Studying extreme climate events using machine learning and rare event algorithms

Alessandro Lovo

THÈSE
en vue de l'obtention du grade de Docteur, délivré par
l'ÉCOLE NORMALE SUPÉRIEURE DE LYON

École Doctorale N°52
Physique et astrophysique (PHAST)

Discipline : Physique

Soutenue publiquement le 15/10/2024, par :

Alessandro Lovo

Studying extreme climate events using machine learning and rare event algorithms

Étude d'événements climatiques extrêmes avec des approches d'apprentisage automatique et des algorithmes d'événements rares

---

Devant le jury composé de :

FABLET, Ronan	Professeur des universités IMT Atlantique	Rapporteur
HASSANZADEH, Pedram	Professeur The University of Chicago	Rapporteur
BLANCHET, Juliette	Directrice de recherche Université Grenoble Alpes	Examinatrice
MONTELEONI, Claire	Directrice de recherche INRIA Paris & University of Colorado	Examinatrice
SIMONNET, Eric	Chargé de recherche Université Côte d'Azur	Examinateur
HERBERT, Corentin	Chargé de recherche-HDR École normale supérieure de Lyon	Examinateur
BOUCHET, Freddy	Directeur de recherche École normale supérieure de Lyon	Directeur de thèse

Studying extreme climate events using machine learning and rare event algorithms

Alessandro Lovo

Defended: October 15, 2024
Last compiled: January 15, 2025

And thence we emerged to see the stars

Free translation of the last line of Dante's Inferno

Acknowledgements

Thanks

First and foremost, I would like to thank my supervisor, Freddy Bouchet, for his guidance, especially in identifying which of my ideas had potential and which were dead ends. A heartfelt thank goes to my *co-encadrant* Corentin Herbert, for his insight, kindness, availability and patience in discussing topics from the most abstract levels to the nitty-gritty details.

Special thanks go to my reviewers, Pedram Hassanzadeh and Ronan Fablet, for the patience of reading this manuscript and for their valuable comments. I am also grateful to them and the other members of the jury for their appreciation of my work and for their many thought-provoking questions and comments at the time of my defense. Those comments and questions are very interesting ideas for future research.

I would also like to thank my supervisors ad-interim: Peter Ditlevsen and Johannes Lohmann during my time in Copenhagen and Henk Dijkstra in Utrecht. My gratitude extends to the teams in Copenhagen and Utrecht, you have all been very welcoming and, though I had to change apartment many times, you all made me feel at home. Be blessed *hyggelighed/gezelligheid*.

I would also like to thank the staff of the computing centers I interacted with: Emmanuel Quemener in Lyon, Roman Nuterman in Copenhagen and Michael Kliphuis in Utrecht. You helped me to familiarize myself with the different high performance computing centers and clusters and were patient the many times I asked for help after being stuck on obscure bugs.

I want to express my gratitude towards my collaborators: Bastien Cozian, Clément Le Priol and George Miloshevich in Lyon, and Alfred Hansen in Utrecht. A special thank goes to Amaury Lancelin, writing a paper together with you was a very pleasant experience, and I really enjoyed our many discussions. Continue like this! An even more special thank goes to Valeria Mascolo. Not only did we coauthor a paper - and what a paper! - but our work in Lyon has been closely entwined. Of course, my gratitude goes beyond work collaboration. During this three years you have been a true friend, someone I could totally rely on and not feel afraid to ask for comfort. I wish you all the best for your future.

My thanks extend to Clara Hummel, Ignacio Del Amo, Jade Ajagun-Brauns, Kolja Kypke, Sacha Sinet and Valérian Jacques-Dumas, for the wonderful time we shared during our secondment periods, and in general to all the other Critical Earth fellows. It really did feel like a *fellowship*, in the Tolkieninan sense.

I would also like to thank Jules Guioth, Alessandro Sozza, Brivaël Collin and Louis Saddier: it was nice to share the office in Lyon with you.

A well deserved thanks goes to my family, for the love and steadfast support they showed me during my whole life and especially these last three years. A particular mention to my sister Nicoletta, my father Roberto, my aunt Laura and my stepmother Claudia for convincing the stubborn ox that I am to start a psychological journey. I don't think I

could have handled the stress of redacting this manuscript without it.

And finally, I would like to thank YouTuber *Artifexian* for his insane multi-year-spanning physically accurate worldbuilding series. With the excuse of creating a fictional world, you deliver surprising pearls of knowledge on the Earth sciences, from geology to climate and biology. Your videos were the spark that inspired me to work in the geoscience community.

Funding

My work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 956170 (Critical Earth). Computational resources were provided by the *Centre Blaise Pascal* in Lyon, and the High Performance Computing Centers *Pôle Scientifique de Modélisation Numérique (PSMN)* in Lyon, *HPC* in Copenhagen and *Lorenz* in Utrecht.

Abstract

Extreme events in weather and climate are among the most detrimental effects of the Climate Crisis. Extreme heatwaves, for instance, have been responsible for significant excess mortality. Moreover, as the climate warms, there is a risk, still unsatisfactorily quantified, that extreme events could make us cross Tipping Points in the Earth System, leading to abrupt changes in the current climate. It is thus of paramount importance to improve our understanding of such extreme events and our ability to forecast them. However, by their nature, extreme events are rare, so there are very few instances in observational data and simulating them with state-of-the-art climate models can be very expensive. To counter this lack of data issue, rare event algorithms can be applied to significantly improve the efficiency in simulating extreme events. Such algorithms need an estimate of the probability of occurrence of the event conditioned on the state of the system, and this is exactly what the prediction task provides.

This thesis develops in two main directions. The first is to use machine learning (ML) to estimate from long climate model simulations the probabilities of extreme heatwaves over France. In particular, through a hierarchy of increasingly complex ML models, the trade-offs between amount of data, performance and interpretability of the predictions are investigated. The second is to apply a rare event algorithm to the study of the abrupt collapse of the Atlantic meridional overturning circulation (AMOC). Finally, these two pieces are put together to investigate how coupling machine learning and rare event algorithms may improve our ability to sample and predict extreme events.

Resumé

Les phénomènes météorologiques et climatiques extrêmes sont parmi les effets les plus néfastes de la crise climatique, causant des surmortalités importantes, comme lors des vagues de chaleur extrême. Le réchauffement climatique augmente également le risque, encore mal quantifié, de franchir des points de basculement entraînant des changements brusques du climat. Il est donc crucial d'améliorer notre compréhension et notre capacité de prévision de ces événements extrêmes. Cependant, en raison de leur rareté, les observations sont limitées et les simulations climatiques de pointe sont coûteuses. Pour pallier ce manque de données, des algorithmes d'événements rares peuvent améliorer l'efficacité des simulations. Ces algorithmes nécessitent une estimation de la probabilité d'occurrence de l'événement en fonction de l'état du système, qui est exactement ce que la prédiction fournit. Cette thèse explore deux directions principales. La première utilise l'apprentissage automatique (ML) pour estimer, à partir de données issues de simulations climatiques, les probabilités de vagues de chaleur extrême en France. En examinant une hiérarchie de modèles ML de complexité croissante, elle étudie les compromis entre la quantité de données, la performance et l'interprétabilité des prédictions. La deuxième direction applique un algorithme d'événements rares à l'étude de l'effondrement brutal de la circulation méridienne de retournement de l'Atlantique (AMOC). Enfin, la thèse combine ces deux approches pour étudier comment le couplage de l'apprentissage automatique et des algorithmes des événements rares peut améliorer notre capacité à échantillonner et à prédire les événements extrêmes.

Contents

Chapter 1

Introduction

Extreme weather and climate events are among the most detrimental effects of the Climate Crisis [Seneviratne et al., 2021]. Floods, droughts, storms and heatwaves are causing significant damage to human society, both in terms of human lives and economic losses [IPCC, 2021b]. For instance, the 2003 European heatwave caused the death of over 15 thousand people [Fouillet et al., 2006], while in 2010 massive floods over Pakistan destroyed 1.8 million homes and caused 40 billion US dollars in damage [Lau and Kim, 2012]. Impacts extend to whole ecosystems, with excess heat and drought increasing the likelihood of wildfires [Karoly, 2009; Porfiriev, 2014], and in general causing higher stress on plants and animals [Stevens-Rumann et al., 2018]. A fitting example in this case is that of the series of 'mega-fires' that hit Australia between September 2019 and March 2020, collectively burning over 7 million hectares of forest [Collins et al., 2021].

The events described above are all temporary, but as the changing climate destabilizes many components of the Earth system, such temporary fluctuations may act as triggers and gateways for substantial regime shifts, with important long-term consequences [Goosse et al., 2002; Drijfhout et al., 2013].

Consequently, there is significant interest in enhancing our understanding of weather and climate extremes, and the first research question is:

1. What is the probability of these extreme events? Namely, how rare are they?

Indeed, linking the amplitude of an extreme event with its probability is possibly one of the most relevant matters for policymakers, and answering this question both in scenarios with and without climate change is a fundamental aspect of attribution studies [Shepherd, 2016]. However, in this work we will not treat climate change explicitly, and though question 1 will be partially addressed in chapter 5, it will mostly sit in the background, acting as motivation. Indeed, the tools we develop throughout all of this thesis can ultimately be used to answer this question (see section 1.5 and chapters 5 and 6 for more details).

If question 1 ponders extreme events somewhat from afar, equally crucial is to look at them more up close, which means asking:

2. How and how accurately can we forecast these events?

3. What are the sources of predictability?

From a purely academic perspective, understanding the dynamics of extreme events is an interesting problem per se, but by combining it with the prediction task, it becomes automatically more quantitative and much more relevant for society. Answering questions 2 and 3 will be the main goal of this thesis. In particular, we will focus on the case study of extreme heatwaves over France, but we argue that the methods developed can be easily transferred to heatwaves over other regions of the globe and to many other types of extreme events.

Now, several techniques have already been employed in the literature to study extreme events (see section 1.2 for more details), and in this work we will use machine learning, which is one of the most promising for the forecast task [Miloshevich et al., 2023a]. The natural follow-up question is then:

4. How can we use machine learning to effectively forecast extreme events, while at the same time getting insight into the dynamics that lead to them?

In chapter 2 we will show that it is relatively straightforward to use neural networks to obtain a skillful prediction, but, on the other hand, it is much harder to understand what led the network to its decisions (see section 1.3.1 and chapter 4). To answer question 4, we will then use state-of-the-art explainability tools as well as develop new ones (chapters 2 and 4), but also build new neural network architectures which are explicitly designed to provide interpretable predictions (chapters 2 to 4 and 6).

Finally, the main technical obstacle that hinders research on extreme events is that of the lack of data. Indeed, as extreme events are also rare, we have very few instances in observational records and simulating them with climate models is expensive (see sections 1.4 and 1.5). This issue becomes even more relevant when we want to use machine learning techniques, as they notoriously require a lot of data (see section 1.3.2). Thus, the final big question that will be addressed in this thesis is:

5. How can we cope with the issue of lack of data?

One direction is to simplify as much as possible the architecture of the neural networks used, and in chapters 3 and 4 we will show that indeed this can be a very effective strategy. The other option is to employ rare event algorithms, which allow us to use climate models efficiently to get many samples of our extreme event of interest, without the need for unfeasibly long control runs. We will first test this second option in chapter 5, where we study extreme weakenings of the Atlantic meridional overturning circulation (AMOC) in an intermediate complexity ocean model. Then, in chapter 6, we will lay the theoretical foundations for building a synergy between machine learning and rare event algorithms (see section 1.5 and particularly section 1.5.5 for more details), and test them on a toy model.

In the following of this chapter I provide a (relatively quick) overview of the state of the art concerning rare events and machine learning in the climate community, putting

into perspective the work done in my three years of PhD and presented in this manuscript. In particular, section 1.1 will give an overview of extreme events in weather and climate as well as a short summary of the current physical understanding of the case studies of this thesis, namely heatwaves and, to a minor extent, tipping points. Section 1.2 will give an overview of the statistical objects we are interested in when dealing with extreme events, as well as the main ways to compute them. In section 1.3 we will dive a bit deeper in the field of machine learning, with a particular focus on the matter of interpretability. Then, following the idea that, to tackle the lack of data issue, we need to run climate models together with rare event algorithms, we will provide a brief overview of the possible ways to model weather and climate (section 1.4) and of the main algorithms that can be used to improve sampling of rare events (section 1.5). Finally, section 1.6 will give a more detailed summary of the contribution of this thesis.

1.1 Extreme events in weather and climate

When we say *extreme*, we are referring to an event for which a particular observable reaches uncommonly high values. Consequently, extreme events are *rare* in nature, and the more extreme an event is, the less likely it is to occur. Now, not all rare events are also extreme, for instance observing exactly the mean seasonal temperature every day for the whole summer is a very rare event, but it is not particularly interesting to study. However, rare events may be dangerous not only because of their immediate impact, but also because they could destabilize current equilibria of the climate system, which could lead to long term, irreversible changes, i.e. tipping points [Goosse et al., 2002; Drijfhout et al., 2013]. Thus, in the following of this work, we will focus on events which have a significant immediate or delayed effect, and we will use the terms *extreme event* and *rare event* more or less interchangeably, with the former putting the accent on the event amplitude and the latter on its probability.

Among the different types of extreme events, heatwaves are the ones most easily attributable to climate change [Seneviratne et al., 2021]. Indeed, consistently with the shift in the average global temperatures, events with a fixed probability of occurrence are expected to see their amplitude increase linearly with respect to degrees of global warming, while heatwaves with a fixed, sufficiently high amplitude will see their frequency increase exponentially [Seneviratne et al., 2012]. Since heatwaves will be a major case study in this thesis, let us briefly discuss the current understanding of the physics that leads to them.

1.1.1 The physics of extreme heatwaves

Providing a precise definition of heatwave is a highly debated matter in the literature, to the point that almost every study uses a slightly different one [Perkins, 2015]. We will provide the precise definition used in this thesis in chapters 2 and 3, but for now, since the discussions that follow are mostly qualitative, we can stay rather vague and say that heatwaves are “prolonged periods of abnormally hot weather relative to the expected

conditions at a given time and place" [Barriopedro et al., 2023].

Now, according to the review of Perkins [2015], there are three major drivers of extreme heatwaves, where with *driver* we mean components of the climate that significantly affect the chance or the amplitude of a heatwave.

The first is the presence of a persistent anticyclone over the region of interest. This usually happens as a blocking event [Charney and DeVore, 1979], where a high pressure system in the meanders of the jet stream gets cut off from the normal eastward flow and sits over the same location for an anomalously long time, from several days to a few weeks [Egger, 1978]. The stationary anticyclone extends vertically up to the 250 hPa geopotential height level [Meehl and Tebaldi, 2004], and the sinking air heats up adiabatically. Also, the high pressure keeps the sky clear of clouds, increasing solar radiation input and thus diabatic heating [Vallis, 2017; Barriopedro et al., 2023].

This link with anticyclones means that the areas affected by extreme heatwaves are generally at the size of the synoptic scale, or roughly 1000 km [Bouchama, 2004; Barriopedro et al., 2011]. Moreover, atmospheric circulation connects the weather across the globe and is responsible for important teleconnection patterns [Miloshevich et al., 2023c], which can cause extreme events to happen at the same time in different places [Lau and Kim, 2012; Kornhuber et al., 2020], increasing even further the stress on society.

Another potentially relevant factor related to atmospheric dynamics is the horizontal advection of warm air. However, while there is a consensus on the importance of anticyclones [Barriopedro et al., 2023; Perkins, 2015], the role of horizontal advection is still debated, and may be relevant only in specific geographical regions. For instance, Schumacher et al. [2022] finds that air previously heated over the North Pacific was a determining factor for the 2021 Canadian heatwave, but, on the contrary, Zschenderlein et al. [2019] shows that, over Europe, the hot air masses responsible for heatwaves didn't experience significant heating before being already over the region of the heatwave.

The second important driver is the soil moisture of the interested area. When the soil is wet, the upward heat flux at the surface is dominated by latent heat, with evaporation cooling the surface [Alexander, 2011]. On the other hand, when the soil is dry, the incoming solar radiation is absorbed through sensible heat, leading to increased surface temperatures and consequently higher near-surface air temperatures (fig. 1.1). Moreover, especially over regions of low orography [Stéfanon et al., 2014], the lack of evaporation inhibits cloud formation, which acts as a reinforcing feedback, preventing precipitation and drying the soil even further [Miralles et al., 2012]. This mechanism is in principle valid all over the globe, but it is especially important at the mid-latitudes, in the so-called *transitional zone* between wet and dry climates, where the water content of the soil fluctuates the most [Seneviratne et al., 2006]. In the preindustrial and present-day climate, the Mediterranean region is such a transitional zone, but as the climate warms, such zone moves northward, encompassing central and Eastern Europe [Seneviratne et al., 2006]. And indeed, low soil moisture was a key factor for the European heatwave of 2003 [Fischer et al., 2007] and the Russian one of 2010 [Hauser et al., 2016].

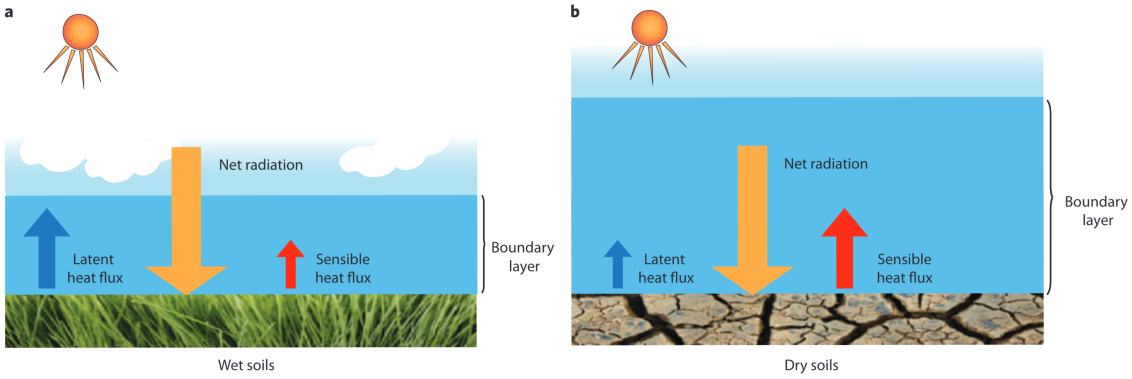


Figure 1.1: Schematics of the effect of soil moisture on the radiation budget at the surface. Figure taken from [Alexander, 2011]

If the characteristic timescale of the movement of cyclones and anticyclones is of the order of days and weeks, soil moisture evolves much slower, on the seasonal to annual timescale, so that precipitation levels in winter and spring are a determinant factor for the chance of extreme heatwaves in summer [Durre et al., 2000; Quesada et al., 2012; Mascolo et al., 2024b; Vautard et al., 2007].

This is particularly bad news for Europe. As the climate changes due to anthropogenic greenhouse gas emissions, the Mediterranean area is getting and will continue to get dryer [IPCC, 2021a], especially once, by the end of the century, alpine glaciers will mostly disappear [Haeberli et al., 2019; Beniston, 2012]. Together with its susceptibility to blocking events as the jet stream exits the Atlantic Ocean [Woollings, 2010], this will turn Europe more and more into a hotspot for extreme heatwaves [Rousi et al., 2022; Tripathy et al., 2023].

The third and final driver is the ocean, through coupled atmospheric-ocean variability phenomena on the interannual-to-decadal timescale. These include El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), and the Atlantic Multidecadal Variability (AMV) [Zhou and Wu, 2016; Mascolo et al., 2024b]. The effect of these drivers can vary from region to region and is in general less understood with respect to the previous two [Perkins, 2015].

To summarize, heatwaves are influenced by phenomena that involve different climate components, i.e. atmosphere, land and ocean, and at different spatial and temporal scales. The tree drivers presented above are the ones that are most generally relevant for heatwaves at mid-latitudes, but there is still a lot of research to be done to identify the factors that are specific for each region of the globe [Perkins, 2015]. Moreover, although there is a general of understanding of *what* the different drivers are, how much each of them contributes to the chances of heatwaves and how they interact is still unsatisfactorily quantified [Perkins, 2015]. In this thesis, especially in chapters 3 and 4, we will develop tools that can be used to answer the first part of this question.

1.1.2 Tipping points

So far the extreme events we discussed are (increasingly less) rare temporary states that the climate system visits. Once the event is over, the climate goes back to its previous state with little long term consequences. However, as we continue to pump greenhouse gases into the atmosphere, we may find ourselves in the condition where after a rare fluctuation the climate doesn't come back to normal: we have crossed a tipping point [Goosse et al., 2002; Drijfhout et al., 2013].

Tipping points can be broadly described as "critical thresholds at which a tiny perturbation can qualitatively alter the state or development of a system" [Lenton et al., 2008]. As for heatwaves, the recent popularity of the subject of climate tipping points has sprouted many definitions [van Nes et al., 2016; Russill, 2015], to the point that some behaviors may be either considered as a tipping point or not depending on the definition. For instance, it is possible to achieve a higher response to small perturbations by a simple increase in the sensitivity of the system, but the main body of literature associates tipping points with the notion of bi- or multi-stability, and thus the idea of transitions between different attractors [e.g. Lohmann et al., 2024]. This way, we add to sudden change the notion of irreversibility, where, even if the forcing is removed, the system doesn't recover [Lenton and Ciscar, 2013]. In this work, we agree with this view, and later in this section we will provide a more mathematically precise definition for the different types of tipping we are interested in.

According to this view, multi-stability is the enabling factor for tipping points, and, itself, it is the result of positive feedbacks in the climate system. Components of the climate system that exhibit such feedbacks are called *tipping elements*, and studying these feedbacks is the main way to gain a physical understanding of where the tipping point may be.

For instance, for the tipping element of the Atlantic meridional overturning circulation (AMOC), the main positive feedback is due to the advection of highly saline water from the tropics into the North Atlantic [Stommel, 1961]. As this water cools it becomes denser and sinks, initiating the overturning cycle that sustains the whole circulation and enables the salty water to reach the Arctic in the first place. Increased heat and freshwater input in the Arctic due to climate change have then the potential to dilute this salty water and hinder its sinking. This would, in turns, weaken the circulation and consequently advect less salt into the Arctic, which makes the circulation even weaker. If the external freshwater forcing is too high, it could lead to a complete collapse of the AMOC [Rahmstorf, 2002] (see chapter 5 of this thesis for more details).

Another example is that of ice sheets. For example, in Greenland the main feedback is that between the elevation of the glacier and the temperature at which it is exposed [Boers and Rypdal, 2021]. A bigger, taller glacier will be in contact with colder air and thus will tend to grow, while a smaller glacier will 'feel' higher temperatures and thus will tend to shrink. As the Arctic warms, the ice sheets recede and at the same time the snow line moves upward, which means larger and larger portions of the glacier will be exposed

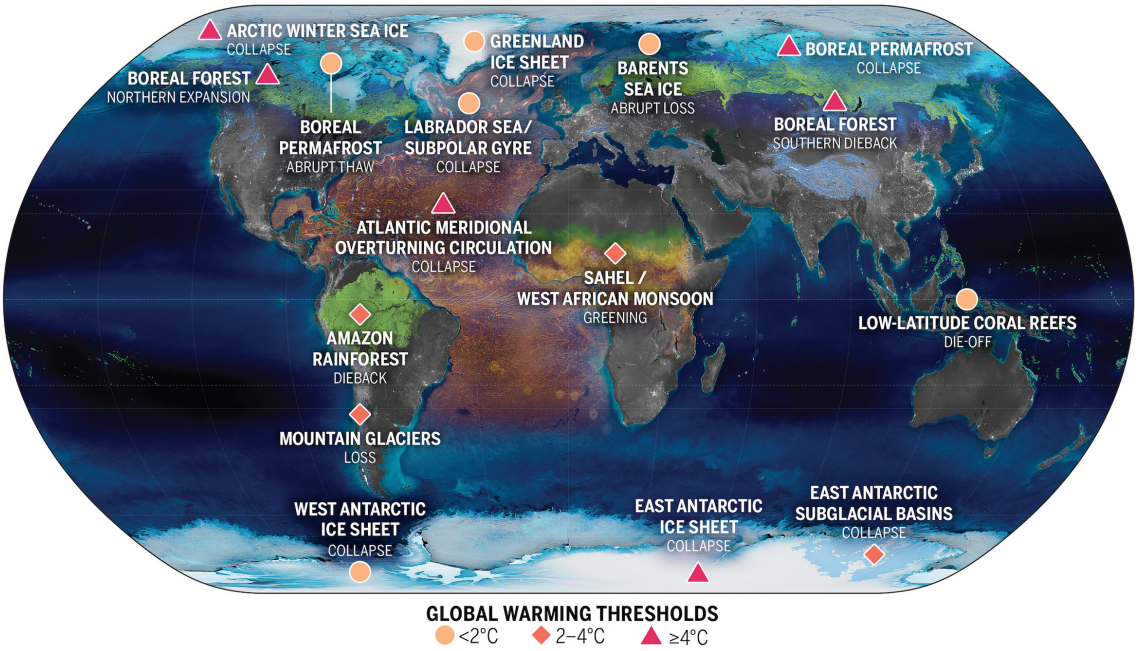


Figure 1.2: Climate tipping elements and their estimated critical thresholds in terms of degrees of warming above preindustrial. Figure taken from [Armstrong McKay et al., 2022]

to higher temperatures and as well receive rain rather than snow, which further increases melting.

Finally, a third example of tipping element is the Amazon rain forest, where a significant portion of precipitation is recycled through evapotranspiration from the forest itself [Cox et al., 2004]. Though climate change will impact the rainforest [Malhi et al., 2009], this time the main threat comes from deforestation [Boers et al., 2017]. Indeed, as more trees are removed, less water is recycled into the atmosphere, with the result of a longer dry season and a potential regime shift from rain forest to savanna.

The tipping mechanisms we have described so far imply change in a forcing parameter that compromises the stability of the current attractor of the system, causing it to shift to a new attractor. This is what is called bifurcation- or B-tipping, and it relies on the concept of saddle-node bifurcations in dynamical system theory [Guckenheimer and Holmes, 2013]. Due to the relative simplicity of the mathematics behind it, this type of tipping is the most common in the literature, to the point where some authors consider it the only true definition of tipping point. Moreover, it has the advantage of providing a framework for the detection of Early Warning Signal (EWS), a field which experienced a recent explosion in the literature [Ditlevsen and Ditlevsen, 2023; Boulton et al., 2014; van Westen et al., 2024b; Boers and Rypdal, 2021].

However, this method comes with strong assumptions, like that of the system having only two timescales: a very fast component that acts as noise and a very slow one that tracks the equilibrium manifold [Ritchie et al., 2021]. This can be questionable when applied to the climate system, due to its strongly multiscale nature [Arto et al., 2014], that

doesn't allow such a clean separation. This assumption is relaxed when one takes into the account the speed at which the forcing is increased, giving rise to the field of rate- or R-tipping. In this case, depending on the interplay between the rate of forcing and the typical timescales of the system, one can have effects like causing tipping before the equilibrium threshold is reached [e.g. O'Keeffe and Wieczorek, 2020], or, on the contrary, allowing for overshoots beyond the equilibrium threshold if the forcing is removed fast enough [e.g. Bochow et al., 2023; Ritchie et al., 2021].

And finally, the third possibility is noise- or N-tipping, where there is no change in the forcing parameters, but rather a rare fluctuation in the internal variability of the system brings it to a new attractor [e.g. Castellana et al., 2019; Cini et al., 2024]. Of the three, this is the most concerning, as it requires good models of the noise processes involved, and it can happen with little warning, causing a tipping point in a regime still deemed safe according to equilibrium analysis.

Through the lens of N-tipping we then gain a new perspective on rare events. Indeed, they are important not only because of their immediate impact, but also because they could act as triggers of transitions between different attractors in the Climate System. However, this type of tipping is also the least studied, as sampling such extreme fluctuations can be prohibitive in state-of-the-art climate models. With the recent theoretical development of rare event algorithms (see section 1.5) it is becoming possible to overcome this issue, but the number of publications is still small [Castellana et al., 2019; Cini et al., 2024]. In chapter 5 of this thesis, we will contribute to filling this gap, applying a rare event algorithm to an intermediate complexity ocean model for the study of the noise-induced collapse of the AMOC.

In this section we often used climate change as a motivation for the study of extreme events. However, in this thesis we focus mainly on methodological developments, that are treated in a stationary climate framework. Thus, we will not discuss changes in extreme event properties, with the argument that a first way to assess climate change is by studying different stationary climates [IPCC, 2021b]. Moreover, a better understanding of the processes in themselves will be extremely valuable for the assessment of how the Climate Crisis is affecting or may affect them in the future.

1.2 Studying rare events

To efficiently study rare events, we need to start by making the question more precise, namely: "What property of a particular rare event do we wish to estimate?". If we are interested in the probability of the rare event or in the average dynamics that leads to them, we are approaching the problem *a-posteriori*. Namely, we are interested in properties conditioned on the fact that the rare event happened. If instead we are interested in predicting the occurrence of an extreme event, then we need *a-priori* statistics, where we condition on the current state of the world and estimate the probability that the event will

happen. These two categories of statistics are fundamentally different (see chapter 3), and should not be confused. In this section we briefly describe them and provide and overview of the ways to compute them.

1.2.1 A-posteriori statistics

Return times

The most common statistics which is associated with rare events is that of the *return time* τ_r, which is the inverse of the probability of the event and quantifies how often we expect it to happen. This is the key quantity that, for instance, insurance companies are after when planning their policies, as it measures the risk posed by the rare event. Often, an extreme event is defined as a scalar observable A exceeding a threshold a, which is called the *return level*. It is then natural to monitor the return time as the threshold a is varied, producing what are called *return time plots* [e.g. Ragone et al., 2018; Ragone and Bouchet, 2021], which relate the amplitude of the event with its probability. In other words, the return time plot is another way of representing the far tail of the distribution $p(A)$ of the event amplitude

$$\frac{\Delta t}{\tau_r(a)} = \mathbb{P}(A \geq a) = \int_a^{+\infty} \mathrm{d}A\, p(A), \tag{1.1}$$

where Δt is the time difference between two independent samples of A. For instance, if we are interested in summer temperature maxima, Δt will be one year.

The standard way of computing return time plots comes from Extreme Value Theory (EVT), which extracts and extrapolates the tail properties of A from a sufficiently long time series [Coles et al., 2001]. Within this framework there are the two approaches of Block Maxima (BM) and Peak Over Threshold (POT). The first one consists in chopping the time series into blocks, which have to be long enough to be considered statistically independent and identically distributed, and then computing the maximum value of A in each block. Then, one can use this data to fit a Generalized Extreme Value (GEV) distribution, which allows extrapolating beyond the most extreme observed values. The second approach, instead, looks directly at the values of A that are above the threshold a, and, if a is large enough, then one can make the hypothesis that the exceedance of the threshold is a Poisson process, which immediately gives access to the return time. Subsequently, one can fit the Generalized Pareto (GP) distribution to the threshold exceedances to gain additional information on the tail of A beyond a [Pickands, 1975].

Applying these techniques is, at least in principle, relatively straightforward, and is thus often the first go-to when one approaches the study of rare events. However, when one goes into the details, things can get tricky. Particularly in the climate community, observational time series are very rarely *sufficiently long*, and due to the multiscale nature of the climate, and recent climate change, it is quite hard to satisfy the requirement of stationary data. Moreover, the multiscale nature of the climate has the second consequence of imbuing time series with long-term correlations, which violate the hypothesis of statistical independence.

For these reasons, often EVT estimations come with huge associated error bars, with uncertainties increasing rapidly as we move deeper into the tail of A [Coles et al., 2001; Le Priol et al., 2024]. Despite this big margin of error, such extrapolations might still fail spectacularly when dealing with very extreme events. One such example is the Canadian heatwave of 2021, which was way outside the EVT confidence intervals and was thus deemed impossible [Fischer et al., 2023].

Furthermore, extrapolation beyond the observed values of A gives information only about the return time. It does not sample new extreme events and thus cannot give insight about the dynamics of the event. Since the most extreme, 'record shattering', yet unsampled events are the most detrimental [Robine et al., 2008], there is interest in developing new tools to have more precise estimations as well as dynamical information. One such category of tools is rare event algorithms (REAs), which we will discuss later in section 1.5.

Composite maps

Return times focus simply on the tail properties of A, but, especially as scientists, we are interested not only in computing the probability of an extreme event, but also in gaining an understanding of the dynamics that led to it. To answer this question, the simplest, and thus most commonly used, tool is that of *composite maps* [e.g. Grotjahn and Faure, 2008; Teng et al., 2013; Miloshevich et al., 2023c; Noyelle et al., 2024]. If we suppose we have a set of climate variables that we can collectively call $X \in \mathbb{R}^d$, then the composite map can be defined as the average state of X a given amount of time τ before an extreme event happened. Namely,

$$C(\tau) = \mathbb{E}(X(t-\tau)|A(t) \geq a), \tag{1.2}$$

where $\mathbb{E}$ is the expected value, and is in practice estimated as an empirical average over the data.

Again, this method is very simple and gives some preliminary information, but it is important to acknowledge its limitations as well. The first is that, contrary to the previous EVT approach to estimate probabilities, it cannot go beyond the most extreme event $A_{\max}$ in the data. In fact, in order to have a meaningful empirical average, we can only study events with a threshold a significantly below $A_{\max}$.

Besides direct computation of composite maps, there have been recent developments that allow to extrapolate beyond the observed instances. For instance, a very recent work uses spatial Extreme Value Theory methods together with machine learning to investigate European heatwaves [Koh et al., 2024]. In a similar fashion, in chapter 3 we will show that for certain types of events, extreme heatwaves included, composite maps scale nicely with respect to the threshold a, which allow us to compute them on less extreme, more plentiful events, and then infer properties on the rarest ones, effectively mitigating this fundamental drawback of composite maps.

Despite this, composite maps still capture only average relations between X and A. They don't give information about what *caused* the event, and, being averages, they don't give a physically consistent storyline. Indeed, as will be explained in detail in chapter 3, to access causal information we need to move to a-priori statistics.

1.2.2 A-priori statistics: the committor function

The main focus of this thesis is on the prediction problem. Since the work of Lorenz [1963] we know that the atmosphere is a chaotic system, where, although the governing equations are in principle deterministic, slight differences in initial conditions eventually lead to qualitatively different macroscopic states. Hence, since we can never know the *exact* set of initial conditions, any forecast of events that happen beyond the Lyapunov time, which for atmospheric phenomena such as heatwaves is of the order of a few days [Lorenz, 1963; Ragone et al., 2018; Ragone and Bouchet, 2021], will necessarily be probabilistic [Lucente, 2021].

When extreme events are involved, accurately quantifying their likelihood is crucial, and current forecast methods, especially at a sub-seasonal to seasonal scale, are not very good at it. For instance, predictions for the 2021 Canadian heatwave consistently underestimated its amplitude, from a month and a half up to just a week before the event [Domeisen et al., 2023; Lin et al., 2022].

Now, the proper tool for probabilistic forecasts is the *committor function*, a mathematical object initially introduced by Onsager [Onsager, 1938] in the framework of dynamical system theory. For a stochastic process $X(t)$ in phase space Ω, we define the first hitting time $\tau'_{\mathcal{A}}$ of set $\mathcal{A} \subset \Omega$ as

$$\tau'_{\mathcal{A}}(x) := \inf\{t : X(t) \in \mathcal{A} | X(0) = x\}. \tag{1.3}$$

Then, given two disjoint sets $\mathcal{A}$ and $\mathcal{B}$, the committor function q is defined as the probability of hitting $\mathcal{A}$ before $\mathcal{B}$.

$$q(x) = \mathbb{P}(\tau'_{\mathcal{A}}(x) < \tau'_{\mathcal{B}}(x)). \tag{1.4}$$

If $\mathcal{A}$ and $\mathcal{B}$ are two attractors of the system, the committor is a key object for identifying the transition paths between the two attractors and the overall the transition probability [Weinan et al., 2005]. This makes the committor a particularly useful object in the field of molecular dynamics [Li and Ma, 2014; Thiede et al., 2019] and the study of protein folding [Belkacemi et al., 2022; Ren et al., 2005], where these rare transitions between attractors are fundamental for the successful realization of complex chemical reactions.

In the climate community, there is interest in using the definition of eq. (1.4) for the study of tipping points [Jacques-Dumas et al., 2024; Finkel et al., 2021], where two attractors are clearly defined. On the other hand, for extreme fluctuations like heatwaves, there is only a single attractor $\mathcal{A}$, and $\mathcal{B}$ is an atypical state that we want to reach within a specified time frame τ [Lucente et al., 2019]. One possibility is then to adapt eq. (1.4) by making the sets $\mathcal{A}$ and $\mathcal{B}$ time dependent. Equivalently, if we define set $\mathcal{B}$ as the observable

$A(t)$ exceeding the threshold a, we can drop first hitting times, and rather express the committor simply as a conditional probability:

$$q(x) = \mathbb{P}(A(t) \geq a | X(t-\tau) = x). \tag{1.5}$$

With this definition, we can immediately see that, compared to composite maps (eq. (1.2)), the a-priori statistics of the committor function reverses the conditioning. This shows that a-priori and a-posteriori statistics are fundamentally different, but also closely linked. We will discuss thoroughly about this in chapter 3.

Now, we have shown that the committor is *the* proper object to use for prediction, but, as we will see in section 1.5, it is also necessary for running rare event algorithms efficiently [Lucente et al., 2022b], which makes the committor an even more interesting object of study. Unfortunately, committor functions are extremely hard to compute, being objects with the same dimensionality of the phase space, which for climate can be very high. In the following subsection we will give a brief overview of the possible strategies one might try.

1.2.3 Computing committor functions

The first option to compute committor functions applies when we have access to the governing equations of the dynamical system. In geophysics, this for example happens when one uses simple conceptual models of specific subsystems of the climate (see section 1.4). Then one can observe that the committor function solves the backward Fokker-Planck equation, which, if the system is simple enough, can be solved analytically or numerically. For more complex systems, solving the backward Fokker-Planck equation might still be feasible, but it requires more advanced techniques. For instance, Li et al. [2019] use a combination of importance sampling and machine learning (see section 1.5) in a system with roughly a thousand degrees of freedom.

If the dynamical equations are too complex or not accessible, but we still can run the dynamics, then the simplest option is that of direct numerical simulations. Namely, to compute the committor function at point x we initialize N trajectories at $X = x$, propagate them forward, and count the fraction that displays the rare event of interest. However, when $q(x)$ is small this method is highly inefficient, and if we want to compute the committor for many points in a high dimensional space it quickly becomes unfeasible. Fortunately, there are several approaches to optimally generate new data that will help the estimate of the committor more efficiently. These methods fall under the umbrella term of rare event algorithms (REAs), and we will discuss them in more detail in section 1.5.

Finally, if none of the above apply, we have to work directly with data. This is, for example, the case when we work with observations, reanalysis datasets or climate model output without access to the climate model itself. A first possible approach is that to use analogue Markov chains to construct from the data an effective dynamics. This is, for instance, the basic idea behind Stochastic Weather Generators (SWGs), which have many uses besides computing committor functions [e.g. Yiou, 2014; Yiou et al., 2013, 2023; Miloshevich et al., 2023b]. In our case, once we have a simpler effective dynamics, we can

try to estimate the committor directly from the Markov chain [Lucente et al., 2022a,b] or apply the Galerkin approximation of the Koopman operator [Thiede et al., 2019; Strahan et al., 2021], which expands the infinite dimensional transition (Koopman) operator over a finite number of basis functions and thus simplifies the subsequent computations.

Another option is to extract the committor directly from the raw data, and one of the most promising tools for this task is machine learning [Lucente et al., 2019, 2022a; Miloshevich et al., 2023a]. This is the direction we will take in this thesis.

1.3 Machine learning for Weather and Climate

In this thesis, we will use machine learning techniques to compute committor functions. However, machine learning applications to weather and climate problems are much broader in scope [Chantry et al., 2021; Huntingford et al., 2019; Schneider et al., 2022]. So, let us for a moment take a step back and widen our field of view on the matter.

Machine learning is probably the scientific field which has known the fastest explosion in the recent decades. Moreover, with the appearance of chatbots and AI assistants, it is already revolutionizing the world at a societal level. Indeed, the recent proliferation of big data, together with significant increases in computing power has led to a golden age for Artificial Intelligence (AI) (see for instance `https://www.forbes.com/sites/joemckendrick/2019/10/23/artificial-intelligence-enters-its-golden-age/` and `https://www.weforum.org/podcasts/radio-davos/episodes/ai-chat-gpt-haptik/`).

The key strength of machine learning is the ability to extract useful information from complex high dimensional data. One of the early examples where this ability proved particularly valuable is that of image recognition, which has sparked the whole field of computer vision, and now is, for instance, making possible the existence of self-driving cars. The most common type of data that one deals with in climate science is represented on latitude-longitude grid. This is essentially an image, where instead of RGB channels, we have the different climate variables that can act as effective 'colors'. So, using machine learning in climate sciences is particularly appealing, as we can leverage the developments that have already been made in other areas and use off-the-shelf products like Convolutional Neural Networks (CNNs) [Miloshevich et al., 2023a; Jacques-Dumas et al., 2022].

Machine learning techniques can be broadly divided into three categories. While unsupervised learning, with a particular focus on clustering [Fang et al., 2021; Toms et al., 2021], and, to a minor extent, reinforcement learning [Zhou et al., 2021; Morcego et al., 2023] have successfully been applied to weather and climate problems, the overwhelming majority of works use supervised learning [Bochenek and Ustrnul, 2022; Chantry et al., 2021; Huntingford et al., 2019; Schneider et al., 2022]. The types of applications to weather and climate problems also cover a broad range [Schneider et al., 2022; Huntingford et al., 2019], from downscaling to have accurate high resolution precipitation maps [Sachindra et al., 2018], to prediction of extreme events [Miloshevich et al., 2023a; Jacques-Dumas

et al., 2022; Lopez-Gomez et al., 2022], assessment of the best places to install new solar panels [López Gómez et al., 2020], and many others specifically for tackling climate change [Rolnick et al., 2022]. In particular, a significant portion of the literature focuses on improving numerical weather forecasts and climate modeling [Chantry et al., 2021; Lai et al., 2024], with big tech companies like Google or Amazon investing significant resources (see for instance `https://techmonitor.ai/technology/ai-and-automation/big-tech-coming-for-weather`). This is a fascinating research direction, but it is more closely related to weather and climate modeling rather than to committor functions, so we will cover it briefly in section 1.4.2.

As we already pointed out, among the many applications of machine learning in climate science, the one we are most interested in within the scope of this thesis is that of *prediction* problems. And, even restricting to this particular task, there are still many interesting options, though most works provide a deterministic answer, which is not particularly useful for our need to estimate the committor function. Fortunately, there has been recent progress in this direction, with methods that quantify the uncertainty in the prediction [Haynes et al., 2023] or use ensembles to provide a distribution of possible answers [Asadollah et al., 2022]. Even better, some works phrase the problem as a *probabilistic prediction* in the first place [Miloshevich et al., 2023a; Watson, 2022; Pan et al., 2022].

Furthermore, machine learning has been used for prediction over various timescales, from nowcasting [e.g. Ravuri et al., 2021], to short-range weather prediction [e.g. Giffard-Roisin et al., 2020], to sub-seasonal to seasonal (S2S) forecasts [e.g. Delaunay and Christensen, 2022] and up to decadal climate predictions [e.g. Ham et al., 2019]. However, applications specifically to rare events are still scarce, mainly because of the issue of lack of data [Miloshevich et al., 2023a].

To summarize, the field is rich in solutions that can be applied to the task of committor estimation, but two problems remain mostly unaddressed: whether the committor we compute is human-understandable and whether enough data is available to obtain an accurate estimate.

1.3.1 Interpretability

Emboldened by advances in computing resources, the progress in machine learning has led to more and more complex neural network architecture, where having a million or even a billion parameters is nothing special. This makes it extremely hard to decipher the decision process of the network, which behaves, *de facto*, as a *black box*. This compromises trust in the machine learning model, but most importantly acts as an obstacle to gaining an (even qualitative) understanding of the underlying processes, which, especially in the climate community, is far more valuable than simple predictive accuracy [Rudin, 2019; Barnes et al., 2022].

In fact, a similar reasoning is true for climate model projections that try to predict the various effects of climate change. The multimodel mean of our favorite observable (average global temperature, sea level rise, etc.) is only a small part of the answer. Much

of the research is devoted to giving a proper confidence interval to that bare number, to performing different experiments with different models to understand which phenomena are well captured, and which one show biases [IPCC, 2013, 2021a]. In summary, the value of interpretability is well established.

In recent decades, this is becoming the case also in the machine learning community [Murdoch et al., 2019]. Thus, there has been a vast proliferation of methods, which fall under the umbrella term of Explainable Artificial Intelligence (XAI) [Holzinger et al., 2022], that try to explain why neural networks reach a particular decision. Within this framework the vast majority of methods are *post-hoc* approaches, which use an already trained model and provide an explanation of its working either for a particular sample or for the whole dataset [Murdoch et al., 2019]. When the input data is image-like, as is often the case for climate, the fundamental explainability tool is the saliency map, which highlights the importance of each pixel, i.e. the value of each climate variable in each geographic location, for the prediction [Haar et al., 2023]. Such method and its many variants have been quite successfully applied to climate science, elucidating model prediction and sometimes leading to the discovery of new science [McGovern et al., 2019; Toms et al., 2020, 2021; Davenport and Diffenbaugh, 2021; Zhang et al., 2024].

However, there are some major drawbacks, the first being that the explanation necessarily makes a compromise between fidelity to the original machine learning model and interpretability, which can either lead to it being too simplistic or very hard to decipher [Barnes et al., 2020; Martin et al., 2022]. Moreover, sometimes different explanation methods can lead to contrasting conclusions [Mamalakis et al., 2022a], which is a major concern. Indeed, if one uses a single method, the explanation obtained may be misleading. Finally, most saliency map methods are fundamentally flawed in the sense that they show which pixels are the most important, but don't give any insight into *how* the information they contain is used by the network (see fig. 1.3).

Figure 1.3: Example of saliency map explanation for an image classifier. As the two saliency maps look very similar, there is no way to tell why the network would classify this image as a dog or a musical instrument. Image taken from [Rudin, 2019].

For these reasons there is a growing trend in, rather than trying to explain complex model prediction, use models which are interpretable in the first place, namely they are *white boxes* [Rudin, 2019; Barnes et al., 2022]. A fitting example of a white box is that of symbolic regression [Brunton et al., 2016]. In this case, the machine learning model

builds an equation describing the data, picking its terms from a library of simple algebraic operations. After training, the model prediction is perfectly transparent, as it is literally a mathematical expression. Unfortunately this particular method works only for systems that are simple enough and needs to be properly regularized, otherwise it produces equations with so many terms that they are not much better than a black box [Grundner et al., 2024].

Another possibility is to use more standard machine learning techniques, but with a focus on simplifying the architecture of the neural network, for example using decision trees [Rudin, 2019], or comparing the input data with a set of learned prototypes [Barnes et al., 2022], or devising clever ways to perform dimensionality reduction [Murdoch et al., 2019]. Work in this latter direction is one of the main contributions of this thesis, where, through the idea of optimal projection of the committor function, I develop the Intrinsically Interpretable Neural Network (IINN) architecture (chapter 2).

Despite being in its early stages in the literature, the white box approach appears very promising. Many studies have found only minimal reductions in performance compared to more complex black boxes [Rudin, 2019], while providing users with valuable insights that were previously unavailable. Moreover, white box models generally involve simpler architectures, which require fewer data to be trained properly, and thus are particularly competitive in contexts where data is scarce.

1.3.2 The issue of lack of data

Machine learning methods are notoriously data hungry. Indeed, one the key factors that enabled the success of AI chatbots like OpenAI's chatGPT is that they are trained on enormous amounts of data, of which *small* subsets are all of GitHub's public repositories and all of Wikipedia's articles [OpenAI et al., 2024].

In the climate community we are not so fortunate, as comprehensive datasets like reanalyses are of good quality only after the advent of the satellite era (1970) [Hersbach et al., 2020; Uppala et al., 2005], meaning we only have 50 years of data. Alternatively, weather models can perform re-forecasts on the period covered by reanalysis datasets, using multiple ensemble members, and thus artificially extending the amount of data available up to a few centuries [e.g. Hagedorn et al., 2008]. Concerning state-of-the-art climate models used by the Intergovernmental Panel on Climate Change (IPCC), the typical length of control runs is usually close to 1000 years [Eyring et al., 2016], although some recent projects like LongRunMIP [Rugenstein et al., 2019] are pushing for multi-millenial simulations.

In absolute terms, this might seem like a lot of data. However, when we are interested in rare events we can clearly see that it is not. For instance, if we want to study the 1% most extreme average summer temperatures, in a 1000-year-long control run we will have only 10 instances. Our machine learning models then may have seen too few of these events to represent them properly [Watson, 2022]. This has indeed been verified to be the case for extreme heatwaves, where many centuries of data are needed to achieve good predictions on not-so-rare events, and even when training on an 8000-year-long control run, there is

still margin for improvement [Miloshevich et al., 2023a].

There are essentially two ways to address this problem, which will both be considered in this thesis. The first is to use simpler machine learning models, which brings us back to the advantages of white boxes. The second is to generate more data using climate models, in conjunction with rare event algorithms to focus the computational effort on the extreme events we aim to study. We will expand on this second direction in the following sections, giving first an overview of weather and climate models (section 1.4) and then discussing the possible strategies to improve their efficiency with rare event algorithms (section 1.5).

1.4 Simulating Weather and Climate

1.4.1 The model hierarchy

Modeling the climate on Earth can be done at very different degrees of complexity. At the bottom of the hierarchy, we find conceptual models, and the simplest one is a zero-dimensional, static, planetary radiative energy balance model that can be written as a single equation.

$$4\sigma_B T^4 = S_0(1-\alpha), \tag{1.6}$$

where σ_B is the Stefan-Boltzmann constant, T is the average temperature of the planet, S_0 is the solar energy flux reaching the Earth, and α is the planetary albedo. Despite its disarming simplicity, this model identifies the strongest negative feedback (T^4) in the Earth system, which is responsible for keeping the average planetary temperature inside a very narrow range. When one considers that ice and snow have a very high albedo, this model also highlights one of the strongest *positive* feedbacks, which, together with other mechanisms, amplifies the minute variations in S_0 due to changes in orbital parameters, leading to the alternance between interglacial periods and ice ages [Ajagun-Brauns and Ditlevsen, 2023].

Slightly more complex models usually involve a few equations and focus on a particular subsystem of the climate. Examples include the two-box model for the AMOC [Stommel, 1961] or the many variants of Lotka-Volterra models [Lotka, 1910] that are used to describe the dynamics of ecosystem populations. These types of models can be extremely diverse, as they are developed through physical reasoning, identifying the important quantities and how they interact. Due to their relative simplicity, conceptual models lend themselves very easily to be discussed from an analytical point of view. For instance, this involves finding equilibrium manifolds and their stability as well as evaluating the response of the system to various types of forcing [e.g. Stommel, 1961; Ajagun-Brauns and Ditlevsen, 2023; Mehling et al., 2024].

Putting together conceptual models for all the important components of the climate (atmosphere, ocean, ice sheets, vegetation, carbon cycle, etc.) gets us to the category of Earth system Models of Intermediate Complexity (EMIC) [Claussen et al., 2002]. These models are no longer analytically solvable, but, due to their relative simplicity, they are

quite cheap to integrate numerically, which allows performing very long runs [Weber, 2010]. Together with the fact that these models include dynamical components that act at different timescales, they are particularly suited for paleoclimate studies and long-term climate projections [Weber, 2010].

When we model the climate, we are often interested in long-term averages, and the mathematical way to frame our questions is in terms of a *boundary value problem*. This is also why many conceptual models are developed with a top-down approach, where the physical understanding of the subsystem under study allows physicists to write down the key equations. On the other hand, producing a weather forecast is a very different *initial value problem*, where initial conditions are as, if not more, important than the model itself [Carrassi et al., 2018]. Also, this time the modeling approach is bottom-up, where the governing equations describe the fundamental behavior of fluid motion, thermodynamics, tracking of moisture and so on.

There are models that work like this on a regional scale, but the most common are General Circulation Models (GCMs) which aim to describe the whole circulation of the atmosphere or the ocean. Due to the non-linear nature of Navier-Stokes equations, an analytical solution is not possible, and the only option is to discretize the domain and proceed by direct numerical integration. At the time when the word 'computer' still meant a human with pen and paper, performing the task on a grid with a meaningful spatial resolution was seen as a futuristic fantasy (see fig. 1.4 and `https://www.emetsoc.org/resources/rff/`).

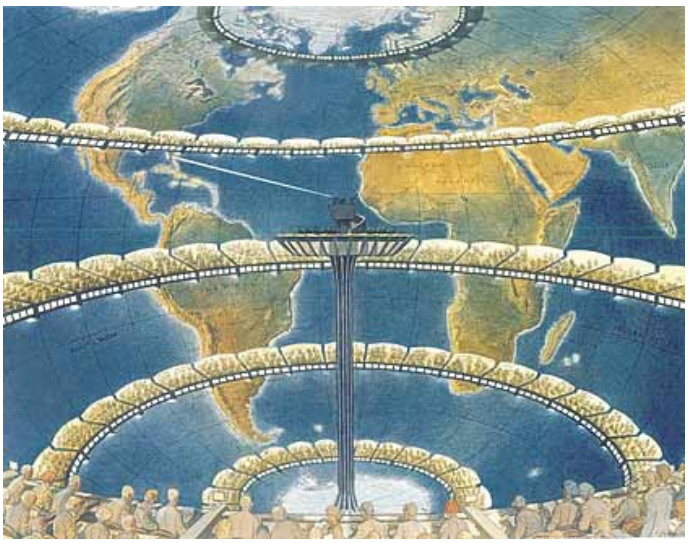


Figure 1.4: In 1922 Lewis Fry Richardson proposed to employ 64000 human calculators who will sit together in a globe-shaped theater to numerically solve the Navier-Stokes equations for the atmosphere and provide weather forecasts. Even if put in place, the computations would have taken longer than the evolution of the weather itself, making the forecast completely useless. Image taken from `https://www.historyofinformation.com/detail.php?id=59`.

However, even now that we have supercomputers, weather predictions are still a hard endeavor, and the key problem is that the atmosphere is a multiscale system. Indeed,

turbulence per se spans all scales from thousands of kilometers to a few meters, but the processes of nucleation that lead to cloud formation happen at the microscopic level. This means that we would have to use a grid that covers the whole globe, with cells the size of a few micrometers, which is absolutely not feasible. In fact, the highest resolution GCMs used for weather predictions have a horizontal grid spacing of roughly 1 km [Wehner et al., 2008], but studies that use GCMs for climate simulations need longer integration times and thus have a maximum resolution around 10 km [Demory et al., 2020; IPCC, 2013].

Consequently, all the processes that happen at a finer scale than the grid of the model are not resolved by the *dynamical core*, but are instead modeled by *sub-grid parameterizations*. These can be developed from physical understanding, from fitting data, or from models with a finer grid that can resolve the process but are not run for the full globe. Beside sub-grid processes, boundary conditions are also parameterized similarly in GCMs. For instance, if we are modeling only the atmosphere, boundary conditions involve the ocean, sea ice and ice caps, solar radiation, vegetation, aerosols and greenhouse gases, and many more. The quality of the sub-grid parameterizations and boundary conditions varies a lot across different GCMs, and often it is possible to run the same model at different degrees of complexity.

When models solve the Navier-Stokes equation for both ocean and atmosphere, and use sophisticated dynamical models for the other processes, we reach the last step of the hierarchy: the so called fully coupled models or Earth System Models (ESMs). Since many components of the Earth system evolve on slow timescales (e.g. ice sheets or the deep ocean), this class of models is mainly used in the context of climate studies. In particular, the experiments performed within the Coupled Model Intercomparison Project (CMIP) [Eyring et al., 2016] are one of the key pillars of the assessment reports of the IPCC [IPCC, 2021a].

1.4.2 Open challenges and AI solutions

Though the field of weather, and to a lesser extent climate, modeling is advancing rapidly [Bauer et al., 2015], there are still some open challenges. The first one is the sheer amount of computational resources that are needed to run state-of-the-art models, which relegates many of them to specialized centers, like the European Centre for Medium-Range Weather Forecasts (ECMWF), which processes roughly 700 PB each day. The second one is that many parameterizations require sophisticated tuning processes, which often involve a lot of heuristics. This leads to the models having biases, that are often corrected as a post-processing step of the climate model output, which is often a non-trivial process. Moreover, different models using different parameterizations lead to significant uncertainties in multimodel ensemble climate projections. For instance, the different approaches to the modeling of clouds are the main contribution to the spread of equilibrium climate sensitivity (i.e. how much would the planet warm if we double CO_2 concentrations with respect to preindustrial levels) [Vial et al., 2013; Ceppi and Nowack, 2021].

In recent years there has been a lot of enthusiasm towards AI solutions [Chantry

et al., 2021]. Indeed, machine learning techniques can be applied to post-processing of model output for bias correction, or to learn new parameterizations from data. Moreover, replacing components of weather or climate models with machine learning emulators can significantly lighten the computational burden and thus speed up computations. This idea has been very recently brought to the extreme, with the realization of fully AI weather models [Lam et al., 2023; Bi et al., 2023; Nguyen et al., 2023], that seem to rival the performance of more standard physics-based numerical simulations. The success of these models has caught the attention of weather forecast centers, with ECMWF launching last year its own Artificial Intelligence Integrated Forecasting System (AIFS) (see `https://www.ecmwf.int/en/newsletter/177/editorial/aifs-launched` and `https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system`). This direction seems very promising, but it still too early to draw clear conclusions. In particular, weather predictions by AI models are often overly smooth, and may not capture well the statistics of extreme events [Bi et al., 2023].

To summarize the most relevant conclusion for this thesis: although AI accelerators and specialized hardware are ever improving the speed of state-of-the-art weather and climate models, running them for a very long time, as is needed for the study of rare events, is still prohibitive. For this reason we necessarily have to make a compromise between model complexity (and thus fidelity to the real climate) and simulation length. For instance, with the resources available at the scale of a university, we can afford to run GCMs with relatively simple parameterizations for a few thousands years.

There is thus interest in finding ways to generate extreme events more efficiently, without the need of unfeasibly long control runs. An overview of the possible methods one might use will be the topic of the next section.

1.5 Rare event algorithms

The problem with studying rare events, is that we need only a very small part of the available data $\{X_i\}_{i=1}^N$. To put it more formally, if $p(x)$ is the stationary measure of the system, we are interested in averages of the observable $h(x)$, which is non-zero only for the extreme events of interest.

$$\mathbb{E}(h) = \int \mathrm{d}x\, h(x)p(x) = \int_{h(x)\neq 0} \mathrm{d}x\, h(x)p(x) \tag{1.7}$$

For instance, if we want the return time τ_r of a heatwave with threshold a, then $h(x)$ will be an indicator function:

$$\frac{\Delta t}{\tau_r} =: \gamma = \mathbb{P}(A \geq a) = \int \mathrm{d}x\, p(x)\mathbb{1}_{A(x)\geq a} \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{A_i \geq a} =: \hat{\gamma}, \tag{1.8}$$

where Δt is the time spacing between the different samples A_i, that we assume independent.

Now, according to the central limit theorem, the variance of the empirical estimator $\hat{\gamma}$ is

$$\mathrm{Var}(\hat{\gamma}) = \frac{1}{N}\mathrm{Var}(\mathbb{1}_{A\geq a}) = \frac{1}{N}\hat{\gamma}(1-\hat{\gamma}) \xrightarrow[\hat{\gamma}\ll 1]{} \frac{\hat{\gamma}}{N}. \tag{1.9}$$

This means that, for a sufficiently rare event, the relative error in estimating its probability, or equivalently its return time, is $1/\sqrt{\hat{\gamma}N}$, which can easily be very large.

We notice that $\hat{\gamma}N$ is the total number of points for which $h(x) \neq 0$, which suggest that we can reduce the variance of our estimator by artificially increasing this number. Indeed, the general idea of rare event algorithms is to generate a new dataset $\{\tilde{X}_j\}_{j=1}^M$, where a significant fraction of points satisfies $h(\tilde{X}_j) \neq 0$, while at the same time computing weights w_j that ensure the new empirical estimation $\breve{\gamma}$ is unbiased.

$$\frac{1}{\breve{\gamma}} = \frac{1}{M}\sum_{j=1}^{M} w_j \mathbb{1}_{\tilde{A}_j\geq a}, \quad \mathbb{E}(\breve{\gamma}) = \gamma. \tag{1.10}$$

And this time, the relative error will be $O(1/\sqrt{M})$, which is much better than before.

In the following we give an overview of the possible ways to implement this strategy.

1.5.1 Importance sampling

The first possible option is that of importance sampling (IS), which consists in drawing the new data $\{\tilde{X}_j\}_{j=1}^M$ from a new distribution $\pi(x)$, different from the stationary measure $p(x)$. Then, the weights w_j will simply be

$$w_j = \frac{p(\tilde{X}_j)}{\pi(\tilde{X}_j)}, \tag{1.11}$$

which guarantee unbiasedness [Kahn and Harris, 1951].

In order for the algorithm to be efficient, we need to choose π such that the rare event we want to study is much more common under π than under $p(x)$. However, this may be a very tricky task, as often we don't have a clear enough picture of $p(x)$ over the region of interest to properly define π.

If the system under study is low dimensional, or we have a general theoretical understanding of the processes that lead to the rare event, we can circumvent this problem by engineering a parametric form for π and iteratively optimizing its parameters, giving rise to what is known as Adaptive Importance Sampling (AIS) [Tokdar and Kass, 2010; Nemoto et al., 2016]. This method has been used successfully in many applications, such as signal processing [Bugallo et al., 2015], chemistry [Geissler and Chandler, 2000] and materials science [Paananen et al., 2021]. However, it may not be very relevant for the high dimensionality of the climate system, as the number of parameters to optimize can quickly explode [Tokdar and Kass, 2010]. Nevertheless, with some clever manipulations, a recent work shows some interesting results on an intermediate complexity ocean model [Annan and Hargreaves, 2010].

When the system is high dimensional, a more suitable variant of importance sampling is that of Sequential Importance Sampling (SIS) [Tokdar and Kass, 2010]. In this case

the idea is to shift the focus from the whole stationary measure $p(x)$ to the transition probabilities $p(x_t|x_{t-1})$, which are generally easier to access. Again, this method has limitations, as it works best for discrete systems and when the transition probabilities are known [Liu and Chen, 1998]. This is the case if one uses analogue Markov chain methods and Stochastic Weather Generators [Lucente et al., 2022b; Miloshevich et al., 2023b], which build an effective dynamics of the system where everything is known. On the other hand, if we wish to work directly with a climate model, SIS has limited applicability, as in this case the true transition probabilities are mostly unknown.

Nevertheless, a variant of SIS, known as Sequential Importance Sampling with Resampling (SISR) [Grassberger, 1997; Liu and Chen, 1998], introduces the idea of increasing the number of simulated trajectories that are moving in a promising direction, which brings us to the next class of algorithms.

1.5.2 Genealogical algorithms

Genealogical algorithms, also known as population dynamics or cloning algorithms [Bouchet et al., 2019b; Garnier and Moral, 2006; Moral and Garnier, 2005; Tailleur and Kurchan, 2007], expand on the SISR paradigm, removing the need to know the transition probabilities. This means that now it is possible to consider the underlying dynamics as a black box that takes as input an initial condition and outputs a trajectory, which makes this class of algorithms particularly appealing for climate applications.

One particular subclass is that of the Interacting Particle System (IPS) [Garnier and Moral, 2006; Moral and Garnier, 2005], where we start with an ensemble of N typical initial conditions (i.e. sampled from $p(x)$) and propagate them forward for a given amount of time τ. Then we perform a selection step, computing a pre-defined score function V for each ensemble member, killing trajectories which have a low score and cloning those that have a high score, such that at the end we still have N trajectories. Then we propagate forward for another time τ and repeat (see fig. 1.5), steadily improving the quality of the ensemble.

As usual, many variants exist, mainly differing in the choice of the score function V and the precise procedure for killing and cloning trajectories. In this thesis, we will use the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) algorithm [Giardinà et al., 2011, 2006; Tailleur and Kurchan, 2007; Wouters and Bouchet, 2016], of which we will discuss the technical details in chapter 5.

This algorithm has already shown very promising results in the climate community, especially for the study of extreme heatwaves [Ragone et al., 2018; Ragone and Bouchet, 2020, 2021], but also for tropical cyclones [Webber et al., 2019], winter precipitations [Wouters et al., 2023] and very recently for extreme arctic sea ice reduction [Sauer et al., 2024] and for the study of the collapse of the AMOC [Cini et al., 2024]. For instance, in [Ragone et al., 2018] the authors are able to sample whole-summer heatwaves over France 1000 times more efficiently than simply running the climate model on its own.

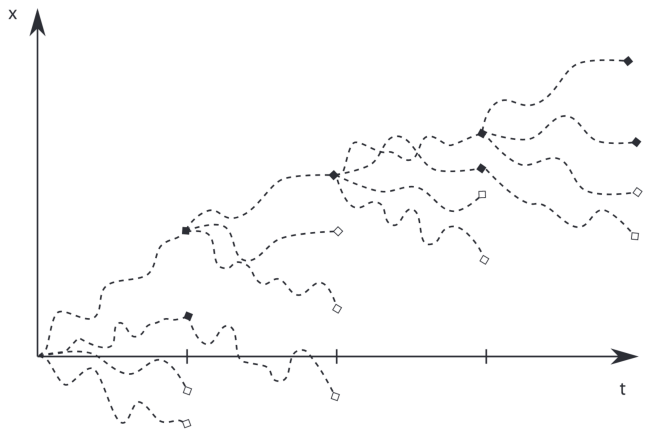


Figure 1.5: Schematic illustration of an IPS genealogical algorithm, with an ensemble of $N = 4$ trajectories. Figure taken from [Wouters and Bouchet, 2016].

1.5.3 Splitting algorithms

If genealogical algorithms can be still described within the framework of importance sampling, a different family of rare event algorithms is that of splitting algorithms. Since these algorithms will not be used in this thesis, I will give a little more detail in their workings, so that the reader can have a more comprehensive picture.

As an example, the Adaptive Multilevel Splitting (AMS) algorithm [Cérou et al., 2011; Rolland et al., 2016; Cérou et al., 2005] works by having an ensemble of N trajectories that are propagated in time until they either fall back to the typical attractor of the system $\mathcal{A}$, or they manifest a rare event by visiting $\mathcal{B}$. If the event is very rare, at the beginning none of the trajectories will reach $\mathcal{B}$. Then, using a score function Q, which increases the closer we are to $\mathcal{B}$, we can steadily improve our ensemble. Indeed, at each iteration, we compute for each trajectory i the maximum value $\mathcal{Q}_i$ of the score function Q and kill the trajectory that has the lowest value. Supposing it was trajectory $i = 1$, we replace it with a new trajectory, branching from one of the remaining $N - 1$ trajectories at the point where the score function had value $\mathcal{Q}_1$ (see fig. 1.6). This way, we are guaranteed to steadily move our ensemble closer to $\mathcal{B}$. If we stop the algorithm when, after R iterations, all trajectories have reached $\mathcal{B}$, then the transition probability is estimated as $\hat{\gamma} = \left(1 - \frac{1}{N}\right)^R$.

One of the main problems with this algorithm is that it may take a long time for trajectories to relax back to $\mathcal{A}$, especially in the last iterations of the algorithm [Cérou et al., 2019]. For this reason a variant of the algorithm has been developed, namely the Trajectory Adaptive Multilevel Splitting (TAMS) [Lestang et al., 2018], which removes set $\mathcal{A}$ altogether and rather stops the integration either when the trajectory hits $\mathcal{B}$ or after a maximum amount of time T.

This latter algorithm has also shown some interesting applications in climate, especially for the study of AMOC tipping [Jacques-Dumas et al., 2024; Castellana et al., 2019].

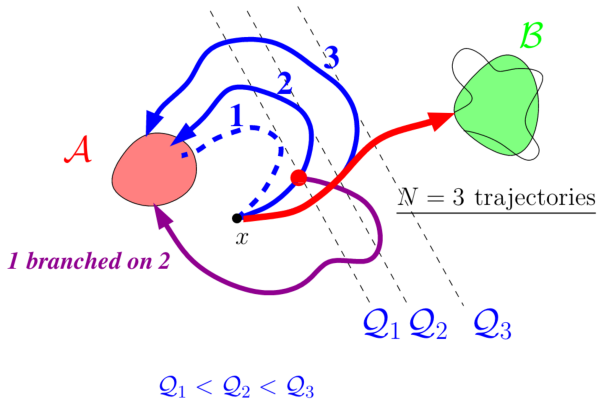


Figure 1.6: Sketch of the workings of one iteration of the AMS algorithm with an ensemble of 3 trajectories and score function $\mathcal{Q}$. Figure taken from [Rolland et al., 2016].

1.5.4 Optimal score functions

As we have seen, all rare event algorithms rely in some way or another on a score function, which is used to push the system in the right direction. Though the results are proven to be asymptotically unbiased for *any* score function [Tokdar and Kass, 2010; Giardinà et al., 2011; Cérou et al., 2011], in practice a bad choice means that convergence will be terribly slow, to the point that there is no gain with respect to simply running the climate model as is [Lucente et al., 2022b].

If we think about it, we would like, at the moment of selecting trajectories, to pick the ones that are more likely to lead to the rare event. It should then come as no surprise that optimal score function for the AMS algorithm is exactly the committor function [Cérou et al., 2019] and for the GKLT one is a combination of the committor and the transition probabilities [Chraibi et al., 2020].

So now we face a conundrum: if we trace back our steps a little, we resorted to rare event algorithms because the lack of data issue meant that our estimate of the committor function was poor, but now we realize that to run rare event algorithms efficiently, we need a good approximation of the committor in the first place.

1.5.5 Coupling machine learning and rare event algorithms

Fortunately rare event algorithms are not so fragile, and indeed most applications use heuristic score functions [Castellana et al., 2019; Cini et al., 2024; Ragone and Bouchet, 2021; Ragone et al., 2018; Zinovjev and Tuñón, 2014; Wouters et al., 2023; Webber et al., 2019; Simonnet et al., 2021; Bouchet et al., 2019a; Rolland et al., 2016]. The idea is then to use the currently available data to get a first estimate of the committor function via machine learning techniques. Then we will use it as score function to run a rare event algorithm and generate new data, which, in turns, can be used to improve the estimate of the committor function (see fig. 1.7). We have just turned the feedback loop from an obstacle into an opportunity.

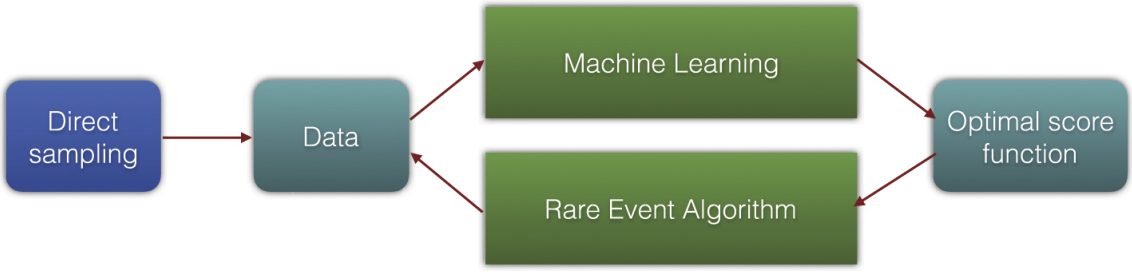


Figure 1.7: Sketch of the coupling process between machine learning and rare event algorithms. Figure taken from [Lucente et al., 2022b].

This approach has already been used successfully in Lucente et al. [2022b], but the authors applied it to a toy model and used the analogue Markov chain method to estimate the committor function. A similar feedback loop was investigated in Nemoto et al. [2016], but again it was applied to very simple models and didn't use techniques that can be strictly called machine learning. In this thesis, instead, we want to build tools that would enable us to run this framework with climate models and neural networks.

1.6 Contribution of this thesis

In this thesis I address the task of estimating committor functions in a regime of lack of data, and to provide a framework for their interpretation, with the long term goal to realize the coupling between machine learning and rare event algorithms described in the previous section. The main contributions of my work are then summarized below.

1.6.1 Interpretability through optimal projection

The first key idea, which I feel is my most *personal* contribution, is that committor functions are useful per se, but are much more useful if the information they provide is human-understandable. And I argue that the only way to achieve this is to drastically reduce the dimensionality of the object we try to understand. For this reason, I developed the framework of optimal projection of the committor function (section 2.2), which consists in finding the best way to represent a reduced version of the committor in a low dimensional space that still retains most of the information encoded in the high dimensional original committor. This framework can then be applied both as a post-hoc explainability method (explored in chapter 2) and as a way to directly compute the committor function in an interpretable way (chapters 3, 4 and 6).

1.6.2 Probabilistic prediction of extreme heatwaves

Using the case study of extreme heatwaves over France, and expanding on the framework of optimal projection, we will devise the simplest non-trivial way to compute the committor (chapter 3). We call this method the Gaussian approximation (GA), and we find that it is a powerful way to combat the lack of data issue. Moreover, it immediately identifies the

important sources of predictability. However, if enough data is available, more complex neural networks are able to extract more information than the Gaussian approximation and thus give better predictions. Chapter 4 is dedicated to finding the sources of information that the Gaussian approximation is not able to capture. To do so we will use post-hoc explainability methods on black box models, as well as a hierarchy of increasingly complex white box models. We will discuss the interplay between complexity, performance and interpretability, and we will show that interpretable models are clearly superior.

1.6.3 Rare event sampling

If the first part of the thesis deals only with machine learning, in the second part we focus on the strategies for improving sampling of rare events. As a first step, in chapter 5 we will run the GKLT rare event algorithm on an intermediate-complexity ocean model to study the noise-induced collapse of the AMOC. Then, in chapter 6, we will investigate the most underdeveloped link in the coupling between machine learning and rare event algorithms, namely how to optimally generate new data, specifically to improve the estimate of the committor function. We will develop a theoretical framework using importance sampling techniques, and then we will test it on a custom toy Two Dimensional Activation Model (TDAM), which I designed specifically to reproduce the situation where information in the bulk of the data may not be very useful for the study of extremes.

1.6.4 Clean coding

A transversal theme running mostly unseen through this thesis is the care that I took when building my code spaces. In a world that values productivity, we are often incentivized to just stick with the first working version and get results fast. This has the downside that the code left by the PhD candidate will be practically unusable by anybody else, and possibly, after a while, also by the very person who wrote it. I experienced this first hand when I had to work with my predecessor's python files. Hence, a significant portion of my time was spent making my code modular, tested and well commented, so that future researchers can benefit from it. All the code is stored in GitHub repositories, which are already public or will become so after their respective papers are published. Here is a brief summary (some links may not work until the repository is published).

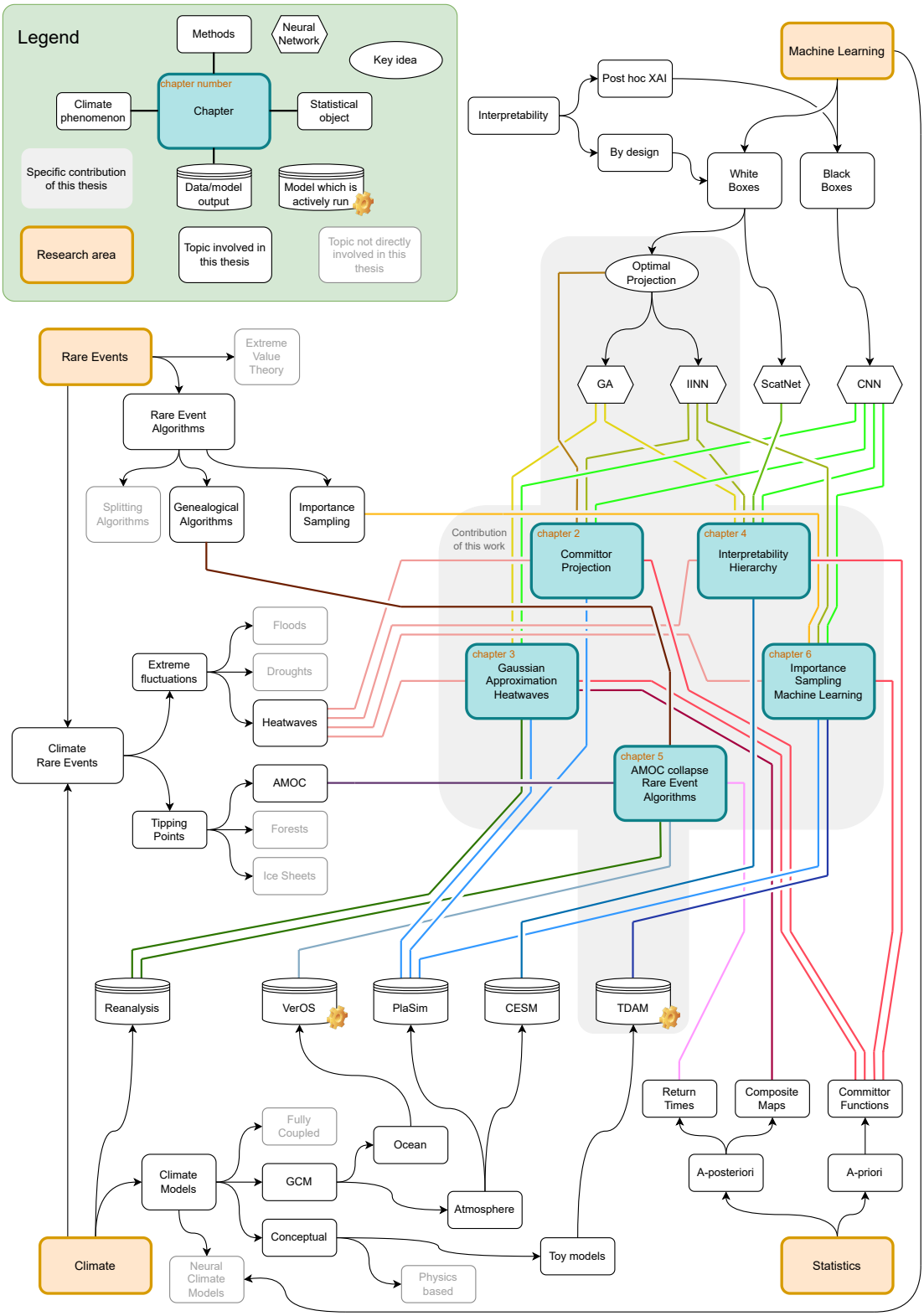
- `https://github.com/AlessandroLovo/RemoteJupyter` - utility for easily running Jupyter notebooks on remote servers. It is already being routinely used by other members of the group in Lyon.

- `https://github.com/AlessandroLovo/general_purpose/` - useful functions for handling data structures like nested dictionaries and easily plot geophysical data and uncertainties.

- `https://github.com/georgemilosh/Climate-Learning` - repository owned by George Miloshevich[1], at the time postdoc in Lyon, containing the main tools for applying machine learning techniques to climate data (used in chapters 2 to 4). Contribution to this repository was one of my biggest coding projects (see for instance `https://github.com/georgemilosh/Climate-Learning/blob/main/PLASIM/Learn2_new.py`).
- `https://github.com/AlessandroLovo/EW2-heatwaves` - code for the second Critical Earth Workshop in Berg en Dal (NL), April 2022.
- `https://github.com/AlessandroLovo/committor_projection` - code containing the framework for committor projection, used in chapter 2.
- `https://github.com/AlessandroLovo/gaussian-approximation-zenodo` - code for reproducing the results of chapter 3.
- `https://github.com/AlessandroLovo/intepretability-hierarchy-zenodo` - code for reproducing the results of chapter 4
- `https://github.com/AlessandroLovo/REA-Veros` - code for applying the GKLT rare event algorithm to climate models. It provides the implementations for testing on the Ornstein-Uhlenbeck process and the coupling with the VerOS ocean model, but can be easily adapted to other models.
- `https://github.com/AlessandroLovo/importance-sampling4parameter-estimation` - python package currently in development for performing optimal importance sampling to improve machine-learned committor functions.

[1] *Centre for mathematical Plasma Astrophysics, Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium*

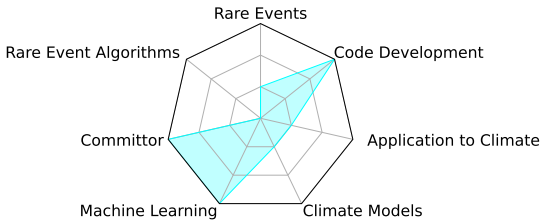
1.6.5 Graphical guide



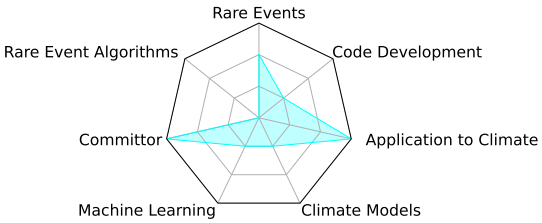
The figure in the previous page acts as a graphical summary of the contributions of this thesis (gray shaded area) and the context in which it sits. 'Faded' blocks are only part of the context and are not actively used in the thesis. Each of the 5 central blocks represents a chapter, which receives data from the bottom, methods from the top, statistical objects of interest from the right and climate phenomena studied from the left. The hexagonal blocks represent the hierarchy of machine learning models used in this thesis. From left to right with increasing degree of complexity: Gaussian approximation (GA), Intrinsically Interpretable Neural Network (IINN), scattering network (ScatNet) and Convolutional Neural Network (CNN). The cylinders represent data sources, either as publicly available datasets like reanalyses [Uppala et al., 2005; Hersbach et al., 2020; Rohde and Hausfather, 2020], or as the output of climate models. Of the latter, we use the Versatile Ocean Simulator (VerOS) [Häfner et al., 2018], the Planet Simulator (PlaSim) [Fraedrich et al., 2005a], the Community Earth System Model (CESM) [Hurrell et al., 2013]. The Two Dimensional Activation Model (TDAM) is a custom-made toy model. The gear symbol on VerOS and TDAM means that these models are actively run, while for the other two models we simply use control runs that were computed for previous works in the group.

Below we show a summary of the focus of each of the following chapters of this manuscript. For each topic, the levels of the radar charts should be interpreted as four different degrees of focus: 0: absent, 1: background, 2: significant, 3: main object of study.

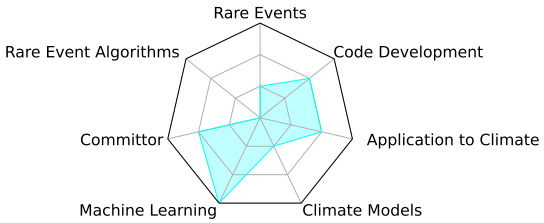
2: Committor Projection



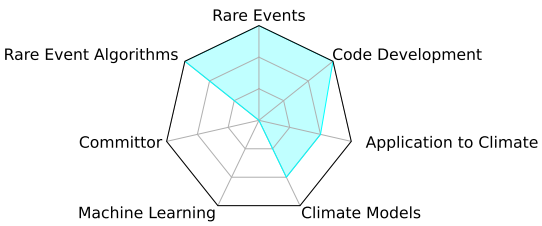
3: Gaussian Approximation Heatwaves



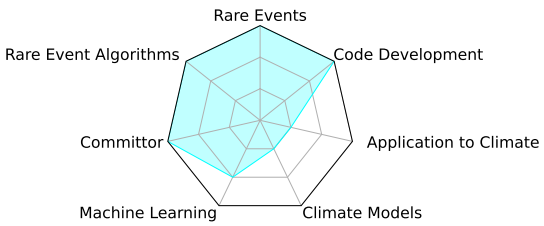
4: Interpretability Hierarchy



5: AMOC collapse Rare Event Algorithms

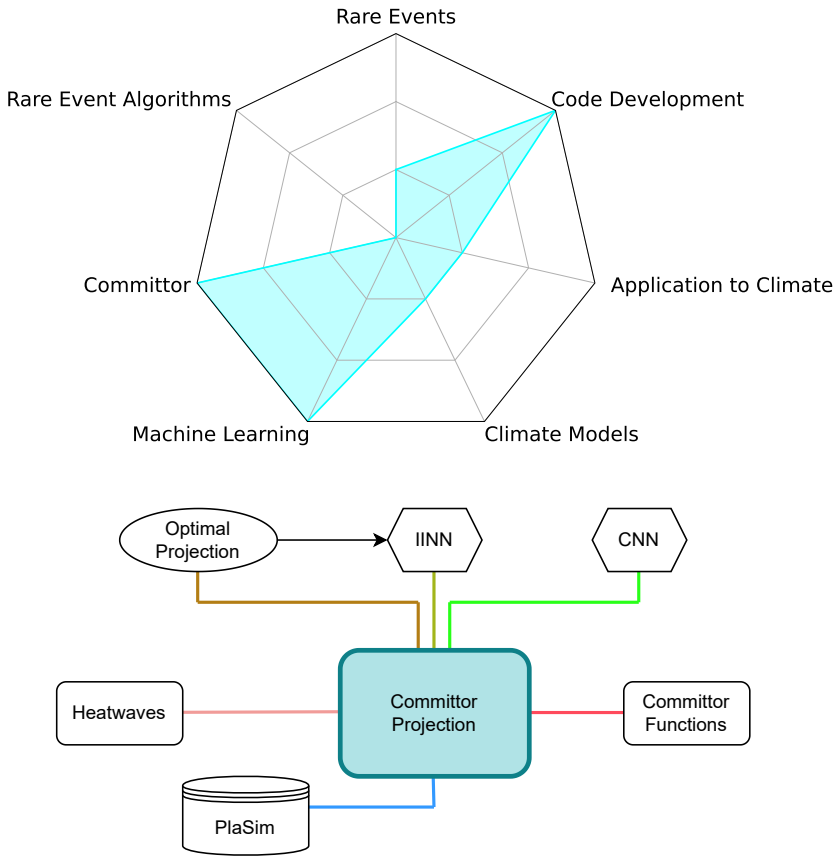


6: Importance Sampling Machine Learning



Chapter 2

Optimal projection of committor functions



The goal of this chapter is to introduce the concept of optimal projection of the committor function, which, as pointed out in the introduction, was a key step in the development of intrinsically interpretable architectures. The results presented in this chapter are mainly methodological, and though we will use as case study the prediction of extreme heatwaves over France, to avoid repetitions I leave most of the discussion of the climate model used and of the underlying physics of heatwaves to chapter 3, where these matters are addressed in detail.

2.1 Probabilistic classification of extreme heatwaves with a Convolutional Neural Network

Even though I just said that I'll leave many details to chapter 3, we still need a minimum of context so that the discussion of the methods we developed makes sense. In particular, to optimally project the committor we need a committor in the first place, and we will compute it using a Convolutional Neural Network.

2.1.1 Predictors and heatwave amplitude

The data used in this work is the output of the simple General Circulation Model called the Planet Simulator (PlaSim) [Fraedrich et al., 2005a,b; Ragone et al., 2018; Miloshevich et al., 2023a], which resolves the atmosphere with a horizontal grid, uniformly spaced in latitude and longitude, of 64×128 pixels and on 10 vertical layers. Sub-grid parameterizations and boundary conditions over land are relatively simplified, and sea ice and sea surface temperature are cyclically prescribed, as well as for incoming solar radiation and greenhouse gases concentrations. This way, the model is able to run in a steady state that reproduces a climate close to the one of the 1990s.

In Miloshevich et al. [2023a], Bastien Cozian[1] used this model to generate an 8000-year-long control run, which is the data we will use in this chapter (and the next as well). In particular, we will focus on the daily averages of 2 m temperature ($T_{2\mathrm{m}}$), 500 hPa geopotential height (Z_{500}) and soil moisture (S), for the three months of summer (June, July and August). Of the first two climate variables we take all grid points above 30 degrees North, while for soil moisture we keep only the 12 pixels over France. Of this data we also take the anomalies with respect to the climatology computed as the day-wise (and of course pixel-wise) average over the whole dataset.

In the end, by combining the three anomaly fields, we are left with a set of $22 \times 128 \times 2 + 12 = 5644$ scalar *predictors*, that we collectively call X.

We then define the heatwave amplitude A as the space and time average of the temperature anomaly:

$$A(t) := \frac{1}{T} \int_t^{t+T} \left(\frac{1}{\mathcal{A}} \int_{\mathcal{A}} T_{2\mathrm{m}}(\vec{r}, u) \mathrm{d}\vec{r} \right) \mathrm{d}u, \tag{2.1}$$

[1] *RTE France*

where T is the duration in days of the heatwave (here we will focus on $T = 14$) and $\mathcal{A}$ is the spatial region of interest, which in our case are the 12 pixels of France.

2.1.2 Estimating the committor function with a neural network

At this point, the question we want to answer is: "Given that the predictors at time $t - \tau$, where $\tau \geq 0$ is the lead time, are x, what is the probability of observing a two-week heatwave that starts at time t?". Namely, we are after the true committor

$$q(x) = \mathbb{P}(A(t) \geq a \mid X(t - \tau) = x). \tag{2.2}$$

In this case, we choose the threshold a to correspond to the 5% most extreme values of A over the whole dataset, which gives $a = 2.76\,\mathrm{K}$. For simplicity, we will also focus here on $\tau = 0$, which means we are after the probability of a two-week heatwave that starts today.

Now, we can use a neural network to write a parametric approximation of the committor $\hat{q}(x; \theta)$, and use the available data to optimize the parameters θ such that $\hat{q}$ is close to the true committor. More precisely, if $\mu(x)$ is the stationary measure of the predictors X, we want to minimize the Kullback-Leibler (KL) divergence

$$KL(q, \hat{q}) = \int \mathrm{d}x \mu(x) \left(q(x) \log \left(\frac{q(x)}{\hat{q}(x; \theta)} \right) + (1 - q(x)) \log \left(\frac{1 - q(x)}{1 - \hat{q}(x; \theta)} \right) \right). \tag{2.3}$$

We can then split the KL divergence into two contributions:

$$KL(q, \hat{q}) = \int \mathrm{d}x \, \mu(x) \left(\mathcal{E}(q(x), \hat{q}(x; \theta)) - \mathcal{E}(q(x), q(x)) \right), \tag{2.4}$$

where

$$\mathcal{E}(w, z) := -w \log z - (1 - w) \log(1 - z) \tag{2.5}$$

is called the point-wise cross entropy between w and z. $\int \mathrm{d}x \, \mu(x) \mathcal{E}(q(x), q(x))$ is the self cross entropy, or simply entropy, of the true committor, and since it does not depend on the estimated committor, we can ignore it during the minimization problem.

What we are left with is the loss we want to minimize:

$$\mathcal{L}(\theta) := \int \mathrm{d}x \, \mu(x) \mathcal{E}(q(x), \hat{q}(x; \theta)). \tag{2.6}$$

However, we don't have access to the true committor (if we had, there would be no need to approximate it!), and neither to $\mu(x)$. Fortunately, we can use our data to solve both problems.

First, we can define the heatwave label

$$Y(t) := \begin{cases} 1 & \text{if } A(t) \geq a \\ 0 & \text{otherwise} \end{cases}, \tag{2.7}$$

and use it to replace the true committor. Then we can replace the integral over the stationary measure $\mu(x)$ by the empirical average over the dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$, which leads us to the empirical loss function

$$\hat{\mathcal{L}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}(Y_i, \hat{q}(X_i; \theta)). \tag{2.8}$$

Another way of reasoning, is by mapping our problem to a very standard one in supervised machine learning: probabilistic classification. In our case we have two classes, identified by the labels Y defined above, namely non-heatwave ($Y = 0$) and heatwave ($Y = 1$).

2.1.3 Proper scoring rule for probabilistic classification

In the literature several losses besides the cross entropy can be used for classification. For example the Brier Score [Brier, 1950]

$$\hat{\mathcal{B}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(Y_i - \hat{q}(X_i; \theta) \right)^2 \tag{2.9}$$

or Matthews Correlation Coefficient [Matthews, 1975]

$$\hat{\mathcal{M}}(\theta) = \frac{N N_1^1 - N_1 N^1}{\sqrt{N_1 N_0 N^1 N^0}}, \tag{2.10}$$

where N_i^j are the entries of the confusion matrix (table 2.1),

	$\hat{Y} = 0$	$\hat{Y} = 1$	total
$Y = 0$	N_0^0	N_0^1	N_0
$Y = 1$	N_1^0	N_1^1	N_1
total	N^0	N^1	N

Table 2.1: The confusion matrix counts the number of true negatives (N_0^0), true positives (N_1^1), false positives (N_0^1) and false negatives (N_1^0).

which are defined based on the predicted heatwave label

$$\hat{Y} := \begin{cases} 1 & \text{if } \hat{q} > 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{2.11}$$

This latter loss, in particular, turns the problem into a *deterministic* classification task [Jacques-Dumas et al., 2022].

However, from a purely theoretical standpoint, it has been proven [Benedetti, 2010] that only the cross entropy, sometimes also called logarithmic score, is the *proper* score for probabilistic classification, where proper means:

1. It is additive with respect to new data
2. Depends only on the probability assigned to events that actually occurred
3. It exhibits an extremant if the assigned probability is held constant on all data points

For instance, Brier Score violates point 2. We refer the interested reader to Benedetti [2010] for further details.

The important conclusion is that the choice of using the cross entropy loss in not arbitrary, but rather well rooted in mathematical reasoning.

2.1.4 Classification or regression?

So far we have seen that it is possible to compute *directly* the committor function by phrasing the problem as a probabilistic classification task. However, there is another option, which is that of probabilistic regression. In this second case we don't estimate directly $\mathbb{P}(A \geq a|X)$, but rather the full conditional probability density function of A given X:

$$\hat{p}(a|x;\theta)\mathrm{d}a \approx p(a|x)\mathrm{d}a = \mathbb{P}(a \leq A < a + \mathrm{d}a|X = x). \tag{2.12}$$

Then, for any particular threshold a, we can immediately get the associated committor function:

$$\hat{q}(x;\theta) = \int_a^{+\infty} \mathrm{d}a'\, \hat{p}(a'|x;\theta). \tag{2.13}$$

A simple way to compute $\hat{p}(a|x;\theta)$ in practice, is to assume that it is a Gaussian distribution, so that the neural network will have to estimate only its first two, state dependent, moments $\hat{\mu}(x;\theta)$ and $\hat{\sigma}^2(x;\theta)$ (see chapters 4 and 6). More advanced possibilities include having a more complex parametric form for $\hat{p}(a|x;\theta)$, for instance including a skewness parameter, or even letting the neural network learn $\hat{p}(a|x;\theta)$ directly, in a process known as quantile regression [Zhang, 2018].

Regardless of the specific choice for expressing $\hat{p}(a|x;\theta)$, when performing probabilistic regression, the neural network will be trained on samples (X_i, A_i) rather than (X_i, Y_i), which has pros and cons. The main advantage is that the heatwave amplitudes A_i contain more information than the heatwave labels Y_i, which then helps the neural network to better discriminate between events that would barely overcome the threshold a and those that would confidently exceed it. However, this extra information may also be distracting. Indeed, the network needs to be good in the overall prediction of the heatwave amplitude, and most of the samples lie in the bulk of the distribution of A, so they might not be particularly useful for the estimation of the tail, which is what we are actually interested in. One way to avoid this, is to use weighted loss functions, which give more importance to samples that are indeed in the tail. We will explore this option in chapter 6.

Now, since in this chapter we are focusing on a methodological work on the committor function, we will stick with the framework of classification, which gives us immediate access to the committor. On the other hand, when performing applications to real heatwaves (chapters 3, 4 and 6) the physical processes are continuous, with no abrupt regime shifts between mild and extreme heatwaves. The choice of the threshold a becomes then quite arbitrary, and so a regression framework is more appropriate. Indeed, in chapter 4 we will see that the committor estimated indirectly by regression and subsequent integration (eq. (2.13)) is more accurate than the one estimated directly by classification.

2.1.5 Network architecture and training parameters

Now that we have all the theoretical ingredients, we can proceed to actually train the network and compute committor functions. Since the input data, i.e. the predictors X, are

image-like, we will use a Convolutional Neural Network (CNN) scheme. As a first step, we organize the input data as a $22 \times 128 \times 3$ tensor where the three climatic variables serve the function of 'color' channels. For soil moisture, we set to zero all grid points outside France. Then we split the data into 10 subsets of equal length, and containing the same number of heatwaves (see section 3.9.2), to perform 10-fold cross validation. For each of the 10 folds, we use the training set to compute the pixel-wise mean and standard deviation and use it to standardize both the training and validation set. This way all the input features entering the network will be of order one, thus facilitating training.

During the first year of my PhD I worked in close collaboration with George Miloshevich[2], who at the time was a postdoc researcher in the group in Lyon. For this reason the precise network architecture is the same as the one in Miloshevich et al. [2023a], quickly summarized in table 2.2. The main core of the network is constituted by three convolutional and pooling layers, followed by two fully connected (dense) layers and a final softmax to ensure that the predicted committor sits between 0 and 1. Batch normalization layers are used to accelerate the learning process and dropout layers act as regularizers to prevent overfitting [Miloshevich et al., 2023a].

The 10 networks (one for each fold) were trained using the *Adam* optimizer with learning rate of 10^{-4}, batch size of 1024 and for a maximum of 40 epochs. As a further measure to prevent overfitting, we perform early stopping, monitoring the validation loss and interrupting training if it doesn't decrease for 5 epochs.

Since we don't perform a hyperparameter optimization step, there is no need for a separate test set, and we will evaluate the networks based on their performance on the validation sets.

[2] *Centre for mathematical Plasma Astrophysics, Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium*

Layer name	kernel size	number of channels	size
Input	-	3	22×128
Convolution	3×3	32	20×126
BatchNormalization	-		
ReLU	-		
SpatialDropout(0.2)	-		
MaxPool	2×2		10×63
Convolution	3×3	64	8×61
BatchNormalization	-		
ReLU	-		
SpatialDropout(0.2)	-		
MaxPool	2×2		4×30
Convolution	3×3	64	2×28
BatchNormalization	-		
ReLU	-		
SpatialDropout(0.2)	-		
Flatten	-	1	3584
Dense	-	1	64
ReLU	-		
Dropout(0.2)	-		
Dense	-	1	2
Softmax	-	1	1

Table 2.2: Architecture of the Convolutional Neural Network used in this work. BatchNormalization layers rescale the incoming data so that each neuron has average activation 0 and standard deviation 1. Dropout(p) and SpatialDropout(p), *only during training* randomly set to zero a fraction p of the neurons of the previous layer, thus acting as a regularization

2.2 Optimal projection of committor functions

The discussion of the performance of the Convolutional Neural Network is already presented in great detail in Miloshevich et al. [2023a], and partially also in chapter 3 of this thesis. Consequently, we will not comment on it here, but rather simply assume that we now have a committor function $\hat{q}(x;\theta)$ that we can play with. In this section we will lay down the mathematical foundations for the optimal projection of the committor function and put them in practice on the one computed by the CNN.

2.2.1 Theoretical framework

The committor is a non-linear function in a very high dimensional space (in our case $\mathbb{R}^{5644}$), which means it is intrinsically difficult to understand. A possible strategy is then to try to project it onto a lower dimensional space. More formally, let's say we have available a committor

$$q : \mathbb{R}^d \to [0,1]. \tag{2.14}$$

We then want to use a projection

$$\varphi : \mathbb{R}^d \to \mathbb{R}^m, \tag{2.15}$$

with $m \ll d$, that will give us a projected (or *reduced*) committor

$$\tilde{q} : \mathbb{R}^m \to [0,1], \tag{2.16}$$

which hopefully will be more easily interpretable.

When we perform the projection, we will reasonably lose some of the information contained in the original committor. We can quantify this loss of information by using the Kullback-Leibler divergence:

$$\mathcal{K}_\varphi := KL(q, \tilde{q}\circ\varphi) = \int \mathrm{d}x\, \mu(x) \left(\mathcal{E}(q(x), \tilde{q}(\varphi(x))) - \mathcal{E}(q(x), q(x))\right), \tag{2.17}$$

as $\mathcal{K}_\varphi/\log(2)$ has the meaning of the number of bits of information lost when using $\tilde{q}\circ\varphi$ as a model for q. As for training neural networks, we want to minimize this loss of information, and, again, we can ignore the second term. Let us, for now, assume that we have a given φ, then,

$$\mathcal{L}_\varphi := \int \mathrm{d}x\, \mu(x) \mathcal{E}(q(x), \tilde{q}(\varphi(x))) \tag{2.18}$$

$$= \int \mathrm{d}f \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \mathcal{E}(q(x), \tilde{q}(f)) =: \int \mathrm{d}f\, \ell(f) \tag{2.19}$$

$$= -\int \mathrm{d}f \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \left(q(x)\log\tilde{q}(f) + (1-q(x))\log(1-\tilde{q}(f))\right). \tag{2.20}$$

We can then minimize $\ell(f)$ with respect to $\tilde{q}(f)$:

$$0 = \frac{\partial \ell(f)}{\partial \tilde{q}(f)} = \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \left(\frac{q(x)}{\tilde{q}(f)} - \frac{1-q(x)}{1-\tilde{q}(f)}\right) \tag{2.21}$$

$$= \mu(\varphi^{-1}(f)) \left(\frac{\bar{q}_{\varphi^{-1}(f)}}{\tilde{q}(f)} - \frac{1-\bar{q}_{\varphi^{-1}(f)}}{1-\tilde{q}(f)}\right), \tag{2.22}$$

where

$$\mu(\mathcal{C}) := \int_{\mathcal{C}} \mathrm{d}x\, \mu(x) \tag{2.23}$$

is the measure of set $\mathcal{C}$, and

$$\bar{q}_{\mathcal{C}} := \frac{1}{\mu(\mathcal{C})} \int_{\mathcal{C}} \mathrm{d}x\, \mu(x) q(x) \tag{2.24}$$

is the average committor over $\mathcal{C}$. From eq. (2.22), we then have that

$$\tilde{q}(f) = \bar{q}_{\varphi^{-1}(f)}, \tag{2.25}$$

which means that the optimal projected committor is the average of the original committor on the iso-levels of the projection function.

Now, to understand what it means to have a good projection, we need to recover the full KL-divergence:

$$\mathcal{K}_\varphi = \int \mathrm{d}f \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \left(q(x) \log \frac{q(x)}{\tilde{q}(f)} + (1 - q(x)) \log \frac{1 - q(x)}{1 - \tilde{q}(f)} \right) \tag{2.26}$$

$$=: \int \mathrm{d}f \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \kappa(q(x), \tilde{q}(f)). \tag{2.27}$$

Let us focus on a single f-slice, and define $\delta(x) := q(x) - \tilde{q}(f)$. If we simplify the notation with

$$q \equiv q(x), \quad \tilde{q} \equiv \tilde{q}(f), \quad \delta \equiv \delta(x),$$

we can then write

$$\begin{aligned}
\kappa(q, \tilde{q}) &= (\tilde{q} + \delta) \log \left(1 + \frac{\delta}{\tilde{q}} \right) + (1 - \tilde{q} - \delta) \log \left(1 - \frac{\delta}{1 - \tilde{q}} \right) \\
&= (\tilde{q} + \delta) \left(\frac{\delta}{\tilde{q}} - \frac{1}{2} \left(\frac{\delta}{\tilde{q}} \right)^2 + O(\delta^3) \right) + (1 - \tilde{q} - \delta) \left(-\frac{\delta}{1 - \tilde{q}} - \frac{1}{2} \left(\frac{\delta}{1 - \tilde{q}} \right)^2 + O(\delta^3) \right) \\
&= \delta - \frac{1}{2} \frac{\delta^2}{\tilde{q}} + \frac{\delta^2}{\tilde{q}} - \delta - \frac{1}{2} \frac{\delta^2}{1 - \tilde{q}} + \frac{\delta^2}{1 - \tilde{q}} + O(\delta^3) \\
&= \frac{1}{2\tilde{q}(1 - \tilde{q})} \delta^2 + O(\delta^3).
\end{aligned}$$

It follows that

$$\mathcal{K}_\varphi \approx \int \mathrm{d}f\, \frac{1}{2\tilde{q}(f)(1 - \tilde{q}(f))} \int_{\varphi^{-1}(f)} \mathrm{d}x\, \mu(x) \left(q(x) - \tilde{q}(f) \right)^2. \tag{2.28}$$

Which means that a good projection minimizes the variance of the original committor on the iso-levels of φ. In particular, the best projection, i.e. one that doesn't lose any information, is any injective function of the committor itself. This way the iso-levels of φ are the same of the ones of q and so q is constant on them. This solution, however, is not what we are looking for, as our projection will be as interpretable as the full committor itself.

Linear projection

To actually learn something, we need to impose a constraint on φ, for example requiring it to be linear:

$$f = \varphi(x) = Mx, \tag{2.29}$$

where M is an $m \times d$ matrix. The advantage of the linear projection is that it is intrinsically interpretable as each of the m coordinates of the projected space represents the correlation between the input field x and a specific pattern (or map) M^μ, $\mu = 1, \dots, m$. Moreover, since each pattern has the same dimension as the data, we can easily visualize it.

An important observation is that we don't care about the full function φ, but only about its iso-levels. In other words, $\mathcal{K}_\varphi$ is invariant by any (possibly non-linear) injective rescaling of φ. For the linear case, the iso-levels are hyper-planes, and we care only about their orientation in $\mathbb{R}^n$, thus we can impose

$$|M^\mu|^2 = \sum_{\nu=1}^{d} (M^\mu_\nu)^2 = 1 \quad \forall \mu = 1, \dots, m. \tag{2.30}$$

2.2.2 Binning

Now that we have a theoretical framework, we want to try to project the predicted committor $\hat{q}$. Since we have a finite amount of data, we cannot actually use the precise iso-levels $\varphi^{-1}(f)$, as too few points will land exactly on that hyperplane to compute a meaningful average. What we will do, is instead resort to binning. In practice, first we compute the projection f_i for each data point X_i. Then we can subdivide $\mathbb{R}^m$ into B bins $\{H_b\}_{b=1}^B$, and for each bin compute the average committor

$$\bar{q}_b = \langle \hat{q}(X_i, \theta) \rangle_{i | f_i \in H_b}, \tag{2.31}$$

and broadcast it to all the data points in that bin

$$\tilde{q}_i := \bar{q}_b \quad \forall i | f_i \in H_b. \tag{2.32}$$

Finally, we can compute the KL-divergence as an empirical average over our dataset:

$$\mathcal{K}_\varphi = \int \mathrm{d}x \, \mu(x) \kappa(\hat{q}(x), \tilde{q}(\varphi(x))) \approx \frac{1}{N} \sum_{i=1}^{N} \kappa(\hat{q}(X_i; \theta), \tilde{q}_i). \tag{2.33}$$

From the understanding that we want to minimize the variance inside each bin, we see that $\mathcal{K}_\varphi$ is, necessarily, a monotonically decreasing function of the number of bins. Intuitively, if we have enough of them, at some point there will be at most one data point in each bin and hence $\tilde{q}_i = \hat{q}(X_i; \theta)$ $\forall i$, yielding $\mathcal{K}_\varphi = 0$.

Fortunately, as we can see from fig. 2.1, there is a broad region where $\mathcal{K}_\varphi$ plateaus. If we stick to projecting to one dimension, namely $m = 1$, we can use Scott's formula [Scott, 1979] for the optimal choice of the bin width:

$$h = 3.49 \sigma_f N^{-1/3} \tag{2.34}$$

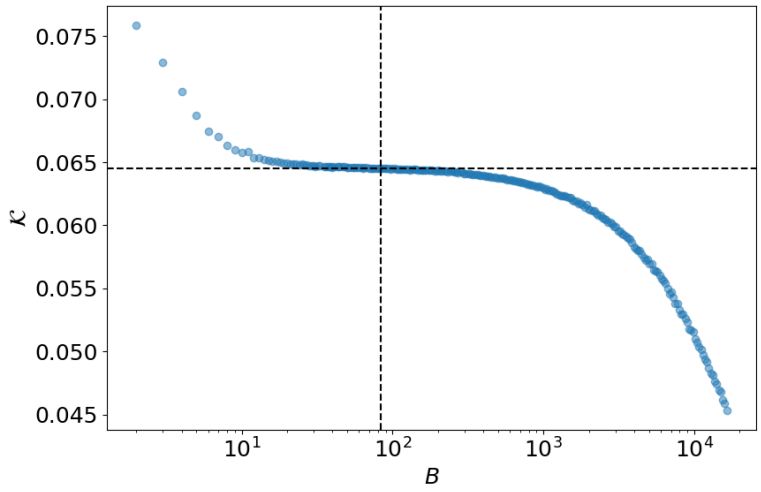


Figure 2.1: $\mathcal{K}_\varphi$ as a function of the number of bins when projecting to a one dimensional space. Dashed lines refer to the values obtained using Scott's formula (eq. (2.34)).

where σ_f is the standard deviation of $\{f_i\}_{i=1}^N$. This formula is based on the assumption of a Gaussian distribution, but this is not a problem as the important thing is to be on the plateau, and we have a decent margin for error.

2.2.3 1-dimensional projections

Let us start by focusing on the case $m = 1$, where bins are consecutive intervals, and we can drop the index μ. The simplest possible experiment that we can perform is to project onto a single grid point, namely setting all but one of the entries of M to zero. If we repeat this for all the grid points we can then plot the result as a map and identify which pixels are the most important for prediction. Since the values of $\mathcal{K}_\varphi$ are not immediately intuitive, we can rescale them to represent the fraction of information lost when neglecting all other pixels. To do so, we can assume that if we don't project onto anything, i.e. $M = 0$, we indeed lose all information. But in this case there is only a single iso-level of φ, which is the whole projected space $\mathbb{R}^1$. Then,

$$\mathcal{K}_0 = \frac{1}{N}\sum_{i=1}^{N} \kappa(\hat{q}(X_i;\theta), \langle \hat{q}(X_i;\theta)\rangle). \tag{2.35}$$

Since this is the maximum amount of information we can lose, it is also the information content of $\hat{q}$, and so when we plot $\mathcal{K}_\varphi/\mathcal{K}_0$ it has indeed the proper meaning of the fraction of information lost when projecting the committor with φ.

In the top row fig. 2.2 we plot these values for each pixel, which shows some interesting patterns. To better understand them, let us take a step back and have a look at the simple a posteriori heatwave statistics of the composite map. As explained in section 1.2, the composite map is the average state of X given that the heatwave happened. Here we can write the empirical average as

$$C = \langle X_i \rangle_{i|Y_i=1}\,. \tag{2.36}$$

Similarly, we can compute the pixel-wise standard deviation S of the same set of data points that led to a heatwave. If we now take the pixel-wise ratio between the two, $K = C/S$, we have a quantification of how different from the global mean (which is 0 since the input fields are normalized), the composite map is at each pixel. We call K the significance map, and we plot it in the bottom row of fig. 2.2. By comparing it with the map of $\mathcal{K}_\varphi$, we can see a stunning similarity, where the regions with higher values of $|K|$ also retain more information when we project onto them.

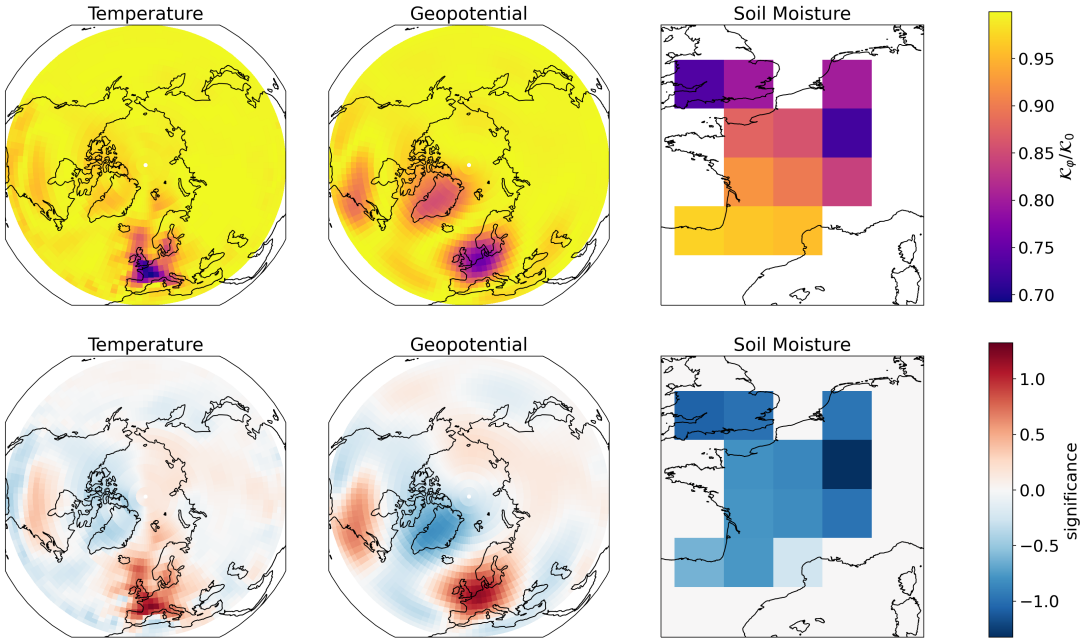


Figure 2.2: Fraction of information lost when projecting on a single grid point (top) and significance map (bottom). Most points yield an almost complete loss of information, but a few regions manage to retain 10 to 30% of it. The best single-pixel projection is to look at the temperature values over France, followed by soil moisture in Northeastern France and finally geopotential height over Western Europe, Greenland and the Eastern United States. Low information loss is highly correlated with high significance.

Now, to get a better intuition of why this is the case, we need to look at the projected committor $\tilde{q}(f)$. From the orange line in the top left panel of fig. 2.3 we can see that when we project onto a pixel of temperature over Siberia (the one which has the highest information loss), the profile of the projected committor is completely flat. In other words, the specific value of f in the projected space doesn't give us any information about the original committor. On the other hand, when we project onto the pixel with the lowest information loss (temperature in Normandy), we can clearly see that low values of f correspond to very low values of $\hat{q}$, while high values of f correspond, on average, to higher values of $\hat{q}$. And now we have all the ingredients to understand the similarity between K and $\mathcal{K}_\varphi$ shown in fig. 2.2. Indeed, if a given pixel is highly significant in the composite map, it means that the values it assumes within the heatwave class are very different from the ones in the non-heatwave class, and, thus, it is a good way to discriminate the two classes.

Since the neural network is good at its job [Miloshevich et al., 2023a], configurations of X that are predicted to lead to a heatwave with probability q actually result in a heatwave a fraction q of the times. This means that the data for which the network predicts a high committor mostly belong to the heatwave class, and thus the particular pixel which was highly significant in the composite map is also a good way to discriminate between low and high committor values, which makes it a good candidate to project the committor on.

This reasoning was rather qualitative for the moment, and that is fine because the goal was to build an *intuition* of what is happening. With some proper hypotheses, we can build a much more quantitative picture, that will give us a better understanding of the committor function and its relationship with composite map. This will be the main object of study of chapter 3, so we won't go further into the details here.

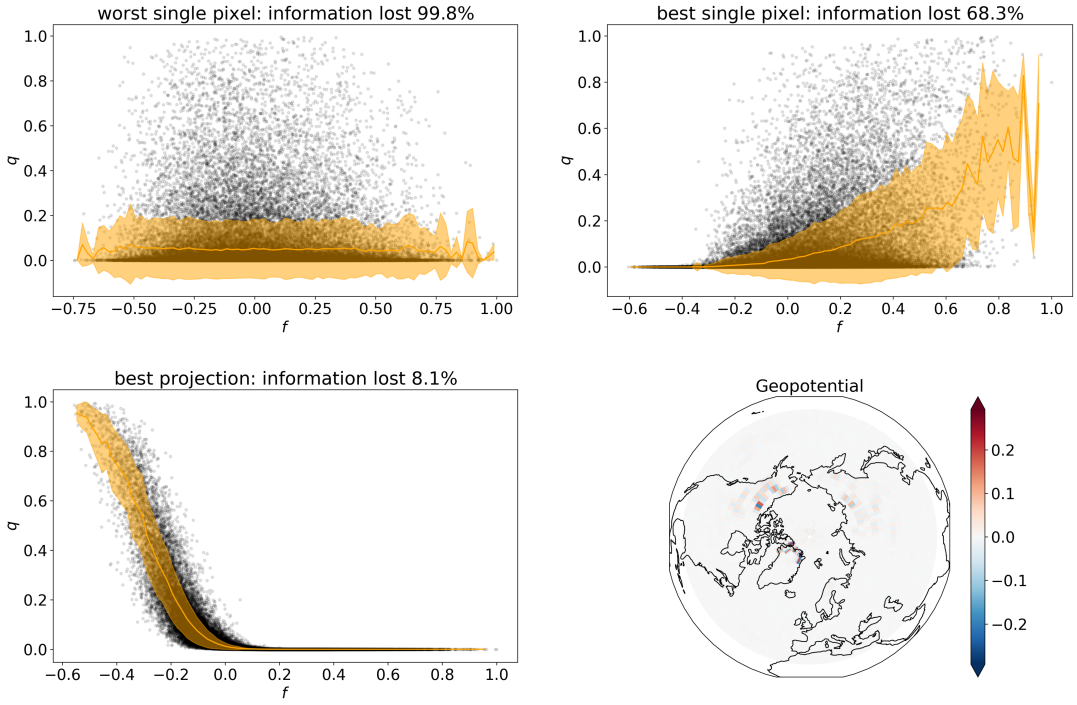


Figure 2.3: The projected committor for three different types of projection: in order, worst single pixel (temperature in Siberia), best single pixel (temperature in the North of France) and SIR projection. In these three plots the black dots are the original committor $\{(f_i, \hat{q}(X_i;\theta))\}_{i=1}^N$ while the orange line is the reduced committor $\tilde{q}(f)$. The shaded area is the standard deviation of $\hat{q}$ inside each f-bin. On the bottom right, the geopotential height part of M_{SIR}, which shows a very noisy pattern. The temperature and soil moisture components are quite similar (not shown).

Going back to the plots in figs. 2.2 and 2.3, we see that even the best pixel retains only 30% of the original information contained by $\hat{q}$, and this was somewhat expected because we are ignoring completely the other 5643 pixels. Ideally, we would like to find a more general projection pattern M, which optimally combines the information content of all pixels and thus leads to a very low $\mathcal{K}_\varphi$. Since we are doing machine learning, the first

thought is to perform gradient descent on $\mathcal{K}_\varphi$ itself. Unfortunately, due to the presence of bins, necessary to compute $\tilde{q}$, $\mathcal{K}_\varphi$ is highly non-differentiable, and so it proves very hard to optimize directly.

However, the intuition we developed from looking at the top row of fig. 2.3, is that we want f to be a good proxy for $\hat{q}$. In a more mathematical sense, there exists a function g such that $\hat{q}(X_i;\theta)$ and $g(M \cdot X_i)$ are highly correlated. Of course, we know that the best g is $\tilde{q}$, but forgetting the precise details of how it is computed removes the problem of differentiability and allows us to apply Sliced Inverse Regression (SIR) [Li, 1991]. Indeed, this dimensionality reduction method bins the data with respect to $\hat{q}$ (hence the 'inverse') and computes M such that the average f of each bin are as far apart as possible (see [Li, 1991] for more details). This way, f will be highly informative of $\hat{q}$.

If we apply SIR to find M and then compute $\tilde{q}$ as explained in section 2.2.2, we get the astonishing result of the bottom left plot of fig. 2.3, where we lose only 8% of the original information. This is indeed remarkable, as it shows that the committor computed by a complex Convolutional Neural Network in a $d = 5644$-dimensional space can be approximated very well by a linear projection to a single *optimal index*, followed by a non-linear, scalar activation function. Which, by the way, looking at the plot of $\tilde{q}$ looks very similar to a sigmoid.

Unfortunately, when we try to visualize the projection pattern M, which yields the optimal index, we are disappointed to find something very obscure (see the bottom right plot of fig. 2.3), where most values are very close to zero and a few points around the polar circle have very high values, with no spatial continuity. This comes from the fact that SIR involves computing the inverse of the $d \times d$ covariance matrix of X, which is ill-conditioned and thus leads to singularities (see also section 3.9.12).

We could continue the analysis digging into the Python package for SIR and adding a regularization term, but it is not worth it. Indeed, the results so far were already very successful in pointing out three key conclusions:

- It is possible to project the committor function to a much lower dimensional space with minimal loss of information.
- Avoiding expressing the projected committor in its theoretically optimal form simplifies greatly the optimization problem.
- Regularization is needed to get a physically interpretable projection pattern.

In particular, the first conclusion suggests that we may skip the CNN entirely, and learn directly from the data a committor which has the form $\tilde{q}(M \cdot X)$, which would yield a prediction that is intrinsically interpretable.

2.3 Intrinsically Interpretable Neural Networks

And so, to conclude this chapter, we propose the framework of the Intrinsically Interpretable Neural Network (IINN), shown in fig. 2.4. This way, the first layer of the network

is in charge of finding the optimal projection pattern(s), while the rest of the network computes the reduced committor $\tilde{q}$. As long as this second block is differentiable, we can apply standard stochastic gradient descent and back-propagation techniques and train the whole network as any other architecture.

Crucially, since this second block maps a low dimensional space to, essentially, the one-dimensional space of the committor, it doesn't matter how complex and obscure it is, we will always be able to interpret it by plotting it end-to-end.

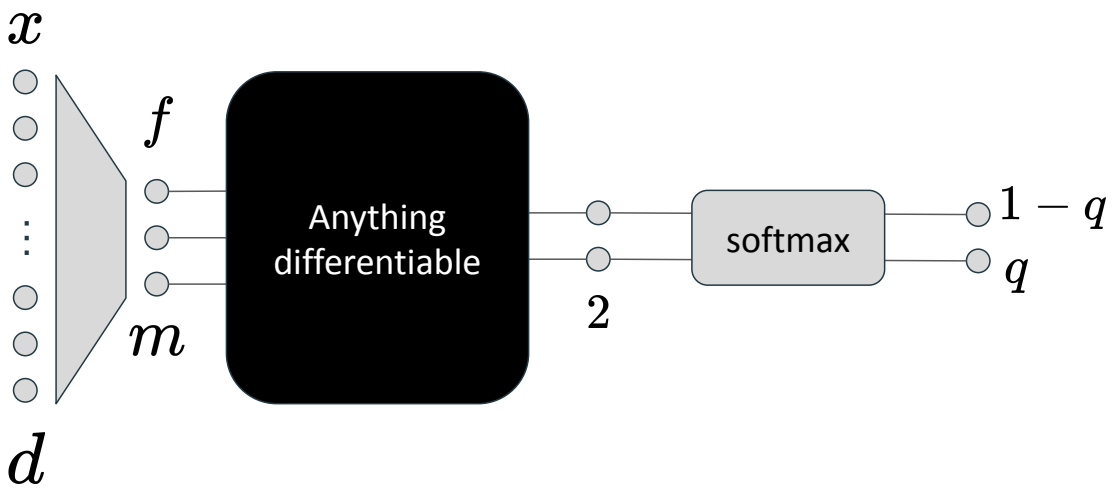


Figure 2.4: Sketch of the IINN architecture for a binary classification task: The first fully connected layer projects $x \in \mathbb{R}^d$ to m optimal indices collectively called f, which are then used by the black box to compute the logits of the committor ($\log(q)$ and $\log(1-q)$), as it is commonly done in probabilistic classification problems.

Also, now the m projection patterns M^μ are encoded in the weights of the first layer, and so it becomes very easy to add a regularization term to the loss function, which would ensure that the projection patterns are interpretable. For instance one could add an L_2 penalty, or a custom-made one that penalizes the spatial gradient of M, thus forcing it to be smooth (see sections 3.9.6 and 3.9.12 for more details).

This flexible IINN architecture will be properly tested the task of predicting heatwaves in chapter 4, but another option is that of prescribing a parametric form for the black box that represents $\tilde{q}$. For instance, in fig. 2.3 we have seen that $\tilde{q}$ resulting from SIR looked like a sigmoid, so why not *impose* it to be a sigmoid? If we do so, we force $m = 1$, and the only trainable parameters are the d entries of M. We have turned the problem into a simple logistic regression. As we will see in chapter 3, we will reach a very similar approach, though from a more rigorous mathematical reasoning, rather than intuition. However, the intuitions presented in this chapter were the reason we decided to undergo precise derivations in the first place, and though these derivations led to the development of tools that outperform the experiments of this chapter, I think they still have pedagogical value.

As a sidenote, in this chapter we presented projection to an m-dimensional space, but in the following of this manuscript we will always use $m = 1$. As we will see in chapter 3,

this is because extreme heatwaves are very 'well-behaving phenomena', which can be forecasted very well using simple regularized linear regression. We did try to train IINNs with $m = 2, 3, 4$, but none were better than the ones with $m = 1$. Nevertheless, there is potential for applying IINNs with $m > 1$ to different case studies. For instance, Delaunay and Christensen [2022] show that the Madden-Julian Oscillation is well represented in a two-dimensional space.

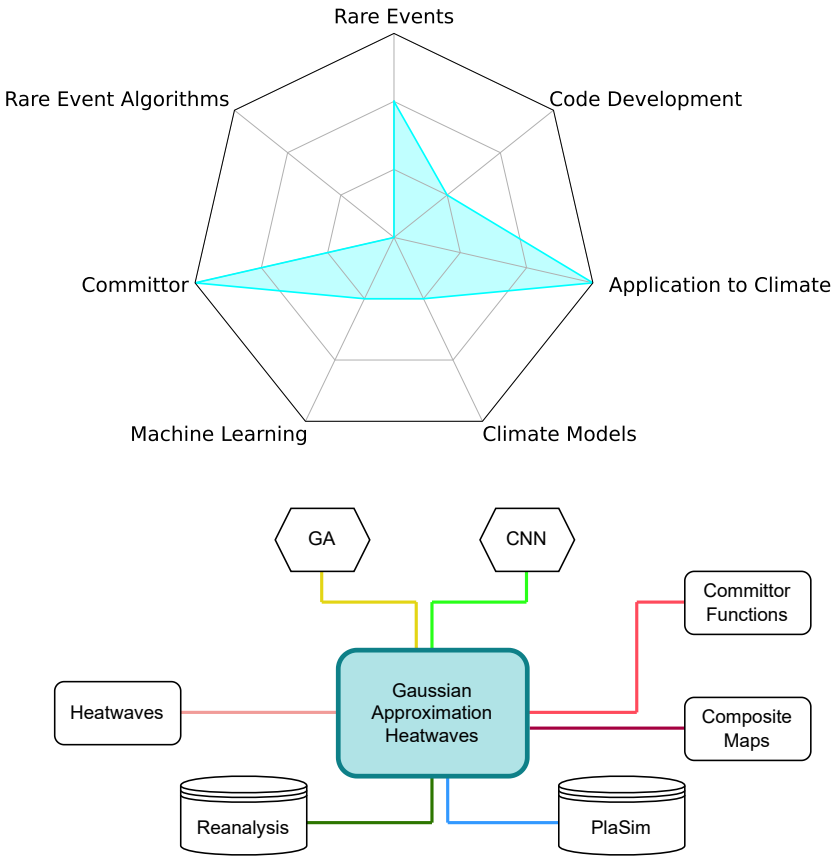
2.4 Conclusions

In this first short chapter, we briefly discussed how to approach heatwave prediction as a probabilistic classification problem. More importantly, we laid the foundations for the concept of optimal projection of the committor function which can either be used as a model-agnostic post-hoc explainability tool, or directly as white box model, in the form of an Intrinsically Interpretable Neural Network. This concept is extremely important when we want to gain an understanding of high dimensional problems, which is commonly the case in climate science.

The idea of optimal projection will run as a thread through the rest of this manuscript, a reference and a guide as we wander through the study of extreme events. Dear reader, this is your Virgil.

Chapter 3

Gaussian Framework and Optimal Projection of Weather Fields for Prediction of Extreme Events



In this chapter, we embark in a mathematical discussion to outline the inherent differences between a priori (committor function) and a posteriori (composite maps) statistics of extreme events. We then refine the intuition of optimal projection of the committor function presented in chapter 2 and develop, from simple mathematical assumptions, a framework to easily compute an interpretable committor directly from data. We will then apply this framework to the study of extreme heatwaves over France, and compare its performance with the Convolutional Neural Networks presented in section 2.1. As we will see, the simplicity of the new method will not only provide interpretability, but also combat the lack of data issue.

What follows is the paper I wrote together with Valeria Mascolo[1], which is currently in review at the American Geophysical Union's Journal of Advances in Modeling Earth Systems (JAMES). This paper is nicely split into the analysis of composite maps and committor functions, with the former being mainly a contribution from Valeria Mascolo and the latter mainly mine. However, our collaboration involved a lot of back and forth, so there is no sharp distinction of who did what.

For consistency and simplicity, the paper is reported here exactly the same as its preprint, available on arXiv at `https://arxiv.org/abs/2405.20903v2`. Hence, expect some minor repetitions of concepts which have already been presented, as well as some shifts in writing style.

Key Points:

- This work presents a new simple framework, called the Gaussian approximation, for a-posteriori and a-priori statistics of extreme events.
- Our method provides an interpretable probabilistic forecast of extreme heatwaves which is competitive with off-the-shelf neural networks.
- The analysis highlights quasi-stationary Rossby waves and low soil moisture as precursors to extreme heatwaves over France.

Abstract

Extreme events are the major weather related hazard for humanity. It is then of crucial importance to have a good understanding of their statistics and to be able to forecast them. However, lack of sufficient data makes their study particularly challenging.

In this work we provide a simple framework to study extreme events that tackles the lack of data issue by using the whole dataset available, rather than focusing on the extremes in the dataset. To do so, we make the assumption that the set of predictors and the observable used to define the extreme event follow a jointly Gaussian distribution. This

[1] *ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France*

naturally gives the notion of an optimal projection of the predictors for forecasting the event.

We take as a case study extreme heatwaves over France, and we test our method on an 8000-year-long intermediate complexity climate model time series and on the ERA5 reanalysis dataset.

For a-posteriori statistics, we observe and motivate the fact that composite maps of very extreme events look similar to less extreme ones.

For prediction, we show that our method is competitive with off-the-shelf neural networks on the long dataset and outperforms them on reanalysis.

The optimal projection pattern, which makes our forecast intrinsically interpretable, highlights the importance of soil moisture deficit and quasi-stationary Rossby waves as precursors to extreme heatwaves.

Plain Language Summary

Extreme weather events such as heatwaves are responsible for large financial and human costs and their impact can only be expected to grow in the future. Understanding such events and being able to predict them is therefore of major interest, but suffers from a fundamental problem of lack of data. In this work we present a new framework which addresses this issue by making simple assumptions on the statistics of weather fields relevant for heatwaves. We validate our method using a very long climate simulation. We find that it provides good approximations of atmospheric conditions prevailing during heatwaves, and good prediction capabilities. It even outperforms existing approaches for short datasets, such as those obtained by combining observations and state-of-the-art weather prediction models, which contain much less extreme events than climate simulations but represent more accurately the dynamics of the atmosphere. This approach explains the observed property that more extreme events are simply stronger versions of less extreme ones, and allows to identify the features of atmospheric patterns which are relevant for making predictions. The method is very general and could be applied for many types of extreme events.

3.1 Introduction

Extreme weather and climate events, often exacerbated by climate change, have led to major disasters in our recent history [Seneviratne et al., 2012]. Heatwaves, in particular, are among the deadliest events. Prolonged exposure to abnormal heat for a certain duration has proven to worsen existing illnesses and to have caused excess deaths during the recent events of the Western European heatwave of 2003 and the Russian heatwave of 2010 [Fouillet et al., 2006; García-Herrera et al., 2010; Barriopedro et al., 2011]. Moreover, losses in the agricultural sector with the subsequent endangerment of the food production system, together with the endangerment of entire ecosystems, allow classifying heatwaves as events

which have critical impacts on the whole society, according to the Intergovernmental Panel on Climate Change [Seneviratne et al., 2021].

The intensification and the proliferation of these extreme events in the current climate call for urgent progress in our understanding of the mechanisms that drive them, and for developing prediction tools to anticipate risks. However, the most extreme events are the rarest. For this reason, those two classical tasks of analysis and prediction for extreme event study suffer from large methodological difficulties associated to a lack of both historical and model data [Miloshevich et al., 2023a]. In this paper we propose a new framework to infer analysis and prediction tools, which is effective with rather short datasets, and efficient for the rare unobserved events up to some approximation we fully characterize. Here, we test thoroughly this framework for extreme heatwaves, but we surmise that it can be applied to a large set of other extreme events.

For the task of understanding which weather conditions led to extreme events, once they have occurred, composite patterns, i.e. maps of averaged dynamical variables conditioned on the outcome of the extreme event, are the most commonly used statistical diagnostic (see for instance [Grotjahn and Faure, 2008; Sillmann and Croci-Maspoli, 2009; Teng et al., 2013; Ratnam et al., 2016; Miloshevich et al., 2023c; Noyelle et al., 2024]). As visible in fig. 3.1 for reanalysis data and two other climate models, the composite patterns associated very extreme events strikingly resemble those for less extreme ones. This fascinating property has not been much commented in the literature before a recent study [Miloshevich et al., 2023c] and has never been explained. Whenever this property is relevant, it means that composite maps for rare events can be computed from typical statistics, even if those rare events have not been observed. This is of huge practical interest, and requires understanding. The Gaussian framework we develop in this paper gives a straightforward and enlightening explanation.

For the second task, prediction of future extreme events based on current weather conditions, composite maps are not useful. We clearly demonstrate and explain this in the present paper. The appropriate statistical concept to make predictions is the probability that an extreme event will occur conditioned on the present state of the climate system, the so-called committor function. However, in order to compute this committor function, one actually has to build a forecasting tool able to estimate this probability. Moreover, the committor function is a function of all the variables which characterize the state of the system, called predictors. For these reasons, it is extremely hard to compute practically and to represent it. Several computations of committor functions have been performed with applications in either geophysical fluid dynamics or in climate sciences [Finkel et al., 2021; Miron et al., 2021; Finkel et al., 2020; Lucente et al., 2019, 2022a,b], using either direct or involved approaches. For climate sciences, methods have been devised using either analogue Markov chains [Lucente et al., 2022a], Galerkin approximations of the Koopman operator [Thiede et al., 2019; Strahan et al., 2021], or neural networks [Lucente et al., 2019; Miloshevich et al., 2023a]. Neural network seems to be the most efficient and versatile tool. As a matter of fact, there is currently a flourishing literature using neural

networks for spatial and temporal predictions of several families of extreme events, such as hurricanes [Racah et al., 2017], tropical cyclones [Giffard-Roisin et al., 2020], droughts [Agana and Homaifar, 2017; Dikshit et al., 2021], and heatwaves [Chattopadhyay et al., 2020; Jacques-Dumas et al., 2023; Miloshevich et al., 2023a]. However, in [Miloshevich et al., 2023a] the authors clearly demonstrate that machine learning for rare extreme events is most of the time performed in a regime of lack of data and gives suboptimal predictions for typical climate datasets. Moreover, deep learning approaches are, in general, very hard to interpret [Bach et al., 2015a; Krishna et al., 2022; Rudin, 2019], and it is extremely difficult to gain some understanding using the forecasting tool.

The main aim of this work is to propose a much simpler alternative method to devise a forecast tool for prediction and to explain the structure of composite maps. This new framework is based on the assumption that the joint probability distribution of the predictors and the extreme event amplitude is Gaussian. Even if this hypothesis is verified only approximately, we show in this paper that the quality of its prediction and its potential for interpretability is extremely high, for extreme heatwaves. We prove that this hypothesis gives a very simple and straightforward explanation of the stability of composite patterns when changing the extreme event amplitude. For the prediction problem, this Gaussian hypothesis leads to a linear regression problem of the heatwave amplitude on the predictor fields. This is in sharp contrast with regression of fields on scalars value, commonly used in climate sciences. In this case, the predictor is a field in very high dimension, and the predicted value is a scalar. The key outcome of this procedure is a regression map, which we call the optimal prediction map for the extreme event. This optimal prediction map is a new concept of this study. It is directly interpretable as it gives, at each geographical location, the importance of the predictor field and its sign to determine the heatwave amplitude. Because of the high dimension of the predictors and because of the not so long dataset length, this regression requires regularization. We analyze thoroughly such optimal prediction maps for extreme heatwaves.

A large part of the work is devoted to the estimate of the accuracy of the results obtained using the Gaussian approximation, compared to the truth. It turns out that this Gaussian approximation is able to give fully interpretable results which compare very well with the truth. For instance, it computes composite maps up to errors of the order of 20 to 30%, depending on the cases. Moreover, this Gaussian approximation requires much fewer data, and it can predict composite maps for unobserved events. For prediction, it should often be preferred to neural networks for short datasets. For instance, we prove to have a prediction skill close to convolutional neural networks on very long datasets and to outperform them on short datasets, like the 80-year long ERA5 reanalysis.

This work is organized as follows. In section 3.2 we give the definition of heatwaves used for this study, we present the two datasets used and the set of predictors. In section 3.3 we show with two theoretical examples that composite maps and committor functions are two different probabilistic objects. We then introduce the Gaussian approximation framework, and we derive the formulae for computing composite maps and committor

functions. Section 3.4 and section 3.5 are dedicated to a methodological study of the Gaussian framework using the climate model PlaSim. Finally, in section 3.7 we apply our methodology to the reanalysis dataset ERA5. In section 3.8 we summarize our findings and give perspectives for future works.

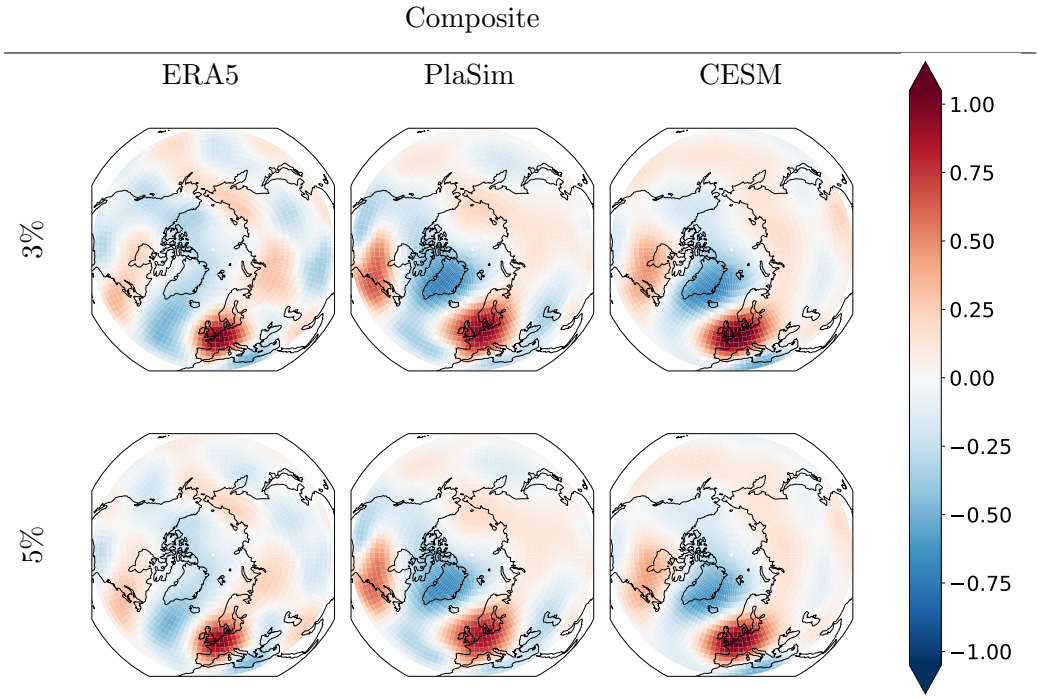


Figure 3.1: First line: maps of 500 hPa geopotential height anomaly for heatwaves over France, defined as situations with the 3% most extreme values of two weeks averaged 2 m temperature anomaly over France (3% composite maps). Second line: the same for a 5% threshold (5% composite maps). The maps are normalized pixel-wise by the climatology standard deviation. Composite maps are estimated respectively on ERA5 (daily data from 1940 to 2022), PlaSim (8000 years of simulation), CESM (1000 years of simulation), datasets. The models reproduce very well ERA5 patterns. Moreover, while the amplitude depends on the threshold defining heatwaves, strikingly the patterns do not. Indeed, we observe in all models and for both thresholds a strong anticyclonic anomaly over Western Europe (which is correctly correlated with the fact that we aim at predicting heatwaves over France). This anticyclonic anomaly is part of a train of a cyclone and an anticyclone which starts over the western part of the United States and continues with a cyclonic anomaly over North Atlantic Ocean for ERA5, while it is northward shifted over Greenland for both PlaSim and CESM.

3.2 Heatwave Definition, Datasets, and Predictors

In this section we provide the definition of heatwaves that will be used in the following (section 3.2.1), we present the datasets (section 3.2.2), and we identify the weather variables of interest (section 3.2.3).

3.2.1 Heatwave Definition

In the literature heatwaves have been defined in a plethora of different ways for different analysis purposes [Perkins, 2015]. Short and long-lasting heatwaves affect differently our society and environment, but long-lasting ones are the most detrimental [Barriopedro et al., 2011]. Despite this, most of the literature on heatwaves focuses on daily events [Seneviratne et al., 2012], as was pointed out in the last assessment report of the Intergovernmental Panel on Climate Change [Seneviratne et al., 2021].

Having a definition which measures independently the persistence and the amplitude of heatwaves is thus of primary interest. The simplest way to achieve this is by monitoring the running average of the air temperature field, and this has been applied to the study of heatwaves of different duration (7 days, two weeks, one month) [Barriopedro et al., 2011; Coumou and Rahmstorf, 2012; Schär et al., 2004]. In this work, following the recent studies of [Gálfi et al., 2019; Gálfi and Lucarini, 2021; Ragone et al., 2018; Ragone and Bouchet, 2021; Jacques-Dumas et al., 2023; Miloshevich et al., 2023a], we use a definition which is based on a time and a spatial average of the 2 m temperature anomaly. We believe that this viewpoint is complementary with the more common definitions [Perkins, 2015] and relevant for our analysis. Such an average-based definition has the advantage of carrying a natural measure of the heatwave amplitude, which can be easily adapted to heatwaves of different duration and intensity or over different regions of the globe. On the contrary, many classical heatwave definitions involve hard thresholds to be reached within specified time frames and are thus less flexible [Perkins, 2015].

Let $\tilde{T}_{2\mathrm{m}}$ denote the daily-averaged 2 m air temperature field, which depends on the location $\vec{r}$ and time t. Given that the statistics of $\tilde{T}_{2\mathrm{m}}$ are affected by the seasonal cycle, we use temperature anomaly $T_{2\mathrm{m}} := \tilde{T}_{2\mathrm{m}} - \mathbb{E}_y(\tilde{T}_{2\mathrm{m}})$ where $\mathbb{E}_y(\tilde{T}_{2\mathrm{m}})$ is the average of $\tilde{T}_{2\mathrm{m}}$ over many years for each calendar day, i.e. the climatology. We thus define the heatwave amplitude A as the space and time average of the temperature anomaly:

$$A(t) := \frac{1}{T}\int_t^{t+T}\left(\frac{1}{\mathcal{A}}\int_{\mathcal{A}} T_{2\mathrm{m}}(\vec{r},u)d\vec{r}\right)du, \tag{3.1}$$

where T is the duration in days of the heatwave and $\mathcal{A}$ is the spatial region of interest. Both parameters, T and $\mathcal{A}$ can be changed according to the event one wishes to study. In this work, T ranges from one day (short event) to one month (long event), but nothing prevents it from going even to longer, seasonal events. The region $\mathcal{A}$ typically extends over distances comparable to the synoptic scale, which, in the mid-latitudes, is about 1000 km. This is the order of magnitude of the spatial correlations in tropospheric dynamics, corresponding to the size of cyclones and anticyclones, and of the jet stream meanders. In this study we choose $\mathcal{A}$ to be the equivalent region of France, which is shown for instance in the last column of fig. 3.2. Moreover, as summer heatwaves have higher impacts, we consider only the months of June, July and August.

Following the studies [Jacques-Dumas et al., 2023; Miloshevich et al., 2023a], we define an extreme heatwave as an event for which the amplitude A exceeds a threshold

a corresponding to rare fluctuations. This threshold can be changed depending on the heatwaves of interest. In this work we will mainly focus on a defined as the 95^{th} quantile of the distribution of A, i.e. we consider as heatwaves the 5% most extreme events in our dataset. For a two-week heatwave, in the PlaSim model (see section 3.2.2), the threshold amounts to $a = 2.76\,\text{K}$. We will also comment briefly on heatwaves that are more or less rare than the 5% most extreme ones.

3.2.2 Datasets

In this work we use two datasets. The first is the output of the intermediate complexity climate model called PlaSim, the second is the ERA5 reanalysis data. We use PlaSim to generate an extremely long dataset, over which to train, optimize and test our Gaussian approximation framework (introduced in section 3.3) with little statistical errors. On the other hand, the simplicity of this climate model means that our results may suffer from potentially large biases with respect to the real climate. Hence, after this validation step we also apply our new methods to ERA5 data, which can be expected to suffer from smaller biases and be a more faithful representation of the actual climate.

PlaSim

The Planet Simulator (PlaSim) [Fraedrich et al., 2005a,b] is an intermediate complexity climate model that has a dynamical core that solves the moist primitive equations [Vallis, 2017] in the atmosphere. The model has a T42 horizontal resolution in Fourier space, that in direct space corresponds to a 64×128 grid of 2.8 degrees both in latitude and longitude, with 10 vertical layers and covering the whole globe. The model uses a relatively simplified parameterization of the sub-grid processes such as radiation, clouds, convection and hydrology over land. For the latter, in particular, PlaSim uses a single-layer bucket model [Manabe, 1969], with soil moisture increased by snow melt and precipitation and depleted by evaporation. Sea ice cover and ocean surface temperature are cyclically prescribed for each day of the year, acting as boundary conditions. By prescribing as well the greenhouse gases concentration and incoming solar radiation, the model is able to run in a steady state that reproduces a climate close to the one of the 1990s.

The fact that PlaSim lacks a dynamic ocean means that, in our study of heatwaves, we cannot investigate the effects of ocean related phenomena such as El Niño [Hafez, 2017; Zhou and Wu, 2016], or the North Atlantic Oscillation [Hafez, 2017; Li et al., 2020]. On the other hand, the representation of the atmosphere of PlaSim is sufficient to properly resolve the large scale dynamics of cyclones, anticyclones and the jet stream, including important teleconnection patterns relevant for heatwaves [Miloshevich et al., 2023c; Fraedrich et al., 2005b]. Moreover, the simplified parameterizations used in PlaSim allow it to run 100 times faster than the models used for CMIP studies, which makes it very suitable to obtain extremely long datasets. Here, we use a dataset consisting of 8000 years. It is the same data that was used for previous work on probabilistic forecast of heatwaves using machine

learning [Miloshevich et al., 2023a]. Mode details on the model setup can be found in [Miloshevich et al., 2023a].

As we will show, our proposed method for studying heatwaves does not need such a long dataset to achieve good performances. However, we also want to perform comparisons with alternative deep learning methods, and those do require as much data as possible [Miloshevich et al., 2023a].

PlaSim resolves the daily cycle and has an output frequency of 3 hours, but we are interested only in daily averages. In particular, we will focus on the anomalies (with respect to the daily, grid point-wise climatology) of 2 m temperature ($T_{2\mathrm{m}}$), 500 hPa geopotential height (Z_{500}) and soil moisture (S).

ERA5

In this manuscript we also present an application of our methodology to the ERA5 dataset [Hersbach et al., 2020]. We use daily data from the public available dataset of the European Centre for Medium-Range Weather Forecasts (ECMWF) service for summer seasons from 1940 to 2022. ERA5 has a resolution of 0.25 degrees in latitude and longitude. We use this fine resolution to compute the average 2 m temperature anomaly over France and hence the heatwave amplitude A eq. (3.1).

On the other hand, since the dataset is quite small, we reduce the number of predictors (see next section) by using only the 500 hPa geopotential height anomaly field and re-gridding it onto the coarser PlaSim grid.

An important remark is that in our study of heatwaves we assume a stationary climate. We thus need to remove the global warming signal from ERA5 data. This is achieved by means of a parabolic detrending of the averaged temperature over France and of zonal averages of the geopotential height. More technical details on the detrending procedure are given in section 3.9.1.

3.2.3 Predictors

To study heatwaves, we focus on a subset of climate variables that we call *predictors* and denote it with X. In particular, for a heatwave that starts at time t, we will be interested in the predictors $\tau \geq 0$ days before the event, i.e. $X(t-\tau)$.

For PlaSim, X will be the stack of the anomalies of 2 m temperature ($T_{2\mathrm{m}}$), 500 hPa geopotential height (Z_{500}) and soil moisture (S). The choice of $T_{2\mathrm{m}}$ is straightforward given its implication in heatwaves, and the potential of simple persistence and advection of temperature to be useful for prediction. The geopotential height anomaly at the middle of the troposphere (Z_{500}) is a good representation of the dynamical state of the atmosphere because of its relation with cyclones and anticyclones in the lower troposphere. At that height, the geostrophic approximation applies and thus Z_{500} gives also a good insight into the wind flow. Finally, it has been shown that low soil moisture acts as an important preconditioning factor for the occurrence of extreme summer temperatures in the mid-

latitudes, by limiting the evaporative cooling of the surface [Perkins, 2015; Miloshevich et al., 2023a; Benson and Dirmeyer, 2021; D'Andrea et al., 2006; Fischer et al., 2007; Hirschi et al., 2011; Lorenz et al., 2010; Rowntree and Bolton, 1983; Schubert et al., 2014; Shukla and Mintz, 1982; Stefanon et al., 2012; Vargas Zeppetello and Battisti, 2020; Zeppetello et al., 2022; Zhou et al., 2019; Vautard et al., 2007].

For the 2 m air temperature and 500 hPa geopotential height fields we will focus on the whole Northern Hemisphere (latitude above 30 degrees North), while soil moisture, instead, is a local variable, and we care only about the values on our region of interest (France). Considering the resolution of the PlaSim model, this will amount to a total of $d = 5644$ scalar predictors.

On the other hand, for ERA5 we use only the 500 hPa geopotential height anomaly field, which yields a total of $d = 2816$ pixels.

For both datasets, as it is commonly done in the machine learning community, we normalize each field value at each grid point independently dividing by its standard deviation. This way, X will be a collection of d (correlated) dimensionless variables with zero mean and unitary standard deviation, which also allows us to easily compare fields with different physical units.

3.3 Optimal Projection, Committor Functions, Composite Maps, and the Case of Gaussian Statistics

As climate scientists, concerned in understanding extreme events, we might ask two classes of questions. The first class is related to prediction or a priori statistics: given the current state of the system (the predictors X), what is the probability to observe an extreme event starting within τ days? The second class of question is related to a posteriori understanding: given that the extreme event actually occurred, what were the probabilities of the system states leading to this event? For instance, composite maps defined as the averaged state given that the event occurred, widely used by climate scientists, are examples of a posteriori statistics. Both a priori and a posteriori statistics are useful and important for the sake of understanding, but only a priori statistics is useful for prediction.

Indeed, the first goal of this section is to stress the difference between a priori and a posteriori statistics. For instance, it is key to understand that in general composite maps do not provide useful information for prediction. At the same time, we define some useful statistical quantities for prediction, namely the committor function (see a definition below). The second goal is to explain the difficulty to compute committor functions, motivating why they are not commonly used. The third and final goal is to devise predictive and simply interpretable statistical models, for instance the regression of the predictors (the state X) on the extreme event observable.

3.3.1 A Posteriori Statistics are Usually not Useful for Prediction

In this subsection we will stress the differences and the links between a posteriori and a priori statistics.

Let's consider two events, F and G, where G happens after F. We will denote with $\mathbb{P}(F|G)$ the *a posteriori* probability of F conditioned on the happening of the future event G. Vice versa, $\mathbb{P}(G|F)$ will be the *a priori* probability of G conditioned on the past event F. In our case the past event will be the predictors being in a particular state $X = x$, while the future event will be the realization of a heatwave $Y = 1$, where Y is the binary random variable

$$Y(t) := \begin{cases} 1 & \text{if } A(t) \geq a \\ 0 & \text{otherwise} \end{cases}, \tag{3.2}$$

and a is the threshold which defines a heatwave and will be the quantile of the distribution of A.

Bayes Formula

When comparing different conditional probabilities, we can make use of Bayes formula:

$$\mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1) = \mathbb{P}(X = x, Y = 1) = \mathbb{P}(Y = 1|X = x)\mathbb{P}(X = x), \tag{3.3}$$

where

- $\mathbb{P}(X = x, Y = 1)$ is the joint probability of being in state x and experiencing a heatwave ($Y = 1$)
- $\mathbb{P}(X = x) =: P_S(x)$ is the *stationary measure* of the predictors, namely the probability of being in state x
- $\mathbb{P}(Y = 1|X = x) =: q(x)$ is the a priori *committor function*: the probability of observing a heatwave, conditioned on being in state x
- $\mathbb{P}(Y = 1) = \int q(x)P_S(x)dx =: p$ is the unconditional (or climatological) probability of having a heatwave, inversely proportional to its *return time*, that tells us how extreme the event is.
- $\mathbb{P}(X = x|Y = 1)$ is the a posteriori probability that the state of the predictors were x given that the heatwave occurred.

Summarizing, Bayes formula clearly shows the difference and the relation between a priori and a posteriori statistics. In the next subsections we will illustrate a proper tool for the prediction task, namely the committor function, and we will illustrate for what composite maps can be used for, namely a posteriori statistics.

Definition of Committor Functions

If one is interested in a prediction task, the proper tool is the committor function $q(x)$, originally introduced in the field of stochastic processes (see section 3.9.5) for studying transitions between attractors [Bolhuis et al., 2000; Lucente et al., 2022a]. In our case we do not have two attractors, but rather a typical state of the climate with no heatwaves ($Y = 0$) and an atypical one ($Y = 1$). In this context the concept of transition gets a bit blurred, and the committor is simply the a priori conditional probability mentioned before. If we expand the notation and introduce back the lead time τ, we can write it as

$$q(x) = \mathbb{P}(A(t) \geq a \mid X(t-\tau) = x), \tag{3.4}$$

where a is the threshold used to define a heatwave. As we will discuss later, committors are extremely hard to compute properly and hence are quite rarely used in the field of climate sciences. However, they are *the* right tool for prediction, and even a very rough estimate of them is better than alternative methods.

Definition of Composite Maps

On the other hand, a commonly used tool in the climate community to study a wide range of events, including the extreme ones, is the *composite map* [Grotjahn and Faure, 2008; Sillmann and Croci-Maspoli, 2009; Teng et al., 2013; Ratnam et al., 2016; Miloshevich et al., 2023c; Noyelle et al., 2024]. It is defined as the average state of the climate τ days before the heatwave happened:

$$C := \mathbb{E}(X(t-\tau)|A(t) \geq a), \tag{3.5}$$

where $\mathbb{E}$ denotes an expectation over event realizations and a is the threshold used to define a heatwave. In practice one would estimate such expectation with an empirical average over all the heatwave events in the dataset, which makes the composite one of the easiest objects to compute and hence motivates its popularity.

It is important to point out that the empirical average will be a good estimate of the true composite provided that the number of heatwave events is enough. This means that, depending on the size of our dataset, a direct estimation of the composite map is useful only for not too rare (extreme) events, because of sampling errors.

Going back to the simpler notation used earlier, we can interpret the composite as the mean of the a posteriori probability distribution

$$C = \mathbb{E}(X|Y=1) := \int x\mathbb{P}(X = x|Y = 1)dx, \tag{3.6}$$

and thus, through Bayes theorem, we can relate it to the stationary measure P_S and the committor function q.

$$C = \int x\frac{\mathbb{P}(X = x)\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 1)}dx = \frac{\int xP_S(x)q(x)dx}{\int P_S(x)q(x)dx}, \tag{3.7}$$

Equation (3.7) clearly shows that the composite is the mean of a distribution proportional to $P_S(x)q(x)$ and thus not equivalent to $q(x)$. In particular, for rare events, we expect $q(x)$ to be peaked for very atypical values of x, namely in the tail of the stationary measure $P_S(x)$. Thus, the composite map may differ significantly from the typical states x associated with a high committor.

Two Simple Examples which Illustrate that Composites Might be Useless for Prediction

Now that we have defined the important quantities of interest, we will use some examples to highlight the difference between composites and committor, and in particular how the first may not give us any useful insights on the second.

As a first example, let us assume that our predictor is one dimensional ($X \in \mathbb{R}$), with stationary measure given by a simple normal distribution $P_S(x) \propto \exp\left(-\frac{x^2}{2}\right)$. Similarly, let the committor function be another Gaussian distribution centered in $x^* > 0$ and with standard deviation σ: $q(x) = \mathbb{P}(Y = 1|X = x) \propto \exp\left(-\frac{(x-x^*)^2}{2\sigma^2}\right)$. This means that the probability of a heatwave is maximum when we are in state $X = x^*$. We will now compute the composite, and show that it is different from x^*.

From eq. (3.7) we know that the composite is the mean of a distribution proportional to $P_S(x)q(x)$, and with some trivial algebraic manipulations, we find that

$$P_S(x)q(x) \propto \exp\left(-\frac{x^2}{2} - \frac{(x-x^*)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2}\left(1+\frac{1}{\sigma^2}\right)\left(x - \frac{x^*}{\sigma^2+1}\right)^2\right).$$

Hence, the composite is

$$C = \frac{x^*}{\sigma^2+1},$$

which is strictly smaller than the condition where the heatwave probability is highest. An important consequence is that the probability of having a heatwave when we are in the composite state may be vanishingly small depending on the values of x^* and σ, showing the low predictive power of the composite map:

$$\frac{q(C)}{q(x^*)} = \exp\left(-\frac{1}{2}\left(\frac{x^*}{\sigma+\sigma^{-1}}\right)^2\right).$$

As a second example, let us consider $X = (X_1, X_2) \in \mathbb{R}^2$ with $P_S(x)$ being a distribution that correlates the two components X_1 and X_2, for instance a bi-variate Gaussian with mean $(0,0)$ and covariance matrix $\begin{pmatrix}\sigma_1^2 & \phi \\ \phi & \sigma_2^2\end{pmatrix}$. We will then consider a committor $q(x) = q(x_1)$ that depends only on the first component. Without going into the details (available in section 3.9.3), it will be clear that the composite map will have a non-zero x_2 component, thanks to the correlation ϕ between x_1 and x_2. However, we know that the committor depends only on x_1, and so the composite will be misleading if we are interested in prediction, as it will draw our attention to variables that do not contain *any* information about the probability of having a heatwave.

In conclusion, the composite map is an average that takes into account both the probability of having a heatwave starting from state x *and* the probability of *being* in state x (eq. (3.7)). This is good to study the statistics of our extreme event, but if we want to know if there is going to be a heatwave tomorrow, we do not care how rare it was to have had today's weather.

3.3.2 Committor Functions and Optimal Projection

Now that we have a clear mathematical understanding of committor functions as the proper tool for prediction, we can move to the problem of computing them in practice. In this subsection we will point out why this is such a complex task as well as provide a way to evaluate how good any approximation of the true committor is. Finally, we will propose the framework of optimal projection of the committor, which will mitigate the problem of high dimensionality as well as make the committor much more interpretable.

Complexity of Committor Functions

The committor is a function that maps every point of the phase space x to a number $q(x)$ between 0 and 1 that quantifies the likelihood of having a heatwave. A naive way of estimating the committor would be to initialize many trajectories at the point x and count how many actually lead to a heatwave. This method is called direct numerical simulation, and, if rather inefficient, it is still doable for simple stochastic processes in low dimensional spaces.

In our case, however, $x \in \mathbb{R}^d$, with $d = 5644$ for PlaSim and $d = 2816$ for ERA5 and the dynamics is described by a rather complex climate model. One could argue that we do not need to explore the whole $\mathbb{R}^d$ space, but only the much lower dimensional manifold of *physical* states, which, under ergodic conditions, would be properly sampled by an extremely long trajectory. This argument is absolutely correct, but the task of a thorough and precise sampling of the committor still remains out of reach, even with the help of supercomputers.

Given the importance of committor functions, there is incentive in finding efficient ways to get a reasonable approximation of the committor, potentially also limiting the search to only the physical states that are most likely to yield a heatwave. This makes the task feasible, but far from simple, and attempts have been made using machine learning [Miloshevich et al., 2023a], rare event algorithms [Ragone et al., 2018] or both [Lucente et al., 2022b].

In this work, we strive to find an approach which is far simpler than all the aforementioned, yet still leads to a good enough approximation of the committor.

Evaluation of Approximations of the Committor Function

To quantify how good an approximation $\hat{q}$ of the true committor q is, we need a sort of distance between the two. Since committors are probabilities, the natural object to use is

the Kullback-Leibler divergence

$$KL(q,\hat{q}) = \int P_S(x)\left(q(x)\log\left(\frac{q(x)}{\hat{q}(x)}\right) + (1-q(x))\log\left(\frac{1-q(x)}{1-\hat{q}(x)}\right)\right)dx, \tag{3.8}$$

which quantifies the amount of information lost when using $\hat{q}$ instead of q. Expanding the logarithm and removing the terms that depend only on the true committor, we are left with the cross entropy loss.

$$CE(q,\hat{q}) = -\int P_S(x)\left(q(x)\log\hat{q}(x) + (1-q(x))\log(1-\hat{q}(x))\right)dx. \tag{3.9}$$

Now, since we do not have access to neither the true committor q nor the stationary measure $P_S(x)$, we can replace the first with the heatwave labels Y and the integral over the second with the average over our dataset $\mathcal{D}$. We obtain then the empirical cross entropy loss

$$\mathcal{L} = -\left\langle Y(t)\log\hat{q}(X(t)) + (1-Y(t))\log\left(1-\hat{q}(X(t))\right)\right\rangle_{(X(t),Y(t))\in\mathcal{D}}, \tag{3.10}$$

which is proven to be the only proper score for a probabilistic forecast [Benedetti, 2010].

$\mathcal{L}=0$ is the perfect prediction, but $\mathcal{L}$ can be arbitrarily large. To have a reference we can consider the climatological committor, that comes from assuming the only information we have is that we are studying the p-eth most extreme heatwave, for example setting the threshold a to be the 95^{th} quantile of the distribution of A means $p=0.05$. With only this information, the climatological committor is the constant p, and the associated empirical cross entropy is

$$\begin{aligned}\mathcal{L}_{\text{clim}} &= -\left\langle Y(t)\log p + (1-Y(t))\log(1-p)\right\rangle_{(X(t),Y(t))\in\mathcal{D}} \\ &= -p\log p - (1-p)\log(1-p).\end{aligned} \tag{3.11}$$

Finally, we can define the *normalized log score* $\mathcal{S}$ as in [Miloshevich et al., 2023a], that will quantify the skill of our prediction:

$$\mathcal{S} := 1 - \frac{\mathcal{L}}{\mathcal{L}_{\text{clim}}}. \tag{3.12}$$

A value $\mathcal{S}=1$ will mean a perfect prediction, namely $\hat{q}(t)=Y(t)\ \forall t$, and $\mathcal{S}<0$ will mean that our forecast is worse than the climatology.

Optimal Committor Projection

Now that we have the tools for evaluating committor approximations, we can tackle the problem of the high dimensionality of $q:\mathbb{R}^d\to[0,1]$. The key idea is to write a surrogate committor $q_\varphi = \tilde{q}\circ\varphi$, which first applies a projection $\varphi:\mathbb{R}^d\to\mathbb{R}^m$ to a space with dimension $m\ll d$, and then represents the committor in this reduced space with function $\tilde{q}:\mathbb{R}^m\to[0,1]$. We want to perform this decomposition in an *optimal* way, which means minimizing the cross entropy defined above, i.e., losing as little information as possible about the original committor.

It is relatively easy to see that, for a given projection function φ, the best committor representation is the average of the original committor on the iso-levels of φ

$$\tilde{q}^*(f) = \mathbb{E}_{x\in\varphi^{-1}(f)} q(x). \tag{3.13}$$

Moreover, the information loss comes from mapping very different values of the original committor onto the same iso-level. Ideally, then, the optimal projection would be the one that has the same iso-levels of q, namely q itself (up to any monotonic rescaling). Of course this is not desirable, as we simply shifted the problem from computing q to computing φ. To have something useful, we need to constrain the search space of φ, for example to linear maps.

Even with these simplifications, the general problem remains hard to treat in practice. In the next subsection, we will show the case of Gaussian statistics, which gives an analytic way to compute the optimal linear projection, as well as the reduced committor.

3.3.3 The Case of a Joint Gaussian Distribution

In this section we present the theory for what we call the Gaussian approximation. We describe the theoretical idea and derive analytically the expressions for the composite map and the committor function.

The Gaussian approximation consists in assuming that the predictor X at time $t-\tau$ and the heatwave amplitude A at time t follow a jointly Gaussian distribution

$$(X(t-\tau), A(t)) \sim \mathcal{N}\left(0, \Sigma(T,\tau)\right), \tag{3.14}$$

where X is thought of as a d-dimensional vector, and represents all grid-point values of either a single field or stacked fields. The joint distribution has mean zero because both X and A are anomalies, and it is then solely characterized by the $d+1$ dimensional covariance matrix $\Sigma(T,\tau)$, that depends on the heatwave duration and the lead time.

To simplify the notation, we assume that we work at fixed T and τ, and thus drop the dependencies on them. We can then write Σ as a block matrix of the form $\begin{bmatrix}\Sigma_{XX} & \Sigma_{XA}\\ \Sigma_{AX} & \Sigma_{AA}\end{bmatrix}$, where $\Sigma_{XX} = \mathbb{E}(XX^\top)$ is the $d\times d$ covariance matrix of X, $\Sigma_{XA} = \Sigma_{AX}^\top = \mathbb{E}(XA)$ is the $d\times 1$ correlation map between X and A and $\Sigma_{AA} = \mathbb{E}(A^2)$ is the scalar variance of A.

Composite Maps Within the Gaussian Approximation

Under the Gaussian assumption, the composite map can be computed analytically as

$$C_{\mathcal{G}} = \mathbb{E}[X|A\geq a] = \int x \frac{\int_a^{+\infty}\mathbb{P}(x,A)dA}{\int_a^{+\infty}\mathbb{P}(A)dA}dx = \eta\left(\frac{a}{\sqrt{2\Sigma_{AA}}}\right)\frac{\Sigma_{XA}}{\sqrt{\Sigma_{AA}}}, \tag{3.15}$$

with

$$\eta(z) = \sqrt{\frac{2}{\pi}}\frac{e^{-z^2}}{\operatorname{erfc}(z)}, \tag{3.16}$$

where erfc(•) is the complementary error function and the subscript $\mathcal{G}$ reminds that the composite is evaluated under the Gaussian assumption. The detailed computation is shown in section 3.9.4.

From eq. (3.15), we can clearly see that the composite is directly proportional to the correlation map, with the proportionality constant depending only on the threshold a. This has the important implication that the average state of the climate τ days before a heatwave looks like the τ-lagged correlation between the fields and the heatwave amplitude, *regardless* of how extreme the heatwave is. In other words, the composite of a more extreme event has exactly the same pattern as a less extreme one, but amplified according to the function η. The fact that we do observe this effect in the actual data (fig. 3.1) suggests a good validity of our Gaussian approximation. We test it more thoroughly in section 3.4. Moreover, it gives us access to composites of very extreme events, where the direct estimation as the average over the (very small) heatwave set would suffer from huge sampling errors. On the other hand, the correlation map Σ_{XA} is estimated on the whole dataset and thus does not have this issue.

The function η is plotted in fig. 3.14, and has the interesting property that $\eta(z) \sim \sqrt{2}z$ as $z \to \infty$, which means that for very extreme heatwaves the composite map tends to the simple linear regression of X against A.

$$C_{\mathcal{G}} \xrightarrow[a \gg \sqrt{\Sigma_{AA}}]{} a\frac{\Sigma_{XA}}{\Sigma_{AA}} = a\xi, \quad \xi = \arg\min_{\xi} \mathbb{E}\left((X - A\xi)^2\right). \tag{3.17}$$

Committor Functions Within the Gaussian Approximation

By definition, the committor is the integral of the a priori distribution of A conditioned on knowing X:

$$q(x) = \mathbb{P}(A \geq a | X = x) = \int_a^{+\infty} \mathbb{P}(A = a | X = x) dA. \tag{3.18}$$

Under the assumption of a joint Gaussian distribution for (X, A), the conditional distribution of A given X is also Gaussian. In particular, it has mean $\mu(x)$ that scales linearly with x and constant variance σ^2:

$$\mu(x) = \Sigma_{XX}^{-1}\Sigma_{XA} \cdot x, \qquad \sigma^2 = \Sigma_{AA} - \Sigma_{AX}\Sigma_{XX}^{-1}\Sigma_{XA}. \tag{3.19}$$

For the details of this computation see section 3.9.4. In fact, $\mu(x) = \tilde{M}^\top x$ is precisely the linear regression of A against X:

$$\begin{aligned} \tilde{M} &:= \Sigma_{XX}^{-1}\Sigma_{XA} = \arg\min_M \left(M^\top \Sigma_{XX} M - 2M^\top \Sigma_{XA}\right) \\ &= \arg\min_M \mathbb{E}\left((A - M^\top X)^2\right). \end{aligned} \tag{3.20}$$

Then, to obtain the full committor, we just have to compute the Gaussian integral in eq. (3.18), which gives

$$q_{\mathcal{G}}(x) = \frac{1}{2}\mathrm{erfc}\left(\frac{a - \tilde{M}^\top x}{\sqrt{2}\sigma}\right). \tag{3.21}$$

This result can be viewed in light of the framework of optimal committor projection presented in section 3.3.2. In this case, the optimal projection of the high dimensional committor is onto the normalized projection pattern

$$M = \frac{\Sigma_{XX}^{-1}\Sigma_{XA}}{|\Sigma_{XX}^{-1}\Sigma_{XA}|}, \tag{3.22}$$

which condenses all the important information of the high dimensional vector x into the scalar variable $f = M^\top x$. Then the committor in the projected space is simply

$$\tilde{q}(f) = \frac{1}{2}\mathrm{erfc}\left(\alpha + \beta f\right), \tag{3.23}$$

with

$$\alpha = \frac{a}{\sqrt{2}\sigma}, \quad \beta = -\frac{|\tilde{M}|}{\sqrt{2}\sigma}. \tag{3.24}$$

The two operations of linear projection and reduced committor can also be viewed as the architecture of a simple one layer perceptron with the custom activation function $\tilde{q}$. In comparison to other neural network architectures (such as convolutional ones) that may be trained on the same task [Miloshevich et al., 2023a], this approach is far simpler, and depends on a much smaller number of parameters.

In addition, we would like to stress that the method is *interpretable* by design: with complex neural networks one may need sophisticated explainable AI techniques to understand *why* they are outputting a particular probability [McGovern et al., 2019; Toms et al., 2020; Delaunay and Christensen, 2022], while in our case the answer is straightforward, namely, it is computing the optimal index f. Furthermore, since the projection pattern M has the same dimension as the predictor X, we can plot it as a map, representing the relative importance of each pixel in our predictor, and providing potential insight in the physical dynamics leading to extreme heatwaves.

Another interesting point to pay attention to is the difference of the two linear regressions for the composite (eq. (3.17)) and for the committor (eq. (3.20)). In the first case, we are doing d independent linear regressions of each pixel in X against the heatwave amplitude A, while for the committor we have a single optimization, regressing A against X. This shows once again the fundamental difference between a posteriori and a priori statistics.

In the following sections, we apply the Gaussian approximation to actual data, see to what extent the assumption of Gaussianity holds and what useful information we are able to extract.

3.4 Validation of the Gaussian Approximation for the Computation of Composite Maps for Extreme Heatwaves

Composite maps are very interesting to understand weather situations that actually led to extreme events (a-posteriori statistics). They are actually defined as the average of weather variables conditioned on the future occurrence of the extreme event.

In section 3.4.1 we show and compare qualitatively composite maps evaluated empirically and using the Gaussian approximation. In section 3.4.2 we quantify the error made under the Gaussian approximation, and we distinguish systematic and sampling errors. Subsequently, in section 3.4.3, using the Gaussian approximation, we give an explanation of the puzzling independence of the empirical composite maps patterns from the threshold a used to define an extreme heatwave. Finally, in section 3.4.4 we discuss in more detail the effect on the quality of the Gaussian approximation of both the dataset length and the threshold defining extreme events, and conclude that the Gaussian approximation is the best way to estimate composite maps in a regime of lack of data.

In section 3.6, we will use these results to make a physical analysis of extreme events, by varying the heatwave duration T and the lead time τ.

In this section we use the PlaSim dataset with 8000 years of data and predictors $X = (T_{2m}, Z_{500}, S)$ (see section 3.2.2). We show an application of our methodology to the ERA5 dataset in section 3.7.

3.4.1 Comparing Empirical Composite Maps with Composite Maps Computed Within the Gaussian Approximation

We now compare the composite maps computed either directly from the data or using the Gaussian approximation, showing that the two are qualitatively very similar, with a relative error of the order of 20%. We consider 14-day heatwaves ($T = 14$), looking at the composites for the first day of the heatwave (lead time $\tau = 0$) and we first focus on the 5% most extreme heatwaves ($a = 2.76\,\mathrm{K}$).

Composite maps C are averages of the predictors X conditioned on the occurrence of a heatwave: $C = \mathbb{E}[X(t)|A(t) \geq a]$ (see section 3.3.1). We first estimate this conditional expectation as an empirical average

$$C_{\mathcal{D}} = \frac{1}{N}\sum_{\mu=1}^{N} x_\mu, \quad \text{where } \{x_\mu\}_{\mu=1}^{N} = \{X(t-\tau)|A(t) \geq a\}.$$

Figure 3.2 shows the empirical composite maps for the three predictors X (top row). We observe a positive anomaly of both 2 m temperature and 500 hPageopotential height over France and Western Europe, which is expected since we are conditioning over events that are happening over the French region. In the PlaSim grid, France is identified as the 12 pixels shown for the soil moisture field. Soil moisture anomaly displays negative values, as the soil tends to be drier than usual when heatwaves happen. In the rest of the Northern Hemisphere, we see teleconnection patterns in the temperature and geopotential height field, in particular a cyclone over Greenland and an anticyclone over the mid and eastern United States.

All these important features are also visible in the composite map $C_{\mathcal{G}}$ computed with the Gaussian approximation (using eq. (3.15)), represented in fig. 3.2 (middle row), to the point that the only visible discrepancy with the empirical map is slightly darker shades of soil moisture. Indeed, if we take the difference between the two estimates of the composite

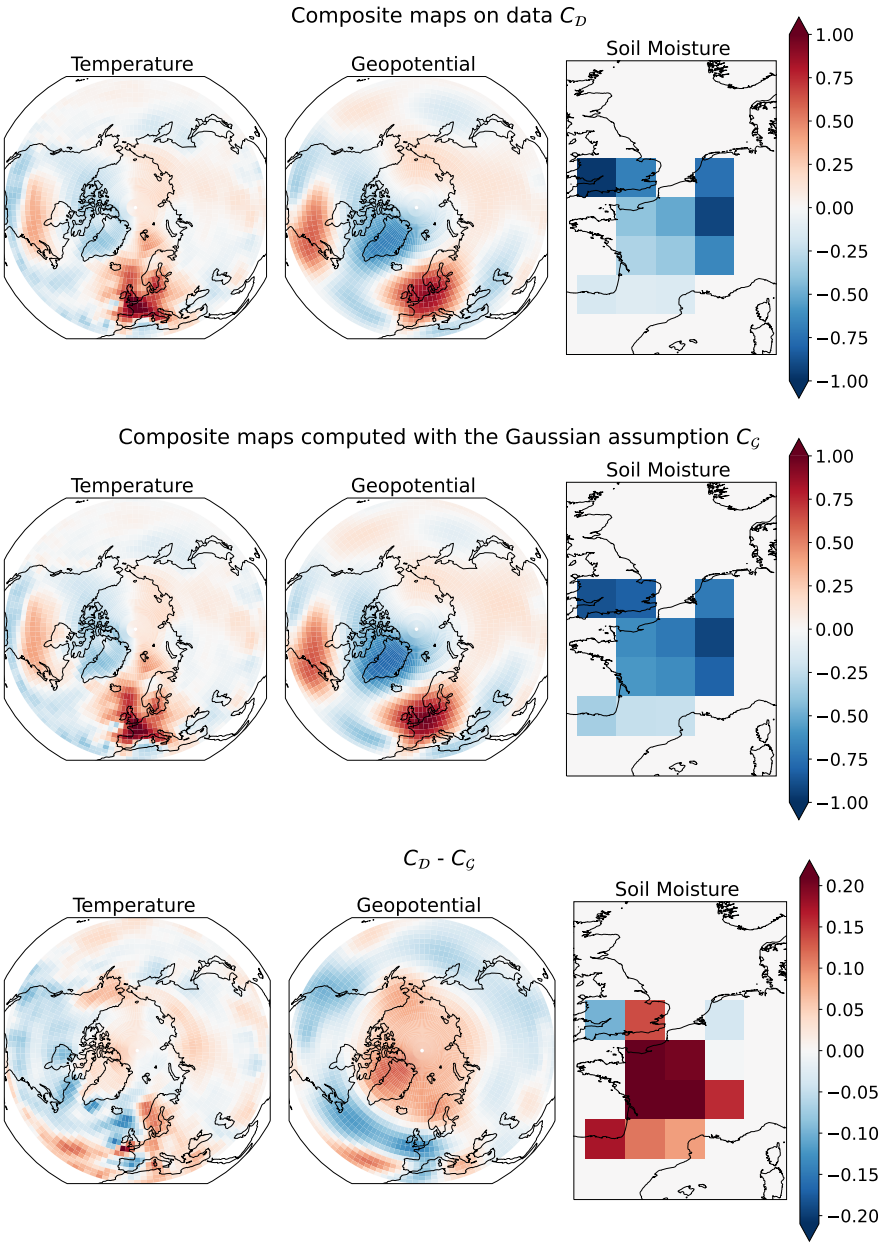


Figure 3.2: Composite maps of normalized 2 mtemperature, 500 hPa geopotential height, and soil moisture anomalies, conditioned on events with the 5% most extremes 14-day temperature over France. Composite maps are computed either directly from PlaSim data (first line), or under the Gaussian approximation (second line). The third line shows the difference between the two. The salient features of both temperature and geopotential are well captured by the Gaussian approximation, with errors of the order of 20% at most.

(fig. 3.2, bottom row), most of the weight is concentrated on the soil moisture field. However, non-trivial patterns are also visible in the temperature and geopotential fields. The latter, in particular, shows a wave zero pattern, with positive values around the polar region and negative ones in the mid-latitudes. The amplitude of the difference, read on the color bar, is on the order of 20% of the amplitude of the composite. To have a more quantitative measure, we compute the ratio $\mathcal{R}$ between the L2 norms of the difference between the two composites and the empirical one:

$$\mathcal{R} = \frac{|C_{\mathcal{D}} - C_{\mathcal{G}}|}{|C_{\mathcal{D}}|}. \tag{3.25}$$

In evaluating the norms, we took into account that we consider grid-cells of different areas. For the parameters considered in this section, the norm ratio is $\mathcal{R} = 0.21$, in agreement with our visual estimate (in section 3.6 we will investigate how this metric varies with the heatwave duration T and the lead time τ).

In the next section, we analyze in more detail the sources of the difference between the two estimates. We will then give an explanation of the striking independence of the pattern from the extreme event threshold a in section 3.4.3.

3.4.2 Quantification of the Quality of the Gaussian Approximation for Composite Maps of Extreme Heatwaves

In the previous section, we showed that the empirical composite map $C_{\mathcal{D}}$ and the Gaussian composite map $C_{\mathcal{G}}$ differ at most by 20% (fig. 3.2, bottom row). A natural interpretation of this difference is that it is an error due to the fact that the Gaussian assumption is not exactly satisfied, and therefore the Gaussian composite map is only an approximation of the true composite map. Indeed, we can investigate the validity of this assumption by visualizing the joint and marginal distributions of the heatwave amplitude A and the predictors at the grid-point level, for regions of low or high error (see section 3.9.9). For instance, we show in fig. 3.14 that the assumption is poorly satisfied for soil moisture at a grid point over France, where the error is large, while it is a much better assumption for geopotential over Greenland, where the error is small.

However, another source of discrepancy between the two composites is the sampling error affecting $C_{\mathcal{D}}$ due to the limited number of heatwaves in the dataset over which we perform the empirical average. Indeed, if we focus on a single pixel i, and call $\{x_\mu\}_{\mu=1}^{N} = \{X^i(t-\tau)|A(t) \geq a\}$ the subset of heatwave events, the central limit theorem tells us that

$$\sqrt{N_{\text{eff}}}\frac{C^i - C^i_{\mathcal{D}}}{\sigma(C^i_{\mathcal{D}})} \xrightarrow[N\to\infty]{} \mathcal{N}(0,1), \tag{3.26}$$

where C^i is the true composite, $C^i_{\mathcal{D}} = \frac{1}{N}\sum_{\mu=1}^{N} x_\mu$ is the empirical one,

$$\sigma(C^i_{\mathcal{D}}) = \sqrt{\frac{1}{N}\sum_{\mu=1}^{N}(x_\mu - C^i_{\mathcal{D}})^2}$$

is the standard deviation of the heatwave set and N_{eff} is the number of *effectively* independent heatwaves. If all the x_μ were actually independent, we would have $N_{\text{eff}} = N$, but from our definition of heatwave (eq. (3.1)), it is very likely that a series of consecutive days will be all heatwave events, and thus far from independent. In this paper we decide to fix N_{eff} to the number of years with at least one heatwave (equals to 2627 years for 5% most extreme heatwaves of duration $T = 14$ days and lead time $\tau = 0$). The motivation beside this choice can be found in section 3.9.8.

Equation (3.26) tells us, then, that the distance between the empirical composite and the true one will be of the order of $\frac{\sigma(C^i_{\mathcal{D}})}{\sqrt{N_{\text{eff}}}}$, and thus if the Gaussian composite $C^i_{\mathcal{G}}$ falls much farther than $\frac{\sigma(C^i_{\mathcal{D}})}{\sqrt{N_{\text{eff}}}}$ from the empirical one, we can safely say that is also far from the true composite. In other words, we can define the statistical significance of the error we make as

$$s_i = \frac{\sqrt{N_{\text{eff}}}|C^i_{\mathcal{G}} - C^i_{\mathcal{D}}|}{\sigma(C^i_{\mathcal{D}})}. \tag{3.27}$$

To obtain a global metric for the whole composite map, we can consider the fraction of area $\mathcal{F}$ that have a significance above 2. This allows us to say that, with 95% confidence, a fraction $\mathcal{F}$ of the region of interest has a systematic error, not explainable by the finite size effect of the empirical composite. For the parameters studied here, we obtain the value $\mathcal{F} = 0.37$ (in section 3.6 we will investigate how this metric varies with the heatwave duration T and the lead time τ).

This allows us to conclude that the Gaussian composite suffers from a statistically significant error over roughly half the domain. In spite of this, it gives a reasonable approximation of the empirical composite, within an error of order 20%. However, having 8000 years of data to work with is not common in the climate community, especially when working with observational data or complex model simulations, and we can expect that when data is scarce, the error due to the Gaussian approximation becomes smaller than the sampling error in the empirical composite. In section 3.4.4 we will address this point on the dataset length and identify a regime where the Gaussian composite gives a better estimation of the true one than the empirical composite.

3.4.3 Composite Maps do not Depend Much on the Extreme Event Threshold

This section aims firstly at giving an explanation for the striking independence of composite maps pattern from the threshold a. Secondly, we show how the norm of the empirical composite maps scales with the threshold a and that this scaling is very close to the one predicted from the Gaussian composite.

In section 3.3.3 we explained that the composite map pattern does not depend on the extreme event threshold a. The independence of the pattern of the empirical composite maps from the threshold a is explained by the Gaussian composite, eq. (3.15). In this equation we see that the threshold intervenes only in the scaling of the pattern and not on the structure of the pattern itself, which is precisely what we observe in the estimated

composite maps. Indeed, in fig. 3.15, we show the difference between the empirical composite and the Gaussian one (evaluated using eq. (3.15)) for the three fields, namely (from the left) 2 m air temperature anomaly, 500 hPa geopotential height anomaly and soil moisture anomaly evaluated for a corresponding to the 1% most extreme temperature 14-day anomaly of A for PlaSim dataset. As predicted by the theory, the observed 500 hPa geopotential height pattern is the same as the one from fig. 3.2. To give a quantitative measure of the error, in fig. 3.16, we evaluate the error using the norm ratio defined in eq. (3.25) for different thresholds a, showing that the error is around the 20%, thus of the same amplitude of the one obtained for a threshold at 5%.

A natural follow-up question regards the scaling presented in eq. (3.15). In fig. 3.3 we plot the norm of the empirical composite maps as a function of the threshold a. The gray line corresponds to the total one, the colored lines are the field-wise norms. The dashed line represents the theoretical scaling η of eq. (3.15). The behavior is very well captured by the 2 m air temperature anomaly, and less well captured by the soil moisture anomaly field. The departure of the empirical scaling from the theoretical one for large values of a might be also due to sampling error.

Due to independence of the composite maps on the parameter a we will omit the sensitivity analysis of this parameter in favor of the other two, which are proven to provide different responses for heatwaves, namely the heatwave duration T and the lead time τ (see section 3.6).

3.4.4 Effect of Dataset Length on Estimation of Composite Maps

This section aims at motivating the usage of the Gaussian composite when the estimation of the true composite is highly affected by sampling issues, i.e. when we are in a regime of scarcity of data. For datasets' length of 200 years, the same order of magnitude of ERA5 reanalysis dataset, the Gaussian composite performs much better than the empirical one, for events more extremes than 5%.

Firstly, we use the empirical composite $C_{\mathcal{D}}$ computed on the whole 8000 years dataset as an estimate of the true composite. Then we take a subset $\mathcal{P}$ of our data and compute over it the empirical composite $C_{\mathcal{P}}$ and the Gaussian one $C_{\mathcal{G}}^{\mathcal{P}}$.

In fig. 3.4 we see the values of the empirical norm ratio $\mathcal{R}_{\mathcal{P}} = \frac{|C_{\mathcal{P}} - C_{\mathcal{D}}|}{|C_{\mathcal{D}}|}$ (solid lines) and the Gaussian one $\mathcal{R}_{\mathcal{G}} = \frac{|C_{\mathcal{G}}^{\mathcal{P}} - C_{\mathcal{D}}|}{|C_{\mathcal{D}}|}$ (dashed lines), for datasets $\mathcal{P}$ of different lengths. To get confidence intervals, we repeat the experiment 8 times for each dataset length, with 8 independent batches of data.

The Gaussian composites over 1000 years and over 200 years of data show a monotonic increase (in log scale) as function of the heatwave threshold a. The latter shows a plateau for values of p ranging from 50% to 1%, meaning that the error made for typical events is comparable to fairly extreme ones. This is not valid for the composite over 1000 years as there is a constant and more rapid worsening of the Gaussian norm ratio. It is interesting to notice that in the very tail of the distribution of A, thus for small values of p, we achieve very similar values of the norm ratio in both datasets. The spread of the norm ratio among

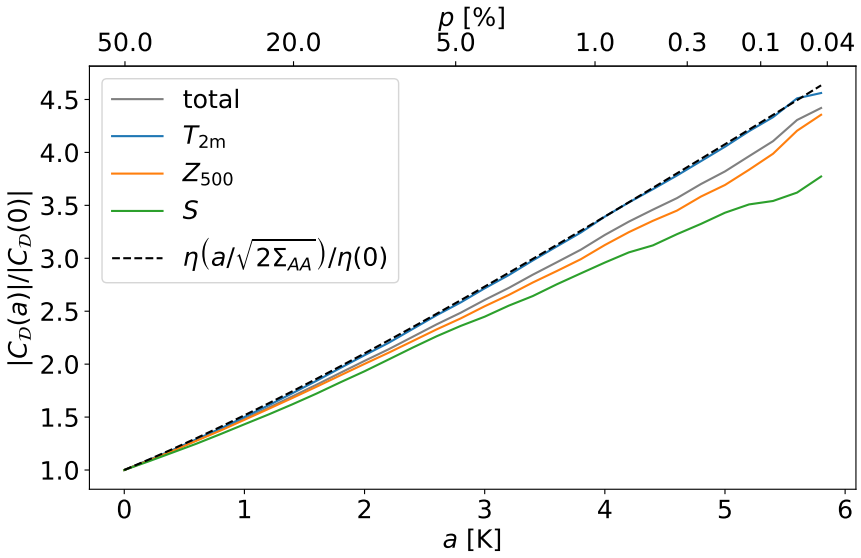


Figure 3.3: Normalized norm of the empirical composite map as function of a, the threshold to define a heatwave. The normalization is the norm of the empirical composite evaluated at $a = 0$. The dashed line represents the theoretical scaling of the composite maps using the Gaussian approximation, see eq. (3.15). The scaling of the empirical composite is not far from the Gaussian one (gray curve) and it is precise for the 2 m air temperature (blue curve). The bottom of x-axis a is the threshold value used to define a heatwave event from the distribution of the temperature anomaly over France, A, for heatwaves of 14 days of duration. On the top of the x-axis, p is the respective percentile value corresponding to a given a.

the batches is more pronounced for less extreme events than for the most extreme ones. In the case of the Gaussian composite, the main source of error is systematic, as we use the full dataset $\mathcal{P}$ to evaluate the Gaussian composite and not a small subset which depends on the threshold (eq. (3.15)).

The empirical composite norm ratio for 200 years of data stays almost constant until $p = 5\%$, after which it starts increasing both in the mean and in the spread of data. For the empirical composite norm ratio over 1000 years we see a less evident constant behavior and a more pronounced minimum of the norm ratio around $p = 5\%$, both in the mean and in the standard deviation. Similar to the 200 years line, there is a worsening of the norm ratio as a increases. It is remarkable that both composites for very small values of p never attain the same value as it happens for the Gaussian ones. Indeed, there is always a constant gap between the two solid lines. As we select fewer and fewer data on the right side of the plot, we see an increase of the spread of the data, mostly due to sampling issues.

Focusing on both composites for 200 years datasets, until $p = 5\%$ both the empirical and the Gaussian have the same values of the norm ratio. For more extreme events, the norm ratio of the empirical one increases drastically, mostly due to the more and more limited data available in the tail, reaching 100% of error at the 0.3% most extreme heatwaves. This is not the case for the Gaussian approximation, whose values of the norm ratios still

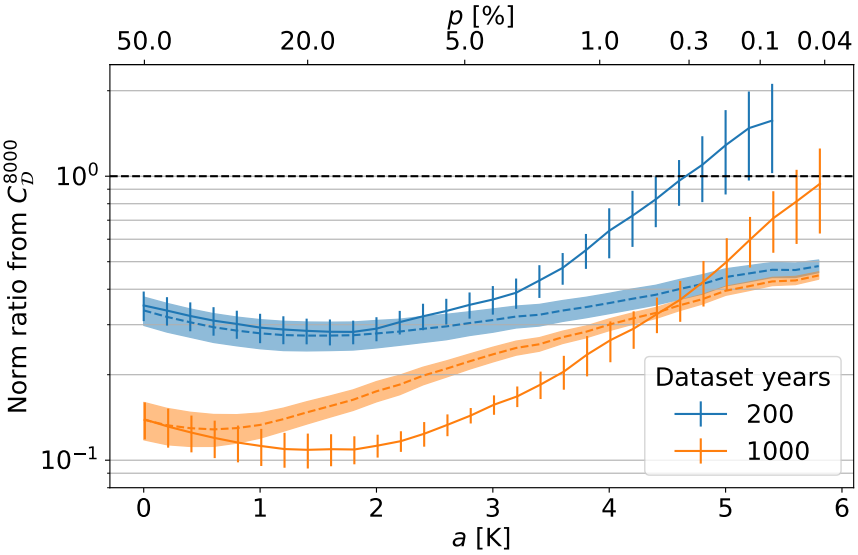


Figure 3.4: Norm of the relative error of the conditional average (composite map) using the Gaussian approximation (dashed lines with shading) and of the composite maps using only a part of the full PlaSim dataset (solid lines with error bars). The error is relative to the empirical conditional average (composite map) evaluated over the full PlaSim dataset of 8000 years. The orange color indicates a 200 years long dataset. The blue color a 1000 years one. Shading or error bars indicates the one standard deviation spread obtained from 8 independent batches of either 200 or 1000 years. On the bottom of x-axis a is the threshold value used to define a heatwave event from the distribution of the 2 m temperature anomaly over France, A. On the top of the x-axis, p is the respective percentile value corresponding to a given a. The higher its value, the lower the value of the threshold a, the less extreme are the heatwaves considered. The relative error for dataset of 1000 years is always lower than the one obtained for 200 years simply because of higher amount of available data. The difference is more remarkable, and stays quite stable as a increases, in the relative error obtained with the empirical composite than with the Gaussian approximation. This is not surprisingly because the Gaussian composite uses the information of the full dataset, not just of the subset of the heatwave events (see eq. (3.15)). All the curves show an increase in the relative error as a increases due to the lack of data. When we are in this regime, the relative error obtained with the Gaussian composite is lower than the one obtained with the empirical composite. This happens for a p value of around 0.2% for datasets of 1000 years length, and of 5% for datasets of 200 years length. For less extreme events, the Gaussian composite performs worst or similarly than the empirical one.

increase but much more slowly. Here we can see the power of the Gaussian approximation on smaller datasets.

Indeed, when we are in a regime of scarcity of data, which naturally arises when one wants to study very extreme heatwaves, calculating composite maps using empirical data poses sampling issue. Our methodology overcomes this issue by relying on an estimate of the composite which uses the whole dataset. To confirm this, we see that on longer datasets, such the 1000 years one, where we already have a sufficient amount of data to have a good estimate of the empirical composite, the Gaussian approximation is not a better estimate than computing the composite directly. At least for events up to $a = 4.5\,\mathrm{K}$, after which, due to the sampling issue, the Gaussian estimation performs better than the empirical composite.

3.5 Validation of the Gaussian Approximation for Computing Committor Functions on Climate Datasets

In section 3.3 we defined committor functions and optimal projection patterns, both generally and within the Gaussian approximation. In this section, we apply the Gaussian approximation of the committor on climate data, the PlaSim dataset described in 3.2.2, and compare its skill with the prediction from a neural network. We then proceed to study the optimal projection pattern, which is given by eq. (3.22). However, we will see in this section that the mathematical expression eq. (3.22), is not directly applicable to high dimensional climate data, where the datasets are usually too short. Indeed, in section 3.5.2 we show that regularization is necessary to have physically meaningful projection patterns. In sections 3.5.3 and 3.5.4 we will show the effect of lack of data on the performance. In the first case lack of data will come from reduced dataset lengths, and in the second from more extreme events.

We illustrate this for the task of predicting heatwaves, but we assume it will generalize well to other prediction problems in climate.

3.5.1 Skill of the Gaussian Approximation Compared to Prediction with Neural Networks

We first apply the Gaussian approximation of the committor, defined in eq. (3.21), to the forecast of the 5% most extreme two week heatwaves ($T = 14$), predicted at lead time $\tau = 0$, using the full PlaSim dataset. To have a robust estimate of the performance of our method, we repeat the experiment 10 times in a k-fold cross validation process (see section 3.9.2). Doing so we get an average validation normalized log score of 0.455 ± 0.010. We can say that the score is much better than the climatology ($S = 0$), but it is very tricky to quantify the maximum achievable score, as $S = 1$ is absolutely unrealistic due to the chaotic nature of the climate system.

However, we can compare to other methods, for instance the prediction using a deep convolutional neural network [Miloshevich et al., 2023a]. This network takes as input

the stack of predictors and produces an estimate of the committor. It is trained on a probabilistic binary classification of the labels Y, i.e. it directly minimizes the loss $\mathcal{L}$ defined in eq. (3.10). More details about the network's architecture can be found in [Miloshevich et al., 2023a]. Such a network yields a validation score of $S_{CNN} = 0.465 \pm 0.007$.

This is a remarkable result, as the Gaussian approximation is much simpler than a deep neural network, but is able to achieve a result that is only 2% (or less than a standard deviation) worse.

3.5.2 Regularization of the Projection Pattern

The simplicity of the Gaussian committor comes with the added benefit of being an interpretable forecast, as we can look at the projection pattern M to obtain some insight into the dynamics leading to a heatwave.

Unfortunately, a direct plot of M looks like the first row of fig. 3.5, from which we cannot extract any meaningful information as no well-defined patterns emerge. This is due to the fact that the covariance matrix Σ_{XX} is very high dimensional ($d^2 \sim 10^7$) and is estimated with a relatively low number of data points ($8000 \times 0.9 \times (90 - T + 1) \sim 10^6$). Hence, it will be nearly singular, causing problems when we compute the inverse in eq. (3.22).

A simple solution is the standard Tikhonov regularization, that corresponds to adding an L_2 penalty to the minimization problem:

$$M_\epsilon \propto (\Sigma_{XX} + \epsilon \mathbb{I})^{-1} \Sigma_{XA} = \arg\min_M \left((A - M^\top X)^2 + \epsilon |M|^2 \right), \tag{3.28}$$

where $\mathbb{I}$ is the identity matrix.

However, in our case we can better enforce interpretability of the pattern M by requiring it to be spatially smooth. Namely, we will penalize the squared norm of the spatial gradient, H_2, that we can compute as the weighted sum of the square differences between values of adjacent pixels in the map M. We can then write $H_2(M) = M^\top W M$ (see section 3.9.6 for the exact formula of matrix W), and hence the regularized pattern will be

$$M_\epsilon \propto (\Sigma_{XX} + \epsilon W)^{-1} \Sigma_{XA} = \arg\min_M \left((A - M^\top X)^2 + \epsilon H_2(M) \right). \tag{3.29}$$

Note that if we tweak the projection pattern M, we should also update the formulas for the coefficients α and β in eq. (3.24). This is relatively straightforward and is discussed in section 3.9.7.

Varying ϵ yields the different maps shown in fig. 3.5, where indeed we see that the regularization makes the patterns progressively smoother. Unsurprisingly, we note that a higher regularization comes at the price of a lower skill score S (see also table 3.1). It is then up to the user to decide what is a good compromise between performance and interpretability of the pattern. In our case, we argue that the best pattern is the one in the center row of fig. 3.5 ($\epsilon = 1$), as it is smooth enough that we can see some clear structures in the 500 hPa geopotential height field, while a higher regularization does not improve its physical understanding. At this value of the regularization coefficient, the

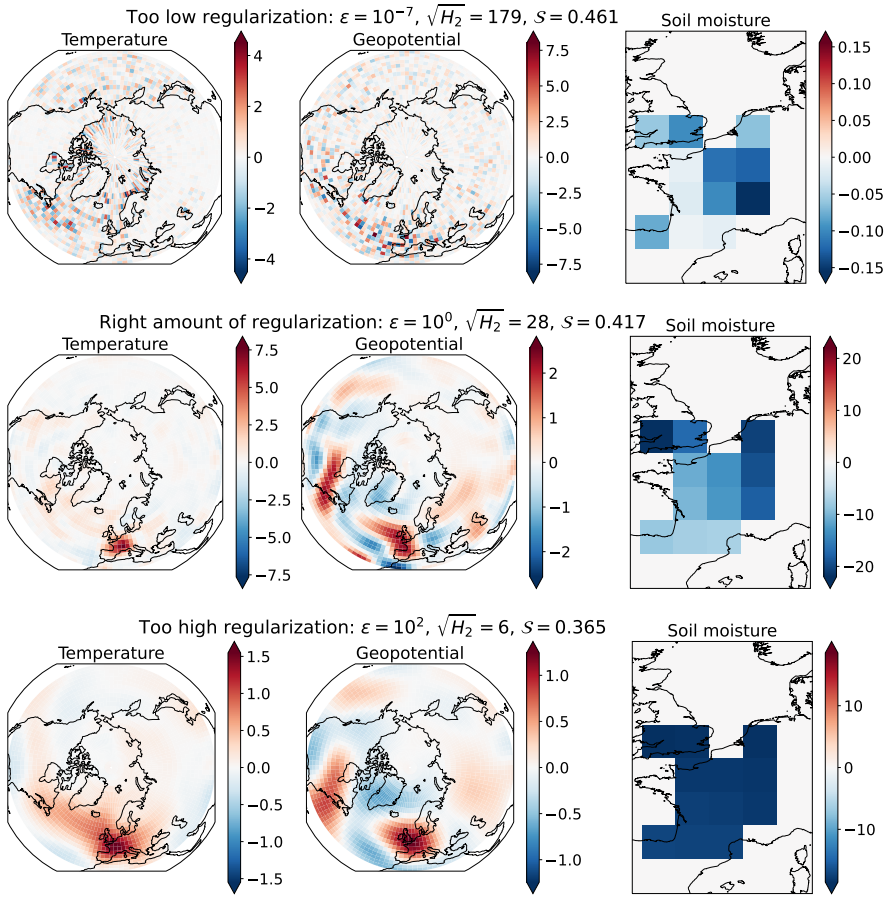


Figure 3.5: Choice of regularization for optimal physical content of the projection map M, using PlaSim data. Each line features the projection map M at different values of the regularization coefficient ϵ. Each map M is represented as its three field components: 2 m air temperature, geopotential height at 500 hPa and soil moisture anomalies (trained on 7200 years of data, for one of the 10 folds). On top of the figures we report the values of ϵ, of the norm of the gradient $\sqrt{H_2}$ and of the normalized log score $\mathcal{S}$. The intermediate value, $\epsilon = 1$, is the best compromise with a very high predictive skill and an excellent readability of the physical fields.

average validation score is 0.418 ± 0.006: three standard deviations or 8% worse than the non-regularized case, and five standard deviations or 10% worse than the neural network.

It is important to point out that after proper regularization the skill of the prediction is still much better than climatology, while providing physical insight on the dynamics leading to heatwaves. This latter point is further discussed in section 3.6.

3.5.3 Performance on Smaller Datasets

So far we have applied the Gaussian approximation to an extremely long 8000 year dataset. Such datasets are uncommon in the climate community, especially when dealing with observations or high resolution simulations. To study the effect of the amount of data on the performance of our method, we apply it to gradually smaller and smaller subsets of our climate model output.

In the left panel of table 3.1, we can see the behavior of the normalized log score S of the Gaussian committor, as a function of the regularization coefficient and the size of the training set. The first important thing to notice is that the score is not very sensitive to the amount of training data, showing that our method is well suited also for small datasets. By looking at the dependence with respect to ϵ, we see that when we have a lot of data, a stronger regularization means a poorer prediction skill. On the other hand for small datasets the best performance is achieved at a finite value of ϵ. This can be explained by the fact that as we have fewer and fewer data to estimate a constant size covariance matrix, it will become more and more singular, thus requiring a stronger regularization. Also, a smoother pattern is more likely to generalize well when training and validation data are very small.

In any case, we remind that choosing the proper regularization coefficient is not just a matter of score, but also of physical interpretability of the projection pattern, as explained in the previous section. From a qualitative look at projection maps at different values of T, τ and ϵ, $\epsilon = 1$ seemed to be a universally good compromise for the PlaSim dataset. Hence, if not specified differently, in the remainder of this work we will always consider $\epsilon = 1$.

On the right panel of table 3.1 we see the comparison with the skill of the neural network in the form $1 - \mathcal{S}/\mathcal{S}_{CNN}$, which shows that as the dataset gets smaller, the CNN loses its advantage, being outperformed when crossing the 1000 years threshold. An important caveat here is that the many hyperparameters of the CNN where optimized for the biggest dataset [Miloshevich et al., 2023a], and then kept constant for the experiments when training on fewer data. This potentially makes the comparison between the neural network and our method not completely fair. In fact, some experiments (not shown in this work), suggest that by optimizing hyperparameters such as the learning rate and batch size used for training the neural network allow it to prevail even when training only on 450 years of data. The Gaussian approximation, however, is still better when working with 200 years or less, even considering the optimization. So, the qualitative behavior displayed in table 3.1 still holds, and can be ultimately attributed to the higher complexity of the CNN (roughly a million parameters) with respect to the Gaussian approximation (roughly a few thousands of parameters).

Summarizing, our method is well suited to work in a regime of lack of data due to short datasets, where complex neural networks struggle.

3.5.4 More Extreme Heatwaves

A question complementary to the one of smaller datasets is the one of more extreme heatwaves, as they both result in very few samples of the event of interest.

First, the Gaussian approximation provides a committor that depends on the heatwave threshold a only through the parameter α. This means that, similarly to the composite maps, the projection pattern M will be the same for all heatwaves independently on how extreme they are. It is thus extremely easy and cheap to get a new committor estimate for a different value of a.

Normalized log score

years of training	$\epsilon = 10^{-2}$	10^{-1}	10^{0}	10^{1}	10^{2}
7200	0.43	0.43	0.42	0.40	0.37
3600	0.43	0.42	0.41	0.39	0.37
1800	0.42	0.42	0.41	0.39	0.36
900	0.44	0.43	0.43	0.41	0.38
450	0.43	0.42	0.42	0.40	0.37
180	0.36	0.38	0.39	0.39	0.37

$1 - \mathcal{S}/\mathcal{S}_{CNN}$

$\epsilon = 10^{-2}$	10^{-1}	10^{0}	10^{1}	10^{2}
0.07	0.08	0.10	0.14	0.20
0.03	0.05	0.07	0.11	0.17
0.01	0.02	0.04	0.09	0.15
-0.03	-0.03	-0.01	0.03	0.10
-0.09	-0.08	-0.07	-0.02	0.05
0.01	-0.05	-0.07	-0.07	-0.02

Table 3.1: Left table: normalized log score of the Gaussian approximation (the higher, the better), versus training dataset length and the regularization coefficient ϵ. With small datasets, intermediate ϵ values are optimal, while vanishing ones are for large datasets. The apparent peak in performance for 900 years of training is not significant. The Gaussian approximation's skill is already nearly optimal for small datasets. Right table: comparison with the skill of the neural network (ϵ affects only the Gaussian approximation). Brown colors mean the CNN performs better, while blue hues mean the Gaussian approximation is better. When the neural network has much data to learn, it can leverage its expressivity potential to outperform the Gaussian approximation. With small datasets, the added complexity of neural networks is detrimental to its score. Both panels are based on PlaSim data.

On the other hand, since the neural network we consider is trained on a classification task, as we change a, the labels $Y(t)$ change as well, and hence the whole network needs to be retrained every time. Although transfer learning can reduce the computational cost and avoid retraining from scratch, it still a more complex task than computing a new Gaussian committor. Furthermore, as we focus on more and more extreme heatwaves, the imbalance between the $Y = 0$ and $Y = 1$ classes becomes more and more relevant, eventually hindering the performance of the network (see the gray error-band in fig. 3.6). On the contrary the smaller size of the heatwave class affects the performance of the Gaussian approximation only in its variance, while the mean normalized log score $\mathcal{S}$ has a very weak dependence on the amplitude of the heatwave (blue line in fig. 3.6). This, in turn, suggests that our Gaussian approximation is sufficient to capture well the relationship between the predictors and the heatwave amplitude A even in the most extreme tails of the distribution.

In this section we showed that the Gaussian approximation can be a simple, but powerful, tool for the prediction of extreme heatwaves. Compared to other methods, such as deep neural networks, it does not need as much data to be properly trained. This makes it particularly suited for short datasets, which is typically the case in the climate community. This direction is further expanded in section 3.7, where we apply our method to the ERA5 reanalysis data. Furthermore, and crucially, it is usually very hard to interpret the prediction performed by a deep neural network, while the Gaussian approximation, through the optimal projection pattern, is interpretable *by design*. The study of the

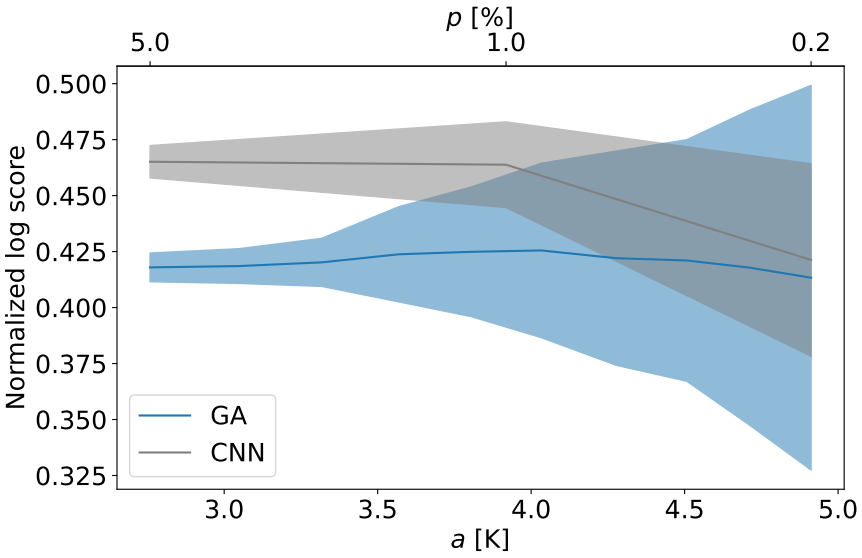


Figure 3.6: Normalized log score of the Gaussian approximation (blue) and CNN (gray) when varying the heatwave threshold a (bottom horizontal axis) or, equivalently, its climatological probability p (top horizontal axis). The solid line is the mean over the 10-fold cross validation process, while the shaded area represents one standard deviation. The experiment is performed with 7,200 years of training with PlaSim data. The regularization coefficient for the Gaussian approximation is kept at the optimal $\epsilon = 1$ value. The CNN is always best but because of the lack of data for rare events, its relative skill decreases with a. The skill of the Gaussian approximation is not much sensitive to the rareness of the event.

projection pattern opens the possibility for insight on the physical processes behind the event under study, and we expand on this in section 3.6.

3.6 Committor Functions and Optimal Projection for Extreme Heatwaves

In sections 3.4 and 3.5 we computed composite maps and committor functions for extreme heatwaves. However, in these sections the focus was mainly methodological, with attention to performance and the technical details that influence it. In this section we complement the previous analysis by focusing instead on the physical insight that our method provides on extreme heatwaves. To do so we will compare composite maps and optimal projection patterns at different values of the heatwave duration T and the lead time τ.

3.6.1 Comparison Between Composite Maps and Projection Patterns

In section 3.3 we showed that a-priori and a-posteriori statistics are fundamentally different. Here we proceed to further include some physical reasoning that arises when comparing the two types of statistics. In fig. 3.7 we have the side by side comparison, at

different values of the lead time τ, of the Gaussian composite map $C_{\mathcal{G}}$ with the projection pattern M needed for the computation of the committor. As explained in section 3.3.1, the composite map captures the *correlations* between the heatwave amplitude A and the predictors X, while the committor, and thus the projection pattern M, focuses on what is really important for the *prediction*.

A clear example of this is the difference between the 2 m temperature anomaly field in the composite map and in the projection pattern. From fig. 3.7, we can see that the composite shows many teleconnection features, for example over North America, while in the projection map virtually all the weight is over France. This suggests that the relationship between heatwaves and these temperature teleconnections is only of correlation, not causation. Similarly, the 500 hPa geopotential height field anomaly shows a very strong anticyclone over Greenland in the composite maps, which is not present in the projection patterns.

Another remarkable difference between $C_{\mathcal{G}}$ and M is the relative magnitude of the fields. By looking at the colorbars at the bottom of the figure, we see that, in the composite, all the fields have roughly the same order of magnitude, and this makes sense as we work with normalized data and the composite is representative of the typical heatwave event. On the other hand, from the projection patterns we observe that the values of soil moisture are 4 to 10 times higher than the ones of temperature and geopotential, showing that the soil moisture anomaly field is far more important for prediction than one might assume by just looking at the composite.

If we now focus on what happens when we change the lead time τ, we see that in the composites there is essentially just a fading of the structure of the 2 m temperature and 500 hPa geopotential height anomalies apparent at $\tau = 0$, with some minor qualitative changes, such as the connection of the two high pressure systems over the Atlantic at $\tau = 5$. On the other hand, the soil moisture anomaly component remains almost unchanged. This increased prominence of soil moisture as the lead time increases is even more pronounced for the projection pattern M, showing that soil moisture is the key factor for long term heatwave forecast.

Finally, from the evolution of the projection map for the 500 hPa geopotential height field, we see a clear shift of focus from the North-eastern Atlantic at $\tau = 0$ to the United States at $\tau = 5$. At $\tau = 10$ the most prominent feature in the 500 hPa geopotential height projection pattern is a small cyclone over the continental US, something which can barely be seen at all in the composite. These changes in the projection pattern give us insight into the dynamics of atmospheric circulation that leads to heatwaves over France, in particular the dynamics of the jet stream.

3.6.2 Effects of Changing T and τ

In this section we analyze more quantitatively how the performance of the Gaussian approximation is affected by the heatwave duration T and the lead time τ, and what physical conclusions we can derive from it. We will first perform this sensitivity analysis

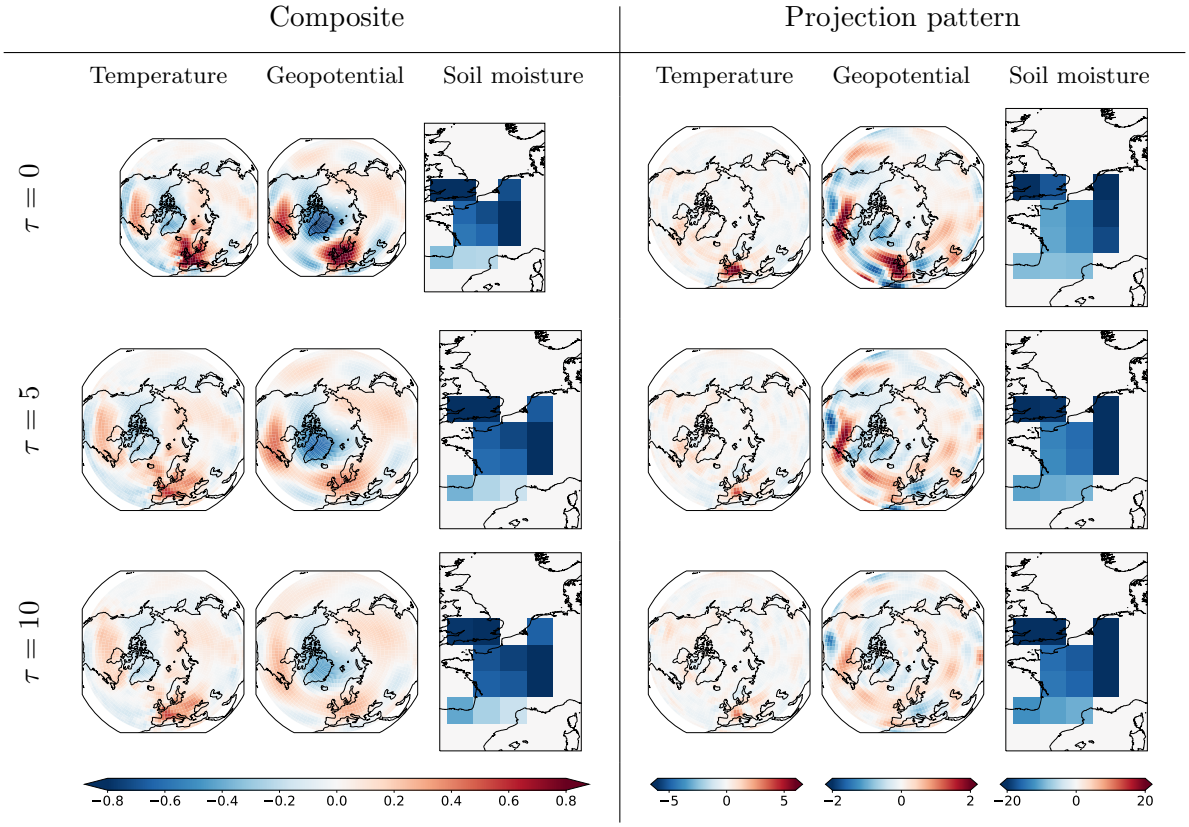


Figure 3.7: Left columns: Gaussian composite maps, temperature, geopotential height at 500hPa and soil moisture, for three different values of the lag time τ. Right column: the optimal projection pattern for prediction, within the Gaussian approximation ($\epsilon = 1$). As expected the two sets of maps are different, characterizing either a-posteriori statistics or best prediction patterns. The composite features hemispheric scale patterns dominated by zonal wave-number zero and zonal wave number three modes. For long lead times, the zonal wave-number zero pattern clearly dominates. The soil moisture composite pattern does not change much with the lag time. The information needed for making an optimal projection, as seen through the projection pattern, is at a finer scale, less global, with a strong meridional structure. Temperature contributes weakly and only through its local values to the projection pattern.

Fraction of area with error above 2σ

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.52	0.50	0.46	0.34	0.27	0.19	0.15	0.08	0.04	0.02	0.02
3	0.52	0.44	0.41	0.28	0.22	0.14	0.11	0.05	0.02	0.01	0.02
7	0.45	0.41	0.34	0.26	0.18	0.15	0.10	0.05	0.03	0.02	0.01
14	0.37	0.30	0.23	0.17	0.11	0.08	0.06	0.05	0.03	0.01	0.01
30	0.15	0.10	0.08	0.07	0.05	0.04	0.03	0.01	0.01	0.01	0.02

Table 3.2: Fraction of significant area in the conditional average (composite map) computed using the Gaussian approximation. The significance in assessed using the fraction of area which is above two standard deviations from the composite map evaluated over the 8000 years PlaSim data (see eq. (3.27)). The threshold used for defining a heatwave is $a = 2.76K$, corresponding to the 5% most extreme values of the distribution of the 2 m temperature anomaly over France, A. The table shows the dependency of the norm of the relative error on T, the heatwave duration, and τ, the lead time. Low value (dark red) means less areas beyond two standard deviations from the empirical composite, thus the Gaussian composite better reproduces the empirical one and the statistical error is a systematic error not due to the size of the dataset. Significant areas are monotonic increasing as T and τ increase.

on the composite maps (a-posteriori statistics) in section 3.6.2 and then for committor functions (a-priori statistics) in section 3.6.2.

Composites

In table 3.2, we see the fraction $\mathcal{F}$ of area for the Gaussian composite that has a significance above 2, as defined in eq. (3.27). The table shows a monotonic trend, with fast and imminent heatwaves having more non-Gaussian features with respect to long and delayed ones. Indeed, for higher values of the heatwave duration T, we expect the statistics of A to be more Gaussian, as we average over a larger number of days. Instead, when we increase the lead time we can think that the chaotic nature of the weather makes the states that led to a heatwave more different from one another. So, both the empirical and the Gaussian composite will tend to 0 as τ increases. Moreover, the higher differences between the states over which we take the empirical average increase the standard deviation. Thus, the significance of each pixel as in eq. (3.27) naturally decreases with τ.

On the other hand if we look at the values for the norm ratio (eq. (3.25)) displayed in table 3.3, we see a rather non-monotonic behavior. In fact, we can gain more understanding if we plot the norm ratio for the three climate variables independently (tables 3.7 to 3.9), which shows that the main contribution to the norm ratio comes from the 500 hPa geopotential height field.

This overall non-monotonic trend can be explained as a competition between the non-linear chaotic dynamics of the weather, that makes the real composite stray more from its Gaussian approximation as τ increases, with the loss of memory that averages out the

Norm ratio

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.25	0.28	0.29	0.26	0.27	0.27	0.28	0.26	0.22	0.21	0.21
3	0.24	0.25	0.26	0.24	0.25	0.25	0.26	0.23	0.21	0.19	0.19
7	0.22	0.23	0.24	0.25	0.27	0.29	0.28	0.24	0.23	0.22	0.19
14	0.20	0.23	0.26	0.28	0.28	0.27	0.27	0.26	0.24	0.22	0.22
30	0.21	0.24	0.26	0.28	0.28	0.27	0.27	0.26	0.25	0.25	0.25

Table 3.3: Norm of the relative error of the conditional average (composite map) evaluated using the Gaussian approximation. Relative to the composite value obtained through empirical conditional average over the 8000 years PlaSim dataset. The threshold used for defining a heatwave is $a = 2.76K$, corresponding to the 5% most extreme values of the distribution of the 2 m temperature anomaly over France, A. The higher the value (bright yellow) the worst the Gaussian composite approximates the empirical one. Lower values (dark red) denotes a lower value of the error. The table shows the dependency of the norm of the relative error on T, the heatwave duration, and τ the lead time. There is a non-monotonic trend which is due to the different atmospheric fields using in the conditional average. Events which are long-lasting and far in time behave more closely to Gaussian distributed events.

non-linear effects, bringing the empirical composite closer to the Gaussian one. This also would explain why geopotential dominates the norm ratio, as, of the three fields, it is the one with the most non-linear dynamics.

Committor

Similarly to what has been done for the composite maps, we can look at how the skill of the prediction is affected by the heatwave duration T and the lead time τ. In the left panel of table 3.4, we can see that the prediction skill decreases monotonically with τ at any level of T. For shorter lead times the skill is best when dealing with shorter heatwaves, while for longer delays, the skill is higher for longer-lasting events. In the limit of $T = 1$ and $\tau = 0$, we are forecasting a one-day heatwave that starts today, so we might just look outside the window and see if it is hot. And indeed there is perfect correlation between the temperature anomaly over France and the heatwave amplitude A. However, one day heatwaves are very erratic events, which become very hard to predict for longer lead times. On the other hand, longer lasting events are non-trivial to predict for very short delays, but are more influenced by processes with long timescales such as the dynamics of soil moisture, and hence maintain some predictability at higher values of τ [Miloshevich et al., 2023a].

On the right panel of table 3.4, we see the skill comparison with the neural network, which is able to capture non-linear and non-Gaussian structures in the data. We can see that our Gaussian committor struggles the most for shorter heatwaves and, more importantly, around $\tau = 5$. We can interpret this region of struggle as the one where the prediction is

Normalized log score

T [days] \ τ [days]	0	5	10	15	20	30
1	0.89	0.27	0.14	0.11	0.09	0.08
7	0.53	0.25	0.18	0.14	0.13	0.12
14	0.42	0.26	0.20	0.18	0.17	0.16
30	0.34	0.26	0.23	0.21	0.21	0.20

$1 - \mathcal{S}/\mathcal{S}_{CNN}$

τ [days]	0	5	10	15	20	30
	-0.00	0.26	0.22	0.19	0.18	0.15
	0.11	0.21	0.17	0.14	0.10	0.08
	0.10	0.17	0.13	0.09	0.06	0.04
	0.07	0.09	0.05	0.03	0.00	-0.00

Table 3.4: Left table: normalized log score of the Gaussian approximation (the higher, the better), versus heatwave duration T and lag time τ. In all cases, we focus on the 5% most extreme heatwaves. As the prediction task gets harder, the skill decreases monotonically with the lag time, faster for shorter heatwaves. Right table: comparison with the skill of the neural network. The CNN is always better, but more so for shorter heatwaves and around $\tau = 5$. This is the regime where the dynamics is more non-linear, and thus the neural network complexity has a better opportunity to make a difference.

most *dynamical*, rather than *statistical*. Namely, where mere linear correlations are not enough and the complex and non-linear dynamics of the atmosphere plays a significant role.

3.7 Application to the ERA5 Reanalysis Dataset

In this article we presented a methodology for estimating composite maps and committor functions using a theoretical framework that we called the Gaussian approximation (see section 3.3) and we tested it over a very long simulation dataset obtained from the climate model PlaSim. The results are really promising.

A key point that we showed in the previous sections is that our Gaussian framework is particularly suited for short datasets. In the case of the composite map (see section 3.4.4), the empirical average is performed over too few samples to be very accurate. For the committor (see section 3.5.3), the alternative approach of deep neural networks struggles with the lack of data. It is then natural to try to apply our method the ERA5 reanalysis data [Hersbach et al., 2020], and in this section we show that indeed for this dataset the Gaussian approximation is the best option.

3.7.1 Composites

In this subsection we compute composite maps on the ERA5 dataset, using both the empirical average and the Gaussian approximation. Because the dataset is much shorter than the PlaSim dataset, we do not know the ground truth as precisely as in Section 3.4. We can nevertheless compare the two estimates and see if they qualitatively agree.

Figure 3.8 shows the empirical composite and the composite evaluated within the Gaussian framework for the geopotential height anomaly at 500 hPa. They are both evaluated for $T = 14$, $\tau = 0$ and for heatwaves corresponding to the 5% most extreme

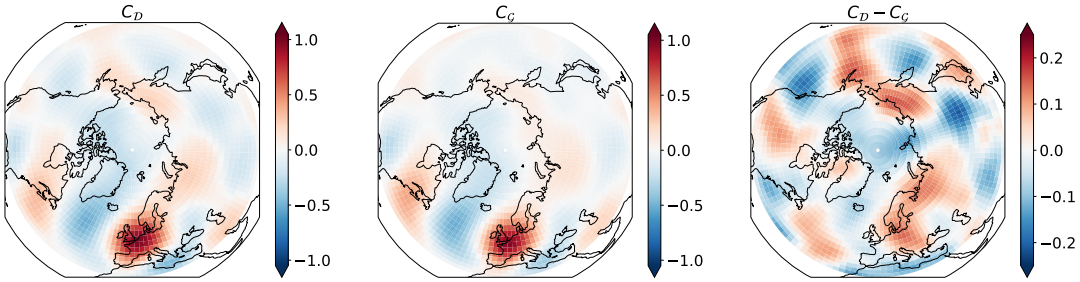


Figure 3.8: Composite maps of normalized 500 hPa geopotential height for 5% most extremes 14-day temperature anomaly over France ($T = 14$). Composite maps are computed either directly from ERA5 data (left map), or under the Gaussian approximation (central map). The right map shows the difference between the first two. The salient features of geopotential are well captured by the Gaussian approximation, with errors of the order of 25% at most.

value in the distribution of A. The two composites look qualitatively very similar. In both cases we see a clear wave train which starts from the western part of the United States and Canada with an anticyclonic anomaly, continues over the North Atlantic Ocean and finally terminates over Western Europe with another anticyclonic anomaly, stronger than the rest of the wave pattern. This is consistent with the fact that we condition on the temperature anomaly over France. The overall wave structure is well represented by the Gaussian composite, even if it puts a higher weight over the Western Europe anticyclone (fig. 3.8, right panel). The difference between the two composites is larger over Asia and over the Pacific Ocean. Unlike the case of PlaSim data, the difference does not have an annular mode structure but contains a visible wave number 6 component. The largest differences between the two composites are on the order of 20%.

As for PlaSim data, we analyzed how the two indices $\mathcal{R}$ (norm ratio, defined in eq. (3.25)) and $\mathcal{F}$ (fraction of area where differences between Gaussian and empirical composites are significant, defined in eq. (3.27)) vary with the parameters T and τ for ERA5 data. Table 3.5 shows the norm ratio as a function of T and τ for the 5% most extreme heatwaves. We see that, similarly to PlaSim, there is a non-monotonic trend with the lowest values for T between 1 and 14 and τ between 0 and 6. Outside this range, the norm ratio has rather high values which are the sign of a great mismatch between the two composites. However, $\mathcal{F}$ (table not shown) assumes values which are almost never above 1% and very often below 0.1%, meaning that we cannot rule out that any discrepancy between the Gaussian and empirical composite is simply a sampling error.

3.7.2 Committor

In this subsection we deal with the computation of committor functions on the reanalysis dataset. After the necessary technical adaptations to work on this dataset, we compare the prediction skill of the Gaussian approximation with the one of neural networks, which shows the first is clearly better.

Before discussing any result, we need to define a protocol for the choice of the proper

Norm ratio

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.29	0.38	0.47	0.59	0.81	0.89	0.88	0.88	0.78	0.74	0.73
3	0.33	0.40	0.48	0.69	0.90	0.87	0.86	0.81	0.82	0.72	0.81
7	0.35	0.39	0.53	0.73	0.79	0.80	0.83	0.84	0.84	0.76	0.81
14	0.35	0.46	0.61	0.73	0.82	0.86	0.85	0.87	0.83	0.79	0.81
30	0.62	0.81	0.89	0.90	0.84	0.81	0.82	0.82	0.82	0.81	0.83

Table 3.5: Norm of the relative error of the conditional average (composite map) evaluated using the Gaussian approximation. Relative to the composite value obtained through empirical conditional average over the ERA5 dataset. The threshold a used for defining a heatwave corresponds to the 5% most extreme values of the distribution of the temperature anomaly over France, A. The higher the value (bright yellow) the worst the Gaussian composite approximates the empirical one. Lower values (dark red) denotes a lower value of the error. The table shows the dependency of the norm of the relative error on T, the heatwave duration, and τ the lead time.

regularization coefficient ϵ, the only hyperparameter of our method. When working with 8000 years of PlaSim data, we had to choose empirically $\epsilon = 1$ to have interpretability in the projection patterns, and this interpretability came at the cost of a lower skill score. On the other hand, on the reanalysis dataset, and more generally when working with small datasets (table 3.1), the value ϵ_{best} of the regularization coefficient that yields the highest skill score also provides an interpretable projection map.

The reanalysis dataset consists of 83 years of data. To have a meaningful cross validation we take the 80 years from 1943 to 2022 and split them in 5 balanced folds (see section 3.9.2). This way we train on 64 years and validate on 16.

With this choice, for the 5% most extreme two-week heatwaves ($T = 14$) at $\tau = 0$, we obtain a skill score of $\mathcal{S} = 0.16 \pm 0.07$. This number is considerably lower than the skill we have measured for PlaSim (section 3.5.1). To understand why, we can investigate the impact on the skill score for the PlaSim dataset of the reduced number of predictors (using only the 500 hPa geopotential height field as for reanalysis data) and of the amount of data (training on a subset of the same size as the reanalysis data). As can be seen from

	Predictor fields	
years of data	$T_{2\text{m}}, Z, S$	Z
8000	0.418 ± 0.006	0.23 ± 0.01
80	0.33 ± 0.07	0.18 ± 0.04

Table 3.6: Skill score on PlaSim, when using different amount of data and different predictor fields.

table 3.6, both the reduced number of predictor fields and the smaller dataset severely impact the skill score. However, even combining the two effects, the performance remains

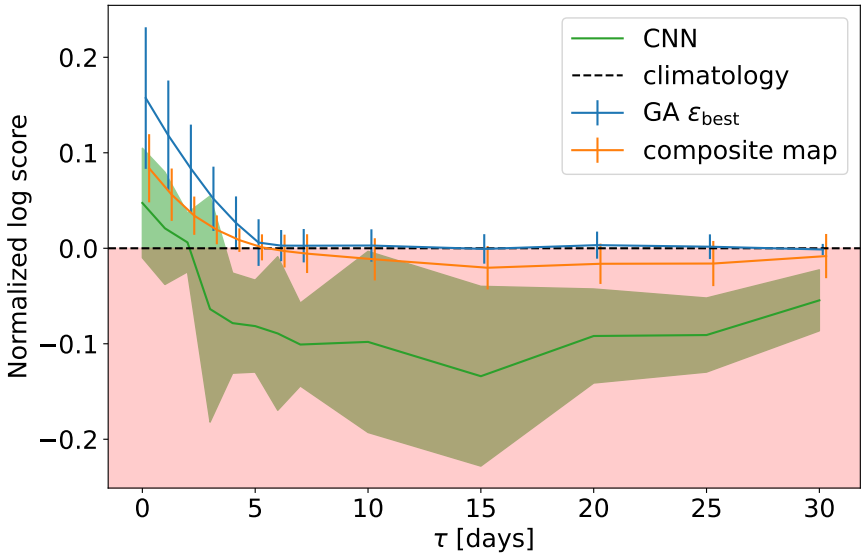


Figure 3.9: Skill score of different prediction techniques for reanalysis data (using geopotential height at 500 hPa anomaly as the only predictor, $T = 14$) changing the lead time. In green the convolutional neural network, in blue the Gaussian approximation, both at their best values for hyperparameters. In orange the Gaussian approximation when using the composite map as projection pattern. Error bars or shaded area indicates the variation among the 5 folds. The red shaded zone below 0 indicates where the prediction is worse than the climatology. The Gaussian approximation is always the best, and gives results better than the climatology only for $\tau \leq 5$.

slightly better than for the reanalysis data. This suggests that the more realistic data of ERA5 have more complexity and variability with respect to PlaSim, and thus it is harder to make a skillful prediction.

Nevertheless, we argue that the result achieved for reanalysis data, albeit humble, is the best we can do. To support this claim, in fig. 3.9 we compare it to the skill of other methods at different values of the lead time τ. In green is the performance of a convolutional neural network with a similar architecture to the one used for PlaSim. It always performs worse than the Gaussian approximation (in blue), and already at $\tau = 3$ days it is consistently below the climatology. On the other hand, the Gaussian approximation manages to extend the predictability margin a few more days. For $\tau \geq 6$ days the latter becomes useless too, and, interestingly, $\epsilon_{\text{best}} \to \infty$, yielding a uniform projection pattern.

In the regime where the prediction is still skillful, the projection patterns look remarkably similar to the composite maps (fig. 3.11), so it is natural to try to project onto the composite itself. This is the orange line in fig. 3.9, which, despite having a smaller error bar than the optimal projection pattern M, on average yields a worse performance. This once again highlights the fact that composite maps are not the proper tool for prediction.

Now that we showed that the Gaussian approximation is the best option for very small datasets, we can investigate what happens when we vary the heatwave duration. From

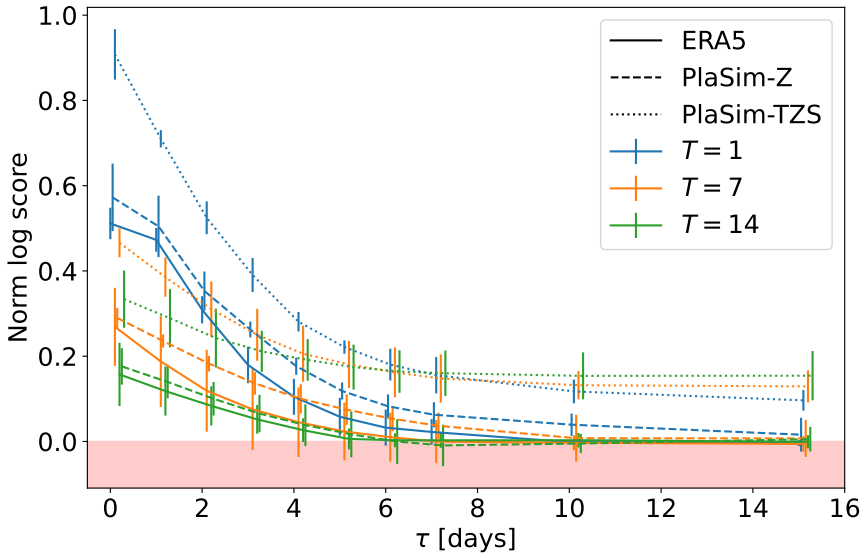


Figure 3.10: Skill score of the Gaussian committor as a function of τ for different values of the heatwave duration T, and three different datasets: ERA5 with only geopotential height at 500 hPa (solid line), 80 years of PlaSim data with 2 m temperature, geopotential height at 500 hPa and soil moisture (dotted lines) and only with geopotential height at 500 hPa (dashed line). PlaSim has a consistently higher predictability than ERA5, and the addition of the slow evolving soil moisture greatly extends the predictability horizon. In the absence of this slow variable, predictability decreases with the heatwave duration.

fig. 3.10 we see that, at any fixed value of τ, the prediction skill decreases with increasing heatwave duration T (solid lines), with shorter heatwaves having a longer predictability horizon. The comparison with 80 years of PlaSim data with only the 500 hPa geopotential height as predictor (dashed lines), shows that predicting heatwaves is harder on the more realistic data. This can be an effect of the lower spatial resolution of the PlaSim model, which yields a more sluggish and less chaotic atmospheric dynamics, and, hence, better predictability. This hypothesis is further reinforced by the fact that, on average, training on PlaSim requires a lower regularization coefficient than the one on reanalysis data (fig. 3.11).

Finally, the dotted lines in fig. 3.10 represent the skill when still training on 80 years of PlaSim data, but with all three predictors. For short lead times and heatwave duration, the increase in skill comes mainly from the direct information of the 2 m temperature field, but the more interesting effect is for longer delays. Here, almost all the predictive power resides in the soil moisture field, and is able to extend the predictability horizon significantly. This effect is enhanced for longer lasting heatwaves. As was already pointed out in [Miloshevich et al., 2023a], soil moisture acts as a slow modulator of the chance of a heatwave, that is still able to give some useful information when the fast predictors, such as the 500 hPa geopotential height, are beyond their de-correlation timescale.

Summarizing, the higher complexity of the ERA5 dataset, its reduced length, and the absence of soil moisture as a slow predictor, all these aspects make so that the forecast

skill is much lower than the one for PlaSim. However, the prediction performed by the Gaussian approximation proved to be the best available option, with results that are still remarkable.

3.7.3 Physical discussion

Now that we discussed composite maps and committor functions from the point of view of performance, we proceed to focus more on the physics-oriented analysis of composite maps and optimal projection patterns.

In fig. 3.11, we show the comparison between composite maps and projection patterns computed on reanalysis data and on 80 years of PlaSim data using only the 500 hPa geopotential height as predictor. Interestingly, we observed that both composite maps and projection patterns do not change much with respect to the heatwave duration T (not shown). This is only partially explained by consecutive days with high temperature contributing both to short and long heatwaves. In the future it may be worth investigating this further, but in this work we simply exploit it to discuss the patterns only for a single value of T and still provide a relatively comprehensive picture. In particular, we show the results for 1-day heatwaves, as they display a clearer evolution of both composites and projection patterns with the lead time τ.

As already mentioned before, one of the main differences between PlaSim and ERA5 is a generally higher signal-to-noise ratio in PlaSim, that manifests itself in higher norms of the composite maps (left columns) and less smooth projection patterns (right column). At $\tau = 0$, most of the weight of both composite maps and projection patterns is concentrated around France. More precisely, with an anticyclone over France and Central Europe to ensure clear skies and a cyclone north of Portugal to advect warm African air northward. This cyclone is more localized in the ERA5 projection map than in composite maps. As the lead time increases, this dipole structure stretches westward into the Atlantic Ocean. In the reanalysis dataset there is a clear emergence of a wave-train pattern, and C and M look rather similar. On the other hand, PlaSim's projection patterns stray considerably from the composite maps, and the physics that they hint at is harder to explain. For both datasets, and in both composite maps and projection patterns, a rather strong anticyclonic anomaly is always present over France, getting fainter as τ increases, but remaining always a prominent feature. This suggests that even for very short $T = 1$ heatwaves the most *common* (composites) and the most *likely* (committor) causes of the extreme events are connected with quasi-stationary weather states.

Concerning the reanalysis dataset, the similarity between composite maps and projection patterns may tempt us to use the composite as a prediction tool. However, although both composites and projection maps display the dynamics of a stationary Rossby wave, a careful examination shows a different weight distribution in the projection patterns, for example at $\tau = 6$ the focus is more over North America than in the composite map. The lower prediction skill achieved with the composite map compared to the optimal projection pattern, already discussed above (fig. 3.9), suggests that such differences matter

for prediction even if they appear small at first sight. Furthermore, the similarity between the two maps can most likely be attributed to the need for a relatively high regularization coefficient, required to have a prediction that generalizes well when trained on such a short dataset. More technical details are available in section 3.9.12.

This section has three main conclusions. Firstly, given the size of the ERA5 dataset (and of any other real-data dataset), it is hard to go beyond the Gaussian approximation for both analysis of the averaged weather conditions that led to heatwave events (composite maps) and for a probabilistic forecast of heatwaves, as shown from fig. 3.8 and fig. 3.9. Secondly, the difference between the empirical and the Gaussian composite maps, shown in fig. 3.8 (right panel) has a different wave number with respect to the one observed for PlaSim, for heatwave of the same duration and intensity, (see fig. 3.2 bottom row, central map, which exhibits a wave zero pattern). However, we cannot exclude that this mismatch is due to sampling error. Thirdly, the reduced size of the dataset forces us to strongly regularize the optimal projection patterns, which makes them visually similar to the composite map. However, even if they do not provide any additional qualitative information, they do provide more precise quantitative information, leveraged for prediction skill. Finally, the comparison with the data from PlaSim suggests the importance of predictor fields other than the 500 hPa geopotential height, which can significantly improve the prediction skill. For the reanalysis dataset, this opens the possibility to use also ocean variables, like sea surface temperature.

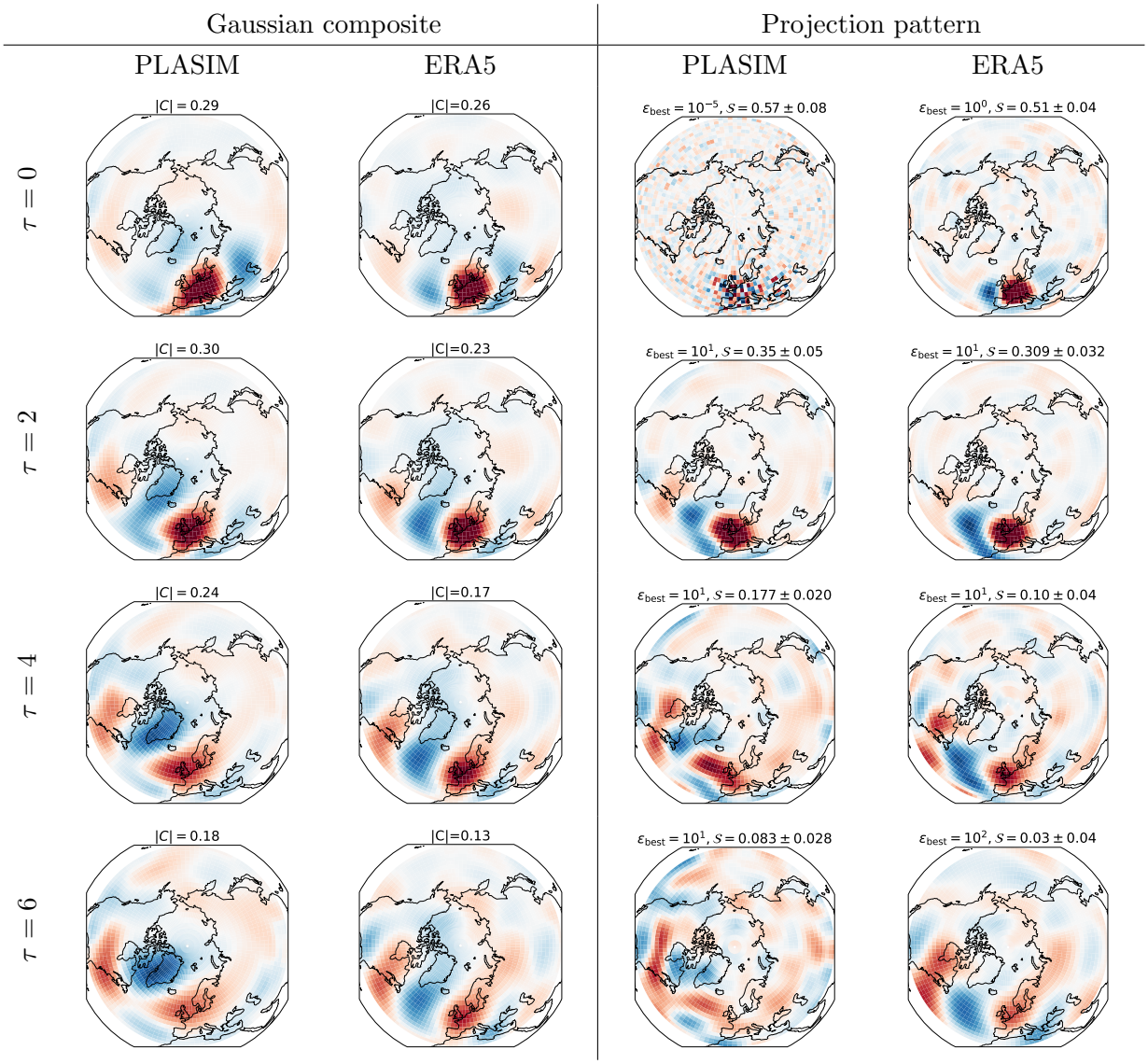


Figure 3.11: Comparison between composite maps and projection patterns of ERA5 and PlaSim (80 years, geopotential height at 500 hPa only), at $T = 1$ and different values of τ. All maps are shown as normalized to unitary L2 norm. The L2 norm of the actual composite maps is reported on top of them, while for the projection pattern we display the regularization coefficient and the skill score. For ERA5 composites and projection patterns *look* qualitatively similar. However, this is a result of the small size of the dataset, which forces us to use high values of the regularization coefficient.

3.8 Conclusions and Perspectives

In this work, we stressed the important difference between the statistics of climate and weather conditions which led to an extreme event (a-posteriori statistics, for instance composite maps) and the prediction in the future of an extreme event given some knowledge (predictors) of the state of the climate system (a-priori statistics, for instance committor functions). We have highlighted the second as the proper set of tools for any prediction task. At the same time, we provided a simple framework to give easy access to these tools, which is effective even with short datasets of length of the order of several decades to several centuries. In the context of extreme heatwaves over France, we evaluated our method on a very long time series of climate model output data and successfully applied it to a reanalysis dataset.

Concerning a-posteriori statistics, with our Gaussian framework, we were able to provide an explanation of why the composite maps of very extreme heatwaves look qualitatively similar to the ones of less extreme ones. We made this statement quantitative, showing that composite maps are the same up to a rescaling by a non-linear function of the threshold that defines heatwaves. This opens the possibility to estimate composite maps of extremely rare events, even ones that have never been observed in the dataset. For PlaSim data, the computation of composite maps using the Gaussian approximation gives results which are valid up to an error (in L2 norm) of the order of 20 to 30%. We also stress that the deviations from the Gaussian prediction are statistically significant, showing that the statistics is actually not Gaussian and that information beyond the Gaussian approximation can be computed with dataset length of the order of a thousand years or more. On the other hand, on the much shorter reanalysis dataset, errors are larger, but entirely compatible with the imperfect sampling of the empirical composite, and one cannot compute statistically significant deviations from the composite map obtained within the Gaussian approximation.

However, if one is interested in *predicting* heatwaves instead of studying their statistics a-posteriori, composite maps are not the proper tool. The right one is the committor function, and our framework gives probably the easiest non-trivial access to this very complex object. Our method gives very good prediction skill, and is particularly competitive with more complex alternatives, such as neural networks, when working with small datasets, which are very common in the climate community. In fact, for the 80-year long ERA5 dataset, the Gaussian approximation proved to be the method with the highest predictive skill.

As demonstrated in [Miloshevich et al., 2023a], too short datasets prevent optimal use of neural networks in many applications in climate sciences. This issue is particularly salient for rare events, for instance extreme events. In this respect, we see the Gaussian framework developed in this paper as a key solution to make the first relevant prediction. It should play an important role in future studies. For rare events, going beyond the results of the Gaussian approximation may require to have datasets with more rare events. One way is to sample exceptionally rare extreme events using the recently developed rare event

simulation techniques, that are able to multiply by several orders of magnitude the number of observed heatwaves with PlaSim model [Ragone et al., 2018] and with CESM (the NCAR model used for CMIP experiments) [Ragone and Bouchet, 2021]. A perspective is to couple these rare event simulations with the Gaussian framework presented in this paper or other machine learning forecast. We have already coupled machine learning simulations to rare event algorithms, for simple academic models [Lucente et al., 2022b]. Coupling the rare event simulations with machine learning is a very interesting perspective to solve the key fundamental issue of lack of data in the science of climate extremes.

Moreover, beside pure skill, our method provides an optimal index for prediction, which, once properly regularized, makes it easy to interpret our results, giving insight in the dynamics behind our subject of study. This optimal prediction map is one of the key results of this paper. It makes the Gaussian approximation appealing even for applications on long enough datasets so that its skill can be outperformed by neural networks, which are often hard to understand.

From the point of view of understanding the underlying physics, in the case of extreme heatwaves over France, we found that both composite maps and optimal projection maps display a quasi-stationary pattern, that does not depend much on the lead time. In particular, the development of a Rossby wave-train over the Atlantic Ocean plays an important role for the short term prediction. This appears very clearly in the reanalysis data, while PlaSim has a strong competing contribution from a wave number 0 pattern. For longer lead times, instead, the analysis on PlaSim data and the comparison with ERA5 confirmed the key importance of slow drivers, such as soil moisture. The natural next step is then to include these slow drivers in the study on the reanalysis dataset, maybe even using ocean-based variables like sea surface temperature.

As further perspectives, we argue that a deeper analysis at the physical level of optimal projection patterns is needed, turning the qualitative insights presented in this work to more quantitative statements. Moreover, we took as an example extreme heatwaves over France: it would be interesting to apply our method to heatwaves on different geographical locations or to different types of extreme events altogether. Another very interesting direction is, in the cases where the Gaussian approximation is outperformed by neural networks, to interpret where this extra skill comes from. Finally, we suggest that our method can be used as a better baseline than the mere climatology when testing more sophisticated tools for probabilistic prediction.

Open Research Section

Data from the reanalysis dataset ERA5 [Hersbach et al., 2020], publicly available at `https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5` are used in this study. We also use data from a long simulation (8000 years) of the PlaSim climate model [Fraedrich et al., 2005a]. Details of the model setup can be found in [Miloshevich et al., 2023a]. We provide all the data necessary for reproducing the figures

on Zenodo (https://zenodo.org/doi/10.5281/zenodo.11400868) or GitHub (https://github.com/AlessandroLovo/gaussian-approximation-zenodo)

Acknowledgements

V. Mascolo and A. Lovo have contributed equally to the work performed in this study and should both be considered as the main author. V. Mascolo and A. Lovo have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 956396 (EDIPI) and 956170 (Critical Earth), respectively. The authors thank the computer resources provided by the Centre Blaise Pascal at ENS de Lyon. We are grateful to Emmanuel Quemener for his help with the platform. This work was granted access to the HPC/AI resources of CINES under the allocations 2018-A0050110575, 2019-A0070110575, 2020-A0090110575 and 2021-A0110110575 made by GENCI. We thank B. Cozian, C. Le Priol, and A. Lancelin for scientific discussions and suggestions.

All the authors declare no conflict of interest.

3.9 Supporting Information

3.9.1 Detrending of ERA5

In this manuscript we present an application of our methodology to the ERA5 dataset [Hersbach et al., 2020]. In this section we go over the technical details of handling this dataset.

We start from taking daily averages of the hourly data from the public available dataset of the ECMWF service (https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5) for summer seasons from 1940 to 2022. We decide to use 2 m air temperature for defining the heat wave amplitude and, differently from what has been used for PlaSim, only the 500 hPa geopotential height field as set of predictors X. The reason is that when we will use ERA5 for forecast of heatwaves, it will be impossible to work with the same amount of fields as for PlaSim but with only 83 years of data. In this line of thought, we also performed a regridding of the data over the PlaSim grid to considerably reduce the numbers of features. We then remove the seasonal cycle, so that we can work with anomalies.

In this work we aim at studying the response of climate models in stationary conditions, thus we detrend the ERA5 dataset to remove the climate change signal. For the temperature field, given that we used it just to define our heat wave amplitude A, we performed a spatial average over the France region, and detrended its seasonal mean with a quadratic fit. The time series of the seasonal mean is shown in fig. 3.12, where the orange line is the trend that we removed for each summer. We tried other sophisticated detrending methodologies, but this one was simultaneously the most simple and effective one.

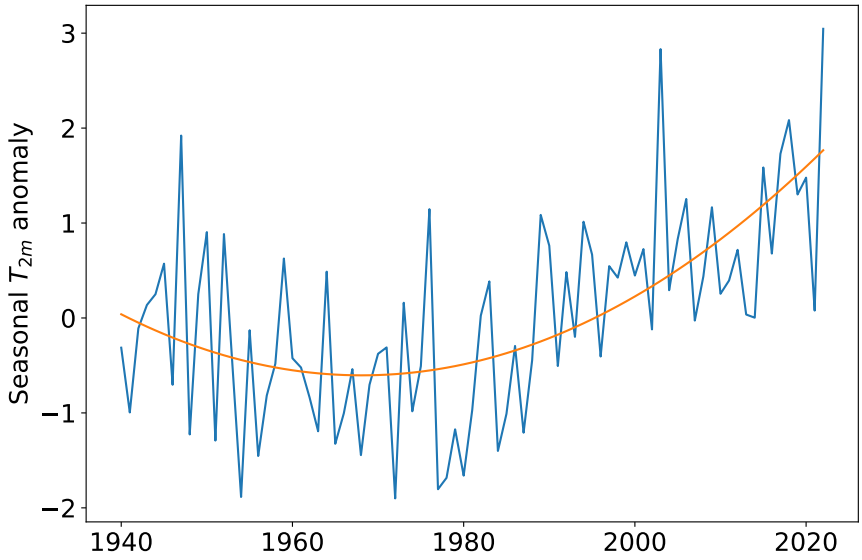


Figure 3.12: Seasonal T_{2m} anomaly averaged over France for the ERA5 dataset. The orange curve is the trend fitted via a second order polynomial.

A similar protocol was also applied for the detrending of the geopotential height field. However, given that we noticed a latitudinal dependence of the trend, we decided to independently detrend via a quadratic fit the seasonal and zonal means of geopotential height. Figure 3.13 shows the contour plot of the trend that we removed as a function of the latitude and year. Indeed, this trend is non-monotonic trend at mid and high latitudes, while it is monotonously increasing at lower latitudes. Given that this non monotonicity is present at the beginning of the dataset, our guess is that it might depend on the quality of the data available before the satellite era.

3.9.2 Balanced K-fold Cross Validation

The standard K-fold cross validation process consists in splitting the dataset $\mathcal{D}$ into K disjoint subsets of equal length $\{\mathcal{F}_k\}_{k=1}^K$, that can be called *folds*. Then, for each $k = 1, \dots, K$ we define training and validation sets as

$$\mathcal{T}_k = \bigcup_{i \neq k} \mathcal{F}_i, \quad \mathcal{V}_k = \mathcal{F}_k. \tag{3.30}$$

To make a *balanced* K-fold cross validation, we ask that the $\mathcal{F}_k$ all contain the same amount of heatwaves. This is essentially equivalent to the classical technique of stratified K-fold cross validation [Hastie et al., 2001], but in our case, to avoid contamination between the different folds, we force data belonging to the same summer to end up in only one of the folds.

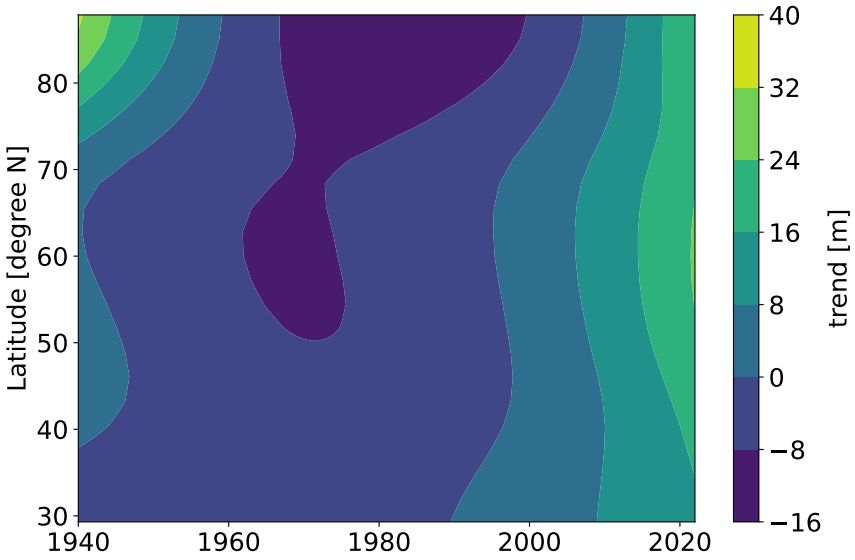


Figure 3.13: Contour plot of the 500 hPa geopotential height trend for ERA5 dataset as function of years and latitude. At latitudes the trend is non-monotonic, while it is monotonically increasing in time at lower latitudes.

3.9.3 Detailed Calculations of the Composite Map in a 2-Dimensional Gaussian System with Committor Depending Only on One Variable

In the second example of section 3.3.1, we use intuition to say that if we have two correlated variables but the committor depends only on one, the composite will still be non-zero for the variable upon which the committor does *not* depend. Here we give a formal proof.

According to the assumption of zero mean and Gaussianity for the two variables, we can write the stationary measure as

$$P_S(x_1, x_2) \propto \exp\left(-\frac{1}{2}(x_1, x_2)\begin{pmatrix}\sigma_1^2 & \phi \\ \phi & \sigma_2^2\end{pmatrix}^{-1}\begin{pmatrix}x_1 \\ x_2\end{pmatrix}\right) = \exp\left(-\frac{1}{2}(ax_1^2 + bx_2^2 - 2cx_1x_2)\right), \tag{3.31}$$

where

$$(a, b, c) = \frac{1}{\sigma_1^2\sigma_2^2 - \phi^2}(\sigma_2^2, \sigma_1^2, \phi). \tag{3.32}$$

Then, according to eq. (3.7), the composite value for X_2 is

$$C_2 \propto \int x_2 P_S(x_1, x_2) q(x_1) dx_1 dx_2, \tag{3.33}$$

$$\propto \int dx_1 q(x_1) \int dx_2 x_2 e^{-\frac{1}{2}(ax_1^2 + bx_2^2 - 2cx_1x_2)}, \tag{3.34}$$

$$= \int dx_1 q(x_1) e^{-\frac{1}{2}\left(a - \frac{c^2}{b}\right)x_1^2} \int dx_2 x_2 e^{-\frac{1}{2}b\left(x_2 - \frac{c}{b}x_1\right)^2}, \tag{3.35}$$

$$\propto \int dx_1 q(x_1) e^{-\frac{1}{2}\left(a - \frac{c^2}{b}\right)x_1^2} \frac{c}{b} x_1. \tag{3.36}$$

Now, first we notice that $C_2 \propto c \propto \phi$, so if there is no correlation between x_1 and x_2 we get the expected result that the composite is zero. Otherwise, for a generic committor q, $C_2 \neq 0$. A particular case for which $C_2 = 0$ is when q is an even function. However, this means that the committor must give equal probability to x_1 and $-x_1$, and thus cannot focus on a single tail of the distribution of x_1.

3.9.4 Detailed Calculation of the Composite Maps and of the Committor Function in the Gaussian Approximation Framework

As already presented in the main text, the Gaussian approximation relies on the hypothesis that, at each pixel, the field and the heatwave amplitude follow a jointly Gaussian distribution, namely

$$(X, A) \sim \mathcal{N}(0, \Sigma), \tag{3.37}$$

with Σ being the covariance matrix, $\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XA} \\ \Sigma_{AX} & \Sigma_{AA} \end{bmatrix}$. The joint multivariate Gaussian distribution, for given values of X and A, is generally written in the form:

$$\mathbb{P}(x, a) = \frac{1}{Z} \exp\left(-\frac{\left(x^T \Lambda_{XX} x + 2x^T \Lambda_{XA} a + \Lambda_{AA} a^2\right)}{2} \right), \tag{3.38}$$

where $Z = \sqrt{(2\pi)^d \det(\Sigma)}$ is the normalization constant d is the dimension of the stack of X and A, $\Lambda = \Sigma^{-1}$, $\Lambda_{XA} = \Lambda_{AX}$. We took advantage of the fact that a is a scalar quantity. Equation (3.15), can be obtained via the following calculation:

$$C = \mathbb{E}[X|A \geq a] = \frac{1}{\mathbb{P}(A \geq a)} \int_a^{+\infty} \left(\int x \mathbb{P}(x, a') \mathrm{d}x \right) \mathrm{d}a' , \tag{3.39}$$

$$= \frac{1}{\mathbb{P}(A \geq a)} \int_a^{+\infty} \mathbb{P}(a') \mathbb{E}[X|A = a'] \mathrm{d}a' , \tag{3.40}$$

$$= \frac{\int_a^{+\infty} \mathbb{P}(a') a' \mathrm{d}a'}{\mathbb{P}(A \geq a)} \frac{\mathbb{E}[XA]}{\Sigma_{AA}}, \tag{3.41}$$

$$= \eta\left(\frac{a}{\sqrt{2\Sigma_{AA}}} \right) \frac{\mathbb{E}[XA]}{\Sigma_{AA}}, \tag{3.42}$$

with

$$\eta(z) = \sqrt{\frac{2}{\pi}} \frac{e^{-z^2}}{\mathrm{erfc}(z)}, \tag{3.43}$$

where $\mathrm{erfc}(\bullet)$ is the complementary error function. Previously, we have used that:

$$\mathbb{E}[X|A = a] = \int x \mathbb{P}(x|a) \mathrm{d}x = \frac{\int x \mathbb{P}(x, a) \mathrm{d}x}{\mathbb{P}(a)} = \frac{a}{\Sigma_{AA}} \mathbb{E}[XA], \tag{3.44}$$

and this completes the proof.

For the committor, the important point is finding the expression for the conditional probability $\mathbb{P}(A = a|X = x)$. This is done by taking a slice at $X = x$ from of $\mathbb{P}(X, A)$ and

expressing it as a function of a:

$$\mathbb{P}(A = a|X = x) \propto \exp\left(-\frac{1}{2}\left(2(\Lambda_{XA}^{\top}x)a + \Lambda_{AA}a^2\right)\right), \tag{3.45}$$

$$\propto \exp\left(-\frac{1}{2}\Lambda_{AA}\left(a + \Lambda_{AA}^{-1}\Lambda_{XA}^{\top}x\right)^2\right), \tag{3.46}$$

which is the expression for a one dimensional Gaussian distribution with variance $\sigma^2 = \Lambda_{AA}^{-1}$ and mean $\mu(x) = -\Lambda_{AA}^{-1}\Lambda_{XA}^{\top}x$. Now from the expressions for inverting a block matrix like Σ, we know that

$$\Lambda_{AA} = \left(\Sigma_{AA} - \Sigma_{AX}\Sigma_{XX}^{-1}\Sigma_{XA}\right)^{-1}, \tag{3.47}$$

and

$$\Lambda_{XA} = -\Sigma_{XX}^{-1}\Sigma_{XA}\Lambda_{AA}. \tag{3.48}$$

Remembering again that Λ_{AA} is a scalar, we immediately get eq. (3.19),

$$\mu(x) = \Sigma_{XX}^{-1}\Sigma_{XA} \cdot x, \qquad \sigma^2 = \Sigma_{AA} - \Sigma_{AX}\Sigma_{XX}^{-1}\Sigma_{XA}. \tag{3.49}$$

After this, getting the full committor is a simple one dimensional Gaussian integral, which is already well explained in the main text.

3.9.5 Committor Function for a Stochastic Process

Let's consider a stochastic process $X(t)$ on a phase-space Ω. The *first hitting time* $\tau_{\mathcal{V}}^{'}$ of the set $\mathcal{V} \subset \Omega$, given that the trajectory started at x, is defined as:

$$\tau_{\mathcal{V}}^{'}(x) := \inf\{t : X(t) \in \mathcal{V} \mid X(0) = x\}. \tag{3.50}$$

The committor function q is defined as the probability that the first hitting time of the set $\mathcal{C}$ is smaller than the first hitting time of set $\mathcal{B}$, given the initial conditions x, where $\mathcal{B}, \mathcal{C} \subset \Omega$, $\mathcal{B} \cap \mathcal{C} = \emptyset$:

$$q(x) := \mathbb{P}(\tau_{\mathcal{B}}^{'}(x) > \tau_{\mathcal{C}}^{'}(x)). \tag{3.51}$$

Sets $\mathcal{B}$ and $\mathcal{C}$ can be two attractors of the system or for instance one could correspond to a typical state of the system around which it fluctuates and another one to an atypical state which is visited when rare fluctuations arise. In the context of this paper, we are interested in the second case, where we define the fluctuations of interest based on an observable, namely the heatwave amplitude, defined in eq. (3.1), reaching a given threshold a. It is then natural to rewrite the definition of the committor function as in eq. (3.4).

3.9.6 Spatial Gradient Regularization

To compute the spatial gradient of the projection pattern M, we need to consider that we are working in a spherical geometry, which has two effects. If Λ and Φ are respectively latitude and longitude, the gradient in the local flat geometry x-y (with x pointing eastward and y northward) is

$$\begin{cases} \frac{\partial}{\partial x} & = \frac{\partial \Phi}{\partial x}\frac{\partial}{\partial \Phi} = \frac{1}{\cos\Lambda} \times \frac{\partial}{\partial \Phi} \\ \frac{\partial}{\partial y} & = \frac{\partial \Lambda}{\partial y}\frac{\partial}{\partial \Lambda} = 1 \times \frac{\partial}{\partial \Lambda} \end{cases}. \tag{3.52}$$

The other effect is that the area of a grid cell is

$$d\mathcal{A} = dxdy = \cos\Lambda d\Lambda d\Phi. \tag{3.53}$$

If for simplicity we assume we are dealing with only one climate variable, the total squared spatial gradient of M is

$$H_2(M) = \int \left(\left(\frac{\partial M}{\partial x} \right)^2 + \left(\frac{\partial M}{\partial y} \right)^2 \right) dxdy, \tag{3.54}$$

$$= \int \cos\Lambda \left(\left(\frac{1}{\cos\Lambda} \frac{\partial}{\partial \Phi} \right)^2 + \left(\frac{\partial}{\partial \Lambda} \right)^2 \right) d\Lambda d\Phi, \tag{3.55}$$

$$= \int \left(\frac{1}{\cos\Lambda} \left(\frac{\partial}{\partial \Phi} \right)^2 + \cos\Lambda \left(\frac{\partial}{\partial \Lambda} \right)^2 \right) d\Lambda d\Phi. \tag{3.56}$$

In our case, however, M spans three climate variables, and is sampled on a uniform grid in latitude and longitude. This means we can write the projection pattern as a tensor $M^{\lambda\phi f}$, with indices $\lambda = 1, \ldots, n_\lambda = 22$ for latitude, $\phi = 1, \ldots, n_\phi = 128$ for longitude and $f = 1, \ldots, n_f = 3$ for distinguishing the fields. The discrete version of the gradient is thus

$$\begin{aligned} H_2(M) = \sum_{f=1}^{n_f} \Bigg[& \sum_{\lambda=1}^{n_\lambda - 1} (\cos\Lambda_\lambda) \sum_{\phi=1}^{n_\phi} (M^{(\lambda+1)\phi f} - M^{\lambda\phi f})^2 \\ & + \sum_{\lambda=1}^{n_\lambda} (\cos\Lambda_\lambda) \sum_{\phi=1}^{n_\phi} \left(\frac{M^{\lambda((\phi \mod n_\phi)+1)f} - M^{\lambda\phi f}}{\cos\Lambda_\lambda} \right)^2 \Bigg], \end{aligned} \tag{3.57}$$

where the first row is the meridional gradient and the second row the zonal one, considering also the periodic term. To be precise, we should add the multiplicative term $\Delta\Lambda\Delta\Phi$, but since it is a constant that we can include in the regularization coefficient ϵ, we can ignore it

If we now collapse all the indices of M into a single one $i = i(\lambda, \phi, f)$, it is quite obvious that we can write

$$H_2(M) = M^\top W M = \sum_{ij} W_{ij} M_i M_j. \tag{3.58}$$

To get the expression for W, we can first notice that it is symmetric: $W_{ij} = U_{ij} + U_{ji}$, and, by matching terms, we get

$$\begin{aligned} U_{ij} = & \left(\frac{\cos\Lambda_\lambda + \cos\Lambda_{\lambda-1}}{2} + \frac{1}{\cos\Lambda_\lambda} \right) \delta_{i(\lambda,\phi,f)j(\lambda,\phi,f)} + \\ & - (\cos\Lambda_\lambda) \delta_{i(\lambda+1,\phi f)j(\lambda,\phi,f)} - \frac{1}{\cos\Lambda_\lambda} \delta_{i(\lambda,(\phi \mod n_\phi)+1,f)j(\lambda,\phi,f)}. \end{aligned} \tag{3.59}$$

For simplicity of notation, we assumed a null contribution when one of the indices goes out of range or (in the case of soil moisture) points to a grid cell with no data.

3.9.7 Regularized Gaussian Committor

To have the proper coefficients α and β when we deal with a regularized pattern, we can notice that the assumption that X and A follow a jointly Gaussian distribution implies

that, for *any* M, $F = M^\top X$ and A also follow a jointly Gaussian distribution. We can then use the same formulas of eqs. (3.19) and (3.24), but applied to the 2 by 2 covariance matrix between F and A

$$\hat{\Sigma} = \begin{pmatrix} \sigma_F^2 & \mathbb{E}[FA] \\ \mathbb{E}[FA] & \sigma_A^2 \end{pmatrix} =: \hat{\Lambda}^{-1}, \tag{3.60}$$

and simply

$$q_{\mathcal{G}}(x) = \frac{1}{2}\mathrm{erfc}\left(\hat{\alpha} + \hat{\beta} M^\top x\right), \tag{3.61}$$

with

$$\hat{\sigma}^2 = \sigma_A^2 - \left(\frac{\mathbb{E}[FA]}{\sigma_F}\right)^2, \quad \hat{\alpha} = \frac{a}{\sqrt{2}\hat{\sigma}}, \quad \hat{\beta} = \frac{\mathbb{E}[FA]}{\sqrt{2}\hat{\sigma}\sigma_F^2}. \tag{3.62}$$

3.9.8 Effective Number of Independent Heatwaves

As we said in section 3.4.2, estimating the effective number of independent heatwaves can be challenging. The standard way of computing an effective data size for a time-series is the one presented in [Santer et al., 2000], where one uses the lag-1 autocorrelation coefficient r to rescale the total number of data points: $N_{\text{eff}} = N(1-r)/(1+r)$. However, when we consider heatwave events, they are not evenly spaced in time, so the whole approach does not make sense.

We can, tough, easily provide some bounds by observing that surely $N_{\text{y}} \leq N_{\text{eff}} \leq N_{\text{all}}$, where $N_{\text{all}} = N$ is the total number of heatwaves and N_{y} is the number of years that have at least a heatwave. Assuming that heatwaves at least a year apart are independent is definitely reasonable, if rather conservative. In fact, if we indeed compute the lag-1 autocorrelation coefficient for the time-series of $A(t)$, which gives $r = 0.9896$, and then estimate the decorrelation time of A as $\tau_{\text{decorr}} = (1+r)/(1-r)$, we get $\tau_{\text{decorr}} = 191$ days. Namely, it takes half a year to lose memory of the heatwave amplitude, and thus N_y is not only a lower bound for N_{eff}, but likely also very close to it.

If we apply this to our study of 14-day heatwaves we have $N_{\text{y}} = 2627 \lesssim N_{\text{eff}} \leq N_{\text{all}} = 30800$. Considering that we work with 8000 years of data, N_{y} tells us that there is a heatwave at least once every three years, and a year with a heatwave, on average, has $N_{\text{all}}/N_{\text{y}} \approx 12$ days for which $A(t) \geq a$.

3.9.9 Visualization of the Error Between Empirical and Gaussian Composites on Two Grid-Points

From fig. 3.2, we see that the biggest error we make when using the Gaussian composite is for the soil moisture variable. To investigate why this is the case, we show in fig. 3.14 (left) the joint and marginal distributions of the heatwave amplitude A (on the y-axis) and of one pixel of soil moisture S^i (on the x-axis). For comparison, we show the same for a pixel of the 500 hPa geopotential height Z^j in fig. 3.14 (right). While the marginal distributions of A and Z^j are approximately Gaussian (as it is shown from the black curve

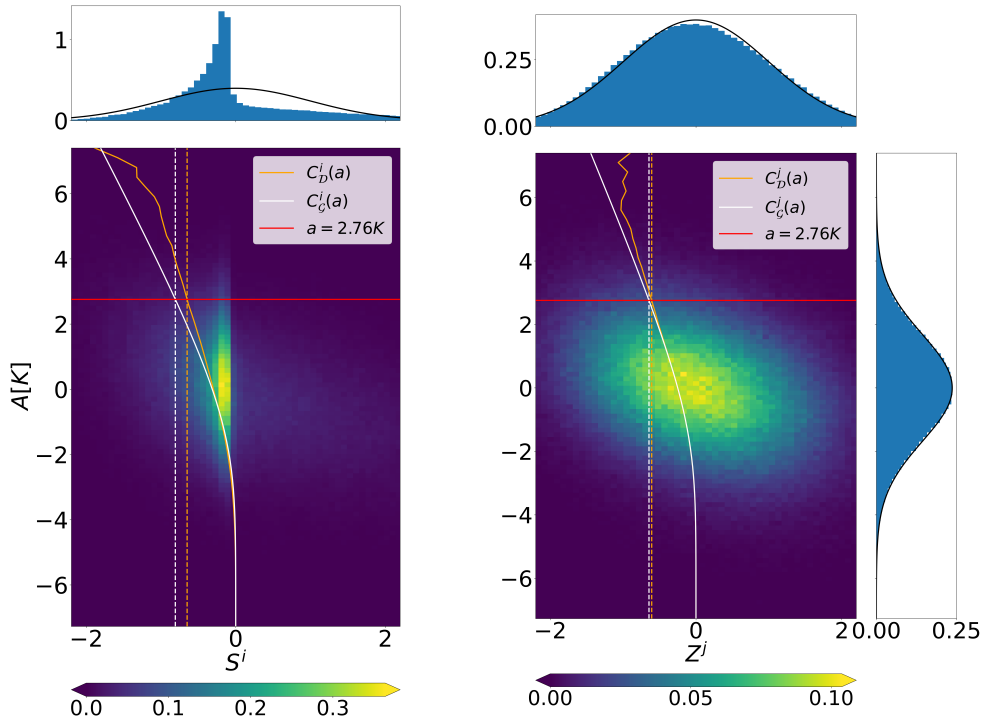


Figure 3.14: Comparison of the quality of the Gaussian and empirical composite map, for two grid-points and at different values of the heatwave threshold a. We show the results for a pixel over France of soil moisture S^i (left) and one over Greenland of 500 hPa geopotential height Z^j (right). For each panel, the main plot is the joint probability density function (PDF) of S^i or Z^j and the heatwave amplitude A, with marginal distributions displayed on top and to the side. In these plots the black line is a Gaussian fit. In the main plots, the orange line is the empirical composite as a function of a, while the white line is the one estimated through the Gaussian approximation. The red line is the threshold value of $a = 2.76K$ corresponding to the 5% most extreme heatwaves. The dotted vertical white and orange lines indicate the values of the empirical and Gaussian composites at this particular value of a. For Z^j the error is much smaller.

on the marginal plots of the figure), the one of S^i is clearly not: it is strongly skewed towards negative values of soil moisture anomaly, and exhibits fat tails. This is also reflected in the plot of the joint distribution of A and S^i on the one hand and Z^j on the other hand (heat maps in fig. 3.14), as only the latter has the shape of an ellipsoid.

In both panels of fig. 3.14, the orange curve shows the behavior of the empirical composite $C_{\mathcal{D}}$ as we change the threshold a, while the white curve is the behavior of the Gaussian composite $C_{\mathcal{G}}$. By construction, they both tend to 0 as $a \to -\infty$, since soil moisture S^i and 500 hPa geopotential height Z^j both have zero mean as they are anomalies.

When a is very small, then, the Gaussian composite provides a good, yet useless, approximation of the empirical composite as both are very close to zero. As the threshold increases beyond this trivial region, for soil moisture the two curves start to diverge already at $a \approx 0$, and thus show a significant distance when they reach the 95^{th} quantile of A, $a = 2.76$ K (red line). On the other hand, the approximation holds quite well in the case

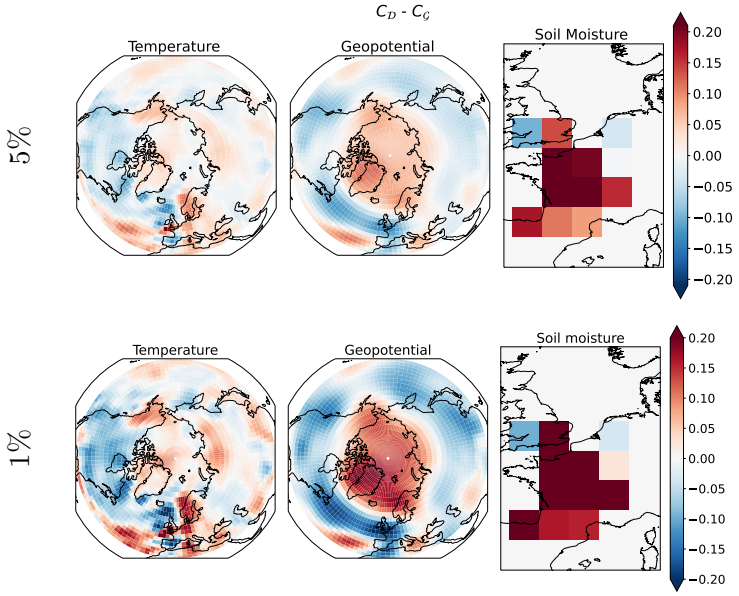


Figure 3.15: Maps of the difference between empirical composite maps and the ones estimated from the Gaussian approximation, for the 5% and 1% most extreme 14-day heatwaves over France. The patterns of this difference do not depend on the threshold a, varying only in intensity. The salient features of both temperature and geopotential height are well captured by the Gaussian approximation, with errors of the order of 30% at most.

of geopotential height, and the two curves separate significantly only when the sampling error kicks in (around $a = 5\,\mathrm{K}$) and throws off the empirical estimate.

3.9.10 Error Between Empirical and Gaussian Composites at Different Heatwave Thresholds

In fig. 3.15 we show the difference between the empirical composite and the Gaussian one (evaluated using eq. (3.15)) for the three fields of 2 meter air temperature, 500 hPa geopotential height and soil moisture evaluated for a corresponding to the 5% and 1% most extreme 14-day heatwaves. The striking result is that the pattern observed changes only in magnitude between extreme and very extreme events.

To give a quantitative measure of the error, in fig. 3.16, we evaluate the error using the norm ratio defined in eq. (3.25), for different threshold level a, showing that the error is around the 20% for the 5% most extreme events. Figure 3.16 gives more details on the behavior of the norm of the error shown in fig. 3.15 for different values of a. On the y-axis we represent the norm ratio introduced in section 3.4.1 (eq. (3.25)), which measures how distant (in norm) the Gaussian composite is from the empirical composite (normalized by the norm of the empirical composite). We calculated this norm for the three fields independently, and for the whole stack of them (gray line). The norm ratios of the 500 hPa geopotential height and the one of the 2 m air temperature stay pretty close to the norm of the stack, showing values below 0.3 even for events which represents the 1% most extreme

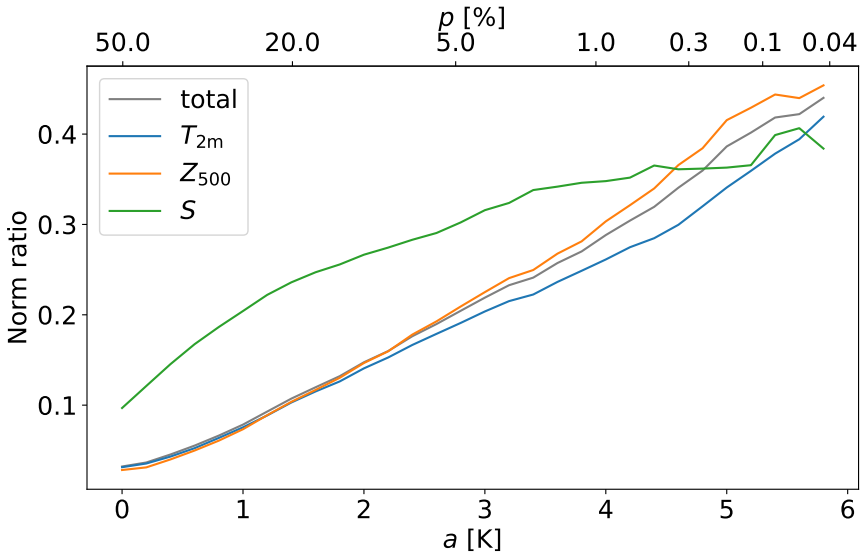


Figure 3.16: Norm ratio (see eq. (25) of the main text) of the difference between the empirical composite map and the Gaussian approximated one as a function of the threshold value used to define a heatwave event a. The total norm ratio is in gray, while the colors represent the norm ratio for each of the three fields (namely 2 m temperature anomaly, 500 hPa geopotential height anomaly and soil moisture anomaly). For events which are the 1% most extreme ones of the PlaSim dataset, the relative error is always below the 30%. The bottom x-axis a is the threshold value used to define a heatwave event from the distribution of the 14-day heatwave amplitude A. On the top x-axis, p is the respective percentile value corresponding to a given a.

ones in the dataset (the x-axis on the top shows the respective percentile of rareness of the a, which is on the bottom x-axis). Soil moisture has a different behavior, showing higher values of the norm ratio for much less extreme events. This is possibly due to a violation of the Gaussian approximation assumption, as we showed for a single pixel in fig. 3.14.

3.9.11 Field-Wise Norm Ratio of Composite Maps at Different Values of T and τ

In tables 3.7 to 3.9, we show the norm ratio as defined in eq. (3.25), but computed independently for the three climate variables of the PlaSim dataset. Values related to temperature peak at $T = 1$ and for small delay time. Soil moisture exhibits a clear monotonic trend with respect to T, while not being very sensitive to the lead time τ. Finally, the geopotential height field shows the more complex structure, with the highest errors happening at intermediate values of both T and τ.

Soil moisture has only 12 pixels, compared to the 2816 of the other two fields, so its contribution to the total norm ratio (table 3.3) is negligible. Outside the small region of low values of T and τ, the norm ratio of the temperature field is almost constant, so the structure visible in table 3.3 is mostly due to the geopotential height field.

Temperature

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.26	0.30	0.31	0.22	0.21	0.22	0.22	0.21	0.19	0.19	0.19
3	0.25	0.29	0.26	0.19	0.19	0.20	0.20	0.19	0.18	0.17	0.18
7	0.22	0.24	0.21	0.19	0.19	0.20	0.20	0.19	0.18	0.18	0.16
14	0.19	0.20	0.20	0.19	0.20	0.20	0.21	0.21	0.19	0.18	0.18
30	0.18	0.19	0.20	0.20	0.20	0.20	0.20	0.19	0.19	0.20	0.20

Table 3.7: Values of the norm ratio between Gaussian and empirical composites computed for 2 m temperature anomaly.

Geopotential

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.25	0.26	0.28	0.30	0.31	0.33	0.36	0.34	0.29	0.26	0.27
3	0.24	0.23	0.26	0.28	0.29	0.31	0.34	0.31	0.27	0.25	0.25
7	0.21	0.23	0.27	0.30	0.33	0.38	0.37	0.32	0.30	0.30	0.25
14	0.21	0.24	0.30	0.34	0.34	0.33	0.33	0.34	0.32	0.28	0.28
30	0.22	0.25	0.29	0.32	0.33	0.33	0.34	0.32	0.30	0.31	0.31

Table 3.8: Values of the norm ratio between Gaussian and empirical composites computed for 500 hPa geopotential height anomaly

Soil moisture

T [days] \ τ [days]	0	3	6	9	12	15	18	21	24	27	30
1	0.10	0.11	0.12	0.12	0.13	0.13	0.13	0.13	0.14	0.14	0.14
3	0.14	0.12	0.12	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12
7	0.22	0.20	0.19	0.18	0.18	0.17	0.16	0.16	0.15	0.15	0.14
14	0.30	0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.20	0.19
30	0.40	0.38	0.36	0.35	0.33	0.31	0.30	0.28	0.27	0.25	0.24

Table 3.9: Values of the norm ratio between Gaussian and empirical composites computed for soil moisture anomaly.

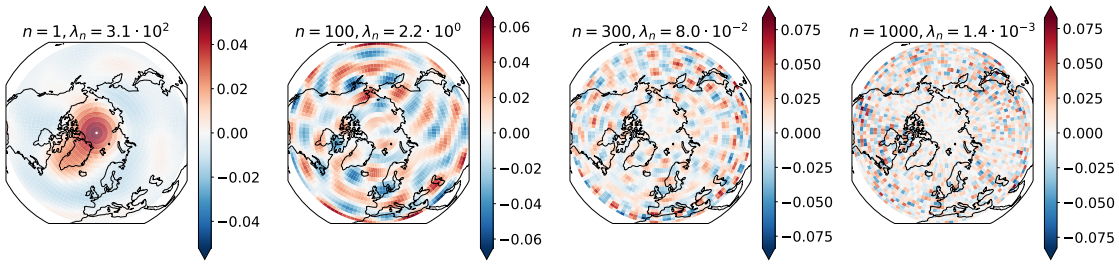


Figure 3.17: Four EOFs e_n. The higher the n the smaller the spatial scales of the characteristic features represented. On top of each plot we report the value of λ_n.

3.9.12 Asymptotic Behavior of the Regularized Projection Pattern

In this section we discuss the effect of the regularization coefficient ϵ on the optimal projection pattern M_ϵ, and in particular why a highly regularized projection pattern may look similar to the composite map.

To do so we move to the basis of Empirical Orthogonal Functions (EOFs) [Hannachi et al., 2007], which diagonalizes the covariance matrix Σ_{XX}. We call its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots \geq \lambda_d$, and the corresponding eigenvectors e_n. Here we will use as an example the prediction of $T = 14$ day heatwaves at delay time $\tau = 0$, performed on the ERA5 reanalysis dataset. In this context there are $d = 2816$ degrees of freedom, corresponding to the pixels of the 500 hPa geopotential height anomaly field.

The first important point is that as n increases, the variance λ_n explained by e_n decreases, and so does decrease the typical spatial size of the features that appear in e_n (fig. 3.17). In particular, features with a typical size of the order of the synoptic scale are represented around $n = 100$. Moreover, almost two thirds of the EOFs ($n > 1000$) explain less than 0.05 % of the variance and are extremely noisy.

Second, the Gaussian composite map is proportional to the correlation map Σ_{XA} (eq. (3.15)), and when we write it in the EOF basis,

$$C_{\mathcal{G}} \propto \Sigma_{XA} = \sum_{n=1}^{d} c_n e_n, \tag{3.63}$$

it is dominated by EOFs at low values of n (see black lines in fig. 3.18), and thus it appears spatially smooth.

Third, in the EOF representation the non-regularized ($\epsilon = 0$) projection map is written as

$$M_0 = \sum_{n=1}^{d} M_n^0 e_n \propto \Sigma_{XX}^{-1} \Sigma_{XA} \propto \sum_{n=1}^{d} \frac{c_n}{\lambda_n} e_n, \tag{3.64}$$

and λ_n goes to zero much faster than c_n (dashed dark blue lines in fig. 3.18), resulting in M_0 being dominated by large n, high spatial frequency modes. This is what makes the non-regularized pattern utterly non-interpretable.

If we perform L_2 regularization, the aim of the regularization coefficient is to prevent the contribution of these high frequency modes to explode, making them proportional to

their values in the composite map:

$$M_\epsilon \propto (\Sigma_{XX} + \epsilon\mathbb{I})^{-1}\Sigma_{XA} \propto \sum_{n=1}^{d} \frac{c_n}{\lambda_n + \epsilon} e_n \approx \sum_{n=1}^{n_\epsilon - 1} M_n^0 e_n + \frac{1}{\epsilon} \sum_{n=n_\epsilon}^{d} c_n e_n, \tag{3.65}$$

where $\lambda_{n_\epsilon} \approx \epsilon$. It is then clear that as ϵ increases $n_\epsilon \to 1$, and the projection map smoothly converges to the composite map (left panel of fig. 3.18).

On the other hand, when we perform H_2 regularization, as we do in this work, we regularize with matrix W, which doesn't share the same eigenvectors of Σ_{XX}. We can in any case write W in the EOF basis as

$$W = \sum_{mn} W_{nm} e_n e_m^\top. \tag{3.66}$$

If we compute the terms W_{mn}, we notice that $W_{nn} \gg \max_{m \neq n} |W_{nm}|$. We can then say that the W is almost diagonal and thus

$$M_\epsilon \propto (\Sigma_{XX} + \epsilon W)^{-1}\Sigma_{XA} \approx \sum_{n=1}^{d} \frac{c_n}{\lambda_n + \epsilon W_{nn}} e_n. \tag{3.67}$$

This lets us apply a similar reasoning to the one explained above for L_2 regularization, where W_{nn} is the norm of the spatial gradient of EOF e_n, which, considering the spatial structure of the EOFs (fig. 3.17), clearly increases with n. For this reason, when we increase ϵ, we remove the high spatial frequencies faster than we would with L_2 regularization (right panel of fig. 3.18). On the other hand, for very high regularization, the approximation of W being diagonal falls apart, and the high frequencies are brought back to achieve a spatially uniform pattern, similarly to the Fourier representation of a square wave. So there is no asymptotic convergence to the composite map (brown curve). However, for intermediate values of ϵ (yellow, orange and red curves), the projection pattern is smoothed in a similar way as with L_2 regularization, and thus it may look similar to the composite map.

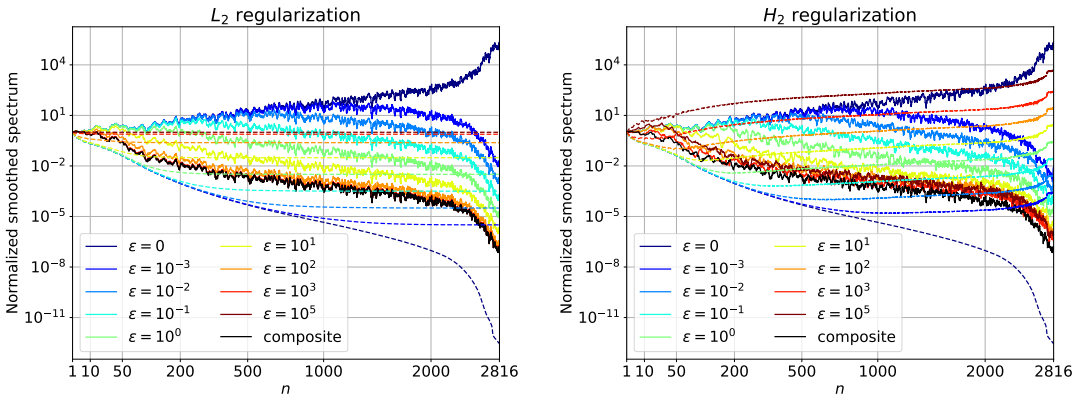
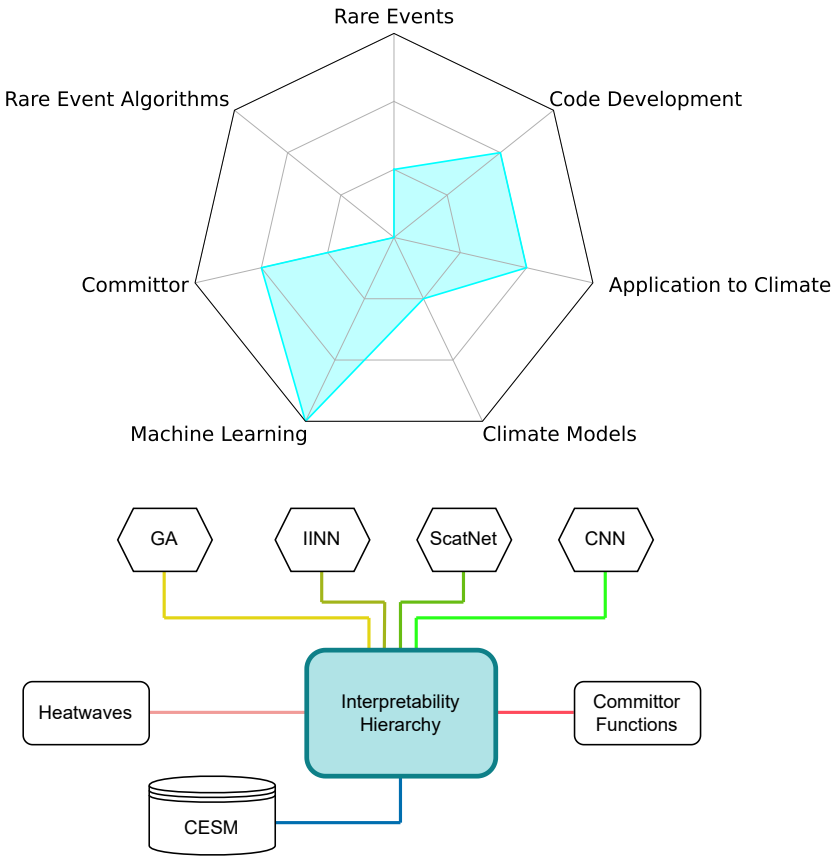


Figure 3.18: EOF spectra $|M_n^\epsilon|$ of the projection pattern at different values of the regularization coefficient ϵ (solid lines) when penalizing the L_2 norm of the pattern (left) or its spatial gradient (right). All spectra are normalized so that the term at $n = 1$ has unitary values. To ease the visualization, the spectra have been smoothed with a running average. On the left panel the dashed lines represent the spectra of the regularized covariance matrix: $\lambda_n + \epsilon$. On the right panel they represent the diagonal part of the gradient regularized covariance matrix in the original EOF basis, i.e. $\lambda_n + \epsilon e_n^\top W e_n$. The black line is the spectrum c_n of the Gaussian composite map. For high n (i.e. EOFs with small spatial scales), the values of λ_n decay faster than those of c_n, which makes the non-regularized pattern extremely noisy. For L_2 regularization, increasing values of ϵ progressively reduce the contribution of EOFs at high n, and the projection pattern converges to the composite. On the other hand, H_2 regularization directly penalizes the spatial gradient of the projection pattern, so the small scales are first suppressed and then brought back to achieve the spatially uniform pattern. These spectra are presented for $T = 14$ and $\tau = 0$ on the ERA5 dataset, where $\epsilon_{\text{best}} = 10$ (light green lines).

Chapter 4

Interpretability of Prediction in a Hierarchy of Machine Learning Models for Extreme Heatwaves



In chapter 3, we compared the performance of the Gaussian approximation with that of standard Convolutional Neural Networks (CNNs) for the task of predicting extreme heatwaves over France. In this chapter we explore in more detail the balance between complexity, performance and interpretability. In particular, in chapter 3 we saw that, when enough data is available, the CNNs are able to extract more information from the data with respect to the Gaussian approximation. However, what this extra information was remained an open question. This chapter is devoted to answering this question. To do so, we will use a hierarchy of increasingly complex machine learnings models, and compare their performance in predicting extreme heatwaves over France. Then we will use a combination of post-hoc explainability tools and intrinsically interpretable models, such as the Intrinsically Interpretable Neural Network presented in section 2.3, to pinpoint what information the more complex networks are capturing that the Gaussian approximation is missing.

Another remark is that in chapter 3 we used data from the PlaSim model and selected temperature, 500 hPa geopotential height and soil moisture as predictors. This was done to have a nice parallel between composite maps and optimal projection patterns, but, in practice, temperature doesn't provide additional predictability once we have the geopotential height field in our set of predictors [Miloshevich et al., 2023a]. Thus, since in this chapter we focus specifically on finding the important sources of predictability, we will use only 500 hPa geopotential height and soil moisture. Moreover, instead of the PlaSim model, we will use the higher resolution and more physically accurate data generated with the Community Earth System Model (CESM) [Hurrell et al., 2013].

What follows is a draft of the paper I wrote together with Amaury Lancelin[1][2]. The paper was submitted to the American Meteorological Society's Artificial Intelligence for the Earth Systems (AIES) journal for review, and the related preprint is available on arXiv at `https://arxiv.org/abs/2410.00984`, with minor differences with respect to the version included in this manuscript. Amaury Lancelin's contribution was in the use of scattering networks and in the post-hoc explanation of the CNNs via the expected gradient method. The rest was my work, as well as the general direction of this project.

Abstract

Extreme weather events, especially heatwaves, pose significant risks to human health, ecosystems, and infrastructure. Accurate prediction of these rare events is essential for effective early warnings and mitigation but remains challenging due to their infrequency. This paper investigates the potential of machine learning (ML) techniques for forecasting extreme heatwaves, focusing on direct probabilistic forecasts from initial conditions and comparing models of increasing complexity.

[1] *LMD/IPSL, CNRS, ENS, Université PSL, École Polytechnique, Institut Polytechnique de Paris, Sorbonne Université, Paris, France*

[2] *RTE France*

More precisely, we evaluate a hierarchy of ML models that ranges from a global Gaussian approximation (GA) to deep Convolutional Neural Networks (CNNs), with the intermediate steps of a simple Intrinsically Interpretable Neural Network (IINN) and a model using the Scattering Transform (ScatNet). Our findings reveal that while CNNs provide higher accuracy, their black-box nature severely limits interpretability. To address this, we leverage recent Explainable Artificial Intelligence (XAI) tools to gain some insight into their predictions. In contrast, ScatNet achieves similar performance to CNNs while providing greater transparency, identifying key scales and patterns in the data that drive predictions.

This study underscores the potential of interpretability in ML models for climate science, demonstrating that simpler models can sometimes rival the performance of more complex counterparts, all the while being much easier to understand. This gained interpretability is crucial for building trust in model predictions and uncovering new scientific insights, ultimately advancing our understanding and management of extreme weather events.

Significance Statement

The purpose of this work is to test increasingly complex machine learning models on the task of forecasting extreme heatwaves and explain their prediction. This is important to quantify what additional information the more complex models are able to capture. We find that answering this question with a black-box model and explainability techniques is not efficient. Indeed, the explanations are mainly qualitative and don't add much to what was known from the simplest linear models. On the other hand, by using an inherently interpretable architecture design, we are able to precisely quantify the sources of additional information while maintaining the same predictive skill of the black-box model.

4.1 Introduction

Extreme events in weather and climate are responsible for the most detrimental observed and projected impacts of climate change [Seneviratne et al., 2012]. Heatwaves, in particular, caused significant increase in mortality in 2003 over Western Europe [Fouillet et al., 2006], in 2010 in Russia [Barriopedro et al., 2011] and in 2021 in Canada [Henderson et al., 2022]. Moreover, the impact of extreme heatwaves extends to losses in the agricultural sector as well as the endangerment of ecosystems [Seneviratne et al., 2021].

It is thus of paramount importance to improve our understanding of these events and to be able to accurately forecast them sufficiently in advance. However, the events with the highest impact are the rarest, which means they are very few in observational records and properly simulating them with climate models is expensive [Miloshevich et al., 2023a; Ragone et al., 2018].

There are several statistical tools that are tailored to the study of extreme events. For instance Extreme Value Theory extrapolates from existing limited data [Ghil et al., 2011;

Keellings and Waylen, 2014] and rare event algorithms make simulating very extreme events considerably more efficient [Ragone et al., 2018; Ragone and Bouchet, 2021]. However, these tools mainly focus on the statistics of extremes (such as accurately estimating their return times), and are not very useful for prediction.

We thus argue that there is great potential for improving prediction of extreme events. Moreover, rare event algorithms require specific score functions to work efficiently, and the optimal score function is connected to the committor function [Chraibi et al., 2020; Rolland et al., 2016], which is the conditional probability of an extreme happening in the future, given the current state of the world. The committor function is closely entwined with prediction, and thus, improving our prediction of extreme events would also allow us to simulate them more efficiently.

There have been recent advances in estimating committor functions, such as analogue Markov chain methods [Lucente et al., 2022a] or Stochastic Weather Generator [Miloshevich et al., 2023b]. However, the most promising direction is the one using machine learning techniques.

Indeed, the field of Artificial Intelligence has known an exponential explosion in the recent decades, with significant advances in the climate community. In particular, the prediction problem has been tackled with tools ranging from simple neural networks focusing on specific tasks [Asadollah et al., 2021; Khan et al., 2021; Petersik and Dijkstra, 2020], to full foundation models that forecast the global weather with accuracy comparable to state of the art numerical simulations [Bi et al., 2023; Lam et al., 2023; Nguyen et al., 2023].

This remarkable progress has seen machine learning systems become more and more complex and achieve more and more accurate predictions. However, the added complexity means that these new models require vast amounts of data and extensive resources to be trained. Moreover, they behave as black boxes, so, even though their prediction is highly accurate, researchers may struggle to trust them.

Indeed, being able to explain why the model reached a particular decision can be far more valuable than a few percents of improved accuracy [Rudin, 2019]. For this reason, there has been an intense proliferation of post-hoc explainability methods [Murdoch et al., 2019], mainly imported from the computer science community, which aim at providing the user with human-understandable insight.

These methods have been very successful in elucidating model predictions, which sometimes even led to the discovery of new science [McGovern et al., 2019; Toms et al., 2020; Barnes et al., 2020]. However, they are fundamentally flawed as the explanations provided only approximate the true prediction of the model [Montavon et al., 2018; Yang et al., 2024] and are either understandable but simplistic or accurate but hard to analyze [Murdoch et al., 2019]. Even worse, sometimes different explanation methods can yield contrasting explanations [Mamalakis et al., 2022a,b]. On the other hand, a much better, though underdeveloped, direction is that of developing machine learning models which are interpretable by design and thus don't require convoluted and often unstable explainability

methods to be understood [Murdoch et al., 2019].

Indeed, it has been shown that such models can have minimal reduction in performance with respect to complex black box models [Barnes et al., 2022; Rudin, 2019], which means that the complexity of the latter is often unjustified. The drawback is that interpretable models are usually harder to design and train, shifting the load from computers to the brain power of researchers [Murdoch et al., 2019; Rudin, 2019; Yang et al., 2024].

Fortunately, for the task of forecasting extreme heatwaves, recent work has shown that very simple models like regularized linear regression can be very effective [Mascolo et al., 2024a]. In this work, we expand on this finding by training a hierarchy of increasingly complex machine learning models, from simple linear regression to intermediate interpretable architectures to fully fledged deep Convolutional Neural Networks (CNNs).

We quantify in which conditions a more complex model actually performs better, and when this is the case we try to understand what *extra* information these models are able to capture with respect to their simpler counterparts. Indeed, we show that post-hoc explanations of the CNN show something very similar to what was already clear from linear regression, and thus are not very useful. On the contrary, when we use an interpretable variant of scattering networks [Bruna and Mallat, 2013], we achieve the same skill of the CNN, but we are able to pinpoint much more precisely the source of improvement with respect to the linear regression baseline.

4.2 Data and methods

4.2.1 Data

We use the 1000 year-long control run of the Community Earth System Model (CESM) version 1.2.2 [Hurrell et al., 2013], already used in Ragone and Bouchet [2021]. Only atmosphere and land components are dynamic, while sea surface temperature, sea ice and greenhouse gas concentrations are prescribed to reproduce a stationary climate that resembles the one of the year 2000. The model is run with a 0.9° resolution in latitude, 1.25° in longitude and 26 pressure levels. Previous studies [Ragone and Bouchet, 2021; Miloshevich et al., 2023c] have shown that, with this setup, CESM is able to accurately reproduce the relevant atmospheric phenomena connected to heatwaves, for instance planetary teleconnection patterns [Miloshevich et al., 2023c].

Since we focus on summer heatwaves, we will use daily averaged data for the months of June, July and August.

Of the total 1000 years of data available, we keep the last 200 for testing, while the first 800 are split in training and validation according to a 5-fold cross validation process.

4.2.2 Heatwave amplitude

In the literature there are many different definitions of what a heatwave is [Perkins, 2015], with most of them involving hard thresholds that need to be overcome for a specific

amount of consecutive days. In this work, we instead follow the definition used in Gálfi et al. [2019]; Gálfi and Lucarini [2021]; Ragone et al. [2018]; Ragone and Bouchet [2021]; Jacques-Dumas et al. [2023]; Miloshevich et al. [2023a]; Mascolo et al. [2024a], which gives a continuously varying heatwave amplitude A and allows easily to control for the heatwave duration and intensity.

If $T_{2\mathrm{m}}$ is the 2 m temperature anomaly, we define the heatwave amplitude at $A(t)$ as

$$A(t) := \frac{1}{T}\int_t^{t+T}\left(\frac{1}{\mathcal{A}}\int_{\mathcal{A}} T_{2\mathrm{m}}(\vec{r},u)d\vec{r}\right)du, \tag{4.1}$$

where T is the heatwave duration in days and $\mathcal{A}$ is the geographical region of interest. In this work, $\mathcal{A}$ will be the region of France, which has a size of roughly 1000 km and thus sits nicely at the scale of cyclones and anticyclones, which is the relevant one for large-scale atmospheric dynamics. Also, studying heatwaves involving a whole country can be very relevant for policymakers [Barriopedro et al., 2011; Fouillet et al., 2006]. Considering that longer lasting heatwaves have a much higher impact than shorter ones [Seneviratne et al., 2012, 2021; Meehl and Tebaldi, 2004; Amengual et al., 2014; Ragone et al., 2018], we focus here on two-week heatwaves ($T = 14$).

4.2.3 Predictors

To forecast a heatwave at time t, we will use a set of predictors $X(t-\tau)$, where τ is the lead time. Since our data has only land and atmosphere dynamical components, we will use as predictors the geopotential height anomaly at 500 hPa (Z500) over the whole Northern Hemisphere and soil moisture anomalies (SM) over France.

While the geopotential height anomaly at the middle of the troposphere gives good insight into the wind flow (thanks to the geostrophic approximation), soil moisture acts as an important modulator of the likelihood of heatwaves at mid-latitudes, as it controls the evaporative cooling potential of the surface [Perkins, 2015; Miloshevich et al., 2023a; Benson and Dirmeyer, 2021; D'Andrea et al., 2006; Fischer et al., 2007; Hirschi et al., 2011; Lorenz et al., 2010; Rowntree and Bolton, 1983; Schubert et al., 2014; Shukla and Mintz, 1982; Stefanon et al., 2012; Vargas Zeppetello and Battisti, 2020; Zeppetello et al., 2022; Zhou et al., 2019; Vautard et al., 2007]. To facilitate the work of neural networks, each scalar predictor is independently standardized to zero mean and unitary standard deviation.

Since this study focuses more on methodology, we will focus on $\tau = 0$, namely on predicting the average temperature of the next two weeks.

4.2.4 Probabilistic regression

The climate system is chaotic, so any forecast should be probabilistic. More precisely, here we want to predict the distribution $\hat{p}(A|X)$ of the heatwave amplitude A, given the current state of the predictors X. Then we can immediately compute probabilities of the heatwave amplitude exceeding a given threshold a: $q(X) = \mathbb{P}(A > a|X)$. These

probabilities are the committor function for an event more extreme than a, and are the fundamental tools for prediction of extreme events [Mascolo et al., 2024a].

To quantify the goodness of our prediction, we will then use three metrics: two for a regression task, and one for the task of classifying the 5% most extreme events.

The first two are the Negative Log Likelihood (NLL)

$$NLL = -\frac{1}{N}\sum_{i=1}^{N} \log \hat{p}(A_i|X_i), \tag{4.2}$$

and the Continuous Ranked Probability Score (CRPS)

$$CRPS = \frac{1}{N}\sum_{i=1}^{N} \int_{-\infty}^{+\infty} \left(\mathbb{1}_{a \geq A_i} - \int_{-\infty}^{a} \hat{p}(a'|X_i) da' \right)^2 da, \tag{4.3}$$

where $\mathbb{1}$ is the indicator function.

The third one is the Binary Cross Entropy (BCE)

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} \left(\mathbb{1}_{A_i < a_5} \log \left(\int_{-\infty}^{a_5} \hat{p}(a'|X_i) da' \right) + \mathbb{1}_{A_i \geq a_5} \log \left(\int_{a_5}^{+\infty} \hat{p}(a'|X_i) da' \right) \right), \tag{4.4}$$

where $a_5 = 3.11\,\mathrm{K}$ is the threshold that defines the 5% most extreme heatwaves.

To have a reference, we use the zero-order prediction given by the climatology $\hat{p}_{\mathrm{clim}}(A)$ which doesn't use any information about the predictors, but rather fits the distribution of A on the training set and uses it to compare with the data in the validation and test sets. For the first two metrics we compute $\hat{p}_{\mathrm{clim}}(A)$ using a kernel density estimate, while for the BCE, we simply assume $\int_{a_5}^{+\infty} \hat{p}_{\mathrm{clim}}(a')da' = 0.05$.

Once we have the values of the climatological metrics, for a prediction performed by a $\{MODEL\}$, we can compute the skill score of a $\{METRIC\}$ as

$$\{METRIC\}S = 1 - \frac{\{METRIC\}_{\{MODEL\}}}{\{METRIC\}_{\mathrm{clim}}} \tag{4.5}$$

Using these skill scores, a perfect prediction will give a value of 1, while a prediction that is worse than the climatology will result in negative values.

4.3 The model hierarchy

To perform probabilistic regression, we will approximate the conditional distribution of the heatwave amplitude $\hat{p}(A|X)$ as a Gaussian distribution with mean $\hat{\mu}(X;\theta)$ and variance $\hat{\sigma}^2(X;\theta)$. We will use increasingly complex models to parameterize $\hat{\mu}$ and $\hat{\sigma}$. A summary of this model hierarchy is presented in table 4.1.

4.3.1 Gaussian approximation

The simplest option after the climatology is to perform a linear regression of A against X. The underlying assumption of this method is that the joint distribution of X and A is

Method	$\hat{\mu}(X;\theta)$	$\hat{\sigma}(X;\theta)$	trainable parameters	non-trainable parameters	hyperparameters
GA	$M \cdot X$	σ	27 425	0	1
IINN	$g_\mu(M \cdot X)$	$s(g_\sigma(M \cdot X))$	55 058*	0	10
ScatNet	$\beta_\mu \cdot \phi(X)$	$s(\beta_\sigma \cdot \phi(X))$	19 930*	656 640*	5
CNN	$g_\mu(X)$	$s(g_\sigma(X))$	684 000*	0	10

Table 4.1: Complexity of the hierarchy of probabilistic prediction models. Values marked with * denote the number of parameters at the optimum with respect to hyperparameters. M is the projection pattern, g_μ and g_σ are general non-linear functions parameterized by neural networks, ϕ is the scattering transform, β_μ and β_σ are projection patterns in the transformed space, and s is the softplus function ($s(x) = \log(1 + \exp(x))$), which ensures that $\hat{\sigma}(X;\theta) > 0$.

a multivariate Gaussian [Mascolo et al., 2024a], and results in a prediction with constant variance $\hat{\sigma}(X;\theta) = \sigma$ and $\hat{\mu}(X;\theta) = M \cdot X$. As explained in Mascolo et al. [2024a], M has the same dimensions of X and is an optimal projection pattern, that condenses all the important information for heatwave prediction into the scalar index $F = M \cdot X$. Training is performed in one step, with

$$\begin{cases} M &= \arg\min_M \frac{1}{N} \sum_{i=1}^{N} (A_i - M \cdot X_i)^2 + \epsilon H_2(M) \\ \sigma^2 &= \mathrm{Var}[A] - \frac{(\mathbb{E}[FA])^2}{\mathrm{Var}[F]} \end{cases}, \tag{4.6}$$

where $\mathbb{E}[\bullet]$ and $\mathrm{Var}[\bullet]$ denote respectively expectation and variance of $\bullet$.

For high dimensional climate data, we need to regularize the projection pattern M, and this is achieved by penalizing the norm of its spatial gradient $H_2(M)$, which forces M to be spatially smooth. See Mascolo et al. [2024a] for more details.

For this method the trainable parameters are $\theta = \{M, \sigma\}$ and the single hyperparameter is the regularization coefficient ϵ.

4.3.2 Intrinsically Interpretable Neural Network

To go a step further from the Gaussian approximation, we still project onto an optimal index $F = M \cdot X$, but then we apply a (relatively small) fully connected neural network to compute $\hat{\mu}$ and $\hat{\sigma}$. We call this method Intrinsically Interpretable Neural Network (IINN) [Lovo et al., 2023], as the prediction is decomposed in a linear projection, of which we can visualize the projection pattern M (as for GA) and two non-linear functions $\mathbb{R} \to \mathbb{R}$ which are as well very easy to visualize.

In this case, training is performed by gradient descent of the CRPS loss with the added regularization $+\epsilon H_2(M)$. The trainable parameters are the components of M and the weights and biases of the following network, while hyperparameters are ϵ, the architecture of the network (number of layers and neurons per layer), plus the usual hyperparameters concerning training, like learning rate and batch size.

4.3.3 Scattering Transform

scattering networks were introduced by Mallat [2012] as an alternative to traditional Convolutional Neural Networks (CNNs). They aim to create a representation of the data that is stable to local deformations and translations while still preserving essential information. It computes a set of features by applying convolutions with a set of *fixed* wavelet filters to the input signal (such as an image). Given this set of features, we can then apply a classifier or a regressor to obtain the final prediction, depending on the task we are interested in. The regressor could be a simple linear layer, since the features are already extracted. One can also combine the Scattering Transform with an extra CNN on top of it, as it is done in Oyallon et al. [2019].

Since its introduction, the Scattering Transform has shown promising results for various physical fields and tasks [Cheng et al., 2024], especially when the data have limited training samples or when the task requires robustness to deformations and translations. It has been applied, for instance, to quantum chemical energy regression and the prediction of molecular properties [Eickenberg et al., 2018], and to astrophysics through the statistical description of the interstellar medium [Allys et al., 2019]. This work is, to the best of our knowledge, the first time that scattering networks are applied to atmospheric circulation.

The main steps of the scattering transform are as follows:

- Wavelet Transform and non-linearity: The input signal is convolved with a set of wavelet filters to extract frequency and phase information at different scales. The set of wavelet filters is arranged to tile the Fourier space, which is discretized by scales and orientations. The discretization involves two key hyperparameters: the number of scales J and the number of orientations L. Consequently, each wavelet filter is located in a distinct position in Fourier space, defined by a scale $j \in \{1, \dots, J\}$ and an orientation $l \in \{1, \dots, L\}$. The result of the wavelet transform is then passed through a point-wise non-linear operator, namely the modulus operator.

- Pooling and Aggregation: To achieve translation invariance, a downsampling operation is applied to the transformed coefficients. This helps in reducing the spatial resolution of the representation while retaining the essential information. We apply a local averaging operator (typically a Gaussian smoothing function), followed by an appropriate downsampling by a factor 2^J.

- Recursive Operation: The same wavelet transform, non-linearity, and pooling steps can be repeated a second time to create a two-layered scattering representation. Each layer captures different levels of abstraction and invariance to transformations. It's not useful to compute higher order scatterings, because their energy is negligible [Bruna and Mallat, 2013].

One can observe the resemblance in the way CNNs operate. However, a major difference with CNNs is that convolution filters in scattering networks are not learned but fixed from the beginning (to wavelet filters). We let the reader refer to Bruna and Mallat [2013]

and Oyallon et al. [2019] for further details on the scattering transform. We used the implementation offered by the Python package *Kymatio* [Andreux et al., 2020].

In our work, the idea was to keep the overall architecture as simple as possible, so we chose to predict $\hat{\mu}(X;\theta)$ and $\hat{\sigma}(X;\theta)$ by simply applying a fully connected layer to the (flattened) features obtained from the concatenation of the scattering transform of the 500 hPa geopotential height anomaly field and the raw pixels of soil moisture over France. We call the resulting model *ScatNet*.

The architecture hyperparameters to tune are the number of scales J and the number of orientations L at each scale for the wavelet filters. Additionally, there is the maximum order of the scattering, which can be either 0, 1, or 2 [see Bruna and Mallat, 2013]. After hyperparameter optimization, we fixed $J = 3$, $L = 8$, and the maximum order to 1. The hyperparameters related to the training phase are learning rate and batch size.

4.3.4 Convolutional Neural Network

The most complex model in this study is a Convolutional Neural Network (CNN), for which $\hat{\mu}(X;\theta)$ and $\hat{\sigma}(X;\theta)$ are the result of convolutional layers followed by fully connected ones. In this case, we feed the input X to the neural network as a two-channel image, with the two 'colors' corresponding to the geopotential height and soil moisture fields, the latter being set to 0 outside France. The network is trained by gradient descent of the CRPS loss, and the hyperparameters of this model are its architecture plus learning rate and batch size.

4.3.5 Hyperparameter optimization

Hyperparameters of IINN, ScatNet and CNN are optimized using a Bayesian search algorithm provided by the *optuna* [Akiba et al., 2019] Python package. The best combination is the one that gives the highest validation BCES. We tried also optimizing with respect to the other metrics and the results do not change significantly.

The regularization coefficient ϵ of GA and IINN is treated separately, as it controls how smooth, and thus interpretable, the projection pattern M is. At the best configuration of hyperparameters found by *optuna*, we systematically try logarithmically spaced values of ϵ. The skills of both GA and IINN display a broad plateau for intermediate values of ϵ, with poor performance for either too high or too low values (see figs. 4.7 and 4.8). We then choose ϵ as the one that leads the smoothest projection pattern without loss of performance.

4.4 Performance

In table 4.2, we show the performance of the hierarchy of models on the test set at the optimal values for hyperparameters. Since at the end of the k-fold cross validation process we have 5 trained models, we evaluate all of them on the test set, which allows us to obtain error bars on the skill of the networks.

As expected, the more complex architectures perform better than the simple ones, but GA and IINN are very close, showing that once we project the high dimensional predictors X onto the scalar variable $F = M \cdot X$, allowing for non-linearity in the forms of $\hat{\mu}(F)$ and $\hat{\sigma}(F)$ doesn't improve significantly the performance. Similarly, ScatNet and CNN have very comparable skills, suggesting that the important part of the prediction is capturing the local structures in the data, which can be achieved as easily by a deterministic wavelet transform as with more complex learned convolutional filters. Once again, after the important features are extracted, there appears to be no benefit in the more complex parametrization of $\hat{\mu}$ and $\hat{\sigma}$ provided by the CNN.

		Metric		
		CRPSS	NLLS	BCES
Model	GA	0.2864 ± 0.0009	0.2169 ± 0.0009	0.293 ± 0.001
	IINN	0.287 ± 0.002	0.217 ± 0.002	0.291 ± 0.003
	ScatNet	0.3097 ± 0.0007	$\mathbf{0.246 \pm 0.003}$	$\mathbf{0.314 \pm 0.005}$
	CNN	$\mathbf{0.310 \pm 0.003}$	0.245 ± 0.007	0.311 ± 0.008

Table 4.2: Test skills (the higher, the better) of the different models, shown as mean and standard deviation over the 5 folds. In bold the best performing model according to each of the three metrics.

4.4.1 Training on a smaller dataset

Before, we observed that more complex models have a better performance. However, this result is valid when training on 640 years (and validating on 160). Often, climate datasets are much shorter, and machine learning techniques are notoriously data-hungry. Thus, we perform a second experiment using a total of 80 years (size comparable to reanalysis datasets), 64 for training and 16 for validation, as usual optimizing hyperparameters to maximize validation skill.

The results presented in table 4.3 show a reversal of the ranks, with the Gaussian approximation clearly outmatching all other methods. In Mascolo et al. [2024a] the authors already point out the remarkable robustness of the Gaussian approximation to lack of data, but here we see that even the relatively similar method of the IINN suffers a lot from the dataset size, especially for the classification task, with a BCES comparable to that of the CNN. The main reason lies in the simplicity of GA, while the other methods are prone to overfitting on the very small validation sets during the hyperparameter optimization phase (fig. 4.7).

However, these results seem quite surprising for ScatNet since it has fewer trainable parameters than GA. In fact, the optimization method might play a role. Specifically, stochastic gradient descent in a lack of data regime could lead to worse results than finding the explicit solution to the optimization problem (if this explicit solution exists). The same remark applies to IINN. To verify this hypothesis, we also tried to estimate $\hat{\mu}(x)$ with a

direct linear regression from the scattering coefficients (minimizing Mean Square Error (MSE)) and setting the same constant $\hat{\sigma}$ as in GA. By doing so, this modified ScatNet achieves results that, when training on 64 years of data, are a few percents better than the Gaussian approximation (not shown here).

		Metric		
		CRPSS	NLLS	BCES
Model	GA	$\mathbf{0.250 \pm 0.004}$	$\mathbf{0.165 \pm 0.006}$	$\mathbf{0.262 \pm 0.006}$
	IINN	0.241 ± 0.008	0.14 ± 0.02	0.22 ± 0.02
	ScatNet	0.230 ± 0.004	0.05 ± 0.01	0.227 ± 0.002
	CNN	0.22 ± 0.02	0.09 ± 0.03	0.22 ± 0.03

Table 4.3: Test skills (the higher, the better) of the different models when trained on the smaller 80-year dataset, shown as mean and standard deviation over the 5 folds. In bold the best performing model according to each of the three metrics.

4.4.2 A remark on regression and classification

In this work, the different neural network architectures were trained to minimize the CRPS loss, but very similar results can be obtained minimizing the NLL loss. Similarly, the Gaussian approximation minimizes a regression metric (see eq. (4.6)). On the other hand, training on BCE yields worse performance (even for the BCES metric itself). Indeed, a classification task separates the training data in just two classes (heatwave and non-heatwave), neglecting the information about the heatwave amplitude. For the particular problem of heatwaves over France, the choice of the threshold a_5 to distinguish the two classes is arbitrary, and there is no obvious regime shift between mild and very extreme heatwaves [Mascolo et al., 2024a].

4.5 Interpretability

Now that we know how the different networks perform, we will go up the hierarchy trying to get an understanding of why the models provided their predictions.

4.5.1 GA and IINN

Both GA and IINN perform a linear projection followed by a scalar function, so they are fully interpretable. In fig. 4.1, we show the projection patterns and the prediction in the projected space.

The projection patterns highlight low values of soil moisture and an anticyclone over Central Europe, with a wave train spanning the North Atlantic. These patterns are very robust across the folds and look very similar between GA and IINN, with the major difference being a greater importance of soil moisture in the former. This result is consistent with previous studies on heatwaves over France [Miloshevich et al., 2023a; Mascolo et al.,

2024a] and with the main physical understanding of heatwave mechanisms at mid-latitudes [Perkins, 2015; Barriopedro et al., 2023; Miloshevich et al., 2023c].

The projected space clearly shows that the Gaussian approximation is performing a linear fit of the data with constant predicted variance. On the other hand, for IINN we observe that $\hat{\mu}(F = M_{\text{IINN}} \cdot X)$ has two different slopes depending on whether $F > 0$, showing also a larger variance for hotter events.

However, the values of the two slopes are not very robust across the 5 folds, sometimes being very similar, and as well the trend in the predicted variance is not stable. Considering that the projection pattern looks very similar to the one of the Gaussian approximation, and together with the fact that the IINN doesn't have a better performance, we can conclude that we don't gain anything with the added complexity of the IINN.

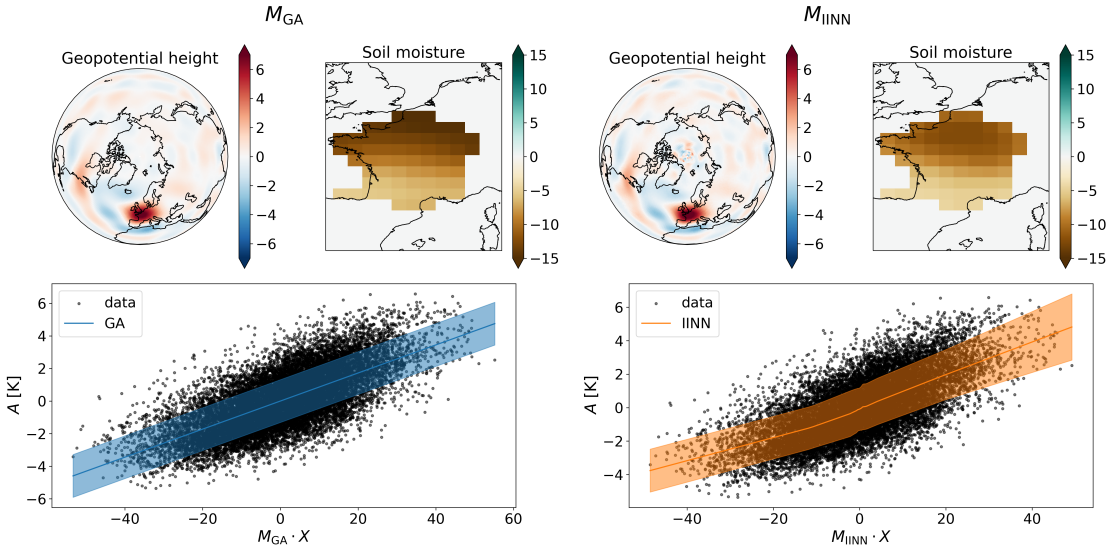


Figure 4.1: Projection patterns (top) and projected space (bottom) for GA (left) and IINN (right). In the bottom plots, the black dots are the test data, the continuous line is the predicted $\hat{\mu}(X)$ and the shading corresponds to $\pm\hat{\sigma}(X)$. We show among the 5 models the one with the highest skill.

4.5.2 CNN

The previous methods were simple enough that it is possible to relate the trained weights to the final prediction in a rather intuitive way, which allowed us to explain the prediction at once for any input X (global interpretability). This is totally not the case for the CNN, which behaves like a black box. To explain its prediction we need to use post-hoc explainable AI techniques, and most will give an explanation on a point by point basis (local interpretability).

Local interpretability: Expected Gradients

Local interpretability focuses on individual predictions, offering explanations that enhance understanding of feature contributions for each input (i.e., the initial conditions leading to heatwaves). Global interpretation techniques often overlook these insights.

In some cases, then, local interpretability helps to identify the model's strengths, weaknesses, errors, and biases, providing feedback for improvement. Methods for achieving local interpretability include feature attribution and counterfactuals. Feature attribution, which includes techniques like saliency maps, assigns values to measure the importance of each input feature. Counterfactuals explain a prediction by examining which features would need to be changed to achieve a desired prediction.

Several studies have already applied local XAI techniques to explain CNN predictions in geoscience applications [e.g. Barnes et al., 2022; Toms et al., 2021], and, in this work, we focus on feature attribution methods, as numerous approaches are available and specifically suited for deep neural networks. Gradient-based methods such as Deconvolution [Zeiler and Fergus, 2014], Guided Backpropagation [Springenberg et al., 2014], and Grad-CAM [Selvaraju et al., 2017] have been developed to perform feature attribution for CNNs. To address the limitations of gradient-based approaches, axiomatic methods such as Layer-wise Relevance Propagation [Bach et al., 2015b], Taylor Decomposition [Montavon et al., 2017], and Deep LIFT [Li et al., 2021] have been introduced. A state-of-the-art axiomatic approach, *Integrated Gradients* [Sundararajan et al., 2017], adds two new key axioms: sensitivity and implementation invariance. We choose the latter approach for its desirable properties and robustness. For a review of various methods and their evaluation in the context of climate science, refer to Bommer et al. [2023].

In this context, our explanatory analysis specifically focuses on the 500 hPa geopotential height anomaly input for predicting $\hat{\mu}(X)$. Results for soil moisture and the prediction of $\hat{\sigma}(X)$ are shown in figs. 4.10 and 4.11.

If we consider one input x, which in our case is the concatenation of the Z500 and SM fields, Integrated Gradients computes pixel attribution at pixel i using the following formula:

$$\text{IntegratedGrads}_i(x) := \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F_\theta\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha, \tag{4.7}$$

where $F_\theta(x)$ represents one of the outputs of the neural network for which we seek insight, namely $\hat{\mu}(x)$ or $\hat{\sigma}(x)$, or directly the heatwave probability given the initial condition x. In our case, $F_\theta(x) = \hat{\mu}(x)$. x' is a baseline input, namely the mean input on the training set, which in our case it's zero since our data is normalized.

Importantly, the sum of $\text{IntegratedGrads}_i(x)$ over all pixels i should yield the predicted value $F_\theta(x)$, according to the completeness axiom (see Sundararajan et al. [2017] for further details). Thus, $\text{IntegratedGrads}_i(x)$, which can be either positive or negative, quantifies the contribution of pixel i to the predicted output $F_\theta(x)$. We used an extension of this method called *Expected Gradients* [Erion et al., 2021], which reformulates the integral in eq. (4.7) as an expectation and combines it with sampling reference values from a background dataset

(taken as the entire training set here). We used the implementation provided in the SHAP Python package [Lundberg and Lee, 2017].

To illustrate the feature importance derived from the CNN predictions on relevant cases, we randomly selected examples of heatwaves (with $A_i > a_5 = 3.11\,\mathrm{K}$) from the test dataset. The resulting Expected Gradient Feature Importance (EGFI) maps for the CNN, along with the associated Z500 initial conditions, are shown, respectively, in the second and first rows of fig. 4.2. In practice, EGFI maps are smoothed by a Gaussian filter for visualization purposes (see fig. 4.9).

As we are interested in understanding what the CNN has learned beyond the GA, we compared the EGFI maps of the CNN with the ones computed for the GA. As the GA model predicts $\hat{\mu}(X)$ linearly, its EGFI maps are exactly equal to the input x multiplied by the GA projection pattern M_{GA}. We show these maps in the third row of fig. 4.2. Furthermore, in the fourth row, we show the difference between the EGFI maps of the GA and those of the CNN.

Several observations applicable to both the CNN and GA models can be drawn from the feature importance maps. First, as anticipated, there is a pronounced positive feature importance over France in the presence of positive geopotential anomaly (anticyclones). Two additional relevant observations include the positive contribution to the predicted temperature attributed to the positive geopotential anomaly within the storm-track region proximal to Northeastern America (columns 1 and 2) and the negative geopotential anomaly observed in the Labrador Sea (columns 1, 3, and 4). Both phenomena are likely correlated with the position of the Jet Stream.

The comparison between the CNN and the GA models reveals that the patterns of feature importance are strikingly similar, with the regions of importance largely overlapping. However, the CNN model displays a more localized and intense response in these specific areas. This indicates that the CNN model might be capturing more complex interactions or localized phenomena that the GA model, due to its linear nature, cannot represent.

Extracting more detailed information from these plots is challenging because this local, input-dependent approach only allows for a qualitative analysis. A global feature importance map could be generated by averaging the absolute values of the local maps over the whole dataset, but this may average out important predictability sources that are valid only for specific inputs. Another method based on *optimal input* is detailed in the following section. However, both approaches are limited, as they only highlight the areas of the input deemed important by the CNN model. They do not elucidate how this information is utilized to make predictions [Rudin, 2019]. A particularly relevant question is which scales are significant and how different scales interact. An ad-hoc attempt to address this issue for any black-box model (including CNNs) is discussed in Kasmi et al. [2023]. In Section 4.5.3 of this work, we partially address this question with the ScatNet model.

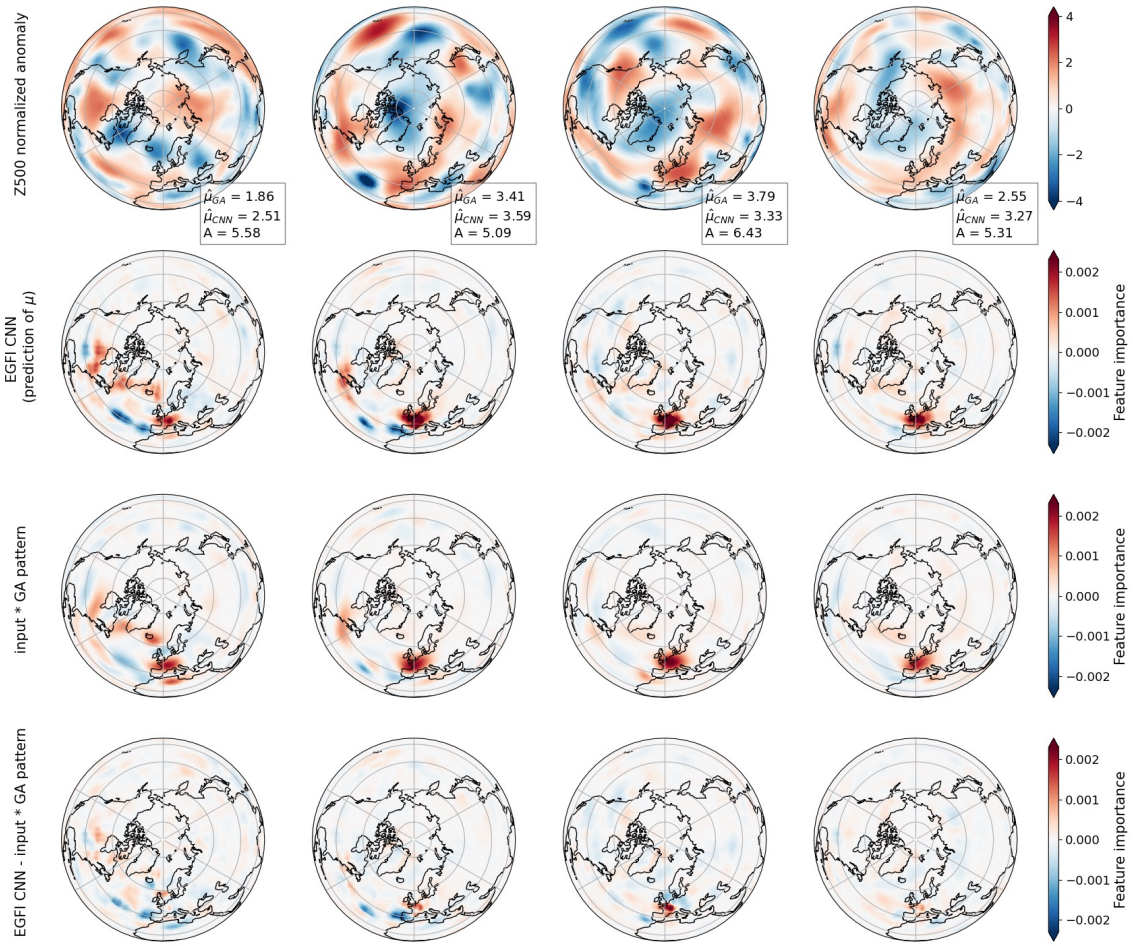


Figure 4.2: Top row: Several normalized Z500 (no units) initial conditions X associated with A above the 95^{th} percentile (heatwaves). Second row: Expected Gradient Feature Importance (EGFI) of the CNN predictions on these inputs. Third row: point-wise multiplication between inputs and GA projection pattern. Since $\hat{\mu}_{GA}(X)$ is linear in X, this amounts to compute EGFI for the GA prediction. Fourth row: EGFI CNN minus EGFI GA for each input. We show among the 5 models the one with the highest skill.

Global interpretability: Optimal Input

Another possibility to investigate the prediction of the CNN is trying to compute the input S that yields the highest predicted heatwave amplitude. We do so using Backward Optimization [Olah et al., 2017], initializing our input S_0 with a given data point and then performing gradient descent of the loss $\ell = -\hat{\mu}_{\mathrm{CNN}}(S)$, not on the weights of the CNN but on the coordinates of S. However, doing so yields very noisy maps and absurdly high predicted heatwave amplitudes (see fig. 4.12) We, instead, want our optimized input S to be something physically realistic, and to do so we add regularization terms to constrain the L2 norm and the roughness of S:

$$\ell = -\hat{\mu}_{\mathrm{CNN}}(S) + \lambda_2 \left(|S| - n_0\right)^2 + \lambda_r \left(\sqrt{H_2(S)} - r_0\right)^2, \tag{4.8}$$

where n_0 and r_0 are the average values of L2 norm and roughness of the data (see fig. 4.13).

If we run the optimization for all the points in the test dataset and then take the mean and standard deviation, we get the results in fig. 4.3, and an average $\hat{\mu}_{\mathrm{CNN}}(S) = 14.8 \pm 0.4\mathrm{K}$ (see fig. 4.14 for the full histogram), which is extremely high. The average optimal input looks again very similar to the projection pattern of the Gaussian approximation, while the variance is mostly located on the northern coast of Canada. Given the relative magnitudes of mean and standard deviation, we can conclude that the optimal input is very robust with respect to the choice of the initial seed S_0.

To explain why the optimal input resembles the projection pattern of the Gaussian approximation, we can write the prediction of the CNN as a perturbation on top of the prediction of the GA:

$$\hat{\mu}_{\mathrm{CNN}}(X) = \hat{\mu}_{\mathrm{GA}}(X) + (\hat{\mu}_{\mathrm{CNN}}(X) - \hat{\mu}_{\mathrm{GA}}(X)) = M_{\mathrm{GA}} \cdot X + \hat{\mu}_{\mathrm{pert}}(X). \tag{4.9}$$

Since the GA gives already very good predictions, $\hat{\mu}_{\mathrm{pert}}(X)$ will be a small non-linear perturbation. Now, when we follow the gradient to optimize S,

$$\frac{\partial}{\partial S}\hat{\mu}_{\mathrm{CNN}}(S) = M_{\mathrm{GA}} + \frac{\partial}{\partial S}\hat{\mu}_{\mathrm{pert}}(S). \tag{4.10}$$

Importantly, the first term doesn't depend on S, while the second one will act differently during the optimization as S evolves and as different initialization seeds S_0 are used. It is thus reasonable that once we take the average over all the test set, this term averages to 0. However, it will be responsible for the pattern observed in the variance of the optimal input.

So, with this method, we were able to shed some light on what in the CNN is beyond the Gaussian approximation, albeit encoded in the variance. To put our focus more onto $\hat{\mu}_{\mathrm{pert}}(X)$, we can add to the loss ℓ another regularization term $+\lambda_{\mathrm{orth}} \left(\hat{\mu}_{\mathrm{GA}}(S) - \hat{\mu}_{\mathrm{GA}}(S_0)\right)^2$, which forces the optimization to move in an orthogonal direction with respect to M_{GA}.

Doing so yields the results in fig. 4.4, with an average $\hat{\mu}_{\mathrm{pert}}(S) = 10 \pm 1\mathrm{K}$. The pattern of the variance didn't change much with respect to fig. 4.3, confirming its dependence on the perturbation term. On the other hand, the mean pattern is radically different,

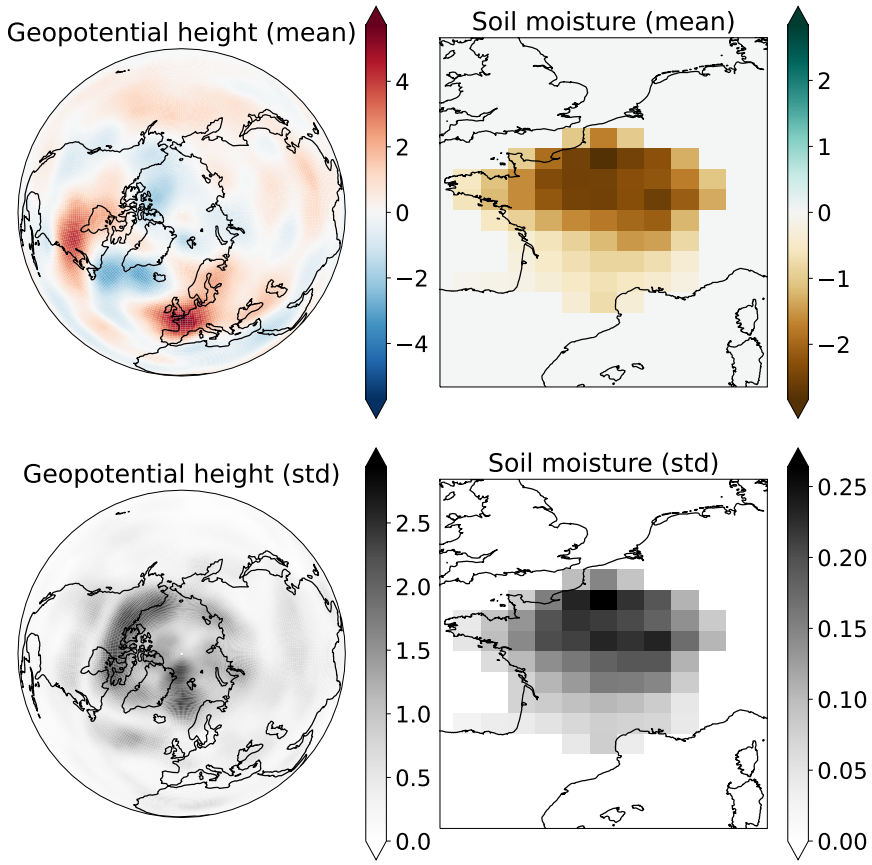


Figure 4.3: Average (top) and standard deviation (bottom) of the optimal inputs S that maximize the heatwave amplitude $\hat{\mu}_{\text{CNN}}(S)$ predicted by the CNN, across the different seeds S_0 taken from the test dataset.

though not very significant. Nevertheless, it highlights among non-linear enhancers of the heatwave condition a North-East to South-West gradient in soil moisture. A possible explanation is that Northeastern France is generally wetter, so when even that region gets dry the heatwave will be more intense. Indeed, the simple optimal input in fig. 4.3 and, to some extent, the projection pattern of the Gaussian approximation (fig. 4.1) already partially focus on Northeastern France. The patterns in the geopotential height field are more chaotic, with the most prominent feature being an anticyclone over Scotland. At this stage, however, we don't provide a physical interpretation.

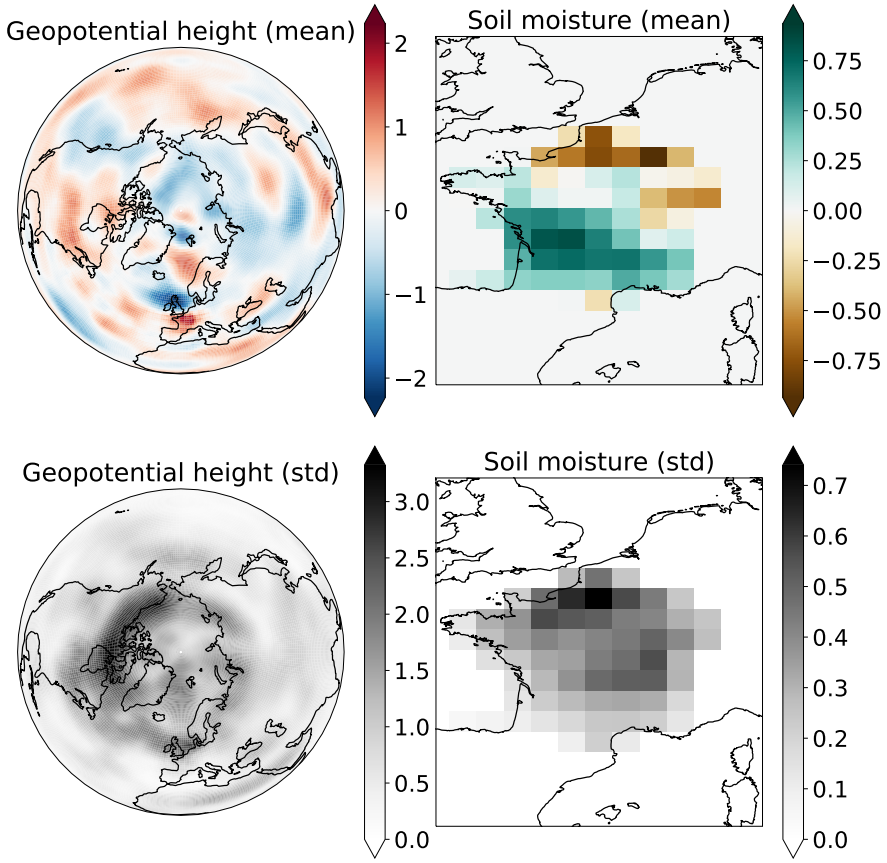


Figure 4.4: Average (top) and standard deviation (bottom) of the optimal inputs S that maximize the heatwave amplitude $\hat{\mu}_{\mathrm{CNN}}(S)$ predicted by the CNN while keeping fixed the prediction $\hat{\mu}_{\mathrm{GA}}(S)$ of the Gaussian approximation, across the different seeds S_0 taken from the test dataset.

4.5.3 ScatNet

Our next focus is on interpreting the ScatNet model, our intermediary complexity model that may provide insights into what could be learned beyond linearity.

At the end of the scattering transform, we obtain a feature map with the shape $\left(\frac{N_{lat}}{2^J}, \frac{N_{lon}}{2^J}, 1 + J \times L\right)$, where N_{lat} and N_{lon} are the number of latitude and longitude

points in the input image, J is the number of scales, and L is the number of orientations. The first feature map corresponds to the zeroth-order scattering transform, which is just the low-pass-filtered input. The subsequent $J \times L$ feature maps correspond to the first-order scattering transform. These maps are generated by convolving the input field with wavelet filters at each scale $j \in \{0, 1, 2\}$ and orientation $\theta \in \{0, \ldots, 7\}$ (since $J = 3$ and $L = 8$), followed by applying the modulus operator and pooling.

Similarly to what was done with the GA and IINN models, we can project the learned weights onto spatial maps for each channel of the feature maps. For the weights to be proper (global) *feature importance* maps, we normalize them to account for the different ranges of values of each feature map after the scattering transform. The formula for the feature importance FI_i for each feature i in the feature map is given by:

$$FI_i = \mathbb{E}(|X_i - \mathbb{E}(X_i)|) \cdot \beta_i \tag{4.11}$$

where β_i represents the weight associated with feature i in the final linear layer of the network. The expectation is taken as the mean over the test dataset. This definition corresponds to the mean absolute value of EGFI for the ScatNet model (see section 4.5.2) for feature i, but retaining the information of the sign of the weight.

Here we focus on interpreting the scattering features of the 500 hPa geopotential height anomaly (Z500) for the prediction of $\hat{\mu}(X)$. The first thing we can examine is the feature importance maps corresponding to the zeroth order features (or *coarse field*). These are obtained by simply applying a Gaussian low-pass filter to the input Z500 field, allowing us to directly compare the resulting projection pattern with the one computed by the GA model, as shown in fig. 4.5.

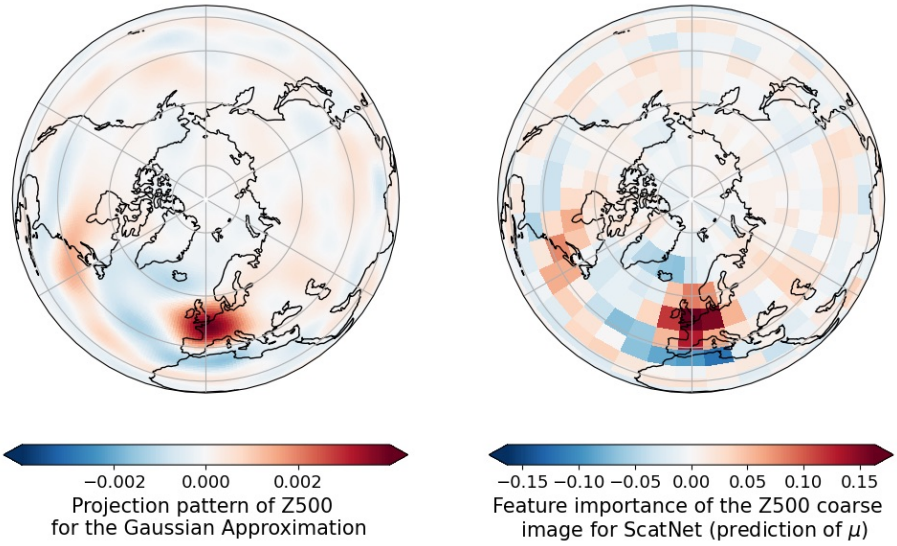


Figure 4.5: Projection patterns of the GA (left) and the feature importance of coarse Z500 field for the prediction of $\hat{\mu}(X)$ with ScatNet (right). We show among the 5 models the one with the highest skill.

The two patterns are nearly identical. This is not surprising, but rather reassuring. It

can be interpreted as the fact that the coarse graining by a factor 2^J acts as a regularization equivalent to the penalization of the spatial gradient that we used to smooth the GA projection pattern. The question is then to know whether we can achieve a predictive performance similar to the GA model with a lower resolution Z500 field. It amounts to know which is the effective scale for a linear prediction.

To answer this question, we define $\text{ScatNet}_{\text{coarse}}$, the simplistic variant of ScatNet where only the zeroth order features (the coarse Z500 field) is computed, rather than proceeding to the first order. This neural network then predicts $\hat{\mu}(X)$ and $\hat{\sigma}(X)$ linearly from the coarse Z500 field and from all soil moisture pixels across France. In table 4.4, we compare its predictive performance with both our simplest model and the more complex one, namely the GA and CNN models respectively.

		Metric		
		CRPSS	NLLS	BCES
Model	GA	0.2864 ± 0.0009	0.2169 ± 0.0009	0.293 ± 0.001
	$\text{ScatNet}_{\text{coarse}}$	0.2862 ± 0.0005	0.203 ± 0.001	0.291 ± 0.001
	CNN	0.310 ± 0.003	0.245 ± 0.007	0.311 ± 0.008

Table 4.4: Test skills (the higher, the better) of the different models. We compare $\text{ScatNet}_{\text{coarse}}$ to our most simple approach (GA) and our more complex one (CNN).

We show that using $\text{ScatNet}_{\text{coarse}}$ we achieve a performance barely lower than the GA model that uses the Z500 field at a higher resolution. This suggests that the performance gains from the ScatNet and CNN models are due to processing finer-scale information. To further validate this statement, table 4.5 compares the relative global feature importance for different scales of the Z500 coarse field, and the pixels of soil moisture. These global feature importance values are derived by calculating the feature-wise mean of the absolute feature importance values across the entire test dataset and then, independently for each scale, summing over geographical locations and different orientations.

	$j = 0$	$j = 1$	$j = 2$	coarse field	soil moisture
Relative FI [%]	5.2 ± 0.3	10.4 ± 0.5	20.0 ± 1.0	51.0 ± 2.4	13.5 ± 1.0

Table 4.5: Relative feature importance of various scales, expressed as percentages. The first three columns represent the relative mean absolute feature importance on first-order feature maps, summed across all orientations for scales $j = 0, 1$ and 2. The fourth column shows the mean feature importance on the zeroth-order Z500 feature map. The final column presents the mean feature importance summed across all pixels of soil moisture in France. We show the average values across the 5 models, using the dispersion between them to provide confidence intervals.

We observe that over 60% of the feature importance is attributed to the coarse Z500 field and soil moisture, which is the same information accessible to the GA model. The coarse Z500 field cannot distinguish structures occurring below 2^3 pixels (approximately 900 km at the equator's latitude), which corresponds roughly to the synoptic scale. It is

then not surprising that most of the predictive power arises from this scale. We also note that the soil moisture projection pattern is quite noisy (not shown) due to the absence of regularization, yet it still exhibits relatively high feature importance.

The remaining 35% of feature importance is derived from the first-order scattering features, with larger scales contributing more. During the scattering transform process, when we convolve our Z500 field with a wavelet filter located at scale j and orientation θ, we are essentially applying a band-pass filter concentrated around the wavelength of 2^j pixels. However, this band-pass filter is non-local in Fourier space, meaning that it affects a range of wavelengths and orientations rather than a single one, and thus the different scales and orientations are not completely independent. Nevertheless, table 4.5 indicates that additional information at the sub-synoptic scale (450 km and below) has an important role for prediction.

Finally, we present in fig. 4.6 the feature importance patterns derived from all the first-order scattering features at scale $j = 2$. For scale $j = 0$ and $j = 1$, the feature importance maps are very similar, but with less feature importance (see fig. 4.16).

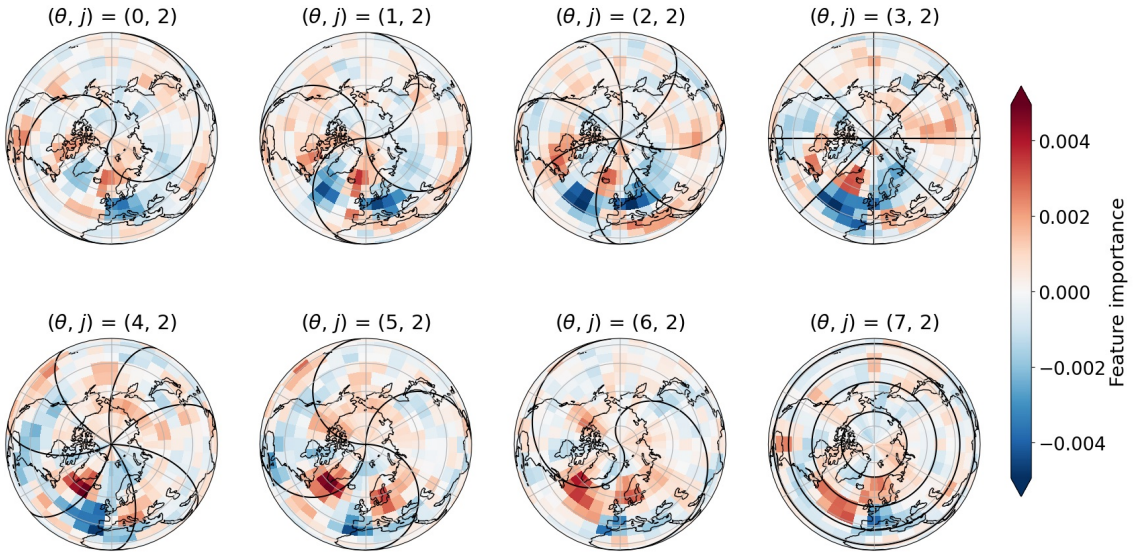


Figure 4.6: Feature importance of first order features for the prediction of $\hat{\mu}(X)$ with ScatNet at scale $j = 2$ for each orientation. Maps at finer scale are very similar, but with less feature importance (see Figure S10 of the Supplementary Material). The black filaments represent the orientation of the wavelet in spatial space, with the approximate wave vector being orthogonal to the filaments. We show the average maps across the 5 models.

To interpret these maps, let us consider for example the map corresponding to $\theta = 3$ (top right). The blueish structure in the Atlantic Ocean indicates that zonal oscillations at a scale of 2^2 pixels in this location contribute to a decrease in the predicted value of $\hat{\mu}(x)$, and consequently, to a lower probability of a heatwave. Another relevant case is that of the map at $\theta = 7$ (bottom right), which corresponds to meridional oscillations at a scale of 2^2 pixels. The reddish structure, located, as in the previous example, in the middle of the Atlantic Ocean, indicates that this time the oscillations increase the likelihood of a

heatwave.

It should be noted that areas exhibiting the same hue in this figure do not represent structures of the size of these blobs. Rather, they denote regions in which structures approximately 400 km in scale produce equivalent effects, irrespective of their exact location within the region, with darker shades signifying a more pronounced impact.

Despite these qualitative remarks, the maps of fig. 4.6 remain challenging to interpret. Nonetheless, some general observations can be made. Firstly, the patterns appear to evolve continuously from one orientation to another. This can be partially explained by the overlap of the Fourier supports of wavelet filters from one orientation to the next [Bruna and Mallat, 2013]. Additionally, it is noteworthy that most of the feature importance is concentrated around France and the Atlantic Ocean, which aligns with the expected behavior of the global circulation with westward moving weather patterns.

Finally, across all maps, it is noteworthy that the feature importance surrounding France predominantly exhibits negative values. This indicates that oscillating structures over France are unfavorable for the occurrence of heatwaves, and it aligns with the qualitative meteorological understanding that associates persistent anticyclones with the occurrence of heatwaves [Perkins, 2015; Barriopedro et al., 2023].

4.6 Discussion

By testing increasingly complex architectures, we were able to quantify to which extent higher complexity leads to a better performance. As pointed out in Rudin [2019], the benefits of complexity are often overrated, and indeed, when data is scarce, the best performance was achieved for the simple linear regression of the Gaussian approximation.

This simple model remained competitive with the more complex ones even with much larger amounts of data. Moreover, when the scattering network uses only the coarse-grained input, it becomes itself a simple linear regression, and it performed very similarly to the Gaussian approximation while using 64 times less scalar features. This highlights the potential for even further simplification of the regression task without loss of performance from the GA baseline.

For the task of forecasting heatwaves, the slightly increased complexity of the Intrinsically Interpretable Neural Network didn't give any benefits, neither from the point of view of performance nor of additional insight into the physical processes. However, the design of the architecture is very interesting, reimagining a simple fully connected neural network with a bottleneck in terms of optimal linear projection of the data. It is then easy to relax the bottleneck from a single optimal index to a set of m optimal indices. This could be particularly useful when studying problems where the effective dynamics can be well expressed non-linearly in a low-dimensional space. A possible example could be the study of the Madden-Julian Oscillation [Delaunay and Christensen, 2022]. Compared to standard principal component analysis, the IINN would give linear projections that are tailored to the regression task rather than the ones that best explain the variance of the

input data.

We did try to train IINN architectures with more than one projection pattern on the heatwave task, but the results were not better than the ones presented in this work. Together with the high skill of the Gaussian approximation presented here and in Mascolo et al. [2024a], this hints at the fact that heatwaves are a highly linear process.

Going beyond the skill of the linear model thus requires several centuries of data to properly train more flexible models. However, in the case of Convolutional Neural Networks, the interpretation of where the added performance comes from is hard to grasp. When using post-hoc explainability methods, we would like something that gives a global answer for all different inputs, but our analysis with the optimal input found something very similar the Gaussian approximation projection pattern, which thus gives no information on what causes the *extra* skill of the CNN. Indeed, post-hoc explanations are necessarily incomplete, as they need to drastically simplify a complex non-linear black box into something understandable by a human [Rudin, 2019]. By using a hierarchy of models, we were able to quantitatively expose how incomplete these explanations are.

One might argue that the incompleteness is obvious because we require a too simple answer when we ask for something that works for all inputs. However, together with the point-by-point analysis of Expected Gradients, we tried performing K-means clustering on the optimal inputs, similarly to what is done in Toms et al. [2021], but we didn't find any significant multi-modal behavior in the data. In the end, these methods confirm that there is something significant in the CNN beyond the linear model, but don't really enable the user to understand what.

The architecture of the ScatNet proposed in this work solves this issue, allowing for global interpretability like GA and IINN. Importantly, the ScatNet has the same skill of the CNN, so we can say that there is nothing significant in the predictions of the CNN that ScatNet misses.

Our analysis shows that the linear model contributes to 65% of the prediction of the ScatNet, with the remaining information being encoded in oscillations in the 500 hPa geopotential height anomaly field at scales around 400 km. Even more interestingly, we are able to visualize the individual contributions of oscillations in different directions and geographical locations. In particular, North-South oscillations in East Atlantic promote heatwaves while East-West ones in the same area inhibit them. The precise physical interpretation of these effects goes beyond the scope of this paper, but we feel this is a very promising direction for the discovery of new physics.

Also, in our work the scattering transform is applied to the geopotential height field by treating it as a two-dimensional image with pixels of constant size in latitude and longitude. A natural improvement is then to use scattering transform on a sphere [McEwen et al., 2021] to better account for Earth's geometry.

Finally, a general remark is that in this work we approximated, across all models, the conditional distribution of the heatwave amplitude $\mathbb{P}(A|X)$ as a Gaussian distribution, of which we estimate the mean and the standard deviation. This may be a factor

limiting performance in highly non-linear problems where the conditional distributions are potentially skewed. A possible fix could be to use a non-parametric approach such as quantile regression [Zhang et al., 2018] or other types of parametric approaches, for instance relying on Extreme Value Theory [Cisneros et al., 2024]. However, this has the effect of complicating network architectures and thus hindering interpretability, which was the main focus of this work.

4.7 Conclusions

Recent developments in machine learning have led to a proliferation of increasingly complex models being applied to weather and climate tasks. Most of these models behave as black boxes, and thus require additional explanation tools to elucidate their decision-making process. Although it is commonly accepted that these explanations are incomplete [Murdoch et al., 2019], often they are presented as comprehensive insight, which can be misleading [Rudin, 2019]. On the other hand, inherently interpretable models can be a much better tool for expanding our knowledge with the help of machine learning [Rudin, 2019; Barnes et al., 2022].

In this work we tested three increasingly complex interpretable models and a black-box one on the task of forecasting extreme heatwaves. Our findings can be summarized in three main conclusions: first, when data is scarce the simplest model (GA) has the highest skill. Second, even with enough data the black-box model (CNN) does not outperform the best interpretable model (ScatNet), and its explanations don't tell us much beyond what was already clear from the GA model. Third, analysis of the ScatNet model easily shows that the gain in performance from the GA model comes from oscillations in the geopotential height field, mainly over the North Atlantic and with a wavelength around 400 km.

With this work, we highlight once again the power of interpretable models, and we identify scattering networks (ScatNets) as a very promising tool to be applied in climate science. In particular, ScatNets could be integrated as an interpretable module [Yang et al., 2024] into more complex models, serving as a preprocessing step to give a sparse representation of a geophysical field. For instance, it could potentially act as the initial layer in an encoder-decoder architecture. Furthermore, the mean Scattering Spectrum could be used as a penalization term in loss functions for deep learning weather and climate emulators to promote physically consistent forecasts.

Acknowledgements

Author Lovo received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 956170 (Critical Earth). Author Lancelin is funded by RTE, the French national electricity operator.

Data Availability Statement

Code is available on GitHub (`https://github.com/AlessandroLovo/intepretability-hierarchy-zenodo`), together with all the data needed to reproduce the results in this paper. Due to its size, it was not possible to publish the 1000-year-long control run of CESM. However, the details and code of how this control run was obtained are explained in Ragone and Bouchet [2021] and available on Zenodo (`https://doi.org/10.5281/zenodo.4763283`).

4.8 Supporting Information

4.8.1 Pareto plots

For the interpretable models of the Gaussian approximation (GA) and Intrinsically Interpretable Neural Network (IINN), we have a visualizable projection pattern M. For M to be interpretable, we require it to be spatially smooth, by adding a regularization term to the loss function. As explained in Mascolo et al. [2024a], this can be achieved either penalizing directly the roughness H_2 of the projection pattern (solid lines in figs. 4.7 and 4.8) or indirectly penalizing its L_2 norm (dashed lines in figs. 4.7 and 4.8). In any case, we need to find the proper value of the regularization coefficient ϵ which gives a good compromise between skill and interpretability. Since there are two objectives to optimize at the same time, we call the plots in figs. 4.7 and 4.8 Pareto plots.

Focusing on the top row of fig. 4.7, we see that both IINN and GA, for both types of regularization, have a broad plateau of good skill, with performance drops on both sides. On the right side, ϵ is very small, and thus the pattern is very rough. For GA (blue curves), this severely impacts the validation performance, as the model is overfitted to the training data. On the other hand, IINN (orange curves) can somewhat compensate overfitting with the network following the linear projection, namely by predicting a higher variance. More interesting is the behavior on the left side. Again, for GA the pattern is now too smooth and so the model isn't flexible enough to capture the important information. IINN, however, has a weird behavior, where increasing too much the regularization coefficient leads to rougher patterns (the curves bend back towards the right). This is simply explained by the fact that while GA is trained in one step [Mascolo et al., 2024a], IINN is optimized with gradient descent starting from randomly initialized weights. Since we use an early stopping protocol on the validation skill to avoid overfitting, when ϵ is too high, the main evolution of the pattern is to make it smooth rather predictive. This means that early stopping kicks in too early and the model is checkpointed back to one of the early epochs where the projection pattern was still very rough.

By comparing the two types of regularization, we can see that L_2 regularization yields slightly higher validation skills. However, the Pareto optimum is to be as close as possible to the top left corner of the plot (high skill, low roughness), and thus we choose to use H_2 regularization.

The second row of fig. 4.7 shows the Pareto plots when the metrics are evaluated on the test set. Any skill gain that more complex (IINN with respect to GA) or less regularized architectures had in the validation set is completely lost on the test set. This suggests that these marginal gains were effects of overfitting on the validation set during the hyperparameter optimization phase.

In fig. 4.8, we display the same results for the networks trained on the shorter 64-year-long datasets. The conclusions about IINN and GA are very similar to the ones for the longer dataset, and the main difference is the greatly reduced performance of CNN and ScatNet, which now display the worst skills in the hierarchy.

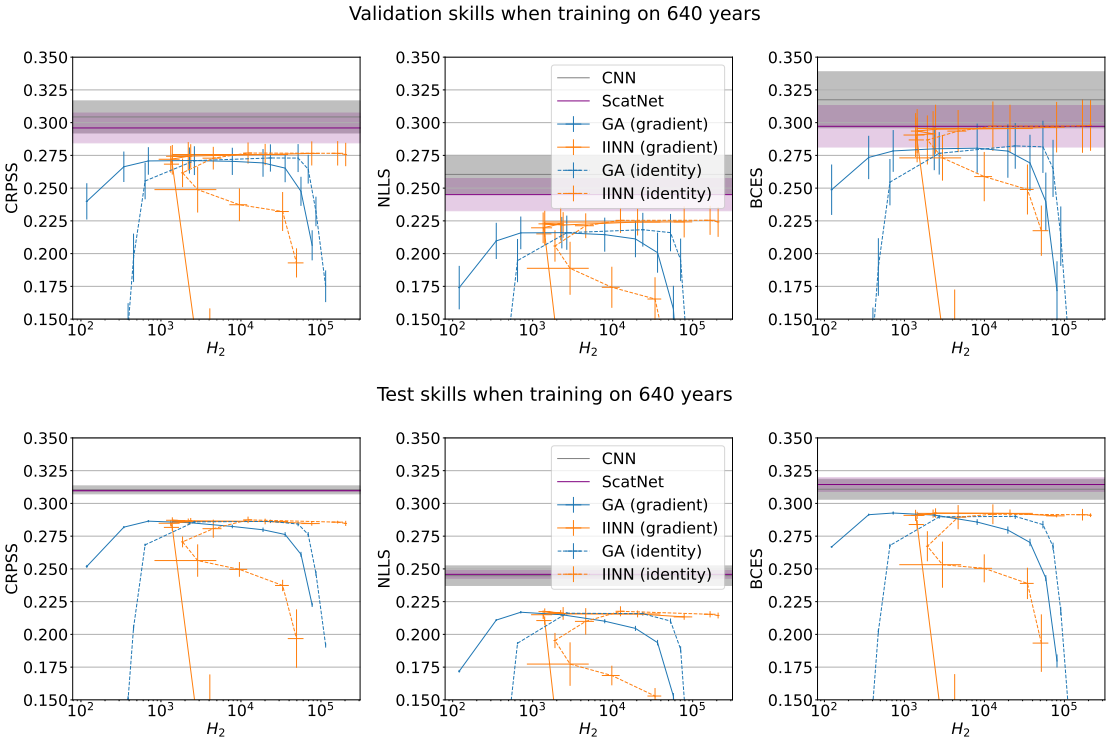


Figure 4.7: Pareto plots when training on the 800-year-long dataset. Skill on the y-axis (higher is better) for the three metrics and roughness of the projection pattern on the x-axis (lower is better). CNN and ScatNet don't have a projection pattern, so we report their skill as constant across values of H_2. For IINN and GA we show the results both with L2 regularization ('identity') and gradient regularization ('gradient'). The latter gives better results and is the one used in the main text of this paper. For validation, the error bars and shading correspond to the 5 different models being evaluated on the 5 different 160-year-long validation sets. For test, the error bars and shading are due only to the 5 different models being evaluated on the common 200-year-long test set.

4.8.2 CNN: Local Interpretability

In fig. 4.2, we show several examples of local Expected Gradient Feature Importance (EGFI) maps. For visual purposes, we chose to smooth these maps by a Gaussian filter with width $\sigma = 1.3$ pixels. We also normalize the maps by the sum of feature importance

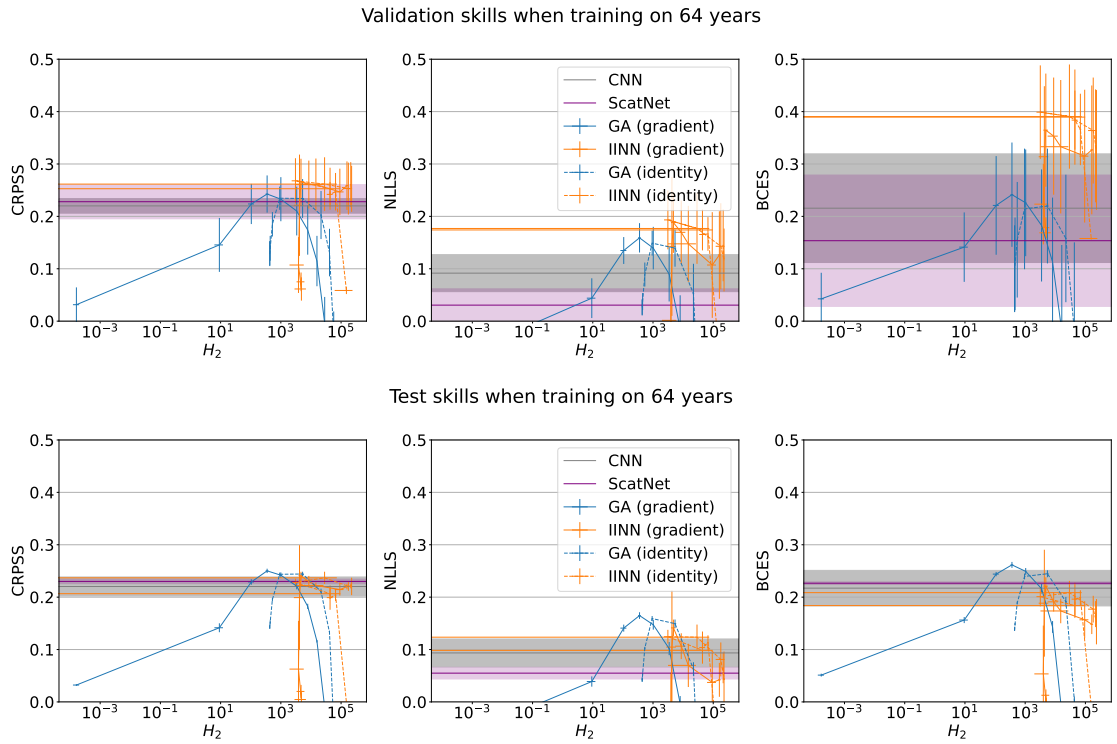


Figure 4.8: Same as fig. 4.7, but when training on the smaller 64-year datasets

(that is equal to $\hat{\mu}$ according to the completeness axiom [Sundararajan et al., 2017]) to have a better visual comparison between different predictions. In fig. 4.9, we show an example of a raw EGFI map (middle) and its Gaussian-smoothed version (right).

Furthermore, in fig. 4.2, we showed the EGFI maps for the prediction of $\hat{\mu}(X)$ with the CNN only for Z500. In fig. 4.10, we complement this figure by showing the EGFI maps for the soil moisture input. Also, as done in the main text, we show the point-wise multiplication between the input and the GA projection pattern, which amounts to computing the EGFI for the GA prediction. We can see that the GA projection pattern is very smooth and has a clear spatial structure. The EGFI maps for the CNN prediction show a quite similar structure, but with a large amount noise.

Finally, as our models predict both the mean and the standard deviation of the heatwave amplitude, we can also compute the EGFI for the standard deviation. Figure 4.11 presents the EGFI maps of Z500 for predicting $\hat{\sigma}(X)$ using the CNN. The EGFI maps exhibit significant noise. Notably, for specific inputs, such as the first two on the left in the top row, a structure far from France, located in the Northeastern United States, significantly influences the predicted standard deviation. This teleconnection indicates that strong positive anomalies in the geopotential height in the Northeastern United States are linked to higher uncertainty in predicting the heatwave amplitude in France. This connection seems reasonable, given the eastward propagation of weather patterns.

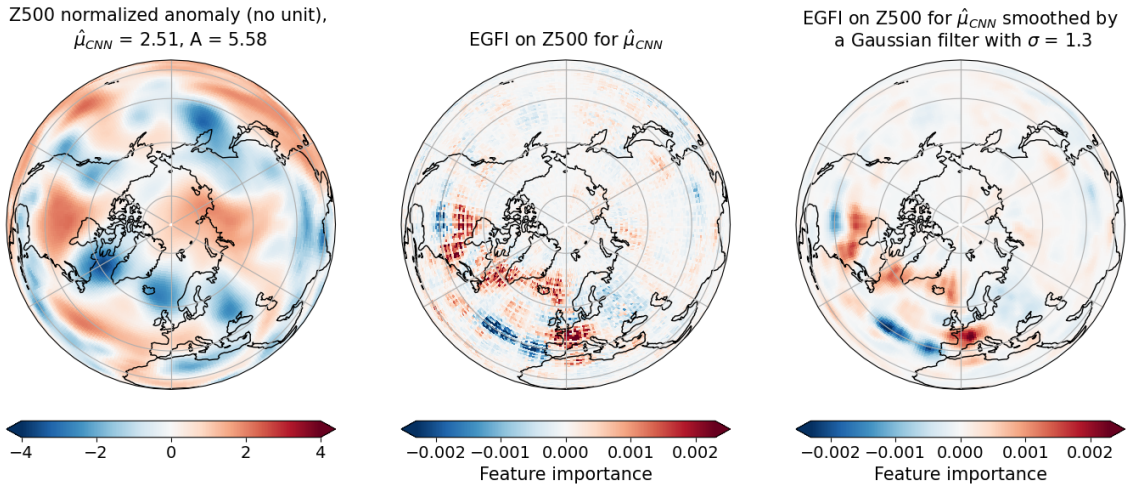


Figure 4.9: Raw Expected Gradient Feature Importance (EGFI) (middle) and Gaussian-smoothed EGFI (right) for the prediction of $\hat{\mu}(X)$ with the CNN. The smoothing helps in highlighting the broader areas of influence with less noise. For the Gaussian filter, we set $\sigma = 1.3$ on qualitative grounds. We also show the corresponding Z500 input (left) from the test set, selected from the heatwaves examples. We show among the 5 CNN models the one with the highest skill.

4.8.3 CNN: Optimal input

In fig. 4.12 we show the effect of direct backward optimization of $\hat{\mu}_{\text{CNN}}$ on a sample point from the test data. Such optimization follows the complex non-linearities of the CNN and thus leads to a very rough optimized input, and an extremely unrealistic predicted heatwave amplitude of 36 K. To counteract this effect and have more physically realistic maps, we add to the simple loss $\ell = -\hat{\mu}_{\text{CNN}}(S)$ terms constraining the roughness and L_2 norm of S. In particular, when we plot the histogram of these quantities for test data, we find that the distributions are relatively narrow (fig. 4.13). Then, we modify the loss ℓ as

$$\ell = -\hat{\mu}_{\text{CNN}}(S) + \lambda_2 \left(|S| - n_0\right)^2 + \lambda_r \left(\sqrt{H_2(S)} - r_0\right)^2, \tag{4.12}$$

where n_0 and r_0 are respectively the average L_2 norm and roughness of the test data, and λ_2 and λ_r are the two regularization parameters.

To complement the information presented in the main text about the average optimal input, we present in figs. 4.14 and 4.15 the distribution of the predicted heatwave amplitude. From fig. 4.14 we can see that, despite having imposed some physical constraint in the form of roughness and L_2 norm, the distribution of the predicted heatwave amplitude (orange) is still very far from the actual heatwave events in the test data (blue). Forcing the optimization to move orthogonally to the Gaussian approximation projection pattern (fig. 4.15) yields lower predicted values. From this latter plot we can see that the variance of $\hat{\mu}_{\text{CNN}}$ is much higher with respect to the previous plot, but this is due to the fact that moving orthogonally to the projection pattern leads $\hat{\mu}_{\text{GA}}$ to stay constant during the optimization, and thus retain the distribution it had for the prediction on test data.

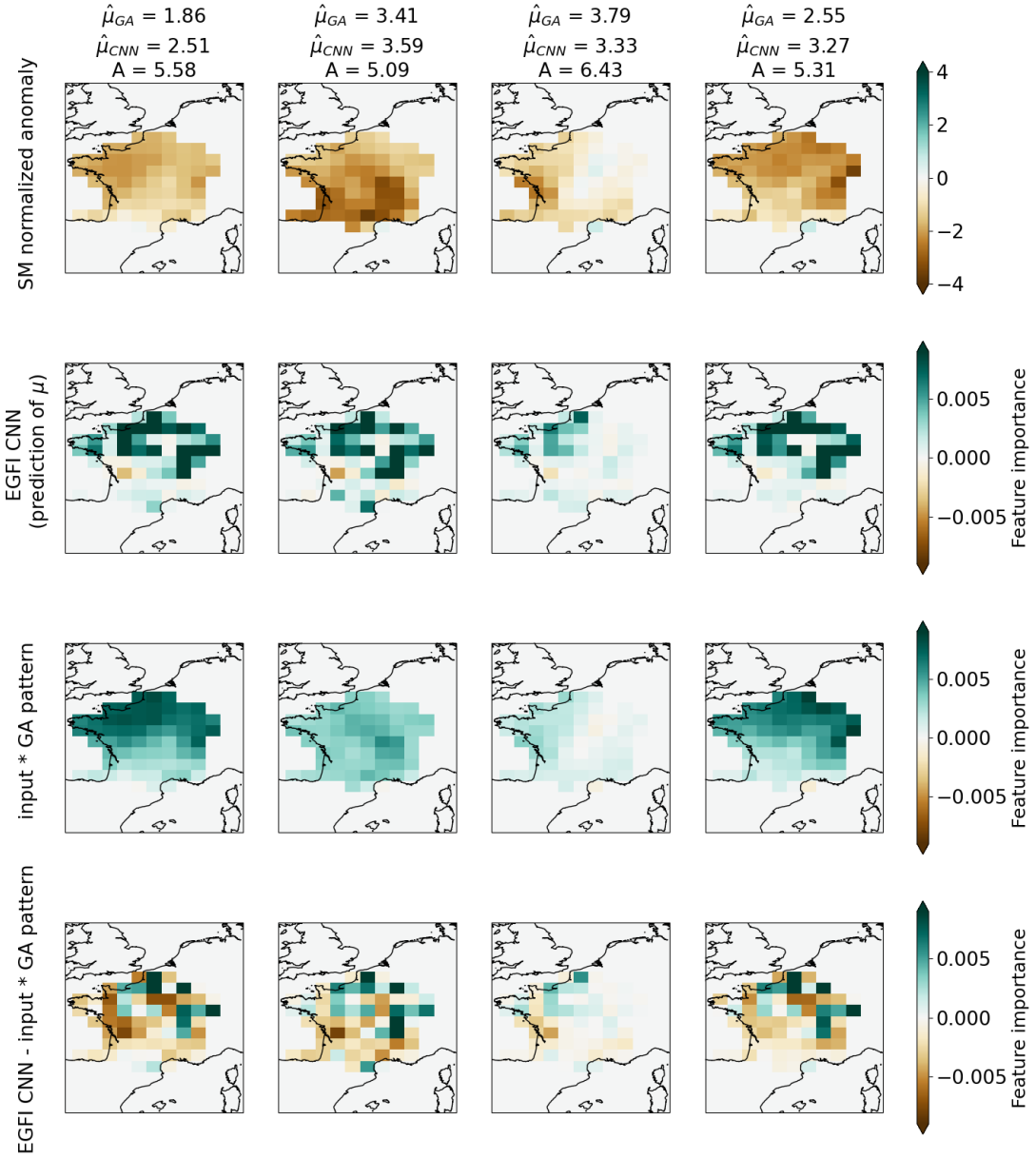


Figure 4.10: Top row: Several normalized soil moisture initial conditions X (inputs) associated with A above the 95^{th} percentile (heatwaves). Second row: Expected Gradient Feature Importance (EGFI) on soil moisture of the CNN predictions. Third row: point-wise multiplication between inputs and GA projection pattern. Since $\hat{\mu}_{GA}(X)$ is linear in X, this amounts to compute EGFI for the GA prediction. Fourth row: EGFI CNN minus EGFI GA for each input. We show among the 5 models the one with the highest skill. Here, the EGFI maps are not smoothed with a Gaussian filter.

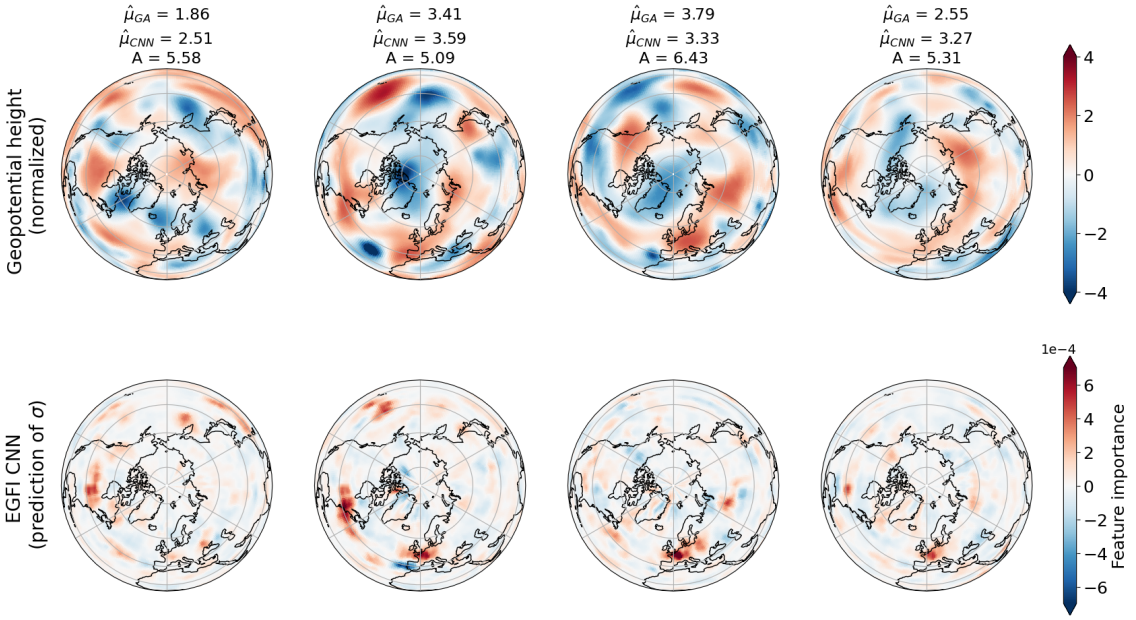


Figure 4.11: Top row: Several normalized Z500 initial conditions X (inputs) associated with A above the 5th percentile (heatwaves). Second row: Expected Gradient Feature Importance (EGFI) on Z500 of the CNN predictions of $\hat{\sigma}(X)$. We show among the 5 models the one with the highest skill.

Indeed, $\hat{\mu}_{\text{pert}} = \hat{\mu}_{\text{CNN}} - \hat{\mu}_{\text{GA}}$, which is what is freely optimized, retains a similar variance to that of $\hat{\mu}_{\text{CNN}}$ without any orthogonality constraint.

4.8.4 Scatnet

In fig. 4.16, we present the feature importance of first-order features for the prediction of $\hat{\mu}(X)$ using ScatNet across various scales and for orientation $\theta = 0$. The feature importance maps are proportional across scales, with a reduced amplitude for smaller scales j. This proportional behavior is consistent across other orientations, though not depicted here.

The question arises as to why these maps are proportional. Initially, it is not apparent that the model should learn identical patterns at different scales, as there are no intrinsic reasons for this. However, considering that wavelet filters are non-local in the frequency domain and their supports overlap, we expect these patterns to be close in some way, either in spatial pattern or amplitude (or both combined).

In fig. 4.6, we observed that as we transition between orientations, the spatial pattern of feature importance changes continuously. Here we observed, when transitioning from a larger scale j to a smaller one, that the spatial pattern remains the same but with diminished amplitude. Although we do not have a clear explanation for these behaviors, we interpret them as indications of robustness, as the model does not appear to be learning mere noise.

Additionally, we observed the robustness of feature importance maps across each independent model used in the 5-fold cross-validation (not shown here), further supporting

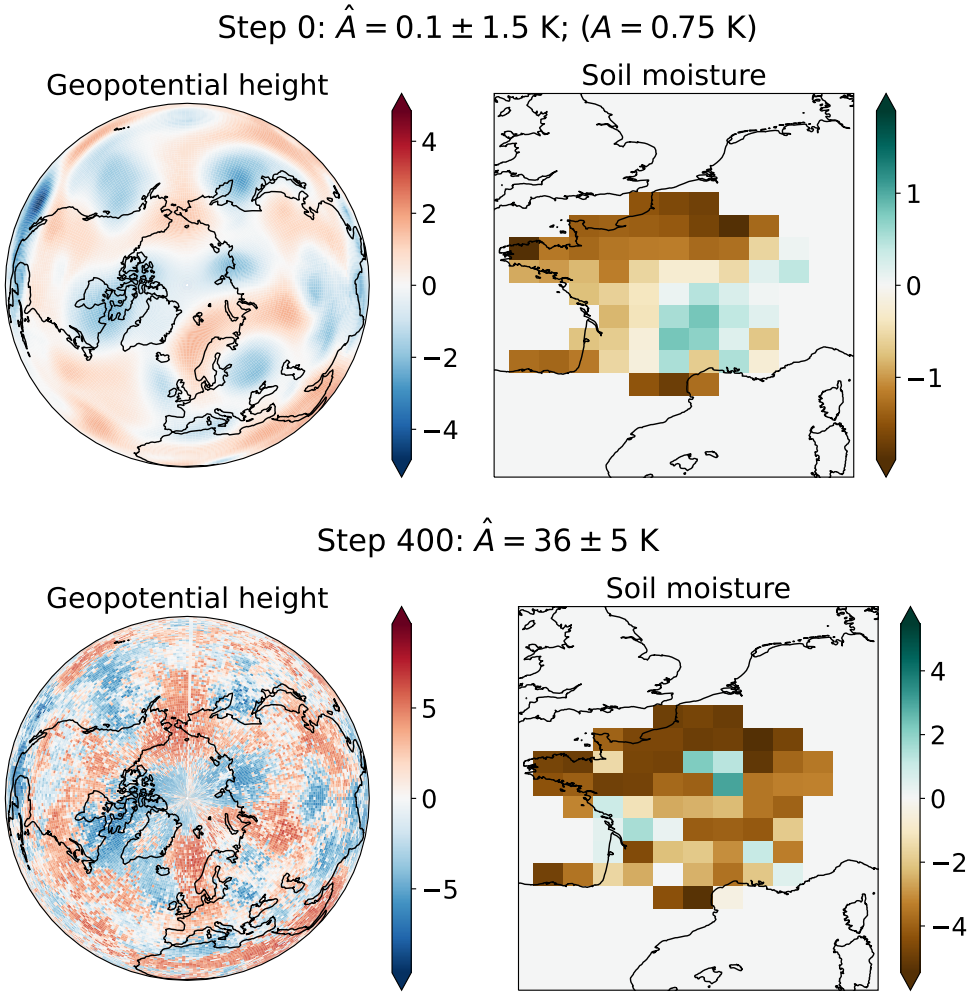


Figure 4.12: Example of the effect of the unconstrained backward optimization of $\hat{\mu}_{\text{CNN}}$ on a typical snapshot from the test set (top), after 400 iterations of gradient descent (bottom). For both figure we report the prediction of the CNN and for the first also the true value of the heatwave amplitude A.

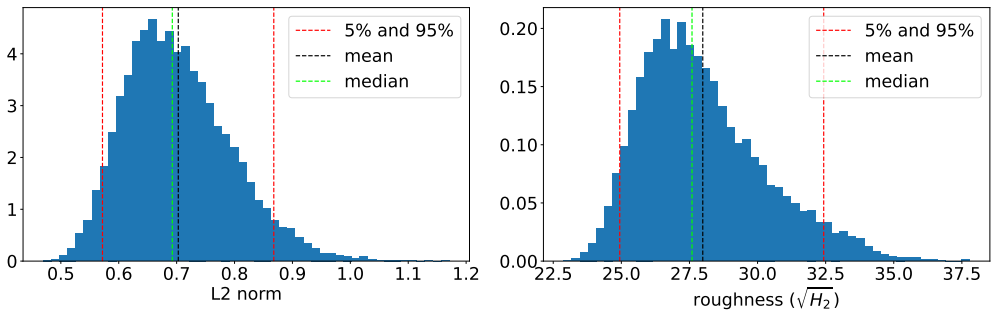


Figure 4.13: Histograms of the L_2 and H_2 norms of the test data. Vertical lines show the mean, median and 5% and 95% quantiles

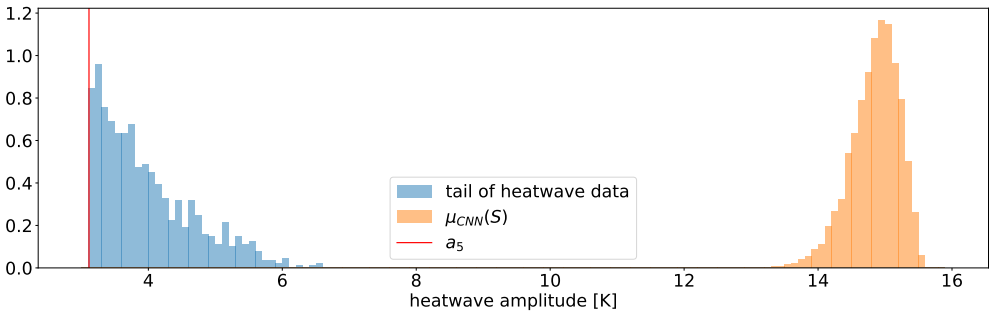


Figure 4.14: Histograms of the heatwave amplitude in the tail (5% most extreme events) of the distribution of test data (blue) and of the mean heatwave amplitude $\hat{\mu}_{\mathrm{CNN}}(S)$ predicted by the CNN on the optimized inputs S (orange).

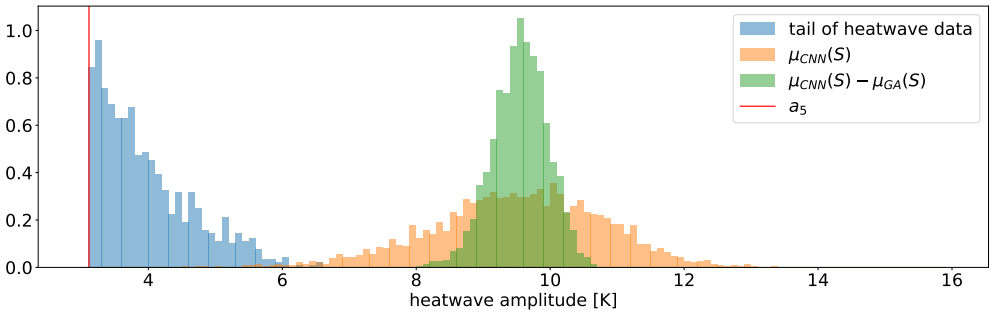


Figure 4.15: Histograms of the heatwave amplitude in the tail (5% most extreme events) of the distribution of test data (blue) and of the mean heatwave amplitude $\hat{\mu}_{\mathrm{CNN}}(S)$ predicted by the CNN on the optimized inputs S in an orthogonal direction to the projection pattern of the Gaussian approximation (orange). In green the difference between the prediction of CNN and GA, where the latter didn't change during the optimization process.

the model's reliability.

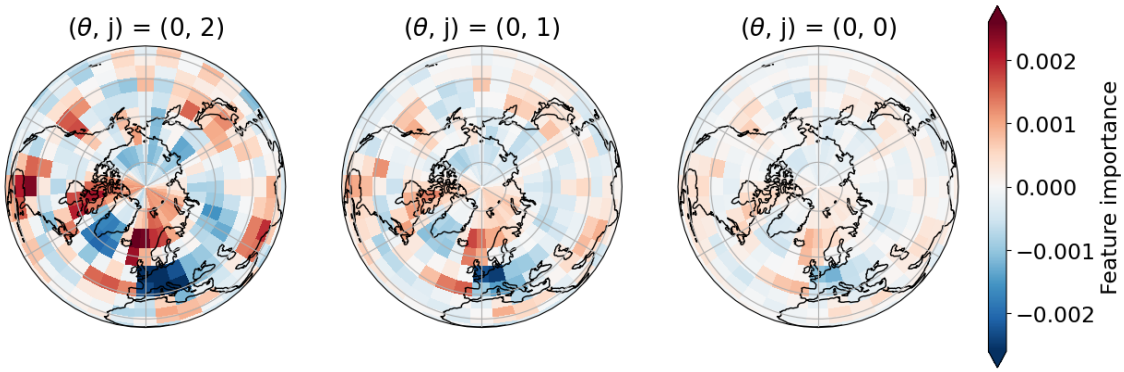
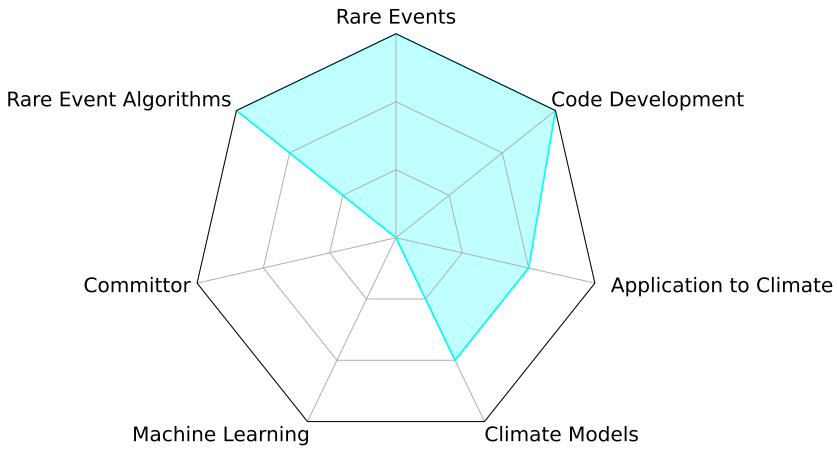
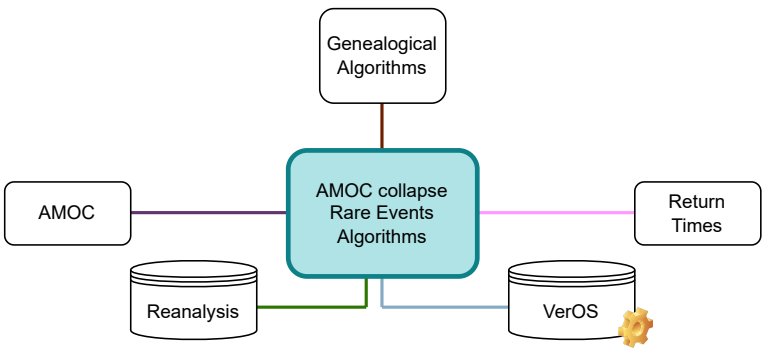


Figure 4.16: Feature importance of first order features for the prediction of $\hat{\mu}(X)$ on geopotential height with ScatNet for each scale and for orientation $\theta = 0$. For any given orientation θ, the feature importance maps at different scales j are proportional, but with less amplitude for smaller scales. See fig. 4.6 for maps at different orientations for $j = 2$. We show here the average maps across the 5 models.

Chapter 5

Simulating the noise induced collapse of the Atlantic meridional overturning circulation with a rare event algorithm





In the previous chapters 3 and 4, we saw that a way to mitigate the lack of data issue inherent to the study of rare events is that of using simpler statistical models. However, this approach has its limitations, and indeed we always focused on the 5% most extreme heatwaves. These types of events do satisfy the common consensus for calling an event extreme [Perkins, 2015; Seneviratne et al., 2012], but they have a return time of the order of a few years (see section 3.9.8), which means they are not particularly rare. The choice to study these not-so-rare extremes was forced, because otherwise we wouldn't have had enough data to properly train the networks.

As we pointed out in the introduction, if we are interested in very rare events, one of the most feasible options is that of using rare event algorithms to efficiently over-sample the tail of the distribution. In this chapter, we explore this research direction, outlining the theory of the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) rare event algorithm, seeing it work in practice on the simple Ornstein-Uhlenbeck process, and finally applying it to the Versatile Ocean Simulator (VerOS) model to investigate the noise induced collapse of the Atlantic meridional overturning circulation (AMOC).

The results presented in this chapter cover the work done during my two secondment periods at the University of Copenhagen (UCPH) (in collaboration with Johannes Lohmann[1] and Peter Ditlevsen[1]) and at Utrecht University (UU) (in collaboration with Alfred Bendtzon Hansen[1] and Henk Dijkstra[23]).

A virtually twin project was carried out independently and at the same time by Matteo Cini[45] [Cini et al., 2024], showing a successful noise induced tipping of the AMOC in the PlaSim-LSG model. In my work, the VerOS model didn't display any collapse. Nevertheless, this negative result is still valuable, and we will discuss the attempts made, as well as investigate the reasons for the higher resilience of the AMOC in our setup.

A condensed summary of the results presented in this chapter was included in a paper first-authored by Johannes Lohmann, recently submitted to Proceedings of the Royal Society A (PRSA). The preprint is available on arXiv at `https://arxiv.org/abs/2410.16277`.

5.1 Introduction

The Earth's oceans are laced with a complex network of ocean currents, which transport heat and salt across the planet [Rahmstorf, 2002]. The meridional gradient of insolation, and consequently temperature, causes a natural poleward heat transport. However, the Atlantic Ocean is an exception, where the meridional heat flux is oriented northward across all latitudes [Srokosz et al., 2021]. In a somewhat simplified manner, warm surface water

[1] *Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark*

[2] *Institute for Marine and Atmospheric research Utrecht, Department of Physics, Utrecht University, Utrecht, The Netherlands*

[3] *Centre for Complex Systems Studies, Department of Physics, Utrecht University, Utrecht, The Netherlands*

[4] *Department of Physics, Università degli studi di Torino, Turin, Italy*

[5] *National Research Council of Italy, Institute of Atmospheric Sciences and Climate (CNR-ISAC), Bologna, Italy*

from the Indian Ocean rounds the southern tip of Africa, enters the Gulf of Mexico, and then spreads eastward over the North Sea. Through evaporation, this water gets saltier, and, as it cools in the Arctic, it becomes denser and sinks, feeding a deep cold current that moves southward hugging the eastern coast of the Americas [Rahmstorf, 2002]. As a whole, this system is called the Atlantic meridional overturning circulation (AMOC), and is commonly characterized via the zonally integrated volume transport.

The AMOC is responsible for the mild climates of northwestern Europe, which can be up to 6 °C warmer than locations at the same latitude but on the Pacific coast [Weijer et al., 2019]. Also, it draws heat away from the Southern Ocean and, by connecting the deep ocean with the surface, it is an important regulator of the carbon cycle [Nielsen et al., 2019].

Due to anthropogenic greenhouse gas emissions, the AMOC is projected to weaken [IPCC, 2013], and already some worrying trends have been observed. For example, data from buoys and satellite measurements shows a 15% reduction of the AMOC strength at 26° N in the period 2008-2012 compared to 2004-2008 [Smeed et al., 2014]. However, observations are sparse and span a very short period of time, so it is hard to draw precise conclusions. Similarly, climate projections have big uncertainties, with different models providing different answers and on different timescales [IPCC, 2013]. Adding to this the intrinsic uncertainty of emission scenarios, the projected weakening of the AMOC comes with huge error bars.

Nevertheless, evidence from paleoclimate records shows that the AMOC did experience abrupt fluctuations, known as Dansgaard-Oeschger (DO) events [Dansgaard et al., 1993], transitioning in a matter of decades from a vigorous overturning circulation, similar to the one observed today, to a much weaker one or even to a complete shutdown. These events could be the effect of quasi-periodic oscillations [Vettoretti et al., 2022] or of very rare fluctuations in the natural variability of the climate system [Ditlevsen et al., 2007]. Ice-core records also show a temporal correlation between volcanic eruptions and the onset of DO events [Lohmann and Svensson, 2022], which provides a complementary explanation. Regardless of the precise nature of the triggering mechanism, DO events suggest that the AMOC, at least in the past, was in a bistable regime.

The gradual weakening of the AMOC becomes then even more worrying, as it could be the herald of the fact that we are heading for a tipping point, where a small amount of extra warming would cause the AMOC to abruptly shut down.

In case of an AMOC collapse in the near future, Europe would come to experience a much colder climate, but repercussions would reverberate throughout the whole globe. Indeed, there would be more Arctic sea ice, a more northern jet stream and a southward shift of the Intertropical Convergence Zone (ITCZ), with consequent changes in global precipitation patterns [Bellomo et al., 2023; Liu et al., 2020]. Furthermore, in a world with a weaker AMOC, a warmer Southern Ocean would increase Antarctic melting rates [Holden et al., 2010], and, more worryingly, the reduced ventilation on the deep ocean would decrease its carbon uptake [Nielsen et al., 2019], further exacerbating the climate

crisis.

Due to its high impact, the shutdown of the AMOC has been intensively studied, and it is likely *the* canonical tipping problem in climate [Lenton et al., 2008]. Early studies by Stommel [Stommel, 1961] have shown that the advection of salt by the AMOC can create a positive feedback loop, with consequent emergence of two coexisting stable states, either with a vigorous or a weak overturning circulation.

As the climate warms, the Greenland Ice Sheet (GIS) melts, dumping increasingly large amounts of fresh meltwater into the Arctic, right where the major convection points of the AMOC are situated [Rahmstorf, 2002]. The meltwater lowers the density of the surface ocean, hindering its sinking, which is a major driver of the overturning pump. In the simple picture of the box model by Stommel, once the freshwater forcing reaches a critical value, the vigorous state becomes unstable and the AMOC collapses. We just experienced a *saddle-node bifurcation tipping*.

The multi-stability of the AMOC and the presence of bifurcations has been shown in models across different levels of complexity, from 2-box models [Stommel, 1961] and 5-box models [Cimatoribus et al., 2014], to intermediate complexity models [Rahmstorf et al., 2005], to fully fledged General Circulation Models (GCMs), albeit run at low resolution [van Westen and Dijkstra, 2023; Jackson and Wood, 2018; Jackson et al., 2023]. For box models, the different attractors can be computed analytically, while for GCMs, the standard way is to perform hosing experiments [Jackson et al., 2023], where the freshwater forcing is gradually increased until collapse, then steadily decreased until recovery of the system, drawing a hysteresis loop.

However, studies on DO events suggest that the bistability regime occurs at specific levels of CO_2 concentrations in the atmosphere [Menviel et al., 2020], pinpointing the region of spontaneous DO oscillations between 190 and 225 ppm [Vettoretti et al., 2022]. For comparison, preindustrial levels of CO_2 concentration are around 280 ppm and current ones are at 420 ppm and increasing. Hysteresis experiments on state-of-the-art climate models are very expensive and thus have not been run systematically [Castellana et al., 2019], while analysis of observational data is insufficient to tell whether the present day AMOC is bistable or not in the real world. Tough a recent study on Early Warning Signals from sea surface temperature measurements does suggest we are approaching a tipping point [Ditlevsen and Ditlevsen, 2023], others claim the results to be unreliable [van Westen et al., 2024b].

Moreover, further uncertainties come from the different physical processes that interact with the AMOC. There is considerable uncertainty on the projected melting rate of Greenland [Cazenave, 2006], and the Greenland Ice Sheet is not the only source of freshwater in the Arctic. For example, increased river runoff through the Canadian Archipelago can be a substantial contributor, and would act on a faster timescale with respect to the melting of the ice sheets [Rennermalm et al., 2006; Prowse et al., 2006]. Even more complex is the interaction with phenomena that don't affect directly the freshwater forcing. For instance, there is evidence that all-year Arctic sea ice is necessary for the collapse of the AMOC

[Nadeau et al., 2019; van Westen et al., 2024a], so the loss of sea ice due to global warming could improve the stability of the AMOC in its current strong state. Also, recent studies have identified winds in the Southern Ocean as a major dynamical driver of the AMOC [Nikurashin and Vallis, 2011, 2012; Jochum and Eden, 2015; Webb et al., 2021], so changes in atmospheric circulation due to climate change could come back to influence the response of the AMOC through this additional channel.

In the end, there is still no clear consensus on the multi-stability of the present day AMOC, and even less on whether the tipping point is close or far from present day conditions.

Nevertheless, our uncertainty on the matter together with the implications of an actual collapse of the AMOC make this topic crucial for adaptation and mitigation policies, and thus call for further research.

The vast majority of literature on tipping of the AMOC either focuses on hosing hysteresis experiments [Jackson et al., 2023] or on Early Warning Signals [Boulton et al., 2014]. Both rely heavily on the framework of bifurcation tipping, assuming the system is in quasi-equilibrium and that collapse happens due to a loss of stability of the current vigorous AMOC state as forcing parameters change. More recently, the condition of being at equilibrium has been relaxed by studying the effects of the rate at which the freshwater forcing is increased [Alkhayuon et al., 2019]. However, not many studies have addressed the possibility of a noise induced collapse, namely driven not by changes in forcing but rather solely by the internal variability of the climate. Even if such events are deemed rare, they could effectively anticipate the bifurcation tipping point, collapsing the AMOC at a stage where equilibrium analysis still told us we were safe.

One of the main reasons why noise induced tipping is understudied, is indeed the rarity of such events, which makes them extremely expensive to sample via ordinary numerical simulations. This is where rare event algorithms come into play. This broad family of algorithms is specifically designed to efficiently sample very rare events, while at the same time being able to properly address their probability of occurrence. Rare event algorithms have already been successfully applied to the climate system [Finkel et al., 2021], in particular for sampling extreme heatwaves [Ragone et al., 2018; Ragone and Bouchet, 2021; Gessner et al., 2021].

On the specific problem of noise induced AMOC collapse, very recently rare event algorithms have been applied to a 5-box model [Castellana et al., 2019] and to an intermediate complexity GCM [Cini et al., 2024].

In this chapter we will do something very similar to the latter, applying the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) algorithm on the Versatile Ocean Simulator (VerOS). We will first describe the VerOS model, and its stability landscape for the AMOC. We will then proceed with a theoretical overview of rare event algorithms in general and the GKLT one in particular, validating it on the Ornstein-Uhlenbeck process. Finally, we will apply the GKLT algorithm to the VerOS model, and attempt to sample very rare spontaneous collapses.

5.2 VerOS and the stability landscape of the AMOC

Before looking for transitions between attractors, we need to have an idea of the stability landscape of the AMOC as represented in the VerOS model. Fortunately, an extensive study [Lohmann et al., 2024] has characterized it in detail. In this section, we will briefly describe the VerOS model and summarize the results of Lohmann et al. [2024] which are relevant for this work.

5.2.1 The Versatile Ocean Simulator model

To simulate the dynamics of the AMOC, we use the Versatile Ocean Simulator (VerOS) [Häfner et al., 2018]. It is an intermediate complexity ocean-only model, which allows us to generate very long simulations at a contained computational cost. Moreover, VerOS is written entirely in Python, which makes it very easy to use compared with more standard Fortran models. The model is highly customizable, and since this project was done in close collaboration with Johannes Lohmann, we use the same model settings. These details and many important features of the model are well described in Lohmann and Ditlevsen [2021]; Lohmann et al. [2024]. In the following, I will present only the most relevant for the understanding of the results shown in this chapter.

VerOS solves the primitive equations for the global ocean with a finite difference method. In our settings, it has 40 vertical layers of increasing thickness (from 23 m at the surface to 274 m at the bottom) and a horizontal grid with a constant longitudinal resolution of 4° and a variable latitudinal one that increases from 5.3° at the poles to 2.1° at the equator. The model domain ends at 80°N, so there is no connection of the Atlantic and Pacific oceans through the Arctic. Unstable density profiles are removed by a gradual increase of the vertical diffusivity, which is a smoother variant of the more crude convective adjustment used in many intermediate complexity atmospheric models, [Vallis, 2017].

Being an ocean-only model, VerOS doesn't have a dynamical atmosphere, but rather it uses climatologies derived from ERA-40 [Uppala et al., 2005] as boundary conditions. In particular, heat exchange between the ocean surface (uppermost layer of the grid) and the atmosphere is modeled with a first order Taylor expansion [Barnier, 1998]

$$Q(T_i) = Q(T_i^{\text{obs}}) + \left.\frac{\partial Q}{\partial T}\right|_{T_i^{\text{obs}}} \left(T_i^{\text{obs}} - T_i\right), \tag{5.1}$$

where, at grid cell i, T_i is the sea surface temperature predicted by the model, while the climatological sea surface temperature T_i^{obs}, heat flux $Q(T_i^{\text{obs}})$ and its derivative are computed from ERA-40.

In the real world, the surface salinity S of the ocean is controlled by evaporation and precipitation, plus river runoff close to coast. On the other hand, in the model, we use something much simpler, namely relaxing the surface salinity to a climatological value S^{obs}, again estimated from ERA-40. More precisely, the salinity flux at the surface is modeled as

$$\phi_i = \frac{h}{\tau_S}\left(S_i^{\text{obs}} - S_i\right), \tag{5.2}$$

where $h = 23\,\mathrm{m}$ is the thickness of the upper ocean layer and τ_S is the characteristic timescale of the relaxation.

This latter parameter is of paramount importance for studying the tipping of the AMOC. Indeed, a short relaxation time will destroy any considerable salinity anomalies, effectively shutting down the salt advection feedback, which is the main mechanism that maintains a multistable AMOC. In this work, as well as in Lohmann et al. [2024], we use $\tau_S = 2\,\mathrm{years}$.

Finally, to characterize the AMOC, we take the zonal integral of the velocity field in the Atlantic Ocean. This gives a two-dimensional field v as a function of depth (z) and latitude (λ), which can be expressed as $v(\lambda, z) = \nabla \times \psi(\lambda, z)\hat{\phi}$, where $\psi(\lambda, z)$ is the stream function and $\hat{\phi}$ is the versor pointing east.

We are then interested in the overturning component of the stream function:

$$\psi_o(\lambda, z) = -\int_{z_0}^{z} v(\lambda, z')\mathrm{d}z', \tag{5.3}$$

where z_0 is the depth of the ocean.

What we will call the AMOC strength is the maximum of ψ_o in the North Atlantic ($\lambda > 50°N$) and at a depth $z > 500\,\mathrm{m}$ to exclude the wind driven maximum. It has the dimensions of $\mathrm{m^3 s^{-1}}$, but in the climate community it is often measured in Sverdrup ($1\,\mathrm{Sv} = 10^6 \mathrm{m^3 s^{-1}}$).

5.2.2 The complex stability landscape of the AMOC

Classical studies on the collapse of the AMOC rely on hosing experiments, where meltwater from the Greenland Ice Sheet is inputted as a freshwater forcing in the North Atlantic (green area in fig. 5.1). By slowly varying this control parameter, one can perform a hysteresis analysis. In many conceptual box models like Stommel's [Stommel, 1961] the AMOC has the standard saddle-node bifurcation tipping point highlighted in the black square in panel (**B**) of fig. 5.2. For high values of the freshwater parameter there is a mono-stable weak AMOC, and for low values a mono-stable strong AMOC. For intermediate values, there are two existing stable states. Then, during a hosing experiment, as the freshwater forcing in gradually increased, the system tracks the upper branch up to where it becomes unstable and then relaxes to the lower branch. Similarly, as freshwater forcing decreases the system will track the lower branch, going back to a vigorous state at values of the control parameter which are much lower than they were at the first tipping point. This simple hysteresis loop neatly describes the difficulty of coming back once such a tipping point is crossed.

Box models like Stommel's are well named *conceptual*, as they give a simple qualitative description of the tipping of the AMOC. On the other hand, the analysis performed in Lohmann et al. [2024] yields a far more complex multi-stability landscape for the AMOC in the VerOS model, with up to 9 coexisting attractors (fig. 5.2). This rich landscape is thought to be due to the spatial extent of the system [Bastiaansen et al., 2022; Rietkerk

et al., 2021] which creates intermediate tipping points as reorganizations of local structures, for instance the subpolar gyre. There is also reason to believe that the chaotic nature of a realistic atmosphere (which VerOS lacks) would result in the merging of close by attractors and thus a simplification of the multi-stability landscape.

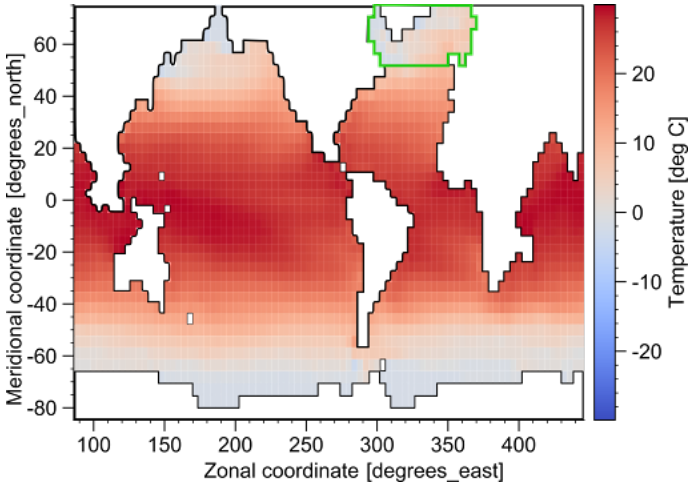


Figure 5.1: Geometry of the ocean basin of the VerOS model with the resolution used in this work. The background color is a snapshot of sea surface temperature and the highlighted green area is the one where freshwater forcing is applied. Image courtesy of Johannes Lohmann.

Noise induced transitions

As already pointed out in the introduction of this chapter, the object of this work is not to perform a hosing experiment, but rather to study noise-induced collapses of the AMOC. Note that at this stage the VerOS model is fully deterministic, so "noise-induced" might look misleading. What we mean is that the transition would be caused entirely by internal (chaotic) variability, and not by changes in the freshwater forcing parameter. To do so we want to place ourselves in a relatively simple region of the stability landscape, where there are no intermediate attractors between a vigorous and a weak AMOC.

At the time when I conducted my experiments, the point on the purple branch at a freshwater forcing $F = 0.3525\,\mathrm{Sv}$ and an equilibrium average AMOC strength around $8.3\,\mathrm{Sv}$ (red arrows in fig. 5.2) seemed a good candidate. At that time, the cyan and light green branches immediately below the purple one were not known, and the chosen point seemed to have a clear, yet narrow, path between the end of the orange branch and the beginning of the dark green one, straight towards the collapsed states of the yellow branch at $3\,\mathrm{Sv}$ (red vertical lines in fig. 5.2).

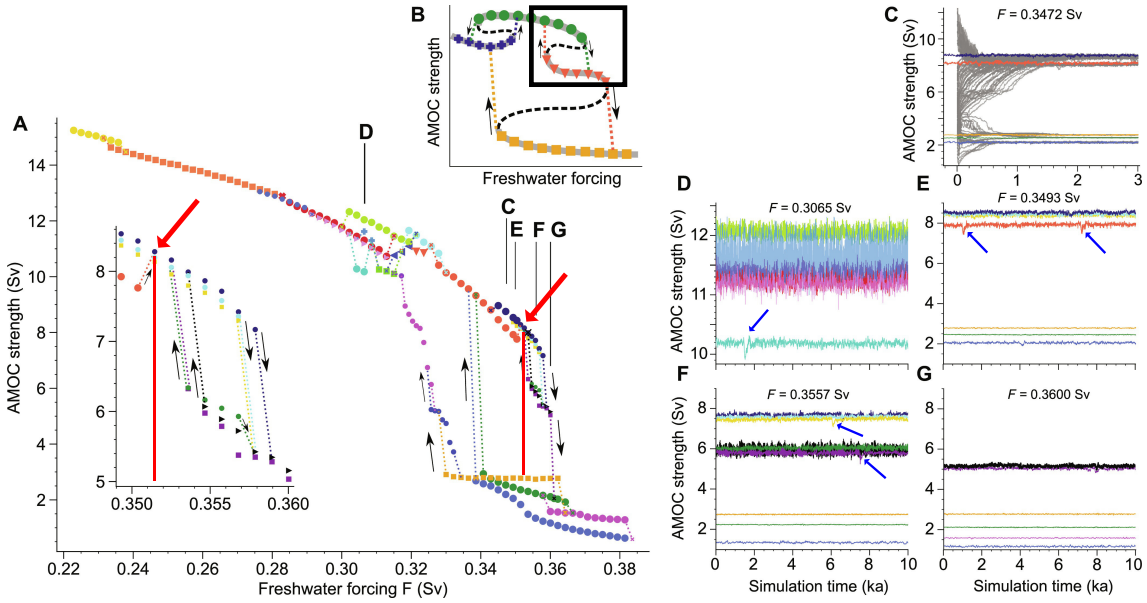


Figure 5.2: This is an annotation of Figure 3 from Lohmann et al. [2024]. Panel (**A**) shows the complex bifurcation diagram of the AMOC as a function of the freshwater forcing control parameter. Each point corresponds to the mean of the AMOC strength over 1000 years. The inset is a zoom of the region close to the major tipping point. The red arrows and vertical lines indicate the point from which the simulations in this work are initialized. (**B**) is a schematic representation of the concept of intermediate tipping points, with the black square highlighting the classical scheme of a standard bifurcation tipping point. Panel (**C**) shows simulations with perturbed initial conditions converging to different attractors. Panels (**D** to **G**) show long equilibrium simulations for coexisting attractors at four different values of the control parameter. In these panels, the blue arrows indicate abrupt spontaneous weakening of the AMOC, that relax back to the attractor in a matter of a few hundred years. More details on the figure can be found in Lohmann et al. [2024].

5.3 Rare event algorithms

As we have mentioned in the introduction, rare events can be a very important object of study, as even if they happen with a low probability, they have huge impacts. This extends beyond the field of climate science. For example rare events can be extremely important for chemical reactions [Chandler, 2005], complex protein folding processes [Noé et al., 2009; ten Wolde and Frenkel, 1997], or even buffer overflows in online queuing systems [Heidelberger, 1995].

The main challenge with studying these events comes from their very nature of being rare, which makes it difficult, or expensive, to properly sample them. The idea of rare event algorithms is to tamper with the system to artificially increase the occurrence of the events we want to study. Afterwards, the intervention of the algorithm can be taken into account and properly corrected to obtain unbiased results.

While it is, in principle, possible to act directly on the equations that generate the dynamics and push the system towards the region of phase space we are interested in, this is not usually viable for complex climate models, and it makes unbiasing operations very hard. The alternative option is to act only on initial conditions, which has the advantage that generated trajectories are physically consistent storylines. Furthermore, this choice comes with the technical advantage to be able to treat the system as a black box.

The first possible approach is that of importance sampling [Rubinstein and Kroese, 2016; Rubino and Tuffin, 2009], which twists the stationary measure of the system to over-sample specific rare events. While conceptually simple, this method can be very sensitive to how the measure is twisted, and thus requires good prior knowledge of the system.

A more robust and model-agnostic approach is the one of genealogical algorithms [Moral and Garnier, 2005]. In this broad class of methods, the user doesn't need to directly act on the stationary distribution of the system, but rather they have to provide a score function that is maximal when the rare event of interest happens. The algorithm will then use an ensemble of trajectories, and iteratively clone or kill members to increase the score function, thus approaching the rare event. While for standard importance sampling the perturbation to the stationary measure of the system is applied in one step, for genealogical algorithms it is achieved iteratively, with each small perturbation pushing the system in the right direction. The iterative nature of these algorithms gives more control over the trajectories and can thus be very effective for studying extremely rare events. A family of such algorithms is that of Interacting Particle System (IPS) [Moral and Garnier, 2005; Garnier and Moral, 2006], where an ensemble of trajectories is propagated for a given time step, after which the most promising members are cloned, while the least promising ones are killed.

An orthogonal approach to importance sampling and IPS is that of Adaptive Multilevel Splitting (AMS) [Cérou et al., 2011; Rolland et al., 2016], where the trajectories in the ensemble are left to evolve up until they reach specific states, or for a fixed amount of time

[Lestang et al., 2018]. The iterations of the algorithm then do not go step by step with the dynamics of the system, but rather by iteratively re-simulating the worst member.

Both approaches have benefits and drawbacks, and are thus suited for different types of problems. In this work we use the Giardinà-Kurchan-Lecomte-Tailleur algorithm [Giardinà et al., 2011], belonging to the family of IPS methods. In the following of this section we will describe the details of this algorithm and validate it on the Ornstein-Uhlenbeck process.

5.3.1 The Giardinà-Kurchan-Lecomte-Tailleur algorithm

There are many good descriptions of the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) algorithm [Giardinà et al., 2006; Garnier and Moral, 2006; Wouters and Bouchet, 2016; Lestang et al., 2018]. However, the ones provided in application papers are very summary and are often relegated to supplementary material [Ragone and Bouchet, 2020; Ragone et al., 2018; Cini et al., 2024], while the original one [Giardinà et al., 2011] is too long and detailed for our purposes. Hence, here I try to provide a compromise between the two, giving a complete, yet synthetic, description of the algorithm, with a particular focus on its practical implementation.

General formulation

Let us assume we have a (stochastic) system $\mathrm{d}X = f(X,t)\mathrm{d}t + \sigma(X,t)\mathrm{d}B(t)$ with $X \in \mathbb{R}^d$, and an observable $O(X(t)) \in \mathbb{R}$ in which we are interested. We can then identify a region $\mathcal{A} \subset \mathbb{R}^d$ of the phase space that is not often visited by the system and has particular values of the observable O. We then define a score function $V(\{X\},t)$ that is a proxy for the probability of trajectory $\{X\}$ reaching $\mathcal{A}$. At this stage, these definitions are very vague, as there are many variants of the algorithm. For example, we may want to reach $\mathcal{A}$ eventually, or within a specified time interval.

For simplicity here, we will assume that the score function is either a simple function v of the observable at time t, i.e. $V(\{X\},t) = v(O(X(t)))$, or its time integral $V(\{X\},t) = \int_0^t v(O(X(s)))\mathrm{d}s$. In our case, we want to study the collapse of the AMOC, and a natural observable is the AMOC strength itself $O = \mathrm{AMOC}$, possibly time-averaged. Then, the simplest score function choice is using simply $V(X,t) = -\mathrm{AMOC}(X(t))$. This choice makes the assumption of persistence to approximate probabilities, namely: "the AMOC is more likely to collapse in the future if it is already weak right now". Whether this choice is justified will be discussed later, and for the moment, let us continue with the theoretical description of the algorithm, which is the same for any V.

After we have defined the score function, we initialize N ensemble members with trajectories $\{X_j^{(0)}\}$, $j = 1, \dots, N$, each ending at time $t_j^{(0)}$. For simplicity and without loss of generality, we can assume $t_j^{(0)} = t_0 = 0$, and, since the dynamics is Markovian, we can ignore everything that happens before t_0. Finally, to complete the initialization of our ensemble we compute the initial scores $V_j^{(0)} = V(\{X_j^{(0)}\})$, and we assign to each trajectory an unbiasing weight $\pi_j^{(0)} = 1$. This weight will track the likelihood to observe trajectory j

when running the climate model with no rare event algorithm, and it will be crucial to have access to the unbiased probabilities of events sampled by the biased trajectories.

Once we have our initial ensemble, we need to select a resampling time τ and a selection strength k. Then, for each iteration, we will integrate forward in time each ensemble member producing candidate trajectories (denoted with $\tilde{\bullet}$), and select which ones will survive to the next iteration based on their scores. More precisely, for $i = 1, \dots, I$:

1. Extend each trajectory integrating forward in time for one resampling step τ, obtaining $\{\tilde{X}_j^{(i)}\}$, $j = 1, \dots, N$ that run from t_0 to $t_i = i\tau$.

2. Compute for each trajectory its score function at the end of the resampling step:

$$\tilde{V}_j^{(i)} = V(\{\tilde{X}_j^i\}, t_i)$$

3. Compute for each trajectory a weight according to the improvement of its score in the last resampling step:

$$\tilde{w}_j^{(i)} = \exp\left(k\left(\tilde{V}_j^{(i)} - V_j^{(i-1)}\right)\right)$$

and then normalize the weights so that they sum up to N:

$$w_j^{(i)} = \frac{1}{Z^{(i)}}\tilde{w}_j^{(i)}, \quad Z^{(i)} = \frac{1}{N}\sum_{j=1}^{N}\tilde{w}_j^{(i)}$$

4. Resample N new trajectories by cloning or killing according to the weights. More precisely:

   (a) Compute a tentative number of clones for each trajectory: $\tilde{m}_j^{(i)} = \lfloor w_j^{(i)} + u_j^{(i)} \rfloor$, where $\lfloor \bullet \rfloor$ is the floor function and $u_j^{(i)} \sim \mathcal{U}(0,1)$ are N independent random numbers distributed uniformly between 0 and 1.

   (b) Adjust the number of clones so that they sum to N. In particular, if $\Delta N^{(i)} = \sum_{j=1}^{N} \tilde{m}_j^{(i)} - N > 0$, then we randomly select $\Delta N^{(i)}$ values of j among the ones for which $\tilde{m}_j^{(i)} > 0$ and reduce by one the number of clones of these trajectories. On the other hand, if $\Delta N^{(i)} < 0$, then we randomly clone $|\Delta N^{(i)}|$ trajectories selecting again from the ones for which $\tilde{m}_j^{(i)} > 0$. Finally, we are left with definitive number of clones $m_j^{(i)}$ that do add up to N.

   (c) Create the parent mapping

$$\begin{aligned} &p^{(i)} : \{1, \dots, N\} \to \{1, \dots, N\}, \\ &p^{(i)}(j) = \min\{l \text{ such that } \sum_{n=1}^{l} m_n \geq j\} \end{aligned} \tag{5.4}$$

   which links each trajectory to the one it was cloned from.

(d) Actually perform the cloning and killing: $\{X_j^{(i)}\} = \{\tilde{X}^{(i)}_{p^{(i)}(j)}\}$. Similarly, update the scores and unbiasing weights:

$$V_j^{(i)} = \tilde{V}^{(i)}_{p^{(i)}(j)}, \quad \pi_j^{(i)} = \frac{\pi^{(i-1)}_{p^{(i)}(j)}}{w^{(i)}_{p^{(i)}(j)}}$$

We point out that in the original paper by Giardinà et al. [Giardinà et al., 2011], the weights in point 3 are computed not as the improvement of the score function but as the score function itself at the end of the resampling step. However, in [Garnier and Moral, 2006], the authors suggest using the improvement of the score function. Another point is that if the model is fully deterministic, the cloning step 4.d will yield trajectories that will evolve in exactly the same way. To avoid this we introduce a small perturbation to the trajectories just after the cloning step, which allows them to diverge. The details of this point are further discussed in section 5.4.

Computing averages over a biased ensemble

Given an observable $U(\{X\})$, one can approximate its average over the stationary measure of the system with the empirical average over many realizations, where each trajectory has the same weight. When we run the rare event algorithm, we proceed in the same way, but we need to account for the fact that we biased the trajectories by cloning and killing them, and this is achieved by the unbiasing weights $\pi_j^{(i)}$:

$$\bar{U}_N = \frac{1}{N}\sum_{j=1}^{N} \pi_j^{(I)} U(\{X_j^{(I)}\}) \sim_{N\to\infty} \mathbb{E}\left[U(\{X\})\right], \tag{5.5}$$

where $\mathbb{E}$ is the expectation over the stationary measure. Since the trajectories are not independent, the central limit theorem doesn't apply, and thus the typical error can be larger than $1/\sqrt{N}$. In practice, convergence will be faster for observables closely linked to the score function V used for biasing the ensemble, while for others it may be more efficient to run a simple unbiased Monte Carlo estimation.

Note that if we are interested in computing probabilities rather than averages, we can simply think of a probability as the expectation of an indicator function:

$$\mathbb{P}\left(U(\{X\}) > u\right) \equiv \mathbb{E}\left(\mathbb{1}_{U(\{X\})>u}\right) \tag{5.6}$$

Now, if we unravel the expression for the unbiasing weights, we get

$$\pi_j^{(I)} = \left(\prod_{i=1}^{I} Z^{(i)}\right) \exp\left(-k\left(V(\{X_j^{(I)}\}_{0\le t\le I\tau}) - V(\{X_j^{(I)}\}_{t=0})\right)\right), \tag{5.7}$$

where we had a telescopic canceling in the exponential. This shows that the bias we introduced with the algorithm is proportional to the exponential of the improvement of the score function over the whole algorithm run for each trajectory.

This algorithm was initially developed to compute large deviation rate functions [Giardinà et al., 2006], and in fact there is a large deviation principle for the variation of the score function during the whole algorithm run

$$v_T = V(\{X\}_{0\leq t\leq T}) - V(\{X\}_{t=0}), \tag{5.8}$$

which can be written as in Ragone et al. [2018]:

$$\mathbb{P}\left(v_T \geq a\right) \asymp_{T\to\infty} e^{-T\mathcal{I}(a)}, \tag{5.9}$$

where $\asymp$ denotes log-equivalence and $\mathcal{I}$ is the large deviation rate function, which can be computed as the Legendre-Fenchel transform of the scaled cumulant generating function $\lambda(k)$: $\mathcal{I}(a) = \sup_k\{ka - \lambda(k)\}$. The GKLT algorithm provides an efficient estimation of such scaled cumulant generating function at the value k (and its neighborhood [Ragone and Bouchet, 2020]) used as selection strength.

$$\lambda(k) = \lim_{I\to\infty} \frac{1}{I\tau} \sum_{i=1}^{I} \log Z^{(i)} \tag{5.10}$$

A remark on finite size ensembles

In the previous subsection, we showed that our empirical averages are well-behaved in the limit $N \to \infty$. However, in practice, the computational cost of running a climate model, albeit simple, is still very high. This means that our ensemble will necessarily be quite small, usually with at most a few hundred trajectories [Ragone et al., 2018]. In this work, we wanted to run many experiments, which led us to an even smaller ensemble size of $N = 50$. With ensembles this small, we expect large errors in the averages that we compute. Even worse, when estimating probabilities, we have no theoretical guarantee that our estimate will be smaller than 1, and, in fact, we did get probabilities even larger than 2.

Proper choice of score function, resampling time and selection strength

The algorithm described above has four 'hyperparameters': the ensemble size N, the resampling time τ, the selection strength k and the score function V.

Choosing the ensemble size is a matter of compromise between accuracy and computational cost, while the other parameters require some degree of understanding of the system on which we want to run the algorithm.

The resampling time should be of the order of the Lyapunov time of the system [Wouters and Bouchet, 2016]. Long enough to allow clones of the same ensemble member to separate sufficiently, but not so long that they relax back to the attractor. Namely, we don't want segments of trajectories to completely lose memory of their initial condition.

Concerning the selection strength k, it should be chosen based on how extreme of an event we want to observe: the more extreme the event, the higher the selection strength. More precisely, if we assume the random variable v_T defined in eq. (5.8) has a normal

distribution with mean μ and variance σ^2, then to observe values of the order of a, an indicative value of k should be

$$k \sim \frac{a - \mu}{\sigma^2} \tag{5.11}$$

Equivalently, if we want to push the system n standard deviations from the mean μ, we should set $k \sim n/\sigma$ [Le Priol et al., 2024].

However, the higher k is, the more trajectories will be killed at each iteration, greatly reducing the diversity of the final ensemble. At the extreme this may lead to all trajectories sharing a single 'parent', and we call this *ensemble collapse* or *extinction*. It becomes then clear that the ensemble size limits how heavily we can push the system, and, thus, the values of a that we will be able to sample.

Finally, the score function can be the most crucial parameter of the algorithm. It can make the difference between a more or less efficient sampling of the event of interest or a complete failure to observe the event at all. In fact, it is possible to prove that there is an *optimal* score function [Chraibi et al., 2020], which is closely related to the committor function and gives the most efficient sampling of our rare event of interest. Unfortunately, our ignorance of the committor function is often the reason we want to run the rare event algorithm in the first place. A possible solution, which was the general goal of this thesis, is to create a loop where machine learning is used on existing data to compute a first estimate of the committor function, which is then used to run the rare event algorithm. It, in turns, will efficiently generate new data, that would allow us to refine our estimate of the committor, thus closing the loop.

In this chapter, however, the goal was just to implement and test the Giardinà-Kurchan-Lecomte-Tailleur algorithm on the VerOS model, so the score function was chosen to be simply the observable itself, in line with many studies that make use of this rare event algorithm [e.g. Ragone et al., 2018; Cini et al., 2024; Le Priol et al., 2024].

5.3.2 Implementation

The practical implementation of the GKLT algorithm is available at `https://github.com/AlessandroLovo/REA-Veros`. Producing this code repository was a significant portion of the work presented in this chapter, particularly so because I set myself the goal to produce a clear, well commented and flexible code. As it is, the code supports running on single machines or on clusters. In fact, during my work I used three different high performance computing clusters (in Lyon, Utrecht and Copenhagen). This allowed me to parallelize my work greatly, running three experiments at the same time, which was very useful, since each experiment could take up to a week to run. On the other hand, the different protocols on the three clusters required my code to be even more flexible. Moreover, I designed the code to support running virtually any model, not just the VerOS model, with minimal changes and precise instructions on how to implement them. I thus hope that my efforts will be useful for future scholars working with the GKLT algorithm.

5.3.3 Validation on the Ornstein-Uhlenbeck process

In this subsection, I describe a quick benchmark of the GKLT algorithm on the Ornstein-Uhlenbeck process. This will help the reader familiarize with the algorithm, and it was essential to debug my rather complex codebase.

The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck (OU) process describes the evolution of the scalar variable $X \in \mathbb{R}$ with the simple stochastic differential equation

$$dX = -\lambda(X - \mu)dt + \sigma dW. \tag{5.12}$$

The system is simple enough to be solvable analytically, and if the distribution of the system at time t is Gaussian with mean $m(t)$ and variance $v(t)$, then it will evolve according to

$$\begin{cases} \dot{m} = -\lambda(m - \mu) \\ \dot{v} = \sigma^2 - 2\lambda v. \end{cases} \tag{5.13}$$

For simplicity, in the following we will assume $\mu = 0$ and $\lambda = \sigma = 1$, moreover, we will initialize the system with $X(0) = 0$. With these assumptions, $m(t) = 0$ and $v(t) = \frac{1}{2}\left(1 - e^{-2t}\right)$. To have a parallel with what we want to do with the AMOC, we will look for the probability that $X(T) \leq a$, for which the theoretical solution is

$$\mathbb{P}(X(T) \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi v(t)}} e^{-\frac{x^2}{2v(t)}} \mathrm{d}x = \frac{1}{2}\mathrm{erfc}\left(-\frac{a}{\sqrt{1 - e^{-2T}}}\right), \tag{5.14}$$

where erfc(•) is the complementary error function. Without loss of generality, we will focus on a total integration time $T = 2$.

Running the GKLT algorithm on the Ornstein-Uhlenbeck process

In [Wouters and Bouchet, 2016], the authors present a thorough analysis of how the different parameters of the GKLT algorithm can affect its performance in sampling rare events, applied to the Ornstein-Uhlenbeck process, with these exact choices for μ, λ, σ and T. In the following, I will present just a few examples, using the score function $V(\{X\}, t) = -X(t)$, a selection strength $k = 4$ and a resampling time $\tau = 0.1$, which means $I = 20$ iterations.

From fig. 5.3, we can see the backward reconstructed trajectories $\{X_j^{(I)}\}$ for an $N = 1000$-member ensemble. Going back in time, more and more trajectories share the same ancestors, resulting in reduced ensemble diversity. In this case, the ensemble size is big enough to avoid extinction, as described in section 5.3.1. At the end of the simulation, the rare event algorithm shifted the distribution by roughly 3 standard deviations. If we evaluate eq. (5.11) for the suggested selection strength, we would have expected the ensemble to shift $k\sqrt{v(T)} = 2.8$ standard deviations, which is consistent with what we actually find.

If we use eq. (5.5) to compute the unbiased probabilities of being below threshold a, more precisely,

$$\mathbb{P}(X(T) \leq a) = \frac{1}{N} \sum_{j=1}^{n} \pi_j^{(I)} \mathbb{1}_{X_j^{(I)}(T) \leq a}, \tag{5.15}$$

we find that the rare event algorithm allowed us to sample events with a probability as small as 10^{-8} (fig. 5.4), at the same computational cost of a control run that only managed to sample events with probability greater that 10^{-4}. An impressive ten thousand times improvement. By repeating the experiment 6 times, we can confirm that the probabilities we obtain are indeed unbiased with respect to the analytical result. Importantly, the estimated probabilities fluctuate less around the theoretical result when $a \in [-2.5, -1.8]$, i.e. in the bulk of the shifted distribution (fig. 5.3). The control run is much more effective at sampling less rare events, as a rare event algorithm run with higher selection strength would be more effective to sample even more extreme events [Wouters and Bouchet, 2016].

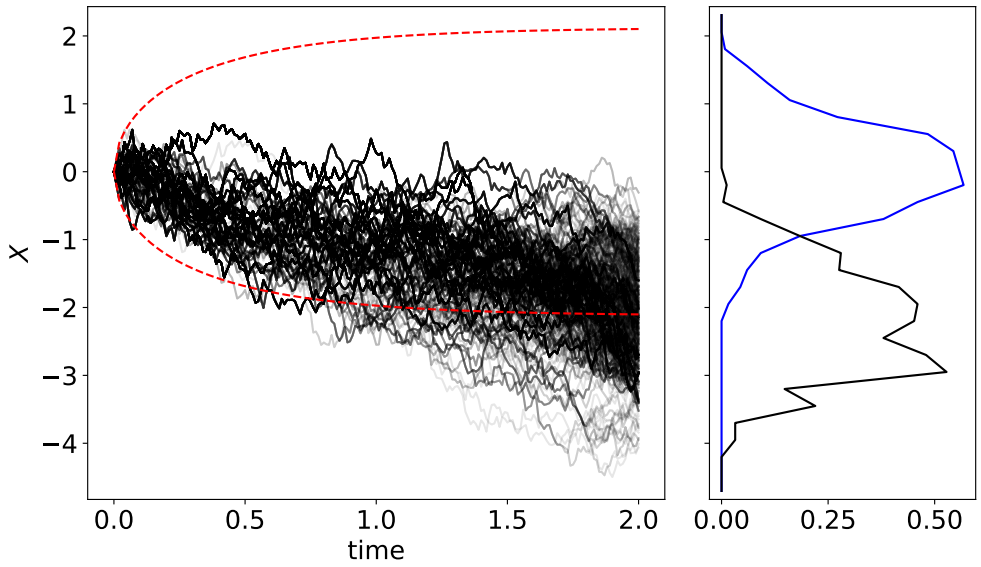


Figure 5.3: Rare event algorithm run on the Ornstein-Uhlenbeck process, at selection strength $k = 4$ and resampling time $\tau = 0.1$, performed on an ensemble of $N = 1000$ members. On the left: backward reconstructed trajectories in black and theoretical confidence interval $(\pm 3\sqrt{v(t)})$ in red. On the right: histogram of the ensemble members at the end time $T = 2$ in black. In blue the histogram of a control run with the same number of members at $T = 2$. The rare event algorithm produced a shift of the distribution of around 3 standard deviations.

One aspect which is often neglected in many application papers of the GKLT algorithm, is the incremental gain that happens at each iteration. Indeed, one would expect that since at each resampling step we kill and clone trajectories, the longer we run the algorithm, the more rare events we will observe. This is true only up to a certain point.

In fig. 5.5, we show not the backward reconstructed trajectories, but the state of the ensemble at the end of each integration time step, highlighting the killed and cloned members. From the top panel of the figure we can see clearly that each resampling step

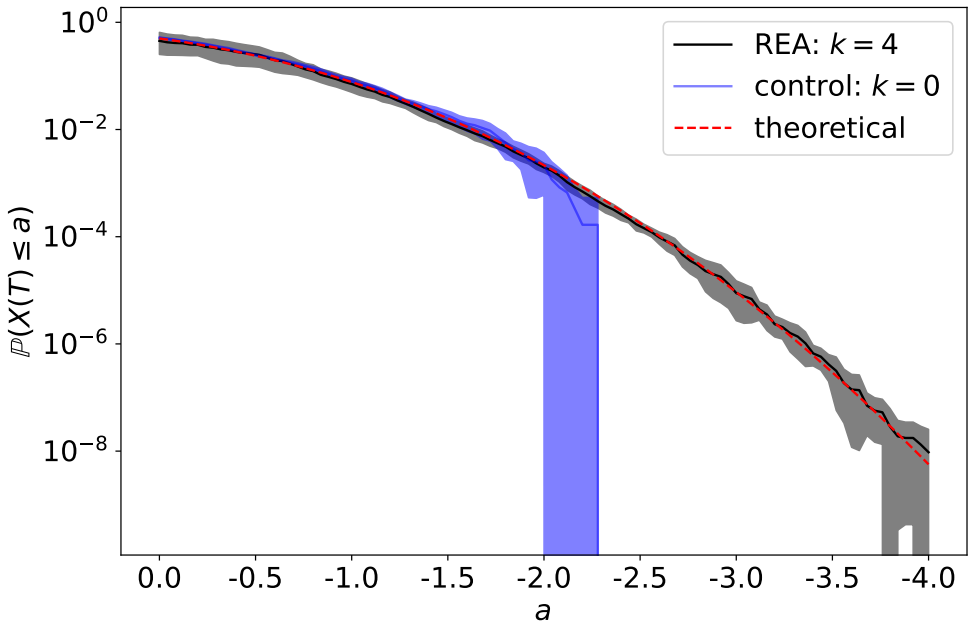


Figure 5.4: Probabilities of being below threshold a at time $T = 2$. In black for the rare event algorithm and in blue for the control run. In both cases the experiment is repeated 6 times, the solid line represents the mean over the 6 realizations and the shading corresponds to 1 standard deviation. The red dashed line is the theoretical result as given in (5.14). The rare event algorithm allows to efficiently sample very rare events, giving an unbiased estimate of the probabilities.

causes the distribution to brusquely shift downward, while during the integration step the ensemble relaxes back upward toward the center of the attractor at $x = 0$. As the ensemble moves away from the attractor, this upward drift becomes stronger, and after $t = 1.1$ it compensates the downward kicks, resulting in a plateauing effect. The interesting point is what happens to the estimated probabilities. Since we have seen that the best estimate are for events in the bulk of the shifted distribution, we can monitor, at each iteration i, the unbiased probability $p_{1/2}^{(i)}$ of being below the median $a_{1/2}^{(i)}$ of the current distribution. In other words,

$$p_{1/2}^{(i)} = \mathbb{P}(X(i\tau) < a_{1/2}^{(i)}) = \frac{1}{N}\sum_{l=1}^{N/2} \pi_{j_l}^{(i)}, \quad X_{j_1}^{(i)} \leq X_{j_2}^{(i)} \leq \ldots \leq X_{j_l}^{(i)} \leq \ldots \leq X_{j_N}^{(i)}. \tag{5.16}$$

In the bottom panel of fig. 5.5, we can see that this probability saturates as well, and the events at time $t = 2$ are no rarer than the ones at $t = 1$.

This is because the members at iteration $i - 1$ that are more likely to be cloned (and thus have a large weight $w_j^{(i-1)}$) reached farther from the attractor, and thus feel a stronger pull towards it. This means that at iteration i, they will be more likely to be killed, and so all the progress that they achieved is lost. On the other hand the members that are more likely to survive at iteration i, had a modest weight $w_j^{(i-1)}$, and so, overall, the unbiasing weights $\pi_j^{(i)}$ don't keep decreasing indefinitely as the algorithm progresses.

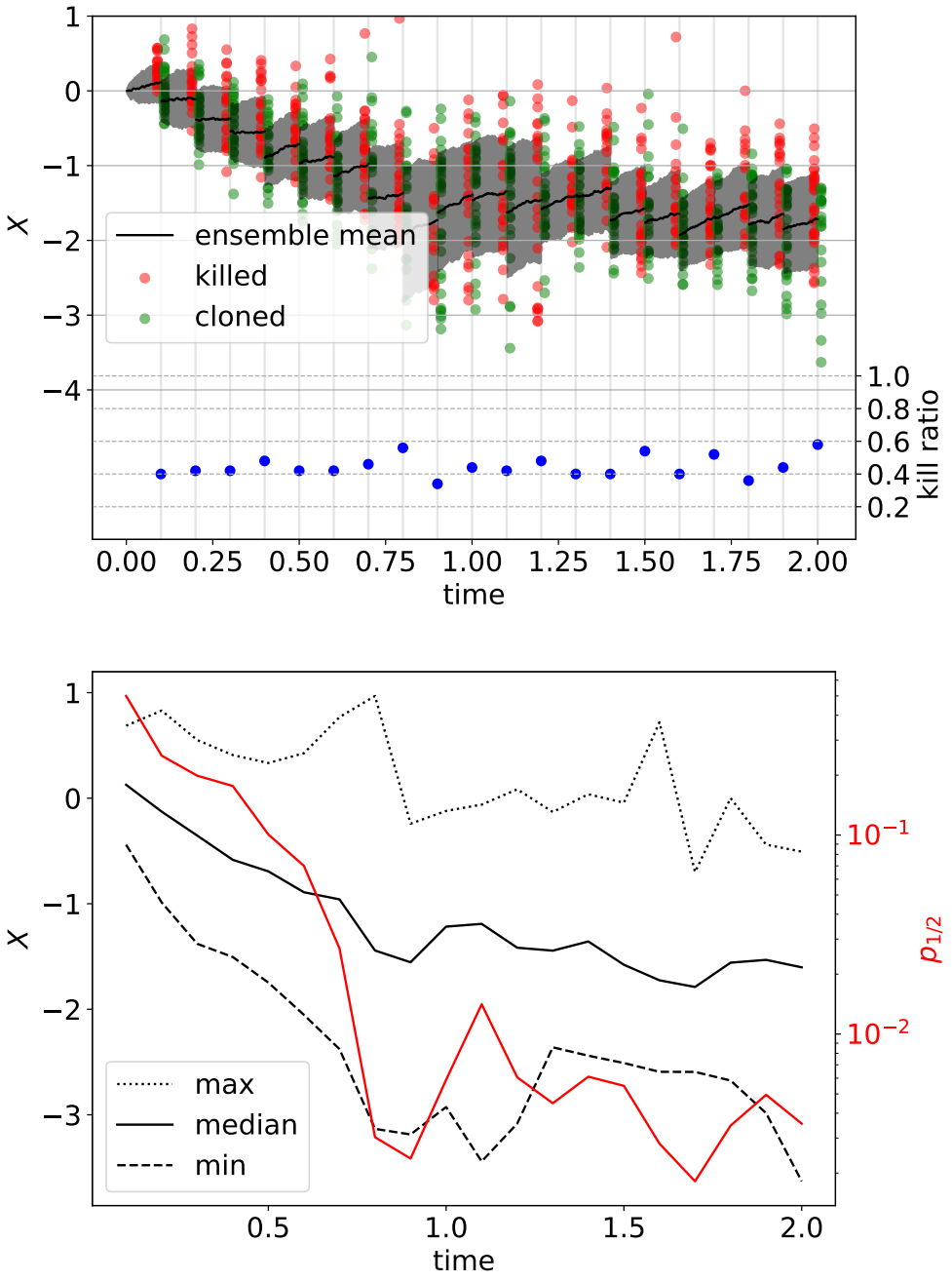


Figure 5.5: Evolution of an $N = 50$ member ensemble during a rare event algorithm run on the Ornstein-Uhlenbeck process with selection strength $k = 4$. In the top panel, the black line represents the ensemble mean and the shaded area corresponds to one standard deviation. Each vertical line is at a resampling step (every $\tau = 0.1$). In red are the ensemble members killed and in green the ones that survived to the next iteration. The blue points on the bottom (right y-axis) show the fraction of ensemble members killed at each resampling step. In the bottom panel, the black lines (left y-axis) represent the minimum, median and maximum values of the ensemble before each resampling step. The red line (right y-axis) is the unbiased probability $p_{1/2}$ of being below the median, as given by (5.16). After $t = 1.1$ the ensemble reaches a plateau, where the downward kick at each resampling step is compensated by the upward drift during integration of the process. At the same time $p_{1/2}$ stops decreasing, saturating around 0.005.

5.4 Attempting to tip the AMOC in the VerOS model

Now that we are familiarized with the workings of the GKLT algorithm on a case where we know what to expect, we can proceed to apply it to the VerOS model. When we worked with the OU process, there was a single stable attractor, and we were trying to push the system away from it. For the AMOC, we know that there are many attractors, but we chose the value of freshwater forcing to be in a region where the stability landscape is not too complicated (see section 5.2.2). If we assume that there are only two stable attractors, the one from which we start and the one we want to reach, then the rare event algorithm will be needed just to push the system over the potential barrier. After that, trajectories will spontaneously relax to the target attractor.

We will start by characterizing the neighborhood of the starting point using a long control run. With this we'll be able to properly choose the hyperparameters of the rare event algorithm. Consequently, we'll try to actually run the rare event algorithm. As we will see, we won't be able to observe a transition to the collapsed state, so we will proceed to derive an atmospheric noise (section 5.5) to add to ocean-only VerOS model, to increase its variance and speed up its dynamics. This still won't be enough to reach the target attractor, but will give us interesting insight on the response of the AMOC to such atmospheric forcing (section 5.6).

5.4.1 Control run

We start our investigation of the AMOC in the VerOS model with a 20 thousand year long control run (which took roughly two weeks to compute) at our desired freshwater forcing value. In the left panel of fig. 5.6, we show the time series and histogram of the yearly and 5-year average AMOC strength. The first will be important later, when adding atmospheric noise to the system will allow us to work with shorter timescales (see section 5.6), while the second is more relevant to the slower responding ocean-only system.

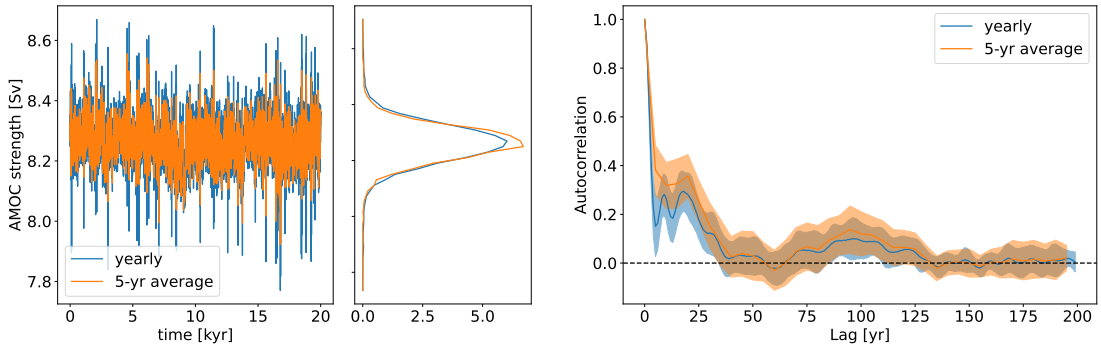


Figure 5.6: Time series, histogram and autocorrelation of the yearly (blue) and 5 year (orange) average AMOC strength, for a 20-thousand-year-long control run. For the last plot, the time series is split into 10 segments of 2000 years, autocorrelation is computed for each of them, and we show the mean with shading corresponding to one standard deviation.

Both the yearly and 5-year average sit at an equilibrium value of 8.26 Sv, and the 5-year average has a standard deviation of 0.06 Sv, just slightly smaller than the one of the yearly average (0.07 Sv), hinting at the fact that the main timescales of the system are larger than 5 years (see fig. 5.15 for the power spectrum of the control run). Indeed, the right panel shows the autocorrelation of the AMOC strength, and there is significant correlation for a lag time up to 25 years, with a second peak at the centennial timescales. This points to the presence of multiple timescales in the system, which, as we will see, can cause problems when running the rare event algorithm.

5.4.2 Making clones diverge

The VerOS model is, per se, fully deterministic. However, when we run the rare event algorithm, we want clones of the same trajectory to branch off. This is achieved by, at the time of cloning, adding a small noise perturbation to both the temperature T and salinity S fields. Understanding how this perturbation impacts the trajectories is an important and complementary information to the control run, which will allow us to make informed choices on the hyperparameters of the rare event algorithm.

More precisely, the perturbation is defined as

$$\begin{aligned} T &\mapsto T + T\epsilon_T \eta \\ S &\mapsto S + \epsilon_S \eta, \end{aligned} \tag{5.17}$$

where η is sampled from a normal distribution with mean 0 and variance 1, independently for each field and grid point. The noise amplitudes are chosen to be $\epsilon_T = 0.001$ and $\epsilon_S = 0.002\,\mathrm{g\,kg^{-1}}$. Typical salinity values are around $34.5\,\mathrm{g\,kg^{-1}}$ with fluctuations of the order of $0.2\,\mathrm{g\,kg^{-1}}$, while temperature values have a mean of 5 °C with a standard deviation of the order of 10 °C, though the temperature distribution is strongly skewed due to the cold deep ocean. Since temperatures in the model are expressed in degrees Celsius, the choice of multiplicative noise was a simple way to have bigger perturbations close to the surface and negligible ones in the deep ocean, which is more realistic than a uniform noise.

The effect that this perturbation has on the branching of the clones is shown in fig. 5.7, where we follow 50 clones after the perturbation is applied at time 0. We can see that already after only 5 years the standard deviation of the ensemble is around 30% of the standard deviation of the control run. However, after this first jump, the standard deviation increases much slower, appearing to saturate at around 60% of the standard deviation of the control run. This suggests that longer timescales (centennial) are needed to fully separate the trajectories.

5.4.3 Choosing the hyperparameters of the rare event algorithm

As score function for the rare event algorithm, we will simply take the 5-year average of the AMOC strength, changed of sign to push the system toward weaker states.

$$V(\{X\}, t) = -\frac{1}{5\,\mathrm{yr}} \int_{t-5\,\mathrm{yr}}^{t} \mathrm{AMOC}\left(X(t')\right) \mathrm{d}t'. \tag{5.18}$$

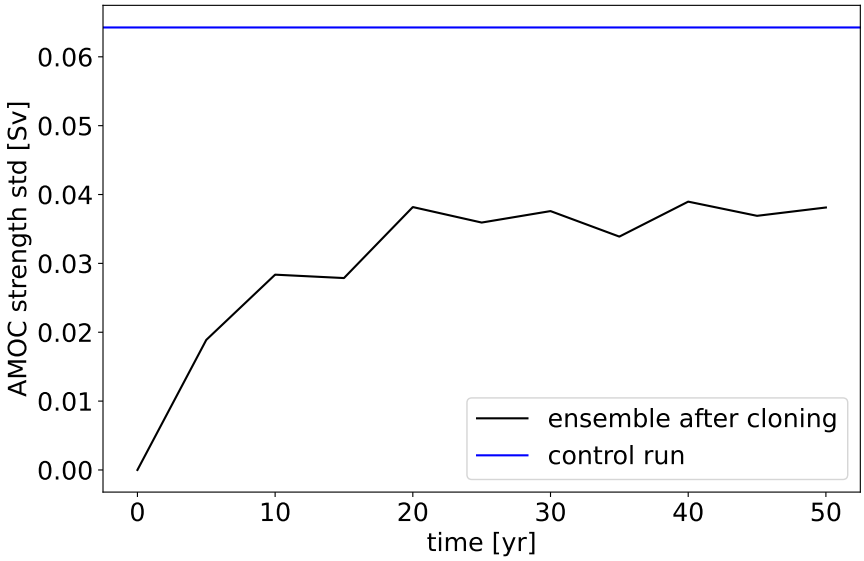


Figure 5.7: Evolution of the standard deviation of the 5-year average of the AMOC strength in an ensemble of 50 clones, after the perturbation explained in section 5.4.2 is applied at time 0. In blue the standard deviation of the 5-year average of the AMOC strength in the control run.

Secondly, to choose the resampling time of the rare event algorithm, we can rely on the autocorrelation plot in fig. 5.6 and the clone branching in fig. 5.7. When we did our benchmark on the Ornstein-Uhlenbeck process, the resampling time $\tau = 0.1$ that we used, corresponded to an autocorrelation of 0.9. To have something similar for the Versatile Ocean Simulator model, will require us to use a very short resampling time, below 5 years, which will complicate matters with our choice of the score function. Still, at $\tau = 5\,\mathrm{yr}$ the autocorrelation is around 0.4 and the clones are already starting to branch off. We will try later to use $\tau = 5\,\mathrm{yr}$, but another option is to leverage correlations at the centennial timescale. To do so we'll use $\tau = 50\,\mathrm{yr}$, which gives plenty of time for clones to separate (fig. 5.7).

Finally, in fig. 5.8, we show the suggested values of k to push the system one standard deviation from the mean (as given by eq. (5.11)), as a function of the total time for which we want to run the rare event algorithm. These values don't depend much on the total integration time, and since already eq. (5.11) is meant to give only an order of magnitude for selection strength, we can conclude that k should be of order $10\,\mathrm{Sv}^{-1}$.

Now that we have an idea of good candidates for resampling time and selection strength, we can try to actually run the rare event algorithm. As pointed out in the header of this chapter, I wasn't able to observe any true collapse of the AMOC. We expected it to be a hard task, since extremely long control runs didn't show any collapse, but not impossible, as there were temporary abrupt weakenings (see blue arrows in fig. 5.2). In order to maximize our chances, we chose to run as many experiments as possible, exploring the hyperparameter space, which meant using a relatively small ensemble size of $N = 50$.

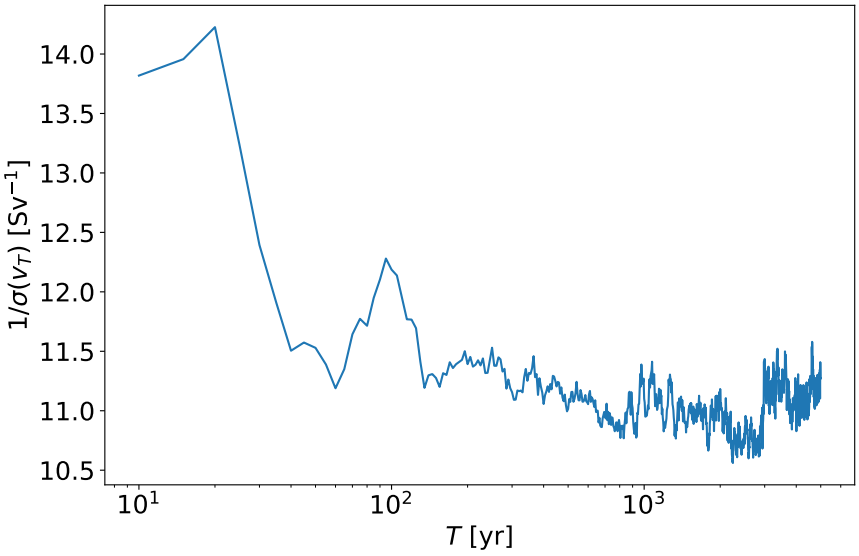


Figure 5.8: Values of the suggested k given from eq. (5.11) to push the system 1 standard deviation from the mean as a function of the total time T for which the rare event algorithm is run. Computed from the 5-year average of the control run in fig. 5.6.

5.4.4 The rare event algorithm can excite temporary weakenings of the AMOC

The temporary weakenings of the AMOC present in the very long runs of fig. 5.2 occur very sporadically, making them indeed *rare* events. Moreover, with the score function we use, biasing according to the value of the 5-year averaged AMOC strength, we are likely to sample them, and the hope is that they will be gateways to a more substantial collapse.

As we show in fig. 5.9, we were indeed were able to excite two of these temporary weakenings, which are triggered when the system is already roughly 3 standard deviations from the mean. Unfortunately, these fluctuations *are* temporary, and the system recovers, albeit not completely, despite the rare event algorithm constantly 'pushing' downward. Another issue is that when such an abrupt fluctuation happens, almost all the ensemble members that didn't jump downward are killed. Such decimation drastically reduces the variety of the ensemble, and may lead to the loss of potentially promising members that could have had a more substantial collapse later. In fact, I think this is one of the main issues that prevented the ensemble to make consistent progress, and, in hindsight, it could have been partially mitigated by using larger ensembles.

5.4.5 Discovery of a new stable attractor

As an interesting aside, we tried to run the ensemble after one of these temporary fluctuations, without the rare event algorithm. As shown in fig. 5.10, the ensemble did recover from the abrupt fluctuation, but didn't climb back up to the initial attractor from which we started the simulation. This suggests the presence of a new branch of attractors,

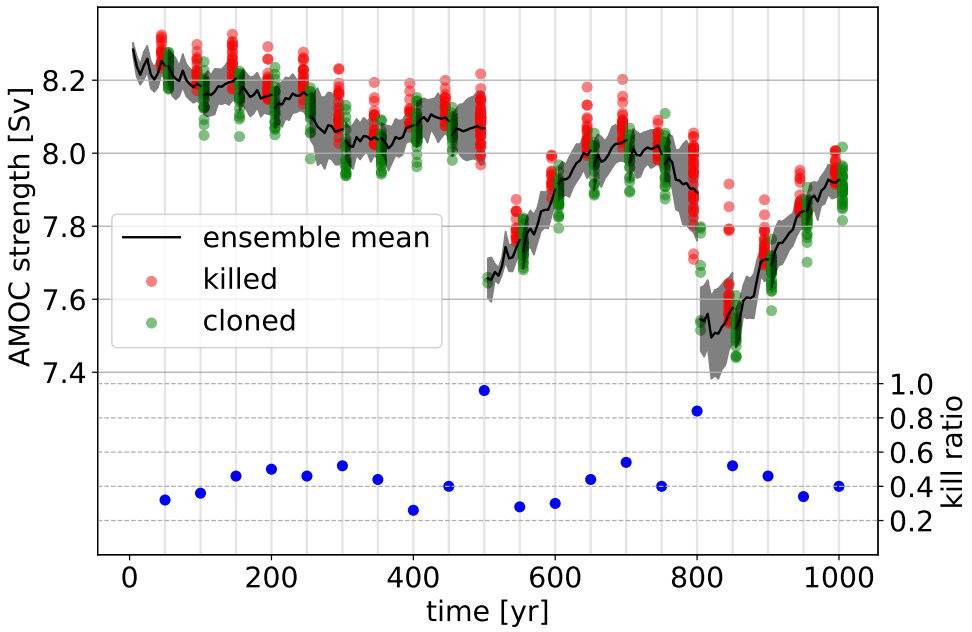


Figure 5.9: Evolution of the ensemble (5-year average AMOC strength) during a rare event algorithm run with $k = 20\,\mathrm{Sv}^{-1}$, featuring two abrupt temporary fluctuations. The black line (left y-axis) represents the ensemble mean and the shaded area corresponds to one standard deviation. Each vertical line is at a resampling step (every 50 years). In red are the ensemble members killed and in green the ones that survived to the next iteration. The blue points on the bottom (right y-axis) show the fraction of ensemble members killed at each resampling step. At $t = 500\,\mathrm{yr}$ and $t = 800\,\mathrm{yr}$, only 2 and 8 out of the 50 ensemble members survive, resulting in the collapse of the ensemble.

with AMOC strength around 8 Sv, in the already very complex stability landscape of fig. 5.2. Moreover, it shows that rare event algorithms are a viable tool to jump between attractors, which is what we wanted to do in the first place, and achieved, albeit on a much more humble scale.

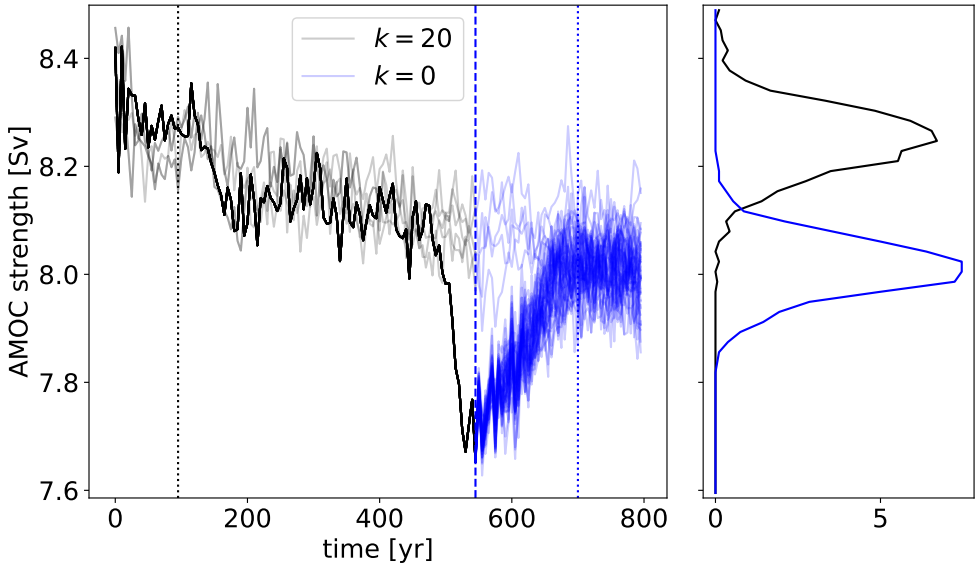


Figure 5.10: Discovery of a new stable attractor with the rare event algorithm. On the left, in black, backward reconstructed trajectories of the 5-year average AMOC strength after 10 iteration of the REA with selection strength $k = 20\,\mathrm{Sv}^{-1}$. The first resampling step happens after 100 years (dotted line), then every $\tau = 50\,\mathrm{yr}$. The run terminates with an ensemble collapse similar to the ones in fig. 5.9, with the darkest trajectory consisting of 45 identical clones. In blue, the 50 ensemble members are propagated forward for 250 years, with no rare event algorithm. All ensemble members recover from the temporary fluctuation in 150 years, and then stabilize around a different value of the AMOC strength. On the right, histograms of the distribution of the 5-year average AMOC strength during the first 100 years (in black, before the dotted line in the left panel) representing the original attractor, and during the last 100 years (in blue, after the dotted line in the left panel) representing the new attractor. The black histogram is computed on the ensemble before the first resampling step, i.e. not only on the visible black trajectories.

So far we were able to weaken the AMOC in the VerOS model, but not by much and over a very long time. This can hardly be called an *abrupt* collapse, even if after millennia of fluctuations and incomplete relaxations we eventually reach our target attractor at 3 Sv. The most promising option to push for a quicker collapse is thus by reducing the resampling time of the rare event algorithm.

5.4.6 Shorter resampling time isn't viable in the deterministic case

When we chose the resampling time to be $\tau = 50\,\mathrm{yr}$, we argued that it was necessary to ensure a good separation between the clones. However, the analysis in section 5.4.2 suggests

that we should be fine also with $\tau = 5\,\mathrm{yr}$. Unfortunately, the results were disappointing.

In fig. 5.11 we show a rare event algorithm run with resampling time $\tau = 5\,\mathrm{yr}$ and selection strength $k = 16\,\mathrm{Sv}^{-1}$. After a pleasing decrease in the first 400 years, the ensemble reaches a plateau around 8 Sv, where we now know there is a new attractor (fig. 5.10). What is interesting, is that the estimated probabilities don't decrease, but rather explode reaching the very nonphysical values of $p_{1/2} = 4$. By looking carefully at the plot, we can see that many abrupt surges of $p_{1/2}$ occur shortly after sudden drops in the minimum of the ensemble (for example at $t = 370\,\mathrm{yr}$ or $t = 790\,\mathrm{yr}$). The explanation is that when there is a drop in the minimum, most of the weight of the ensemble goes to those members with the lowest AMOC strength, as we have seen happening in fig. 5.9. As a consequence, the members that didn't fluctuate acquire a very small expected number of clones $w_j^{(i)} < 1$, which gives them, if they survive, a potentially very big unbiasing factor $\pi_j^{(i)}$. This is not a problem when these lucky survivors are a minority in the ensemble. However, since the resampling time is short and clones don't branch out enough, eventually the members that performed that fluctuation recover from it and are all killed. Which means that now the lucky survivors become *all* the ensemble, thus causing a delayed upward spike in the probabilities. This mechanism is very similar to the one that caused the runs on Ornstein-Uhlenbeck process to reach a plateau (fig. 5.5), but here it is exacerbated by abrupt fluctuations and clones not branching out sufficiently.

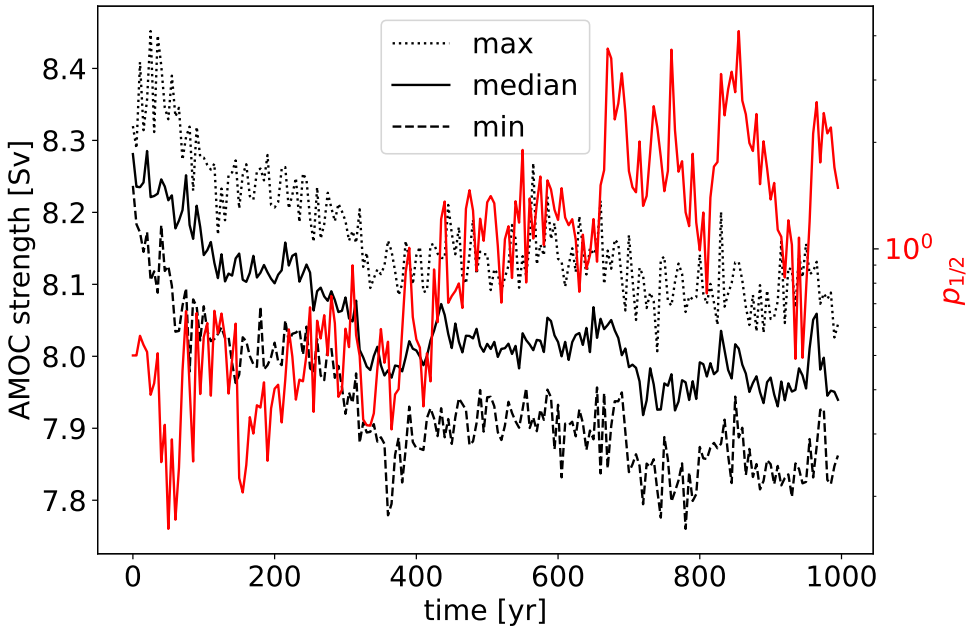


Figure 5.11: Evolution of the 5-year average of the AMOC strength during a rare event algorithm run with selection strength $k = 16\,\mathrm{Sv}^{-1}$ and resampling time $\tau = 5\,\mathrm{yr}$. In black (left y-axis) the minimum, median and maximum values of the ensemble before each resampling step. In red (right y-axis) the unbiased probability $p_{1/2}$ of being below the median, as given by eq. (5.16). After an initial decrease in AMOC strength, the ensemble plateaus around 8 Sv and $p_{1/2}$ reaches nonphysical values above 1, showing that the algorithm is not working well.

To move past this problem, we should force clones to separate faster, and, while we could simply try to crank up the noise amplitudes in eq. (5.17), there is a much more interesting and physically sensible solution. Namely, to give the VerOS model an atmosphere.

5.5 A more realistic atmospheric noise product for VerOS

In the previous section, we were not able to tip the AMOC in the VerOS model. One possible explanation is that the slow dynamics of the system prevented us to use short enough resampling times. Another viable one is that a noise-induced tipping event is just too rare, and will probably require a much larger ensemble to avoid collapsing to a single trajectory when the selection strength k is high. However, this will make the simulations significantly more expensive to run.

Both problems are important, and a possible solution is to increase the variance of the AMOC and the speed of the dynamics by adding atmospheric noise. This should allow us to reach the tipping point more easily.

To do so, the ideal option would be to couple the VerOS model to a dynamic atmospheric component. However, a much simpler option is to compute an atmospheric noise model offline and later add it to VerOS as a stochastic component. During my Secondment in Utrecht, we opted for the latter, and I collaborated with Alfred Hansen from the University of Copenhagen, following the work he performed in his master thesis [Hansen, 2023] to extract a noise model from the Berkeley Earth [`https://berkeleyearth.org/data/`; Rohde and Hausfather, 2020] sea surface temperature (SST) record and insert it into the VerOS model.

In practice, we will derive a time dependent grid-point-wise sea surface temperature noise ϵ_i^T that we insert into the equation for the heat flux (eq. (5.1)), by substituting T_i^{obs} with $\tilde{T}_i^{\text{obs}} = T_i^{\text{obs}} + \sigma\epsilon_i^T$. Here we added the scalar parameter σ, that allows us to artificially rescale the noise amplitude with respect to the one derived from the data ($\sigma = 1$). Setting $\sigma = 0$ will suppress the noise and give the results discussed in the previous section.

5.5.1 Derivation of the noise model

The Berkeley Earth data has a monthly resolution and spans from 1870 to present day. After interpolating it onto the grid used by VerOS, we mask it to be zero outside the Atlantic Ocean. I tried also to have noise over all oceans, as Hansen did in his thesis, but it didn't make much of a difference, and since we are focusing on the tipping of the AMOC, it seemed better to keep the noise only over the Atlantic Ocean, also with the foresight of looking for drivers of the tipping point if we managed to actually observe one.

After masking, we want to remove the seasonal cycle and global warming trend. To do so, we subtract from the data, for each grid point independently, the 10-year month-wise running mean. The choice of 10 years comes from the fact that atmospheric phenomena are at most decadal [Williams et al., 2017], and anything with a slower dynamics likely comes from the ocean. Since we want an *atmospheric* noise, we don't want modes of the real

ocean to be included. Also, we removed the seasonal cycle because it is already accounted in the climatological T^{obs} derived from ERA-40 (section 5.2.1).

The detrended data is then decomposed in Empirical Orthogonal Functions (EOFs) [Hannachi et al., 2007], i.e. the spatial modes $e^{(j)}$ that diagonalize the covariance matrix of the SST field. Then, for each month n (at time t_n), we can write

$$SST_i(t_n) = \sum_{j=1}^{J} \tilde{c}_n^{(j)} e_i^{(j)}. \tag{5.19}$$

To simplify matters, we keep only the first 32 modes, which, combined, account for 90% of the total variance. The first 8 are shown in fig. 5.12. In particular, we can see that the first two modes are strong temperature oscillations at the deep water formation sites in the Baffin Bay and in the Denmark Strait [Rahmstorf, 2002]. Mode 3 and 6 capture the phenomenon of the Atlantic Niño [Lübbecke et al., 2018], while 4 and 5 correspond to the region of the subpolar gyre. Mode 8 is related to the East Atlantic Pattern, which is especially active during summer [Gastineau and Frankignoul, 2015].

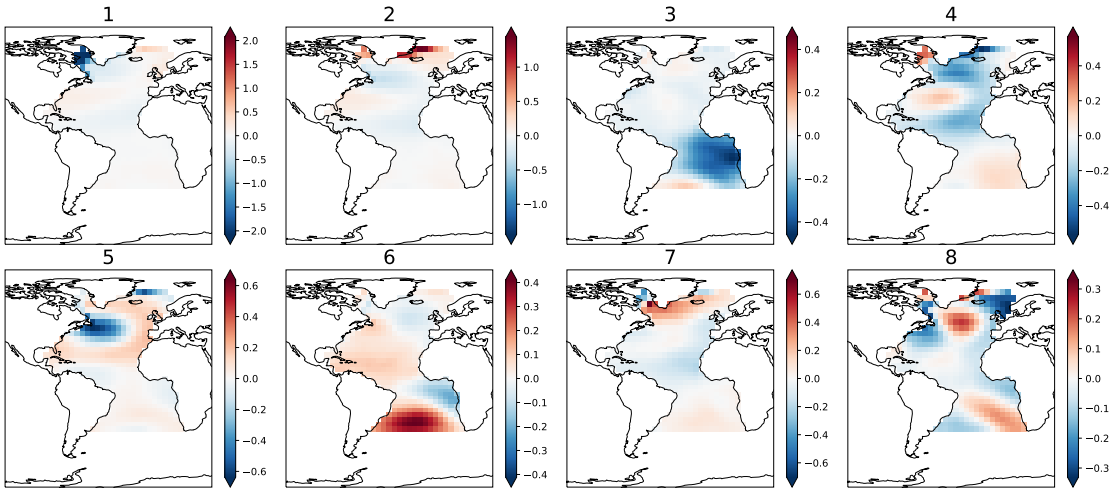


Figure 5.12: The first 8 EOFs of SST anomaly. The Atlantic Ocean extends in latitude from the tip of Africa at 35°S to the northern edge of the model domain at 80°N. The main phenomena are captured, like fluctuations in the Baffin Bay, the Atlantic Niño, and the subpolar gyre.

In the decomposition of eq. (5.19), we choose to save the information about the explained variance in the norm of the EOFs $e^{(j)}$, so that $\tilde{c}_n^{(j)}$ have zero mean and unitary variance. Subsequently, $\tilde{c}_n^{(j)}$ is modeled as an autoregressive process of rank $R^{(j)}$:

$$\tilde{c}_n^{(j)} \approx c_n^{(j)} = \sigma^{(j)} \xi_n^{(j)} + \sum_{r=1}^{R^{(j)}} \rho_r^{(j)} c_{n-r}^{(j)}, \tag{5.20}$$

where $\xi_n^{(j)}$ is white noise.

The maximum lag $R^{(j)}$ is computed following the standard procedure by Box and Jenkins [Box et al., 2015]. This involves computing the partial autocorrelation function

$PACF(l)$, which is the standard autocorrelation function at lag l from which the effect of all lags $l' < l$ has been removed. For a perfect autoregressive process of rank R, $PACF(l)$ is equal to 0 $\forall l > R$, so we determine $R^{(j)}$ as the maximum lag l at which the PACF is significantly different from 0 with a significance p-value of $\alpha = 10^{-4}$.

Subsequently, the autoregressive coefficients $\sigma^{(j)}$ and $\rho_r^{(j)}$ are fitted from the time series of $\tilde{c}_n^{(j)}$ with the Yule-Walker equations [Yule, 1927; Walker, 1931]. The values of these coefficients are shown in fig. 5.12, where we can clearly see regular spikes every 12 months, which ensure that there is no seasonal cycle (see right panel of fig. 5.14).

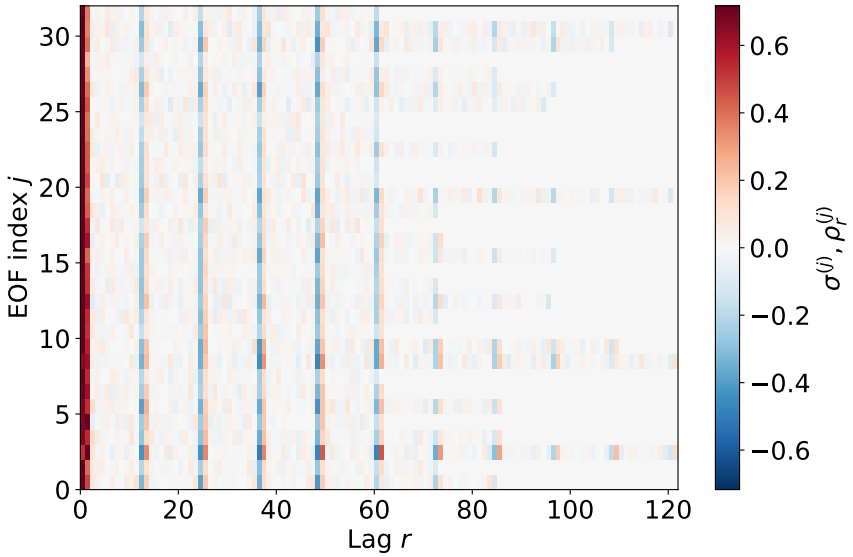


Figure 5.13: Yule-Walker coefficients for the 32 EOFs of SST anomaly: $\rho_r^{(j)}$ for $r = 1, \ldots, 122$ and $\sigma^{(j)}$, plotted at $r = 0$. There are clear spikes every 12 months, which ensure that there is no seasonal cycle (see right panel of fig. 5.14)

When we want to run the noise model in VerOS, we initialize it setting $c_n^{(j)} = 0 \, \forall n < 0$. Then, at the beginning of every month $n \geq 0$, we compute the next coefficients $c_{n+1}^{(j)}$ according to eq. (5.20), and combine them into a map of sea surface temperature:

$$\epsilon_i^T(t_{n+1}) = \sum_{j=1}^{32} c_{n+1}^{(j)} e_i^{(j)}. \tag{5.21}$$

Finally, for each time step t inside month n, the actual noise map $\epsilon_i^T(t)$ is computed as the linear interpolation between $\epsilon_i^T(t_n)$ and $\epsilon_i^T(t_{n+1})$.

With the initialization described above, the noise model will have a warm-up time before all the lag terms actually have an effect. Given that the maximum lag $R^{(j)}$ is 121, the warm-up time is just over 10 years. Unfortunately, I noticed only at the stage of redacting this manuscript, that the implementation by Alfred Hansen of the addition of the noise model into VerOS was meant for long runs. On the other hand, when running the rare event algorithm, said implementation yields a re-initialization of the noise at every resampling step. For runs with a short resampling time, this means that the noise will

be always in its warm-up phase. For this reason, in fig. 5.14, we show the power spectra of each EOF component during the first year of warm-up period and once steady state is achieved. Both spectra show a clear peak at the decadal timescale, though a few EOFs have higher main frequencies, for example 4 years for $j = 3$, which is the Atlantic Niño. The decadal peak is much broader in the warm-up phase, and there are no band gaps at the harmonics of the yearly cycle. However, considering the results shown in the following sections I think this effective bug in the noise model is just a minor detail.

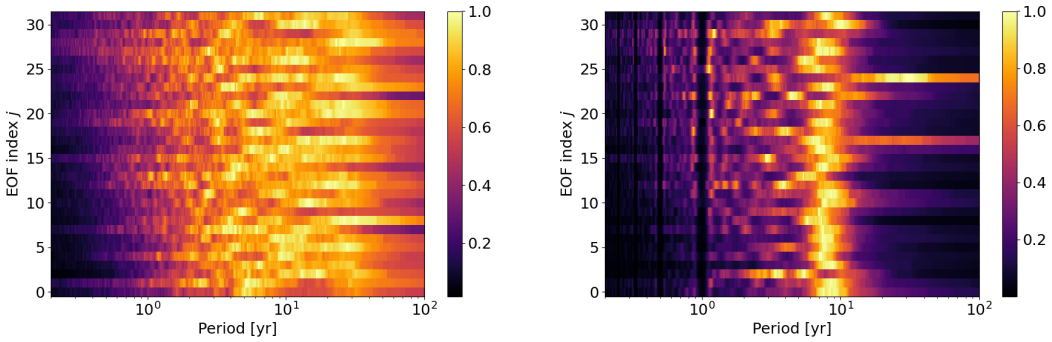


Figure 5.14: Normalized power spectra for each EOF, during the first year of warm-up (left) and in steady state (right). For visualization purposes the spectra have been smoothed with a uniform kernel of width $0.06\,\mathrm{yr}^{-1}$. The spectra are computed from an 800-year simulation, for the left plot re-initializing the system every year, for the right plot every 50 years. This choice is to have a parallel with the resampling times used for the rare event algorithm. Both spectra show a clear decadal peak. For the steady state one (right), the peak is sharper and there are clear troughs at the 1-year frequency and its harmonics. These troughs are the effect of having removed the seasonal cycle.

5.6 Rare events in the stochastic model

Now that we have a proper stochastic parameterization of the atmosphere inserted into VerOS, we can proceed to run the rare event algorithm. However, first, it is worth investigating the response of the system to the added noise.

In fig. 5.15, we compare the time series and power spectra of the 1-year averaged AMOC strength with and without noise. The variance of AMOC strength fluctuation increases dramatically, getting closer to the one displayed by properly coupled ocean-atmosphere models [Cini et al., 2024]. The oscillations occur at the decadal timescale, and this is not surprising, considering the power spectra of the noise itself (fig. 5.14). In different experiments, the precise period of these oscillations varied between 8 and 11 years, and, as can be seen from the power spectra, it is the main mode of variability, while the centennial peak disappears. Interestingly, this result is quite the opposite to the one Hansen finds when using an all-oceans noise [Hansen, 2023]. In his case, the slower modes are amplified more than the decadal ones. Moreover, he works at a higher value of freshwater forcing, and in his case, the decadal oscillation is already very dominant in the model without noise.

This suggests that, in our case, there may be resonance between the deterministic model and the noise at the decadal timescale, that selectively amplifies it. On the other hand, for the all-oceans noise and higher freshwater forcing used by Hansen this is not the case.

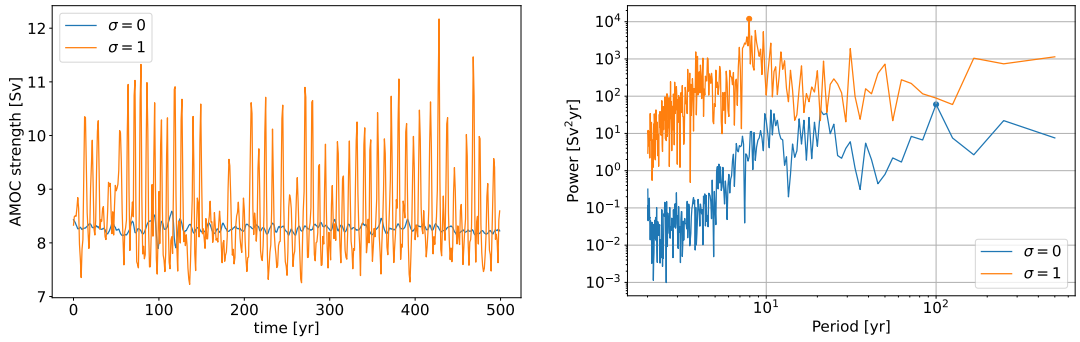


Figure 5.15: Time series (left) and power spectra (right) of the yearly averaged AMOC strength for 500-year long control runs without ($\sigma = 0$) and with ($\sigma = 1$) sea surface temperature noise. While the noise-free run is dominated by a centennial timescale, with secondary peaks at 10 and 25 years, the run with noise exhibits a very strong 8-year oscillation, while showing no centennial activity.

5.6.1 Non-trivial response to the noise amplitude

If the appearance of strong decadal oscillations in the time series of the AMOC strength was expected from a simple linear response theory point of view, the response of the system to the changes in the noise amplitude is something far from linear.

In fig. 5.16, we show a 500-year portion of control runs at different values of the noise amplitude σ. As expected, the variance of the AMOC strength fluctuation increases with σ. However, already for $\sigma = 2$ the system starts to move towards states that, on average, have a more vigorous AMOC. For $\sigma = 5$, the system oscillates around a completely different attractor, one with an average AMOC strength around 11 Sv. Note that there is no such attractor in the noise-free stability landscape in fig. 5.2. This suggests that only for small noise amplitudes can the noise be thought as just a small perturbation on top of the deterministic system. For more substantial amounts of noise, the whole stability landscape changes. In fact, we cannot rule out the possibility that the landscape did change already for $\sigma = 1$, with, for example, some deterministic attractors becoming unstable under the influence of the noise.

Now that we have more knowledge about the effect of noise, we can finally try to apply the rare event algorithm. Unfortunately, as was pointed out already at the beginning of this chapter, we still weren't able to observe an abrupt collapse of the AMOC.

5.6.2 The rare event algorithm is still ineffective

One of the reasons for adding noise to the system, was to be able to run the rare event algorithm with shorter resampling time. So, we run the algorithm with $\tau = 1\,\text{yr}$ and as

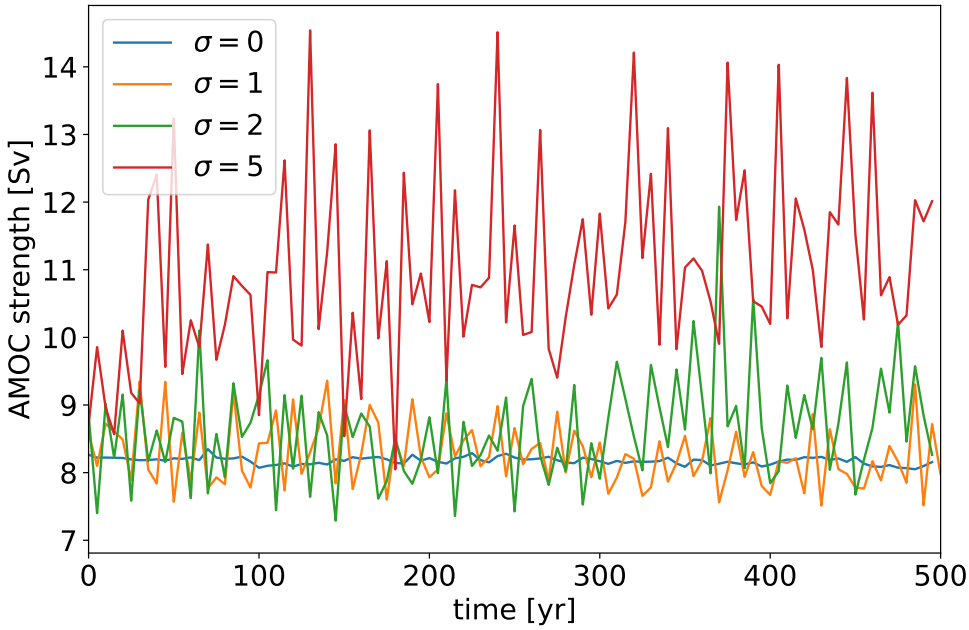


Figure 5.16: 5-year averaged AMOC strength for 500-year long control runs at different values of noise amplitude σ. $\sigma = 0$ is the run without noise, $\sigma = 1$ uses the noise product extracted from the Berkeley Earth dataset and higher values of σ artificially and uniformly rescale this noise product. For small noise amplitudes, the system fluctuates with increased variance around the noise free attractor. For larger noise amplitudes, the system shifts to stronger AMOC strengths.

well shift the observable and the score function from the 5-year mean to the 1-year mean.

In fig. 5.17, we show an example of a run with selection strength $k = 20\,\mathrm{Sv}^{-1}$. Compared to what we saw in fig. 5.11, now the algorithm is indeed sampling extremely rare events, with estimated probabilities as low as 10^{-72}. However, these probabilities are not very meaningful, as the ensemble is not really progressing toward a collapsed state. In fact, the strong 10-year oscillations interfere with the algorithm. Namely, during the downward slope the algorithm acts amplifying the oscillations, selecting members for which the AMOC strength decreases faster, while on the upward slope the algorithm has a dampening effect. Since the algorithm is constantly pushing downward and selecting trajectories, the probabilities get smaller and smaller, but by the end of a 10-year oscillation, the system loses memory and thus no qualitative long term progress is observed.

A possible solution could be to take longer resampling times and averages of the AMOC strength, to smoothen out these oscillations. But this would bring us back to the problems pointed out for the deterministic system. In any case, we did try longer resampling times (10, 20, 50 years), but the results were disappointing in all cases.

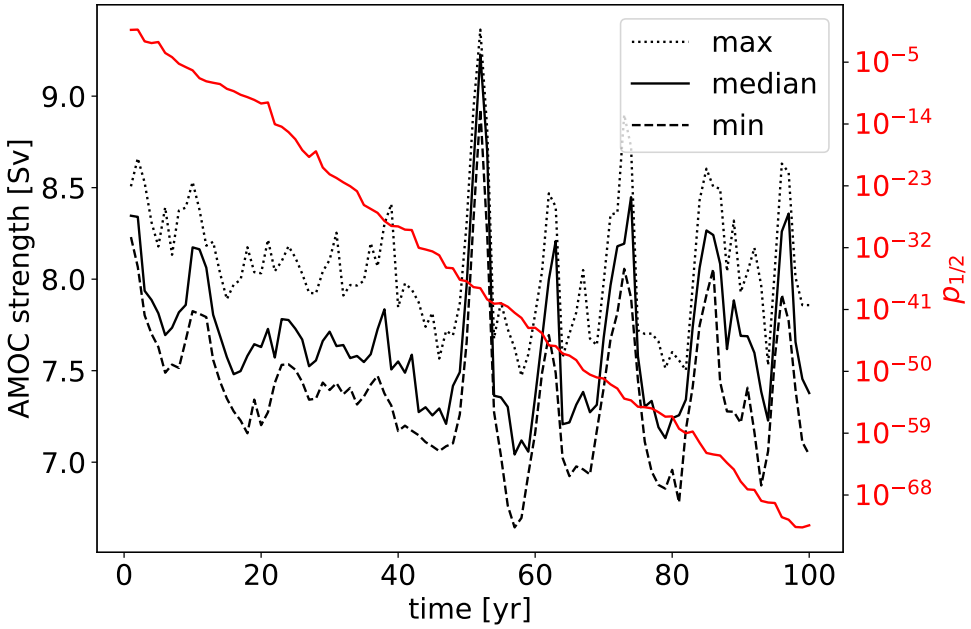


Figure 5.17: Rare event algorithm run on the VerOS model with added atmospheric noise. Resampling time $\tau = 1\,\mathrm{yr}$ and selection strength $k = 20\,\mathrm{Sv}^{-1}$. In black (left y-axis) the minimum, median and maximum values of the 1-year averaged AMOC strength before each resampling step. In red (right y-axis) the unbiased probability $p_{1/2}$ of being below the median. Despite sampling extremely rare events, the ensemble doesn't move significantly toward a collapsed state.

5.7 Discussion

The aim of this project was to use rare event algorithms to observe a noise induced tipping of the AMOC. Unfortunately, it was not possible to do so. Nevertheless, we were able to improve our understanding of the AMOC in the VerOS model, by discovering a new branch of stable attractors and characterizing the response of the deterministic system to sea surface temperature noise.

It is interesting, however, to ponder why we couldn't observe any collapse, and in particular to compare to the work of Cini et al. [Cini et al., 2024]. First, in Cini et al. [2024] the ocean is simulated with the Large Scale Geostrophic Ocean (LSG) model, and coupled to the Planet Simulator (PlaSim) [Fraedrich et al., 2005a] to represent a realistic atmosphere, with all its relevant climate variables. Most importantly, Cini et al. [2024] identifies wind stress anomalies in the North Atlantic as the main trigger of spontaneous AMOC collapse. In our case, instead, in the deterministic VerOS model wind stress is prescribed and when we added atmospheric noise, it was only through sea surface temperature.

Since there are very few studies on the noise-induced collapse of the AMOC, the mechanism of tipping are not well understood. It then becomes very interesting to have studies with different types of models and settings. For example, the fact that without wind stress noise I could not observe any collapse of the AMOC strengthens even more the conclusions made of Cini et al. [2024].

Another key point that emerges from Cini et al. [2024] and the comparison with our work, is the need of a short resampling time, of the order of 1 year. This is critical to avoid losing memory of the progress made in previous iterations when the fast evolving atmosphere is involved. In this work, resampling every year wasn't possible in the deterministic model, as clones needed more time to properly separate. However, it was possibly also not extremely relevant, as we have seen that, even with the addition of atmospheric noise the main timescales are of the order of 10 years.

Adding atmospheric noise also caused the whole stability landscape to shift. As we saw, artificially increasing the amplitude of the sea surface temperature noise led the AMOC to reinvigorate. So it is possible that this particular noise made a spontaneous collapse even more unlikely than in the deterministic system.

Finally, another obstacle was clear from the beginning, namely the fact that the stability landscape has too many attractors. Indeed, different attractors with similar values of the AMOC strength, can have quite different circulation patterns [Lohmann et al., 2024], and thus can trap the system in a local minimum that is not along the path towards the fully collapsed state.

This last point may call for the necessity to use a less trivial score function for the rare event algorithm. Namely, not simply the AMOC strength itself, but rather a more tailored transition coordinate, potentially related to the specifics of the circulation patterns.

5.8 Conclusions

In this work I have successfully implemented the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) rare event algorithm, and coupled it to the Versatile Ocean Simulator. Though this is not strictly novel work, as implementations of the GKLT algorithm are already available, I argue that my code is more flexible and better commented and documented. In this regard, a future direction could be the integration into the `stochrare`(`https://github.com/cbherbert/stochrare`) repository by Corentin Herbert and Thibault Lestang, which has a broader scope concerning rare events, but lacks the possibility to work on complex climate models and on high performance computing clusters.

The use case of trying to tip the AMOC in the VerOS model yielded negative results. However, we were able to find suitable explanations. Both technical in the nature of the noise, and physical in the presence of too many intermediate attractors and in the shift of the deterministic stability landscape upon forcing with atmospheric noise. There were some interesting intermediate results, like the discovery of a new branch of attractors. In any case, these partial results, as well as all the problems and limitations I encountered and documented, will prove very useful for other people continuing in this research direction.

As a personal note, this work allowed me to gain hands-on experience with rare event algorithms and climate models, from a theoretical, physical and technical point of view. This experience could prove, by itself and independently of the results, very valuable for my future work.

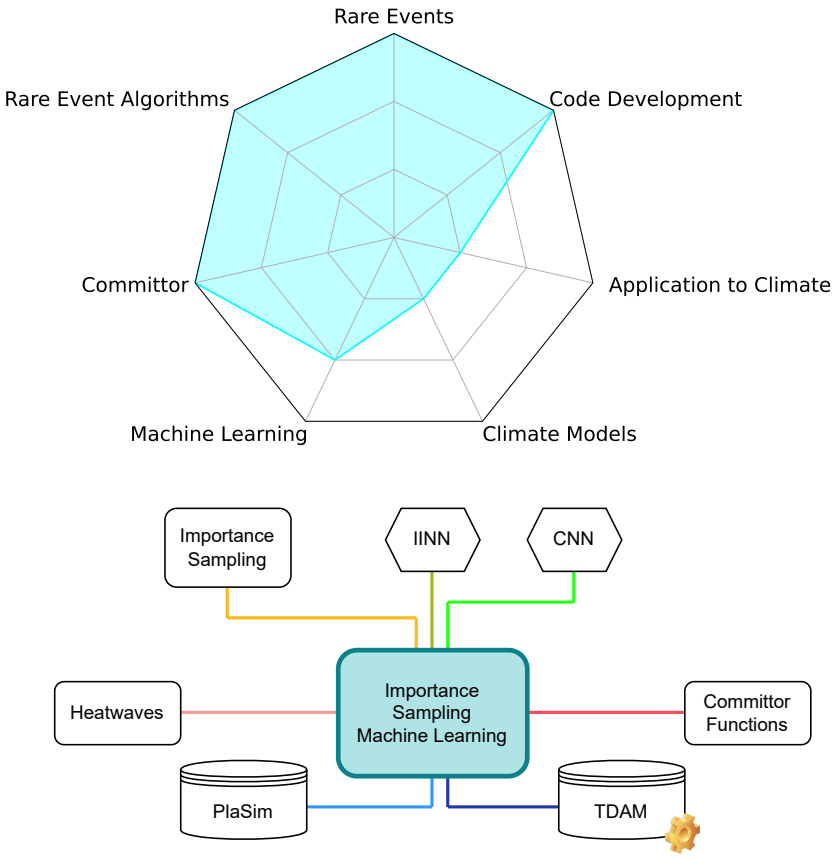
Finally, the objective of my thesis was the coupling of rare event algorithms with machine learning, so working on rare event algorithms on their own was a fundamental stepping stone.

5.9 Acknowledgements

This work was supervised, together with Freddy Bouchet and Corentin Herbert, by Peter Ditlevsen and Johannes Lohmann during my time in Copenhagen, and by Henk Dijkstra in Utrecht. I would like to thank the high performance computing centers in Copenhagen (HPC), Utrecht (Lorenz) and Lyon (PSMN) for the resources provided and their staff for the support with technical issues. Particular thanks go to Roman Nuterman from the University of Copenhagen, who helped me familiarize with working on clusters. I would also like to thank Alfred Hansen from the University of Copenhagen for the collaboration on extracting atmospheric noise from reanalysis data. Finally, heartfelt thanks to Francesco Ragone and Matteo Cini for the valuable discussions.

Chapter 6

Importance Sampling for Rare-Event-Oriented Parameter Estimation



The long term goal of this thesis was to develop a coupling between machine learning and rare event algorithms (see fig. 1.7). Ideally, the coupling goes both ways: machine learning provides a score function (the committor) to be used in a rare event algorithm, and rare event algorithms efficiently generate new data to improve the performance of machine learning. While the first direction is thoroughly understood, the second one is not as obvious as it might seem. Indeed, in Lucente et al. [2022b], the authors claim to have coupled machine learning and rare event algorithms, but they actually didn't investigate how the newly generated data improves the committor estimate.

The goal of this chapter is then to answer this question, and shed some light on this last missing link in the feedback loop between machine learning and rare event algorithms. More precisely, we will use importance sampling techniques to try to find the *optimal* way to generate new data, specifically in order to improve the committor estimated by a neural network.

I will start by presenting the results of my early studies on simulated resampling for Convolutional Neural Networks, with the goal to build an intuitive understanding of what we aim to achieve. Then, I will move to the discussion of the first results of the ongoing optimal resampling project. Contributors to the latter are Tony Lelièvre[1] and Julien Reygner[1] for the theoretical part, myself and Amaury Lancelin[2][3] for the implementation and testing, Clément Le Priol[2] and Valeria Mascolo[4] for the coupling with climate models and finally Corentin Herbert[4] and Freddy Bouchet[2] for the general direction of the project.

6.1 Introduction

As has been pointed out many times in this manuscript, extreme weather and climate events, and in particular heatwaves, have a very high impact on society and are thus worth studying. However, they are also rare events, which means we suffer severely from the issue of lack of data. We then want to mitigate this issue using rare event algorithms. Again, we know at this point that rare event algorithms require a score function and that the optimal one is related to the committor function [Chraibi et al., 2020; Rolland et al., 2016]. However, if the event we want to study lasts for very long, then we don't really need the committor function [Ragone et al., 2018; Ragone and Bouchet, 2021; Wouters and Bouchet, 2016].

Indeed, in Ragone and Bouchet [2021] the authors efficiently estimate the return times of whole-summer heatwaves with the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) algorithm, using as score function the heatwave amplitude itself. This is the same trick that we

[1] *Centre d'Enseignement et de Recherche en Mathématiques et Calcul Scientifique, École des Ponts ParisTech, 6 et 8, Avenue Blaise Pascal, Cité Descartes—Champs sur Marne, 77455 Marne la Vallée Cedex 2, France*

[2] *LMD/IPSL, CNRS, ENS, Université PSL, École Polytechnique, Institut Polytechnique de Paris, Sorbonne Université, Paris, France*

[3] *RTE France*

[4] *ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France*

used in chapter 5 to study the collapse of the Atlantic meridional overturning circulation (AMOC), and it is based on the fact that for long events, persistence is a good estimate of the committor. And indeed, the very extreme heatwaves simulated in Ragone and Bouchet [2021] do not feature absurd daily records, but rather temperatures that are consistently slightly above average.

However, this approach is not suited to the study of shorter heatwaves (from a few days to two weeks), which can still have significant impacts (see for instance the 2021 Canadian heatwave [Lin et al., 2022; Henderson et al., 2022], which lasted roughly a week). Indeed, persistence of temperature is useful only when we are very close or already inside the event, and for short heatwaves this gives us a window of maximum two weeks for meaningfully running the rare event algorithm. However, from chapter 5 we know that the GKLT algorithm can sample very extreme events if it can run a sufficient number of iterations. And as well we know that the resampling time should be of the order of the Lyapunov time of the system, as to give time to the cloned trajectories to branch out. Considering that in the atmosphere the Lyapunov time is of the order of a few days, this would severely limit the number of iterations that we can perform, and consequently how rare of an event we can sample. A confirmation of this problem was shown in Lestang [2018], where the author tried, and failed, to use a persistence based score function to sample short events of extreme drag on an object immersed in a turbulent flow.

That is why, for the study of short heatwaves, we really need a better approximation of the committor function, for instance the one obtained using neural networks, which can give significant predictability even a month before the event happens (see chapter 3 and [Miloshevich et al., 2023a]). However, now we are back to the problem of lack of data, which means our estimate of the committor from the available data might not be very good. And we just ruled out using the GKLT algorithm to generate new events. Moreover, this type of algorithm is good at estimating global properties such as return times, and though we could still use the newly generated data to train a neural network, it is not designed for estimating the state dependent committor function $q(x)$.

What we need is a different kind of rare event algorithm, one with the explicit goal of generating new data to optimally improve our estimate of the committor. We then resort to the class of importance sampling algorithms, which are more suited to our problem as, contrary to genealogical algorithms like GKLT, don't require the system to steadily advance in time, and thus are not limited by the length of the event.

Indeed, importance sampling algorithms are the main go-to for the study of rare events, collecting great success in many fields of science, from signal processing [Bugallo et al., 2015], to chemistry [Geissler and Chandler, 2000], to condensed matter physics [Paananen et al., 2021] and also in the climate community [Annan and Hargreaves, 2010]. The simple concept of tilting the stationary measure to focus the computational power in specific areas of the phase space has sprouted many variants [Tokdar and Kass, 2010], with the recent emergence of Adaptive Importance Sampling methods which iteratively improve the tilting based on the generated data. This is essentially what we are after.

However, many of these methods are efficient for estimating low-dimensional objects, like moments of the distribution of an observable (0-dimensional objects) or maybe the whole density of states of a physical system [Paananen et al., 2021] (still a 1-dimensional object). What we are after is the committor $q(x)$, which has the same dimensionality as the phase space! Recently, importance sampling techniques have been developed specifically to estimate committor functions, often with the help of machine learning, which is very close to what we are after. However, such methods either work only for low-dimensional systems [Li et al., 2019], or they actively modify the dynamics [Lin and Ren, 2024; Evans et al., 2022], or resort to variational principles that require access to the equations of the system [Rotskoff et al., 2022; Kang et al., 2024]. None of these options is relevant for climate models, where the only thing we can do is selectively choose initial conditions and let the system evolve, unaltered, as a black box. We will then need to develop a custom algorithm.

In the following of this chapter we will first provide some empirical evidence that there is an optimal way to add new data to improve the machine learning estimate of the committor function, specifically for extreme heatwaves (section 6.2). We will then approach the problem from a more fundamental point of view, developing an optimal resampling algorithm (section 6.3) and testing it on a toy model (section 6.5).

6.2 Empirical evidence of skill improvement from optimal new data injection

On the road towards developing an efficient resampling method to improve our estimate of extreme climate events, it is important to have intermediate steps to facilitate understanding (and debugging). Indeed, running climate models is not only expensive, but also rather tedious and time-consuming, so we want to resort to using them only when we are confident that the pipeline will work properly. There are then two options: the first is to use toy models, which are fast and easy to run and give us a full understanding of the processes we are trying to model. This option is investigated in the following sections. However, toy models ignore the complexity of climate data, which could have a significant impact on the performance of the algorithm. That is why, in this section, we investigate the second option, which is simulating resampling by using an already existing very long climate model output dataset, gradually showing more data to the neural networks, and monitoring how the estimated committor improves.

More precisely, we will use the 8000-year-long PlaSim output dataset described in chapter 3. Similarly, we will use the Convolutional Neural Network architecture described in section 2.1 and used in chapter 3 for the probabilistic classification task of predicting the 5% most extreme two-week heatwaves over France. The network takes as inputs the stacked fields of 2 m temperature, 500 hPa geopotential height and soil moisture anomalies at time $t-\tau$ and returns the committor function q, which is the probability that a heatwave starts at time t and lasts at least $T = 14$ days. For more details on the architecture of the network and the definition of a heatwave, we refer to chapters 2 and 3 respectively.

Now, what we will do is take the 8000-year-long dataset and perform 10-fold cross-validation. For each fold we will have a total training dataset with N data points (corresponding to 7200 years) and a validation one with N_v data points (800 years). Instead of training on all the data, we will first show the network only a fraction p_0 of the training set. Then we will use the trained network to compute the committor on the remaining $(1-p_0)N$ data points, and select based on this $p_1(1-p_0)N$ data points to add to the initial small training set. We will then continue training the network on the new dataset containing $(p_0 + p_1(1-p_0))N$ points and compare the change in the validation loss. The idea is then to empirically find the optimal way of selecting new data to improve the network.

Since we want to address the sensitivity of the committor to the addition of new data, we'll perturb only slightly the original dataset, thus using $p_1 = 0.01$. In fig. 6.1, we show the results for an experiment run with $p_0 = 0.3$, and where the criterion for selecting new data is whether the predicted committor falls into a specified interval $[q_{\min}, q_{\max}]$. By looking at the green curve, we can see from the rightmost point that simply adding $p_1(1-p_0)N$ random data points (the last committor range is $[0, 1]$ which poses no condition on the predicted committor) doesn't cause a decrease in the validation loss. Indeed, after seeing $p_0 \cdot 7200 = 2160$ years of data, the additional 50 years of data with the same distribution don't cause any effect. On the other hand, when we actually select based on the value of the predicted committor we do see a small improvement, which is maximum when $q \in [0.1, 0.2]$.

To gain some insight into how the added data changes the predictions of the neural network, we can plot separately the contributions to the validation loss that come from heatwave and non-heatwave data. This is represented respectively by the orange and blue curves in fig. 6.1, where we can see that we have an overall improvement of prediction on non-heatwave data and a worsening of prediction on heatwave data. This is essentially due to the neural network becoming more conservative and predicting overall lower values of the committor. After all, there are 19 times more non-heatwave data than heatwaves, and in my early studies I have seen that the network becoming more conservative is a general trend as its training progresses. In any case, going more in depth on this matter is not particularly relevant here.

More interesting is to repeat the experiment at different stages of training of the network, namely at different values of p_0. Indeed, in fig. 6.2, we see that the smaller p_0 is, the higher the benefit of adding more training data. What is remarkable, is that the more data the network has seen, the more it improves when seeing data with a lower predicted committor. In other words, an inexperienced network benefits more from data that is very likely to lead to heatwaves, while a more experienced network may benefit more from learning the subtleties of states where the heatwave probability is much lower.

Now, all this analysis has given the important result that there seems to be an optimal non-trivial way to add new data, which depends on how good the network is. However, the method proposed here is mainly a qualitative proof of concept rather than an actual recipe. Indeed, when we add data for which the predicted committor sits in $[q_{\min}, q_{\max}]$,

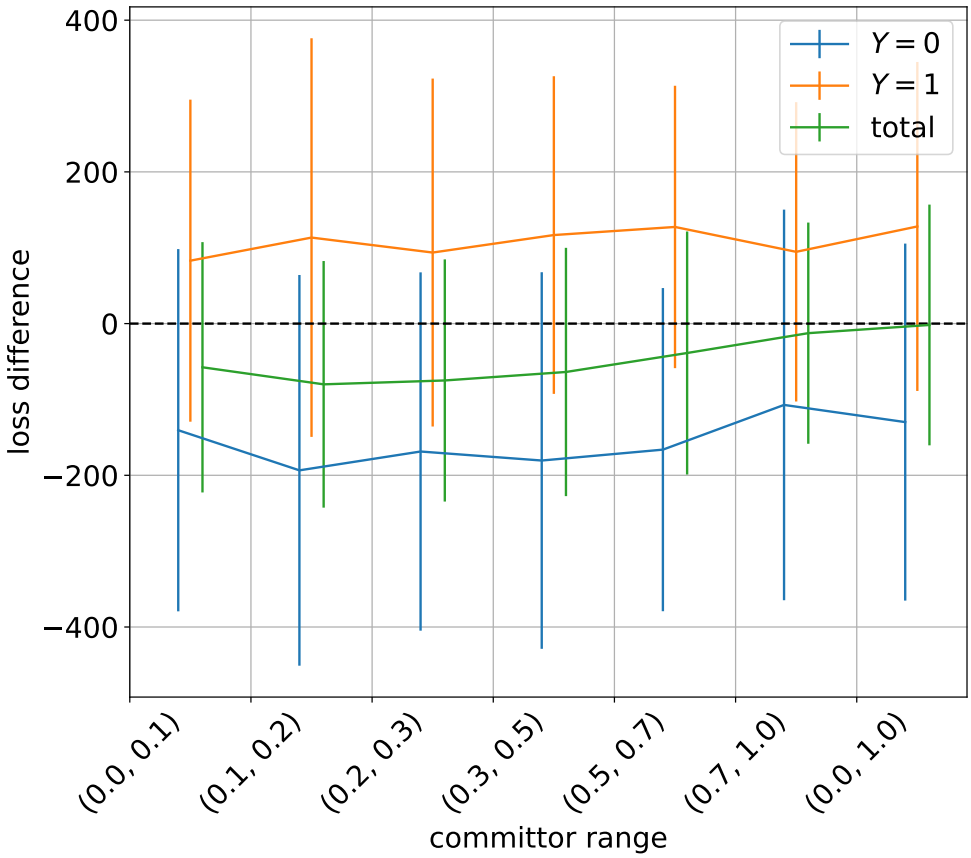


Figure 6.1: Difference in the validation loss (lower is better) before and after adding new training data in specified committor windows (x-axis). The last committor interval is to control for the effect of simply adding more training data. The blue line is the total contribution of the non-heatwave data ($Y = 0$), while the orange line is the contribution of the heatwave data ($Y = 1$). The green line is the sum of the two, i.e. the overall change in the loss function. Error bars represent one standard deviation across the 10 folds. In this experiment $p_0 = 0.3$ and $p_1 = 0.01$.

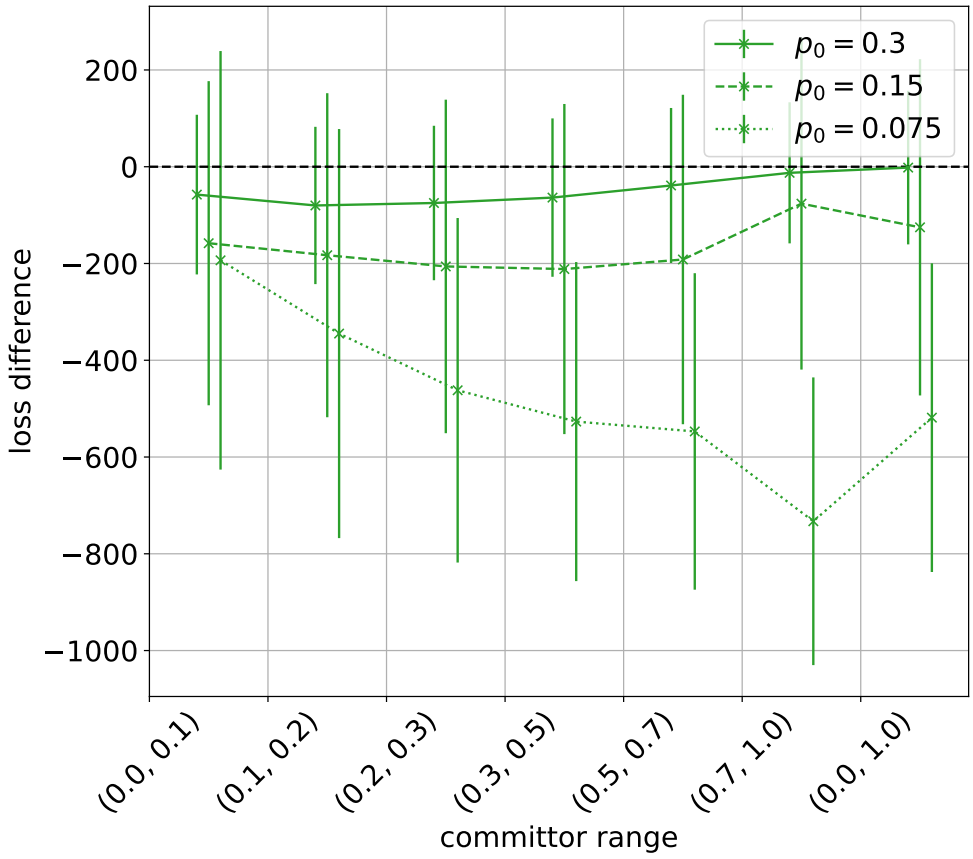


Figure 6.2: Overall validation loss difference before and after adding new training data in specified committor intervals, for networks initially trained on different amounts of data. The interval that yields the highest improvement moves to lower values of the committor the more data the network has seen. Error bars represent one standard deviation across the 10 folds.

the average fraction of actual heatwaves in this additional dataset also belongs to this interval, as the network is good at its job and so the predicted committor is *consistent*. This means that in the additional dataset, the fraction of heatwaves can be different from the 5% of the total dataset. Since the network learns the proper probabilities to assign to the events from the composition of the training data, changing this composition will inevitably introduce biases in the new predictions. And it is not immediately clear how to fix this problem. One option would be to alter the data selection process and force the new dataset to contain exactly 5% heatwaves. This wouldn't be a problem in our *simulated* resampling experiment, but when we actually do real resampling running a climate model, we don't know a-priori which X would lead to a heatwave, and so there is no way to enforce this constraint.

Moreover, although the huge error bars associated with these preliminary results could be mitigated by choosing a larger p_1, this method involves tons of heuristics, and it is difficult to come up with a general, robust protocol to apply when we run the climate model.

For these reasons, in the following section we will take a step back, and discuss the problem of optimal resampling from a more theoretical point of view.

6.3 Theoretical framework for optimal importance sampling

In the previous section, we worked directly on a committor learned through probabilistic classification. As pointed out in section 2.1 and chapter 4, another option is instead to work with the framework of probabilistic regression to predict the conditional distribution of the heatwave amplitude A, and compute the committor later. Since this option is better suited for concrete applications to the prediction of heatwaves, in the following of this chapter we will focus on probabilistic regression.

However, as we discussed in section 2.1, one of the main disadvantages of probabilistic regression is that the network may be distracted by information in the bulk of the distribution that is not relevant for the behavior of the tails, which is what we are most interested in. To obviate this problem, we propose the use of a weighted loss function, which will force the network to focus on the tails.

6.3.1 A tail-oriented loss function

Let us assume that we have a set of predictors $X \in \mathbb{R}^d$ that we use to forecast the observable $A \in \mathbb{R}$. In particular, X is distributed according to the true distribution $P_0(x)$, and A to the true conditional distribution $P_0(a|x)$. Together, the pair (X, A) follows the true joint distribution $P_0(x, a)$. In the context of application to climate, X would be for example a set of climate variable that follows the stationary distribution, and A the heatwave amplitude.

Our goal is to estimate with a parametric function $P(a|x;\theta)$ the conditional distribution $P_0(a|x)$, with a particular focus on the tail of A. Namely, we are mostly interested in, for

instance, extreme heatwaves.

Now, if the latter condition didn't apply, we could simply minimize the Kullback-Leibler (KL) divergence between the true distribution $\mu = P_0(x, a)$ and the estimator $\nu = P(x, a; \theta) = P_0(x)P(a|x; \theta)$, i.e.

$$\mathrm{KL}(\mu, \nu) = \int \mathrm{d}\mu \, \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right), \tag{6.1}$$

To focus on the tail, we want to add a weight function $w(a)$. For simplicity, in the rest of this work we will assume that

$$w(a) = \mathbb{1}_{a>a_0} = \begin{cases} 1 & \text{if } a > a_0 \\ 0 & \text{otherwise} \end{cases}, \tag{6.2}$$

but all the following discussion generalizes to any $w : \mathbb{R} \to [0, +\infty)$.

Ideally we would want to compare the weighted distributions $\mu = w(a)P_0(x, a)$ and $\nu = w(a)P(x, a; \theta)$, but we cannot do so with the simple KL divergence. Indeed, since the weighted distributions are no longer normalized, a simple cheat in the optimization would be to concentrate all the mass of P in the region where $w(a)$ is high. This is not what we are after.

We can then use something similar to what is called *penalized weighted likelihood* in [Pelenis, 2014] and *censored likelihood* in [Diks et al., 2011]:

$$\mathcal{H}(\mu, \nu) := \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - \frac{\mathrm{d}\mu}{\mathrm{d}\nu} + 1\right) \mathrm{d}\nu, \tag{6.3}$$

which is non-negative and reaches 0 only when $\mu = \nu$.

Substituting for μ and ν and remembering that $\mathrm{d}P_0(x, a) \equiv P_0(x, a)\mathrm{d}x\mathrm{d}a$, we get

$$\begin{aligned}\mathcal{H}(\theta) &= \mathcal{H}\left(w(a)P_0(x, a),\, w(a)P(x, a; \theta)\right) \\ &= \int P_0(x, a)\mathrm{d}x\mathrm{d}a \, w(a) \log\left(\frac{P_0(a|x)}{P(a|x; \theta)}\right) + \int P_0(x)\mathrm{d}x\mathrm{d}a \, w(a) \left(P(a|x; \theta) - P_0(a|x)\right).\end{aligned}$$

Now, if we drop the terms that don't depend on θ, we get what we call the H loss:

$$h(\theta) = -\int P_0(x, a)\mathrm{d}x\mathrm{d}a \, w(a) \log(P(a|x; \theta)) + \int P_0(x)\mathrm{d}x\mathrm{d}a \, w(a)P(a|x; \theta) \tag{6.4}$$

In the second term, we notice that we can perform the integral over a of the parametric conditional distribution, defining the normalization term

$$F(x; \theta) := \int \mathrm{d}a \, w(a)P(a|x; \theta). \tag{6.5}$$

This term will ensure that the parametric conditional distribution has the proper mass in the weighted region (e.g. above the threshold a_0 if $w(a)$ is given by eq. (6.2)), while the first term is the weighted Negative Log Likelihood, which will ensure that, *in the weighted region*, the estimator is close to the true distribution. Putting all the pieces together, the final expression for the H loss is

$$h(\theta) = -\int P_0(x, a)\mathrm{d}x\mathrm{d}a \, w(a) \log(P(a|x; \theta)) + \int P_0(x)\mathrm{d}x \, F(x; \theta) \tag{6.6}$$

In practice, however, we don't have access to the true distribution P_0, so we replace it with the empirical estimate from a dataset $\mathcal{D} = \{(X_i, A_i)\}_{i=1}^N$. Then,

$$P_0(x,a) \simeq \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i)\delta(a - A_i), \quad P_0(x) \simeq \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i), \tag{6.7}$$

which leads us to the empirical estimation of the H loss:

$$\hat{h}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \left(w(A_i)\log(P(A_i|X_i;\theta)) + F(X_i;\theta)\right) \tag{6.8}$$

Notice that the observed heatwave amplitudes A_i contribute only to the first term, and if $w(a)$ is indeed an indicator function, then this first term will be zero $\forall\, A_i < a_0$. If we set a_0 to focus on the tail of the unconditional distribution of A, then this means that only a very small part of the dataset will contribute to the weighted Negative Log Likelihood. We thus expect high variance in the estimation of the optimal parameters $\hat{\theta}$, and this motivates using importance sampling to generate new data and mitigate the variance of the estimator.

6.3.2 Importance sampling

Let us assume we have a function $L : \mathbb{R}^d \to [0, +\infty)$, with the properties that

$$\int P_0(x)\mathrm{d}x\, L(x) = 1 \tag{6.9}$$

$$L(x) > 0 \quad \forall x \text{ such as } \int P_0(a|x)w(a)\mathrm{d}a > 0, \tag{6.10}$$

then we call *importance sampling* the process of drawing M samples $\tilde{X}_1, \ldots, \tilde{X}_M$ distributed according to $P_0^L(\tilde{x}) := P_0(\tilde{x})L(\tilde{x})$, and then generating $\tilde{A}_1, \ldots, \tilde{A}_M$ according to the original conditional distribution $P_0(\tilde{a}|\tilde{x})$.

Then, in the limit $M \to +\infty$, we get a new estimate of the H loss:

$$h^L(\theta) = -\int P_0^L(\tilde{x},\tilde{a})\mathrm{d}\tilde{x}\mathrm{d}\tilde{a}\, \frac{1}{L(\tilde{x})} w(\tilde{a})\log(P(\tilde{a}|\tilde{x};\theta)) + \int P_0^L(\tilde{x})\mathrm{d}\tilde{x}\, \frac{1}{L(\tilde{x})} F(\tilde{x};\theta). \tag{6.11}$$

The first property of L (eq. (6.9)) ensures that P_0^L is a probability distribution, while the second is a sort of ergodicity condition, that ensures that all states of x that can contribute to the weighted negative log-likelihood term are sampled, at least in the limit $M \to +\infty$ [Paananen et al., 2021]. Thanks to these two properties, $h^L(\theta) = h(\theta)$.

Once again, however, we have to work with finite amounts of data, both in the original dataset and in the resampled one. In this context, since $P_0(x)$ will be approximated by a sum of delta functions, we have the important point that the resampled points $\tilde{X}_j$ will be chosen among the set of the original X_i. This means that we don't gain any additional variability in the predictors. What we do gain is variability in the $\tilde{A}_j$ as this time they are not sampled but rather simulated by a stochastic climate model initialized at $\tilde{X}_j$. Note

that if the climate model is per se deterministic, we can apply a small random perturbation to the initial conditions to make it effectively stochastic (see for instance section 5.4.2).

Now, since the normalization term depends only on X, there is no advantage in estimating it with the resampled data with respect to using the original data points. So, when we write the empirical version of h^L, we get

$$\hat{h}^L(\theta) = -\frac{1}{M}\sum_{j=1}^{M}\frac{1}{L(\tilde{X}_j)}w(\tilde{A}_j)\log(P(\tilde{A}_j|\tilde{X}_j;\theta)) + \frac{1}{N}\sum_{i=1}^{N}F(X_i;\theta). \tag{6.12}$$

The question now is: what is the optimal function L that allows to resample data in such a way that the θ that minimizes $\hat{h}^L(\theta)$ gives the lowest possible $h(\theta)$?

6.3.3 Optimal resampling function

Let us assume that the minimizer $\hat{\theta}^L$ of $\hat{h}^L(\theta)$ is close to true minimizer θ^* of $h(\theta)$.

Then, we can expand the true loss at this value as

$$h(\hat{\theta}^L) \simeq h(\theta^*) + \frac{1}{2}(\hat{\theta}^L - \theta^*)^\top \nabla^2 h(\theta^*)(\hat{\theta}^L - \theta^*), \tag{6.13}$$

where ∇ denotes taking the gradient with respect to θ and the first order term is missing because θ^* is the minimizer of the true loss.

Now, if we consider that $\hat{\theta}^L$ minimizes $\hat{h}^L(\theta)$,

$$0 = \nabla\hat{h}^L(\hat{\theta}^L) \simeq \nabla\hat{h}^L(\theta^*) + \nabla^2\hat{h}^L(\theta^*)(\hat{\theta}^L - \theta^*), \tag{6.14}$$

which gives

$$\hat{\theta}^L - \theta^* = -\left(\nabla^2\hat{h}^L(\theta^*)\right)^{-1}\nabla\hat{h}^L(\theta^*) \tag{6.15}$$

Then, denoting with $\mathbb{E}$ the expectation in the limit $M \to +\infty$, and assuming independently that $N \to \infty$, we get

$$\mathbb{E}\left(\nabla^2\hat{h}^L(\theta^*)\right) = \nabla^2 h(\theta^*), \tag{6.16}$$

and similarly,

$$\mathbb{E}\left(\nabla\hat{h}^L(\theta^*)\right) = \nabla h(\theta^*) = 0. \tag{6.17}$$

Now, let us define

$$\ell(\tilde{X}_j, \tilde{A}_j;\theta) = -\frac{1}{L(\tilde{X}_j)}w(\tilde{A}_j)\log(P(\tilde{A}_j|\tilde{X}_j;\theta)) \tag{6.18}$$

as the weighted Negative Log Likelihood contribution of each sample.

Then,

$$\hat{h}^L(\theta) = \frac{1}{M}\sum_{j=1}^{M}\ell(\tilde{X}_j, \tilde{A}_j;\theta) + \frac{1}{N}\sum_{i=1}^{N}F(X_i,\theta) =: \bar{\ell}(\theta) + \bar{F}(\theta) \tag{6.19}$$

And eq. (6.17), tells us that

$$\mathbb{E}\left(\nabla\bar{\ell}(\theta)\right) = -\mathbb{E}\left(\nabla\bar{F}(\theta)\right) = -\nabla\bar{F}(\theta), \tag{6.20}$$

where the last step is due to the fact that the expectation $\mathbb{E}$ is performed over resampled data and the normalization term uses the original data.

Then, according to the central limit theorem,

$$\sqrt{M}\nabla\hat{h}^L(\theta^*) = \sqrt{M}\left(\nabla\bar{\ell}(\theta^*) - \mathbb{E}\left(\nabla\bar{\ell}(\theta^*)\right)\right) \sim_{M\to+\infty} \mathcal{N}(0, K^L(\theta^*)), \tag{6.21}$$

where $\mathcal{N}(\mu, K)$ denotes a multivariate normal distribution with mean μ and covariance matrix K, and

$$K^L(\theta^*) = \mathbb{E}\left(\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)^\top\right) \tag{6.22}$$

is the covariance matrix of the random vector $\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)$.

Putting the pieces together in eq. (6.15), we get

$$\sqrt{M}\left(\hat{\theta}^L - \theta^*\right) \sim_{M\to+\infty} \mathcal{N}\left(0, \left(\nabla^2 h(\theta^*)\right)^{-1} K^L(\theta^*) \left(\nabla^2 h(\theta^*)\right)^{-1}\right) \tag{6.23}$$

And finally, inserting this result into eq. (6.13),

$$\mathbb{E}\left(h(\hat{\theta}^L)\right) \simeq h(\theta^*) + \frac{1}{2M}\mathrm{Tr}\left[\left(\nabla^2 h(\theta^*)\right)^{-1} K^L(\theta^*)\right], \tag{6.24}$$

where $\mathrm{Tr}[\bullet]$ denotes the trace of matrix $\bullet$.

Now, the term we want to minimize with the choice of L is

$$V^L = \mathrm{Tr}\left[\left(\nabla^2 h(\theta^*)\right)^{-1} K^L(\theta^*)\right]. \tag{6.25}$$

Remembering the definition of $K^L(\theta^*)$ in eq. (6.22), we get

$$\begin{aligned}
V^L &= \mathrm{Tr}\left[\left(\nabla^2 h(\theta^*)\right)^{-1} \mathbb{E}\left(\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)^\top\right)\right] \\
&= \mathbb{E}\left(\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*) + \nabla\bar{F}(\theta^*)\right)\right) \\
&= \mathbb{E}\left(\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)\right)\right) + \\
&+ \left(\nabla\bar{F}(\theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla\bar{F}(\theta^*)\right) + 2\left(\nabla\bar{F}(\theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \mathbb{E}\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)\right) \\
&= \mathbb{E}\left(\left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla\ell(\tilde{X}, \tilde{A}, \theta^*)\right)\right) + \\
&- \left(\nabla\bar{F}(\theta^*)\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla\bar{F}(\theta^*)\right)
\end{aligned}$$

where in the second to last step we exploit the fact that the hessian $\nabla^2 h(\theta^*)$ is symmetric and in the last step we used once again $\mathbb{E}\left(\nabla\ell(\theta)\right) = \mathbb{E}\left(\nabla\bar{\ell}(\theta)\right) = -\nabla\bar{F}(\theta)$.

Now, the second term in the expression for V^L doesn't depend on the choice of L, so we can drop it. To develop the first term V_1^L, we remember that the expectation $\mathbb{E}$ is with respect to the resampled data $(\tilde{X}, \tilde{A}) \sim L(\tilde{x})P_0(\tilde{x}, \tilde{a})$, so

$$V_1^L = \int L(\tilde{x})P_0(\tilde{x},\tilde{a})\mathrm{d}\tilde{x}\mathrm{d}\tilde{a}\,\left(\frac{w(\tilde{a})}{L(\tilde{x})}\right)^2 \left(\nabla \log(P(\tilde{a}|\tilde{x};\theta^*))\right)^\top \left(\nabla^2 h(\theta^*)\right)^{-1} \left(\nabla \log(P(\tilde{a}|\tilde{x};\theta^*))\right)$$
$$= \int \frac{P_0(\tilde{x},\tilde{a})}{L(\tilde{x})}\mathrm{d}\tilde{x}\mathrm{d}\tilde{a}\, G(\tilde{x},\tilde{a};\theta^*)$$

where

$$G(x,a;\theta) = w(a)^2 \left(\nabla \log(P(a|x;\theta))\right)^\top \left(\nabla^2 h(\theta)\right)^{-1} \left(\nabla \log(P(a|x;\theta))\right) \tag{6.26}$$

Finally, if we define

$$g(x;\theta) = \int P_0(a|x)\mathrm{d}a\, G(x,a;\theta), \tag{6.27}$$

it is easy to prove that the optimal resampling function L is given by

$$L^*(x) = \frac{\sqrt{g(x;\theta^*)}}{\int P_0(x)\mathrm{d}x\,\sqrt{g(x;\theta^*)}} \tag{6.28}$$

Taking a step back from the crude calculations, we can interpret L^* as essentially proportional to the norm of $w(a)\nabla \log P(a|x;\theta^*)$, according to the metric given by the inverse of the hessian. In other words, in some loose sense, $L^*(x)$ is a measure of the sensitivity of the predicted probability distribution to the parameters θ at the point x. It is then reasonable that we want to resample more on the points where this sensitivity is higher.

6.3.4 Optimal resampling of the validation set

Since $P(a|x;\theta)$ will be parameterized with a neural network, we don't need just a training set, but also a validation one to be able to do early stopping and prevent overfitting. However, the validation set will also experience the same high variance problems, so there is interest in performing optimal resampling on it as well.

If we assume once again to have N_v points in the original validation set and to resample M_v new points using a resampling function $L_v(x)$, then the expression of the new loss would be exactly the same as in eq. (6.12):

$$\hat{h}^{L_v}(\theta) = -\frac{1}{M_v}\sum_{j=1}^{M_v} \frac{1}{L_v(\tilde{X}_j)} w(\tilde{A}_j)\log(P(\tilde{A}_j|\tilde{X}_j;\theta)) + \frac{1}{N_v}\sum_{i=1}^{N_v} F(X_i;\theta). \tag{6.29}$$

This time however, we don't want to optimize θ, nor to minimize $\hat{h}^{L_v}(\theta)$, but rather to resample the data in such a way that $\hat{h}^{L_v}(\theta)$ is as close as possible to the true loss $h(\theta)$. Namely, we want to evaluate how good θ is as accurately as possible. This is an important point, because if our goal was to minimize $\hat{h}^{L_v}(\theta)$, then there would be the easy cheat to put most of the weight on the X_i that the network predicts best. Clearly not what we want.

To find the optimal L_v, then, we can apply the central limit theorem directly to the empirical loss. With respect to many resampling iterations, $\hat{h}^{L_v}(\theta)$ will have a variance $\frac{1}{M}V^{L_v}$, with

$$V^{L_v} = \mathbb{E}\left(\frac{1}{L_v(\tilde{X})}w(\tilde{A})\log(P(\tilde{A}|\tilde{X};\theta))\right)^2 \tag{6.30}$$

$$= \int \frac{P_0(\tilde{x},\tilde{a})}{L(\tilde{x})}\mathrm{d}\tilde{x}\mathrm{d}\tilde{a}\ \left(w(\tilde{a})\log(P(\tilde{a}|\tilde{x};\theta))\right)^2. \tag{6.31}$$

To minimize it, the optimal resampling function will be

$$L_v^*(x) = \frac{\sqrt{g_v(x;\theta)}}{\int P_0(x)\mathrm{d}x\,\sqrt{g_v(x;\theta)}}, \tag{6.32}$$

with

$$g_v(x;\theta) = \int P_0(a|x)\mathrm{d}a\ \left(w(a)\log(P(a|x;\theta))\right)^2. \tag{6.33}$$

6.3.5 Resampling algorithm

When we want to run the algorithm with actual data, we don't have access to the true distribution P_0, and neither to the true loss h nor its true minimizer θ^*. We then need to replace these quantities with empirical estimates, and using $P_0(x) \simeq \frac{1}{N}\sum_{i=1}^{N}\delta(x - X_i)$, $h(\theta) \simeq \hat{h}(\theta)$ and $\theta^* \simeq \hat{\theta}$, feels very natural and doesn't pose major problems.

On the other hand, this is not the case for the conditional distribution $P_0(a|x = X_i) \simeq \delta(a - A_i)$, as when we generate new $\tilde{A}_j$ we are using the true distribution in the form of running the climate model. In particular, this dichotomy means that the second condition on L (eq. (6.10)) is no longer guaranteed. Indeed, approximating the true conditional distribution with a delta on the data means that $g(X_i;\theta) = G(X_i, A_i;\theta)$. This, in turns, makes the resampling function depend also on the values of A, and in particular $L^*(X_i) \propto w(A_i)$. Therefore, if $w(A_i) = 0$, then X_i will never be resampled, even in the limit $M \to \infty$, but it could have been a good starting point for generating heatwaves and was just unlucky on the one single realization we observed.

All of this matter has the potential to introduce biases in the resampling procedure, however it cannot be fully eliminated, as $P_0(a|x)$ is exactly the object we want to estimate. A partial mitigation would be to use a weighting function $w(a)$ that is always strictly positive, but this can complicate the calculations, especially of the normalization term $F(x;\theta)$, where we have to integrate $w(a)P(a|x;\theta)$. Another option could be to approximate $P_0(a|x)$ with another data driven method, for instance the analogue Markov chain [Lucente et al., 2019, 2022b]. However, the mathematical soundness of this approach should be investigated in detail, as it is not that different from using $P(a|x;\theta)$ as a surrogate of $P_0(a|x)$, and this latter assumption would lead to tautologies in the loss functions that would make the algorithm ineffective.

As already said, at this stage we value simplicity, and so we will stick with $w(a) = \mathbb{1}_{a>a_0}$ and $P_0(a|x = X_i) \simeq \delta(a - A_i)$, but we are aware of the downsides of this choice.

Now that we discussed the caveats of using empirical estimates, we can flesh out the practical algorithm that will be used in the following of this work.

1. Perform a control run with the climate model and split the available data into a training set $\mathcal{D} = \{(X_i, A_i)\}_{i=1}^{N}$ and a validation one $\mathcal{D}^v = \{(X_i^v, A_i^v)\}_{i=1}^{N_v}$.

2. Optimize $\hat{h}(\theta)$ on the training set, early stopping on the validation set, which yields the optimal parameter vector $\hat{\theta}$.

3. Resample the training set

   (a) Compute the hessian of the empirical loss with respect to the parameters $H = \nabla^2 \hat{h}(\hat{\theta})$.

   (b) For each data point (X_i, A_i) in $\mathcal{D}$, compute the gradient of the weighted Negative Log Likelihood $y_i = -w(A_i)\nabla \log(P(A_i|X_i;\hat{\theta}))$.

   (c) Compute the function G for every data point $G_i = y_i^\top H^{-1} y_i$.

   (d) Compute the optimal resampling function $L_i = N\frac{\sqrt{G_i}}{\sum_k \sqrt{G_k}}$

   (e) Draw from $i = 1, \ldots, N$, M samples with repetition and weights L_i, establishing the parent mapping function $\iota(j)$ which indicates which i was chosen at the j-th draw (see for instance eq. (5.4) for a practical way to implement it).

   (f) Simulate with the climate model new values $\tilde{A}_j$ according to $P_0(a|x = \tilde{X}_j = X_{\iota(j)})$.

4. Resample the validation set

   (a) For each data point (X_i, A_i) in $\mathcal{D}^v$, compute the weighted Negative Log Likelihood $l_i = -w(A_i)\log(P(A_i|X_i;\hat{\theta}))$.

   (b) Compute the optimal resampling function $L_i^v = N\frac{l_i}{\sum_k l_k}$

   (c) Draw from $i = 1, \ldots, N_v$, M_v samples with repetition and weights L_i^v, establishing the parent mapping function $\iota^v(j)$ which indicates which i was chosen at the j-th draw.

   (d) Simulate with the climate model new values $\tilde{A}_j$ according to $P_0(a|x = \tilde{X}_j = X_{\iota^v(j)})$.

5. Find $\hat{\theta}^L$ which minimizes the new loss $\hat{h}^L(\theta)$ on the training set, where the weighted Negative Log Likelihood term is estimated with the resampled data and the normalization term $\bar{F}(\theta)$ is estimated with the original data. During the optimization, perform early stopping on the validation set, similarly using the resampled data to estimate the weighted Negative Log Likelihood term and the original one to estimate the normalization term.

In principle, one could then iterate from step 2 multiple times. However, since we never sample new points X, but rather select among the existing ones, necessarily at each

iteration of the algorithm we reduce the diversity in the sampling of the predictors. Hence, the number of iterations one can perform is limited. A better use of the computational resources is then simply to increase the number M of resampled points or even the length of the initial control run, so that the neural network that we train on the initial data is already relatively accurate.

In the next section, we will discuss some interesting computational challenges that arise when implementing this algorithm in practice.

6.4 Implementation

The algorithm presented above was implemented as a Python package in the GitHub repository `https://github.com/AlessandroLovo/importance-sampling4parameter-estimation`, which is one of the main contributions to this work on my part.

The parametric models for $P(a|x;\theta)$ are implemented as neural networks using the Keras-Tensorflow library, and are trained by stochastic gradient descent with the Adam optimizer. This seemingly minor point had its unique challenges in the context of this work, as the loss functions in eqs. (6.12) and (6.29) have two terms which are computed from two different datasets.

Moreover, the algorithm presented above was the results of many iterations between theory and experiments on toy models, which resulted in a lot of variants of the resampling algorithm. This was an additional challenge from the coding perspective.

Besides these purely technical points, converting the algorithm from theory to code highlighted some interesting points that are worth discussing in a more detail.

6.4.1 Parametric expression of the conditional distribution

The algorithm requires a parametric expression of the conditional distribution $P(a|x;\theta)$. This is the common tool used in all probabilistic forecast problems, so there are standard ways to achieve this. In this work we chose the simplest approach, which is to assume that the conditional distribution is a Gaussian distribution of which we estimate its mean and variance. Namely,

$$P(a|x;\theta) \simeq \mathcal{N}(\mu(x;\theta), \sigma^2(x;\theta)), \tag{6.34}$$

which is the same approach that we used in chapter 4.

Other approaches would be, for example, to parameterize the conditional distribution in a more complex form with more moments, or to approximate it directly with quantile regression [Haugen et al., 2018]. These methods are more flexible, but also more unstable, and they require a lot of data to be properly effective. Since we want to run our algorithm exactly because we don't have enough data, we chose the simplest approach.

6.4.2 The hessian matrix may not be positive definite

To compute the optimal resampling function L for the training set, we need the hessian matrix of the H loss with respect to the parameters of the network $\theta \in \mathbb{R}^P$. In theory, since $\hat{\theta}$ is the minimizer of the H loss, the hessian matrix should be positive definite, which later ensures that $L \geq 0$. However, in practice, the neural network optimizes itself with stochastic gradient descent and early stopping, so there is no guarantee that $\hat{\theta}$ is even a local minimizer. Indeed, convergence to a true local minimum might lead to overfitting, so, in general, $\nabla \hat{h}(\hat{\theta}) \neq 0$. Besides, even if this was the case, since neural networks commonly use ReLU activation functions, the hessian would be at most semi-positive definite, which still poses problems when we need to invert it.

Then, the obvious solution is to regularize the hessian matrix H at the moment of inversion:

$$H \mapsto H + \epsilon \mathbb{I}, \quad \epsilon = \max(\epsilon_{\text{abs}}, -(1 + \epsilon_{\text{rel}})\lambda_1), \tag{6.35}$$

where $\mathbb{I}$ is the identity matrix, ϵ_{abs} and ϵ_{rel} are the regularization parameters and $\lambda_1 \leq \lambda_2 \leq \dots \lambda_P$ are the eigenvalues of H. This method ensures $L \geq 0$, but requires to compute the spectrum of H, which can potentially be expensive.

6.4.3 The hessian matrix is big

Even if we can regularize the hessian matrix as described above, there might be problems simply in storing it into memory. Indeed, H is a $P \times P$ matrix, and if we consider, for instance, the ScatNet architecture presented in chapter 4 for heatwave prediction, $P \approx 2 \cdot 10^4$, which means that in 32-bits-floats, H will take roughly 2 GB of memory. And ScatNet was the architecture with the smallest amount of parameters. Handling such large matrices may not be an issue if one is working on the CPU, but to properly leverage automatic differentiation, one should be working on the GPU, which usually has a much more limited memory.

That is to say, the computational resources available can pose an important constraint on the architecture of the network. For instance, on my laptop the algorithm crashed already when $P \sim 10^4$.

Since the hessian is the only object that scales as P^2, a possible more resource-friendly approach would be to completely replace the hessian matrix with the identity matrix. Indeed, a geometric way to view the matter is that L_i is the norm of $w(A_i)\nabla \log P(A_i|X_i;\hat{\theta})$, under the metric of H^{-1}. Replacing H by $\mathbb{I}$ would mean shifting to the euclidean norm.

Even though this replacement is not optimal according to the theory explained in section 6.3.3, practical tests on toy models (see later) show no significant impact on the results, which is a rather interesting finding in and of itself.

6.4.4 M is finite

When we perform the resampling step, choosing M new points, the expectation of the number of clones of point X_i is $n_i^\infty := \mathbb{E}(n_i) = \frac{M}{N} L(X_i)$, which guarantees that the loss in eq. (6.12) is unbiased. However, since M is finite, drawing with repetition from the initial pool of N points will give a number of clones that is distributed roughly as a Poisson process with mean n_i^∞. For a Poisson process, variance and mean are equal, so the number of clones will have fluctuations of the order of $\sqrt{n_i^\infty}$. This variance, in turns, will propagate to the performance of the network trained on the resampled data, and will make it more difficult to have robust results.

To counter this problem we apply the same trick that is used for rare event algorithms when cloning and killing trajectories. As explained in section 5.3.1, instead of drawing with repetition, we choose the number of clones as

$$n_i = \begin{cases} \lfloor n_i^\infty \rfloor + 1 & \text{with probability } n_i^\infty - \lfloor n_i^\infty \rfloor, \\ \lfloor n_i^\infty \rfloor & \text{with probability } 1 - (n_i^\infty - \lfloor n_i^\infty \rfloor), \end{cases} \tag{6.36}$$

where $\lfloor \bullet \rfloor$ is the integer part of $\bullet$.

This way we still have $\mathbb{E}(n_i) = n_i^\infty$, but this time the variance is

$$(n_i^\infty - \lfloor n_i^\infty \rfloor)\,(1 - (n_i^\infty - \lfloor n_i^\infty \rfloor)) \leq \frac{1}{4}. \tag{6.37}$$

6.5 Testing on a not-so-simple toy model

The end goal of this algorithm is to be applied to climate models and help to improve our understanding of the most extreme events. However, to speed up the development process, we started working with toy models.

We first started with a simple linear model where $X \in \mathbb{R}^d$ and $A \sim \mathcal{N}(\beta \cdot X, \sigma)$, where the goal is to estimate β and σ. However, we soon realized that such a model is not very relevant for resampling extreme events. Indeed, since everything is linear and Gaussian, we can estimate perfectly the distribution of the tails from the knowledge of the bulk, as explained in chapter 3. There is then no interest in resampling new data in the tails, and indeed the only improvement after resampling is simply because the neural network has more data to work with.

So, the toy model we are after has the following desiderata:

- The model is sufficiently simple
- The behavior in the tail of the distribution of A is different from the one in the bulk

In the following, I will present a toy model that satisfies this conditions, and later we will use it to perform a proof of concept of the algorithm presented above.

6.5.1 Description of the toy model

In a project involving many people, the development of this toy model was fully my responsibility. A running theme of this thesis is the use of linear projection of high dimensional data as a simple and interpretable dimensionality reduction method. It would then feel natural that I interpret the second desiderata of the toy model as: "The optimal projection pattern that describes the behavior in the tail of the distribution of A is different from the one in the bulk".

With this in mind, and after several attempts, I came up with the following toy model, which I unimaginatively named the Two Dimensional Activation Model (TDAM).

1. We start from $X \in \mathbb{R}^d$ distributed according to a multivariate Gaussian distribution with mean 0 and the identity as covariance matrix.

2. Using two projection patterns p_1 and p_2, we project linearly the data to a two-dimensional space: $f(x) = (f_1, f_2) = (p_1 \cdot x, p_2 \cdot x)$, where $p_1 \cdot p_2$ controls the correlation between the two indices f_1 and f_2.

3. We generate A drawing from a Gaussian distribution with mean $u(f_1, f_2)$ and standard deviation σ, where

$$u(f_1, f_2) = c_1(f_1 - f_1^0) + c_2 \frac{1}{1 + e^{-\omega(f_2 - f_2^0)}}, \tag{6.38}$$

   and $f_1^0, f_2^0, c_1, c_2, \omega$ and σ are parameters of the model.

As we can see, the first term in the definition of u gives a constant gradient c_1 of the expected heatwave amplitude A with respect to the first index f_1. When $\omega|f_2 - f_2^0| \gg 1$, the second term is roughly constant, and so to predict heatwaves we only need f_1. Which means, thinking in the high dimensional space of X, that we only need to project the data on the direction of p_1. On the other hand when $f_2 \approx f_2^0$, the second term becomes highly sensitive to f_2. In particular,

$$\frac{\partial}{\partial f_2} u(f_1, f_2)\bigg|_{f_2 = f_2^0} = \frac{c_2 \omega}{4}, \tag{6.39}$$

and so now the best direction onto which to project the data is $c_1 p_1 + \frac{c_2 \omega}{4} p_2$. The parameter ω controls how smooth the transition between these two regimes is, while, by properly choosing f_2^0, we can set the regime shift to happen in the tail of the distribution of A, as it is illustrated in fig. 6.3.

6.5.2 Proof of concept of the resampling algorithm

Now that we have a toy model that allows us to generate data extremely fast, we can test our resampling algorithm. To do so, we will need to choose how to parameterize $P(a|x;\theta)$. For the linear toy model, the algorithm didn't work because we easily describe bulk and tail with the same very simple parametric model of a linear regression. In our case

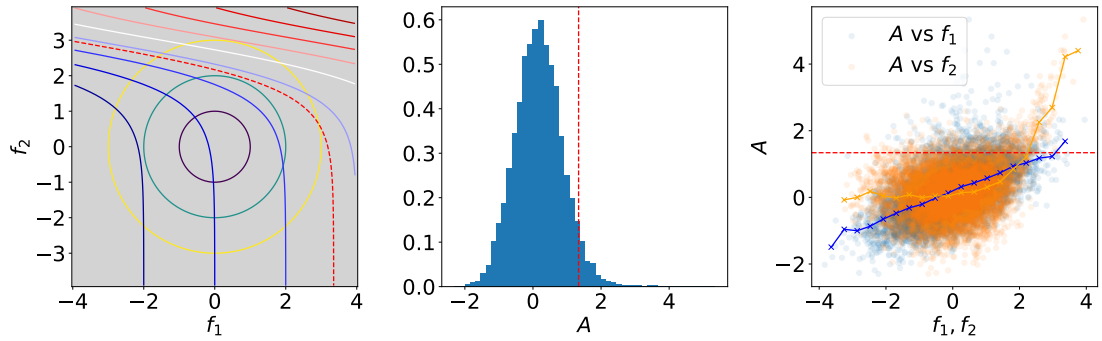


Figure 6.3: Visualization of the TDAM model, with parameters $c_1 = 0.4, c_2 = 6, f_1^0 = 0, f_2^0 = 3, \omega = 1.5, \sigma = 0.5$ and $p_1 \cdot p_2 = 0$. In the first panel, the blue-red lines are the contours of u, overlaid on the distribution of f_1 and f_2, which is represented by the three circles, corresponding to 1,2 and 3 standard deviations in the uncorrelated bivariate Gaussian distribution of f. The second panel shows the histogram of A, obtained from 10^4 samples. The third panel show the sampled points either as couples (f_1, A) (blue) or (f_2, A) (orange). The blue and orange lines show the average value of A conditioned on either f_1 or f_2 respectively. While the blue line increases linearly, the orange one is constant at 0 in the bulk (meaning that the value of f_2 gives no information on the value of A), and it increases rapidly for high values of f_2, which give the highest values of A as well. In all plots the dashed red line represents a_0 chosen such that 5% of the values of A are above it.

we face a similar problem, where if we make our parametric model sufficiently expressive, it will learn the two projection patterns p_1 and p_2 and the function u without need for resampling.

We will then use an Intrinsically Interpretable Neural Network (IINN) architecture (see section 2.3 and chapter 4) as shown in fig. 6.4. To limit the expressivity of the model, we set the bottleneck to have only $m = 1$ neurons, which means that the network will be able to learn a single projection pattern, and thus won't be able to capture the full picture of the toy model.

As pointed out before, the number of parameters of neural networks can quickly get out of hand, causing problems with the estimation of the hessian matrix and most importantly resulting in high variance in the results. To reduce these technical distractions to the minimum we simplify matters as much as possible, choosing to work directly in a $d = 2$-dimensional space, with $p_1 = (1, 0)$ and $p_2 = (0, 1)$. The weights of the first layer of the network will encode the learned projection pattern $\hat{p}$, which we can plot as a vector in the $d = 2$-dimensional space.

We will then sample from the toy model $N = 10^4$ training points and $N_v = 5 \cdot 10^3$ validation points and run the algorithm described in section 6.3.5 with $M = N$ and $M_v = N_v$. For the H loss we use the definition with $w(a) = \mathbb{1}_{a>a_0}$ with $a_0 = 1.34$, which leaves 5% of the data above the threshold.

When we train on the resampled data we could either reinitialize the weights of the network or simply continue from the weights at the end of the training on the initial data.

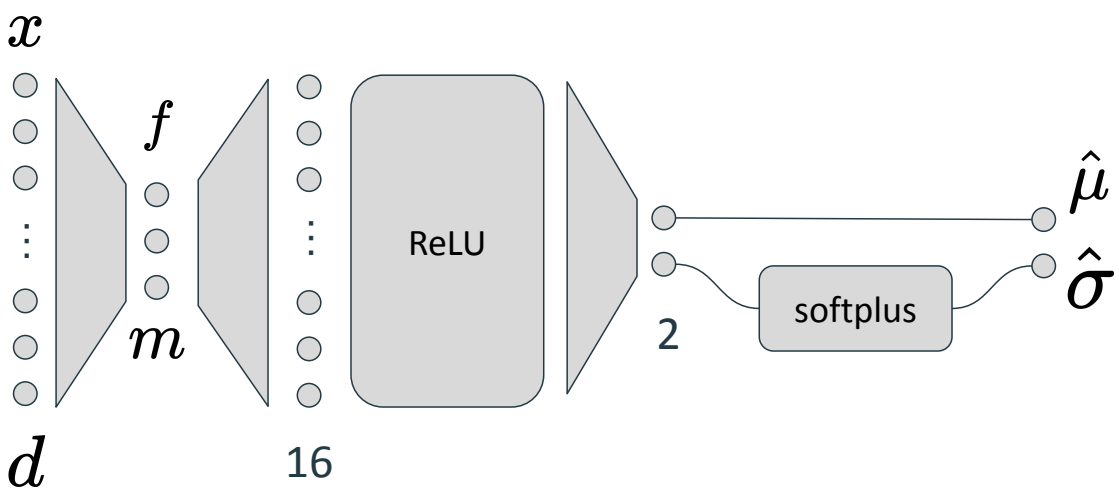


Figure 6.4: Schematics of a very simple Intrinsically Interpretable Neural Network (IINN) architecture. The data is first projected onto m optimal indices and then passed through a single hidden layer with 16 neurons and ReLU activation, followed by an output layer with 2 neurons. The softplus activation ($\text{softplus}(z) = \log(1 + \exp(z))$) on the second output neuron ensures that the predicted standard deviation is positive.

Since our architecture is very simple, we choose the former option.

In any case, after resampling, the network sees double the amount of data as before, and this could lead to an improvement by itself. To quantify this effect, for each experiment, we will also resample data in two other ways. The first is completely uniformly, that is, $L(X_i) = 1$ which means that for each X_i we will generate a second A_i. The second is uniformly on the data that had $A_i > a_0$, i.e. $L(X_i) = 20w(A_i)$.

In fig. 6.5 we compare the learned projection vector $\hat{p}$ after training on the initial data and after training on the data resampled in the three different ways. As we can see, the uniform resampling didn't displace $\hat{p}$ in the right direction, while the other two methods, that sample preferentially in the tail, did align $\hat{p}$ with the gradient of u in the tail.

Interestingly, uniform resampling above a_0 and our approach yield a very similar result, and indeed, if we compare the distribution of the resampled data between our algorithm and uniform resampling above a_0 (fig. 6.6), we can see that they also look similar, both managing to sample events which are very far in the tail of the original distribution of A.

So far these results are very qualitative, and prone to consistent fluctuations between different repetitions of the same experiment. To have something more quantitative and more robust we can simulate an independent test set with $N_t = 10^5$ points and look at how the H loss changes after resampling. To be more thorough, we will repeat the whole experiment 10 times to have an estimation of confidence intervals, and we won't just look at the H loss defined above with threshold a_0, but rather at a class of H losses defined on a variable threshold a_1. As a_1 increases, the focus will be more and more on the very extreme tail of A.

In fig. 6.7 we plot for this class of H losses the ratio of their value before resampling with

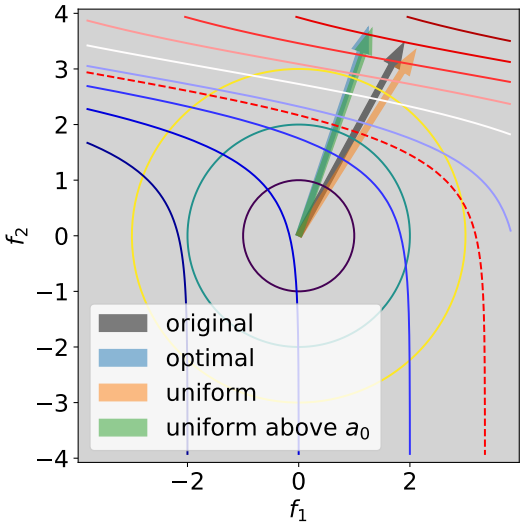


Figure 6.5: The learned projection vector $\hat{p}$ plotted on top of the contours of u (see fig. 6.3 for the details). In black $\hat{p}$ after training on the initial data, while in color $\hat{p}$ after training on the data resampled three different ways: in blue with our algorithm, in orange with uniform resampling and in green with uniform resampling of the data which had $A_i > a_0$. Uniform resampling doesn't achieve anything. Uniform resampling above a_0 and our approach yield a very similar improvement.

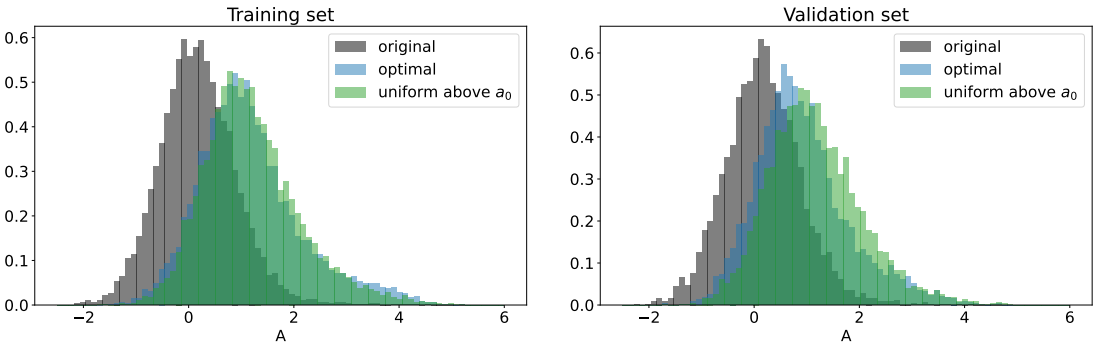


Figure 6.6: Comparison of the distribution of A in the original and resampled training and validation datasets.

the one after resampling, such that ratios above 1 mean an improvement in performance. Unsurprisingly, and consistently to what we observed on the projection vector $\hat{p}$, uniform resampling doesn't help when we are interested in the tails. On the other hand, the curves of uniform resampling above a_0 and our algorithm are virtually identical.

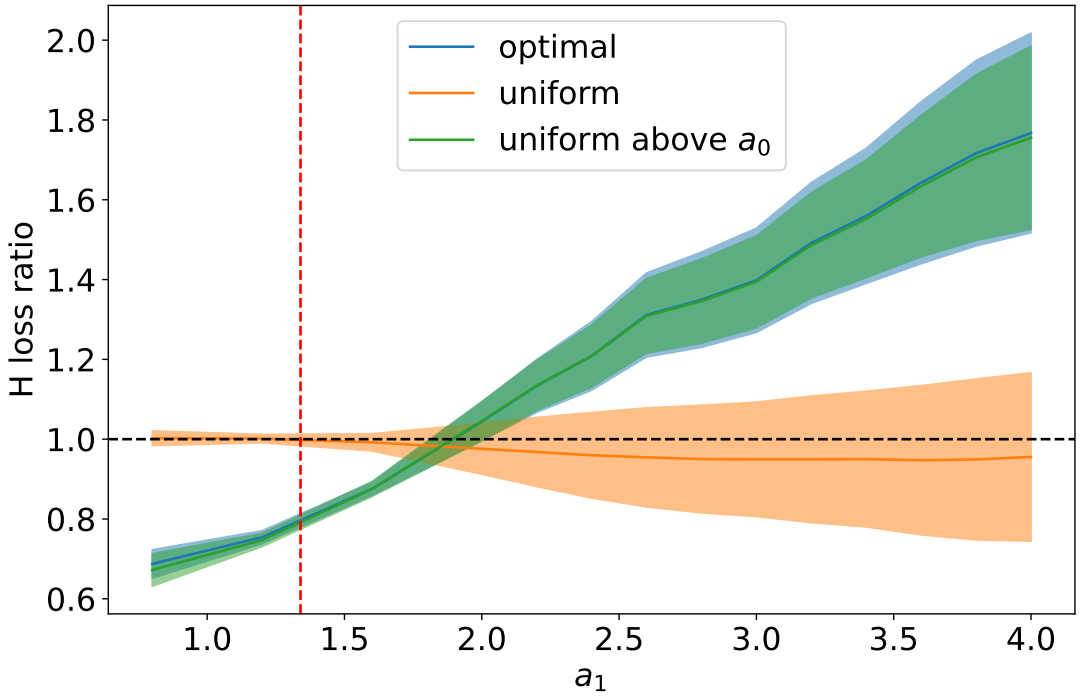


Figure 6.7: Ratio of the test H loss computed with threshold a_1 before resampling with the one computed after ($\hat{h}_{\text{before}}/\hat{h}_{\text{after}}$), as a function of different thresholds a_1 and for the three different resampling methods. A ratio above 1 means that the H loss after resampling is lower than before, so the prediction is better. The dashed red line is the threshold a_0 used for training and resampling. The shaded area corresponds to one standard deviation across 10 runs.

If we put aside this fact for a moment and observe how they behave, we notice that these resampling methods worsen the performance when $a_1 \sim a_0$, and significantly improve it for $a_1 > 2$. A possible explanation for this behavior is that, as already pointed out in section 6.3.5, the choice of $w(a) = \mathbb{1}_{a>a_0}$ leads to biases in the resampled loss. Indeed, since we use only the resampled data to estimate the weighted Negative Log Likelihood term, and the resampled data contains no X_i for which the original A_i was below the threshold a_0, we are missing a non-negligible contribution to the loss from the points x that were discarded but still had a significant chance to produce an $\tilde{A}_j > a_0$. It is then clear that the points which contribute the most to this bias are those for which the true average heatwave amplitude $u(x)$ prescribed by the model is of the order of one standard deviation below a_0. It is then no longer surprising that the loss of performance ends at $a_1 \approx a_0 + \sigma$. Fortunately, this bias and loss of performance problem is an effect of the hard edge of $w(a)$, which means that it can be mitigated by choosing a smoother function, or

completely ignored if we are interested in events far enough away from a_0.

Now it is time to address the elephant in the room, namely the fact that our complicated, and in theory optimal, algorithm behaves exactly the same as a simple uniform resampling of all the X_i for which the original A_i was above a_0. For starter, in section 6.4.3, we already pointed out that substituting the hessian matrix with the identity matrix didn't change the results. This should have already been an alarm bell that the algorithm proposed was perhaps unnecessarily complex, and now this result is striking confirmation.

The algorithm was developed assuming a convex landscape with respect to the parameters θ, where improvement means getting closer to the local (and global) minimum θ^*. In practice, when dealing with neural networks the landscape is highly rugged, with many local minima with very similar performance. When we train on the resampled data, the landscape changes and stochastic gradient descent will carry the optimization to a $\hat{\theta}^L$ that is completely different from the optimum $\hat{\theta}$ computed on the original data. Then all the optimization we computed on $\nabla \log P(A_i|X_i;\hat{\theta})$ to get us closer to θ^* is rather pointless, and the only factor that matters is $w(A_i)$. One could say that this can be mitigated by not re-initializing the networks when we train on the resampled data, but actually performing the experiment shows that this doesn't change anything. Indeed, even if we start at $\hat{\theta}$, the gradient computed on the new data will likely kick us into a new local minimum.

Of course, the effort put into developing our algorithm is not wasted, as when all the hypotheses are satisfied, it is in fact the best theoretical optimum. For example, if $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are linear in x, then the landscape is convex, with a single global minimum. And we have seen from chapter 3 that linear models can be surprisingly powerful when applied to high dimensional climate datasets. Moreover, the theoretical tools that we used to develop this algorithm give us important insights into the nature of the problem, for instance on the matter of the bias introduced at the moment of resampling.

6.6 Ongoing research and future work

This project is in its early stages, and at the moment there are still theoretical and practical kinks that need to be addressed. The code itself is still in the development phase, so I am not absolutely certain that there are no bugs lurking between the lines. Indeed, despite involving eight people for roughly a year, it was for all of us a side quest to be done in the free time between other projects, so progress was quite slow.

The final goal of applying optimal importance sampling to climate models sits then reasonably far in the future, but interesting intermediate steps are going to happen soon. In particular, Tony Lelièvre and Julien Reygner will continue on the theoretical side, while Clément Le Priol is already experimenting on climate models, skipping the phase of improving our estimate of the committor function with importance sampling and directly running the GKLT algorithm using as score function the heatwave amplitude predicted by the Gaussian approximation (see chapter 3). This is much better than persistence and has the potential of being good enough to let us sample very extreme short heatwaves.

In the end, the results obtained so far, although not groundbreaking, are fundamental stepping stones to our goal, and, though the road is long, we can see that we are walking in the right direction.

Chapter 7

Conclusions and Perspectives

Committor functions are the proper tool for any probabilistic forecast, but prove particularly relevant when dealing with extreme events, as they are also a fundamental ingredient for efficiently sampling them, through the use of rare event algorithms. However, computing committor functions is a hard task: they are very high dimensional objects that are difficult to interpret and require prohibitive amounts of data to be computed with sufficient precision.

In this work, we attacked the problem of interpretability developing a framework for explaining committor functions based on their optimal projection onto a much lower dimensional space, which approximates the committor as the composition of a linear projection and a non-linear activation function. When the input X of the committor function is a stack of climate variables, both components are interpretable. Indeed, the coefficients of the linear projection can be easily plotted as one or more maps and the coordinates in the projected space have the clear meaning of the correlation of the input with those maps. Moreover, the non-linear activation function will also be interpretable, because, being low dimensional, we can easily visualize it.

At first, in chapter 2 we used this method as a post-hoc explainability tool for the committor computed with deep Convolutional Neural Networks (CNNs), which allowed us to shed some light on the most important features that the network picked up from the data. More precisely, by projecting on a single grid point, we were able to easily quantify its predictive power if used alone for the prediction task. Although the theoretical framework provides a way to compute the *optimal* projection pattern, actually finding one that involves all grid points proved very impractical. For this reason, we further expanded on the concept of optimal projection, translating it into the architecture of the Intrinsically Interpretable Neural Network (IINN), which allowed us to leverage standard machine learning techniques to learn the optimal projection patters, as well as the non-linear activation function, directly from data. The result is a network which gives us a committor that is automatically interpretable.

With the addition of a more theoretical reasoning, in chapter 3 we used the assumption of joint Gaussianity between predictors X and extreme value observable (in our case the

heatwave amplitude A) to produce the simplest non-trivial form for the committor function. The success of this Gaussian approximation (GA) was remarkable. Indeed, the performance of this new method, although in general not as good as that of CNNs when many centuries of training data are available, was still very good, especially for longer lasting heatwaves and for predictions made two weeks or more in advance. In fact, there was barely any difference in skill between the two methods in this regime, which we argue is probably the most relevant for anticipating coming heatwaves, as for shorter delays standard numerical weather prediction models already perform very well.

In the previous paragraph we used the phrase 'when many centuries of training data are available'. This is rather uncommon in the weather and climate community, which makes the Gaussian approximation even more relevant. Indeed, due to its simplicity, this method is more robust against the lack of data issue, and when only a few decades of data are available (which is the case for reanalysis datasets), it utterly outperforms CNNs, extending the predictability horizon of extreme heatwaves.

Furthermore, the inherent interpretability of the Gaussian approximation means that we can immediately identify the physical drivers of the extreme event under study. For the problem of heatwaves over France we highlighted the importance of low soil moisture, followed by the presence of a strong anticyclone over Western Europe, with a Rossby wave-train extending into the Atlantic. This result was not a surprise, as it agrees with the general understanding of the dynamics of mid-latitude heatwaves presented in chapter 1 and in Perkins [2015]; Barriopedro et al. [2023]. However, our method enables us to turn a general qualitative description into a precise quantitative assessment of the contributions of each of these factors.

In chapters 3 and 4 we showed that, purely based on performance, CNNs appear to be able to capture useful information that goes beyond the Gaussian approximation. However, when we applied state-of-the-art explainability methods to pinpoint what exactly this extra information was, we weren't able to identify it, finding instead that the part of the network that we are able to explain is remarkably similar to the Gaussian approximation. The proper way to go beyond the Gaussian approximation proved to be the use of more complex, but still inherently interpretable, architectures. In particular, we had great success with scattering networks (ScatNets), which are based on a wavelet transform of the input data and showed the same skill as CNNs.

With this method we found that the *extra* information comes from oscillations in the geopotential height field at a sub-synoptic scale (roughly 300 to 500 km), mainly located over Europe and the North Atlantic. Moreover, we found that these oscillations, depending on their spatial orientation, can have different effects on the forecasted heatwave amplitude. For instance meridional oscillations over the North Atlantic are linked with more severe heatwaves while zonal oscillations over the same region with milder ones. I think that this preliminary finding has a very high potential to lead to the discovery of new physics. In any case, besides identifying them qualitatively, we were able to quantify that these oscillations account for 35% of the total information used by the ScatNet. The remaining

65% was already highlighted by the Gaussian approximation, which once again confirms the impressive power of very simple methods.

One of the main reasons for the great success of the Gaussian approximation at the task of predicting extreme heatwaves, is that we focused on events which are spatially averaged over the considerably large region of all of France, and also involve a time average of two weeks. This contributes to make the statistics of the heatwave amplitude very close to Gaussian, allowing us to learn in the bulk of the distribution and successfully extrapolate to the (near) tails. However, this may not be relevant for all types of extreme events. For instance, in the study of the collapse of the Atlantic meridional overturning circulation (AMOC) we look for yet unobserved transitions to a different attractor, which means the system has to venture very far from its typical state. Then, extrapolation from the available data, that only samples the current state of the circulation, may not be very accurate to study the tipping point. In this case, the only option is to actively run climate models, and rare event algorithms can significantly alleviate the computational burden of sampling such rare events.

For this reason, in chapter 5 we developed a flexible framework that can run the Giardinà-Kurchan-Lecomte-Tailleur (GKLT) rare event algorithm on any climate model. In particular, we applied it to the intermediate complexity Versatile Ocean Simulator (VerOS) to sample extreme weakenings of the AMOC. The model showed a remarkable resilience of the vigorous AMOC state, and though we did manage to tip the model to another attractor, this second attractor still featured a quite strong overturning circulation. Although this result is not as exciting as observing a full collapse, it is still useful to pinpoint the important mechanisms of a noise induced tipping of the AMOC. For instance, we tested the VerOS model either in ocean-only mode or with a very rudimentary atmosphere, tracking only the sea surface temperature. On the other hand, in a similar study [Cini et al., 2024], the authors used exactly the same rare event algorithm but on a model where the atmosphere is represented by a fully fledged General Circulation Model. Our negative result, then, corroborates the finding in Cini et al. [2024] that wind stress anomalies may be the main drivers of noise-induced AMOC tipping.

Another insight that we can glean from the negative results of chapter 5, is the importance of using the proper score function for rare event algorithms, which brings us to the overarching idea of this manuscript. This was to couple machine learning and rare event algorithms, where the former computes the score function in the form of the committor, and the latter provide more data to improve the estimates of the neural networks. Such coupling had already been performed by Dario Lucente in his PhD thesis [Lucente, 2021], but he was missing the last step of the feedback loop, namely checking how the newly generated data would improve the estimate of the committor function. Chapter 6 of this thesis focused precisely on this missing link, where we used an importance sampling scheme, specifically designed to optimally improve the committor function. Since we now know that the Gaussian approximation can be a very powerful tool for estimating committor functions, and that it is already very good with little data, we devised a toy model that

explicitly violates the Gaussian hypothesis, displaying in the tail a very different behavior than the one in the bulk. The results we obtained are very promising, as the theoretical framework we developed was indeed able to optimally resample data and improve the committor estimated by a simple Intrinsically Interpretable Neural Network. However, we also obtained the same improvement with a much simpler heuristic resampling strategy, which suggests that in many cases intuition can still be very valuable. Nevertheless, we wouldn't have been able to draw this conclusion without comparing with the theoretically optimal way to do things.

In any case, the development of this optimal resampling algorithm was a key step for finally closing, at least from a theoretical point of view, the feedback loop between machine learning and rare event algorithms. As this project is ongoing, there are still theoretical and technical aspects to be addressed. For instance, the current version of the algorithm improves the committor, but it also introduces biases in the prediction of mildly extreme events. This bias vanishes if one looks only at very extreme events, but eliminating it completely is a desirable improvement. Once such technical issues are addressed, future theoretical work could then move from using simple importance sampling to more complex rare event algorithms, like GKLT, which is more relevant when running climate models. On a parallel note, methodological work could continue by using machine-learned committor functions to run rare event algorithms in fully fledged climate models, and indeed this direction is currently being explored within the group, where the committor learned with the Gaussian approximation is being used to sample short heatwaves in the PlaSim model.

On a similar note, such an approach may be beneficial also for the study of the collapse of the AMOC. Indeed, as in many other works, we have used the value of the AMOC itself as a proxy for the probability of collapse, with the reasoning that a weaker AMOC is more likely to weaken even further. As we have seen in chapter 5, the presence of many intermediate attractors between the vigorous and collapsed state may mean that such proxy is not good enough, and we should rather compute an approximation of the committor function more explicitly, based, for instance, on the full ocean circulation pattern in the Atlantic. Then, it feels natural to use machine learning techniques to try to learn such a committor, though this might be a very hard task if all the data we have is very close to the initial attractor. Another option is that of using a completely different type of algorithm. For example, my collaborators at the University of Copenhagen recently obtained some interesting results with an edge-tracking approach, similar to that used in Mehling et al. [2024].

The work presented in this manuscript was mainly methodological, and we argue that an important perspective is to use these newly developed tools to actually do physics. For instance, we have seen that the Gaussian approximation highlights stationary Rossby waves with a particular frequency. Do we actually see them in real heatwave events? Or, five to ten days before the heatwave hits, the main source of predictability in the geopotential height field comes from a north-south dipole over the United States: is this a detector of the

position of the jet stream? What is the physical meaning of the sub-synoptic oscillations that are identified by the scattering network? Answering these questions has great potential for the discovery of new physics and for the improvement of our general understanding of extreme heatwaves, which will be ever more valuable as the climate warms. Furthermore, another natural research direction is to apply the Gaussian approximation or scattering networks to other types of prediction problems, in the climate community and beyond, but especially in contexts where data is scarce.

And finally, one general, *transversal*, and highly quotable conclusion that can be drawn from this manuscript is that **simple methods can be surprisingly powerful and complexity is often overrated**. With the ever faster development of Artificial Intelligence, we are incentivized to always use state-of-the-art products, which in the recent years has meant increasingly complex and opaque neural networks. As pointed out in Rudin [2019], this can cause severe problems of trust in the models used, and my work joins the critique to post-hoc explainability methods. On the other hand, developing intrinsically interpretable architectures may be harder for researchers, but has the potential to greatly improve our understanding of the processes underlying our object of study. This field is currently severely underdeveloped in the climate community, with only a few recent studies using interpretable models [e.g. Barnes et al., 2022; Chakraborty et al., 2021]. With the rise of AI weather models harbingered by papers like Lam et al. [2023]; Bi et al. [2023]; Nguyen et al. [2023], the field of weather and climate will soon be flooded with huge amounts of data potentially very hard to understand. I thus argue that work in interpretability may become especially beneficial to the weather and climate community in the coming decades.

Bibliography

Agana, N. A. and Homaifar, A. (2017). A deep learning based approach for long-term drought prediction. In *SoutheastCon 2017*, pages 1–8.

Ajagun-Brauns, J. and Ditlevsen, P. (2023). Investigating the dynamics of cusp bifurcations: A conceptual model for glacial-interglacial cycles. In *EGU General Assembly Conference Abstracts*, pages EGU–14342.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Alexander, L. (2011). Extreme heat rooted in dry soils. *Nature Geoscience*, 4(1):12–13.

Alkhayuon, H., Ashwin, P., Jackson, L. C., Quinn, C., and Wood, R. A. (2019). Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2225):20190051.

Allys, E., Levrier, F., Zhang, S., Colling, C., Blancard, B. R.-S., Boulanger, F., Hennebelle, P., and Mallat, S. (2019). The RWST, a comprehensive statistical description of the non-gaussian structures in the ISM. *Astronomy & Astrophysics*, 629:A115.

Amengual, A., Homar, V., Romero, R., Brooks, H. E., Ramis, C., Gordaliza, M., and Alonso, S. (2014). Projections of heat waves with high impact on human health in Europe. *Global and Planetary Change*, 119:71–84.

Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., Zarka, J., Mallat, S., Andén, J., Belilovsky, E., et al. (2020). Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, 21(60):1–6.

Annan, J. D. and Hargreaves, J. C. (2010). Efficient identification of ocean thermodynamics in a physical/biogeochemical ocean model with an iterative Importance Sampling method. *Ocean Modelling*, 32(3):205–215.

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., and Lenton, T. M. (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science*, 377(6611):eabn7950.

Arto, I., Capellán-Pérez, I., Filatova, T., González-Eguinob, M., Hasselmann, K., Kovalevsky, D. V., Markandya, A., Moghayer, S. M., and Tariku, M. B. (2014). Review of existing literature on methodologies to model non-linearity, thresholds and irreversibility in high-impact climate change events in the presence of environmental tipping points.

Asadollah, S. B. H. S., Khan, N., Sharafati, A., Shahid, S., Chung, E.-S., and Wang, X.-J. (2021). Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. *Stochastic Environmental Research and Risk Assessment*.

Asadollah, S. B. H. S., Sharafati, A., and Shahid, S. (2022). Application of ensemble machine learning model in downscaling and projecting climate variables over different climate regions in Iran. *Environmental Science and Pollution Research*, 29(12):17260–17279.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015a). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015b). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Barnes, E. A., Barnes, R. J., Martin, Z. K., and Rader, J. K. (2022). This Looks Like That There: Interpretable Neural Networks for Image Tasks When Location Matters. *Artificial Intelligence for the Earth Systems*, 1(3).

Barnes, E. A., Mayer, K., Toms, B., Martin, Z., and Gordon, E. (2020). Identifying Opportunities for Skillful Weather Prediction with Interpretable Neural Networks.

Barnier, B. (1998). Forcing the Ocean. In Chassignet, E. P. and Verron, J., editors, *Ocean Modeling and Parameterization*, pages 45–80. Springer Netherlands, Dordrecht.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., and García-Herrera, R. (2011). The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe. *Science*, 332(6026):220–224.

Barriopedro, D., García-Herrera, R., Ordóñez, C., Miralles, D. G., and Salcedo-Sanz, S. (2023). Heat Waves: Physical Understanding and Scientific Challenges. *Reviews of Geophysics*, 61(2):e2022RG000780.

Bastiaansen, R., Dijkstra, H. A., and von der Heydt, A. S. (2022). Fragmented tipping in a spatially heterogeneous world. *Environmental Research Letters*, 17(4):045006.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

Belkacemi, Z., Gkeka, P., Lelièvre, T., and Stoltz, G. (2022). Chasing Collective Variables Using Autoencoders and Biased Trajectories. *Journal of Chemical Theory and Computation*, 18(1):59–78.

Bellomo, K., Meccia, V. L., D'Agostino, R., Fabiano, F., Larson, S. M., von Hardenberg, J., and Corti, S. (2023). Impacts of a weakened AMOC on precipitation over the Euro-Atlantic region in the EC-Earth3 climate model. *Climate Dynamics*, 61(7):3397–3416.

Benedetti, R. (2010). Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211.

Beniston, M. (2012). Is snow in the Alps receding or disappearing? *WIREs Climate Change*, 3(4):349–358.

Benson, D. O. and Dirmeyer, P. A. (2021). Characterizing the Relationship between Temperature and Soil Moisture Extremes and Their Role in the Exacerbation of Heat Waves over the Contiguous United States. *Journal of Climate*, 34(6):2175–2187.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538.

Bochenek, B. and Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere*, 13(2):180.

Bochow, N., Poltronieri, A., Robinson, A., Montoya, M., Rypdal, M., and Boers, N. (2023). Overshooting the critical threshold for the Greenland ice sheet. *Nature*, 622(7983):528–536.

Boers, N., Marwan, N., Barbosa, H. M. J., and Kurths, J. (2017). A deforestation-induced tipping point for the South American monsoon system. *Scientific Reports*, 7(1):41489.

Boers, N. and Rypdal, M. (2021). Critical slowing down suggests that the western Greenland Ice Sheet is close to a tipping point. *Proceedings of the National Academy of Sciences*, 118(21):e2024192118.

Bolhuis, P. G., Dellago, C., Geissler, P. L., and Chandler, D. (2000). Transition path sampling: Throwing ropes over mountains in the dark. *Journal of Physics: Condensed Matter*, 12(8A):A147.

Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., and Höhne, M. M.-C. (2023). Finding the right ai method–a guide for the evaluation and ranking of explainable ai methods in climate science. *arXiv preprint arXiv:2303.00652*.

Bouchama, A. (2004). The 2003 European heat wave. *Intensive Care Medicine*, 30(1):1–3.

Bouchet, F., Rolland, J., and Simonnet, E. (2019a). Rare Event Algorithm Links Transitions in Turbulent Flows with Activated Nucleations. *Physical Review Letters*, 122(7):074502.

Bouchet, F., Rolland, J., and Wouters, J. (2019b). Rare Event Sampling Methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8):080402.

Boulton, C. A., Allison, L. C., and Lenton, T. M. (2014). Early warning signals of Atlantic Meridional Overturning Circulation collapse in a fully coupled climate model. *Nature Communications*, 5(1):5752.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Brier, G. W. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3.

Bruna, J. and Mallat, S. (2013). Invariant Scattering Convolution Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937.

Bugallo, M. F., Martino, L., and Corander, J. (2015). Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535.

Castellana, D., Baars, S., Wubs, F. W., and Dijkstra, H. A. (2019). Transition Probabilities of Noise-induced Transitions of the Atlantic Ocean Circulation. *Scientific Reports*, 9(1):20284.

Cazenave, A. (2006). How Fast Are the Ice Sheets Melting? *Science*, 314(5803):1250–1252.

Ceppi, P. and Nowack, P. (2021). Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, 118(30):e2026290118.

Cérou, F., Delyon, B., Guyader, A., and Rousset, M. (2019). On the Asymptotic Normality of Adaptive Multilevel Splitting. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):1–30.

Cérou, F., Guyader, A., Lelièvre, T., and Pommier, D. (2011). A multiple replica approach to simulate reactive trajectories. *The Journal of Chemical Physics*, 134(5):054108.

Cérou, F., LeGland, F., Del Moral, P., and Lezaud, P. (2005). Limit theorems for the multilevel splitting algorithm in the simulation of rare events. In *Proceedings of the Winter Simulation Conference, 2005.*, page 10 pp.

Chakraborty, D., Başağaoğlu, H., and Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, 170:114498.

Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647.

Chantry, M., Christensen, H., Dueben, P., and Palmer, T. (2021). Opportunities and challenges for machine learning in weather and climate modelling: Hard, medium and soft AI. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200083.

Charney, J. G. and DeVore, J. G. (1979). Multiple flow equilibria in the atmosphere and blocking. *Journal of the atmospheric sciences*, 36(7):1205–1216.

Chattopadhyay, A., Nabizadeh, E., and Hassanzadeh, P. (2020). Analog Forecasting of Extreme-Causing Weather Patterns Using Deep Learning. *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001958.

Cheng, S., Morel, R., Allys, E., Ménard, B., and Mallat, S. (2024). Scattering spectra models for physics. *PNAS nexus*, 3(4):103.

Chraibi, H., Dutfoy, A., Galtier, T., and Garnier, J. (2020). Optimal input potential functions in the interacting particle system method. *arXiv:1811.10450 [math, stat]*.

Cimatoribus, A. A., Drijfhout, S. S., and Dijkstra, H. A. (2014). Meridional overturning circulation: Stability and ocean feedbacks in a box model. *Climate Dynamics*, 42(1):311–328.

Cini, M., Zappa, G., Ragone, F., and Corti, S. (2024). Simulating AMOC tipping driven by internal climate variability with a rare event algorithm. *npj Climate and Atmospheric Science*, 7(1):1–10.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme australian wildfires. *Spatial Statistics*, page 100811.

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M.-F., Weber, S., Alcamo, J., Alexeev, V., Berger, A., Calov, R., Ganopolski, A., Goosse, H., Lohmann, G., Lunkeit, F., Mokhov, I., Petoukhov, V., Stone, P., and Wang, Z. (2002). Earth system models of intermediate complexity: Closing the gap in the spectrum of climate system models. *Climate Dynamics*, 18(7):579–586.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An Introduction to Statistical Modeling of Extreme Values*, volume 208. Springer.

Collins, L., Bradstock, R. A., Clarke, H., Clarke, M. F., Nolan, R. H., and Penman, T. D. (2021). The 2019/2020 mega-fires exposed Australian ecosystems to an unprecedented extent of high-severity fire. *Environmental Research Letters*, 16(4):044029.

Coumou, D. and Rahmstorf, S. (2012). A decade of weather extremes. *Nature Climate Change*, 2(7):491–496.

Cox, P. M., Betts, R. A., Collins, M., Harris, P. P., Huntingford, C., and Jones, C. D. (2004). Amazonian forest dieback under climate-carbon cycle projections for the 21st century. *Theoretical and Applied Climatology*, 78(1):137–156.

D'Andrea, F., Provenzale, A., Vautard, R., and De Noblet-Decoudré, N. (2006). Hot and cool summers: Multiple equilibria of the continental water cycle. *Geophysical Research Letters*, 33(24).

Dansgaard, W., Johnsen, S. J., Clausen, H. B., Dahl-Jensen, D., Gundestrup, N. S., Hammer, C. U., Hvidberg, C. S., Steffensen, J. P., Sveinbjörnsdottir, A. E., Jouzel, J., and Bond, G. (1993). Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, 364(6434):218–220.

Davenport, F. V. and Diffenbaugh, N. S. (2021). Using Machine Learning to Analyze Physical Causes of Climate Change: A Case Study of U.S. Midwest Extreme Precipitation. *Geophysical Research Letters*, 48(15):e2021GL093787.

Delaunay, A. and Christensen, H. M. (2022). Interpretable Deep Learning for Probabilistic MJO Prediction. *Geophysical Research Letters*, 49(16):e2022GL098566.

Demory, M.-E., Berthou, S., Fernández, J., Sørland, S. L., Brogli, R., Roberts, M. J., Beyerle, U., Seddon, J., Haarsma, R., Schär, C., Buonomo, E., Christensen, O. B., Ciarlo, J. M., Fealy, R., Nikulin, G., Peano, D., Putrasahan, D., Roberts, C. D., Senan, R., Steger, C., Teichmann, C., and Vautard, R. (2020). European daily precipitation according to EURO-CORDEX regional climate models (RCMs) and high-resolution global climate models (GCMs) from the High-Resolution Model Intercomparison Project (HighResMIP). *Geoscientific Model Development*, 13(11):5485–5506.

Diks, C., Panchenko, V., and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.

Dikshit, A., Pradhan, B., and Alamri, A. M. (2021). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of The Total Environment*, 755:142638.

Ditlevsen, P. and Ditlevsen, S. (2023). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications*, 14(1):4254.

Ditlevsen, P. D., Andersen, K. K., and Svensson, A. (2007). The DO-climate events are probably noise induced: Statistical investigation of the claimed 1470 years cycle. *Climate of the Past*, 3(1):129–134.

Domeisen, D. I. V., Eltahir, E. A. B., Fischer, E. M., Knutti, R., Perkins-Kirkpatrick, S. E., Schär, C., Seneviratne, S. I., Weisheimer, A., and Wernli, H. (2023). Prediction and projection of heatwaves. *Nature Reviews Earth & Environment*, 4(1):36–50.

Drijfhout, S., Gleeson, E., Dijkstra, H. A., and Livina, V. (2013). Spontaneous abrupt climate change due to an atmospheric blocking–sea-ice–ocean feedback in an unforced climate model simulation. *Proceedings of the National Academy of Sciences*, 110(49):19713–19718.

Durre, I., Wallace, J. M., and Lettenmaier, D. P. (2000). Dependence of Extreme Daily Maximum Temperatures on Antecedent Soil Moisture in the Contiguous United States during Summer. *Journal of Climate*, 13(14):2641–2651.

Egger, J. (1978). Dynamics of Blocking Highs. *Journal of the Atmospheric Sciences*, 35(10):1788–1801.

Eickenberg, M., Exarchakis, G., Hirn, M., Mallat, S., and Thiry, L. (2018). Solid harmonic wavelet scattering for predictions of molecule properties. *The Journal of Chemical Physics*, 148(24).

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631.

Evans, L., Cameron, M. K., and Tiwary, P. (2022). Computing committors via Mahalanobis diffusion maps with enhanced sampling data. *The Journal of Chemical Physics*, 157(21):214107.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

Fang, Y., Chen, H., Lin, Y., Zhao, C., Lin, Y., and Zhou, F. (2021). Classification of Northeast China Cold Vortex Activity Paths in Early Summer Based on K-means Clustering and Their Climate Impact. *Advances in Atmospheric Sciences*, 38(3):400–412.

Finkel, J., Abbot, D. S., and Weare, J. (2020). Path Properties of Atmospheric Transitions: Illustration with a Low-Order Sudden Stratospheric Warming Model. *Journal of the Atmospheric Sciences*, 77(7):2327–2347.

Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., and Weare, J. (2021). Learning Forecasts of Rare Stratospheric Transitions from Short Simulations. *Monthly Weather Review*, 149(11):3647–3669.

Fischer, E. M., Beyerle, U., Bloin-Wibe, L., Gessner, C., Humphrey, V., Lehner, F., Pendergrass, A. G., Sippel, S., Zeder, J., and Knutti, R. (2023). Storylines for unprecedented heatwaves based on ensemble boosting. *Nature Communications*, 14(1):4643.

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., Lüthi, D., and Schär, C. (2007). Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave. *Journal of Climate*, 20(20):5081–5099.

Fouillet, A., Rey, G., Laurent, F., Pavillon, G., Bellec, S., Guihenneuc-Jouyaux, C., Clavel, J., Jougla, E., and Hémon, D. (2006). Excess mortality related to the August 2003 heat wave in France. *International Archives of Occupational and Environmental Health*, 80(1):16–24.

Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., and Lunkeit, F. (2005a). The Planet Simulator, towards a user friendly model. *Meteorologische Zeitschritf*, 14(3):299–304.

Fraedrich, K., Kirk, E., and Lunkeit, F. (2005b). Portable University Model of the Atmosphere (PUMA). *Meteorologische Zeitschritf*, 14(6):735–745.

Gálfi, V. M. and Lucarini, V. (2021). Fingerprinting Heatwaves and Cold Spells and Assessing Their Response to Climate Change Using Large Deviation Theory. *Physical Review Letters*, 127(5):058701.

Gálfi, V. M., Lucarini, V., and Wouters, J. (2019). A large deviation theory-based analysis of heat waves and cold spells in a simplified model of the general circulation of the atmosphere. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(3):033404.

García-Herrera, R., Díaz, J., Trigo, R. M., Luterbacher, J., and Fischer, E. M. (2010). A Review of the European Summer Heat Wave of 2003. *Critical Reviews in Environmental Science and Technology*, 40(4):267–306.

Garnier, J. and Moral, P. D. (2006). Simulations of rare events in fiber optics by interacting particle systems. *Optics Communications*, 267(1):205–214.

Gastineau, G. and Frankignoul, C. (2015). Influence of the North Atlantic SST Variability on the Atmospheric Circulation during the Twentieth Century. *Journal of Climate*, 28(4):1396–1416.

Geissler, P. L. and Chandler, D. (2000). Importance sampling and theory of nonequilibrium solvation dynamics in water. *The Journal of Chemical Physics*, 113(21):9759–9765.

Gessner, C., Fischer, E. M., Beyerle, U., and Knutti, R. (2021). Very Rare Heat Extremes: Quantifying and Understanding Using Ensemble Reinitialization. *Journal of Climate*, 34(16):6619–6634.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W.,

Shebalin, P., Vrac, M., Witt, A., and Zaliapin, I. (2011). Extreme events: Dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3):295–350.

Giardinà, C., Kurchan, J., Lecomte, V., and Tailleur, J. (2011). Simulating Rare Events in Dynamical Processes. *Journal of Statistical Physics*, 145(4):787–811.

Giardinà, C., Kurchan, J., and Peliti, L. (2006). Direct Evaluation of Large-Deviation Functions. *Physical Review Letters*, 96(12):120603.

Giffard-Roisin, S., Yang, M., Charpiat, G., Kumler Bonfanti, C., Kégl, B., and Monteleoni, C. (2020). Tropical Cyclone Track Forecasting Using Fused Deep Learning From Aligned Reanalysis Data. *Frontiers in Big Data*, 3.

Goosse, H., Renssen, H., Selten, F. M., Haarsma, R. J., and Opsteegh, J. D. (2002). Potential causes of abrupt climate events: A numerical study with a three-dimensional climate model. *Geophysical Research Letters*, 29(18):7–1–7–4.

Grassberger, P. (1997). Pruned-enriched Rosenbluth method: Simulations of theta polymers of chain length up to 1 000 000. *Physical Review E*, 56(3):3682–3693.

Grotjahn, R. and Faure, G. (2008). Composite Predictor Maps of Extraordinary Weather Events in the Sacramento, California, Region. *Weather and Forecasting*, 23(3):313–335.

Grundner, A., Beucler, T., Gentine, P., and Eyring, V. (2024). Data-Driven Equation Discovery of a Cloud Cover Parameterization. *Journal of Advances in Modeling Earth Systems*, 16(3):e2023MS003763.

Guckenheimer, J. and Holmes, P. (2013). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, volume 42. Springer Science & Business Media.

Haar, L. V., Elvira, T., and Ochoa, O. (2023). An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 117:105606.

Haeberli, W., Oerlemans, J., Zemp, M., et al. (2019). The future of alpine glaciers and beyond. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press New York.

Hafez, Y. (2017). On the Relationship between Heat Waves over the Western and Central Europe and NAO, SOI, El-Nino 3.4 in Summer 2015. *Journal of Geoscience and Environment Protection*, 5(4):31–45.

Häfner, D., Jacobsen, R. L., Eden, C., Kristensen, M. R. B., Jochum, M., Nuterman, R., and Vinter, B. (2018). Veros v0.1 – a fast and versatile ocean simulator in pure Python. *Geoscientific Model Development*, 11(8):3299–3312.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S. (2008). Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Monthly Weather Review*, 136(7):2608–2619.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572.

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

Hansen, A. B. (2023). *Multistability and Tipping Points in Very High-Dimensional Systems and Mplications for Abrupt Climate Change*. PhD thesis, Københavns Universitet.

Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Haugen, M. A., Stein, M. L., Moyer, E. J., and Sriver, R. L. (2018). Estimating Changes in Temperature Distributions in a Large Ensemble of Climate Simulations Using Quantile Regression. *Journal of Climate*, 31(20):8573–8588.

Hauser, M., Orth, R., and Seneviratne, S. I. (2016). Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia. *Geophysical Research Letters*, 43(6):2819–2826.

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., and Ebert-Uphoff, I. (2023). Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems*, 2(2).

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85.

Henderson, S. B., McLean, K. E., Lee, M. J., and Kosatsky, T. (2022). Analysis of community deaths during the catastrophic 2021 heat dome. *Environmental Epidemiology*, 6(1):e189.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hirschi, M., Seneviratne, S. I., Alexandrov, V., Boberg, F., Boroneant, C., Christensen, O. B., Formayer, H., Orlowsky, B., and Stepanek, P. (2011). Observational evidence

for soil-moisture impact on hot extremes in southeastern Europe. *Nature Geoscience*, 4(1):17–21.

Holden, P. B., Edwards, N. R., Wolff, E. W., Lang, N. J., Singarayer, J. S., Valdes, P. J., and Stocker, T. F. (2010). Interhemispheric coupling, the West Antarctic Ice Sheet and warm Antarctic interglacials. *Climate of the Past*, 6(4):431–443.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022). Explainable AI Methods - A Brief Overview. In Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W., editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, pages 13–38. Springer International Publishing, Cham.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., and Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12):124007.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S. (2013). The Community Earth System Model: A Framework for Collaborative Research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC (2021a). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA.

IPCC (2021b). Summary for policymakers. In Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1–31. Cambridge University Press, Cambridge, UK and New York, NY, USA.

Jackson, L. C., Alastrué de Asenjo, E., Bellomo, K., Danabasoglu, G., Haak, H., Hu, A., Jungclaus, J., Lee, W., Meccia, V. L., Saenko, O., Shao, A., and Swingedouw, D. (2023). Understanding AMOC stability: The North Atlantic Hosing Model Intercomparison Project. *Geoscientific Model Development*, 16(7):1975–1995.

Jackson, L. C. and Wood, R. A. (2018). Hysteresis and Resilience of the AMOC in an Eddy-Permitting GCM. *Geophysical Research Letters*, 45(16):8547–8556.

Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., and Bouchet, F. (2022). Deep Learning-Based Extreme Heatwave Forecast. *Frontiers in Climate*, 4.

Jacques-Dumas, V., van Westen, R. M., Bouchet, F., and Dijkstra, H. A. (2023). Data-driven methods to estimate the committor function in conceptual ocean models. *Nonlinear Processes in Geophysics*, 30(2):195–216.

Jacques-Dumas, V., van Westen, R. M., and Dijkstra, H. A. (2024). Estimation of AMOC transition probabilities using a machine learning based rare-event algorithm.

Jochum, M. and Eden, C. (2015). The Connection between Southern Ocean Winds, the Atlantic Meridional Overturning Circulation, and Indo-Pacific Upwelling. *Journal of Climate*, 28(23):9250–9257.

Kahn, H. and Harris, T. E. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30.

Kang, P., Trizio, E., and Parrinello, M. (2024). Computing the committor with the committor to study the transition state ensemble. *Nature Computational Science*, 4(6):451–460.

Karoly, D. J. (2009). The recent bushfires and extreme heat wave in southeast Australia. *Bull Aust Meteorol Oceanogr Soc*, 22(1):10–13.

Kasmi, G., Dubus, L., Saint-Drenan, Y.-M., and Blanc, P. (2023). Assessment of the reliablity of a model's decision by generalizing attribution to the wavelet domain. In *XAI in Action: Past, Present, and Future Applications*.

Keellings, D. and Waylen, P. (2014). Increased risk of heat waves in Florida: Characterizing changes in bivariate heat wave risk using extreme value analysis. *Applied Geography*, 46:90–97.

Khan, N., Shahid, S., Ismail, T. B., and Behlil, F. (2021). Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change. *Stochastic Environmental Research and Risk Assessment*, 35(7):1335–1353.

Koh, J., Steinfeld, D., and Martius, O. (2024). Using spatial extreme-value theory with machine learning to model and understand spatially compounding weather extremes.

Kornhuber, K., Coumou, D., Vogel, E., Lesk, C., Donges, J. F., Lehmann, J., and Horton, R. M. (2020). Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nature Climate Change*, 10(1):48–53.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. (2022). The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective.

Lai, C.-Y., Hassanzadeh, P., Sheshadri, A., Sonnewald, M., Ferrari, R., and Balaji, V. (2024). Machine learning for climate physics and simulations.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

Lau, W. K. M. and Kim, K.-M. (2012). The 2010 Pakistan Flood and Russian Heat Wave: Teleconnection of Hydrometeorological Extremes. *Journal of Hydrometeorology*, 13(1):392–403.

Le Priol, C., Monteiro, J. M., and Bouchet, F. (2024). Using rare event algorithms to understand the statistics and dynamics of extreme heatwave seasons in South Asia.

Lenton, T. M. and Ciscar, J.-C. (2013). Integrating tipping points into climate impact assessments. *Climatic Change*, 117(3):585–597.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6):1786–1793.

Lestang, T. (2018). *Numerical Simulation and Rare Events Algorithms for the Study of Extreme Fluctuations of the Drag Force Acting on an Obstacle Immersed in a Turbulent Flow*. Theses, Université de Lyon.

Lestang, T., Ragone, F., Bréhier, C.-E., Herbert, C., and Bouchet, F. (2018). Computing return times or return periods with rare event algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(4):043213.

Li, J., Zhang, C., Zhou, J. T., Fu, H., Xia, S., and Hu, Q. (2021). Deep-lift: Deep label-specific feature learning for image annotation. *IEEE transactions on Cybernetics*, 52(8):7732–7741.

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, M., Yao, Y., Simmonds, I., Luo, D., Zhong, L., and Chen, X. (2020). Collaborative impact of the NAO and atmospheric blocking on European heatwaves, with a focus on the hot summer of 2018. *Environmental Research Letters*, 15(11):114003.

Li, Q., Lin, B., and Ren, W. (2019). Computing committor functions for the study of rare events using deep learning. *The Journal of Chemical Physics*, 151(5):054112.

Li, W. and Ma, A. (2014). Recent developments in methods for identifying reaction coordinates. *Molecular Simulation*, 40(10-11):784–793.

Lin, B. and Ren, W. (2024). Deep Learning Method for Computing Committor Functions with Adaptive Sampling.

Lin, H., Mo, R., and Vitart, F. (2022). The 2021 Western North American Heatwave and Its Subseasonal Predictions. *Geophysical Research Letters*, 49(6):e2021GL097036.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

Liu, W., Fedorov, A. V., Xie, S.-P., and Hu, S. (2020). Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate. *Science Advances*, 6(26):eaaz4876.

Lohmann, J., Dijkstra, H. A., Jochum, M., Lucarini, V., and Ditlevsen, P. D. (2024). Multistability and intermediate tipping of the Atlantic Ocean circulation. *Science Advances*, 10(12):eadi4253.

Lohmann, J. and Ditlevsen, P. D. (2021). Risk of tipping the overturning circulation due to increasing rates of ice melt. *Proceedings of the National Academy of Sciences*, 118(9):e2017989118.

Lohmann, J. and Svensson, A. (2022). Ice core evidence for major volcanic eruptions at the onset of Dansgaard–Oeschger warming events. *Climate of the Past*, 18(9):2021–2043.

Lopez-Gomez, I., McGovern, A., Agrawal, S., and Hickey, J. (2022). Global Extreme Heat Forecasting Using Neural Weather Models.

López Gómez, J., Ogando Martínez, A., Troncoso Pastoriza, F., Febrero Garrido, L., Granada Álvarez, E., and Orosa García, J. A. (2020). Photovoltaic power prediction using artificial neural networks and numerical weather data. *Sustainability*, 12(24):10295.

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

Lorenz, R., Jaeger, E. B., and Seneviratne, S. I. (2010). Persistence of heat waves and its link to soil moisture memory. *Geophysical Research Letters*, 37(9).

Lotka, A. J. (1910). Contribution to the Theory of Periodic Reactions. *The Journal of Physical Chemistry*, 14(3):271–274.

Lovo, A., Herbert, C., and Bouchet, F. (2023). Interpretable probabilistic forecast of extreme heat waves. Technical Report EGU23-14493, Copernicus Meetings.

Lübbecke, J. F., Rodríguez-Fonseca, B., Richter, I., Martín-Rey, M., Losada, T., Polo, I., and Keenlyside, N. S. (2018). Equatorial Atlantic variability—Modes, mechanisms, and global teleconnections. *WIREs Climate Change*, 9(4):e527.

Lucente, D. (2021). *Predicting Probabilities of Climate Extremes from Observations and Dynamics*. Theses, Université de Lyon.

Lucente, D., Duffner, S., Herbert, C., Rolland, J., and Bouchet, F. (2019). Machine learning of committor functions for predicting high impact climate events. In *Proceedings of the 9th International Workshop on Climate Informatics: CI 2019*, Paris. NCAR.

Lucente, D., Herbert, C., and Bouchet, F. (2022a). Committor Functions for Climate Phenomena at the Predictability Margin: The Example of El Niño–Southern Oscillation in the Jin and Timmermann Model. *Journal of the Atmospheric Sciences*, 79(9):2387–2400.

Lucente, D., Rolland, J., Herbert, C., and Bouchet, F. (2022b). Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083201.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Malhi, Y., Aragão, L. E. O. C., Galbraith, D., Huntingford, C., Fisher, R., Zelazowski, P., Sitch, S., McSweeney, C., and Meir, P. (2009). Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest. *Proceedings of the National Academy of Sciences*, 106(49):20610–20615.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I. (2022a). Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience. *Artificial Intelligence for the Earth Systems*, 1(4).

Mamalakis, A., Ebert-Uphoff, I., and Barnes, E. A. (2022b). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, 1:e8.

Manabe, S. (1969). CLIMATE AND THE OCEAN CIRCULATION: I. THE ATMOSPHERIC CIRCULATION AND THE HYDROLOGY OF THE EARTH'S SURFACE. *Monthly Weather Review*, 97(11):739–774.

Martin, Z. K., Barnes, E. A., and Maloney, E. (2022). Using Simple, Explainable Neural Networks to Predict the Madden-Julian Oscillation. *Journal of Advances in Modeling Earth Systems*, 14(5):e2021MS002774.

Mascolo, V., Lovo, A., Herbert, C., and Bouchet, F. (2024a). Gaussian Framework and Optimal Projection of Weather Fields for Prediction of Extreme Events.

Mascolo, V., Priol, C. L., D'Andrea, F., and Bouchet, F. (2024b). Compared influence of the Atlantic Multidecadal Variability and of spring soil moisture on summer heat waves in Europe.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

McEwen, J. D., Wallis, C. G., and Mavor-Parker, A. N. (2021). Scattering networks on the sphere for scalable and rotationally equivariant spherical cnns. *arXiv preprint arXiv:2102.02828*.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199.

Meehl, G. A. and Tebaldi, C. (2004). More Intense, More Frequent, and Longer Lasting Heat Waves in the 21st Century. *Science*, 305(5686):994–997.

Mehling, O., Börner, R., and Lucarini, V. (2024). Limits to predictability of the asymptotic state of the Atlantic Meridional Overturning Circulation in a conceptual climate model. *Physica D: Nonlinear Phenomena*, 459:134043.

Menviel, L. C., Skinner, L. C., Tarasov, L., and Tzedakis, P. C. (2020). An ice–climate oscillatory framework for Dansgaard–Oeschger cycles. *Nature Reviews Earth & Environment*, 1(12):677–693.

Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., and Bouchet, F. (2023a). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Physical Review Fluids*, 8(4):040501.

Miloshevich, G., Lucente, D., Yiou, P., and Bouchet, F. (2023b). Extreme heatwave sampling and prediction with analog Markov chain and comparisons with deep learning.

Miloshevich, G., Rouby-Poizat, P., Ragone, F., and Bouchet, F. (2023c). Robust intra-model teleconnection patterns for extreme heatwaves. *Frontiers in Earth Science*, 11:1235579.

Miralles, D. G., van den Berg, M. J., Teuling, A. J., and de Jeu, R. a. M. (2012). Soil moisture-temperature coupling: A multiscale observational analysis. *Geophysical Research Letters*, 39(21).

Miron, P., Beron-Vera, F. J., Helfmann, L., and Koltai, P. (2021). Transition paths of marine debris and the stability of the garbage patches. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(3):033101.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Moral, P. D. and Garnier, J. (2005). Genealogical particle analysis of rare events. *The Annals of Applied Probability*, 15(4):2496–2534.

Morcego, B., Yin, W., Boersma, S., van Henten, E., Puig, V., and Sun, C. (2023). Reinforcement Learning versus Model Predictive Control on greenhouse climate control. *Computers and Electronics in Agriculture*, 215:108372.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Nadeau, L.-P., Ferrari, R., and Jansen, M. F. (2019). Antarctic Sea Ice Control on the Depth of North Atlantic Deep Water. *Journal of Climate*, 32(9):2537–2551.

Nemoto, T., Bouchet, F., Jack, R. L., and Lecomte, V. (2016). Population-dynamics method with a multicanonical feedback control. *Physical Review E*, 93(6):062123.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. (2023). ClimaX: A foundation model for weather and climate.

Nielsen, S. B., Jochum, M., Pedro, J. B., Eden, C., and Nuterman, R. (2019). Two-Timescale Carbon Cycle Response to an AMOC Collapse. *Paleoceanography and Paleoclimatology*, 34(4):511–523.

Nikurashin, M. and Vallis, G. (2011). A Theory of Deep Stratification and Overturning Circulation in the Ocean. *Journal of Physical Oceanography*, 41(3):485–502.

Nikurashin, M. and Vallis, G. (2012). A Theory of the Interhemispheric Meridional Overturning Circulation and Associated Stratification. *Journal of Physical Oceanography*, 42(10):1652–1667.

Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016.

Noyelle, R., Yiou, P., and Faranda, D. (2024). Investigating the typicality of the dynamics leading to extreme temperatures in the IPSL-CM6A-LR model. *Climate Dynamics*, 62(2):1329–1357.

O'Keeffe, P. E. and Wieczorek, S. (2020). Tipping Phenomena and Points of No Return in Ecosystems: Beyond Classical Bifurcations. *SIAM Journal on Applied Dynamical Systems*, 19(4):2371–2402.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature Visualization. *Distill*, 2(11):e7.

Onsager, L. (1938). Initial Recombination of Ions. *Physical Review*, 54(8):554–557.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward,

J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). GPT-4 Technical Report.

Oyallon, E., Zagoruyko, S., Huang, G., Komodakis, N., Lacoste-Julien, S., Blaschko, M., and Belilovsky, E. (2019). Scattering Networks for Hybrid Representation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2208–2221.

Paananen, T., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):16.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., and Lee, J. (2022). Improving Seasonal Forecast Using Probabilistic Deep Learning. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002766.

Pelenis, J. (2014). Weighted scoring rules for comparison of density forecasts on subsets of interest.

Perkins, S. E. (2015). A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale. *Atmospheric Research*, 164–165:242–267.

Petersik, P. J. and Dijkstra, H. A. (2020). Probabilistic Forecasting of El Niño Using Neural Network Models. *Geophysical Research Letters*, 47(6):e2019GL086423.

Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119–131.

Porfiriev, B. (2014). Evaluation of human losses from disasters: The case of the 2010 heat waves and forest fires in Russia. *International Journal of Disaster Risk Reduction*, 7:91–99.

Prowse, T. D., Wrona, F. J., Reist, J. D., Gibson, J. J., Hobbie, J. E., Lévesque, L. M. J., and Vincent, W. F. (2006). Climate Change Effects on Hydroecology of Arctic Freshwater Ecosystems. *AMBIO: A Journal of the Human Environment*, 35(7):347–358.

Quesada, B., Vautard, R., Yiou, P., Hirschi, M., and Seneviratne, S. I. (2012). Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nature Climate Change*, 2(10):736–741.

Racah, E., Beckham, C., Maharaj, T., Ebrahimi Kahou, S., Prabhat, Mr., and Pal, C. (2017). ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ragone, F. and Bouchet, F. (2020). Computation of Extreme Values of Time Averaged Observables in Climate Models with Large Deviation Techniques. *Journal of Statistical Physics*, 179(5):1637–1665.

Ragone, F. and Bouchet, F. (2021). Rare Event Algorithm Study of Extreme Warm Summers and Heatwaves Over Europe. *Geophysical Research Letters*, 48(12):e2020GL091197.

Ragone, F., Wouters, J., and Bouchet, F. (2018). Computation of extreme heat waves in climate models using a large deviation algorithm. *Proceedings of the National Academy of Sciences*, 115(1):24–29.

Rahmstorf, S. (2002). Ocean circulation and climate during the past 120,000 years. *Nature*, 419(6903):207–214.

Rahmstorf, S., Crucifix, M., Ganopolski, A., Goosse, H., Kamenkovich, I., Knutti, R., Lohmann, G., Marsh, R., Mysak, L. A., Wang, Z., and Weaver, A. J. (2005). Thermohaline circulation hysteresis: A model intercomparison. *Geophysical Research Letters*, 32(23).

Ratnam, J. V., Behera, S. K., Ratna, S. B., Rajeevan, M., and Yamagata, T. (2016). Anatomy of Indian heatwaves. *Scientific Reports*, 6(1):24395.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677.

Ren, W., Vanden-Eijnden, E., Maragakis, P., and E, W. (2005). Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *The Journal of Chemical Physics*, 123(13):134109.

Rennermalm, A. K., Wood, E. F., Déry, S. J., Weaver, A. J., and Eby, M. (2006). Sensitivity of the thermohaline circulation to Arctic Ocean runoff. *Geophysical Research Letters*, 33(12).

Rietkerk, M., Bastiaansen, R., Banerjee, S., van de Koppel, J., Baudena, M., and Doelman, A. (2021). Evasion of tipping in complex systems through spatial pattern formation. *Science*, 374(6564):eabj0359.

Ritchie, P. D. L., Clarke, J. J., Cox, P. M., and Huntingford, C. (2021). Overshooting tipping point thresholds in a changing climate. *Nature*, 592(7855):517–523.

Robine, J.-M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J.-P., and Herrmann, F. R. (2008). Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes Rendus Biologies*, 331(2):171–178.

Rohde, R. A. and Hausfather, Z. (2020). The Berkeley Earth Land/Ocean Temperature Record. *Earth System Science Data*, 12(4):3469–3479.

Rolland, J., Bouchet, F., and Simonnet, E. (2016). Computing Transition Rates for the 1-D Stochastic Ginzburg–Landau–Allen–Cahn Equation for Finite-Amplitude Noise with a Rare Event Algorithm. *Journal of Statistical Physics*, 162(2):277–311.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. (2022). Tackling Climate Change with Machine Learning. *ACM Comput. Surv.*, 55(2):42:1–42:96.

Rotskoff, G. M., Mitchell, A. R., and Vanden-Eijnden, E. (2022). Active Importance Sampling for Variational Objectives Dominated by Rare Events: Consequences for Optimization and Generalization. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 757–780. PMLR.

Rousi, E., Kornhuber, K., Beobide-Arsuaga, G., Luo, F., and Coumou, D. (2022). Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. *Nature communications*, 13(1):3851.

Rowntree, P. R. and Bolton, J. A. (1983). Simulation of the atmospheric response to soil moisture anomalies over Europe. *Quarterly Journal of the Royal Meteorological Society*, 109(461):501–526.

Rubino, G. and Tuffin, B. (2009). *Rare Event Simulation Using Monte Carlo Methods*. John Wiley & Sons.

Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo Method*. John Wiley & Sons.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rugenstein, M., Bloch-Johnson, J., Abe-Ouchi, A., Andrews, T., Beyerle, U., Cao, L., Chadha, T., Danabasoglu, G., Dufresne, J.-L., Duan, L., Foujols, M.-A., Frölicher, T., Geoffroy, O., Gregory, J., Knutti, R., Li, C., Marzocchi, A., Mauritsen, T., Menary, M., Moyer, E., Nazarenko, L., Paynter, D., Saint-Martin, D., Schmidt, G. A., Yamamoto, A., and Yang, S. (2019). LongRunMIP: Motivation and Design for a Large Collection of Millennial-Length AOGCM Simulations. *Bulletin of the American Meteorological Society*, 100(12):2551–2570.

Russill, C. (2015). Climate change tipping points: Origins, precursors, and debates. *WIREs Climate Change*, 6(4):427–434.

Sachindra, D. A., Ahmed, K., Rashid, M. M., Shahid, S., and Perera, B. J. C. (2018). Statistical downscaling of precipitation using machine learning techniques. *Atmospheric Research*, 212:240–258.

Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., Parker, D. E., and Taylor, K. E. (2000). Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *Journal of Geophysical Research: Atmospheres*, 105(D6):7337–7356.

Sauer, J., Demaeyer, J., Zappa, G., Massonnet, F., and Ragone, F. (2024). Extremes of summer Arctic sea ice reduction investigated with a rare event algorithm. *Climate Dynamics*.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C. (2004). The role of increasing temperature variability in European summer heatwaves. *Nature*, 427(6972):332–336.

Schneider, R., Bonavita, M., Geer, A., Arcucci, R., Dueben, P., Vitolo, C., Le Saux, B., Demir, B., and Mathieu, P.-P. (2022). ESA-ECMWF Report on recent progress and research directions in machine learning for Earth System observation and prediction. *npj Climate and Atmospheric Science*, 5(1):1–5.

Schubert, S. D., Wang, H., Koster, R. D., Suarez, M. J., and Groisman, P. Y. (2014). Northern Eurasian Heat Waves and Droughts. *Journal of Climate*, 27(9):3169–3207.

Schumacher, D. L., Hauser, M., and Seneviratne, S. I. (2022). Drivers and Mechanisms of the 2021 Pacific Northwest Heatwave. *Earth's Future*, 10(12):e2022EF002967.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., Zhang, X., Alexander, L. V., Allen, S., Benito, G., Cavazos, T., Clague, J., Conway, D., Della-Marta, P. M., Gerber, M., Gong, S., Goswami, B. N., Hemer, M., Huggel, C., van den Hurk, B., Kharin, V. V., Kitoh, A., Klein Tank, A. M. G., Li, G., Mason, S. J., McGuire, W., van Oldenborgh, G. J., Orlowsky, B., Smith, S., Thiaw, W., Velegrakis, A., Yiou, P., Zhang, T., Zhou, T., and Zwiers, F. W. (2012). Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*, 3:109–230.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C. (2006). Land-atmosphere coupling and climate change in Europe. *Nature*, 443(7108):205–209.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Vicente-Serrano, S. M., Wehner, M., and Zhou, B. (2021). Weather and climate extreme events in a changing climate. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766.

Shepherd, T. G. (2016). A Common Framework for Approaches to Extreme Event Attribution. *Current Climate Change Reports*, 2(1):28–38.

Shukla, J. and Mintz, Y. (1982). Influence of Land-Surface Evapotranspiration on the Earth's Climate. *Science*, 215(4539):1498–1501.

Sillmann, J. and Croci-Maspoli, M. (2009). Present and future atmospheric blocking and its impact on European mean and extreme climate. *Geophysical Research Letters*, 36(10).

Simonnet, E., Rolland, J., and Bouchet, F. (2021). Multistability and Rare Spontaneous Transitions in Barotropic β-Plane Turbulence. *Journal of the Atmospheric Sciences*, 78(6):1889–1911.

Smeed, D. A., McCarthy, G. D., Cunningham, S. A., Frajka-Williams, E., Rayner, D., Johns, W. E., Meinen, C. S., Baringer, M. O., Moat, B. I., Duchez, A., and Bryden, H. L. (2014). Observed decline of the Atlantic meridional overturning circulation 2004–2012. *Ocean Science*, 10(1):29–38.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Srokosz, M., Danabasoglu, G., and Patterson, M. (2021). Atlantic Meridional Overturning Circulation: Reviews of Observational and Modeling Advances—An Introduction. *Journal of Geophysical Research: Oceans*, 126(1):e2020JC016745.

Stefanon, M., D'Andrea, F., and Drobinski, P. (2012). Heatwave classification over Europe and the Mediterranean region. *Environmental Research Letters*, 7(1):014023.

Stéfanon, M., Drobinski, P., D'Andrea, F., Lebeaupin-Brossier, C., and Bastin, S. (2014). Soil moisture-temperature feedbacks at meso-scale during summer heat waves over Western Europe. *Climate Dynamics*, 42(5):1309–1324.

Stevens-Rumann, C. S., Kemp, K. B., Higuera, P. E., Harvey, B. J., Rother, M. T., Donato, D. C., Morgan, P., and Veblen, T. T. (2018). Evidence for declining forest resilience to wildfires under climate change. *Ecology Letters*, 21(2):243–252.

Stommel, H. (1961). Thermohaline Convection with Two Stable Regimes of Flow. *Tellus*, 13(2):224–230.

Strahan, J., Antoszewski, A., Lorpaiboon, C., Vani, B. P., Weare, J., and Dinner, A. R. (2021). Long-Time-Scale Predictions from Short-Trajectory Data: A Benchmark Analysis

of the Trp-Cage Miniprotein. *Journal of Chemical Theory and Computation*, 17(5):2948–2963.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Tailleur, J. and Kurchan, J. (2007). Probing rare physical trajectories with Lyapunov weighted dynamics. *Nature Physics*, 3(3):203–207.

ten Wolde, P. R. and Frenkel, D. (1997). Enhancement of Protein Crystal Nucleation by Critical Density Fluctuations. *Science*, 277(5334):1975–1978.

Teng, H., Branstator, G., Wang, H., Meehl, G. A., and Washington, W. M. (2013). Probability of US heat waves affected by a subseasonal planetary wave pattern. *Nature Geoscience*, 6(12):1056–1061.

Thiede, E. H., Giannakis, D., Dinner, A. R., and Weare, J. (2019). Galerkin approximation of dynamical quantities using trajectory data. *The Journal of Chemical Physics*, 150(24):244111.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: A review. *WIREs Computational Statistics*, 2(1):54–60.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002.

Toms, B. A., Barnes, E. A., and Hurrell, J. W. (2021). Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks. *Geophysical Research Letters*, 48(12):e2021GL093842.

Tripathy, K. P., Mukherjee, S., Mishra, A. K., Mann, M. E., and Williams, A. P. (2023). Climate change will accelerate the high-end risk of compound drought and heatwave events. *Proceedings of the National Academy of Sciences*, 120(28):e2219825120.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. a. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

Vallis, G. K. (2017). *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press.

van Nes, E. H., Arani, B. M. S., Staal, A., van der Bolt, B., Flores, B. M., Bathiany, S., and Scheffer, M. (2016). What Do You Mean, 'Tipping Point'? *Trends in Ecology & Evolution*, 31(12):902–904.

van Westen, R. M. and Dijkstra, H. A. (2023). Asymmetry of AMOC Hysteresis in a State-of-the-Art Global Climate Model.

van Westen, R. M., Jacques-Dumas, V., Boot, A. A., and Dijkstra, H. A. (2024a). The Role of Sea-ice Processes on the Probability of AMOC Transitions.

van Westen, R. M., Kliphuis, M., and Dijkstra, H. A. (2024b). Physics-based early warning signal shows that AMOC is on tipping course. *Science Advances*, 10(6):eadk1189.

Vargas Zeppetello, L. R. and Battisti, D. S. (2020). Projected Increases in Monthly Midlatitude Summertime Temperature Variance Over Land Are Driven by Local Thermodynamics. *Geophysical Research Letters*, 47(19):e2020GL090197.

Vautard, R., Yiou, P., D'Andrea, F., de Noblet, N., Viovy, N., Cassou, C., Polcher, J., Ciais, P., Kageyama, M., and Fan, Y. (2007). Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit. *Geophysical Research Letters*, 34(7).

Vettoretti, G., Ditlevsen, P., Jochum, M., and Rasmussen, S. O. (2022). Atmospheric CO2 control of spontaneous millennial-scale ice age climate oscillations. *Nature Geoscience*, 15(4):300–306.

Vial, J., Dufresne, J.-L., and Bony, S. (2013). On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*, 41(11):3339–3362.

Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 131(818):518–532.

Watson, P. A. G. (2022). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17(11):111004.

Webb, D. J., Spence, P., Holmes, R. M., and England, M. H. (2021). Planetary-Wave-Induced Strengthening of the AMOC Forced by Poleward Intensified Southern Hemisphere Westerly Winds. *Journal of Climate*, 34(17):7073–7090.

Webber, R. J., Plotkin, D. A., O'Neill, M. E., Abbot, D. S., and Weare, J. (2019). Practical rare event sampling for extreme mesoscale weather. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(5):053109.

Weber, S. L. (2010). The utility of Earth system Models of Intermediate Complexity (EMICs). *WIREs Climate Change*, 1(2):243–252.

Wehner, M., Oliker, L., and Shalf, J. (2008). Towards Ultra-High Resolution Models of Climate and Weather. *The International Journal of High Performance Computing Applications*, 22(2):149–165.

Weijer, W., Cheng, W., Drijfhout, S. S., Fedorov, A. V., Hu, A., Jackson, L. C., Liu, W., McDonagh, E. L., Mecking, J. V., and Zhang, J. (2019). Stability of the Atlantic Meridional Overturning Circulation: A Review and Synthesis. *Journal of Geophysical Research: Oceans*, 124(8):5336–5375.

Weinan, E., Ren, W., and Vanden-Eijnden, E. (2005). Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chemical Physics Letters*, 413(1):242–247.

Williams, P. D., Alexander, M. J., Barnes, E. A., Butler, A. H., Davies, H. C., Garfinkel, C. I., Kushnir, Y., Lane, T. P., Lundquist, J. K., Martius, O., Maue, R. N., Peltier, W. R., Sato, K., Scaife, A. A., and Zhang, C. (2017). A Census of Atmospheric Variability From Seconds to Decades. *Geophysical Research Letters*, 44(21):11,201–11,211.

Woollings, T. (2010). Dynamical influences on European climate: An uncertain future. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1924):3733–3756.

Wouters, J. and Bouchet, F. (2016). Rare event computation in deterministic chaotic systems using genealogical particle analysis. *Journal of Physics A: Mathematical and Theoretical*, 49(37):374002.

Wouters, J., Schiemann, R. K. H., and Shaffrey, L. C. (2023). Rare Event Simulation of Extreme European Winter Rainfall in an Intermediate Complexity Climate Model. *Journal of Advances in Modeling Earth Systems*, 15(4):e2022MS003537.

Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., and Xiong, H. (2024). Interpretable Machine Learning for Weather and Climate Prediction: A Survey.

Yiou, P. (2014). AnaWEGE: A weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7(2):531–543.

Yiou, P., Cadiou, C., Faranda, D., Jézéquel, A., Malhomme, N., Miloshevich, G., Noyelle, R., Pons, F., Robin, Y., and Vrac, M. (2023). Ensembles of climate simulations to anticipate worst case heatwaves during the Paris 2024 Olympics. *npj Climate and Atmospheric Science*, 6(1):1–8.

Yiou, P., Salameh, T., Drobinski, P., Menut, L., Vautard, R., and Vrac, M. (2013). Ensemble reconstruction of the atmospheric column from surface pressure using analogues. *Climate Dynamics*, 41(5):1333–1344.

Yule, G. U. (1927). VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal*

Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 226(636-646):267–298.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Zeppetello, L. R. V., Battisti, D. S., and Baker, M. B. (2022). The Physics of Heat Waves: What Causes Extremely High Summertime Temperatures? *Journal of Climate*, 35(7):2231–2251.

Zhang, H., Finkel, J., Abbot, D. S., Gerber, E. P., and Weare, J. (2024). Using Explainable AI and Transfer Learning to understand and predict the maintenance of Atlantic blocking with limited observational data.

Zhang, W., Quan, H., and Srinivasan, D. (2018). An improved quantile regression neural network for probabilistic load forecasting. *IEEE Transactions on Smart Grid*, 10(4):4425–4434.

Zhang, Z. (2018). Improved Adam Optimizer for Deep Neural Networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2.

Zhou, S., Williams, A. P., Berg, A. M., Cook, B. I., Zhang, Y., Hagemann, S., Lorenz, R., Seneviratne, S. I., and Gentine, P. (2019). Land–atmosphere feedbacks exacerbate concurrent soil drought and atmospheric aridity. *Proceedings of the National Academy of Sciences*, 116(38):18848–18853.

Zhou, Y. and Wu, Z. (2016). Possible impacts of mega-El Niño/Southern Oscillation and Atlantic Multidecadal Oscillation on Eurasian heatwave frequency variability. *Quarterly Journal of the Royal Meteorological Society*, 142(697):1647–1661.

Zhou, Z.-c., Wu, Z., and Jin, T. (2021). Deep reinforcement learning framework for resilience enhancement of distribution systems under extreme weather events. *International Journal of Electrical Power & Energy Systems*, 128:106676.

Zinovjev, K. and Tuñón, I. (2014). Exploring chemical reactivity of complex systems with path-based coordinates: Role of the distance metric. *Journal of Computational Chemistry*, 35(23):1672–1681.

Zschenderlein, P., Fink, A. H., Pfahl, S., and Wernli, H. (2019). Processes determining heat waves across different European climates. *Quarterly Journal of the Royal Meteorological Society*, 145(724):2973–2989.

List of Tables

List of Figures

List of Acronyms

AGU American Geophysical Union. 48

AI Artificial Intelligence. vii, 13, 16, 19, 20, 207

AIES Artificial Intelligence for the Earth Systems. 108

AIFS Artificial Intelligence Integrated Forecasting System. 20

AIS Adaptive Importance Sampling. 21, 179

AMC analogue Markov chain. 12, 22, 25, 110, 190

AMOC Atlantic meridional overturning circulation. v, x, 2, 6, 8, 17, 22, 23, 26, 141–152, 154, 156, 158, 160–168, 170–176, 179, 205, 206, 240, 241

AMS American Meteorological Society. 108

AMS Adaptive Multilevel Splitting. 23, 24, 150, 239

AMV Atlantic Multidecadal Variability. 5

BCE Binary Cross Entropy. 113, 118

BCES Binary Cross Entropy Skill. 116–118, 127

BM Block Maxima. 9

CESM Community Earth System Model. 29, 52, 91, 108, 111, 132

CLT central limit theorem. 21, 190

CMIP Coupled Model Intercomparison Project. 19, 54, 91

CNN Convolutional Neural Network. viii–x, 13, 29, 32, 33, 35–38, 44, 48, 108, 109, 111, 114–125, 127, 130, 131, 133–139, 178, 180, 203, 204, 237, 240

CPU Central Processing Unit. 193

CRPS Continuous Ranked Probability Score. 113, 114, 116, 118

CRPSS Continuous Ranked Probability Skill Score. 117, 118, 127

DNS direct numerical simulation. 12

DO Dansgaard-Oeschger. 143, 144

EAP East Atlantic Pattern. 169

ECMWF European Centre for Medium-Range Weather Forecasts. 19, 20, 55

EGFI Expected Gradient Feature Importance. 121, 122, 126, 133–137, 240

EMIC Earth system Models of Intermediate Complexity. 17

ENSO El Niño Southern Oscillation. 5

EOF Empirical Orthogonal Function. 103–105, 169–171, 240, 241

ESM Earth System Model. 19

EVT Extreme Value Theory. 9, 10, 109, 131

EWS Early Warning Signal. 7, 144, 145

GA Gaussian approximation. viii, ix, 25, 26, 29, 48, 50–52, 54, 56, 58, 60, 62–66, 68, 70–88, 90–92, 94–96, 98–102, 104, 106, 108, 109, 113, 114, 116–119, 121–123, 125–127, 129–136, 139, 200, 204–207, 237, 239, 240

GCM General Circulation Model. 18–20, 32, 144, 145, 205

GEV Generalized Extreme Value. 9

GIS Greenland Ice Sheet. 144, 147

GKLT Giardinà-Kurchan-Lecomte-Tailleur. x, 22, 24, 26, 27, 142, 145, 151, 154–157, 161, 176, 178, 179, 200, 205, 206, 240

GP Generalized Pareto. 9

GPU Graphics Processing Unit. 193

HL H loss. 185, 186, 193, 196, 197, 199, 241

IINN Intrinsically Interpretable Neural Network. viii, ix, 16, 29, 44–46, 108, 109, 114, 116–119, 126, 129, 130, 132, 133, 196, 197, 203, 206, 239–241

IPCC Intergovernmental Panel on Climate Change. 16, 19

IPS Interacting Particle System. 22, 23, 150, 151, 239

IS importance sampling. vii, x, 12, 21, 23, 26, 27, 150, 178–180, 184–187, 189, 191, 200, 205, 206

ITCZ Intertropical Convergence Zone. 143

JAMES Journal of Advances in Modeling Earth Systems. 48

KL Kullback-Leibler. 33, 38–40, 61, 185

LSG Large Scale Geostrophic Ocean. 142, 175

MJO Madden-Julian Oscillation. 46, 129

ML machine learning. i, v, vii, ix, 2, 3, 10, 12–17, 20, 24–27, 29, 43, 51, 54, 56, 60, 91, 108–110, 112, 114, 116–118, 120, 122, 124, 126, 128, 130–132, 134, 136, 138, 140, 155, 176, 178, 180, 203, 205, 206, 239

MSE Mean Square Error. 118

NAO North Atlantic Oscillation. 5

NLL Negative Log Likelihood. 113, 118, 185–187, 191, 199

NLLS Negative Log Likelihood Skill. 117, 118, 127

NN neural network. viii, 2, 14–16, 25, 26, 33, 35, 38, 43, 48–51, 64, 72–76, 81–83, 85, 90, 91, 110, 116, 118, 120, 127, 129, 178–181, 192, 193, 196, 205, 207, 237

OU Ornstein-Uhlenbeck. x, 27, 142, 145, 151, 156, 157, 160, 161, 163, 167, 240, 241

PACF partial autocorrelation function. 169, 170

PDF probability density function. 35

PlaSim Planet Simulator. 29, 32, 52, 54–56, 60, 65, 66, 69, 71, 72, 74–77, 80–92, 101, 108, 142, 175, 180, 206, 237, 239, 240

POT Peak Over Threshold. 9

PRSA Proceedings of the Royal Society A. 142

REA rare event algorithm. i, v, vii, x, 2, 3, 8, 10, 12, 17, 20, 21, 23–27, 110, 141, 142, 144–146, 148, 150–168, 170–172, 174–176, 178, 179, 194, 203, 205, 206, 239–241

ReLU Rectified Linear Unit. 193, 197

S2S sub-seasonal to seasonal. 11, 14

ScatNet scattering network. x, 29, 108, 109, 111, 114–118, 121, 125–131, 133, 137, 140, 204, 207, 237, 240

SGD stochastic gradient descent. 117, 192, 193, 200

SIR Sliced Inverse Regression. 43–45

SIS Sequential Importance Sampling. 21, 22

SISR Sequential Importance Sampling with Resampling. 22

SST sea surface temperature. 32, 111, 144, 146, 148, 168–170, 172, 175, 205

SWG Stochastic Weather Generator. 12, 22, 110

TAMS Trajectory Adaptive Multilevel Splitting. 23

TDAM Two Dimensional Activation Model. 26, 29, 195, 196, 241

UCPH University of Copenhagen. 142, 176, 206

UU Utrecht University. 142

VerOS Versatile Ocean Simulator. x, 27, 29, 142, 145–149, 155, 161–163, 165–171, 174–176, 205, 240, 241

XAI Explainable Artificial Intelligence. 15, 109, 120