yuan, X. Feng, X. Jiang, et al. (2019). « Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey ». en. In: *Multimedia Tools and Applications* 78.11, pp. 15169–15211. doi: 10.1007/s11042-018-6894-4 (see page 73).

Jeong, D. I., B. Yu, and A. J. Cannon (2021). « Links between atmospheric blocking and North American winter cold spells in two generations of Canadian Earth System Model large ensembles ». en. In: *Climate Dynamics* 57.7, pp. 2217–2231. doi: 10.1007/s00382-021-05801-0.

Jiaxiang, G., M. Shoshiro, M. J. Roberts, R. Haarsma, D. Putrasahan, et al. (2020). « Influence of model resolution on bomb cyclones revealed by HighResMIP-PRIMAVERA simulations ». en. In: *Environmental Research Letters* 15.8, p. 084001. doi: 10.1088/1748-9326/ab88fa (see page 35).

Jones, C. D. and P. Friedlingstein (2020). « Quantifying process-level uncertainty contributions to TCRE and carbon budgets for meeting Paris Agreement climate targets ». en. In: *Environmental Research Letters* 15.7, p. 074019. doi: 10.1088/1748-9326/ab858a (see page 35).

Jones, P. D., M. Hulme, and K. R. Briffa (1993). « A comparison of Lamb circulation types with an objective classification scheme ». en. In: *International Journal of Climatology* 13.6, pp. 655–663. doi: 10.1002/joc.3370130606 (see page 40).

Jordan, M. I. (2004). « Graphical Models ». In: *Statistical Science* 19.1, pp. 140–155. doi: 10.1214/088342304000000026 (see page 63).

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). « An Introduction to Variational Methods for Graphical Models ». en. In: *Machine Learning* 37.2, pp. 183–233. doi: 10.1023/A:1007665907178 (see page 79).

Kaiser, H. F. (1958). « The varimax criterion for analytic rotation in factor analysis ». en. In: *Psychometrika* 23.3, pp. 187–200. doi: 10.1007/BF02289233 (see page 50).

Kalkstein, L. S., G. Tan, and J. A. Skindlov (1987). « An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification ». EN. In: *Journal of Applied Meteorology and Climatology* 26.6, pp. 717–730. doi: 10.1175/1520-0450(1987)026<0717:AEOTCP>2.0.CO;2 (see page 46).

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, et al. (1996). « The NCEP/NCAR 40-Year Reanalysis Project ». en. In: *Bulletin of the American Meteorological Society* 77.3, pp. 437–472. doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2 (see pages 36, 83, 86).

Kay, J. E., C. Deser, A. Phillips, A. Mai, C. Hannay, et al. (2015). « The Community Earth System Model (CESM) Large Ensemble Project: A Com-

munity Resource for Studying Climate Change in the Presence of Internal Climate Variability ». EN. In: *Bulletin of the American Meteorological Society* 96.8, pp. 1333–1349. doi: 10.1175/BAMS-D-13-00255.1 (see page 34).

KELLEY, H. J. (1960). « Gradient Theory of Optimal Flight Paths ». In: *ARS Journal* 30.10, pp. 947–954. doi: 10.2514/8.5282 (see page 52).

Kharin, V. V., F. W. Zwiers, X. Zhang, and M. Wehner (2013). « Changes in temperature and precipitation extremes in the CMIP5 ensemble ». en. In: *Climatic Change* 119.2, pp. 345–357. doi: 10.1007/s10584-013-0705-8 (see page 35).

Kherif, F. and A. Latypova (2020). « Chapter 12 - Principal component analysis ». In: *Machine Learning*. Ed. by A. Mechelli and S. Vieira. Academic Press, pp. 209–225. doi: 10.1016/B978-0-12-815739-8.00012-2 (see page 41).

Kidson, J. W. (1997). « The Utility of Surface and Upper Air Data in Synoptic Climatological Specification of Surface Climatic Variables ». en. In: *International Journal of Climatology* 17.4, pp. 399–413. doi: 10.1002/(SICI)1097-0088(19970330)17:4<399::AID-JOC108>3.0.CO;2-M (see page 40).

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, et al. (2001). « The NCEP–NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation ». In: *Bulletin of the American Meteorological Society* 82.2, pp. 247–268 (see pages 36, 83, 86).

Knutti, R. (2018). « Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World ». en. In: *Climate Modelling: Philosophical and Conceptual Issues*. Ed. by E. A. Lloyd and E. Winsberg. Cham: Springer International Publishing, pp. 325–359. doi: 10.1007/978-3-319-65058-6_11.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010). « Challenges in Combining Projections from Multiple Climate Models ». EN. In: *Journal of Climate* 23.10, pp. 2739–2758. doi: 10.1175/2009JCLI3361.1 (see page 34).

Kohonen, T. (1990). « The self-organizing map ». In: *Proceedings of the IEEE* 78.9, pp. 1464–1480. doi: 10.1109/5.58325 (see page 52).

Kohonen, T. (1995). « The Basic SOM ». en. In: *Self-Organizing Maps*. Ed. by T. Kohonen. Berlin, Heidelberg: Springer, pp. 77–130. doi: 10.1007/978-3-642-97610-0_3 (see pages 52, 54, 55).

Krishnamurti, T. N. (1961). « THE SUBTROPICAL JET STREAM OF WINTER ». EN. In: *Journal of the Atmospheric Sciences* 18.2, pp. 172–191. doi: 10.1175/1520-0469(1961)018<0172:TSJSOW>2.0.CO;2 (see page 22).

Kuhlbrodt, T., C. G. Jones, A. Sellar, D. Storkey, E. Blockley, et al. (2018). « The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate ». en. In: *Journal of Advances in Modeling Earth Systems* 10.11, pp. 2865–2888. doi: 10.1029/2018MS001370 (see page 34).

Kumbhar, A. (2023). *Geomorphology - Cyclones and Anticyclones*. English (see page 21).

Kumler-Bonfanti, C., J. Stewart, D. Hall, and M. Govett (2020). « Tropical and Extratropical Cyclone Detection Using Deep Learning ». EN. In: *Journal of Applied Meteorology and Climatology* 59.12, pp. 1971–1985. doi: 10.1175/JAMC-D-20-0117.1 (see page 28).

Laarhoven, P. J. M. v. and E. H. L. Aarts (1987). *Simulated annealing: theory and applications*. eng. Mathematics and its applications (D. Reidel Publishing Company). Dordrecht, Boston: D. Reidel ; Sold, distributed in the U.S.A., and Canada by Kluwer Academic Publishers (see page 48).

Lamb, H. H. (1948). « On the General Circulation of the Atmosphere in Middle Latitudes: Southern and Northern Hemispheres Compared ». In: *Bulletin of the American Meteorological Society* 29.8, pp. 391–394 (see page 21).

Lamb, H. H. (1972). « British isles weather types and a register of the daily sequence of circu lation patterns 1861-1971 ». In: *Geophysical Memoirs* (see page 40).

Lamb, P. (1985). « Climatic pattern analysis of 3-and 7-day summer rainfall in the central United States: Some methodological considerations and a regionalization ». In: *J. Climate Appl. Meteor* 24, pp. 1325–1343 (see page 50).

Lauer, A., C. Jones, V. Eyring, M. Evaldsson, S. Hagemann, et al. (2018). « Process-level improvements in CMIP5 models and their impact on tropical variability, the Southern Ocean, and monsoons ». English. In: *Earth System Dynamics* 9.1, pp. 33–67. doi: 10.5194/esd-9-33-2018 (see page 34).

Lee, L. A., K. S. Carslaw, K. J. Pringle, G. W. Mann, and D. V. Spracklen (2011). « Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters ». English. In: *Atmospheric Chemistry and Physics* 11.23, pp. 12253–12273. doi: 10.5194/acp-11-12253-2011 (see page 34).

Li, C., F. Zwiers, X. Zhang, G. Li, Y. Sun, et al. (2021). « Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models ». EN. In: *Journal of Climate* 34.9, pp. 3441–3460. doi: 10.1175/JCLI-D-19-1013.1 (see page 35).

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. en. IMS (see page 59).
Liu, Y., E. Racah, Prabhat, J. Correa, A. Khosrowshahi, et al. (2016). *Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets* (see page 28).
Loehle, C. (2018). « The epistemological status of general circulation models ». en. In: *Climate Dynamics* 50.5, pp. 1719–1731. doi: 10.1007/s00382-017-3717-7 (see page 32).
López-Bueno, J. A., M. Á. Navas-Martín, J. Díaz, I. J. Mirón, M. Y. Luna, et al. (2021). « The effect of cold waves on mortality in urban and rural areas of Madrid ». In: *Environmental Sciences Europe* 33.1, p. 72. doi: 10.1186/s12302-021-00512-z (see page 26).
Lorenz, E. N. (1963). « Deterministic Nonperiodic Flow ». EN. In: *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
Lorenz, Edward N (1956). *Empirical orthogonal functions and statistical weather prediction*. Vol. 1. Massachusetts Institute of Technology, Department of Meteorology Cambridge (see page 43).
Lucas-Picher, P., D. Argüeso, E. Brisson, Y. Tramblay, P. Berg, et al. (2021). « Convection-permitting modeling with regional climate models: Latest developments and next steps ». en. In: *WIREs Climate Change* 12.6, e731. doi: 10.1002/wcc.731 (see page 32).
Luo, B., D. Luo, A. Dai, I. Simmonds, and L. Wu (2020). « Combined Influences on North American Winter Air Temperature Variability from North Pacific Blocking and the North Atlantic Oscillation: Subseasonal and Interannual Time Scales ». In: *Journal of Climate* 33, pp. 7101–7123. doi: 10.1175/JCLI-D-19-0327.1 (see page 24).
Lupo, A. R. (2021). « Atmospheric blocking events: a review ». en. In: *Annals of the New York Academy of Sciences* 1504.1, pp. 5–24. doi: 10.1111/nyas.14557 (see page 22).
Lynch, P. (2008). « The origins of computer weather prediction and climate modeling ». In: *Journal of Computational Physics*. Predicting weather, climate and extreme events 227.7, pp. 3431–3444. doi: 10.1016/j.jcp.2007.02.034 (see page 32).
MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. en. Cambridge University Press.
Madden, R. A. (1979). « Observations of large-scale traveling Rossby waves ». en. In: *Reviews of Geophysics* 17.8, pp. 1935–1949. doi: 10.1029/RG017i008p01935 (see page 22).
Maher, N., S. Milinski, L. Suarez-Gutierrez, M. Botzet, M. Dobrynin, et al. (2019). « The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability ». en. In: *Journal of Advances*

in Modeling Earth Systems 11.7, pp. 2050–2069. doi: 10.1029/2019MS001639 (see page 34).
Malhomme, N., B. Podvin, D. Faranda, and L. Mathelin (2025). « Evaluation of atmospheric circulation of CMIP6 models for extreme temperature events using Latent Dirichlet Allocation ». In: *Journal of Climate* -1 (aop). Publisher: American Meteorological Society Section: Journal of Climate. doi: 10.1175/JCLI-D-23-0719.1 (see page 109).
Masson, D. and R. Knutti (2011). « Climate model genealogy ». en. In: *Geophysical Research Letters* 38.8. doi: 10.1029/2011GL046864 (see page 34).
McCarthy, M., N. Christidis, N. Dunstone, D. Fereday, G. Kay, et al. (2019). « Drivers of the UK summer heatwave of 2018 ». In: *Weather* 74.11, pp. 390–396. doi: 10.1002/wea.3628 (see page 26).
McCullagh, P. (2002). « What is a statistical model? » In: *The Annals of Statistics* 30.5, pp. 1225–1310. doi: 10.1214/aos/1035844977 (see page 59).
McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, et al. (2017). « Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather ». EN. In: *Bulletin of the American Meteorological Society* 98.10, pp. 2073–2090. doi: 10.1175/BAMS-D-16-0123.1 (see page 28).
McPhillips, L. E., H. Chang, M. V. Chester, Y. Depietri, E. Friedman, et al. (2018). « Defining Extreme Events: A Cross-Disciplinary Review ». en. In: *Earth's Future* 6.3, pp. 441–455. doi: 10.1002/2017EF000686 (see page 26).
Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer (2000). « The Coupled Model Intercomparison Project (CMIP) ». en. In: *Bulletin of the American Meteorological Society* 81.2, pp. 313–318. doi: 10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2 (see page 33).
Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, et al. (2007). « THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research ». en. In: *Bulletin of the American Meteorological Society* 88.9, pp. 1383–1394. doi: 10.1175/BAMS-88-9-1383.
Meehl, G. A., C. A. Senior, V. Eyring, G. Flato, J.-F. Lamarque, et al. (2020). « Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models ». In: *Science Advances* 6.26, eaba1981. doi: 10.1126/sciadv.aba1981 (see page 35).
Michelangeli, P.-A., R. Vautard, and B. Legras (1995). « Weather Regimes: Recurrence and Quasi Stationarity ». EN. In: *Journal of the Atmospheric Sciences* 52.8, pp. 1237–1256. doi: 10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2 (see pages 43–45).

Mignot, J., F. Hourdin, J. Deshayes, O. Boucher, G. Gastineau, et al. (2021). « The Tuning Strategy of IPSL-CM6A-LR ». en. In: *Journal of Advances in Modeling Earth Systems* 13.5, e2020MS002340. doi: 10.1029/2020MS002340 (see page 33).

Mote, T. L. (1998). « Mid–tropospheric circulation and surface melt on the Greenland ice sheet. Part II: synoptic climatology ». en. In: *International Journal of Climatology* 18.2, pp. 131–145. doi: 10.1002/(SICI)1097-0088(199802)18:2<131::AID-JOC228>3.0.CO;2-S (see page 46).

Müller, G. V., R. Compagnucci, M. N. Nuñez, and A. Salles (2003). « Surface circulation associated with frost in the wet Pampas ». en. In: *International Journal of Climatology* 23.8, pp. 943–961. doi: 10.1002/joc.907 (see page 51).

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, et al. (2004). « Quantification of modelling uncertainties in a large ensemble of climate change simulations ». en. In: *Nature* 430.7001, pp. 768–772. doi: 10.1038/nature02771 (see page 34).

Mwasiagi, J. I. (2011). *Self Organizing Maps: Applications and Novel Algorithm Design*. en. BoD – Books on Demand (see page 54).

Nairn, J., R. Fawcett, and D. Ray (2009). « Defining and predicting excessive heat events, a national system ». In: *Modelling and understanding high impact weather: extended abstracts of the third CAWCR Modelling Workshop*. Vol. 30. Citeseer, pp. 83–86 (see page 27).

Nigam, K., A. K. Mccallum, S. Thrun, and T. Mitchell (2000). « Text Classification from Labeled and Unlabeled Documents using EM ». en. In: *Machine Learning* 39.2, pp. 103–134. doi: 10.1023/A:1007692713085 (see page 69).

North, G. R., R. F. Cahalan, and J. A. Coakley Jr. (1981). « Energy balance climate models ». en. In: *Reviews of Geophysics* 19.1, pp. 91–121. doi: 10.1029/RG019i001p00091 (see page 31).

Notz, D. (2015). « How well must climate models agree with observations? » In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2052, p. 20140164. doi: 10.1098/rsta.2014.0164 (see page 33).

Oldenborgh, G. J. van, S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, et al. (2009). « Western Europe is warming much faster than expected ». English. In: *Climate of the Past* 5.1, pp. 1–12. doi: 10.5194/cp-5-1-2009 (see page 34).

Orbach, J. (1962). « Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. » In: *Archives of General Psychiatry* 7.3, pp. 218–219. doi: 10.1001/archpsyc.1962.01720030064010 (see page 52).

Overpeck, J. T. (1995). « Paleoclimatology and climate system dynamics ». en. In: *Reviews of Geophysics* 33.S2, pp. 863–871. doi: 10.1029/95RG01035 (see page 32).
Palmer, T. N. (2019). « Stochastic weather and climate models ». en. In: *Nature Reviews Physics* 1.7, pp. 463–471. doi: 10.1038/s42254-019-0062-2 (see page 35).
Papagiannaki, K., K. Lagouvardos, V. Kotroni, and G. Papagiannakis (2014). « Agricultural losses related to frost events: use of the 850 hPa level temperature as an explanatory variable of the damage cost ». English. In: *Natural Hazards and Earth System Sciences* 14.9, pp. 2375–2386. doi: 10.5194/nhess-14-2375-2014 (see page 26).
Parker, W. S. (2009). « II—Wendy S. Parker: Confirmation and adequacy-for-Purpose in Climate Modelling ». In: *Aristotelian Society Supplementary Volume* 83.1, pp. 233–249. doi: 10.1111/j.1467-8349.2009.00180.x (see page 33).
— (2020). « Model Evaluation: An Adequacy-for-Purpose View ». en. In: *Philosophy of Science* 87.3, pp. 457–477. doi: 10.1086/708691 (see page 33).
Perkins, S. E. and L. V. Alexander (2013). « On the Measurement of Heat Waves ». EN. In: *Journal of Climate* 26.13, pp. 4500–4517. doi: 10.1175/JCLI-D-12-00383.1 (see page 27).
Philipp, A., P. M. Della-Marta, J. Jacobeit, D. R. Fereday, P. D. Jones, et al. (2007). « Long-Term Variability of Daily North Atlantic–European Pressure Patterns since 1850 Classified by Simulated Annealing Clustering ». EN. In: *Journal of Climate* 20.16, pp. 4065–4095. doi: 10.1175/JCLI4175.1 (see page 48).
Philipp, A., C. Beck, R. Huth, and J. Jacobeit (2016). « Development and comparison of circulation type classifications using the COST 733 dataset and software ». en. In: *International Journal of Climatology* 36.7, pp. 2673–2691. doi: 10.1002/joc.3920 (see pages 46, 60).
Phillips, T. J., G. L. Potter, D. L. Williamson, R. T. Cederwall, J. S. Boyle, et al. (2004). « Evaluating Parameterizations in General Circulation Models: Climate Simulation Meets Weather Prediction ». en. In: *Bulletin of the American Meteorological Society* 85.12, pp. 1903–1916. doi: 10.1175/BAMS-85-12-1903 (see page 34).
Pochon, L.-O. and A. Favre (2022). « Introduction aux algorithmes d' " Expectation Maximisation (EM) " et d'échantillonnage de Gibbs et leur application à la classification de documents » (see page 77).
Podvin, B., L. Soucasse, and F. Yvon (2024). « Analysis of Rayleigh-B\énard convection using latent Dirichlet allocation ». In: *Physical Review Flu-*

ids 9.6, p. 063502. doi: 10.1103/PhysRevFluids.9.063502 (see page 73).
Preisendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*. English. Amsterdam ; New York : New York, NY, U.S.A: Elsevier Science Ltd (see page 41).
programme, U. environment (2024). *IPCC — Intergovernmental Panel on Climate Change* (see page 24).
Řehůřek, R. and P. Sojka (2010). « Software Framework for Topic Modelling with Large Corpora ». English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50 (see page 75).
Rex, D. F. (1950). « Blocking Action in the Middle Troposphere and its Effect upon Regional Climate ». In: *Tellus* 2.4, pp. 275–301. doi: 10.3402/tellusa.v2i4.8603 (see page 43).
Riahi, K., D. P. van Vuuren, E. Kriegler, J. Edmonds, B. C. O'Neill, et al. (2017). « The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview ». In: *Global Environmental Change* 42, pp. 153–168. doi: 10.1016/j.gloenvcha.2016.05.009.
Richman, M. B. (1986). « ROTATION OF PRINCIPAL COMPONENTS ». en. In: *Journal of Climatology* (see pages 49, 50).
— (1990). « Review of Principal Component Analysis in Meteorology and Oceanography ». In: *Bulletin of the American Meteorological Society* 71.2, pp. 212–214 (see page 43).
Ringard, J., B. Dieppois, S. Rome, A. Diedhiou, T. Pellarin, et al. (2016). « The intensification of thermal extremes in west Africa ». In: *Global and Planetary Change* 139, pp. 66–77. doi: 10.1016/j.gloplacha.2015.12.009 (see page 36).
Rodrigues, D., M. C. Alvarez-Castro, G. Messori, P. Yiou, Y. Robin, et al. (2018a). « Dynamical Properties of the North Atlantic Atmospheric Circulation in the Past 150 Years in CMIP5 Models and the 20CRv2c Reanalysis ». EN. In: *Journal of Climate* 31.15, pp. 6097–6111. doi: 10.1175/JCLI-D-17-0176.1 (see page 34).
— (2018b). « Dynamical Properties of the North Atlantic Atmospheric Circulation in the Past 150 Years in CMIP5 Models and the 20CRv2c Reanalysis ». EN. In: *Journal of Climate* 31.15, pp. 6097–6111. doi: 10.1175/JCLI-D-17-0176.1 (see page 87).
Rohde, R. A. and Z. Hausfather (2020). « The Berkeley Earth Land/Ocean Temperature Record ». English. In: *Earth System Science Data* 12.4, pp. 3469–3479. doi: 10.5194/essd-12-3469-2020 (see page 30).

Rossby, C. G. and H. C. Willett (1948). « The Circulation of the Upper Troposphere and Lower Stratosphere ». In: *Science* 108.2815, pp. 643–652. doi: 10.1126/science.108.2815.643 (see page 21).
Rossby, C.-G. (1949). « On the Dispersion of Planetary Waves in a Barotropic: Atmosphere ». en. In: *Tellus* 1.1, pp. 54–58. doi: 10.1111/j.2153-3490.1949.tb01928.x (see page 22).
Rousi, E., K. Kornhuber, G. Beobide-Arsuaga, F. Luo, and D. Coumou (2022). « Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia ». en. In: *Nature Communications* 13.1, p. 3851. doi: 10.1038/s41467-022-31432-y (see page 28).
Saha, S., S. Moorthi, H.-L. Pan, X. Wu, J. Wang, et al. (2010). « The NCEP Climate Forecast System Reanalysis ». en. In: *Bulletin of the American Meteorological Society* 91.8, pp. 1015–1058. doi: 10.1175/2010BAMS3001.1 (see page 36).
Saunois, M. (2022). « Modèles pour l'étude des climats passés, présents et futurs ». In: *Université de Versailles* (see page 31).
Scaife, A. A., T. Woollings, J. Knight, G. Martin, and T. Hinton (2010). « Atmospheric Blocking and Mean Biases in Climate Models ». EN. In: *Journal of Climate* 23.23, pp. 6143–6152. doi: 10.1175/2010JCLI3728.1 (see page 34).
Schiemann, R., P. Athanasiadis, D. Barriopedro, F. Doblas-Reyes, K. Lohmann, et al. (2020). « Northern Hemisphere blocking simulation in current climate models: evaluating progress from the Climate Model Intercomparison Project Phase 5 to 6 and sensitivity to resolution ». English. In: *Weather and Climate Dynamics* 1.1, pp. 277–292. doi: 10.5194/wcd-1-277-2020 (see page 35).
Schmidt, G. A., D. Bader, L. J. Donner, G. S. Elsaesser, J.-C. Golaz, et al. (2017). « Practice and philosophy of climate model tuning across six US modeling centers ». English. In: *Geoscientific Model Development* 10.9, pp. 3207–3223. doi: 10.5194/gmd-10-3207-2017 (see page 33).
Selim, S. Z. and K. Alsultan (1991). « A simulated annealing algorithm for the clustering problem ». In: *Pattern Recognition* 24.10, pp. 1003–1008. doi: 10.1016/0031-3203(91)90097-O (see page 48).
Sheridan, S. C. and C. C. Lee (2011). « The self-organizing map in synoptic climatological research ». en. In: *Progress in Physical Geography: Earth and Environment* 35.1, pp. 109–119. doi: 10.1177/0309133310397582 (see page 54).
Shiogama, H., M. Watanabe, T. Ogura, T. Yokohata, and M. Kimoto (2014). « Multi-parameter multi-physics ensemble (MPMPE): a new approach exploring the uncertainties of climate sensitivity ». en. In: *Atmospheric Science Letters* 15.2, pp. 97–102. doi: 10.1002/asl2.472 (see page 34).

Sillmann, J., M. G. Donat, J. C. Fyfe, and F. W. Zwiers (2014). « Observed and simulated temperature extremes during the recent warming hiatus ». en. In: *Environmental Research Letters* 9.6, p. 064023. doi: 10.1088/1748-9326/9/6/064023 (see page 36).

Solomon, S., Intergovernmental Panel on Climate Change, and Intergovernmental Panel on Climate Change, eds. (2007). *Climate change 2007: the physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. en. Cambridge ; New York: Cambridge University Press (see page 20).

Stehlík, J. and A. Bárdossy (2003). « Statistical comparison of European circulation patterns and development of a continental scale classification ». en. In: *Theoretical and Applied Climatology* 76.1, pp. 31–46. doi: 10.1007/s00704-003-0007-6 (see page 56).

Stott, P. A., D. A. Stone, and M. R. Allen (2004). « Human contribution to the European heatwave of 2003 ». en. In: *Nature* 432.7017, pp. 610–614. doi: 10.1038/nature03089 (see page 26).

Suganya, R. and R. Shanthi (2012). « Fuzzy c-means algorithm-a review ». In: *International Journal of Scientific and Research Publications* 2.11, p. 1 (see page 58).

Swart, N. C., J. N. S. Cole, V. V. Kharin, M. Lazare, J. F. Scinocca, et al. (2019). « The Canadian Earth System Model version 5 (CanESM5.0.3) ». English. In: *Geoscientific Model Development* 12.11, pp. 4823–4873. doi: 10.5194/gmd-12-4823-2019 (see pages 36, 110).

Tatebe, H., T. Ogura, T. Nitta, Y. Komuro, K. Ogochi, et al. (2019). « Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6 ». English. In: *Geoscientific Model Development* 12.7, pp. 2727–2765. doi: 10.5194/gmd-12-2727-2019 (see pages 36, 110).

Thurstone, L. L. (1933). *The theory of multiple factors*. The theory of multiple factors. Oxford, England: Author (see page 50).

Tian, B. and X. Dong (2020). « The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation ». en. In: *Geophysical Research Letters* 47.8, e2020GL087232. doi: 10.1029/2020GL087232 (see page 35).

US Department of Commerce, N. (2024). *National Weather Service*. EN-US (see page 23).

Valle, D., P. Albuquerque, Q. Zhao, A. Barberan, and R. J. Fletcher Jr. (2018). « Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change ». en. In:

Global Change Biology 24.11, pp. 5560–5572. doi: 10.1111/gcb.14412 (see page 73).
Vautard, R. (1990). « Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors ». EN. In: *Monthly Weather Review* 118.10, pp. 2056–2081. doi: 10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2 (see page 44).
Vautard, R., G. J. van Oldenborgh, R. Bonnet, S. Li, Y. Robin, et al. (2023). « Human influence on growing-period frosts like in early April 2021 in central France ». English. In: *Natural Hazards and Earth System Sciences* 23.3, pp. 1045–1058. doi: 10.5194/nhess-23-1045-2023 (see page 26).
Vrac, M., K. Hayhoe, and M. Stein (2007). « Identification and intermodel comparison of seasonal circulation patterns over North America ». In: *International Journal of Climatology* 27.5, pp. 603–620 (see pages 46, 60).
Wallace, J. M. and P. V. Hobbs (2006). *Atmospheric Science: An Introductory Survey*. en. Elsevier (see pages 19, 23).
Wehrli, K., B. P. Guillod, M. Hauser, M. Leclair, and S. I. Seneviratne (2018). « Assessing the Dynamic Versus Thermodynamic Origin of Climate Model Biases ». en. In: *Geophysical Research Letters* 45.16, pp. 8471–8479. doi: 10.1029/2018GL079220 (see page 142).
Weilnhammer, V., J. Schmid, I. Mittermeier, F. Schreiber, L. Jiang, et al. (2021). « Extreme weather events in europe and their health consequences – A systematic review ». In: *International Journal of Hygiene and Environmental Health* 233, p. 113688. doi: 10.1016/j.ijheh.2021.113688 (see page 26).
Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. en. Academic Press (see page 29).
Williams, K. D., A. Bodas-Salcedo, M. Déqué, S. Fermepin, B. Medeiros, et al. (2013). « The Transpose-AMIP II Experiment and Its Application to the Understanding of Southern Ocean Cloud Biases in Climate Models ». EN. In: *Journal of Climate* 26.10, pp. 3258–3274. doi: 10.1175/JCLI-D-12-00429.1 (see page 34).
Winsberg, E. (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press. doi: 10.1017/9781108164290 (see page 33).
Wold, S., K. Esbensen, and P. Geladi (1987). « Principal component analysis ». In: *Chemometrics and Intelligent Laboratory Systems*. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists 2.1, pp. 37–52. doi: 10.1016/0169-7439(87)80084-9 (see pages 41, 96).

Yang, M. .-. (1993). « A survey of fuzzy clustering ». In: *Mathematical and Computer Modelling* 18.11, pp. 1–16. doi: 10.1016/0895-7177(93)90202-A (see pages 55, 57, 58).
Yvon, F. (2022). « Probabilistic Graphical Models, course notes ». In: *Université Paris-Saclay* (see pages 63, 67).
Zadeh, L. A. (1965). « Fuzzy sets ». In: *Information and Control* 8.3, pp. 338–353. doi: 10.1016/S0019-9958(65)90241-X (see pages 56, 58).
Ziehn, T., M. A. Chamberlain, R. M. Law, A. Lenton, R. W. Bodman, et al. (2020). « The Australian Earth System Model: ACCESS-ESM1.5 ». en. In: *Journal of Southern Hemisphere Earth Systems Science* 70.1, pp. 193–214. doi: 10.1071/ES19035 (see pages 31, 36, 110).

A - Extreme events in European countries

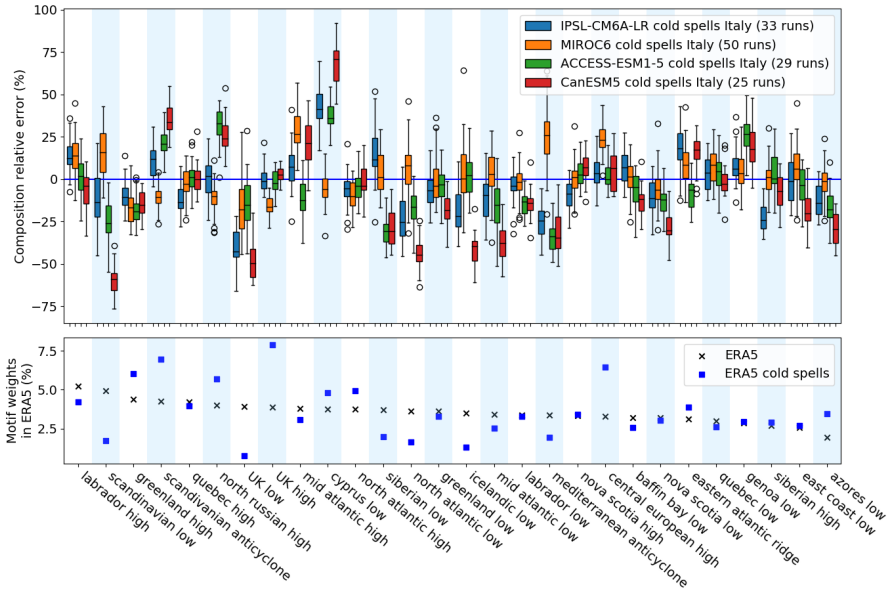


Figure A.1 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Italy. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

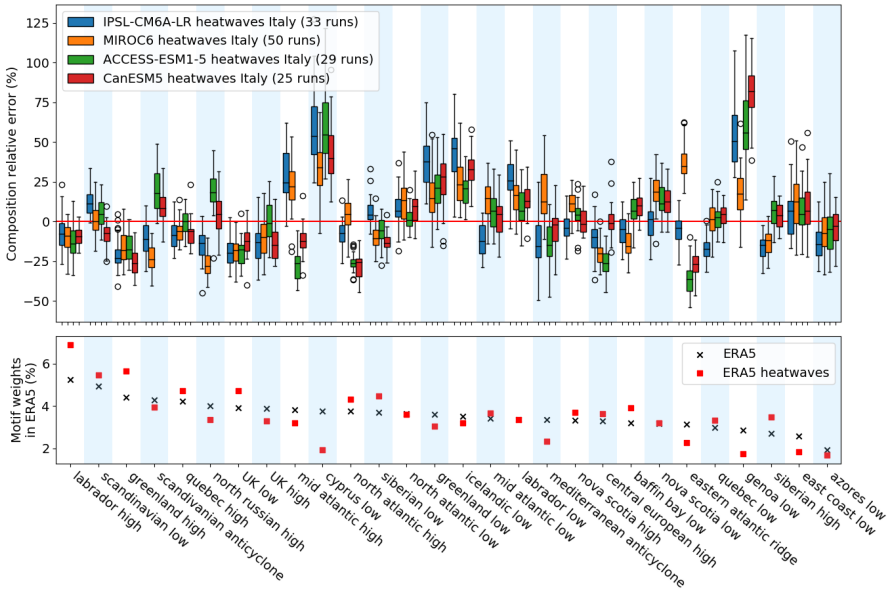


Figure A.2 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Italy. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

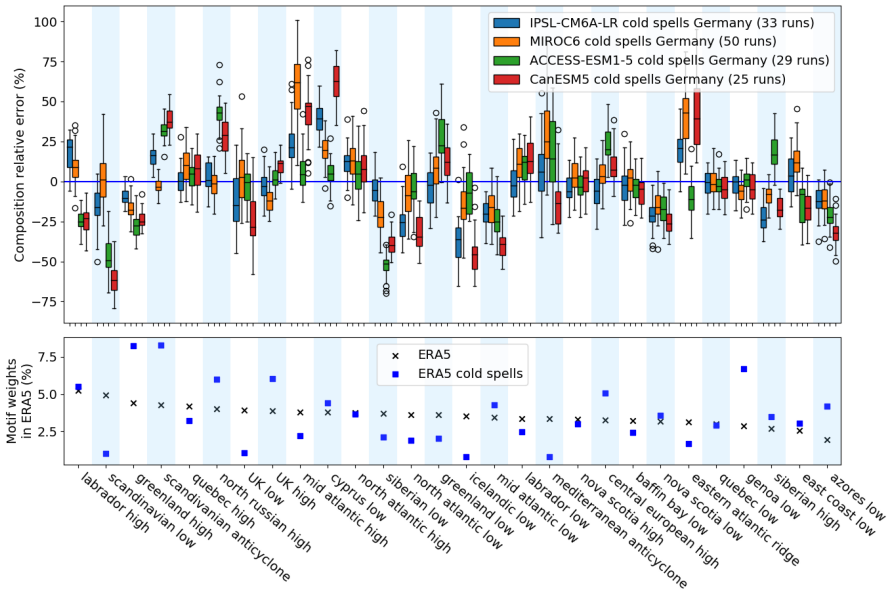


Figure A.3 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Germany. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

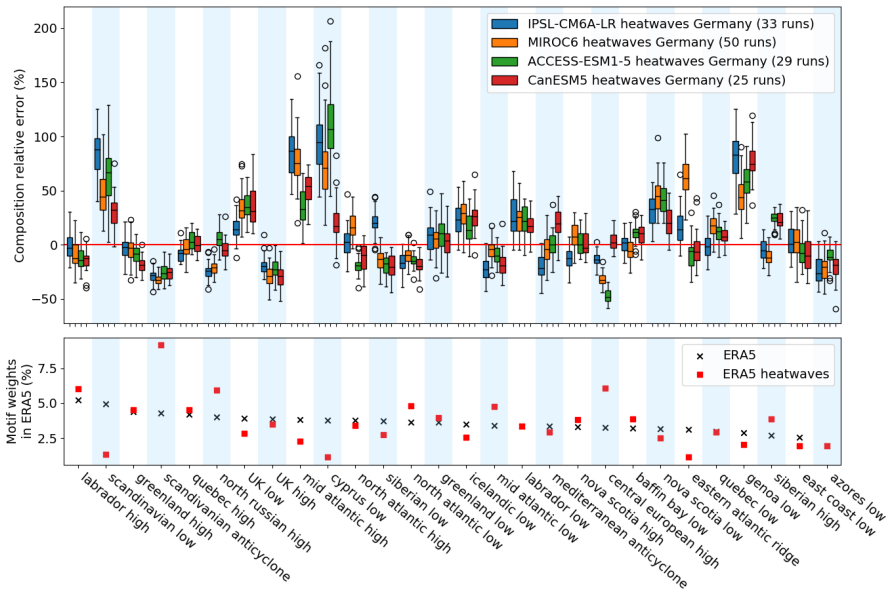


Figure A.4 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Germany. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

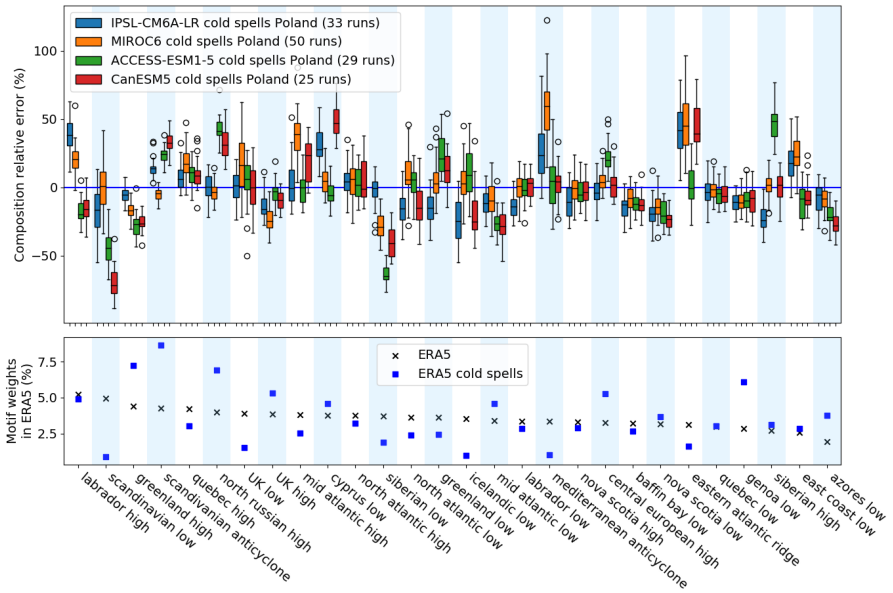


Figure A.5 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Poland. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

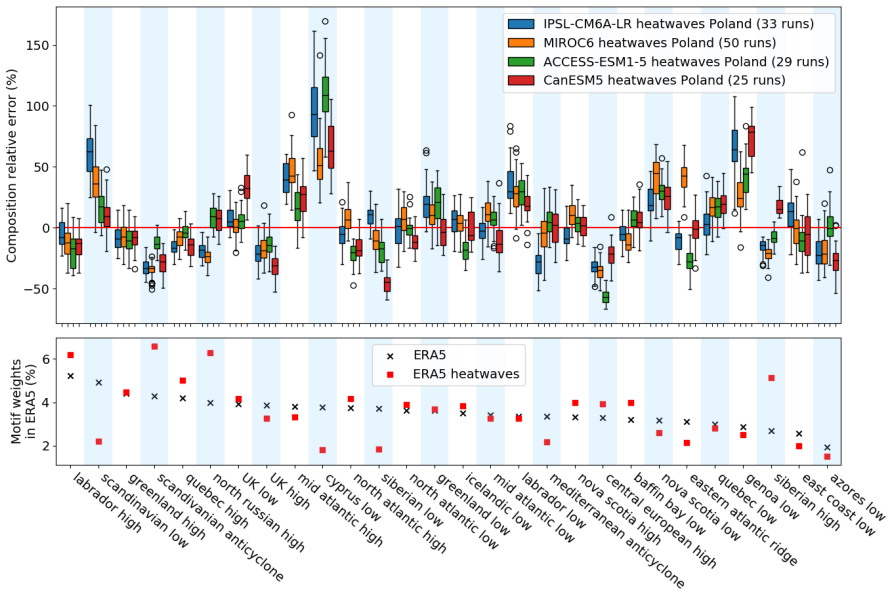


Figure A.6 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Poland. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

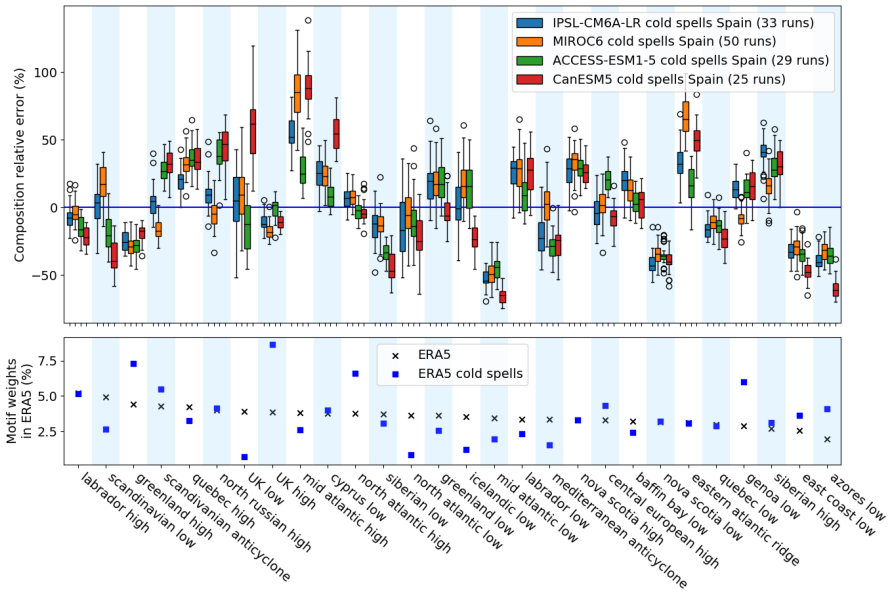


Figure A.7 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in Spain. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

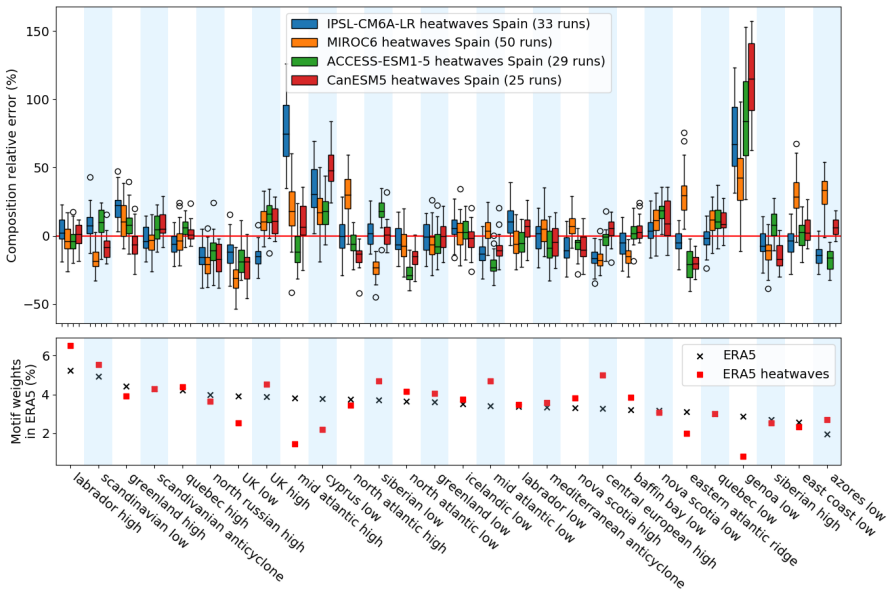


Figure A.8 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in Spain. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.

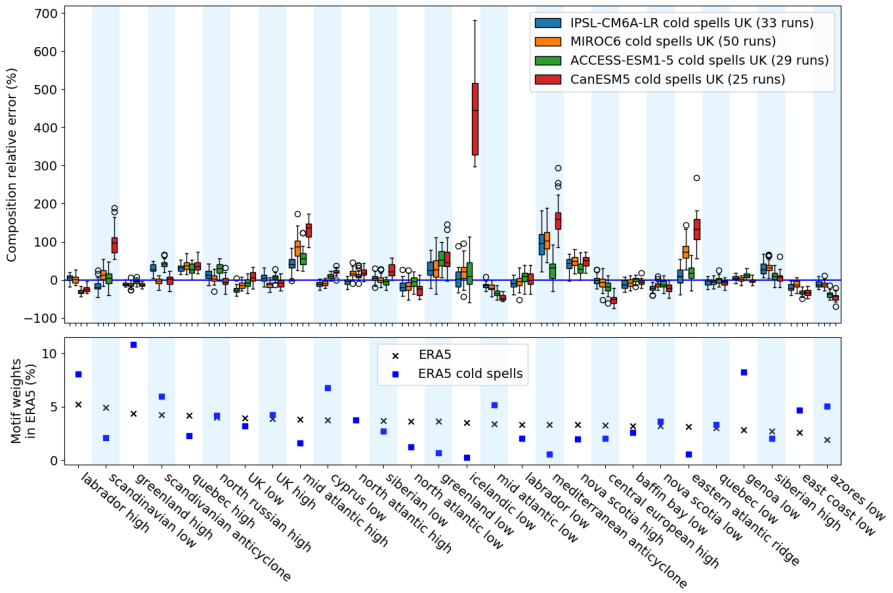


Figure A.9 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Cold Spells** occurring in the UK. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Cold Spells** and in the general case.

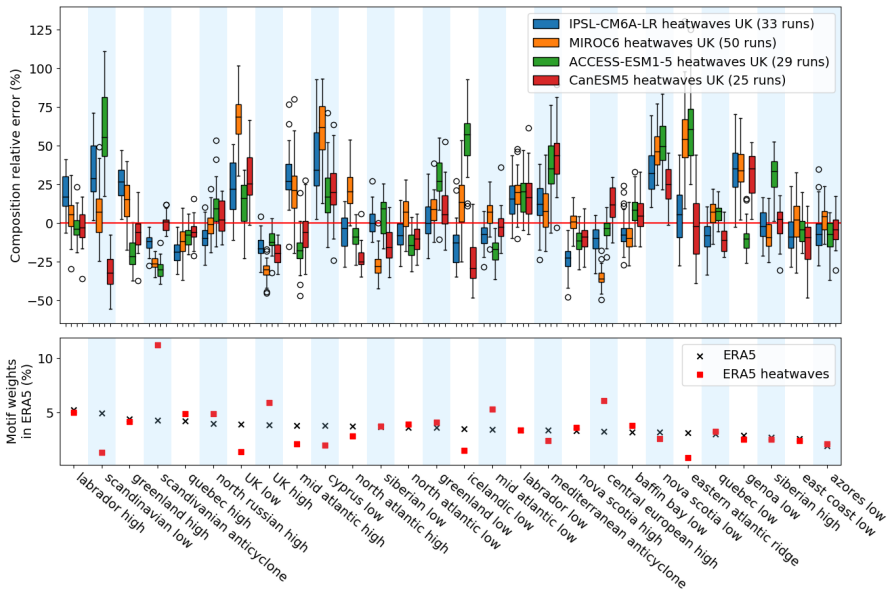


Figure A.10 – Top: Relative error on average motif weight between models and ERA5 reanalysis in the case of **Heatwaves** occurring in the UK. Bottom: average motif weight in the synoptic configuration of ERA5 fields, for **Heatwaves** and in the general case.