


Numéro National de Thèse : 2021LYSEN062

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON
opérée par
l'Ecole Normale Supérieure de Lyon

Ecole Doctorale N° 52
Physique et Astrophysique de Lyon

Discipline : **Physique**

Soutenue publiquement le 09/11/2021, par :
Dario LUCENTE

Predicting probabilities of climate extremes from observations and dynamics

Prédiction des probabilités des extrêmes climatiques à partir des observations et de la dynamique

---

Devant le jury composé de :

HASSANZADEH, Pedram	Professeur assimilé	Rice University	Rapporteur
YIOU, Pascal	Directeur de Recherche	LSCE - CEA Saclay	Rapporteur
ABRY, Patrice	Directeur de Recherche	ENS de Lyon	Examinateur
HARNIK, Nili	Professeur des Universités	Tel Aviv University	Examinatrice
HERBERT, Corentin	CRCN CNRS	ENS de Lyon	Examinateur
LELIÈVRE, Tony	Directeur de Recherche	École des Ponts ParisTech	Examinateur
ROBARDET, Céline	Professeur des Universités	INSA Lyon	Examinatrice
BOUCHET, Freddy	Directeur de Recherche	ENS de Lyon	Directeur de thèse

Acknowledgments

I would like to thank my supervisor Prof. Freddy Bouchet for his invaluable advice, continuous support, and patience during my PhD study. Additionally, I would like to express gratitude to Dr. Corentin Herbert for his treasured support which was really influential in criticizing my results. My gratitude extends to the people I have worked with such as Dr. George Miloshevich and Dr. Joran Rolland, who have allowed me to grow in many scientific aspects. In general, I would like to thank all the team with whom I had the pleasure of sharing both our work and moments of conviviality. I would like also to thank my family, who have been close to me despite the Alps and the pandemic to separate us. Finally, a special thanks goes to my life partner Alice with whom I had the pleasure of sharing this experience. Your closeness has always been a source of comfort especially in these three years. We shared many fantastic moments together and some less beautiful ones, but you have always remained by my side. In the most difficult times, you have been able to spur and encourage me, often giving me the strength that I was not able to find on my own. Without your tremendous support it would have been impossible for me to complete my study. I will never thank you enough for everything you have done and do for me, as well as for all the love you can give me.

My gratitude extends to the "École Normale Supérieure de Lyon" and, more specifically, to the laboratoire de physique for giving me the opportunity to carry out my studies in a welcoming and stimulating environment. My work was funded through the ACADEMICS grant of IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005. Computer time was provided by the "Pôle Scientifique de Modélisation Numérique" and the "Centre Blaise Pascal" in Lyon.

Summary

There is a large interest in predicting the occurrence of high impact climate events such as ENSO (El Niño Southern Oscillation) or rare events, for instance heat waves. Those are *prediction problems at the predictability margin* because the interesting time scale lies at the edge of the mixing time of the system. This thesis aims at introducing the relevant quantity for these prediction problems, the so-called committor function which is the probability for an event to occur in the future, as a function of the current state of the system. Computing the committor in a stochastic model for ENSO illustrates that the transition to strong El Niño regimes can have either intrinsic probabilistic predictability or unpredictability. The second goal is to illustrate how to compute and validate the committor function from observations, by discussing the analogue Markov chain which provides a way for learning effective dynamics from data. Starting from it, a new algorithm is developed, with the scope of computing the committor function more precisely than the other approaches, especially in case of lack of data. Moreover, it is shown, in the context of two stochastic systems, that coupling the learning of the committor with a rare event algorithm improves the performance of the latter. Finally, this methodology is applied to a climate data-set, generated from a climate model, in order to study and predict the occurrence of extreme heat waves. After checking the consistency of the statistical quantities computed by the effective dynamics, a classifier based on the Markov chain is developed, with the capability of classifying heat waves better than other methods.

Contents

1 Introduction **1**

2 Committor Functions **11**
2.1 Introduction 11
2.2 Committor Functions for a Markov chain 13
2.2.1 Markov chains 14
2.2.2 Committor functions: definition and equation 16
2.2.3 Example: the Gambler's ruin 19
2.3 Committor functions for a continuous time Markov process . . 22
2.3.1 Markov processes 22
2.3.2 Committor functions: definition and equation 27
2.3.3 Examples: Wiener and Ornstein-Uhlenbeck processes . 29
2.4 Numerical computation of the committor 34
2.4.1 Direct estimation 34
2.4.2 Committor computation based on a long trajectory . . 35
2.4.3 Committor computation solving the Backward equation 37
2.4.4 Machine learning estimation 38
2.4.5 Validation of a committor estimate 39
2.5 Time dependent committor functions 45
2.5.1 Definition and equation for time dependent committor 46
2.5.2 Relation with cumulative distribution of first hitting time 49
2.5.3 Example: Wiener process 49

3 Committor Functions for Climate Phenomena at the Predictability Margin: The example of ENSO in the Jin and Timmermann model **52**
3.1 Introduction 53

3.2 The Jin and Timmermann model . . . 58
3.3 Statistics of the first exit times for transitions to strong El Niño regimes . . . 64
3.4 Committor function of the Jin and Timmerman model . . . 69
3.4.1 Description of the committor function: deterministic and probabilistic predictability . . . 70
3.4.2 Dynamical characterization of the probabilistically predictable region . . . 74
3.4.3 Committor function computed from long trajectories . 76
3.5 Conclusion . . . 79

4 Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain 81
4.1 Introduction . . . 82
4.2 The committor function . . . 87
4.2.1 Definition of the committor function for a Markov process 88
4.2.2 Direct sampling of the commitor function . . . 89
4.2.3 Estimating the committor function for any point of the phase space . . . 90
4.2.4 Estimation of the quality of an approximate committor function: the Brier score . . . 91
4.3 The analogue Markov chain . . . 92
4.3.1 Definition of the analogue Markov chain . . . 93
4.3.2 Computing the committor function from the analogue Markov chain . . . 96
4.3.3 Applications . . . 98
4.4 Using the learned committor function in Adaptive Multilevel Splitting . . . 112
4.4.1 The Adaptive Multilevel Splitting algorithm and the quality of score functions . . . 112
4.4.2 The learned committor function . . . 117
4.4.3 AMS study for the two dimensional three well model . 118
4.4.4 Application to the Charney-DeVore model . . . 123
4.5 Conclusion . . . 127

5 Predicting extreme events using the analogue method: the heat-wave case 129
5.1 Introduction: heat waves and committor functions . . . 129

5.2 The analogue method for heat waves 135
5.2.1 Stationary vs time periodic analogue Markov chains . . 138
5.2.2 Consistency of synthetic data: return time plot and time averaged temperature statistics 142
5.2.3 Extending return time plots 160
5.3 Committor estimation through the analogue method 162
5.3.1 Committor validation 163
5.3.2 Classifier based on the analogue method 164
5.4 Conclusions and perspectives 177

6 Conclusions **182**

Chapter 1

Introduction

It is beyond doubt that probability and statistics are essential disciplines for science. Indeed, as Poincaré said [Poincaré, 1905], "*science is built up of facts, as a house is built of stones*", and since any data analysis has to be performed using statistical tools, to underestimate their importance would result in a procedural mistake from a scientific point of view. Of course, this also applies to climate sciences, where probability and statistics are nowadays widely used for predictions.

In this context, the adoption of a probabilistic approach has different grounds according to the temporal and spatial scales of the phenomenon under investigation.

For example, weather forecasts aim at accurately predicting the future state of the atmosphere based on its current state. From a mathematical point of view, weather forecasts consist in solving an initial value problem (IVP). Indeed, given an initial condition x_0 representing the current state of the atmosphere, the aim of weather forecasts is to predict the state of the atmosphere at time t, that is $x_t = \phi^t x_0$, where ϕ^t is the flow of the dynamical system. However, in practice this is not feasible due to the chaotic behavior of many dynamical systems. As it is well known, in a chaotic system small differences in the initial conditions are exponentially amplified. This phenomenon was already known to 19th century mathematicians such as Poincaré and Hadamard, but it was Lorenz who first introduced the concept of chaos in relation to climate in 1963 [Lorenz, 1963]. In a famous conference held in 1972 provocatively titled "*Predictability: does the flap of a butterfly's wing in Brazil set off a tornado in Texas?*" [Lorenz, 1972], Lorenz explained that starting from two almost indistinguishable initial conditions (the only

difference being the presence or absence of the flapping of a butterfly's wings) the state of the atmosphere can evolve into macroscopically different states characterized by the presence or absence of a tornado. He therefore concluded that the predictability of the atmosphere could not extend beyond a few days, about two weeks, corresponding to the time it takes for undetectable errors at the smallest scales of the flow to contaminate the large scales [Thompson, 1957, Novikov, 1959, Lorenz, 1969b]. It is worth noting that this limit to predictability occurs even if a perfect model is used for the prediction. In practice, however, the models available for predictions, although very accurate, are not perfect. It should also be taken into account that often it is not possible to observe the initial state of the atmosphere with sufficient accuracy. Due to these limitations, over the past 30 years, major weather forecasting centers such as the National Oceanic and Atmospheric Administration (NOAA) and the European Center for Medium-Range Weather Forecasts (ECMWF) have put aside deterministic forecasts to move to probabilistic forecasts, obtained by means of appropriate techniques such as the ensemble forecast [Kalnay, 2003, Van den Dool et al., 2007, Dijkstra, 2013]. This technique consists in providing a set of predictions obtained by slightly changing both the initial conditions and the parameters of the model. These predictions are then used to estimate the probabilities of future states of the system. As an example, consider the predictions of 24-hr total precipitation provided by the ECMWF shown in Fig. 1.1. The two charts in Fig. 1.1 represent the probability that the 24-hr total precipitation exceeds 1 mm for two different prediction lag times. Both predictions start on 19/08/2021. The upper panel shows the forecast for 20/08/2021, i.e. for a one day lag time. The lower panel refers instead to a forecast for 08/27/2021, i.e. a lag time of 8 days. Comparing the two images, it can be seen immediately that the uncertainty of the forecasts increases as the prediction lag time increases. And indeed, the forecast for the following day is quite accurate, as evidenced by the fact that many locations have a high or low probability that precipitation exceeds the 1 mm threshold in the next twenty-four hours (light-blue and yellow regions, respectively), while for few locations these forecasts are more uncertain (green regions). In the case of the 8-day forecast, the situation changes completely. It is no longer possible to accurately predict the locations where precipitation will be most concentrated and therefore the uncertain forecast regions are more extensive than in the previous case.

From a theoretical point of view, short-term forecasts such as weather forecasts are types of deterministic forecasts where the use of probability is

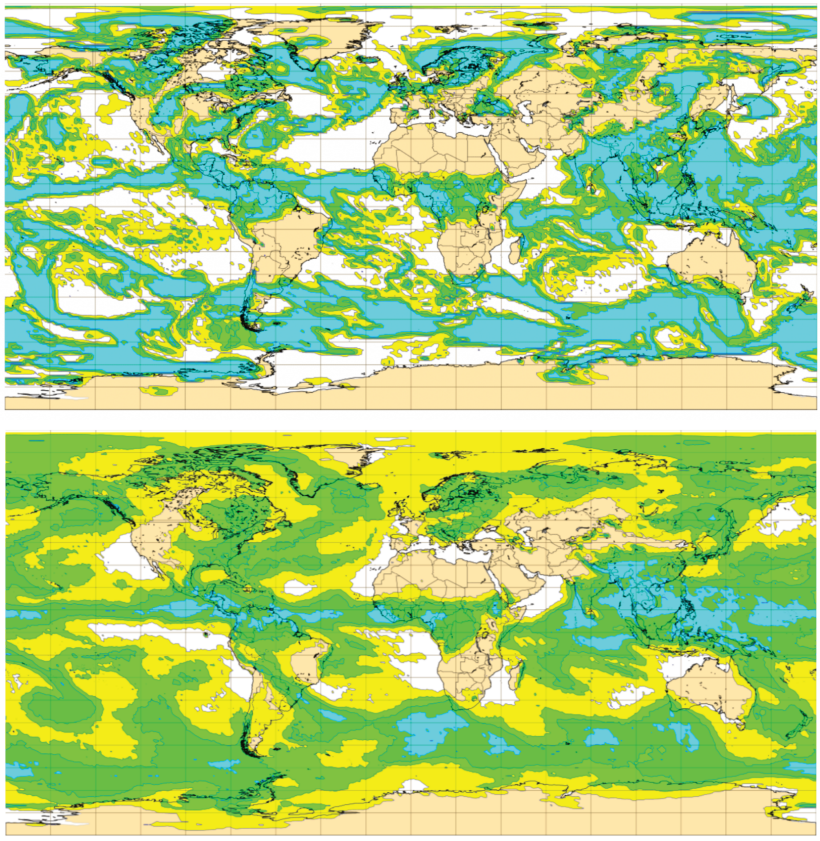


Figure 1.1: ECMWF short term forecasts on 19/08/2021 of the 24-h accumulated precipitation (from ECMWF SEAS5 website). Probability of 24-h accumulated precipitation (rain and snow) exceeding 1 mm for different prediction lag time. Upper panel: lag time 1 day. Bottom panel: lag time 8 days.

linked only to the lack of information about initial conditions and to the unavailability of perfect models.

As already explained, deterministic predictions are valid up to a finite time. For low dimensional dynamical systems, this time scale corresponds to the so called Lyapunov time [Castiglione et al., 2008], while for the atmosphere it corresponds to the time necessary for errors at small scales to become dominant. Beyond this horizon of predictability, the system behaves in all respects as a stochastic process. Thus, while in short-term forecasts the use of probability is due to our ignorance about the system, in the case of medium-term or long-term forecasts, randomness becomes an intrinsic component of the system.

Since climate is commonly defined as the weather averaged over a long period, it can be inferred from the above discussion that climate deals with the statistical properties of some stochastic process $(x_t)_{t>0}$. Over very long time scales, these properties will be independent from the initial condition. This means that, from a mathematical perspective, the relevant concept for climate is the invariant measure of the system. However, such a description is valid only for time scales larger than the mixing time, which corresponds to the time necessary for the system to forget the initial condition.

The situation is even more complex when it comes to medium-range forecasts, because in this case the interesting time scale lies between the deterministic predictability time and the mixing time of the system. This is why in this manuscript such prediction problems will also be referred to as *prediction problems at the predictability margin*. Prediction problems at the predictability margin need the application of a probabilistic method, because, as already mentioned, they cope with time scales beyond the deterministic predictability horizon of the system (e.g. the Lyapunov time) but below its mixing time. By contrast, at the predictability margin, predictions clearly depend on the current state of the system. As an example, consider the seasonal forecasts made by the ECMWF, in Fig. 1.2. It shows the probability that the time averaged 2-meter temperature exceeds the median of the seasonal average computed in the reference period 1993-2016, being the time average intended over a 3 month period. The prediction of Fig. 1.2 starts on 01/05/2021 and it refers to the time averaged temperature in summer (time average over June-July-August), i.e. a prediction lag time of 1 month. The important element to note is that this type of seasonal forecast only makes sense from a probabilistic point of view because, as can be seen from the chart in Fig. 1.2, the probability ranges between 0% and 100%, so that

2m temperature - SEAS5

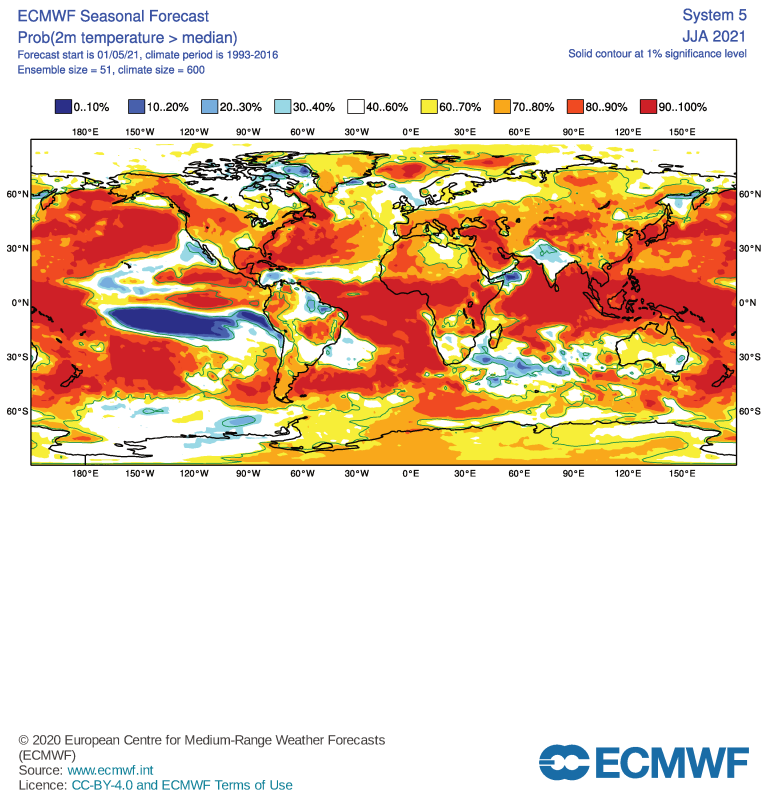


Figure 1.2: ECMWF seasonal forecast on 01/05/2021 of the 2-meter temperature averaged over the summer season (June-July-August) (from ECMWF SEAS5 website). The forecast is plotted in terms of probabilities of exceedance of the median of the seasonal average relative to the observed climate for 1993-2016.

making any kind of deterministic prediction would be impossible.

Prediction problems at the predictability margin represent the main topic of this thesis. The fact that such problems require a probabilistic approach is already known in the climate community [Wilks, 2011]. Indeed, major weather forecasting centers such as NOAA and ECMWF already make use of probabilistic predictions. Furthermore, systems like the weather@home system [Massey et al., 2015] produce a huge amount of high resolution simulations of General Circulation Models in order to sample extreme events. The original part of this work is therefore to introduce the appropriate mathematical formalism to deal with prediction problems at the predictability margin. As a first step, the appropriate mathematical tool to study this class of problems will be introduced. The relevant mathematical concept is called the *committor function*. Broadly speaking, a committor function is the probability for an event to occur in the future, as a function of the current state of the system. For the ECMWF examples this would be the maps (Fig. 1.1 and Fig. 1.2), for all the possible states of the atmosphere. Committor functions have first been introduced in climate sciences in [Lucente et al., 2019], and have been used to study sudden stratospheric warming [Finkel et al., 2020, Finkel et al., 2021] or to understand the flow of ocean debris [Miron et al., 2021]. The interest of giving a name, the committor function, to this otherwise very common and generic concept, is two-fold. First, it allows us to study its mathematical properties and to relate them to key concepts in dynamical systems, for instance the predictability margin. Second, it comes with specific theoretical and computational approaches to compute this important quantity, for instance transition path theory, see for example [Vanden-Eijnden, 2006, Metzner et al., 2006, Metzner et al., 2009] and references therein.

The previous discussion on predictability for climate dynamics is summarized in the diagram of Fig. 1.3.

Once the appropriate mathematical tool for dealing with prediction problems at the predictability margin has been identified, it is natural to wonder how to compute this quantity in practice. This task is extremely complex and becomes even more difficult when extreme events such as heat waves, cold spells or extreme precipitations are considered. By definition, these events have a very low probability of occurring and may not even have been observed in measurements made to date.

The reason for studying rare events is that they play a crucial role in many physics, chemistry, and biology phenomena, for instance when they change

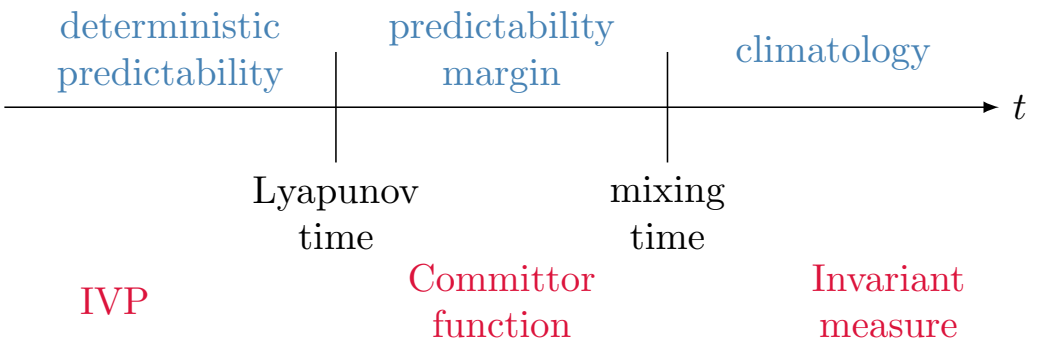


Figure 1.3: Schematic illustration of the concept of *predictability margin*: deterministic predictability is only possible until a finite time (e.g. the Lyapunov time). The associated mathematical problem is an initial value problem (IVP). Long term statistical properties (beyond the mixing time) do not depend on the initial condition, and the corresponding mathematical object is the *invariant measure*. In the intermediate range of timescales, named here as *predictability margin*, the appropriate mathematical concept is the *committor function*, which encodes the probability of a given event to occur, condition on the state of the climate system at the time of the prediction.

the system structure (multistability) or have a huge impact. The study of rare high impact events, such as heat waves, is an important topic in climate science also because the frequency of such events seems to increase due to climate change [Stocker, 2014]. The relevant time scales for many of these rare high impact events fall within the *predictability margin* and therefore the appropriate mathematical concept to study them is the committor function. Indeed, the committor encodes the probability that these events will occur in the future and this information is crucial to mitigate the impact that rare events have on human societies.

In principle, rare high impact climate events could be studied through numerical simulations (this is the case for ECMWF forecasts) but often the computational cost for collecting a large number of observations of a rare event is prohibitive and therefore it is not possible to gather enough information. Over the past 70 years, many algorithms have been designed to overcome this sampling problem [Kahn and Harris, 1951,Cérou et al., 2019b]. Recently this type of algorithm has been applied in climate science to study extreme heat waves [Ragone et al., 2018,Ragone and Bouchet, 2021,Galfi and

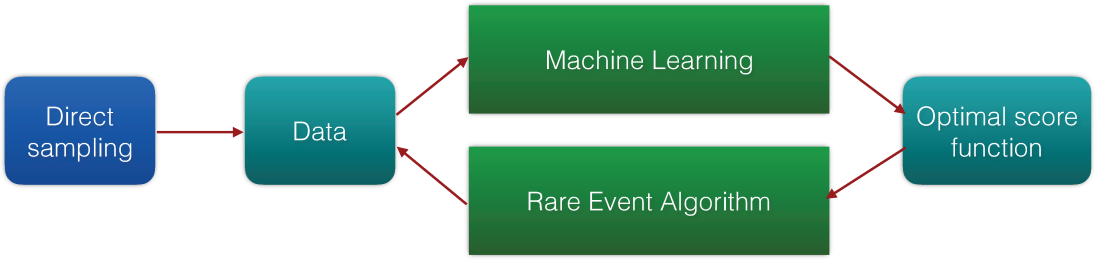


Figure 1.4: Sketch of a feedback control iterative procedure between the rare event algorithm and the machine learning of an approximate optimal score function. The learning of an approximation of the optimal score function makes the algorithm more efficient, and the algorithm provides more data for a better quality of the learning procedure.

Lucarini, 2021]. Rare events algorithms require the use of a score function to iteratively force the system towards atypical configurations in order to sample events, which are difficult to observe with direct numerical simulations, in an efficient way and at low computational cost. Unfortunately, it is well known that the optimal score function is the committor function which is exactly what needs to be computed.

It is therefore evident that it is necessary to find efficient algorithms for computing the committor function even in the presence of a lack of data. In fact, obtaining a first reliable approximation of the committor function is useful not only for studying and predicting the occurrence of the phenomenon under consideration, but it can be used as a score function in rare event algorithms, improving the performance of the latter, allowing in addition to increase the number of observations of the event. Then, the new observations can be used to obtain a more precise estimate of the committor function. The coupling between the data-driven method for estimating the committor function and the rare event algorithms can be iterated until the estimate of the committor function reaches the desired accuracy threshold. Such an iterative scheme is illustrated in Fig. 1.4.

Thus, a relevant purpose of this thesis is to propose and analyze such an algorithm. The method proposed here aims to learn effective dynamics based on dynamical observations. It is a generalization of the analogue method firstly proposed by Lorenz [Lorenz, 1969c, Lorenz, 1969a]. This method has

already been used to build a stochastic weather generator [Yiou et al., 2013, Yiou, 2014,Jézéquel et al., 2018,Yiou and Déandréis, 2019,Yiou and Jézéquel, 2020]. Therefore, the innovative part of this work consists in having used this method to develop a new data-driven approach for computing committor functions. Since this approach has also been successfully coupled to rare event algorithms, this thesis also constitutes a first step in the development of an iterative scheme for the computation of the committor function with rare event algorithms.

To summarize, the original contributions of this thesis are the following:

- it aims at introducing the concept of committor function in climate science,
- it proposes an innovative approach, based on analogue Markov chains, for computing the committor function which, coupled with a rare event algorithm, improves the performance of the latter,
- it applies the new methodology to climate data for studying and predicting extreme heat waves.

The present manuscript is structured as follows: the definition and mathematical properties of the committor function, as well as numerical methods to compute it, approaches for the validation of its estimates and its relation with prediction problems are illustrated in Chapter 2. Chapter 3 studies the dynamics of a low-dimensional stochastic model proposed to explain the decadal amplitude changes of El Niño Southern Oscillation, the Jin and Timmermann model [Timmermann and Jin, 2002, Timmermann et al., 2003]. This model is not aimed at reproducing any precise properties of the real El Niño Southern Oscillation. It is rather used as a paradigmatic example to introduce the concept of a committor function in climate science, and to study its main properties. This study leads to the definition of probabilistic predictability and unpredictability, some concepts that should be useful for other applications. This work with minor modifications, carried out in collaboration with C. Herbert and F. Bouchet, has been submitted to the "Journal of the Atmospheric Sciences" (JAS) [Lucente et al., 2021]. Chapter 4 proposes a novel data-driven approach that efficiently estimates the committor function starting from observed dynamics. The method combines the use of the analogue Markov chain with a spectral characterization of the committor, with the aim of providing an alternative approach for the computation of the latter, which could be useful in the lack of data. In addition,

it is shown that such an approach can be paired with a rare event method with two advantages: the computations can be performed with a minimal prior knowledge and the results are more precise than those obtained with a user-designed score function. A slightly modified version of this work, which is the result of a collaboration with J. Rolland, C. Herbert and F. Bouchet, has been submitted to the "Journal of Statistical Mechanics: Theory and Experiment" (J. Stat. Mech.). In Chapter 5, the new methodology for computing committor functions is applied to a climate data-set, generated by a climate model, to study and predict the occurrence of extreme heat waves. Finally, Chapter 6 highlights the main conclusions of the work and illustrates possible future developments.

Chapter 2

Committor Functions

2.1 Introduction

This chapter is for pedagogical purposes and aims at introducing, in a simple way, the mathematical formalism of the committor function, necessary to rigorously tackle the problems studied in the following chapters.

After being introduced in 1938 by Onsager [Onsager, 1938], the committor function has received particular attention, especially in the context of molecular dynamics [E et al., 2005, Vanden-Eijnden, 2006, Metzner et al., 2006, Bovier, 2006, Metzner, 2008, Metzner et al., 2009, Prinz et al., 2011, Schütte et al., 2011, Bowman et al., 2013, Schütte and Sarich, 2015, Lopes and Lelièvre, 2019, Thiede et al., 2019]. More recently, the committor function has been introduced also in the context of climate sciences [Lucente et al., 2019, Finkel et al., 2020, Miron et al., 2021, Finkel et al., 2021].

To understand the reason for this interest, it is good first of all to define this mathematical concept. A committor function is the probability that an event occurs before another one as a function of the current state of the system. To be slightly more precise, consider two events w_1 and w_2 and let $X(t)$ be the variable describing the state of the system at time t. The committor function for the two events w_1 and w_2 is defined as

$$q(x) = \mathbb{P}(\tau_{w_1}(x) < \tau_{w_2}(x)), \tag{2.1}$$

where $\tau_w(x) = \inf\{t : w \text{ happens } | X(0) = x\}$ is the smallest time at which the event w happens.

Usually, in molecular dynamics the events represent two subsets $\mathcal{A}$ and $\mathcal{B}$ of the phase space, and one is interested in computing the probability

that one set ($\mathcal{B}$) is reached before the other one ($\mathcal{A}$). In climate instead, it is more frequent that one is interested in computing the probability that a given event, far from the typical conditions of the system, will occur within a certain time. In this case, therefore, the committor function describes the probability that the system reaches atypical states before relaxing towards the typical stationary conditions. For example, consider the problem of predicting the probability of occurrence of a heat wave, which will be addressed in Chapter 5. In this case the set $\mathcal{B}$ contains all the possible trajectories in which a heat wave occurs, while the set $\mathcal{A}$ contains all the trajectories for which this event does not occur. Then, the committor function is the probability to reach $\mathcal{B}$ before $\mathcal{A}$, as a function of the current state of the atmosphere, temperature, soil moisture and so on.

Thus, the committor function is a fundamental tool for studying the transitions between events. For example, in transition path theory [Vanden-Eijnden, 2006, Metzner et al., 2006, Metzner, 2008, Metzner et al., 2009], this function is a fundamental block for obtaining the statistics of reactive trajectories (i.e. trajectories starting from $\mathcal{A}$ and arriving in $\mathcal{B}$ or vice versa). It also plays a crucial role in the development of rare event algorithms, which are algorithms designed for sampling rare events at low computational cost. Indeed, these algorithms require to use a score function for forcing the system towards atypical configurations and it is well known that the optimal score function is the committor function (see for instance [Cérou and Guyader, 2007, Rolland and Simonnet, 2015, Lestang et al., 2018]). In climate sciences, instead, it arouses interest as it allows the prediction problems to be treated with a rigorous theoretical apparatus.

Given the importance of the committor function, over the years a theory has been developed in the context of Markov processes. The purpose of this chapter is therefore to introduce the fundamental concepts of the theory of committor functions for Markov processes.

The structure of the chapter is the following: Sec. 2.2 introduces the concept of committor function for a Markov chain. It begins by reviewing the basic notions of Markov chains. Then, it gives the definition of the committor function for a Markov chain and it shows that this function must satisfy a system of linear equations. The committor function for a continuous Markov process is discussed in Sec. 2.3. As for Sec. 2.2, also in Sec. 2.3 there is a general introduction on Markov processes, where all the properties that are necessary to define the committor function are stated. Then, the committor function for this class of stochastic processes is discussed, and it is shown that

the committor is defined by a partial differential equation with appropriate boundary conditions. Sec. 2.4 explains how to compute the committor function numerically. Three different algorithms are described, highlighting their strengths and weaknesses. Moreover, it is explained how to validate an estimate of the committor function. For this purpose, two scores are introduced (the Brier score and the logarithmic score) which quantify the accuracy of the probabilistic predictions made using an approximation of the committor function. Finally, Sec. 2.5 introduces the concept of time-dependent committor function. This mathematical object encodes the probability that an event will happen in the future as a function of the initial conditions. Therefore, it reveals itself as a fundamental tool in the study of prediction problems, especially in the context of climate system where it often happens that it is necessary to predict the future evolution of interesting observables. Thus, Sec. 2.5 precisely defines this object and shows that for a wide class of observables, the time-dependent committor function is nothing but the cumulative distribution function of the first hitting time of a given set $\mathcal{C}$. Each of these sections contains one or more examples, in which the theoretical concepts defined in the section are applied. These examples are useful to better illustrate the theoretical concepts and serve as a test bed to analyze some properties of the committor function before moving on to the study of more realistic and complex systems.

2.2 Committor Functions for a Markov chain

This section introduces the concept of committor function in the context of simple stochastic processes, i.e. Markov chains. The section begins introducing the basic notions of Markov chains. In particular, it aims to introduce the notions of Markov chain and its related transition matrix, the link between the transition matrix and the temporal evolution of probabilities and observables, the concept of invariant distribution and the convergence of the probabilities to this invariant distribution. Then, it provides the definition of a committor function of a Markov chain and it shows that the committor function is the solution of an affine problem. Moreover, it is explained that the committor function is a linear combination of the two leading eigenvectors of a suitable transition matrix and it is shown that this spectral characterization provides an algorithm that efficiently estimates the committor. The section ends by discussing an application of the committor function

in a gambling problem that can be modeled by means of Markov chains.

2.2.1 Markov chains

Let Ω be a countable set and $\{X_t\}_{t\in\mathbb{N}}$ a stochastic process that takes values in Ω. The process $\{X_t\}_{t\in\mathbb{N}}$ is said to be a Markov chain if the following property (Markov property) holds [Norris et al., 1998,Shiryaev, 2004,Boffetta and Vulpiani, 2012,Bowman et al., 2013]

$$\mathbb{P}(X_t = i_t | X_{t-1} = i_{t-1}; \ldots; X_0 = i_0) = \mathbb{P}(X_t = i_t | X_{t-1} = i_{t-1}). \tag{2.2}$$

The Markov property expresses the fact that the probability of the future state of the system depends exclusively on the state of the system at the previous time and is independent of past history. Thus, the main feature of Markov processes is to have a finite memory of length one.

The conditional probabilities that appear in the right hand side of Eq. (2.2) are usually represented by a stochastic matrix $G(t)$, also known as transition matrix. A matrix is called stochastic if

$$\begin{aligned} &G_{ij}(t) \geq 0 \quad \forall\, i,j \in \Omega \text{ and } \forall\, t \in \mathbb{N}, \\ &\sum_{j\in\Omega} G_{ij}(t) = 1 \quad \forall\, i \in \Omega \text{ and } \forall\, t \in \mathbb{N}. \end{aligned} \tag{2.3}$$

The first property states that the elements of a stochastic matrix must be positive as they are probabilities. The second property asserts instead that the sum of the transition probabilities from state i to all the possible states j must be equal to one. Note that the convention $G_{ij}(t-1) = \mathbb{P}(X_t = j | X_{t-1} = i)$ has been adopted and therefore $G_{ij}(t-1)$ is the conditional probability of going from state i at time $t-1$ to state j at time t.

The transition matrix $G(t)$ is sometimes called *generator* since its action on an observable f determines the time evolution of f. To be more precise, let $f(t) = (f_i(t))$ be a column vector that represents an observable of the system at time t. To obtain the observable $f(t+1)$ at time $t+1$ it is sufficient to apply the matrix $G(t)$ to $f(t)$:

$$f(t+1) = G(t+1) f(t). \tag{2.4}$$

The evolution of probabilities is ruled by the adjoint operator of $G(t)$, denoted by $G^\dagger(t)$ ($G^\dagger$ is equivalent to G transposed as G has real positive

elements). Let $p(t) = (p_i(t))$ be a vector of probabilities, i.e. a column vector where each component $p_i(t)$ is the probability that the system is in state $i \in \Omega$ at time t. The probability vector $p(t+1) = (p_i(t+1))$ at time $t+1$ is given by

$$p(t+1) = G^\dagger(t+1)p(t). \tag{2.5}$$

From Eq. (2.4) and Eq. (2.5) it can be deduced that

$$\begin{aligned} f(t+1) &= \left(\prod_{\tilde{t}=1}^{t+1} G(\tilde{t})\right) f(0), \\ p(t+1) &= \left(\prod_{\tilde{t}=1}^{t+1} G^\dagger(\tilde{t})\right) p(0). \end{aligned} \tag{2.6}$$

In general, the transition matrix $G(t)$ depends on time. If the conditional probabilities $\mathbb{P}(X_t = j | X_{t-1} = i)$ do not depend on time t, i.e. $G_{ij}(t) = G_{ij} \, \forall \, t \in \mathbb{N}$, the Markov chain is said to be homogeneous. For an homogeneous Markov chain Eq. (2.6) reads

$$\begin{aligned} f(t+1) &= G^{t+1} f(0), \\ p(t+1) &= \left(G^\dagger\right)^{t+1} p(0), \end{aligned} \tag{2.7}$$

where G^{t+1} and $\left(G^\dagger\right)^{t+1}$ are the matrices G and $G^\dagger$ to the power $t+1$.

For the study of the long-term properties of Markov chains it is useful to introduce the notion of invariant distribution. A probability vector w is called an invariant distribution if

$$\begin{cases} w_i \geq 0 \, \forall \, i \in \Omega, \\ \sum_{i \in \Omega} w_i = 1, \\ G^\dagger w = w. \end{cases}$$

From Eq. (2.3) it is easy to prove that each stochastic matrix G admits a trivial eigenvector $v = v_i$, $v_i = 1 \, \forall i \in \Omega$ with eigenvalue 1[1]. This implies that a solution to the problem $G^\dagger w = w$ always exists but the solution is not always unique. The uniqueness of the invariant distribution can be proven

[1] A Perron-Frobenius theorem for positive operators that preserve probabilities states that eingenvalue 1 is the largest eigenvalue.

for a broad class of Markov chains, the so called ergodic chains [Norris et al., 1998, Shiryaev, 2004, Boffetta and Vulpiani, 2012]. Let G be the transition matrix of a Markov chain whose phase space Ω is finite. The Markov chain is ergodic if

$$\exists t^\star \text{ such that } G_{ij}^{t\star} > 0 \,\forall\, i,j \in \Omega, \tag{2.8}$$

where $t^\star$ does not depend on i and j. Eq. (2.8) means that for any $t > t^\star$ there is a non-null probability of finding the process X_t in any state $j \in \Omega$ regardless the initial state of the process $X_0 = i$.

For an ergodic chain, the eigenvector problem $G^\dagger w = w$ admits a unique solution $w = \pi$ that has the two properties $\pi_i \geq 0 \,\forall\, i \in \Omega$, $\sum_{i\in\Omega} \pi_i = 1$. Furthermore, it can be proven that the invariant distribution π is the limit of G^t for $t \to +\infty$, i.e.

$$\pi_i = \lim_{t\to+\infty} G_{ji}^t, \tag{2.9}$$

which also implies that $\pi_i = \lim_{t\to+\infty} p(t)$[2] [Norris et al., 1998, Shiryaev, 2004].

Ergodic chains play a fundamental role in the theory of Markov chains and their definition will be useful for proving the existence and uniqueness of the committor function.

2.2.2 Committor functions: definition and equation

Let Ω and G be the state-space and the transition matrix of a Markov chain.

Let $\mathcal{A}$ be a subset of Ω, $\mathcal{A} \subset \Omega$, and $\{X_t\}_{0\leq t\leq+\infty}$ a realization of the dynamics. The first hitting time of the set $\mathcal{A}$ is defined as

$$\tau_j^{\mathcal{A}} = \inf\{t : X_t \in \mathcal{A} | X_0 = j\}. \tag{2.10}$$

Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of Ω, i.e. $\mathcal{A}, \mathcal{B} \subset \Omega$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$. The committor function q_j is defined as

$$q_j = \mathbb{P}(\tau_j^{\mathcal{B}} < \tau_j^{\mathcal{A}}). \tag{2.11}$$

Thus, the committor function q_j encodes the probability that a trajectory starting at $X_0 = j$ will reach the set $\mathcal{B}$ before reaching the set $\mathcal{A}$.

[2] As a consequence of the Perron-Frobenius theorem it can be proven that the convergence is exponentially fast, i.e. $p(t) = \pi + O(e^{-t\log|\lambda_2|})$ where λ_2 is the largest eigenvalue in module except $\lambda_1 = 1$.

An equation for q is easily obtained with the following reasoning. Clearly for $j \in \mathcal{A}$ or $j \in \mathcal{B}$ one has $q_j = 0$ or $q_j = 1$, respectively. For any initial state $X_0 = j \notin \mathcal{A} \cup \mathcal{B}$, after the first step three situations could arise: either the Markov chain goes from state j to a state $k \in \mathcal{B}$ and the event $\tau_j^{\mathcal{B}} < \tau_j^{\mathcal{A}}$ occurs, or it goes to a state $k \in \mathcal{A}$ and the event $\tau_j^{\mathcal{B}} < \tau_j^{\mathcal{A}}$ does not occur, or it reach a state $k \notin \mathcal{A} \cup \mathcal{B}$ and then the probability of the event $\tau_j^{\mathcal{B}} < \tau_j^{\mathcal{A}}$ to occur is q_k. Therefore, the probability that a trajectory starting in state j will hit the set $\mathcal{B}$ before $\mathcal{A}$ can be decomposed into the sum of two terms [Bowman et al., 2013, Prinz et al., 2011, Metzner, 2008, Metzner et al., 2009, Bovier, 2006]:

$$\begin{cases} q_j = 0, \text{ if } j \in \mathcal{A}, \\ q_j = 1, \text{ if } j \in \mathcal{B}, \\ q_j = \sum_{k \in \mathcal{B}} G_{jk} + \sum_{k \notin \mathcal{A} \cup \mathcal{B}} G_{jk} q_k, \text{ if } j \notin \mathcal{A} \cup \mathcal{B}, \end{cases} \tag{2.12}$$

where the first term in the right hand side of the last equation in Eqs. (2.12) accounts for the probabilities to have a transition from the state j to a state $k \in \mathcal{B}$ while the second term is related to the probability of visiting a state $k \notin \mathcal{A} \cup \mathcal{B}$ before going to the set $\mathcal{B}$.

For an ergodic Markov chain (Eq. (2.8)), it can be proven that the system (2.12) admits a unique solution [Metzner, 2008, Metzner et al., 2009, Bovier, 2006]. Therefore, for an ergodic chain the committor function can be obtained by solving the linear system in Eq. (2.12).

There is a more elegant formulation of the committor problem which shows that the committor function q is actually an eigenvector of a suitable transition matrix $\tilde{G}$, where the states corresponding to $\mathcal{A}$ and $\mathcal{B}$ are replaced by two absorbing states, one for each set [Prinz et al., 2011].

Consider the modified Markov chain in which the states corresponding to the set $\mathcal{A}$ (resp. $\mathcal{B}$) are grouped together into a single absorbing state[3] denoted by $i_{\mathcal{A}}$ (resp. $i_{\mathcal{B}}$). Let $C(\Omega)$ be the cardinality of the state space and $C(\mathcal{A})$ (resp. $C(\mathcal{B})$) the cardinality of $\mathcal{A}$ (resp. $\mathcal{B}$). The cardinality of the modified state space $\tilde{\Omega}$, i.e. the state space resulting from the aggregation procedure, is $C(\tilde{\Omega}) = C(\Omega) - (C(\mathcal{A}) - 1) - (C(\mathcal{B}) - 1)$. The transition matrix

[3] An absorbing state k is a state from which no transitions to other states than k are allowed. In terms of the transition matrix G, a state k is absorbing if $G_{kk} = 1$ and $G_{kj} = 0$ for $j \neq k$.

$\tilde{G}$ of the modified Markov chain (defined on $\tilde{\Omega}$) is

$$\begin{cases} \tilde{G}_{i_{\mathcal{A}}j} = \delta_{i_{\mathcal{A}},j}, \\ \tilde{G}_{i_{\mathcal{B}}j} = \delta_{i_{\mathcal{B}}j}, \\ \tilde{G}_{ii_{\mathcal{A}}} = \sum_{k\in\mathcal{A}} G_{ik} \quad \text{if} \quad i \neq i_{\mathcal{A}}, i_{\mathcal{B}}, \\ \tilde{G}_{ii_{\mathcal{B}}} = \sum_{k\in\mathcal{B}} G_{ik} \quad \text{if} \quad i \neq i_{\mathcal{A}}, i_{\mathcal{B}}, \\ \tilde{G}_{ij} = G_{ij} \quad \text{if} \quad i,j \neq i_{\mathcal{A}}, i_{\mathcal{B}}, \end{cases} \tag{2.13}$$

where $\delta_{i,j}$ is the Kronecker delta. It is straightforward to find a relationship between the committor function of the original Markov chain q and that of the modified process $\tilde{q}$. Indeed, for $i \notin \mathcal{A} \cup \mathcal{B}$ one has $q_i = \tilde{q}_i$, while $q_k = q_{i_{\mathcal{A}}} = 0$ for $k \in \mathcal{A}$ and $q_k = q_{i_{\mathcal{B}}} = 1$ for $k \in \mathcal{B}$. The committor function $\tilde{q}$ is the solution of the affine problem

$$\begin{cases} \tilde{q}_{i_{\mathcal{A}}} = 0 \quad \text{and} \quad \tilde{q}_{i_{\mathcal{B}}} = 1, \\ \tilde{q}_i = \sum_{j\in\tilde{\Omega}} \tilde{G}_{ij}\tilde{q}_j \quad \text{for} \quad i \neq i_{\mathcal{A}}, i_{\mathcal{B}}, \end{cases} \tag{2.14}$$

that can be written more compactly as $\tilde{G}\tilde{q} = \tilde{q}$ with $q_{i_{\mathcal{A}}} = 0$ and $q_{i_{\mathcal{B}}} = 1$. Therefore $\tilde{q}$ is an eigenvector of the matrix $\tilde{G}$ with eigenvalue 1 which satisfies the correct boundary conditions on $i_{\mathcal{A}}$ and $i_{\mathcal{B}}$.

The matrix $\tilde{G}^\dagger$ has two trivial eigenvectors associated with the eigenvalue 1, which correspond to situations where the probability is concentrated on state $i_{\mathcal{A}}$ or $i_{\mathcal{B}}$. Hence, $\tilde{G}$ also possesses at least two eigenvector associated with eigenvalue 1. It is not difficult to be convinced that if G is ergodic then $\tilde{G}$ and $\tilde{G}^\dagger$ have only two eigenvectors associated with the eigenvalue 1 (i.e. $\lambda_1 = 1$ has multiplicity 2)[4]. Let v^1 and v^2 be two eigenvectors of $\tilde{G}$ associated

[4]The ergodicity of G ensures that, for each pair of states i and j of the Markov chain defined by G, there is a path connecting the two states i and j. Hence, starting from a generic state $i \notin \mathcal{A} \cup \mathcal{B}$ it is always possible to reach the two sets $\mathcal{A}$ and $\mathcal{B}$. Considering the paths that connect a state i with a state $j \in \mathcal{B}$, it is clear that there are two types of paths for going from i to j: those that connect i and j passing through the set $\mathcal{A}$ and those that instead do not pass through $\mathcal{A}$. In the Markov chain defined by $\tilde{G}$, the first type of paths connects the state i to the state $i_{\mathcal{A}}$ while the second one connects i to $i_{\mathcal{B}}$. Since the same reasoning can be applied for $j \in \mathcal{A}$, it follows that, in the Markov chain defined by $\tilde{G}$, any initial state i is transported in $i_{\mathcal{A}}$ or in $i_{\mathcal{B}}$ by the dynamics. Thus, the Markov chain defined by $\tilde{G}$ transports any initial probability distribution p either into a probability distribution concentrated on $i_{\mathcal{A}}$ or on $i_{\mathcal{B}}$ or a combination of them. Therefore, the dimension of the vector space associated to the eigenvalue of $\tilde{G}^\dagger$ (or equivalently of $\tilde{G}$) $\lambda_1 = 1$ is equal to 2.

with $\lambda_1 = 1$, i.e.

$$\begin{cases} \tilde{G}v^1 = v^1, \\ \tilde{G}v^2 = v^2. \end{cases} \tag{2.15}$$

Clearly, any linear combination $v = \alpha v^1 + \beta v^2$ is again an eigenvector of $\tilde{G}$ associated with $\lambda_1 = 1$. The committor function $\tilde{q}$ is the only eigenvector among v's such that $v_{i_\mathcal{A}} = 0$ and $v_{i_\mathcal{B}} = 1$. Therefore one has

$$\tilde{q} = \bar{\alpha} v^1 + \bar{\beta} v^2, \tag{2.16}$$

where

$$\begin{aligned} \bar{\alpha} &= -\frac{v^2_{i_\mathcal{A}}}{v^1_{i_\mathcal{A}} v^2_{i_\mathcal{B}} - v^1_{i_\mathcal{B}} v^2_{i_\mathcal{A}}}, \\ \bar{\beta} &= \frac{v^1_{i_\mathcal{A}}}{v^1_{i_\mathcal{A}} v^2_{i_\mathcal{B}} - v^1_{i_\mathcal{B}} v^2_{i_\mathcal{A}}}. \end{aligned} \tag{2.17}$$

Eq. (2.16), together with the definitions of the coefficients $\bar{\alpha}$ and $\bar{\beta}$ (Eq. (2.17)) provides a practical, fast and robust way, to compute the committor function of a discrete time Markov chain.

2.2.3 Example: the Gambler's ruin

A first example that shows the relevance of the committor function is the problem of *gambler's ruin* [Coolidge, 1909, Norris et al., 1998, Shiryaev, 2004, Boffetta and Vulpiani, 2012, Slade, 2014]. Consider a man who starts playing a betting game and whose starting capital is $X_0 = n$. On each round, one dollar is wagered and the player has a probability p of earning an additional dollar and a probability $q = 1 - p$ of losing his stake. The gambler aims to make a fortune of N dollars before he stops playing. However, during the game the player may also run out of money and be unable to continue betting. Hence, the game ends with either the gambler's ruin or the gambler's win. The problem consists in finding the probability that the player wins by starting with an initial capital $X_0 = n$.

Such a problem can be solved by employing the formalism for the committor function of a Markov chain. Let us consider a Markov chain whose states space $\Omega = \{0, 1, \cdots, N\}$ contains the possible amounts of money the

player may own and whose transition matrix G is defined as

$$\begin{cases} G_{0i} = \delta_{i,0}\,, \\ G_{Ni} = \delta_{i,N}\,, \\ G_{ij} = p\delta_{j,i+1} + (1-p)\,\delta_{j,i-1} \text{ if } i \neq 0, N\,, \\ G_{ij} = 0 \text{ otherwise}\,, \end{cases} \tag{2.18}$$

where $\delta_{i,j}$ is the Kronecker delta. By defining $A = \{0\}$ and $B = \{N\}$, the probability that the player succeeds is nothing but the committor function for these two sets. In this particular case, Eq. (2.12) reads

$$\begin{cases} q_i = pq_{i+1} + (1-p)q_{i-1}\,, \\ q_0 = 0\,, \\ q_N = 1\,. \end{cases} \tag{2.19}$$

Performing the change of variables $z_i = q_i - q_{i-1}$ leads to

$$z_{i+1} = \frac{1-p}{p} z_i = \left(\frac{1-p}{p}\right)^i z_1, \tag{2.20}$$

which implies that

$$q_{i+1} = \begin{cases} \frac{1-\left(\frac{1-p}{p}\right)^{i+1}}{1-\left(\frac{1-p}{p}\right)} q_1 \text{ if } p \neq (1-p)\,, \\ (i+1)q_1 \text{ if } p = (1-p) = 0.5\,. \end{cases} \tag{2.21}$$

Eq. (2.21) has been obtained considering $z_1 = q_1$ and $q_{i+1} = q_1 + \sum_{k=1}^{i+1} z_k$. By imposing the condition $q_N = 1$, the committor function turns out to be

$$q_n = \begin{cases} \frac{1-\left(\frac{1-p}{p}\right)^n}{1-\left(\frac{1-p}{p}\right)^N} \text{ if } p \neq (1-p)\,, \\ \frac{n}{N} \text{ if } p = (1-p) = 0.5\,, \end{cases} \tag{2.22}$$

where the two situations correspond to unfair ($p \neq 1-p$) or fair ($p = 1-p$) games.

For $p = 0.5$ (fair game) the player's probability of winning grows linearly with the starting capital $X_0 = n$ and therefore to have at least a 50% chance of winning the starting capital must exceed half of the stake that the man wants to earn. On the other side, in case of unfair game the committor

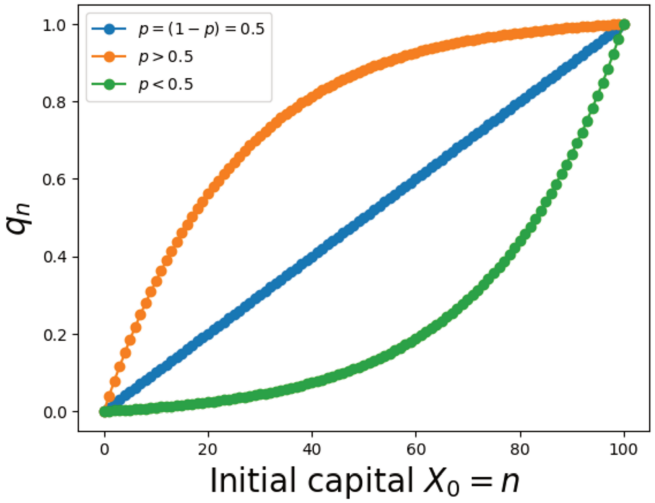


Figure 2.1: Gambler's probability of winning $N = 100\$$ as a function of the initial capital $X_0 = n\$$. The blue, orange and green curves indicate the case of fair play ($p = 0.5$), unfair play in favor of the gambler ($p = 0.51$) or unfair play against the player ($p = 0.49$), respectively.

function q shows an exponential behavior. Therefore, small values of inequity drastically increase ($p > 0.5$) or decrease ($p < 0.5$) the probability of win leading to situations where win ($p > 0.5$) or ruin ($p < 0.5$) is practically certain for many initial amounts.

Fig. 2.1 shows the probability that the player earns $N = 100$ dollars without first getting ruined for either fair or unfair game. The odds of winning a bet have been taken as $p = 0.5$, $p = 0.51$ and $p = 0.49$, respectively. Note that although the inequities are very small the behavior of the committor is completely different. To have a probability of winning equal to half, the man has to start with 50 dollars if the game is fair, with an initial stake of less than 20 dollars if the game is in his favor and an initial capital of about 80 dollars if the game is against him. In the three cases considered, if the player decides to start the game with $X_0 = 50$ dollars, the chances of success are equal to 50%, 85% and 10%, respectively.

2.3 Committor functions for a continuous time Markov process

This section deals with the committor function in the case of continuous time Markov processes. The section begins by introducing the general concepts of Markov processes, such as the propagator, the invariant distribution, and the differential equations that these objects obey. In addition, the notions of stationarity, homogeneity and ergodicity of a Markov process will be introduced. Then the concept of committor function is defined and it is proved that this function is the solution of a partial differential equation with appropriate boundary conditions. The section concludes by discussing two examples in which the equations defining the committor can be solved analytically. These two examples, which are the Wiener and Ornstein-Uhlenbeck processes, will allow understanding the general properties of a committor function and will be preparatory for the study of the committor in more complicated systems.

2.3.1 Markov processes

Let $\Omega \subset \mathbb{R}^N$ be the phase space of the system and $X(t) = (X_1(t), \cdots, X_N(t))$ be a random variable which takes values in Ω. Let $t_n > t_{n-1} > \cdots > t_0$ be an ordered sequence of times and $x_0, \cdots, x_n$ the values assumed by the process at these times, i.e. $x_i = X(t_i)$ for $i = 1, \cdots, n$. The process $X(t)$ is a Markov process if it has the following property [Arnold, 1974, Gardiner et al., 1985, Van Kampen, 1992, Metzner, 2008, Boffetta and Vulpiani, 2012]:

$$\mathbb{P}(x_n, t_n | x_{n-1}, t_{n-1}; \cdots; x_0, t_0) = \mathbb{P}(x_n, t_n | x_{n-1}, t_{n-1}). \tag{2.23}$$

The only differences between Eq. (2.23) and Eq. (2.2) are that the phase space Ω of a continuous time Markov process does not need to be a countable set and the time variable t takes values in $\mathbb{R}$ and not in $\mathbb{N}$. Thus, a Markov process is a process such that the future is completely determined (in a probabilistic sense) by the knowledge of the present state of the system. A Markov process is completely determined by the knowledge of the initial probability distribution $p(x_0, t_0)$ and the propagator $G(x_1, t_1 | x_0, t_0)$. These objects can be two generic distributions and they only need to satisfy the

following relationships

$$p(x_1,t_1) = \int_\Omega \mathrm{d}y\, G(x_1,t_1|y,t_0)p(y,t_0), \tag{2.24}$$

$$G(x_2,t_2|x_0,t_0) = \int_\Omega \mathrm{d}y\, G(x_2,t_2|y,t_1)G(y,t_1|x_0,t_0). \tag{2.25}$$

The relation (2.24) is valid for any stochastic process and it expresses the fact that the one-point distribution $p(x_1,t_1)$ can be obtained by marginalizing the two-point distribution $p(x_1,t_1;x_0,t_0) = G(x_1,t_1|x_0,t_0)p(x_0,t_0)$. The integral equation (2.25) is known as the Chapman-Kolmogorov equation and is a fundamental relation that the conditional probabilities of a Markov process must satisfy. In fact, it can be proven that any two distributions that satisfy Eqs. (2.24),(2.25) define uniquely a Markov process [Gardiner et al., 1985, Van Kampen, 1992].

By making very mild assumptions, the Chapman-Kolmogorov equation can be reformulated into a differential equation for conditional probability. It can be proven (see [Gardiner et al., 1985] for a detailed derivation) that the evolution of $G(x,t_1|y,t_0)$ is ruled by

$$\begin{aligned}\frac{\partial G(x,t_1|y,t_0)}{\partial t_1} = &-\sum_{i=1}^{N}\frac{\partial}{\partial x_i}\left[A_i(x,t_1)G(x,t_1|y,t_0) - \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}D_{ij}(x,t_1)G(x,t_1|y,t_0)\right] + \\ &+\int_\Omega \mathrm{d}z\left[W(x|z,t_1)G(z,t_1|y,t_0) - W(z|x,t_1)G(x,t_1|y,t_0)\right],\end{aligned} \tag{2.26}$$

where

$$W(x|y,t) = \lim_{\Delta t\to 0}\frac{G(x,t+\Delta t|y,t)}{\Delta t}, \tag{2.27}$$

$$A_i(x,t) = \lim_{\epsilon\to 0}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{|x-z|<\epsilon}\mathrm{d}z\,(z_i-x_i)G(z,t+\Delta t|x,t), \tag{2.28}$$

$$D_{ij}(x,t) = \lim_{\epsilon\to 0}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{|x-z|<\epsilon}\mathrm{d}z\,(z_i-x_i)(z_j-x_j)G(z,t+\Delta t|x,t). \tag{2.29}$$

Note that Eq. (2.26) deals with the time evolution of the propagator $G(x,t_1|y,t_0)$ with respect to the final state x and therefore is known as forward differential Chapman-Kolmogorov equation.

Sometimes it is useful to study the temporal evolution of the propagator $G(x, t_1|y, t_0)$ with respect to the initial variable. The equation describing this time development is similar to Eq. (2.26) and can be expressed as

$$\frac{\partial G(x,t_1|y,t_0)}{\partial t_0} = -\sum_{i=1}^{N}\left(A_i(y,t_0)\frac{\partial G(x,t_1|y,t_0)}{\partial y_i}\right) - \frac{1}{2}\sum_{i,j=1}^{N}\left(D_{ij}(y,t_0)\frac{\partial^2 G(x,t_1|y,t_0)}{\partial y_i \partial y_j}\right) + \\ + \int_\Omega \mathrm{d}z W(z|y,t_0)\left[G(x,t_1|y,t_0) - G(x,t_1|z,t_0)\right], \tag{2.30}$$

which is known as backward differential Chapman-Kolmogorov equation.

Having introduced the equations for the time evolution of the propagator of a Markov process, it is time to mention some special classes of these processes.

A Markov process is continuous (i.e. its sample paths are continuous functions of time t with probability one) if, $\forall \epsilon > 0$, one has

$$\lim_{\Delta t \to 0} \int_{|x-z|>\epsilon} \mathrm{d}x\, G(x, t+\Delta t|z,t) = 0. \tag{2.31}$$

This means that the requirement for a Markov process to be continuous is that the probability to observe a finite displacement $\epsilon > 0$ goes to 0 faster than Δt as $\Delta t \to 0$.

Another important property that a Markov process may have is the stationarity. A Markov process is said to be stationary if $X(t)$ and $X(t+\tau)$ have the same statistics for any τ, or in an equivalent way if the joint probabilities of the process are invariant under time translations. Hence, since the Markov process is completely determined by the initial distribution $p(x_0, t_0)$ and the propagator $G(x_1, t_1|x_0, t_0)$ it means that $p(x_0, t_0) = \rho_s(x_0)$ and $G(x_1, t_1|x_0, t_0) = G(x_1, t_1 - t_0|x_0, 0)$, i.e. the propagator only depends on time difference $\tau = t_1 - t_0$. If the initial distribution is not the stationary one but the propagator depends only on the time differences, the process is said to be homogeneous. For a homogeneous Markov process, the functions W, A and D defined in Eqs. (2.27),(2.28),(2.29) do not depend on time t.

As with Markov chains, ergodic processes play an important role in the theory of stochastic processes. In general, a Markov process is said to be ergodic if it is homogeneous and for any subset of the phase space $\mathcal{C} \subset \Omega$,

one has [Gardiner et al., 1985, Vulpiani et al., 2014]

$$\lim_{T\to+\infty}\frac{1}{T}\int_0^T \mathrm{d}t\mathbb{1}_{\mathcal{C}}(X(t)) = \int_{\mathcal{C}}\mathrm{d}x\rho_s(x), \tag{2.32}$$

where $\rho_s(x)$ is the stationary distribution of the process and $\mathbb{1}_{\mathcal{C}}$ the characteristic function of the set $\mathcal{C}$. A sufficient condition for Eq. (2.32) to be valid is that

$$\lim_{\tau\to+\infty} G(x,\tau|y,0) = \rho_s(x). \tag{2.33}$$

Clearly, the stationary distribution $\rho_s(x)$ is a solution of the stationary forward Chapman-Kolmogorov differential equation, i.e.

$$\begin{aligned} 0 = &-\sum_{i=1}^{N}\frac{\partial}{\partial x_i}\left[A_i(x)\rho_s(x) - \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}D_{ij}(x)\rho_s(x)\right] + \\ &+\int_{\Omega}\mathrm{d}z\left[W(x|z)\rho_s(z) - W(z|x)\rho_s(x)\right], \end{aligned} \tag{2.34}$$

$$\tag{2.35}$$

Given the definition of continuous Markov process (Eq. (2.31)), it should be noted that a Markov process is continuous if the quantity $W(x|y,t)$ (Eq. (2.27)) is identically zero, i.e. $W(x|y,t)\equiv 0$. By taking $W(x|y,t)\equiv 0$ into Eqs. (2.26),(2.30) one obtains the so-called Kolmogorov equations

$$\begin{aligned} \frac{\partial G(x,t_1|y,t_0)}{\partial t_1} &= -\sum_{i=1}^{N}\frac{\partial}{\partial x_i}\left[A_i(x,t_1)G(x,t_1|y,t_0) - \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}D_{ij}(x,t_1)G(x,t_1|y,t_0)\right] = \\ &= \mathcal{L}_{fw}[G(x,t_1|y,t_0)], \end{aligned} \tag{2.36}$$

$$\begin{aligned} \frac{\partial G(x,t_1|y,t_0)}{\partial t_0} &= -\sum_{i=1}^{N}\left(A_i(y,t_0)\frac{\partial G(x,t_1|y,t_0)}{\partial y_i}\right) - \frac{1}{2}\sum_{i,j=1}^{N}\left(D_{ij}(y,t_0)\frac{\partial^2 G(x,t_1|y,t_0)}{\partial y_i\partial y_j}\right) = \\ &= \mathcal{L}_{bw}[G(x,t_1|y,t_0)], \end{aligned} \tag{2.37}$$

where the forward and backward Kolomogorov operators are defined as

$$\mathcal{L}_{fw}(\cdot) := \mathcal{L}^{\dagger}(\cdot) = -\sum_{i=1}^{N}\frac{\partial}{\partial x_i}\left[\left[A_i(x,t)\,(\cdot)\right] - \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}\left[D_{ij}(x,t)\,(\cdot)\right]\right], \tag{2.38}$$

$$\mathcal{L}_{bw}\,(\cdot) = -\mathcal{L}\,(\cdot) = -\sum_{i=1}^{N}A_i(x,t)\frac{\partial}{\partial x_i}\,(\cdot) - \frac{1}{2}\sum_{i,j=1}^{N}D_{ij}(x,t)\frac{\partial^2}{\partial x_i\partial x_j}\,(\cdot)\,. \tag{2.39}$$

Note that the backward Kolmogorov operator $\mathcal{L}_{bw}$ is the adjoint of the forward operator $\mathcal{L}_{fw}$, except for a minus sign. Eq. (2.36) is also known as Fokker-Planck equation and the operator $\mathcal{L}^\dagger$ as Fokker-Planck operator.

A Markov process that satisfies Eqs. (2.36),(2.37) is known as diffusion process. A diffusion process can be described either through the Kolmogorov equations for conditional probabilities (Eqs. (2.36),(2.37)) or through suitable stochastic differential equations that describe the temporal evolution of a sample path. To be more precise, let $X(t)$ be a stochastic process which evolves through the stochastic differential equation (Ito form)[5]

$$\dot{X}(t) = a(X(t), t) + \sigma(X(t), t)\xi(t), \tag{2.40}$$

where $\xi(t)$ is a Gaussian white noise, i.e. $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(t_1)\xi_j(t_2) \rangle = \delta_{ij}\delta(t_1 - t_2)$. It is possible to prove [Gardiner et al., 1985] that the propagator of this process $G(x, t|y, t_0)$ satisfies a Fokker-Planck equation (Eq. (2.36)) with $A(x, t) \equiv a(x, t)$ and $D(x, t) = \sigma(x, t)\sigma^T(x, t)$.

In Eqs. (2.38),(2.39) the two operator $\mathcal{L}$ and $\mathcal{L}^\dagger$ have been introduced. Similarly to the case of Markov chains, for a homogeneous diffusion process $X(t)$ one has that the Fokker-Planck operator $\mathcal{L}^\dagger$ describes the evolution of the probability distributions while its adjoint operator $\mathcal{L}$ describes the evolution of the observables. Indeed, let $p(x, t_0)$ be the initial probability distribution of the system. The probability distribution at time t is given by Eq. (2.24). Taking the time derivative of Eq. (2.24) leads to

$$\partial_t p(x, t) = \int_\Omega dy p(y, t_0)\partial_t G(x, t|y, t_0) = \int_\Omega dy p(y, t_0)\mathcal{L}^\dagger \left[G(x, t|y, t_0)\right] = \mathcal{L}^\dagger[p(x, t)], \tag{2.41}$$

where in the last passage the linearity of $\mathcal{L}^\dagger$ has been used. By virtue of Eq. (2.41) if the probability distribution $p(x, t)$ converges to a stationary distribution $\rho_s(x)$ as $t \to +\infty$, the latter satisfies the stationary Fokker-Planck equation, i.e.

$$\mathcal{L}^\dagger[\rho_s(x)] = 0. \tag{2.42}$$

Concerning the observables, let $f(x, t)$ be a function defined as

$$f(x, t) = \int_\Omega \mathrm{d}y g(y) G(y, t|x, 0) = \int_\Omega \mathrm{d}y g(y) G(y, 0|x, -t), \tag{2.43}$$

[5]This thesis deals only with Ito stochastic differential equations.

where the second equality is due to the homogeneity of the process. By differentiating Eq. (2.43) with respect to time t one obtains

$$\partial_t f(x,t) = \int_\Omega \mathrm{d}y g(y) \partial_t G(y,0|x,-t) = \int_\Omega \mathrm{d}y g(y) \mathcal{L}\left[G(x,0|y,-t)\right] = \mathcal{L}[f(x,t)]. \tag{2.44}$$

In the next subsection the committor function will be introduced and it will be shown that it is the solution of a Dirichlet problem involving the backward operator $\mathcal{L}$.

2.3.2 Committor functions: definition and equation

Let $X(t)$ be a homogeneous diffusion process which takes values in the phase space Ω. The process evolves through a stochastic differential equation of the form

$$\dot{X}(t) = a(X(t)) + \sigma(X(t))\xi(t), \tag{2.45}$$

where $\xi(t)$ is a Gaussian white noise.

Let $\mathcal{C} \subset \Omega$ a subset of the phase space. The first hitting time $\tau_{\mathcal{C}}(x)$ is defined as the smallest time for which $X(t)$ belongs to $\mathcal{C}$, as a function of the initial condition $X(0) = x$, i.e.

$$\tau_{\mathcal{C}}(x) = \inf\{t : X(t) \in \mathcal{C} | X(0) = x\}. \tag{2.46}$$

Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of the phase space Ω, i.e. $\mathcal{A}, \mathcal{B} \subset \Omega$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$. The committor function $q(x)$ for the sets $\mathcal{A}$ and $\mathcal{B}$ is the probability that a trajectory starting at point x visits $\mathcal{B}$ first rather than $\mathcal{A}$ first [Onsager, 1938, E et al., 2005, Vanden-Eijnden, 2006, Metzner et al., 2006, Bovier, 2006, Metzner, 2008, Metzner et al., 2009, Prinz et al., 2011,Schütte et al., 2011,Bowman et al., 2013,Schütte and Sarich, 2015,Lopes and Lelièvre, 2019, Thiede et al., 2019]:

$$q(x) = \mathbb{P}(\tau_{\mathcal{B}}(x) < \tau_{\mathcal{A}}(x)). \tag{2.47}$$

As explained in Sec. 2.1 Eq. (2.47) is a general definition and is not restricted exclusively to Markov processes but, in the case of Markov processes, the committor function $q(x)$ is also the solution of a Dirichlet problem involving the adjoint of the Fokker-Planck operator [E et al., 2005, Metzner, 2008, Thiede et al., 2019].

To deal with the first hitting times problems, it is convenient to note that the Fokker-Planck equation (Eq. (2.36)) can be written in the form of a continuity equation

$$\frac{\partial G(x,t|y,0)}{\partial t} = \mathcal{L}^{\dagger}[G(x,t|y,0)] = -\nabla \cdot J(x,t|y,0) \tag{2.48}$$

where $J_i(x,t|y,0) = A_i(x,t)G(x,t|y,0) - \frac{1}{2}\sum_j \frac{\partial}{\partial x_j}\left[D_{ij}(x,t)G(x,t|y,0)\right]$.

In order to derive the differential equation that defines $q(x)$, it is useful to consider the set $\mathcal{S} = \mathcal{A} \cup \mathcal{B}$ as an absorbing set[6]. This means that the propagator $G(x,t|y,0)$ must satisfy a backward Kolmogorov equation (Eq. (2.37)) with the condition $G(x,t|y,0) = 0$ for $y \in \mathcal{S}$. If the process $X(t)$ is ergodic, it is certain that it will hit the frontier of $\mathcal{S}$, denoted by $\partial\mathcal{S}$, if it is observed for a time long enough. However, the committor function is not the probability that the system hits $\mathcal{S}$ but the probability that it hits the part of $\mathcal{S}$ corresponding to $\mathcal{B}$.

Let $z(x;b)|\partial\mathcal{S}(b)|$ be the probability that the process hits $\mathcal{S}$ at point b, where $|\partial\mathcal{S}(b)|$ denotes the area of the surface element at point b. Then, the committor function is nothing but

$$q(x) = \int_{\partial\mathcal{B}} \mathrm{d}b\, z(x;b)|\partial\mathcal{S}(b)|. \tag{2.49}$$

The quantity $z(x;b)$ can be expressed as a function of the current $J(b,t|x,0)$ as follows

$$z(x;b)|\partial\mathcal{S}(b)| = \int_0^{+\infty} \mathrm{d}t\, J(b,t|x,0)\cdot\partial\mathcal{S}(b). \tag{2.50}$$

By considering that $\frac{\partial G(x,t|y,0)}{\partial t} = \mathcal{L}G(x,t|y,0)$ one has

$$\begin{aligned}\int_0^{+\infty} \mathrm{d}t\, \partial_t J(b,t|x,0)\cdot\partial\mathcal{S}(b) &= \mathcal{L}\left[\int_0^{+\infty} \mathrm{d}t\, J(b,t|x,0)\cdot\partial\mathcal{S}(b)\right] = \\ &= \mathcal{L}\left[z(x;b)|\partial\mathcal{S}(b)|\right] = -J(b,0|x,0)\cdot\partial\mathcal{S}(b) = 0 \quad \text{if} \quad x \neq b.\end{aligned} \tag{2.51}$$

Eq. (2.51) shows that $z(x;b)$ is an element of the kernel of the operator $\mathcal{L}$ for $x \neq b$. Since $\mathcal{L}$ is a linear operator one obtains

$$\mathcal{L}[q(x)] = \mathcal{L}\left[\int_{\partial\mathcal{B}} \mathrm{d}b\, z(x;b)|\partial\mathcal{S}(b)|\right] = \int_{\partial\mathcal{B}} \mathrm{d}b\, \mathcal{L}[z(x;b)]|\partial\mathcal{S}(b)| = 0. \tag{2.52}$$

[6]The derivation of the differential equation for the committor closely follows the derivation for the distribution of exit points in [Gardiner et al., 1985].

Thus, it has been shown that the committor obeys to the equation $\mathcal{L}[q(x)] = 0$ for $x \notin \mathcal{S}$. To complete this equation, it is necessary to specify the proper boundary conditions. Clearly, it is certain to exit from the point b if the process starts from b, therefore $z(b;b)|\partial\mathcal{S}(b)| = 1$. Furthermore, since all the points of the set $\mathcal{S}$ are absorbing, for any point $x \in \mathcal{S}$ such that $x \neq b$ the probability that the process will leave the region through b is equal to 0. Therefore, one has that $q(x) = 0$ for $x \in \mathcal{A}$ and $q(x) = 1$ for $x \in \mathcal{B}$. Putting all together, one obtains that the committor function $q(x)$ is the solution of the Dirichlet problem

$$\begin{cases} q(x) = 0 & \text{if} \quad x \in \mathcal{A}, \\ q(x) = 1 & \text{if} \quad x \in \mathcal{B}, \\ \mathcal{L}[q(x)] = 0 & \text{if} \quad x \notin \mathcal{A} \cup \mathcal{B}. \end{cases} \tag{2.53}$$

In the next subsection, two examples where the problem 2.53 can be solved analytically will be considered.

2.3.3 Examples: Wiener and Ornstein-Uhlenbeck processes

In this subsection the committor function is computed for two simply one-dimensional Markov processes: the Wiener process and the Ornstein-Uhlenbeck process.

A Wiener process $X(t) \equiv W(t)$ is defined by the formal equation

$$\dot{W}(t) = \sqrt{2D}\xi(t), \tag{2.54}$$

or, equivalently, its propagator $G(x,t|y,0)$ obeys at the Fokker-Planck equation

$$\partial_t G(x,t|y,0) = D\frac{\partial^2 G(x,t|y,0)}{\partial x^2}. \tag{2.55}$$

From a physical point of view, the Wiener process can be thought of as a process that describes the position of a colloidal particle immersed in a viscous fluid.

The phase space of the system Ω coincides with the real line, that is $\Omega \equiv \mathbb{R}$. Consider the sets $\mathcal{A} = (-\infty, a]$ and $\mathcal{B} = [b, +\infty)$ with $a < b$. The committor function $q(x)$ corresponds to the probability that the process $W(t)$ enters in $\mathcal{B}$ before entering in $\mathcal{A}$ knowing that $W(0) = x$. Since in this case

the Fokker-Planck operator is self-adjoint, $q(x)$ is the solution of the Dirichlet problem

$$D\frac{\partial^2 q(x)}{\partial x^2} = 0, \quad q(a) = 0, \quad q(b) = 1. \tag{2.56}$$

The solutions of the partial differential equation $D\frac{\partial^2 q(x)}{\partial x^2} = 0$ are linear functions of the form $q(x) = \alpha x + \beta$. The two coefficient α and β are found by imposing the boundary conditions $q(a) = 0$ and $q(b) = 1$, obtaining $\alpha = \frac{1}{b-a}$ and $\beta = -\frac{a}{b-a}$. Hence, the committor function $q(x)$ is

$$q(x) = \frac{x-a}{b-a}. \tag{2.57}$$

Eq. (2.57) shows that for the Wiener process the committor function grows linearly from the value 0 in $\mathcal{A}$ to the value 1 in $\mathcal{B}$. The growth is independent on the diffusion coefficient D and the slope only depends on the length of the interval. Furthermore, it can be seen that to have a probability greater than 0.5 to reach first $\mathcal{B}$ rather than $\mathcal{A}$, the process must start beyond the middle of the range, i.e. $x > \frac{a+b}{2}$. Thus, this example is analogous of the fair game in the gambler's ruin problem (see the last expression in Eq. (2.22)), where the odds of winning or losing a bet are equal. This is not surprising since the Wiener process is exactly the generalization of the random walk in a continuous space. To obtain the exponential solution equivalent to that of Eq. (2.22), it is necessary to consider a drift term in Eq. (2.54). This means to consider the process $X(t)$ defined by

$$\dot{X}(t) = \mu + \sqrt{2D}\xi(t), \tag{2.58}$$

which can be thought to describes the position of a colloidal particle immersed in a viscous fluid and driven by a constant force μ. For instance, it could represent the position of a pollen grain in water subject to gravitational force.

Also in this case, the Fokker-Planck operator is self-adjoint and it is equal to $\mathcal{L}^\dagger = -\mu\frac{\partial}{\partial x} + D\frac{\partial^2}{\partial x^2}$. Hence, the committor function is defined by the Dirichlet problem

$$\mu\frac{\partial q(x)}{\partial x} + D\frac{\partial^2 q(x)}{\partial x^2} = 0, \quad q(a) = 0, \quad q(b) = 1, \tag{2.59}$$

whose solution is

$$q(x) = \frac{1 - \exp\left(-\frac{\mu}{D}(x-a)\right)}{1 - \exp\left(-\frac{\mu}{D}(b-a)\right)}. \tag{2.60}$$

It can be noted that Eq. (2.60) reduces to Eq. (2.57) when $\frac{\mu}{D} \to 0$. It means that, when the drift term μ is much smaller than the diffusion coefficient D, the process $X(t)$ roughly behaves like a Wiener process $W(t)$ and the committor function $q(x)$ is almost linear. Instead, in the case $|\mu| \gg D$, the committor function is almost constant except for a region of size $\frac{D}{\mu}$ around a or b, according to $\mu > 0$ or $\mu < 0$, where $q(x)$ changes abruptly from 0 to 1. In fact, if $\mu > 0$ and $\mu \gg D$, $q(x) \sim 1$ for all x such that $(x - a) > \frac{D}{\mu}$. In the opposite limit, i.e. when $\mu \to -\infty$, one has that $q(x) \sim 0$ except for a tiny region around b.

Fig. 2.2 shows the committor function $q(x)$ for the three cases $\mu = 0$, $\mu > 0$ and $\mu < 0$. It can be noted that although the ratio $\frac{\mu}{D}$ is of order $O(1)$,

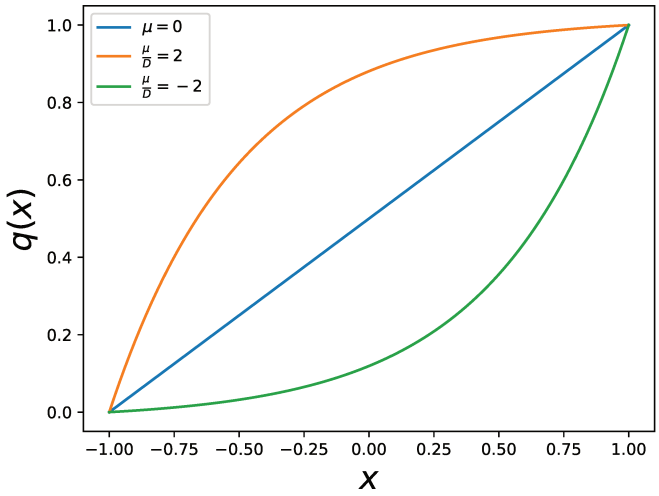


Figure 2.2: Committor function $q(x)$ for the process defined by Eq. (2.58). The sets $\mathcal{A} = (-\infty, -1]$ and $\mathcal{B} = [1, +\infty)$ have been considered. The blue, orange and green curves indicate the case of a pure diffusion ($\mu = 0$), positive drift coefficient ($\mu > 0$) or negative drift coefficient ($\mu < 0$), respectively.

the behavior of the committor is quite different from one case to another. When $\mu = 0$ the committor function is linear, while for $\mu \neq 0$ it is an exponential function. For $\mu > 0$, the particle is driven towards positive and therefore, for many initial positions, it will be more likely to visit $\mathcal{B}$ first than $\mathcal{A}$. In the opposite case ($\mu < 0$) the particle is driven towards the set $\mathcal{A}$ and therefore the probability to reach $\mathcal{B}$ first is small except if the initial position of the particle is chosen in the neighborhood of $\mathcal{B}$. If the initial position of

the particle is chosen to be the middle of the interval $[a, b]$, the probabilities to reach $\mathcal{B}$ before $\mathcal{A}$ are 50%, 85% and 10% for the three cases $\mu = 0$, $\mu = 2D$ and $\mu = -2D$, respectively. The comparison between Fig. 2.2 and Fig. 2.1 and between the values of the committor in the middle point of the interval further highlights the similarity between the two problems, showing that the process $X(t)$ (Eq. (2.58)) is the continuous equivalent of the Markov chain considered in Sec. 2.2.3.

The second example that is considered is the Ornstein-Uhlenbeck process. Historically, it was the first stochastic differential equation to appear in physics, introduced in 1908 by the mathematician Paul Langevin to describe the velocity of a Brownian particle. The stochastic differential equation that describes this process is the following:

$$\dot{X}(t) = -kX(t) + \sqrt{2D}\xi(t). \tag{2.61}$$

In this equation the drift coefficient acts as a restoring force and tries to bring the process towards its equilibrium position as long as $k > 0$, while for $k < 0$ the drift tends to force the system away from the equilibrium. For such a process, the Dirichlet problem defining the committor (see Eq. (2.53)) takes the form

$$-kx\frac{\partial q(x)}{\partial x} + D\frac{\partial^2 q(x)}{\partial x^2} = 0, \quad q(a) = 0, \quad q(b) = 1, \tag{2.62}$$

where the sets $\mathcal{A} = (-\infty, a]$ and $\mathcal{B} = [b, +\infty)$, with $a < b$, have been considered. It is quite simple to show that the solution of Eq. (2.62) is given by

$$q(x) = \frac{\int_a^x \mathrm{d}y \exp\left(\frac{k}{2D}y^2\right)}{\int_a^b \mathrm{d}y \exp\left(\frac{k}{2D}y^2\right)}. \tag{2.63}$$

Thus, the committor function depends on the parameter $\gamma = \frac{k}{2D}$. In the limit $\gamma \to 0$, the process $X(t)$ behaves roughly like a Wiener process and the committor function becomes approximately linear. Instead, for $|\gamma| \to +\infty$, since the integrands of Eq. (2.63) are exponentials and therefore the integrals are dominated by the maximum of the integrands over the interval $[a, b]$, the behavior of the committor strongly depends on the end points of the interval $[a, b]$.

The committor function for the three cases $\gamma \to -\infty$, $\gamma \to +\infty$ and $\gamma \to 0$ is shown in Fig. 2.3. It can be seen that for small values of γ the committor is

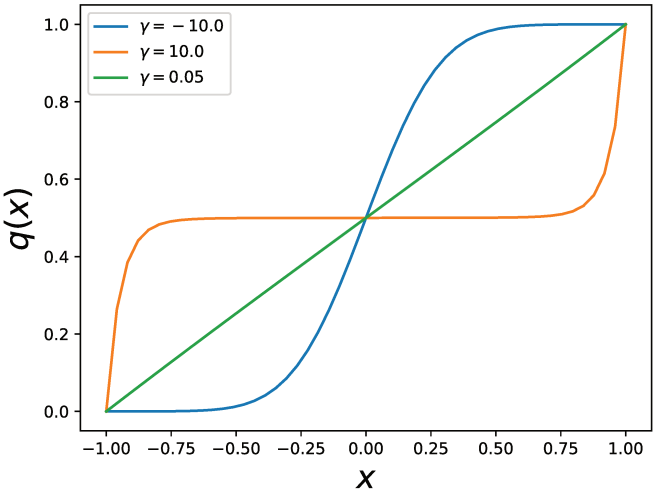


Figure 2.3: Committor function $q(x)$ for the Ornstein-Uhlenbeck process. The sets $\mathcal{A} = (-\infty, -1]$ and $\mathcal{B} = [1, +\infty)$ have been considered. The blue, orange and green curves indicate the three cases $\gamma \ll 0$, $\gamma \gg 0$ and $\gamma \sim 0$, respectively.

indistinguishable from a linear function. For $\gamma \gg 0$, the committor function assumes the value $\frac{1}{2}$ almost everywhere, except for two small region of size $\frac{1}{\gamma}$ around $\mathcal{A}$ or $\mathcal{B}$ where $q(x)$ passes from 0 to $\frac{1}{2}$ and from $\frac{1}{2}$ to 1, respectively. In the limit $\gamma \to -\infty$, the committor function is equal to 0 in a wide interval close to the region $\mathcal{A}$ and it is equal to 1 in the neighborhood of $\mathcal{B}$, while it has a sharp transition from 0 to 1 around the point $x = 0$, which corresponds to the point where the integrand of Eq. (2.63) takes its maximal value. It should also be noted that, for $\gamma > 0$, $q(x)$ is bigger than the linear committor for $x < 0$ while it is smaller than the linear committor for $x > 0$. For $\gamma < 0$ the opposite is true and the committor function is smaller than a linear function for $x < 0$ and it is bigger for $x > 0$. This is consistent with the fact that for $k > 0$ the system is attracted towards equilibrium while in the opposite case the system is moved away from it.

To conclude this section, it is important to note that, for stochastic systems in one dimension of the form

$$\dot{X}(t) = A(X(t)) + \sqrt{2D(X(t))}\xi(t), \tag{2.64}$$

the committor function of two sets $\mathcal{A} = (-\infty, a]$ and $\mathcal{B} = [b, +\infty)$ always

has an explicit formula, i.e.

$$q(x) = \frac{\int_a^x \mathrm{d}y \, [\psi(y)]^{-1}}{\int_a^b \mathrm{d}y \, [\psi(y)]^{-1}}, \tag{2.65}$$

where

$$\psi(y) = \exp\left[\int_a^y \mathrm{d}x \frac{A(x)}{D(x)}\right]. \tag{2.66}$$

2.4 Numerical computation of the committor

In the previous section the theory of the committor function for both Markov chains and Markov processes has been introduced. This section, instead, will be focused on presenting how to compute committor functions with numerical methods. In particular, three methods will be introduced. The first one, called direct estimation, consists in computing the committor from its definition (Eq. (2.47)) by performing Monte Carlo experiments. The second one, which is valid exclusively for ergodic dynamics, still aims to compute the committor from its definition (Eq. (2.47)), but this is done by averaging suitable functions over long trajectories. The third option, valid for diffusion processes, consists in solving numerically the Dirichlet problem which defines the committor (see Eq. (2.53)). Then, some machine learning methods that estimate the committor function from observations are briefly mentioned. Finally, it is explained how to validate an estimation of the committor function. To this purpose, two score functions are introduced: the Brier score and the logarithmic score.

2.4.1 Direct estimation

The direct estimation consists in performing Monte Carlo experiments for computing the committor function of a process $X(t)$. Such a method is valid for both deterministic and stochastic dynamics, since it only requires to be able to generate N different realizations of the process with the same initial condition.

To be more precise, suppose that one needs to compute the committor function $q(x)$ at point x for two sets $\mathcal{A}$ and $\mathcal{B}$. To achieve this result, N realizations of the process $X(t)$ can be initialized at $X(0) = x$ and then

evolved through the dynamics until they reach $\mathcal{A}$ or $\mathcal{B}$. Let $N_{\mathcal{B}}(x)$ denote the number of realizations that ends up in $\mathcal{B}$. An estimation of the committor $q(x)$ is given by

$$q(x) = \frac{N_{\mathcal{B}}(x)}{N}, \tag{2.67}$$

which converges to the exact value in the limit $N \to +\infty$.

If one needs to compute the committor over the entire phase space Ω, a similar procedure can be employed. Indeed, let $\{\mathcal{C}_j\}_{1 \leq j \leq J}$ be a partition of the phase space Ω, i.e. $\mathcal{C}_{j_1} \cap \mathcal{C}_{j_2} = \emptyset$ for $j_1 \neq j_2$ and $\cup_{j=1}^{J} \mathcal{C}_j = \Omega$. Then, for each set $\mathcal{C}_j$, one can generate N different realizations of the process $X(t)$ with initial condition $X(0) = x \in \mathcal{C}_j$. Let $N_{\mathcal{B}}(\mathcal{C}_j)$ be the number of trajectories that reach the set $\mathcal{B}$ first, knowing that they start inside the set $\mathcal{C}_j$. The committor function $q(x)$, for any point $x \in \mathcal{C}_j$, is

$$q(x) = \frac{N_{\mathcal{B}}(\mathcal{C}_j)}{N}, \tag{2.68}$$

which converges to the real committor when $N \to +\infty$ and $J \to +\infty$.

While this procedure is conceptually simple, it has at least two drawbacks. The first one is that this approach is feasible as long as one is able to generate precise initial conditions. Although this can be done in numerical simulations, it may not be the case in real laboratory experiments. The second drawback is related to its computational cost. In fact, since one has to observe N different realizations of the process until they reach one of the two sets, the time for a realization to reach the sets must be sufficiently small. Furthermore, even if the time to hit the two sets is small enough, this method is impractical for high dimensional systems, since the number J of sets $\mathcal{C}_j$ needed to partition the phase space Ω grows exponentially with the dimension.

2.4.2 Committor computation based on a long trajectory

For an ergodic process, being it deterministic or stochastic, the committor function can be computed by using the information contained in one (or more) long trajectory.

Let $\mathbb{1}_{\tau_{\mathcal{B}}<\tau_{\mathcal{A}}}$ be the variable defined as

$$\mathbb{1}_{\tau_{\mathcal{B}}<\tau_{\mathcal{A}}} = \begin{cases} 1 & \text{if} \quad \tau_{\mathcal{B}} < \tau_{\mathcal{A}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.69}$$

From the definition of the committor (Eq. (2.47)) it can be seen that

$$q(x) = \mathbb{E}_x \left[\mathbb{1}_{\tau_{\mathcal{B}}<\tau_{\mathcal{A}}} \right], \tag{2.70}$$

where $\mathbb{E}_x [\cdot]$ denotes the average over the stationary distribution of the process conditioned to the constrain $X(0) = x$. Furthermore, it should be noted that Eq. (2.47) can be rewritten as

$$q(x) = \mathbb{P}(\tau_{\mathcal{B}} < \tau_{\mathcal{A}} | X(0) = x) = \frac{\mathbb{P}(\tau_{\mathcal{B}} < \tau_{\mathcal{A}}; X(0) = x)}{\mathbb{P}(X(0) = x)}, \tag{2.71}$$

where the numerator stands for the joint probability of observing $\tau_{\mathcal{B}} < \tau_{\mathcal{A}}$ and $X(0) = x$, while the denominator is simply given by $\rho_s(x)dx$. Therefore, one has

$$\rho_s(x)q(x)dx = \mathbb{P}(\tau_{\mathcal{B}} < \tau_{\mathcal{A}}; X(0) = x) = \mathbb{E}\left[\mathbb{1}_{\tau_{\mathcal{B}}<\tau_{\mathcal{A}}} \delta(X(0) - x) \right]. \tag{2.72}$$

By using the ergodicity for replacing the statistical average in Eq. (2.72) with temporal averages one obtains

$$\rho_s(x)q(x) = \lim_{T \to +\infty} \frac{1}{T} \int_0^T \delta(X(t) - x) \mathbb{1}_{\tau_{\mathcal{B}}<\tau_{\mathcal{A}}}, \tag{2.73}$$

while the stationary distribution is given by

$$\rho_s(x) = \lim_{T \to +\infty} \frac{1}{T} \int_0^T \delta(X(t) - x). \tag{2.74}$$

By introducing a spatial and temporal discretization, Eqs. (2.73),(2.74) can be used to compute an estimation of the committor function [Lopes and Lelièvre, 2019,Lucente et al., 2019]. As in the previous section, let $\{\mathcal{C}_j\}_{1 \leq j \leq J}$ be a partition of the phase space Ω. Let $\{X_n\}_{1 \leq n \leq N_t}$ be a realization of the process observed at discrete time $t_n = n\Delta t$ with $T = N_t \Delta t$. The probability for the system to be in the set $\mathcal{C}_j$, denoted by $p(\mathcal{C}_j)$, is equal to

$$p(\mathcal{C}_j) = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbb{1}_{X_n \in \mathcal{C}_j}, \tag{2.75}$$

while the probability that it is in the set $\mathcal{C}_j$ and that it will visits the set $\mathcal{B}$ before $\mathcal{A}$, indicated by $p_{\mathcal{B}}(\mathcal{C}_j)$ is simply given by

$$p_{\mathcal{B}}(\mathcal{C}_j) = \frac{1}{N_t}\sum_{n=1}^{N_t} \mathbb{1}_{X_n\in\mathcal{C}_j}\mathbb{1}_{\tau_{\mathcal{B}}(X_n)<\tau_{\mathcal{A}}(X_n)}. \tag{2.76}$$

Finally, by combining Eqs. (2.75),(2.76) the committor function turns out to be

$$q(\mathcal{C}_j) = \frac{p_{\mathcal{B}}(\mathcal{C}_j)}{p(\mathcal{C}_j)} = \frac{\sum_{n=1}^{N_t} \mathbb{1}_{X_n\in\mathcal{C}_j}\mathbb{1}_{\tau_{\mathcal{B}}(X_n)<\tau_{\mathcal{A}}(X_n)}}{\sum_{n=1}^{N_t} \mathbb{1}_{X_n\in\mathcal{C}_j}}. \tag{2.77}$$

Although both the direct estimation of the committor and the computation based on a long trajectory rely on Eq. (2.47), the latter should be preferred since it uses all the informations that a trajectory carries with it. However, as it will be seen in the next chapter, this method may suffer from sampling issues. In fact, a finite trajectory cannot visit the entire phase space Ω. Therefore, meaningful results can be obtained only for the part of the phase where the dynamics spends enough time.

2.4.3 Committor computation solving the Backward equation

It has be seen in Sec. 2.3.2 that for a continuous Markov process (i.e. a diffusion process) the committor function $q(x)$ is the solution of a second order partial differential equation (Eq. (2.53)). Hence, the committor function can be computed by solving Eq. (2.53) numerically. This can be achieved by employing any integration scheme for partial differential equation, for instance by employing the finite difference method which is illustrated here for one dimensional dynamics.

The finite difference method aims to approximate the derivatives of a function with finite difference equations [Smith et al., 1985]. To derive this method, consider the Taylor expansions of the committor function $q(x)$ around the point x:

$$q(x+\Delta x) = q(x) + \frac{\partial q(x)}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 q(x)}{\partial x^2}\Delta x^2 + O(\Delta x^3), \tag{2.78}$$

$$q(x-\Delta x) = q(x) - \frac{\partial q(x)}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 q(x)}{\partial x^2}\Delta x^2 + O(\Delta x^3). \tag{2.79}$$

By adding or subtracting Eqs. (2.78),(2.79) one obtains

$$\begin{aligned}\frac{\partial^2 q(x)}{\partial x^2} &\simeq \frac{q(x+\Delta x)+q(x-\Delta x)-2q(x)}{\Delta x^2},\\ \frac{\partial q(x)}{\partial x} &\simeq \frac{q(x+\Delta x)-q(x-\Delta x)}{2\Delta x}.\end{aligned} \tag{2.80}$$

These two formula suggest that the Dirichlet problem (2.53) can be approximated by a system of linear equations. In fact, let $\mathcal{A} = (-\infty, a]$ and $\mathcal{B} = [b, +\infty)$ be the two sets for which the committor function needs to be computed. It means that the set $(\mathcal{A} \cup \mathcal{B})^c$ where $\mathcal{L}[q(x)] = 0$ is equal to the interval $[a, b]$. To obtain the system of linear equations, consider $L+1$ points $x_l = a + l\Delta x$, with $x_L = b$, and let q_l denotes the committor function computed at point $x = x_l$. Clearly, the boundary conditions for the Dirichlet problem are $q_0 = 0$ and $q_L = 1$. Instead, for $1 \leq l \leq L-1$ one has

$$A_l \frac{q_{l+1} - q_{l-1}}{\Delta x} + \frac{1}{2} D_l \frac{q_{l+1} + q_{l-1} - 2q_l}{\Delta x^2} = 0, \tag{2.81}$$

where $A_l = A(x_l)$ and $D_l = D(x_l)$. By solving Eq. (2.81) one obtains an estimation of the committor that converges to the real one when $\Delta x \to 0$.

Unlike the two methods described in the previous sections, such an approach can only be used for diffusion processes. In spite of this limitation, it has the great advantage that its computational cost does not depend on the intensity of the noise. Therefore, when transitions between sets $\mathcal{A}$ and $\mathcal{B}$ are rare it is preferable to use this method since generating several trajectories ending in $\mathcal{A}$ or $\mathcal{B}$ or a single trajectory long enough to cover a relevant portion of the phase space can take a very long time and may not be feasible in practice [Rolland and Simonnet, 2015].

Solving the problem (2.53) numerically is feasible as long as the system under investigation has few degrees of freedom. When the dynamics takes place in an high dimensional phase space Ω the number of linear equations to be solved grows exponentially with the dimension, highlighting the impossibility of finding a solution with finite computational resources.

2.4.4 Machine learning estimation

The style of this subsection slightly deviates from the style of the chapter. In fact, the purpose here is to mention the possibility of estimating committor functions from observations. A rigorous treatment of the methods

mentioned here is beyond the scope of this thesis. Nonetheless, it is important to mention recent works which are somehow related to the topics of this manuscript.

Many interesting methods have been or are currently being devised to learn committor functions: based on direct machine learning [Pozun et al., 2012], using a characterization of the committor function for diffusions as a solution of a partial differential equation [Khoo et al., 2019, Li et al., 2019], computing the committor function from a finite state Markov chain [Schütte et al., 1999, Prinz et al., 2011, Noé and Rosta, 2019, Tantet et al., 2015], possibly a Markov state model approximation of the dynamics [Ulam, 2004]. Recently a very interesting approach has been considered starting from a Galerkin approximation of the dynamics generator, or the Koopman operator. Finite dimensional approximations of the dynamics generator have been used to identify good reaction coordinates [Froyland et al., 2014, Bittracher et al., 2018], or to evaluate eigenfunctions of the operator [Giannakis et al., 2015, Giannakis, 2019, Williams et al., 2015, Mardt et al., 2018], sometimes with climate applications [Giannakis et al., 2015, Giannakis, 2019]. Recently such direct Galerkin approximation has been used to directly compute committor function, avoiding the burden of discretizing a high dimensional phase space [Thiede et al., 2019, Strahan et al., 2021].

In Chapter 4 a new method for estimating committor functions from dynamical observations will be developed.

2.4.5 Validation of a committor estimate

Having introduced several numerical methods to compute the committor function of a system, it is natural to ask how to evaluate the quality of a particular estimate. First of all, note that the committor function $q(x)$ is the probability of success of a spatially dependent Bernoulli trial where the two possible outcomes are the occurrence of the event $\tau_{\mathcal{B}}(x) < \tau_{\mathcal{A}}(x)$ or $\tau_{\mathcal{A}}(x) \leq \tau_{\mathcal{B}}(x)$ whose probabilities are $q(x)$ and $1 - q(x)$, respectively. Thus, the predictive problem to be faced consists in assigning a probability to two mutually exclusive events. The two mutually exclusive events are the two possible values assumed by the random variable $\mathbb{1}_{\tau_{\mathcal{B}}(x)<\tau_{\mathcal{A}}(x)}$ which for brevity will be denoted by $y(x)$, i.e. $y(x) \equiv \mathbb{1}_{\tau_{\mathcal{B}}(x)<\tau_{\mathcal{A}}(x)}$.

Therefore, the objective of estimating the committor function is to find the best approximation of the joint probability $P(x, y)$ of two events (x, y), with $x \in \Omega$ and $y \in \{0, 1\}$. Note that these random variables are not in-

dependent. Indeed, it has been argued that the conditional distribution of y with respect to x is Bernoulli with parameter $q(x)$, where q is the committor function. Thus, by assuming that the process is stationary (i.e. x is distributed according the stationary distribution $\rho_s(x)$), one has

$$P(x,y) = P(x)P(y|x) = \rho_s(x)\left[q(x)\delta(y-1) + (1-q(x))\delta(y)\right], \tag{2.82}$$

where $\delta(\cdot)$ is the Dirac delta function.

In this context, evaluating an estimate of the committor function requires finding a way to quantify the discrepancy between the predictions made using $P(x,y)$ compared to those obtained using its approximation $\hat{P}(x,y)$. Since usually the committor function is computed using realizations of the process under investigation, or in any case, it is computed only in regions of the phase space mostly visited by the dynamics, it can be assumed without losing generality that also for the distribution $\hat{P}(x,y)$ the variable x is distributed according to the stationary distribution $\rho_s(x)$, i.e.

$$\hat{P}(x,y) = \hat{P}(x)\hat{P}(y|x) = \rho_s(x)\left[\hat{q}(x)\delta(y-1) + (1-\hat{q}(x))\delta(y)\right]. \tag{2.83}$$

The score functions to be used, denoted by $R(P,\hat{P})$, must have two fundamental properties. The first one is that R exhibits an extremant for $P = \hat{P}$, which is equivalent to requiring that $\frac{\partial R}{\partial \hat{P}} = 0$ for $\hat{P} = P$. The second fundamental property is that the function R has an empirical counterpart $\tilde{R}$ that can be computed from observations, since the true committor function is often unknown in real problems.

The request for an empirical counterpart $\tilde{R}$ is what distinguishes scores from distances (or pseudo distances). In fact, one could try to use a distance such as the $L_2^{\rho_s}$ norm or a pseudo distance such as the Kullback-Leibler divergence to compute the error of an estimate of the committor function. However, as will be well illustrated in the next two subsections, these distances cannot be computed from the data. The quantities that can be calculated instead are the scores associated with these two distances. The scores differ from distances only by an additive constant. The diagram in Fig. 2.4, together with the discussions in the following sections help to better understand this important concept.

Two score functions that are widely used in probabilistic forecasts, which are related to the $L_2^{\rho_s}$ norm and the Kullback-Leibler divergence, are the Brier score and the logarithmic score.

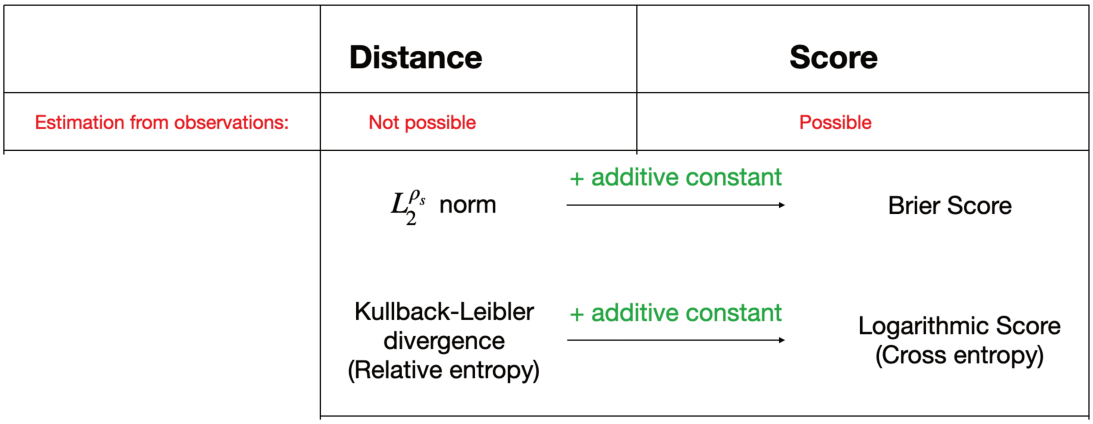


Figure 2.4: Diagram showing the relation between distances and scores. The main difference is that the score can be computed from observations while an estimate of the distance is accessible only when the true committor function is known. Furthermore, each distance is associated with its score function and they differ by an additive constant.

Brier score

Imagine to draw N points (X_n, y_n) with $n = 1, \cdots, N$ from the true probability distribution $P(x, y)$. For simplicity, assume that each point $X_n = x$ has been sampled N_x times, i.e. there are N_x different values for X_n denoted by x_{n_x} with $n_x = 1, \cdots, N_x$, and that for each point x_{n_x} there is a number N_y of independent forecasting instances $y_{n_y}(x_{n_x})$ with $n_y = 1, \cdots, N_y$. Thus, the total number of points N will be $N = N_x N_y$.

Consider an estimate $\hat{P}(x, y)$ of the true distribution $P(x, y)$, which differs from $P(x, y)$ only by the dependence of y on x as expressed in Eq. (2.83). The Brier score quantifies the difference between $\hat{P}(x, y)$ and $P(x, y)$ by averaging the stochastic variable $(y(x) - \hat{q}(x))$ over $P(x, y)$, i.e.

$$B^{\rho_s} = \mathbb{E}_P \left[(y(x) - \hat{q}(x)) \right] = \int \mathrm{d}x\mathrm{d}y \, P(x, y)(y(x) - \hat{q}(x)). \tag{2.84}$$

In its original formulation, proposed by Brier in 1950 [Brier, 1950], this score was designed to quantify the quality of a probabilistic prediction of a single event (i.e. not space-dependent) for a class of n independent forecasting instances. With the notation introduced above, this means considering

as a score the quantity $B_n(x)$ defined as

$$B_{N_y}(x) = \frac{1}{N_y} \sum_{n_y=1}^{N_y} (\hat{q}(x) - y_{n_y}(x))^2, \tag{2.85}$$

The space-dependent Brier score $B_{N_y}(x)$ is thus a random variable, with values between 0 and 1. The random variable $(\hat{q}(x) - y_{n_y}(x))^2$ can only assumes two values: $(\hat{q}(x) - 1)^2$ with probability $q(x)$ and $\hat{q}(x)^2$ with probability $1 - q(x)$. Therefore, the average value of $B_{N_y}(x)$ is

$$\mathbb{E}\left[B_{N_y}(x)\right] = (1-\hat{q}(x))^2 q(x) + \hat{q}(x)^2(1-q(x)) = q(x)(1-q(x)) + (\hat{q}(x)-q(x))^2. \tag{2.86}$$

Eq. (2.86) shows that the average space-dependent Brier score $\mathbb{E}\left[B_{N_y}(x)\right]$ can be decomposed into the sum of two positive terms. The first term represents the variance of a Bernoulli trial with parameter $q(x)$ and it is related with the intrinsic stochasticity of the forecast itself. Indeed, since the variable y is stochastic, there will always be an uncertainty in its prediction. This uncertainty is as great as $q(x)$ is close to the value 0.5 and it vanishes only for $q(x) = 0$ or $q(x) = 1$. The second term, instead, measures how different are the actual committor $q(x)$ and its approximation $\hat{q}(x)$. Such a term is always positive (since it is a square difference) and is equal to 0 if and only if $\hat{q}(x) = q(x)$. From these considerations it is therefore clear that the smaller the score, the more accurate the estimate of the committor function is.

In the limit $N_y \to +\infty$, according to the law of large numbers one has

$$\lim_{N_y \to \infty} B_{N_y}(x) = \mathbb{E}\left[B_{N_y}(x)\right]. \tag{2.87}$$

So far, exclusively how to quantify the quality of the estimate of the committor at a given point x has been discussed. To take into account also the contributions deriving from other regions of the phase space it is sufficient to add the contributions coming from different points:

$$B_N = \frac{1}{N_x} \sum_{n_x=1}^{N_x} \frac{1}{N_y} \sum_{n_y=1}^{N_y} (\hat{q}(x_{n_x}) - y_{n_y}(x_{n_x}))^2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{q}(X_n) - y_n)^2. \tag{2.88}$$

By considering the limits $N_x \to +\infty$ and $N_y \to +\infty$ (or equivalently the limit $N \to +\infty$) into Eq. (2.88) one obtains

$$\lim_{N_x \to \infty} \lim_{N_y \to \infty} B_N = \mathbb{E}_P\left[(y(x) - \hat{q}(x))\right] = B^{\rho_s}. \tag{2.89}$$

Eq. (2.89) shows that B_N is the empirical version of B^{ρ_s}, that is $B_N = \tilde{B}^{\rho_s}$. Furthermore, by using Eq. (2.86), it easy to see that

$$B^{\rho_s} = \| q - \hat{q} \|^2_{\rho_s} + \| \sqrt{q(1-q)} \|^2_{\rho_s}, \tag{2.90}$$

where $\| f \|^2_{\rho_s} = L_2^{\rho_s}(f) = \int \mathrm{d}x \, f^2(x)\rho_s(x)$.

The second term in the right hand-side of Eq. (2.90) depends only on the true committor $q(x)$ and therefore it is constant if one imagines the Brier score as a function of $\hat{q}(x)$. Thus, in all cases in which the real committor function is known, the difference of $\hat{q}(x)$ from $q(x)$ can be studied using the weighted norm $L_2^{\rho_s}$, since it is the only non-constant term in Eq. (2.90).

Therefore, it has been shown that the Brier score can be used for assessing the quality of an estimate of the committor function. In fact, it satisfies the two basic requirements for a score function and it has a simple expression which involves the $L_2^{\rho_s}$ distance. The major limitation, as noted in [Benedetti, 2010], is that it is not appropriate for dealing with rare (or frequent) events. In fact, consider a state x for which the committor function is relatively small, for example $q(x) = 10^{-3}$. Also suppose that two estimates of the committor function $\hat{q}^1(x) = 0$ and $\hat{q}^2(x) = 10^{-3}$ are available. These two approximations lead to the two values $\mathbb{E}\left[B^1_{N_y}(x)\right] = 10^{-3}$ and $\mathbb{E}\left[B^2_{N_y}(x)\right] = (1 - 10^{-3})10^{-3}$. Thus, by using the real committor function the score is reduced by only 0.1% showing that it is very unfair in evaluating forecasts of rare events.

Logarithmic score

The logarithmic score is nowadays widely used both for the validation of probabilistic predictions and in the field of machine learning. From the point of view of information theory, this can be understood given its relationship to relative entropy.

The relative entropy, also known as Kullback-Leibler divergence, was introduced by Kullback and Leibler in 1951 [Kullback and Leibler, 1951] as a measure to quantify how different two distributions are.

Let $P(x, y)$ and $\hat{P}(x, y)$ be two distributions, then the relative entropy is defined as [Kullback and Leibler, 1951, Kullback, 1997]

$$S(P(x,y)||\hat{P}(x,y)) = \int dxdy P(x,y) \log \frac{P(x,y)}{\hat{P}(x,y)}, \tag{2.91}$$

and it measures the information lost when $\hat{P}(x,y)$ is used to approximate $P(x,y)$.

By inserting the expressions (2.82),(2.83) for $P(x,y)$ and $\hat{P}(x,y)$ into Eq. (2.91), one obtains

$$\begin{aligned} S(P(x,y)||\hat{P}(x,y)) &= \int dxdyP(x,y)\log\frac{P(x,y)}{\hat{P}(x,y)} \\ &= \int dx\rho(x)\left[q(x)\log\frac{q(x)}{\hat{q}(x)} + (1-q(x))\log\frac{1-q(x)}{1-\hat{q}(x)}\right] \\ &= C(\hat{q}) + Const(q) = S^{\rho_s}, \end{aligned} \tag{2.92}$$

where

$$C(\hat{q}) = -\int dx\rho(x)[q(x)\log(\hat{q}(x)) + (1-q(x))\log(1-\hat{q}(x))], \tag{2.93}$$

and

$$Const(q) = \int dx\rho(x)[q(x)\log(q(x)) + (1-q(x))\log(1-q(x))]. \tag{2.94}$$

Note that S^{ρ_s} and $C(\hat{q})$ are well defined if and only if $\hat{q}(x) = 0 \Rightarrow q(x) = 0$ and $\hat{q}(x) = 1 \Rightarrow q(x) = 1$. Furthermore, it easy to verify that

$$\underset{\hat{q}}{\text{argmin}}\{S^{\rho_s}\} = \underset{\hat{q}}{\text{argmin}}\{C(\hat{q})\} = q. \tag{2.95}$$

Eq. (2.92) shows that the relative entropy S^{ρ_s} is the sum of two terms. The term $Const(q)$ is minus the spatial average over the stationary distribution of the entropy of a spatial dependent Bernoulli trial with parameter $q(x)$, which is consistent with the interpretation of the committor given above. It does not depend on $\hat{q}$ and corresponds to an additive constant in the equation for S^{ρ_s}. The other term is the cross entropy (or logarithmic score) between $P(x,y)$ and $\hat{P}(x,y)$ and it is related with the mutual information between the two probability distributions. It should be noted that the logarithmic score $C(\hat{q})$ contains all the dependence of S^{ρ_s} on $\hat{q}$ and therefore minimize S^{ρ_s} is equivalent to minimize $C(\hat{q})$.

To be a useful score function, the corresponding empirical score must exist. Unfortunately, the quantity S^{ρ_s} can be computed only when the true committor is known due to the term $Const(q)$ in Eq. (2.92). Nevertheless,

it is possible to provide an empirical version of the logarithmic score $C(\hat{q})$, denoted by $\tilde{C}(\hat{q})$. In fact, consider again N points (X_n, y_n) with $n = 1, \cdots, N$ drawn from the true distribution $P(x, y)$. Let $\tilde{C}_N(\hat{q})$ be the empirical score defined as

$$\begin{aligned}\tilde{C}_N(\hat{q}) &= -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log\left(\hat{q}(X_n)\right) + (1-y_n)\log\left(1-\hat{q}(X_n)\right)\right] = \\ &= -\frac{1}{N}\sum_{n=1}^{N}\int \mathrm{d}x\, \delta(X_n - x)\left[y_n \log\left(\hat{q}(x)\right) + (1-y_n)\log\left(1-\hat{q}(x)\right)\right].\end{aligned} \tag{2.96}$$

When $N \to +\infty$ one has

$$\begin{aligned}\lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}\delta(X_n - x)y_n &= \rho_s(x)q(x), \\ \lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}\delta(X_n - x)(1-y_n) &= \rho_s(x)(1-q(x)),\end{aligned} \tag{2.97}$$

which imply that

$$\lim_{N\to+\infty}\tilde{C}_N(\hat{q}) = C(\hat{q}). \tag{2.98}$$

It has therefore been shown that the logarithmic score $C(\hat{q})$ can be used as a score function as it has an empirical version that can be computed from data and it is also minimal for $\hat{q} = q$. Furthermore, since $C(\hat{q})$ is well defined for $\hat{q}(x)! = 0$ ($\hat{q}(x)! = 1$) unless $q(x) = 0$ ($q(x) = 1$), the logarithmic score tends to discourage trivial assignments such as $\hat{q}(x) = 0$ ($\hat{q}(x) = 1$), thus behaving better than the Brier score in evaluating the predictions of rare (or frequent) events [Benedetti, 2010].

2.5 Time dependent committor functions

So far, the problem of computing the probability that a process enters a certain region of the phase space before another region has been considered. This section is devoted to discuss a slightly different problem which can nevertheless be addressed using the same formalism introduced previously.

The problem to be taken in this section is strictly related to prediction problems. Indeed, it often happens that it is necessary to compute the probability that an observable O of the system exceeds a given threshold ζ within a time T. For instance, O could be the average temperature of a country and one may be interested in computing the probability to have an extreme hot day in the incoming few weeks.

Although it may appear that these prediction problems have nothing to do with committor functions, in this section it will be shown that these probabilities are committor functions of an auxiliary process. Furthermore, it will be explained that in the case in which the observable O does not explicitly depend on time, there is a relationship between the probability of exceeding the threshold and the first hitting time of a region of the phase space. The section concludes by discussing this prediction problem in the context of a two-dimensional Wiener process.

2.5.1 Definition and equation for time dependent committor

Let $X(t)$ be a stochastic or deterministic dynamical system and let $O\left(X(t), t\right)$ be an observable of the system, i.e. a function $O : [0, +\infty) \times \Omega \to \mathbb{R}$. Then, suppose that one is interested in computing the probability $q(x, t_0; \zeta, T)$ that $O\left(X(t), t\right)$ exceeds a threshold ζ within the time interval $[t_0, t_0+T]$, knowing that the process starts at $X(0) = x$, i.e.:

$$q(x, t_0; \zeta, T) = \mathbb{P}\left(\max_{t_0 \leq t \leq t_0+T} \left[O\left(X(t), t\right)\right] > \zeta | X(0) = x\right). \tag{2.99}$$

As previously mentioned $q(x, t_0; \zeta, T)$ is nothing but a committor function for an auxiliary process Y [Lestang et al., 2018]. In fact, consider the process $Y(t) = \left[O\left(X(t), t\right), t\right]$ and let $\mathcal{A}$ and $\mathcal{B}$ be two subsets of the phase space of Y. The definitions of the two sets are

$$\mathcal{A} = \{y = [z, t_0 + T] : z \leq \zeta\}, \tag{2.100}$$

$$\mathcal{B} = \{y = [z, t] : z > \zeta \text{ and } t \in [t_0, t_0 + T]\}. \tag{2.101}$$

Fig. 2.5 shows a schematic illustration of the phase space of $Y(t)$ and it could be useful to visualize the two sets $\mathcal{A}$ and $\mathcal{B}$.

For the process $Y(t)$ one can define the committor function as $\tilde{q}(y) = \mathbb{P}\left(\tau_{\mathcal{B}}(y) < \tau_{\mathcal{A}}(y)\right)$. Furthermore, by noting that $Y(0)$ cannot be arbitrary

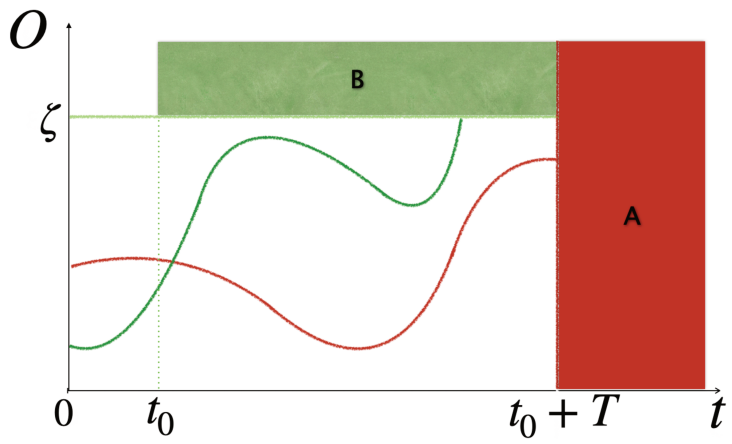


Figure 2.5: Schematic illustration of the phase space of the process $Y(t)$ together with the two sets $\mathcal{A}$ and $\mathcal{B}$. The green line represents a trajectory for which $\tau_{\mathcal{B}} < \tau_{\mathcal{A}}$ while the red line corresponds to a trajectory that reaches the set $\mathcal{A}$ before than $\mathcal{B}$.

but is related to $X(0) = x$ through the relation $Y(0) = [O(x,0),0]$, it can be argued that conditioning a probability with respect to $Y(0) = y$ is equivalent to conditioning it with respect to the value $X(0) = x$[7]. Therefore one has that $\tilde{q}(y) = \tilde{q}(x,0) = \mathbb{P}(\tau_{\mathcal{B}}(x) < \tau_{\mathcal{A}}(x))$. In addition, it should be noted that

$$\mathbb{P}\left(\max_{t_0 \le t \le t_0+T} [O(X(t),t)] > \zeta | X(0) = x\right) = \mathbb{P}(\tau_{\mathcal{B}}(x) < \tau_{\mathcal{A}}(x)), \qquad (2.102)$$

which is equivalent to say that $q(x, t_0; \zeta, T) = \tilde{q}(y)$.

Therefore, it has been shown that the quantity $q(x, t_0; \zeta, T)$ is a committor function and can be computed by using the methods introduced in Sec. 2.4.1,2.4.2.

So far, no assumptions have been made about the observable O. Among the various observables, two classes are of particular interest:

[7]This statement is true as long as the inverse function of O exists at time $t = 0$.

1. instantaneous observables $O\left(X(t)\right)$ which are functions that do not depend explicitly on time,

2. time averaged observables $O\left(X(t),t\right)=\frac{1}{\tilde{T}}\int_{t}^{t+\tilde{T}}\mathrm{d}t\,h\left(X(t)\right)$ where $h\left(X(t)\right)$ is time independent.

Note that the observables which belong to these classes are invariant under the transformation $t\to t+\tau$ provided that the process $X(t)$ is a stationary process. This invariance is important as it allows the introduction of an integral equation for $q(x,t_0;\zeta,T)$. Indeed, consider a stationary Markov process $X(t)$ and an observable which belongs to one of the two classes. Let $G(y,t|x,0)$ denote the propagator of the process. Then, it can be noted that the probability that the observable O exceeds the threshold ζ, knowing the initial condition $X(0)=x$, is equal to the sum over all possibles intermediate states $X(t_0)=y$ of the probabilities that O exceeds the threshold ζ when the process starts at $X(t_0)=y$ times the propagator $G(y,t_0|x,0)$, i.e.

$$q(x,t_0;\zeta,T)=\int\mathrm{d}y\,\mathbb{P}\left(\max_{t_0\leq t\leq t_0+T}\left[O\left(X(t),t\right)\right]>\zeta|X(t_0)=y\right)G(y,t_0|x,0). \tag{2.103}$$

By using the invariance under the transformation $t\to t+\tau$, one can write

$$\mathbb{P}\left(\max_{t_0\leq t\leq t_0+T}\left[O\left(X(t),t\right)\right]>\zeta|X(t_0)=y\right)=\mathbb{P}\left(\max_{0\leq t\leq T}\left[O\left(X(t),t\right)\right]>\zeta|X(0)=y\right)\implies$$
$$\mathbb{P}\left(\max_{t_0\leq t\leq t_0+T}\left[O\left(X(t),t\right)\right]>\zeta|X(t_0)=y\right)=q(y,0;\zeta,T). \tag{2.104}$$

By combining Eqs. (2.103),(2.104) one obtains

$$q(x,t_0;\zeta,T)=\int\mathrm{d}y\,q(y,0;\zeta,T)G(y,t_0|x,0), \tag{2.105}$$

which shows that exists a relationship between the time-dependent committor function at two different times.

In the next subsection, it will be shown that for instantaneous observables the time dependent committor function $q(x,0;\zeta,T)$ is the solution of a partial differential equation analogous to Eq. (2.53) and moreover that it is related with the cumulative distribution of first hitting times of the process $X(t)$.

2.5.2 Relation with cumulative distribution of first hitting time

Let $X(t)$ be an homogeneous continuous Markov process and $O\left(X(t)\right)$ be an observable which depends on time only through the stochastic process $X(t)$. Since $O\left(X(t)\right)$ is time independent, one can define a subset $\mathcal{C} \subset \Omega$ which contains all the points x such that $O(x) > \zeta$, i.e. $\mathcal{C} = \{x : O(x) > \zeta\}$.

Hence, the time-dependent committor function $q(x, 0; \zeta, T)$ is the probability that the first hitting time $\tau_{\mathcal{C}}(x)$ of set $\mathcal{C}$ is smaller than T:

$$q(x, 0; \zeta, T) = \mathbb{P}(\tau_{\mathcal{C}}(x) < T). \tag{2.106}$$

Following the same reasoning that in Sec. 2.3.2 led to the Dirichlet problem for the committor function, it is possible to prove that $q(x, 0; \zeta, T)$ satisfies the following partial differential equation

$$\frac{\partial q(x, 0; \zeta, T)}{\partial T} = -\mathcal{L}[q(x, 0; \zeta, T)], \tag{2.107}$$

with the boundary condition $q(x, 0; \zeta, T) = 1$ if $x \in \mathcal{C}$.

Thus, two other methods have been provided for computing $q(x, 0; \zeta, T)$: the first one consists in solving Eq. (2.107) analytically or numerically, while the second one aims to compute $q(x, 0; \zeta, T)$ by integrating up to time T the first hitting time distribution of the set $\mathcal{C}$. Once $q(x, 0; \zeta, T)$ have been computed, it is possible to obtain $q(x, t_0; \zeta, T)$ by means of Eq. (2.105).

2.5.3 Example: Wiener process

The aim of this section is to apply the formalism of the time-dependent committor function for computing the probability that a Brownian particle leaves a planar domain through a line within a certain time T. Consider a Brownian particle which moves on a planar surface whose position is denoted by $X(t) = (W_x(t), W_y(t))$. The equation of motion are

$$\begin{aligned} \dot{W}_x(t) &= \sqrt{2D}\xi_x(t), \\ \dot{W}_y(t) &= \sqrt{2D}\xi_y(t). \end{aligned} \tag{2.108}$$

Consider the line $x = a$ with $a = 1$, and imagine having to compute the probability that the particle leaves the phase space passing through the line

within a time $T = 10$. Such a problem can be solved with the formalism previously introduced by considering the observable $O(X(t), t) = W_x(t)$ and the threshold $\zeta = a$.

First, note that although the system lies in a two-dimensional space, the two components are independent of each other. Therefore, the time-dependent committor function will only depend on the $x-$component of the initial condition. Moreover, due to the independence of the two components, also the propagator can G be written as the product of a propagator for the $y-$component and a propagator for the $x-$component and the absorbing boundary conditions will be imposed only for the $x-$component propagator.

Let $G_x(W_x(t)|W_x(0))$ and $G_y(W_y(t)|W_y(0))$ be the propagators of the two components. Then, the total propagator G which satisfies the right boundary conditions is [Redner, 2001]

$$\begin{aligned}
&G(X(t), t|X(0), 0) = G_x(W_x(t)|W_x(0))G_y(W_y(t)|W_y(0)),\\
&G_y(W_y(t)|W_y(0)) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(\frac{(W_y(t) - W_y(0))^2}{4Dt}\right),\\
&G_x(W_x(t)|W_x(0)) = \frac{1}{\sqrt{4\pi Dt}} \left[\exp\left(\frac{(W_x(t) - W_x(0))^2}{4Dt}\right) - \exp\left(\frac{(W_x(t) - (2a - W_x(0)))^2}{4Dt}\right)\right].
\end{aligned} \tag{2.109}$$

Then, the probability that the first hitting time of the set $\mathcal{C} = \{(x, y) : x > a\}$ is smaller than T is equal to

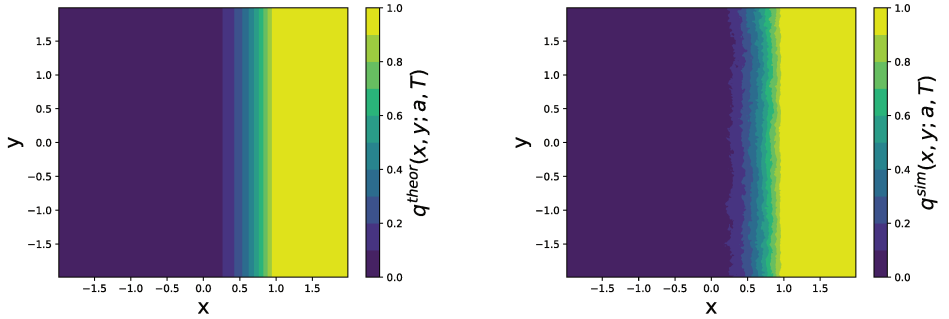
$$\mathbb{P}(\tau_{\mathcal{C}}(x) < T) = 1 - \int_{\mathcal{C}} G(X(T), T|X(0) = (x, y), 0) = 1 - \text{erf}\left(\frac{a - x}{\sqrt{4DT}}\right), \tag{2.110}$$

where $\text{erf}(\cdot)$ denotes the error function.

Since O is an instantaneous observable, the time-dependent committor function is equal to the probability that the first hitting time of the set $\mathcal{C}$ is smaller than T. Thus, one has that $q(x, 0; a, T) = \mathbb{P}(\tau_{\mathcal{C}}(x) < T) = 1 - \text{erf}\left(\frac{a-x}{\sqrt{4DT}}\right)$. Clearly, $q(x, 0; a, T) \to 0$ if $T \to 0$ because if the particle does not have enough time it will not leave the domain. In the opposite limit, i.e. when $T \to +\infty$, $q(x, 0; a, T) \to 1$ given that sooner or later the particle will reach the threshold a. The same considerations hold for the diffusion coefficient D. Indeed, when $D \gg 1$ the particle is much more likely to leave the domain within time T than in the case $D \sim 0$. Considering the dependence of $q(x, 0; a, T)$ on x, one has that if x is close to a then $q(x, 0; a, T)$

is close to 1, while for $x \ll a$ the time-dependent committor is equal to 0. Note that the passage from $q \sim 0$ to $q \sim 1$ takes place in a region that is smaller the smaller the term $4DT$ is.

Fig. 2.6 shows the time-dependent committor function $q(x, y, 0; a, T)$ for the two dimensional Wiener process. The diffusion coefficient has been taken equal to $D = 10^{-2}$. In the left panel it is shown the theoretical committor while the right panel shows the same quantity computed from an ensemble of long trajectories, as described in Sec. 2.4.2. It can be seen from Fig. 2.6a



(a) *Theoretical committor.*

(b) *Committor computed from an ensemble of long trajectories.*

Figure 2.6: Time-dependent committor function $q(x, y, 0; a, T)$ for the two dimensional Wiener process. Left panel: theoretical result for the committor (Eq. (2.110)). Right panel: committor function $q(x, y, 0; a, T)$ computed from numerical simulations.

that as expected the time-dependent committor function $q(x, y, 0; a, T)$ does not depend on y. Furthermore, $q(x, y, 0; a, T)$ is equal to 0 almost everywhere except for a region of size$\sim O(1)$ close to the threshold a. In addition, the growth of $q(x, y, 0; a, T)$ is faster the closer x is to a. By comparing Fig. 2.6a and Fig. 2.6b, it can be seen that the committor function estimated through numerical simulations reproduces the theoretical results quite accurately. This confirms that the numerical methods introduced in Sec. 2.4 can be used in this type of prediction problems, showing once again the generality of the concept of committor function.

Chapter 3

Committor Functions for Climate Phenomena at the Predictability Margin: The example of ENSO in the Jin and Timmermann model

In this chapter, the methods introduced in Chapter 2 are applied for studying the committor function of the Jin and Timmerman model of El-Niño. In this context, it is shown that the ability to predict the probability of occurrence of the event of interest may differ strongly depending on the initial state. The main result is the new distinction between intrinsic probabilistic predictability (when the committor function is smooth and probability can be computed which does not depend sensitively on the initial condition) and intrinsic probabilistic unpredictability (when the committor function depends sensitively on the initial condition).

The content of this chapter is intended to be a paper manuscript and has been submitted to the "Journal of the Atmospheric Sciences" (JAS) [Lucente et al., 2021]. For this reason, the chapter is self-consistent and can be read independently from the rest of the thesis. Therefore, some of the previously described ideas, especially those discussed in the introductory chapter, are discussed again. The reader who is reading the thesis manuscript entirely can therefore skip the introduction of this chapter and start reading from Sec. 3.2.

3.1 Introduction

It has long become clear that statistics and probability are the natural languages for climate: for given boundary conditions, there is a typical state (or several, in case of bimodality), the *climatology*, and fluctuations around typical conditions, referred to as *climate variability*, involving various time and space scales. At first sight, this kind of description may seem orthogonal to the problem of weather forecasting, which consists in predicting the exact state of the atmosphere at a given future time. However, notwithstanding the use of probabilities in numerical weather forecasting for uncertainty quantification, these two approaches meet in several areas of current climate research [Kalnay, 2003, Dijkstra, 2013, Ragone et al., 2018]. For instance, we are often interested in predicting the occurrence of specific fluctuations of the climate system, be it a given mode of climate variability, such as the *El Niño Southern Oscillation* (ENSO) [Philander, 1990], regime changes [Tantet et al., 2015], or rare events such as heat waves [Ragone et al., 2018], sudden stratospheric warming, cold spells, extreme precipitations, or any other event of importance. All these events have a probability of occurring any given year, i.e. with respect to climatological conditions, but one may also be interested in their probability of occurrence conditioned on the state of the climate system at the time of the prediction. For instance, given the global impact of events like ENSO, much efforts have focused on developing methods to forecast it several months in advance [Latif et al., 1994, Clarke, 2008, Chekroun et al., 2011, Ludescher et al., 2014, Feng and Dijkstra, 2017, Nooteboom et al., 2018]. Similarly, one may want to estimate the probability of occurrence of a summer drought based on soil moisture in the spring, the probability of occurrence of a heat wave a few weeks in advance, based on the observed atmospheric circulation, or the probability of an extreme hurricane season, based on sea surface temperature. Such forecasts are extremely challenging, but would be rewarded with proportionally large benefits, given the socio-economic impact of these events at the local and global scales, especially in a climate change context [AghaKouchak et al., 2012, Coumou and Rahmstorf, 2012, Field et al., 2012, Herring et al., 2014]. While it is not clear that this may be reliably achieved for all the above examples, due to their different physical nature, conceptually all these events fall in the same class of prediction problems. The goal of this paper is to discuss the mathematical structure of such climate prediction problems.

Indeed, the mathematical structure of weather forecasting is quite clear:

it consists in solving an initial value problem (IVP). Given an initial condition x_0 belonging to an appropriate phase space, we are interested in computing the trajectory $x_t = \phi^t x_0$, where ϕ^t is the flow of the dynamical system, encapsulating the evolution equations. For many dynamical systems, this description only holds for a finite time in practice, due to sensitive dependence on initial conditions. This limitation was already known from mathematicians in the 19th century, such as Poincaré and Hadamard. For low-dimensional chaotic dynamical systems, this time scale, up to which deterministic forecasts are relevant, corresponds the the *Lyapunov time* [Castiglione et al., 2008]. In the atmosphere, the predictability horizon, about two weeks in practice, corresponds to the time it takes for undetectable errors at the smallest scales of the flow to contaminate the large scales [Thompson, 1957, Novikov, 1959, Lorenz, 1969b]. Early numerical weather prediction attempts fell short of this predictability horizon, both due to model errors and sparsely constrained initial conditions. As models improved and observational data became much denser, owing in particular to the advent of satellite observations, performance rose and skillfull forecasts are now close to the theoretical barrier [Bauer et al., 2015]. Beyond this limit, the dynamics becomes effectively stochastic. Notwithstanding the fact that the relevant phase space may be different for climate dynamics over geological time scales, climate therefore corresponds to the statistical properties of some stochastic process $(x_t)_{t>0}$. Over very long times, we expect those statistical properties to be independent of the initial condition. In other words, the mathematical concept relevant for climate is the *invariant measure* of the system. For lack of better techniques, in practice we still compute these properties by averaging over long times and over realizations using ensembles of trajectories obtained by numerical integration of climate models. In any case, the invariant measure only describes the system for times larger than the *mixing time*, after which the initial condition is forgotten. However, in the applications cited above, the time scale of interest is the intermediate case for which a deterministic forecast is not relevant, but for which some information, more precise than the climate average, might be predicted. We call this range of time scales the *predictability margin*.

Prediction problems at the predictability margin are of a probabilistic nature, because they are concerned with time scales beyond the deterministic predictability horizon of the system (e.g. the Lyapunov time). However, we stress that the Lyapunov time scale, a global quantity, is clearly not the relevant dynamical quantity for this predictability problem. By contrast, at

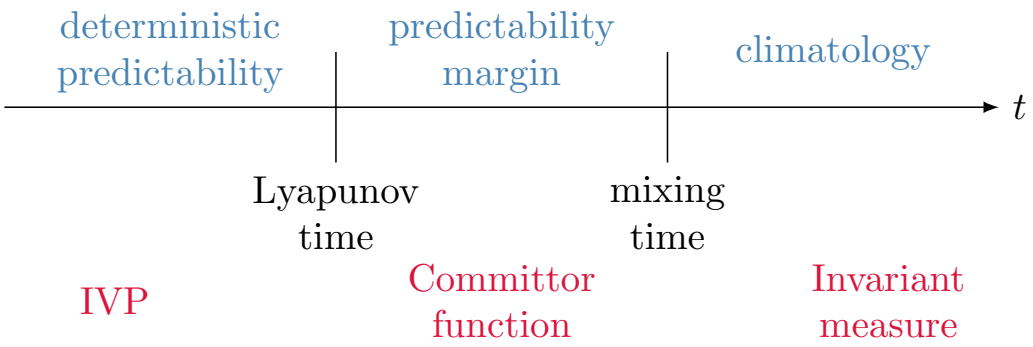


Figure 3.1: Schematic illustrating the concept of *predictability margin*: deterministic predictability is only possible until a finite time (e.g. the Lyapunov time). The associated mathematical problem is an initial value problem (IVP). Long term statistical properties (beyond the mixing time) do not depend on the initial condition, and the corresponding mathematical object is the *invariant measure*. In the intermediate range of timescales, which we call the *predictability margin* here, the appropriate mathematical concept is the *committor function*, which encodes the probability of a given event to occur, condition on the state of the climate system at the time of the prediction.

the predictability margin, the predictability clearly depends on the current state of the system. Then, the question is: what is the relevant mathematical concept for prediction problems at the predictability margin? The relevant mathematical concept is called the *committor function* [E et al., 2005, Vanden-Eijnden, 2006]. This is a very generic concept: a committor function is the probability for an event to occur in the future, as a function of the current state of the system. Committor functions have first been introduced in climate sciences in [Lucente et al., 2019], and has been used to study sudden stratospheric warming [Finkel et al., 2020, Finkel et al., 2021] or to understand the flow of ocean debris [Miron et al., 2021]. The interest of putting a name, the committor function, to this otherwise very common and generic concept, is two-fold. First it allows to study its mathematical properties and to related them to key concepts in dynamical systems, for instance the predictability margin, as we do in the present work. Second, it comes with specific theoretical and computational approaches to compute this important quantity, for instance transition path theory, see for example [Vanden-Eijnden, 2006, Metzner et al., 2006, Metzner et al., 2009] and

references therein. In atmosphere dynamics, a very interesting use of smart ways of estimating the committor function for a simplified model of sudden stratospheric warming is provided by [Finkel et al., 2021].

Many problems in *medium-range forecasting* fall within the *predictability margin* range; to illustrate the interest of committor functions, we will select only one exemple of application, the problem of ENSO prediction, using a very simple model. While, as mentioned above, many studies strive to predict the occurrence of El Niño a few months in advance, we shall address here a slightly different problem, focusing on predicting strong El Niño events on longer time scales. This is also a relevant question from the point of view of climate dynamics; while strong El Niño events have been observed almost periodically since the 1950s, with a return time around 15–20 years, historical data and paleoclimatic proxies indicate that ENSO may exhibit high variability over centennial timescales [Cobb et al., 2003, Khider et al., 2011, McGregor et al., 2013] and beyond [Rickaby and Halloran, 2005, Fedorov et al., 2006, Cobb et al., 2013]. We study the dynamics of a low-dimensional stochastic model proposed to explain the decadal amplitude changes of ENSO, the Jin and Timmermann model [Timmermann and Jin, 2002, Timmermann et al., 2003]. This model is not aimed at reproducing any precise properties of the real El Niño Southern Oscillation. It is rather used as a paradigm example to introduce the concept of a committor function, and to study its main properties. This will lead us to define probabilistic predictability and unpredictability, some concepts that should be useful for other applications.

We show that probabilistic prediction at the predictability margin depends on the initial state, and that probabilistic predictability is encapsulated in the committor function. This property is analogous to classical, deterministic predictability, which is known to depend on the state of the system: some circulation patterns, such as the positive phase of the *North Atlantic Oscillation* (NAO), lead to improved predictability. However, we stress that deterministic and probabilistic predictability are different concepts: probabilistic predictability means that the probability of the event does not depend sensitively on the initial conditions. This corresponds to a region of phase space where the committor function has gentle variations with the initial conditions. In these areas, the event occurs with a probability p that can be easily determined in practice because of these gentle variations. On the contrary, probabilistic unpredictability corresponds to regions of the phase space with a rough committor function. In these regions, the occurence of the

event is also probabilistic. But the probability itself has very rapid variations with the initial conditions, which make the prediction highly dependent on the level of precision with which the initial condition is known. The existence of such features, and especially the new and most interesting *probabilistically predictable* region, should be generic for most prediction problems in climate dynamics.

This paper also discusses relations between qualitative properties of the committor function, finite time Lyapunov exponents, and the stability properties of trajectories with respect to noise perturbations. It also discuss methodological aspects for a data-based approach for the computation of committor functions.

The dynamics of the Jin and Timmerman model, when perturbed by a weak noise, is characterized by rare transitions between a limit cycle and a strange attractor [Roberts et al., 2016, Guckenheimer et al., 2017]. Based on large deviation theory, and with generic hypothesis, the average transition time $\mathbb{E}[\tau_c]$ to see such transitions is expected to change following an Arrhenius law: $\mathbb{E}[\tau_c] \underset{\sigma \to 0}{\asymp} A \exp\left(\Delta V / \sigma^2\right)$, where σ is the noise amplitude. Using numerical simulations, we demonstrate that the Jin and Timmerman transition times do not follow the expected Arrhenius law for a very large range of small noise amplitudes. We conjecture that this very interesting phenomenon might be the first observed counterexample to the expected generic result, for deterministic dynamics perturbed by weak noises. We argue that this is related to the intricated entanglement between the basins of attraction of the limit cycle and the strange attractor.

The paper is organized as follows: in Sec. 3.2 we define the Jin and Timmermann model [Timmermann and Jin, 2002, Timmermann et al., 2003]. In this model, the occurrences of strong ENSO events correspond to noise-induced transitions between a strange attractor and a limit cycle [Roberts et al., 2016, Guckenheimer et al., 2017]. We study in Sec. 3.3 the statistics of such transitions, and we show that they do not obey an Arrhenius law. Finally, in Sec. 3.4 we introduce the committor function, we compute it for the Jin-Timmerman model, and characterize the regions of the phase space with qualitatively different predictability properties. In the regime of intermediate noise amplitude, at the predictability margin, we delineate four regions (see Fig. 3.7): two regions of *deterministic predictability* (where the event occurs with probability 0 or 1), one *probabilistically predictable* region (where a value of the probability $0 < q < 1$ can clearly be predicted with very

mild dependence with respect to initial condition), and finally a region which is unpredictable in practice, because the strong dependence with respect to the initial condition prevents any precise prediction, either deterministic or probabilistic.

3.2 The Jin and Timmermann model

El Niño Southern Oscillation (ENSO) is one of the most important mode of climate variability at the interannual time scales [Philander, 1990]. El Niño events consist in an increase of the Sea Surface Temperature in the eastern equatorial Pacific Ocean, leading at the local scale to reduced thermocline depth, reduced upwelling and reduced nutrient supply, thereby affecting marine life. Such events are correlated with a reorganization of the Walker circulation in the atmosphere, known as the Southern Oscillation. The global phenomenon, referred to as ENSO, has major impacts all over the world. However, the nonlinear coupled atmosphere-ocean dynamics of ENSO makes it very difficult to predict [McPhaden et al., 2015]. Models of various complexities have been constructed to capture the dynamics of El Niño at different levels of realism [Clarke, 2008, Sarachik and Cane, 2010]. In order to introduce and illustrate the concept of committor function in the simplest possible framework, we shall consider here one of the most idealized models, consisting of a low-dimensional stochastic process. This simple dynamical model, introduced by [Jin, 1997a, Jin, 1997b], accounts for the recharge-discharge mechanism which is at the basis of ENSO. This model was later extended by [Timmermann et al., 2003] and was related to the decadal amplitude changes of ENSO [Timmermann and Jin, 2002]. The model describes the evolution of three variables:

1. T_1, the Sea Surface Temperature in the western equatorial Pacific Ocean,
2. T_2, the Sea Surface Temperature in the eastern equatorial Pacific Ocean,
3. h_1, the thermocline depth anomaly in the western Pacific.

Assuming a thermal relaxation towards a radiative-convective temperature T_r, the equations of motion can be written as [Timmermann and Jin, 2002,

Timmermann et al., 2003]:

$$\begin{aligned}
\frac{\partial T_1}{\partial t} &= -\alpha(T_1 - T_r) - \epsilon\beta\tau(1 - \sigma\eta_t)(T_2 - T_1), \\
\frac{\partial T_2}{\partial t} &= -\alpha(T_2 - T_r) + \zeta\beta\tau(1 - \sigma\eta_t)(T_2 - T_{sub}), \\
\frac{\partial h_1}{\partial t} &= r\left(-h_1 - \frac{1}{2}bL\tau\right),
\end{aligned} \tag{3.1}$$

where ϵ and ζ represent the strength of the zonal and vertical advection, T_{sub} denotes the temperature being upwelled into the mixed layer, τ represents the zonal wind stress, L denotes the basin width, b captures the efficiency of wind stress in driving thermocline tilt, $1/\alpha$ measures a typical thermal damping timescale and $1/r$ is the dynamical adjustment timescale of the thermocline depth. The term η_t in the equations for temperatures is a Gaussian white noise with unit variance and the level of stochasticity is controlled by the noise amplitude σ. This term takes into account the fluctuating component of wind stress. In the last equation the noise has not be considered because wave processes are filtered out in the thermocline equations of the model [Timmermann et al., 2003].

The expressions of T_{sub} and τ are

$$\begin{aligned}
T_{sub} &= \frac{T_r + T_{r_0}}{2} + \frac{T_r - T_{r_0}}{2}\tanh\left(\frac{H + h_2 - z_0}{h^*}\right), \\
\tau &= \frac{\mu(T_2 - T_1)}{\beta},
\end{aligned} \tag{3.2}$$

where T_{r_0} is a reference temperature, h_2 is the thermocline departure from its reference value H, z_0 represents the depth at which ζ takes its characteristic value, h^* measures the sharpness of the thermocline. The relation between the eastern and western thermocline depth anomalies is

$$h_2 = h_1 + bL\tau. \tag{3.3}$$

In order to study the dynamical behavior of the system it is useful to perform a change of variables from physical to dimensionless ones [Roberts et al., 2016]. So, we define

$$\begin{aligned}
x &= \frac{T_2 - T_1}{T_0} \quad , \quad y = \frac{T_1 - T_r}{T_0}, \\
z &= \frac{h_1 + H - z_0}{h^*} \quad , \quad \tilde{t} = \frac{t}{t^*},
\end{aligned}$$

where $T_0 = \frac{h^*\beta}{bL\mu}$ and $t^* = \frac{bL}{\beta\zeta h^*}$. After the change of variables, the equations (3.1) read

$$\begin{aligned}
\dot{x} &= \rho\delta(x^2 - ax) + x[x + y + c - c\tanh(x + z)] - D_x(x, y, z)\eta_t, \\
\dot{y} &= -\rho\delta(x^2 + ay) + D_y(x, y, z)\eta_t, \\
\dot{z} &= \delta\left(k - z - \frac{x}{2}\right),
\end{aligned} \tag{3.4}$$

where

$$\begin{aligned}
D_x(x, y, z) &= [(1 + \rho\delta)x^2 + xy + cx(1 - \tanh(x + z))]\sigma, \\
D_y(x, y, z) &= \rho\delta x^2\sigma,
\end{aligned}$$

and the new control parameters δ, ρ, c, k, and a are defined as follows:

$$\delta = \frac{rbL}{\zeta\beta h^*}, \qquad \rho = \frac{\epsilon h^*\beta}{rbL},$$
$$a = \frac{\alpha bL}{\epsilon\beta h^*}, \qquad c = \frac{T_r - T_{r_0}}{2T_0}, \qquad k = \frac{H - z_0}{h^*}.$$

The deterministic version ($\sigma = 0$) of equations (3.4) was widely studied in literature. For some parameter values, the system has only one attractor, a periodic orbit. Figure 3.2-b illustrates such a periodic orbit, with the parameter values $[\delta, \rho, c, k, a] = [0.2625, 0.3224, 2.3952, 0.4032, 6.8927]$ and dimensional normalization constants $[T_0, t^*, h^*] = [2.8182\,°\text{C}, 104.9819\,\text{days}, 62\,\text{m}]$. [Roberts et al., 2016] also analyzed the mechanism through which this limit cycle arises. [Roberts et al., 2016] defined strong El-Niño events for this model as periods in this limit cycle for which the temperature is large. Figure 3.2 shows a qualitative comparison of the eastern Pacific sea surface temperature anomaly for this limit cycle with the El-Niño3 index. Both the measurements and the model display positive temperature anomaly excursions with a return time of approximately 15 years.

Varying the parameter δ, a strange attractor emerges through a period doubling cascade, as shown by [Guckenheimer et al., 2017]. Moreover, [Guckenheimer et al., 2017] show that for some parameter values the limit cycle and the strange attractor coexist. Following [Guckenheimer et al., 2017], we use the parameter values $[\delta, \rho, c, k, a] = [0.225423, 0.3224, 2.3952, 0.4032, 7.3939]$ all along this paper. While [Guckenheimer et al., 2017] considered only the deterministic model ($\sigma = 0$), we also consider later on the stochastic model

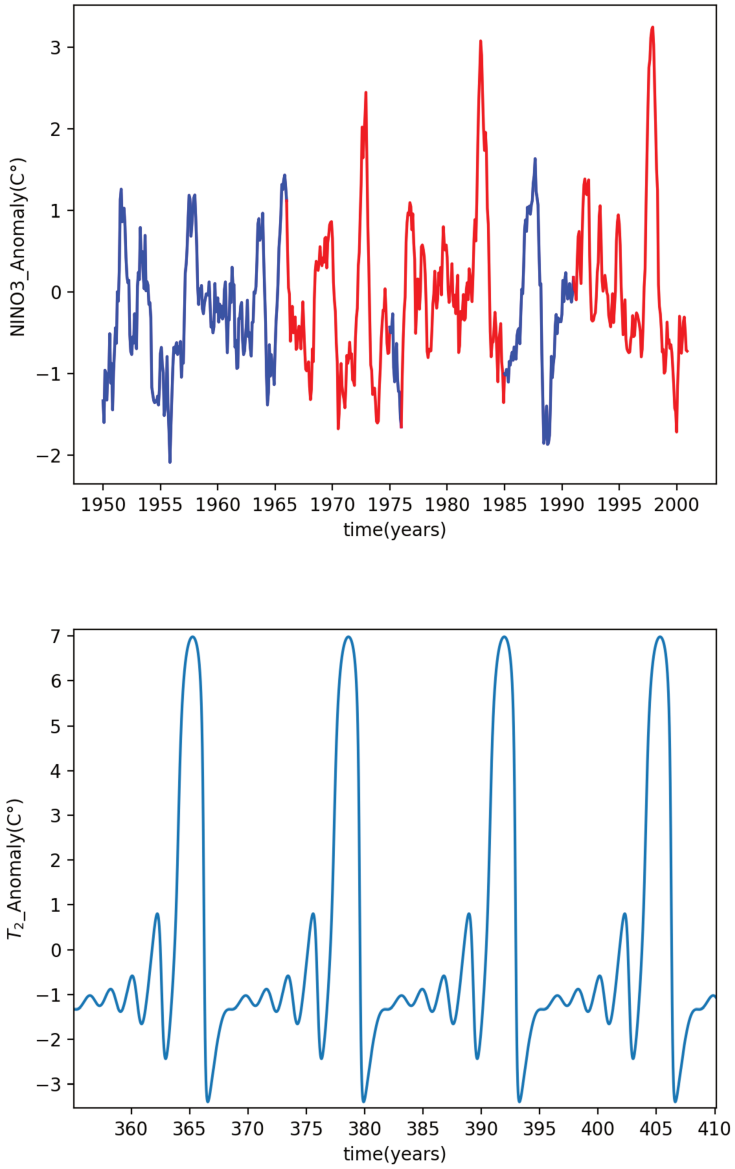


Figure 3.2: Top plot: observed sea surface temperature anomalies over the last decades, spatially averaged over the Niño-3 region. Bottom: eastern Pacific sea surface temperature anomalies simulated with the deterministic Jin and Timmermann model.

($\sigma \neq 0$). For $\sigma = 0$, the strange attractor and the limit cycles are shown in Fig. 3.3. They are intertwined in a complex way.

For this dynamics, we define a strong El-Niño event as any situation when x becomes larger than the threshold $x_c = -1$. As can be seen from Fig. 3.3, this only happens in the limit cycle, and not in the strange attractor. Note

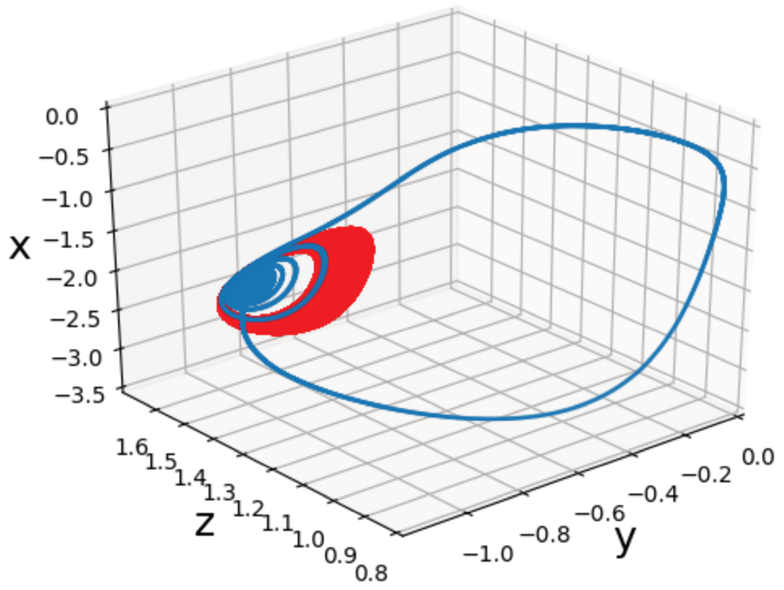


Figure 3.3: The two intertwined attractors of the Jin-Timmermann model (limit cycle in blue and strange attractor in red).

that with this choice of parameters, the return period of strong El Niño events on the limit cycle is around 50 years (the time unit is $t^* = 105$ days and the period is 186 non-dimensional time units), rather than 15 for the parameters studied by [Roberts et al., 2016].

For $\sigma = 0$, when the parameter a is time periodic rather than constant, mimicking a seasonal forcing, [Guckenheimer et al., 2017] observed transitions between the strange attractor and the limit cycle. In the following we consider a constant a, and rather study noise induced transitions between

these two attractors, for $\sigma \neq 0$.

We now discuss qualitatively the effect of the noise level σ. For small $\sigma > 0$, the dynamics can switch from one regime close to the strange attractor to another regime close to the limit cycle. This is evident by looking at

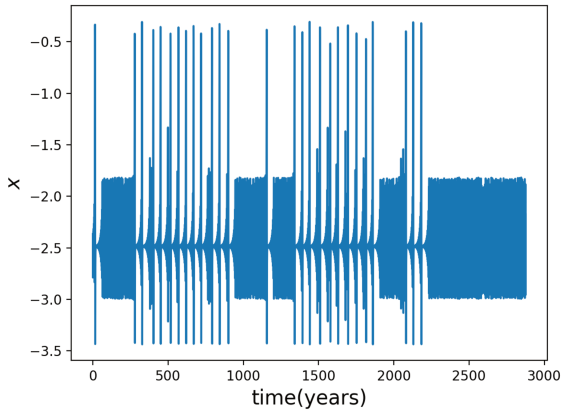


Figure 3.4: Time series of the variable x for $\sigma = 0.00005$. The dynamics exhibits a switching between a regime close to the strange attractor (bounded regions around $x = -2.5$) and a limit cycle regime.

Fig. 3.4, where the time evolution of the variable x for $\sigma = 5 \cdot 10^{-5}$ is shown. As can be seen, the system switches between epochs where the variable x is bounded around the value $x = -2.5$ (strange attractor regime) and epochs where there are large oscillations with a period around 50 years (limit cycle regime). Strictly speaking, for $\sigma \neq 0$, there is not anymore two attractors. However, as illustrated in Fig. 3.4, for small enough σ, the trace of the two deterministic attractors is clearly visible. For small σ, we will thus continue to discuss the strange attractor and the limit cycle, for simplicity. Strong El-Niño events occur only during the limit cycle regime. In the next section, we will study transitions from the strange attractor regime to the limit cycle regime (or equivalently to the strong El Niño regime).

The frequency of transitions from the strange attractor to the strong El-Niño regime increases as the amplitude of the noise increases. For large values of σ, the dynamics is completely dominated by the noise and the distinction between the two attractors becomes meaningless.

3.3 Statistics of the first exit times for transitions to strong El Niño regimes

As discussed in the previous section, we define strong El Niño events as periods of time when $x_c > -1$, which occur along the limit cycle. In this section, we study transitions from the strange attractor regime to the strong El Niño regime, and how their statistics change when the noise amplitude σ is varied.

We consider $\mathbf{X}(\mathbf{t}) = (x(t), y(t), z(t))$ solutions to the stochastic Jin and Timmerman model (3.4). We define *first exit times* from a point $\mathbf{x}$ to the strong El Niño regime as

$$\tau_c(\mathbf{x}) = \inf\{t > 0 : x(t) > x_c \mid \mathbf{X}(0) = \mathbf{x}\}. \tag{3.5}$$

The random variable $\tau_c(\mathbf{x})$ depends both on the realization of the noise and on the initial condition $\mathbf{x}$. The statistics are understood as averages over both the noise realization and the invariant measure of $\mathbf{x}$ over the strange attractor of the deterministic system ($\sigma = 0$), the so called SRB measure. For instance the mean first exit time $\mathbb{E}[\tau_c]$ is defined as

$$\mathbb{E}[\tau_c] = \int \mathrm{d}\mathbf{x}\, \rho_{\mathrm{SRB}}(\mathbf{x}) \mathbb{E}_{\mathrm{noise}}[\tau_{\mathrm{c}}(\mathbf{x})]. \tag{3.6}$$

where $\mathbb{E}_{noise}[\cdot]$ is the expectation with respect to the noise realization and $\mathrm{d}\mathbf{x}\, \rho_{\mathrm{SRB}}(\mathbf{x})$ is the SRB measure.

The SRB measure is defined through time averages of the deterministic dynamics ($\sigma = 0$). In practice, we thus compute a very long trajectory of the deterministic dynamics. We then choose a set of 1000 initial conditions $\mathbf{x}$ taken randomly among all the points of this deterministic trajectory. Then, for any fixed value of $\sigma > 0$, for any initial condition $\mathbf{x}$, we compute the first-passage time τ_c for several noise realizations.

In Fig. 3.5, we show the probability density function $p(\tau_c)$ of τ_c based on this ensemble. The probability density function is close to an exponential: $p(\tau_c) = \lambda e^{-\lambda \tau_c}$. The parameter λ is then equal to the inverse of the mean first exit time: $\lambda^{-1} = \mathbb{E}[\tau_c]$.

Because typically $\tau_c(\mathbf{x})$ is much larger than the relaxation time to the strange attractor, one might expect that for most of the points of the strange attractor the dependence of $\tau_c(\mathbf{x})$ on $\mathbf{x}$ is practically irrelevant. Indeed, we have verified numerically that except for a small region around the transition

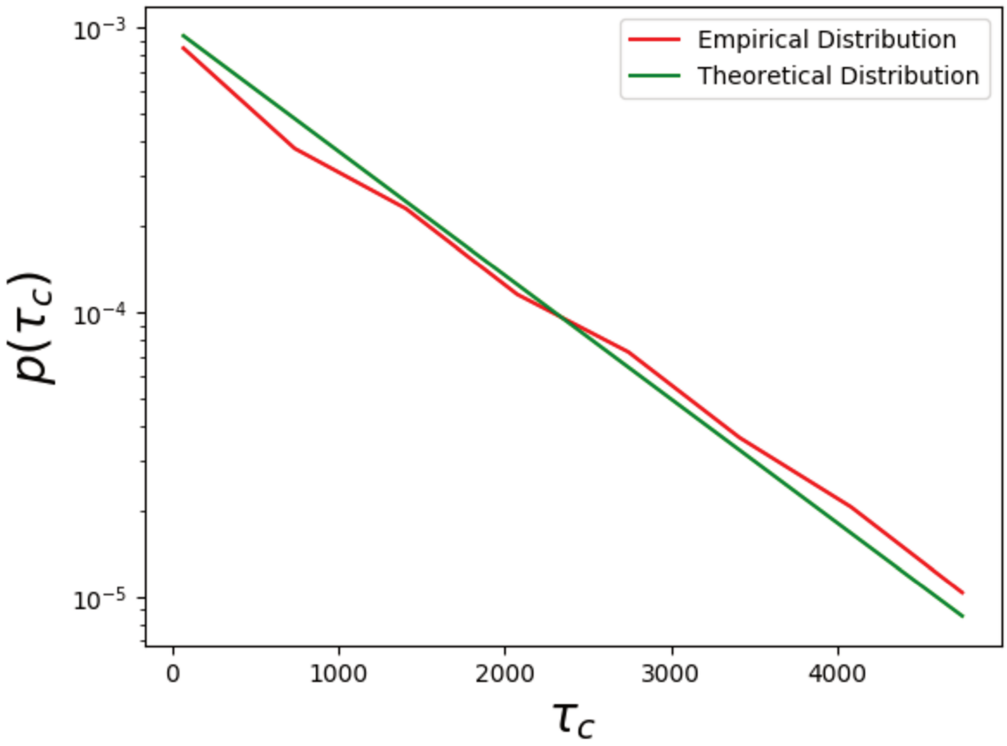


Figure 3.5: Logarithm of the Probability Density Function of the first exit time between the strange attractor and the limit cycle for $\sigma = 5 \times 10^{-5}$ sampled by direct integration (red) of the stochastic differential equation (3.4), and the exponential distribution $p(\tau_c) = \lambda e^{-\lambda \tau_c}$ with $\lambda^{-1} = \mathbb{E}[\tau_c]$ (green).

paths, the statistics are independent from the initial condition, up to numerical accuracy. Hence we have $\mathbb{E}[\tau_c] \simeq \mathbb{E}_{SRB}[\tau_c(\mathbf{x})] \simeq \mathbb{E}_{noise}[\tau_c(\mathbf{x})]$ for generic points $\mathbf{x}$ close to the strange attractors.

As illustrated in Fig. 3.5, the mean first exit time $\mathbb{E}[\tau_c]$ is of the order of 1,000 in non-dimensional time units. The measured value is $\tau_c = 1,039$. As τ_c is much larger than the mixing time of the SRB measure, of order 1, then it is natural to expect that the first exit times should be random and distributed, with a very good approximation, according to a Poisson statistics. The observed exponential distribution is consistent with such a Poison statistics. Similar exponential distributions for first exit times were observed for the deterministic dynamics with periodic modulation of the a coefficient [Guckenheimer et al., 2017].

We now study how the mean first exit time $\mathbb{E}[\tau_c]$ varies when the noise amplitude σ is changed. One generally expects an Arrhenius law:

$$\mathbb{E}[\tau_c] \underset{\sigma \to 0}{\sim} A e^{\frac{\Delta V}{\sigma^2}}. \tag{3.7}$$

Arrhenius laws where first derived by Kramers for gradient dynamics forced by white noise $\dot{\mathbf{x}} = -\mathrm{d}V/\mathrm{d}\mathbf{x} + \sqrt{2\sigma}\eta(t)$, where ΔV (the *potential barrier*) is the difference of potential between the original attractor and the saddle-point separating the basins of attraction of the two attractors (see for example the textbook by [Gardiner et al., 1985]). The Jin and Timmermann model is however not a gradient dynamics, and the function V is not explicit. For such non gradient systems, the exponential factors of the Arrhenius law can be justified through a Laplace principle for a path integral representation of the transition probabilities, or asymptotic studies of Fokker-Planck operators [Graham, 1987], or through large deviation theory [Freidlin and Wentzell, 2012]. The function V is then called the *quasipotential*, which can be computed through a variational problem, or computing viscosity solutions of a Hamilton-Jacobi equation. The sub-exponential prefactor A in Eq. (3.7) can be computed through Eyring-Kramers formulas, derived either for gradient [Bovier et al., 2004] or non-gradient dynamics [Bouchet and Reygner, 2016], for transitions from a point attractor and through a point saddle. Many generalizations exist, for instance for periodically modulated systems [Dykman and Ryvkine, 2005] or systems approaching a bifurcation [Herbert and Bouchet, 2017]. In large dimensional systems related to climate dynamics, effective Arrhenius laws have been observed numerically, for instance in transitions in beta-plane turbulence [Bouchet et al., 2019] or in a simplified climate model with ice-

albedo feedback [Lucarini and Bódai, 2019].

When one studies a transition from a strange attractor rather than a point attractor, or when the saddle set between the two basins of attraction is a not a simple saddle point, then no theorem exists to put the Arrhenius law (3.7) on firm mathematical ground. However it has been argued for a long time [Graham, 1987], that if a finite distance $d > 0$ exists between the strange attractor and the saddle set, then there is a non-zero quasipotential difference $\Delta V > 0$, and an Arrhenius law should generically be expected.

In Fig. 3.6, we show the mean first exit time $\mathbb{E}[\tau_c]$ as a function of the noise amplitude σ. It ranges from transition times of about 25 years ($T = 100$ in

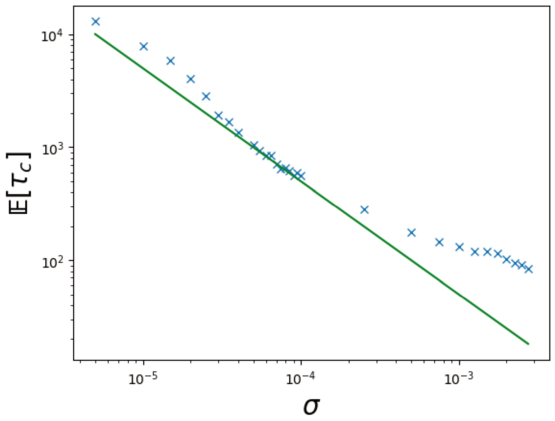


Figure 3.6: Mean first exit time $\mathbb{E}[\tau_c]$ for the transition from the strange attractor regime to the strong Niño event regime, as a function of the noise amplitude σ, in log-log coordinates. In the limit of small noise amplitude $\sigma < 10^{-4}$, $\mathbb{E}[\tau_c]$ seems to be closer to a power-law σ^{-1} (green line) than to the standard Arrhenius law.

non-dimensional time units, for the strongest values of the noise amplitude) to several millenia (about 3000 years for $T = 10000$ in non-dimensional units for weak noise). Fig. 3.6 clearly shows that the mean exit time from the strange attractor to the regime of strong Niño events does not follow an Arrhenius law of the form (3.7). The mean exit time seems much closer to a power law $\mathbb{E}[\tau_c] \propto \sigma^{-1}$.

In order to check the numerical robustness of our result, we computed $\mathbb{E}[\tau_c]$ using two different schemes of integration, and we checked numerical convergence with respect to the time step Δt in the integration schemes. The

first integration scheme is the fourth order Runge Kutta method to which a zero mean gaussian white noise is added. The variance of the noise is proportional to the integration time step Δt. In this way we have a precision of Δt^4 for the deterministic part while we make an error of order $\sqrt{\Delta t}$ for the statistics. The second integration scheme is the stochastic Runge Kutta method which has an error of order Δt for a stochastic dynamics [Roberts, 2012].

Let us note that many other behaviors than exponential ones have been observed for mean exit times. For instance a power-law has been observed for flow reversals in numerical simulations of inviscid turbulent flows [Shukla et al., 2016]. However for this last example, as the dynamics is not a deterministic system with attractors perturbed by weak noise, it was not clear why one should have expected an Arrhenius law in the first place.

We observe a breakdown of the Arrhenius law for the Jin and Timmermann model which is a deterministic system with attractors perturbed by weak noise. This is striking. Indeed we stress again that if a finite distance $d > 0$ exists between the strange attractor and the saddle set, then there is a non-zero quasipotential difference $\Delta V > 0$, and an Arrhenius law should be expected. The distance d might be expected to be generically strictly larger than 0.

We see two possible heuristic explanations for this interesting breakdown. The first explanation might be that a finite distance $d > 0$ and a quasipotential barrier $\Delta V > 0$ between the strange attractors and the basin of attraction limit cycle do actually exist, but they are extremely small. Then the explanation of the observed breakdown in Fig. 3.6 would be that we have not studied small enough values of σ. We note however that we computed first exit times of order $\mathbb{E}[\tau_c] = 5.10^5$ for values of σ as small as 10^{-6}. If this first explanation is valid, this means that the Arrhenius law is practically irrelevant even if it might be mathematically correct.

The second possible explanation might be that there exists no finite distance between the strange attractor and a possible fractal boundary between the basins of attractions. Then for any small values d and v, there always exist points in the strange attractor and in the boundary of the basin of attraction at a distance smaller than d and a quasipotential differences ΔV smaller than v. Many phenomenologies could then be imagined, for instance with a distribution of a large number of transition paths, possibly infinite, leading to a power law or an effective behavior of the first exit times described by any function. Those conjectures are not based on any mathematical re-

sults yet. However the possibility of a breakdown of the Arrhenius law is a very interesting problem, that should be studied further either through theory and mathematics, or through numerical simulations.

3.4 Committor function of the Jin and Timmerman model

In Sec. 3.3, we have shown that, in the stochastic Jin and Timmerman model, transitions between the strange attractor regime and the strong El-Niño regime occur at random times, in the limit of small noise $\sigma \to 0$. In this section, we focus on the associated prediction problem: What is the probability that a strong El Niño event occurs within a given timeframe, given the state of the system at the time of prediction? We will address this question for any finite value of the noise amplitude σ.

We consider solutions $\mathbf{X}(t) = (x(t), y(t), z(t))$ of the stochastic Jin an Timmerman model (3.4). We remind the reader that we identify a strong El Niño with an event when $x > -1$. For a solution $\mathbf{X}(t)$ that starts from $\mathbf{x}$, that is $\mathbf{X}(0) = \mathbf{x}$, we want to predict the probability $q(\mathbf{x})$ that a strong El Niño event occurs within a fixed time T. This is

$$q(\mathbf{x}) = \mathbb{P}\left(\max_{0\le t\le T} x(t) > -1 \mid \mathbf{X}(0) = \mathbf{x}\right). \tag{3.8}$$

Recalling the definition of the first passage time to a strong El Niño regime, Eq. (3.5),

$$\tau_c(\mathbf{x}) = \inf\{t > 0 : x(t) > -1 \mid \mathbf{X}(0) = \mathbf{x}\}, \tag{3.9}$$

we note that $q(\mathbf{x}) = \mathbb{P}[\tau_c(\mathbf{x}) < T]$ is the cumulative distribution function (CDF) of the first-passage time. We now define committor functions and explain that q is a committor function.

Committor functions. For a Markov stochastic process $\{\mathbf{Y}(t)\}$ which takes values in Γ, we define the *first hitting time* of the set C as $\tau_C(\mathbf{y}) = \inf\{t : \mathbf{Y}(t) \in C \mid \mathbf{Y}(0) = \mathbf{y}\}$. For two disjoint subsets $A, B \subset \Gamma$, the committor function $\tilde{q}(\mathbf{y})$ is defined as the probability to hit the set B before hitting the set A:

$$\tilde{q}(\mathbf{y}) = \mathbb{P}(\tau_B(\mathbf{y}) < \tau_A(\mathbf{y})). \tag{3.10}$$

Considering the auxiliary process $\{\mathbf{Y}(t)\}$, with $\mathbf{Y}(t) = (\mathbf{X}(t), t)$, and the two sets

$$\begin{aligned} A &= \{\mathbf{y} = (\mathbf{x}, t) \mid x > -1 \text{ and } t \in [0, T]\} \text{ and} \\ B &= \{\mathbf{y} = (\mathbf{x}, T)\,;\, x \leq -1\}, \end{aligned} \tag{3.11}$$

we see that $q(\mathbf{x}) = \tilde{q}(\mathbf{x}, 0)$. Hence q, in Eq. (3.8) is a committor function.

For an ergodic process, replacing statistical averages by temporal averages in (3.10), and using $\mathbf{y} = (\mathbf{x}, t)$, we have

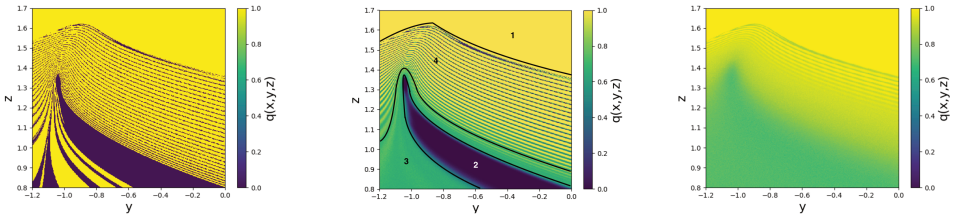
$$\begin{aligned} \rho(\mathbf{x})q(\mathbf{x}) &= \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{d}t'\, \delta\left(\mathbf{X}_{t'} - \mathbf{x}\right) 1_{\{\tau_B \leq \tau_A\}}, \\ \text{and } \rho(\mathbf{x}) &= \lim_{\mathrm{t}\to\infty} \frac{1}{\mathrm{t}} \int_0^{\mathrm{t}} \mathrm{dt}'\, \delta\left(\mathbf{X}_{\mathrm{t}'} - \mathbf{x}\right), \end{aligned} \tag{3.12}$$

where $\rho(\mathbf{x})$ is the stationary distribution function of X, δ is a Dirac delta function, and $1_{\{\tau_B \leq \tau_A\}}$ takes value 1 if $\tau_B \leq \tau_A$ and 0 otherwise. The formulas (3.12) can be used to estimate $q(\mathbf{x})$ from an observed trajectory $\{\mathbf{X}(t)\}$ of the dynamical system. For the sake of completeness, it should be said that when the dynamics is a stochastic differential equation, the committor function $q(\mathbf{x})$ is the solution of the Dirichlet problem [E et al., 2005, Thiede et al., 2019].

To illustrate the concept of predictability margin introduced in Sec. 3.1, we choose the value T=200 in non-dimensional time units, which is slightly larger than the period of the limit cycle (the "natural" periodicity of strong El-Niño events which is 186), and of the order of the Lyapunov time. This choice guarantees that for the deterministic dynamics, $\sigma = 0$, each trajectory starting in one point of the limit cycle almost certainly will reach the threshold $x_c = -1$.

3.4.1 Description of the committor function: deterministic and probabilistic predictability

Figure 3.7 shows the committor function $q(\mathbf{x})$, for different values of σ. As q is a function of 3 variables (x, y, z), we have chosen to represent cuts of q in different planes. We will discuss in detail the cut of q along the plane $x = -2.831$ (Fig. 3.7) and also cuts along the planes $y = -1.1580$ and $z = 1.3409$ (Fig. 3.8 and Fig. 3.9, respectively). To compute the committor function $q(\mathbf{x})$ on the different planes we adopted the following strategy:



(a) $\sigma = 0$ (b) $\sigma = 0.00005$ (c) $\sigma = 0.001$

Figure 3.7: Color plot of the committor function $q(x, y, z)$ in the plane $x = -2.8310$, for three values of the noise amplitude, $\sigma = 0$ (left, deterministic), 0.00005 (middle) and 0.001 (right). Regions with uniform $q = 0$ or 1 values correspond to *deterministic predictability*, smooth regions with $0 < q < 1$ to *probabilistic predictability*, and regions with sensitive dependence on intial conditions to unpredictable parts of phase space.

1. Discretize the plane into $K = L \times L$ cells $\mathcal{C}_k$.
2. For each cell $\mathcal{C}_k$, generate $N = 1000$ trajectories starting from a point $\mathbf{x} \in \mathcal{C}_k$.
3. Count the number of trajectories N_1 that reach the threshold x_c before time T.
4. Estimate $q(\mathbf{x})$ for $\mathbf{x} \in \mathcal{C}_k$ as $\hat{q}(\mathbf{x}) = \frac{N_1}{N}$.

This method is less efficient, from a computational point of view, than the one based on Eq. (3.12). In fact, in the former we only use the information carried by the initial condition while in the latter we use the information carried by a much more significant part of the trajectory. The committor function computed from long trajectories using Eq. (3.12) will be discussed in Sec. 3.4 3.4.3.

Committor Function for the deterministic dynamics ($\sigma = 0$)

In the deterministic case ($\sigma = 0$, Fig. 3.7a), as the future is completely determined by the initial condition, q can only take values 0 or 1. On Fig. 3.7a, we can distinguish three regions corresponding to two very different situations. First, two regions correspond to uniform values of the committor function:

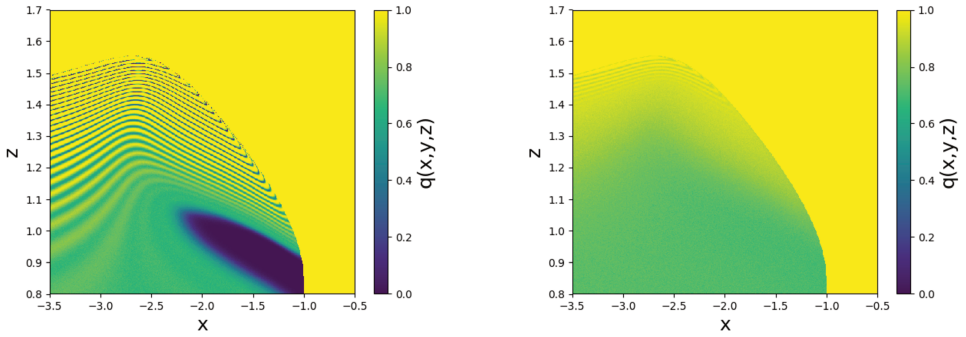
in the yellow area $q = 1$, when trajectories reach the threshold within a time T, corresponding to large values of z; in a thick purple band $q = 0$, when no trajectory reaches the threshold. In those two regions, the occurrence of strong El-Niño events is easily predicted. Everywhere else, we see very fine filaments of alternating yellow and purple values. In this third region, because of the sensitive dependence on the initial conditions, a **small but finite** initial perturbation, of the order of 1% of the values of x or y, leads to a different outcome. In this region, the occurrence of strong El-Niño events is very difficult to predict. A precise definition of this third area is not intrinsic, it depends on the actual precision with which the values of x and y can be measured. However the distinction between areas with easy predictability and areas with difficult predictability, might be crucial at a practical level.

One might ask what relationship exists between the regions outlined above and the basins of attraction of the system. However, this relationship is less trivial than one might expect. Although some regions reflect the structure of the basins of attraction, this is not true in general. In fact, there are points in the basin of attraction of the strange attractor which pass the threshold before reaching the strange attractor, as well as points in the basin of attraction of the limit cycle which do not reach the threshold within the time T.

Committor Function for $\sigma \neq 0$

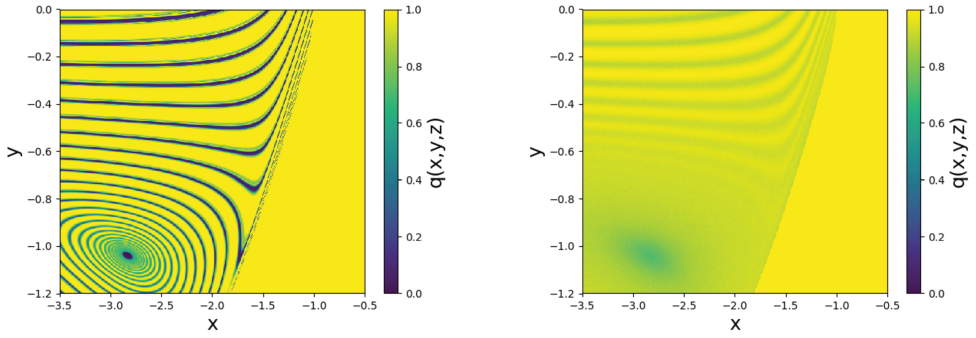
Figures 3.7b and 3.7c show the committor function in the case where a finite noise amplitude $\sigma \neq 0$ is considered. As can be seen by comparing Figs. 3.7a, 3.7b and 3.7c, adding a small noise blurs the visible structures of the deterministic case. For larger noise values ($\sigma = 10^{-3}$), Fig. 3.7c shows that the committor function looks smooth nearly everywhere (mathematically it is smooth everywhere, smooth here is used qualitatively and means with mild variations). This means that the *deterministic predictability is lost* for most initial conditions as ($0 < q < 1$). Then one cannot expect to predict the outcome in the way of a deterministic forecast. However, the occurrence of strong El-Niño events is *probabilistically predictable*: the value of the probability can be determined in practice with an excellent precision as it changes very slowly when one changes the initial conditions. It can also be seen on the figure that the occurrence of strong El-nino events is frequent ($q > 0.6$ almost everywhere). This is an indication that for such a value of σ we are in the noise-dominated regime.

The most interesting case is probably the one with the intermediate noise amplitude $\sigma = 0.00005$. In Fig. 3.7b, we delineate 4 regions: first, two regions of perfect (deterministic) predictability, where the event will occur with probability very close to 1 (region 1) or to 0 (region 2). Second, there exists a *probabilistically predictable region* (region 3) with good (probabilistic) predictability properties, where a value $0 < q < 1$ can clearly be predicted with very mild dependence with respect to the initial conditions. Finally, the region 4, which is unpredictable in practice. In this region, the strong dependence with respect to the initial condition prevents any practical prediction, either deterministic or probabilistic, of the precise value of q. While regions 1, 2 and 4 are reminiscent of their deterministic counterparts (Fig. 3.7a), region 3 is not. Instead, the behavior in this region is similar to the strong noise case shown in Fig. 3.7c. It is a region where the stochasticity is large enough to smooth out the deterministic values of q. The fact that it occurs even at very low noise amplitude is probably related to extremely unstable parts of the phase space, for instance for trajectories passing close to unstable fixed points or orbits. The existence of such features, and especially the new and most interesting *probabilistically predictable region* (region 3), should be generic for most prediction problems in climate dynamics.



(a) $\sigma = 0.00005$ (b) $\sigma = 0.001$

Figure 3.8: Color plot of the committor function $q(x, y, z)$ in the plane $y = -1.1580$, for $\sigma = 0.00005$ (left) and 0.001 (right).



(a) $\sigma = 0.00005$ (b) $\sigma = 0.001$

Figure 3.9: Color plot of the committor function $q(x, y, z)$ in the plane $z = 1.3409$, for $\sigma = 0.00005$ (left) 0.001 (right).

3.4.2 Dynamical characterization of the probabilistically predictable region

In order to understand the reason for which the probabilistically predictable region arises, it can be useful to introduce a quantity which characterizes sensitivity to small perturbations. This quantity is the largest finite time Lyapunov exponent and it is a measure of the sensitivity to initial conditions for the deterministic system. Let us consider two trajectories of the deterministic system ($\sigma = 0$) with initial conditions $\mathbf{x}$ and $\mathbf{x} + \delta\mathbf{x}$ and let $\Delta(t)$ be the value of the euclidean distance between the two trajectories at time t. The largest Lyapunov exponent Λ_L is defined as

$$\Lambda_L = \lim_{t\to+\infty} \lim_{\Delta(0)\to 0} \frac{1}{t} \log\left(\frac{\Delta(t)}{\Delta(0)}\right). \tag{3.13}$$

Since we are dealing with predictions with a time horizon T, we believe it is more appropriate to define a finite-time version of Λ_L. Hence, we compute the largest finite time Lyapunov exponent λ_L as $\lambda_L = \frac{1}{T}\log\left(\frac{\Delta(T)}{\Delta(0)}\right)$. Note that the initial perturbations have to be considered small but finite as we have taken $\mathbf{x}$ and $\mathbf{x} + \delta\mathbf{x}$ into the same cell $\mathcal{C}_k$. Positive values of λ_L mean that the distance between the trajectories grows exponentially. This quantity is shown for the Jin-Timmerman model in the $x = -2.831$ plane in Fig. 3.10. Comparing this figure with Fig. 3.7a is enlightening: the regions of perfect predictability

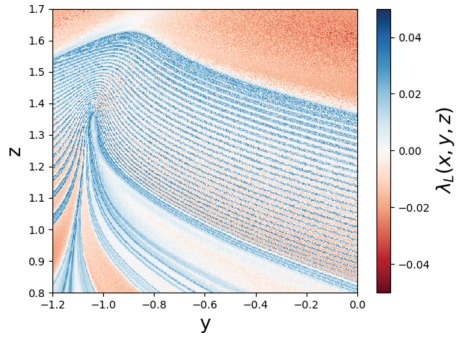


Figure 3.10: Color plot of the maximum finite time Lyapunov exponent λ_L as a function of the initial condition (x, y, z) in the plane $x = -2.831$.

(1 and 2) are associated to values of λ_L which are either negative or close to 0. Moreover, $\lambda_L < 0$ for almost all regions for which the deterministic committor is equal to 1 (see the two thick yellow bands at the bottom left of Fig.3.7a). From this description, it seems that λ_L is not the correct quantity to explain the emergence of the probabilistically predictable region. However, from a careful analysis of Fig. 3.10 it can be noted that there is a region, close to the left boundary of region 2, for which the values of λ_L are positive also for points $\mathbf{x}$ such that $q(\mathbf{x}) = 1$. It means that this region is quite unstable with respect to small initial perturbations. Since this region belongs to the *probabilistically predictable region*, it is reasonable to say that the appearance of region 3 is related with this instability. In fact, although the instability region is a subset of the *probabilistically predictable region*, it should be noted that for $\sigma \neq 0$ the system is perturbed at any time. The ensemble of these small perturbations gives rise to a finite perturbation which could explain the growth of the region of instability.

To reinforce this conjecture, we compute the averaged value of the euclidean distance between two different trajectories $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$, with the same initial conditions but different realisations of the noise:$\langle d_{\max} \rangle = \langle \max_{t\in[0,T]} \|\mathbf{x}_1(t) - \mathbf{x}_2(t)\|^2 \rangle$. Figure 3.11 shows $\langle d_{\max} \rangle$ as a function of the initial condition in the plane $x = -2.831$ for $\sigma = 0.00005$: it can be seen that the typical values of $d_{\max}$ are larger in the probabilistically predictable region than in other regions. This means that trajectories starting in this region are more sensitive to noise induced perturbations than trajectories starting in different regions of phase space. Therefore, the emergence of the

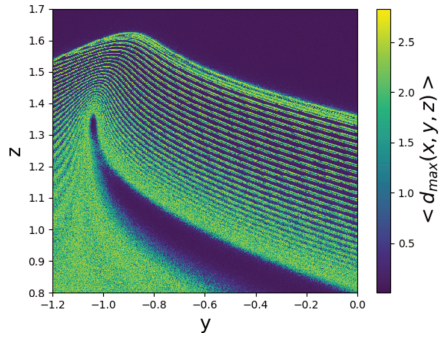


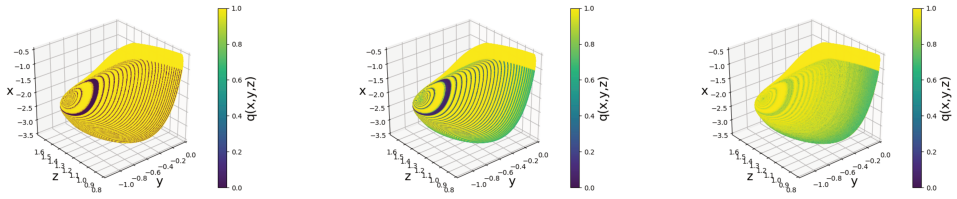
Figure 3.11: Color plot of the average value of the maximum distance $\langle d_{\max} \rangle$ as a function of the initial condition (x, y, z) in the plane $x = -2.831$, computed for $\sigma = 0.00005$.

probabilistically predictable region is qualitatively associated to an instability present in the deterministic system which is accentuated as the amplitude of the perturbations increases.

3.4.3 Committor function computed from long trajectories

In this section we discuss the committor function computed from an ensemble of long trajectories by means of Eq. (3.12). The motivation is that the very precise strategy adopted in Sec. 3.43.4.1 is unlikely to be adapted for real-world problems as well, because it requires a dynamical model and has a very large computational cost. Indeed, it requires to sample an ensemble of trajectories for every point in the phase space. For a data-based approach, it is usually possible to observe only the evolution of a trajectory (or of an ensemble of trajectories) over a very long time. However, individual trajectories, regardless of their length, do not fill the whole phase space. Indeed, they usually concentrate on the region where the invariant measure of the stochastic system is concentrated. Hence, the strategy adopted in Sec. 3.43.4.1 is appropriate for computing the committor function in an arbitrary plane while the use of Eq. (3.12) allows the computation of the committor function in the region where the invariant measure is concentrated.

The committor function computed using Eq. (3.12) is shown in Fig. 3.12, for the same three values of the noise amplitude as in Sec. 3.43.4.1. For this



(a) $\sigma = 0$ (b) $\sigma = 0.00005$ (c) $\sigma = 0.001$

Figure 3.12: Color map of the committor function $q(x, y, z)$ in correspondence to the most visited region of phase space, for $\sigma = 0$ (left, deterministic), 0.00005 (middle) and 0.001 (right).

figure, we used 10 000 trajectories of length $\mathcal{T} = 10^4$ (nondimensional time units) initialized in the strange attractor. Note that $\mathcal{T}$ is 10 times bigger than the first exit time $E[\tau_c]$ for $\sigma = 0.00005$ and 100 times greater than $E[\tau_c]$ for $\sigma = 0.001$ (see Fig. 3.6). This guarantees that the trajectories will be distributed according to the invariant measure of the system.

As already mentioned, Fig. 3.12 shows that the trajectories do not cover all the phase space but they are concentrated in a certain region. Furthermore, we can see that it is more appropriate to call it a manifold rather than a region. In fact, if it were an object of dimension 3, its intersection with a plane should define an area on that plane. Instead, it appears that the intersections between the object and planes are lines rather than areas. This is illustrated in Fig. 3.13 which shows an intersection between the object in Fig. 3.12b and the plane $x = -2.831$ (the same plane as in Fig. 3.7). This leads us to conclude that the trajectories are distributed over a manifold of dimension smaller than 3.

The comparison between Fig. 3.13 and Fig. 3.7b allows us to make another important remark on the committor function on the manifold. As can be seen in Fig. 3.13, the committor function takes values between 0 and 1 at the two ends of the line. For the rest, the line is made up of segments on which q takes the values 0 or 1. It is straightforward to recognize that the two ends of the line belong to the probabilistically predictable region while the central part of the line belongs to region 4. This highlights that the committor function computed from long trajectories provides useful information for many of the states of the system, that is, for all typical states. However, for atypical conditions where we have little or no information, using q to make predictions

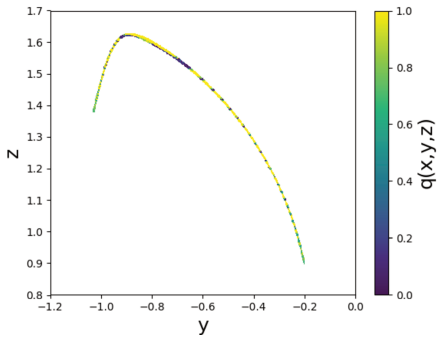


Figure 3.13: Color plot of the committor function for $\sigma = 0.00005$ at the intersection of the manifold with the plane $x = -2.831$.

can lead to erroneous results.

Having made these necessary considerations, we can continue the description of the committor computed from long trajectories. By comparing Fig. 3.12 with Fig. 3.3 we can immediately identify the strange attractor in the manifold on which the committor function is represented. The limit cycle is more difficult to visualize but its presence can be deduced from the spiral behavior present in the region inside the strange attractor and from the shape of the manifold boundaries that follow the shape of the limit cycle in Fig. 3.3.

The qualitative structure of the committor function on the manifold is similar to the one observed on plane cuts: in the deterministic case, we observe regions of perfect predictability, and regions where the sensitivity to initial conditions make it unpredictable in practice. When the noise is sufficiently strong, regions of probabilistic predictability appear, where a finite value of the probability of a strong El Niño event $0 < q < 1$ depends only mildly on the initial conditions (Fig. 3.12c). It can be noted, however, that the intermediate case, analogous to Fig. 3.7b, with coexistence of a region of deterministic predictability, a region of probabilistic predictability, and an unpredictable region, is more difficult to observe in this visualization.

Finally, we underline that the region of unpredictability, made by thin filaments where the committor is a highly fluctuating function, emphasizes again how the two attractors are intertwined in a complex way.

3.5 Conclusion

In this paper, we have introduced a mathematical concept, the *committor function*, encoding the probability that an event occurs within a given time, conditioned on the state of the system at the time of prediction. We believe it is an appropriate concept for many prediction problems in climate science in a range of time scales which we call the *predictability margin*. It corresponds to timescales for which a deterministic description of the system is no longer relevant, because of the sensitive dependence to initial conditions, but for which more precise probabilistic predictions than the climatological one can be made, because the system has not yet forgotten completely the initial condition.

In the context of a simple, low-dimensional stochastic model, the Jin-Timmerman model, in a regime of coexistence of a limit cycle and a strange attractor found by [Guckenheimer et al., 2017], we have shown that noise could induce transitions between the two attractors. These transitions correspond to regime shifts regarding the occurrence of strong El Niño events, which are periodic in the limit cycle, with a return time close to 15 years, and which do not occur at all in the strange attractor (in the deterministic case). In the stochastic case, the occurrence of strong El Niño events therefore becomes random, and the waiting times follow a Poisson statistics.

In this example, we have shown that the probability of occurrence of strong El Niño events had different predictability properties depending on the state of the system at the time of prediction. The most important result is that there exist regions of *probabilistic predictability*, where the event has a finite probability of occurring $0 < q < 1$, and this probability does not depend sensitively on the initial state, and regions of *probabilistic unpredictability* where the probability changes a lot if one changes by a small and finite amount the initial condition. We expect the existence of this dichotomy between probabilistically predictable and probabilistically unpredictable regions to be a generic feature for climate prediction problems at the predictability margin. We stress that this notion depends on the precision with which the initial condition can be assessed.

We have also discussed the methodological aspects for computing the committor functions. For our example, a small stochastic perturbation of a chaotic deterministic system, we have computed the committor function using two approaches. First, by direct sampling of ensembles of initial conditions close to any point in phase space, and second, through a data based

approach using observed trajectories. As soon as the number of degrees of freedom increases, the first method will become impossible to use in practice, because of the numerical cost. The second method may sometimes be associated with sampling issues, as one can get meaningful results only for the parts of the phase space that have been visited many times. Another method mentioned in Sec. 3.4, would be to solve a backward Kolmogorov equation. This method is impractical for systems with more than a few degrees of freedom. To be able to sample efficiently committor functions in large dimensions, more efficient data-based methods will be necessary, relying either on classical statistical methods or machine learning methods [Lucente et al., 2019]. The development of such methods shall be a prerequisite for studying climate prediction problems using more realistic models.

Chapter 4

Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain

In the previous chapters, the committor function was introduced and it was computed for a low-dimensional model for El-Niño showing that the ability to predict the probability of occurrence of extreme events strongly depends on the initial condition. The purpose of this chapter is to propose a new approach, based on the analogue Markov chain, for data-based learning of committor functions. It will be shown that such learned approximate committor functions are extremely efficient scoring functions when used with the Adaptive Multilevel Splitting algorithm. This approach is illustrated in the context of two stochastic systems: a gradient dynamics in a three-well potential and the Charney-DeVore model, which is a paradigmatic toy model of multistability for atmosphere dynamics. For these two dynamics, it is shown that observing few transitions is enough to have a very efficient data-based scoring function for the rare event algorithm. The main advantages of this new approach are that rare events can be simulated with minimal prior knowledge, and results are much more accurate than those obtained with a user-designed scoring function.

This chapter is the result of a collaboration with J. Rolland, C. Herbert and F. Bouchet and has been submitted to the "Journal of Statistical Mechanics: Theory and Experiment" (J. Stat. Mech.). Therefore, this chapter

is self-consistent can be read independently from the rest of the thesis. The reader is advised that the notation of this chapter is slightly different from the rest of the manuscript. This choice was made to make the notation consistent with the one known in the rare event algorithm literature. Here, an arbitrary subset of phase space is called D instead of C, and the first hit time is called T instead of τ, which is used to indicated the duration of reactive trajectories. Many of the topics discussed in the opening sections of this chapter have already been covered in previous chapters. In order to avoid redundancies, the reader of the manuscript can start reading this chapter from Sec. 4.3.

4.1 Introduction

Rare events are often extremely important, either because they have a huge impact, for instance climate extremes [Seneviratne et al., 2021], or because they change completely the structure of the system and shape its history over long times, for instance the dynamics of metastability [Farkas, 1927] and multistability phenomena [Eyring, 1935, Kramers, 1940]. Such rare events are so important in many physics, chemistry, and biology applications that specific tools have been developed to study them, by the statistical mechanics and applied mathematics community: theoretical approaches and dedicated computation algorithms.

In this paper, we are mainly interested in computational approaches for rare events. A key difficulty in numerical computation is that these rare events can be so rare that simulating them directly might be prohibitively expensive. Since the 50' [Kahn and Harris, 1951], rare event algorithms and simulations [Bucklew, 2013] that aim at reducing the computational cost have been devised. They have been used to address many problems in statistical physics, for instance studying percolation [Adams et al., 2008], liquids physics [Allen and Tildesley, 2017], Lyapunov exponents [Tailleur and Kurchan, 2007], dynamical phase transitions [Pérez-Espigares and Hurtado, 2019], first order phase transitions [Rolland et al., 2016], just to cite a few examples among many others. Chemical physics, biochemistry and the study of biomolecules has inspired many new technics, see for example [Bolhuis et al., 2002, Noé et al., 2009, Metzner et al., 2009, Hartmann et al., 2014]. Recent uses in biology models [Donovan et al., 2016] and ecology has also to be noticed.

Recently, rare events have been studied in far from equilibrium systems and non-equilibrium steady states, where one starts from dynamics without detailed balance. Rare event technics have then been extended to scientific fields so far unexpected, with complex dynamics. For instance in study of multistability on turbulence [Laurie and Bouchet, 2015, Bouchet et al., 2019], study of intermittency in turbulence models [Grafke et al., 2013,Grafke et al., 2015, Ebener et al., 2019], transitions to turbulence in pipe and Couette flows [Rolland, 2018, Rolland, 2021, Nemoto and Alexakis, 2018], rogue waves [Dematteis et al., 2018], atmosphere dynamics [Bouchet et al., 2019,Simonnet et al., 2021], climate dynamics [Ragone et al., 2018, Webber et al., 2019,Ragone and Bouchet, 2020,Plotkin et al., 2019,Finkel et al., 2021,Finkel et al., 2020], astronomy [Woillez and Bouchet, 2020, Abbot et al., 2021], among many other examples.

For such non-equilibrium problems, without detailed balances, one can use either computations related to minimum action methods, possibly related to large deviation theory (see for instance [Grafke and Vanden-Eijnden, 2019]), or the vast family of splitting algorithms or cloning algorithms [Kahn and Harris, 1951, Del Moral, 2012, Cérou and Guyader, 2007]. However, for many applications, for instance in turbulence, climate, atmosphere dynamics or astronomy, the system is either deterministic, or may be stochastic, but one has not access to a precise noise statistics, or rare events are not produced directly by the model noise but rather by internal fluctuations. In all these cases, any method that rely on an a-priori given bare action is not appropriated. Then the only possible choices, for rare event algorithms, are splitting algorithms. Those algorithms have indeed been empirically shown to work well for some classes of deterministic chaotic dynamical system [Wouters and Bouchet, 2016, Ragone et al., 2018]. An alternative route for studying rare events, without rare event algorithms, would be to use methods that require only short off equilibrium simulations, for instance through resimulating and milestoning [Noé et al., 2009, Vanden-Eijnden and Venturoli, 2009] or coarse graining of a reduced space of collective variables [Finkel et al., 2021, Finkel et al., 2020]. Such approaches might be very relevant, however only when the system is simple enough or when one knows sufficiently well the system to define a-priori relevant collective variables.

The main aim of this paper is to develop the methodology of splitting algorithms such that they might actually be used, practically, for genuinely complex dynamics. The general principle of splitting algorithms is to perform ensemble simulations, select trajectories prone to produce extremes,

discard other less interesting ones, and resimulate from the interesting ones. The effectiveness of these algorithms strongly relies on the quality of the score function which is used for the selection stage. For complex dynamics, in cases when the dynamics is simple enough or the phenomenology of the dynamics is well understood to devise good score functions, splitting algorithms are wonderful tools. For instance, they were used to compute rare event probabilities, which were totally unreachable with direct numerical simulations, for stochastic partial differential equation [Rolland et al., 2016], atmosphere turbulent flows [Bouchet et al., 2019,Simonnet et al., 2021], or full complexity climate models [Ragone et al., 2018]. However, without a good score function, splitting algorithms might completely fail. If the score function is not too bad, but not very good, splitting algorithms happen to actually produce efficiently rare events, but might suffer from the phenomenon of apparent biases for the estimation of probabilities [Glasserman et al., 1998, Bréhier et al., 2016a]. The aim of this work is to propose a new methodology to solve these problems and to be able to use splitting algorithms in very complex dynamics without a-priori knowledge or understanding of a simple effective description of the dynamics.

For many splitting algorithms, there exists a mathematical characterization of an optimal score function: a score function which minimizes the algorithm variance for the computation of the rare event probability and will be very efficient in practice. For instance, for the Adaptive Multilevel Splitting (AMS) [Cérou and Guyader, 2007], to be used in this article, the committor function is the optimal score function [Cérou et al., 2019a] . The committor function is the probability that a trajectory visits a region $\mathcal{B}$ of the phase space before another region $\mathcal{A}$, as a function of the initial condition [Onsager, 1938] . If $\mathcal{B}$ is the set of rare events of interest, the committor function is then a probabilistic measure of the progress towards the rare event. The committor function is also a central object of transition path theory [E et al., 2005, Weinan and Vanden-Eijnden, 2006, Vanden-Eijnden et al., 2010, Metzner et al., 2006]. A key difficulty is that this optimal score function, the committor function, is actually the rare event probability conditioned on the state of the system. It contains the information one wishes to compute. One has thus no easy access to it.

For similar problems, when one would need to know an approximation of a function to efficiently compute this function itself, it is very natural to consider an iterative procedure: a feedback control between the efficient algorithm to produce the data and the learning of the function itself. The

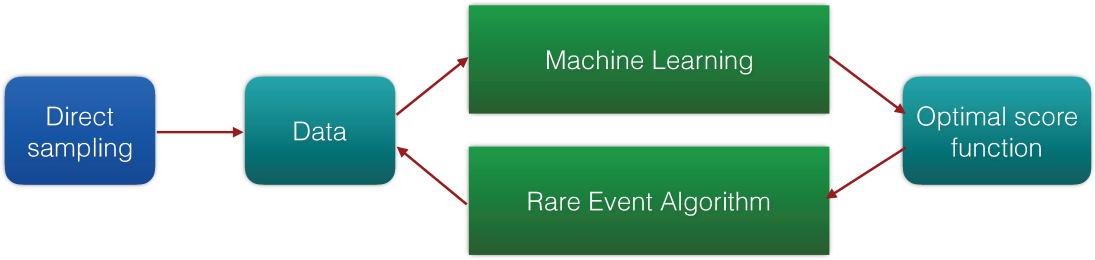


Figure 4.1: Sketch of a feedback control iterative procedure between the rare event algorithm and the machine learning of an approximate optimal score function. The learning of an approximation of the optimal score function makes the algorithm more efficient, and the algorithm provides more data for a better quality of the learning procedure.

learning of an approximation of the optimal score function makes the algorithm more efficient, and the algorithm provides more data for a better quality of the learning procedure. This is for instance the idea behind the Wang and Landau algorithm [Wang and Landau, 2001], in multicanonical methods for equilibrium statistical mechanics, or the idea at the base of adaptive importance sampling [Bugallo et al., 2017]. This feedback iterative procedure is illustrated by figure 4.1. We have already implemented such a feedback iterative procedure for the Giardina–Kurchan cloning algorithm, a specific example of a splitting algorithm [Nemoto et al., 2016]. However, the learning step in this example was extremely simple as the function to be learned was a function over a one-dimensional space. We want to extend this approach to more complex dynamics.

Many interesting methods have been or are currently being devised to learn committor functions: based on direct machine learning [Pozun et al., 2012], using a characterization of the committor function for diffusions as a solution of a partial differential equation [Khoo et al., 2019, Li et al., 2019], computing the committor function from a finite state Markov chain [Schütte et al., 1999, Prinz et al., 2011, Noé and Rosta, 2019, Tantet et al., 2015], possibly a Markov state model approximation of the dynamics [Ulam, 2004]. Recently a very interesting approach has been considered starting from a Galerkin approximation of the dynamics generator, or the Koopman operator. Finite dimensional approximations of the dynamics generator have been

used to identify good reaction coordinates [Froyland et al., 2014, Bittracher et al., 2018], or to evaluate eigenfunctions of the operator [Giannakis et al., 2015, Giannakis, 2019, Williams et al., 2015, Mardt et al., 2018], sometimes with climate applications [Giannakis et al., 2015, Giannakis, 2019]. Recently such direct Galerkin approximation has been used to directly compute committor function, avoiding the burden of discretizing a high dimensional phase space [Thiede et al., 2019, Strahan et al., 2021]. Several computation of committor functions have been performed with applications in either geophysical fluid dynamics or climate applications [Finkel et al., 2021, Miron et al., 2021, Finkel et al., 2020, Lucente et al., 2019, Lucente et al., 2021], using either direct or involved approaches.

The aim of this paper is to test the coupling of data-based learning of approximate score functions with rare event algorithm, in the spirit of figure 4.1. As we are specifically interested in complex dynamics, the learning strategy needs to have the potentiality to scale well in very large dimensions. Moreover, it should be suited for any dynamics, including chaotic deterministic systems or dynamics for which the noise is irrelevant for the process of interest. It also needs to be not too greedy in terms of dataset length. Among all the possible approaches for learning committor functions, the ones based on approximation of the dynamics generator seem to be best suited [Thiede et al., 2019, Strahan et al., 2021].

In this paper we propose a new method based on an approximation of the dynamics generator. For this purpose, we consider a slightly modified version of the analogue method, firstly proposed by Lorenz [Lorenz, 1969c, Lorenz, 1969a]. The idea behind the analogue method can be summed up by Maxwell's sentence [Garnett and Campbell, 1882] "*From like antecedents follow like consequents*". This approach is nowadays used to build stochastic weather generators [Yiou, 2014, Yiou and Déandréis, 2019]. A key remark is that the analogue method defines a Markov chain which an approximation of the dynamics generator of the original dynamics. Then a learned approximate committor function can be computed using classical methods for computing Markov chain committor functions. This new way to compute committor function, based on the analogue Markov chain, is an alternative path that leads to dynamic based estimates of the committor function. We show in this paper that this method is actually very simple, robust, and efficient. We show that the learned committor function, based on the analogue Markov chain, is more precise and efficient than the classical K-nearest neighbors regression, which computes the committor by averaging

the observations of K nearby points.

After having put forward and tested this committor function computation using the analogue Markov chain, we couple it to the Adaptive Multilevel Splitting (AMS) [Cérou and Guyader, 2007]: we directly use the data-based approximate committor function as a score function for the algorithm. We make a precise study that shows that for large enough data sets, the performance of the AMS algorithm is greatly improved. We get rid of the apparent bias phenomena and can compute rare events without a-priori knowledge of the dynamics.

To summarize the previous discussion, the purpose of this work is twofold. On the one hand, we introduce a data-driven approach which can be used to compute the committor function, and which exploits the dynamical information provided by the observed dynamics. On the other hand, we show how it is possible to use this method to build a learned score function for efficient rare event algorithms. We illustrate our approach for two dynamics. First a stochastic gradient dynamics in a three-well potential, in dimension two. Then we study the Charney-DeVore model, which is a paradigmatic toy model of multistability for atmosphere flows [Charney and DeVore, 1979], with six variables. For these two dynamics, we show that having observed a few transitions is enough to have a very efficient data-based score function for the rare event algorithm.

The paper is organized as follows. In Sec. 4.2., we define and discuss the mathematical properties of the committor function, we explain a direct sampling strategy, and define the Brier score which quantifies the quality of an approximate committor function. Sec. 4.3 is devoted to the analogue method and how it can be used to obtain a dynamics-bases estimate of the committor function. Finally, in Sec 4.4 we introduce the AMS rare events algorithm, we use it with a score functions which is the learned analogie Markov chain committor function, and we discuss the improvements given by this approach.

4.2 The committor function

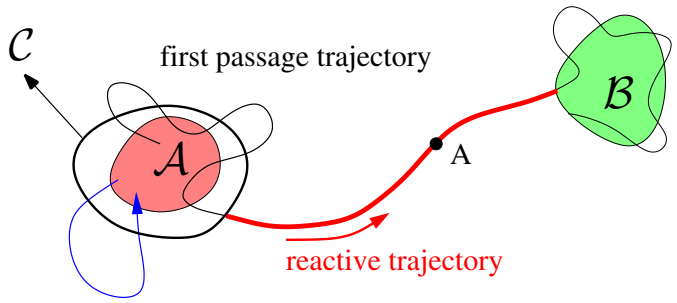


Figure 4.2: An example of first passage trajectory from $\mathcal{A}$ to $\mathcal{B}$ is shown. The transition path, also called reactive trajectory, is highlighted in red.

4.2.1 Definition of the committor function for a Markov process

For a Markov process, a *committor function* [E et al., 2005, Weinan and Vanden-Eijnden, 2006, Vanden-Eijnden et al., 2010, Metzner et al., 2006] is the probability to hit a set $\mathcal{B}$ of the phase space before another set $\mathcal{A}$, conditioned on the knowledge of the initial condition. With adapted definitions of the sets $\mathcal{A}$ and $\mathcal{B}$, it can be the probability of transition between metastable states [Lopes and Lelièvre, 2019], see Fig. 4.2, or to the probability that an event occurs within a given timeframe [Lestang et al., 2018, Lucente et al., 2021].

To give a more precise definition, we consider a discrete time stochastic process on a phase space $\mathcal{X}$. A given realization of the process will be noted as $\{X_n\}_{1\leq n\leq N_t}$, with $X_n \in \mathcal{X}$. The *first hitting time* $T_{\mathcal{D}}(\mathbf{x})$ of a set $\mathcal{D} \subset \mathcal{X}$ is defined as

$$T_{\mathcal{D}}(\mathbf{x}) = \inf\{n : X_n \in \mathcal{D} | X_0 = \mathbf{x}\}. \tag{4.1}$$

The committor function $q(\mathbf{x})$ is the probability that the first hitting time of a set $\mathcal{B}$ be smaller than the first hitting time of set $\mathcal{A}$, as a function of the initial condition, i.e.

$$q(\mathbf{x}) = \mathbb{P}[T_{\mathcal{B}}(\mathbf{x}) < T_{\mathcal{A}}(\mathbf{x})]. \tag{4.2}$$

This definition immediately generalizes for continuous time Markov processes.

If the dynamics is a stochastic differential equation, $q(\mathbf{x})$ is the solution of the Dirichlet problem [E et al., 2005, Thiede et al., 2019]:

$$\mathcal{L}q(\mathbf{x}) = 0 \text{ with } \mathrm{q}(\mathbf{x}) = 0 \text{ if } \mathbf{x} \in \mathcal{A} \text{ and } \mathrm{q}(\mathbf{x}) = 1 \text{ if } \mathbf{x} \in \mathcal{B}, \tag{4.3}$$

with $\mathcal{L}$ the adjoint of the Fokker-Planck operator:

$$\mathcal{L} = \sum_i a_i(\mathbf{x}) \frac{\partial}{\partial x_i}(\cdot) + \sum_{ij} D_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j}(\cdot), \tag{4.4}$$

where $\mathbf{a}$ is the drift coefficient and $\mathbf{D}$ the diffusion coefficient. One way to compute a committor function is to solve this partial differential equation. In practice, such a computation is impossible, using standard techniques as soon as the system has more than a few degrees of freedom. This equation can be used for computing approximate solutions, using machine learning, for systems of dimension of the order of magnitude of ten [Khoo et al., 2019, Li et al., 2019].

4.2.2 Direct sampling of the commitor function

In this section we consider data-based methods for the computation of a committor function. The data consist of sets of trajectories of the stochastic process. The simplest method is to directly use the definition (4.2). In practice, to compute the function at point $\mathbf{x}$, we initialize an ensemble of N trajectories in $X_0 = \mathbf{x}$ and evolve them until they reach $\mathcal{A}$ or $\mathcal{B}$. Let $N_{\mathcal{B}}$ be the number of trajectories that have reached $\mathcal{B}$. Then, the value of the committor function at point $\mathbf{x}$ can be estimated as

$$q(\mathbf{x}) = \frac{N_{\mathcal{B}}}{N}. \tag{4.5}$$

Like the Dirichlet problem (4.3), this method can only be applied if the equations of motion are known, and it is inapplicable for high dimensional systems, as it requires simulating many trajectories for each point of phase space where we want to compute the committor function. The numerical burden thus increases exponentially with the dimension of the system.

For an ergodic process, the committor function $q(\mathbf{x})$ and the stationary distribution function $\rho(\mathbf{x})$ can be computed from an observed trajectory $\{X_n\}$ from the formulas

$$\rho(\mathbf{x})q(\mathbf{x}) = \lim_{N_t \to \infty} \frac{1}{N_t} \sum_{n=0}^{N_t} \delta\left(X_n - \mathbf{x}\right) 1_{\{T_{\mathcal{B}}(X_n) \leq T_{\mathcal{A}}(X_n)\}} \text{ and}$$

$$\rho(\mathbf{x}) = \lim_{N_t \to \infty} \frac{1}{N_t} \sum_{n=0}^{N_t} \delta\left(X_n - \mathbf{x}\right), \tag{4.6}$$

where δ is a Dirac delta function, and $1_{\{T_{\mathcal{B}}(X_n)\leq T_{\mathcal{A}}(X_n)\}}$ takes value 1 if the trajectory visits set $\mathcal{B}$ before set $\mathcal{A}$ starting from X_n, and 0 otherwise. Numerically, $q(\mathbf{x})$ can be computed from (4.6) after spatial and temporal discretization of the process (see for instance [Lucente et al., 2019, Lopes and Lelièvre, 2019, Lucente et al., 2021]). Unlike the previous methods, this approach is applicable even if we do not know the equations of motion. Its numerical cost does not depend on the dimension of phase space, but it only provides estimates of the committor function on points which neighborhood was visited many times by the observed trajectory.

4.2.3 Estimating the committor function for any point of the phase space

In Sec. 4.2.2, we have presented a direct sampling method to estimate the committor function based on data. However, it provides values only on the set of points that was visited along the trajectory. This is also true for the other data-based method that we will present in Sec. 4.3, the *analogue method*. For applications, we may need to estimate the value of the committor function for points which were not in the learning dataset. This may be the case simply for graphical representations of the committor function along a line or on a plane in phase space (e.g. Sec. 4.3.3). Even more importantly, to use the estimated committor function as a score function with the AMS algorithm (Sec. 4.4), we need to be able to compute it for arbitrary points in phase space.

To do so, we will use a *nearest neighbor method* [Altman, 1992]. Let us denote $\{\mathbf{X}_n\}_{1\leq n\leq N_p} \in \mathbb{R}^D$ the learning dataset, for which we have an estimate of the committor $\hat{q}(\mathbf{X}_n)$. For any point $\mathbf{y} \in \mathbb{R}^D$, we search the κ nearest neighbors (using the Euclidean distance $d(\mathbf{y},\mathbf{x})^2 = \sum_{i=1}^D (y_i - x_i)^2$), corresponding to indices $n_j \in [\![1, n]\!]$ in our dataset, for $1 \leq j \leq \kappa$. We then perform a weighted average of the corresponding values of the committor:

$$\hat{q}(\mathbf{y}) = \frac{\sum_{j=1}^{\kappa} w_j \hat{q}(\mathbf{X}_{n_j})}{\sum_{j=1}^{\kappa} w_j}. \tag{4.7}$$

The weights w_j can be chosen uniform: $w_j = 1$ (like in Sec. 4.3.3) or given by a kernel, such as $w_j = e^{-\frac{d(\mathbf{y},\mathbf{X}_{n_j})^2}{\omega^2}}$, where $\omega > 0$ is a kernel width (like in Sec. 4.4), depending on the application.

4.2.4 Estimation of the quality of an approximate committor function: the Brier score

In this section we address the issue of how to quantify the quality of an estimate of the committor function. In what follows, the true committor function is denoted by q while $\hat{q}$ stands for our estimate. As the committor function is the probability of a binary variable, it is natural to look for a score for a forecast of a binary variable. We also require that this score can be computed directly from observations. The Brier score is a natural candidate.

We first consider Y a random variable with binary outcomes, $Y \in \{0, 1\}$, and a Bernoulli distribution: $\mathbb{P}[Y = 1] = q$ and $\mathbb{P}[Y = 0] = 1 - q$. We look for an estimator that quantifies the value of an estimation $\hat{q}$ of q.

One of the simpler quantities having the required properties was proposed in 1950 by Brier [Brier, 1950]. We consider a $\{Y_n\}_{1 \leq n \leq N}$, N independent realizations of the variable Y. The Brier score is defined as

$$B_N = \frac{1}{N} \sum_{n=1}^{N} (\hat{q} - Y_n)^2, \tag{4.8}$$

The Brier score is thus a random variable, with values between 0 and 1.

The random variable $(\hat{q} - Y_n)^2$ takes value $(1 - \hat{q})^2$ with probability q and value $\hat{q}^2$ with probability $(1 - q)$. Then the average value of $B_N(\mathbf{x})$ is

$$\mathbb{E}(B_N) = (1 - \hat{q})^2 q + \hat{q}^2 (1 - q) = q(1 - q) + (\hat{q} - q)^2. \tag{4.9}$$

The expectation of the Brier score $B_N(\mathbf{x})$ is therefore the sum of two terms. The first one, $q(1 - q)$ is related to the stochastic nature of the forecast and is independent of $\hat{q}$, while the second one, $(\hat{q} - q)^2$, is related to the error made in the estimation of q. The closer is the forecast $\hat{q}$ to the real value q, the lower is the Brier score. The Brier score has a fixed lower bound $q(1-q)$. We see that the Brier score is merely a quadratic measure of the error (the second term) plus a constant term (the lower bound). However, while the computation of the quadratic error requires the knowledge of the truth q, the computation of the Brier score does not require the knowledge of q. In the limit $N \to \infty$, we have an ergodic average and $\lim_{N \to \infty} B_N = \mathbb{E}(B_N)$.

We now extend naturally the definition of the Brier score to the case of Markov processes and committor functions. We consider a set of events

$\{(\mathbf{X}_n, Y_n)\}_{1\leq n\leq N}$, where $\mathbf{X}_n$ are points in the phase space distributed according to the invariant measure ρ of the Markov process, $\mathbb{E}\left[(\delta\left(\mathbf{X}_n - \mathbf{x}\right)\right] = \rho(\mathbf{x})$, and Y_n are binary variables which takes the value 1 with probability $q\left(\mathbf{X}_n\right)$ and value 0 with probability $1 - q\left(\mathbf{X}_n\right)$. For instance, the couples $(\mathbf{X}_n, Y_n)$ can be sampled along one or several trajectories of the Markov chain, where $\mathbf{X}_n$ are the states of the Markov chain and Y_n is equal to zero if the first hitting time of $\mathcal{B}$ after n is smaller than the first hitting time of $\mathcal{A}$ after n.

We want to estimate the quality of an approximation $\hat{q}$ of the committor function q. Then the committor function Brier score is defined as

$$BT_N = \frac{1}{N}\sum_{n=1}^{N}\left[\hat{q}\left(\mathbf{X}_n\right) - Y_n\right]^2, \tag{4.10}$$

Extending directly the previous computations, and assuming ergodicity, we have

$$\mathbb{E}(BT_N) = \lim_{N\to\infty} BT_N = \|q - \hat{q}\|_\rho^2 + \left\|\sqrt{q(1-q)}\right\|_\rho^2, \tag{4.11}$$

where $\|f\|_\rho^2 = \int_{\mathcal{D}} f^2(\mathbf{x})\rho(\mathbf{x})\,\mathrm{d}\mathbf{x}$ is L^2 norm weighted according to the invariant measure. Then the committor Brier score is $\|q - \hat{q}\|_\rho^2$, the weighted L^2 norm of the difference $q - \hat{q}$, up to the constant term $\left\|\sqrt{q(1-q)}\right\|_\rho^2$. While the weighted L^2 norm cannot be computed without the knowledge of q and ρ, the Brier score can be directly computed from the data by the ergodic average (4.10).

4.3 The analogue Markov chain

In this section we introduce the analogue method in one of its current versions [Yiou, 2014, Lguensat et al., 2017, Yiou and Déandréis, 2019, Platzer et al., 2021b, Platzer et al., 2021a]. It provides a way to build effective dynamics from the data that can be reused to generate new trajectories of the system under consideration at a lower computational cost. Although more precise definitions will be given throughout the section, we think that briefly illustrating the analogue method in its original form proposed by Lorenz [Lorenz, 1969c, Lorenz, 1969a] in 1969 is both conceptually and historically instructive. Furthermore, this can be seen as a particular case of the method we will present in which only $K = 1$ analogue is considered.

In a nutshell, the idea is the following. Suppose we have access to a time series of observations that we will denote by $\{X_n\}_{1\leq n\leq N_t}$, at times $t_n = n\delta t$ where δt is the sampling time step. Starting from a state $\mathbf{x}$ at time t, we want to predict a possible dynamical evolution at a time $t + l\delta t$. We search among the available data $\{X_n\}_{1\leq n\leq N_t}$ the closest to $\mathbf{x}$, i.e an analogue, which will be denoted by $X_{n_\star}$:

$$X_{n_\star} = \underset{\{X_n\}}{\operatorname{argmin}}\{d(\mathbf{x}, X_n)\}, \tag{4.12}$$

where $d(\cdot)$ is a distance. After identifying the best analogue $X_{n_\star}$, the prediction of $\mathbf{x}(t + \Delta t)$, denoted $\tilde{\mathbf{x}}(t + l\delta t)$, will be

$$\tilde{\mathbf{x}}(t + \Delta t) = X_{n_\star+l}. \tag{4.13}$$

This method was intended by Lorenz as a deterministic prediction. In the following we are rather interested by stochastic predictions, either because the actual dynamics itself is stochastic, or because we understand the analogue method as an approximate effective description of a chaotic dynamics. For stochastic prediction, we will use K analogues rather than a single one.

4.3.1 Definition of the analogue Markov chain

Let $\{X(t)\}_{0\leq t\leq+\infty}$ be a dynamical process that takes values in the phase space $\mathcal{X} \subset \mathbb{R}^D$. The nature of the process, i.e. whether it is deterministic or stochastic, Markovian or not, is irrelevant to the discussion. Suppose that a realization of this process is observed at regular time intervals δt during a total time $\mathcal{T} = N_t\delta t$ and let $\{X_n\}_{1\leq n\leq N_t}$ denote this sampled trajectory made up of N_t points. Each point X_n is in $\mathbb{R}^D$, where D is the dimension of the phase space.

We will build a Markov chain that is a data-based approximation of the initial process, based on a generalization of the Lorenz analogue method. We now define possible transitions starting from an observed state X_n. Rather than considering just a single nearest neighbor of X_n in the observed data, we will use the K nearest neighbors, where K is a positive number. Those K nearest neighbors are denoted $\{\hat{X}_n^{(k)}\}_{1\leq k\leq K}$. After identifying analogues $\{\hat{X}_n^{(k)}\}$, we suppose that we can have a transition between the state X_n and all the possible images of this set of points. These images will be denoted by $\{\hat{X}_{n+1}^{(k)}\}_{1\leq k\leq K}$ and the probability to have a transition between X_n and

$\hat{X}_{n+1}^{(k)}$ is set to $\frac{1}{K}$. An illustration of the analogue Markov chain is shown in Fig. 4.3.

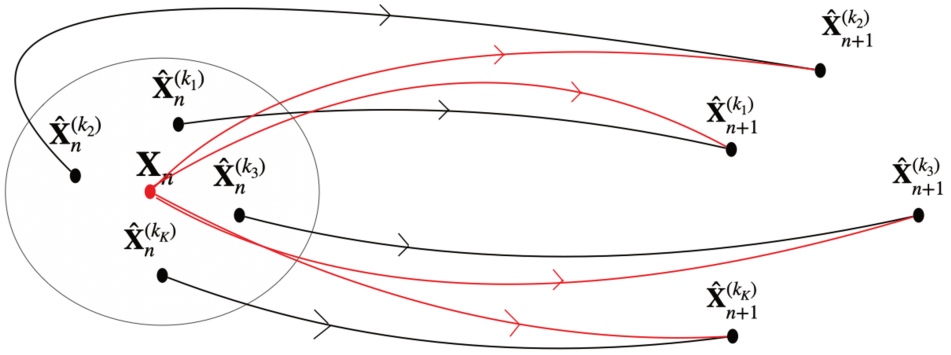


Figure 4.3: Schematic of the analogue method. On the left-hand side of the figure a point X_n surrounded by its analogues $\{\hat{X}_n^{(k)}\}$ is shown (here $K = 4$). On the right-hand side the observed images $\{\hat{X}_{n+1}^{(k)}\}_{1\leq k\leq K}$ of the analogues one time step forward are shown. The transitions observed in the the data are represent by black lines which link the analogues with their corresponding images. Red lines are associated to the possible transitions from the state X_n of the analogue Markov chain.

With this definition, we see that K is both the number of analogues and the number of possible transitions from any state of the Markov chain. One needs K to be large enough to properly approximate all the possible transitions from a given state. At the same time, the larger K, the further the analogue, and the larger the error incurred by using a point further from X_n. The optimal value of K will be a tradeoff between these two effects, as a balance between precision and complexity. In practice, K will be chosen empirically, for instance using cross validation.

The selection of neighbors is subordinated to the choice of a distance. The best distance most probably depends on the system under investigation. Distances will be specified on a case-by-case basis.

The analogue Markov chain is a Markov chain on the finite set of N_t observations. In practice, we introduce a matrix with integer entries, $\mathcal{T} \in \mathcal{M}_{N_t K}(\mathbb{N})$. Each row $n \in [\![1, N_t]\!]$ of $\mathcal{T}$ contains the indices of the K nearest neighbors of the point X_n, i.e. the indices $n_1, \ldots, n_K$ such that $\hat{X}_n^{(k)} = X_{n_k}$

for $1 \leq k \leq K$. We stress that $\mathcal{T}$ is not the transition matrix of the Markov chain, to be described latter. $\mathcal{T}$ is rather a matrix of indices of the states.

Since we cannot associate any transition to the end-point X_{N_t}, this point will be excluded from the possible candidates for the analogues of each point. To summarize, each entry of $\mathcal{T}$ can take values between 1 and $N_t - 1$, i.e. $\mathcal{T}_{nj} \in [\![1, N_t - 1]\!]$ for all n, j such that $1 \leq n \leq N_t$ and $1 \leq j \leq K$.

To generate a synthetic trajectory, we can proceed as follows. We start with a state $s_0 \in [\![1, N_t]\!]$. Then, we generate a random integer k distributed uniformly in the interval $[1, K]$ and the new state will be $s_1 = \mathcal{T}_{s_0 k} + 1$. This procedure is iterated to build the entire trajectory. Through this method we build a Markov chain whose states are $\{X_n\}_{1 \leq n \leq N_t}$, i.e. the learning dataset.

We now describe the transition matrix $G \in \mathcal{M}_{N_t}(\mathbb{R})$. The elements G_{nj} of G are the probability to observe a transition from the state n to the state j. They are given by

$$\begin{cases} G_{nj} = \frac{1}{K} \text{ if } \exists\ k_\star \in [\![1, K]\!] :\ j = \mathcal{T}_{nk_\star} + 1\,, \\ G_{nj} = 0 \text{ otherwise}\,. \end{cases} \tag{4.14}$$

G is an approximation of the propagator $\mathbb{P}(X_j | X_n)$ of the real dynamics.

Given an observable at time t, represented by a column vector $f(t) = f_i(t)$, the observable at time $t+1$ is obtained by applying the operator G to $f(t)$, i.e.

$$f(t+1) = Gf(t). \tag{4.15}$$

Therefore, G plays the same role as the generator of a continuous stochastic process.

Concerning the temporal evolution of probabilities there are two possibilities:

- consider probabilities as row vectors π and let G act to the right, i.e $\pi(t+1) = \pi(t)G$;
- consider column vectors π and let them evolve by applying the adjoint operator $G^\dagger$, i.e. $\pi(t+1) = G^\dagger \pi(t)$.

In this paper, the second choice has been adopted to emphasize the analogy with continuous stochastic processes.

To initialize a trajectory at a point $\mathbf{x}$ that does not belong to the dataset, we search the K nearest neighbors of $\mathbf{x}$ among the available data and we

select as initial condition one of these points with a probability $\frac{1}{K}$. This corresponds to the association of a probability vector $p(\mathbf{x}) = p_i(\mathbf{x})$ to the point $\mathbf{x}$ defined as

$$\begin{cases} p_i(\mathbf{x}) = \frac{1}{K} \text{ if } X_i \text{ is an analogue of } \mathbf{x}\,, \\ p_i(\mathbf{x}) = 0 \text{ otherwise}\,. \end{cases} \tag{4.16}$$

Note that, for simplicity, in equations (4.14) and (4.16) we have assumed that each of the K analogues are chosen with uniform probabilities. We could generalize this choice using analogue dependent weights, for instance computed according to the distances of X_n to its analogues.

4.3.2 Computing the committor function from the analogue Markov chain

Using the analogue Markov chain defined in the previous section, we can compute the committor function q for this Markov chain. A first approach would be to generate trajectories of this Markov chain, and to directly sample the committor function through a Monte Carlo estimation as described in Sec. 4.2. However, we propose a more efficient computation which consists in solving a linear equation that characterizes the committor function of a Markov chain. Solving this linear equation is more precise than the direct approach, as we obtain the exact committor function up to numerical accuracy, without sampling errors. This linear equation will be solved by estimating the leading eigenmodes of a spectral problem, following the algorithm proposed in Ref. [Prinz et al., 2011]. Our paper is the first application of this idea to the analogue Markov chain.

We start from the Markov chain transition matrix G. We consider two sets $\mathcal{A} \subset \mathcal{X}$ and $\mathcal{B} \subset \mathcal{X}$, and we will compute the committor function q which is the probability to reach $\mathcal{B}$ before $\mathcal{A}$. For simplicity, we group together all the states that belong to $\mathcal{A}$ (resp. $\mathcal{B}$) into a single state with index $i_\mathcal{A}$ (resp $i_\mathcal{B}$). We then define an auxiliary process where $\mathcal{A}$ and $\mathcal{B}$ are absorbing states: no transition out of these states is allowed. The corresponding modified transition matrix is $\tilde{G}$, with $\tilde{G}_{i_\mathcal{A} i_\mathcal{A}} = 1$ and for all $j \neq i_\mathcal{A}$, $\tilde{G}_{i_\mathcal{A} j} = 0$, $\tilde{G}_{i_\mathcal{B} i_\mathcal{B}} = 1$ and for all $j \neq i_\mathcal{B}$, $\tilde{G}_{i_\mathcal{B} j} = 0$, while for $i \neq i_\mathcal{A}$, $i \neq i_\mathcal{B}$, $\tilde{G}_{i i_\mathcal{A}} = \sum_{k:X_k \in \mathcal{A}} G_{ik}$ and $\tilde{G}_{i i_\mathcal{B}} = \sum_{k:X_k \in \mathcal{B}} G_{ik}$, and for all other transitions $\tilde{G}_{ij} = G_{ij}$.

For the Markov chain $\tilde{G}$, the committor function is a column vector $q = q_i$ where q_i is the value at the committor function at the state i. q_i is an

approximation of the the committor function of the initial dynamics at point X_i: $q(X_i)$.

For simplicity, we use the same notation for the vector q (associated to the Markov chain) and the function q (associated to the initial dynamics), although they are actually different. In the limit of a large dataset, when the Markov chain fits perfectly the real dynamics, we have asymptotically $q_i \to q(X_i)$.

From the definition $q_i = \mathbb{P}(T_{\mathcal{B}}(i) < T_{\mathcal{A}}(i))$, we have $q_{i_{\mathcal{A}}} = 0$ and $q_{i_{\mathcal{B}}} = 1$. Moreover it is a classical result that $\tilde{G}q = q$ [Schütte et al., 1999, Prinz et al., 2011, Noé and Rosta, 2019, Tantet et al., 2015]. This is a simple consequence of the estimation of q at two successive steps of the Markov chain. The affine problem

$$\tilde{G}q = q \qquad \text{with} \quad q_{i_{\mathcal{A}}} = 0 \qquad \text{and} \quad q_{i_{\mathcal{B}}} = 1 \tag{4.17}$$

then characterizes the committor function, if we assume that G is ergodic.

Following Ref. [Prinz et al., 2011], we note that 1 is the largest eigenvalue of $\tilde{G}$ (a consequence of the Perron-Frobenius theorem for positive operators that preserve probability). Moreover $\tilde{G}^\dagger$ has two trivial eigenstates with eigenvalue 1, corresponding to situations where the full probability vector is concentrated on state $i_{\mathcal{A}}$ or $i_{\mathcal{B}}$, respectively. As a consequence, $\tilde{G}$ has also two eigenstates with eigenvalue 1. If we assume G is ergodic, then the number of eigenstates of $\tilde{G}$ is exactly 2.

This gives a simple algorithm to compute q. We first compute v_1 and v_2 the two leading eigenvectors of $\tilde{G}$ with any standard algorithm. Then q is a linear combination of v_1 and v_2: $q = \alpha v_1 + \beta v_2$, where α and β can be computed from the two conditions $q_{i_{\mathcal{A}}} = 0$ and $q_{i_{\mathcal{B}}} = 1$.

If the initial dynamics is indeed ergodic, we expect that for large enough dataset the Markov chain G will also be ergodic for most of the realizations. However, this might not be the case for some realizations. Such situations could lead to an incorrect computation of q as the solution of equation (4.17) is then not unique. In practice we check *a posteriori* (after running the algorithm) whether $q_i \in [0, 1]$ for all i, which is a necessary condition for q_i to be a probability. Sometimes, for some realizations of the sampling of the analogue Markov chain, rarely and even more rarely for large datasets, q takes values outside the interval $[0, 1]$. We interpret these cases as a sign of breaking of ergodicity. We then exclude these rare realizations, with possible ergodicity breaking of the Markov chain, from the results.

4.3.3 Applications

In this section, we estimate the committor function using the analogue method for two different models: Sec. 4.3.3 deals with a system of dimension 2 while Sec. 4.3.3 concerns a model with 6 degrees of freedom. For each system, we compare the estimated committor to the true committor, and we analyze the behavior of the error as the quantity of data upon which the analogue Markov chain relies varies. Finally, we compare the results of the analogue method with those obtained by the direct method, based on the same amount of data.

Model with two degrees of freedom

Let us consider a non-trivial 2-dimensional dynamics [Bréhier et al., 2016a]. The model is defined by the following stochastic differential equation:

$$\dot{\mathbf{x}} = -\nabla V(\mathbf{x}) + \sqrt{2\epsilon}\boldsymbol{\Xi}(t), \tag{4.18}$$

where $\mathbf{x} = (x, y)$, $\boldsymbol{\Xi} = (\xi_x, \xi_y)$ is a two dimensional gaussian white noise with $\langle \xi_i \rangle = 0$, $\langle \xi_i(t)\xi_j(t') \rangle = \delta_{ij}\delta(t - t')$, and the potential $V(\mathbf{x})$ is

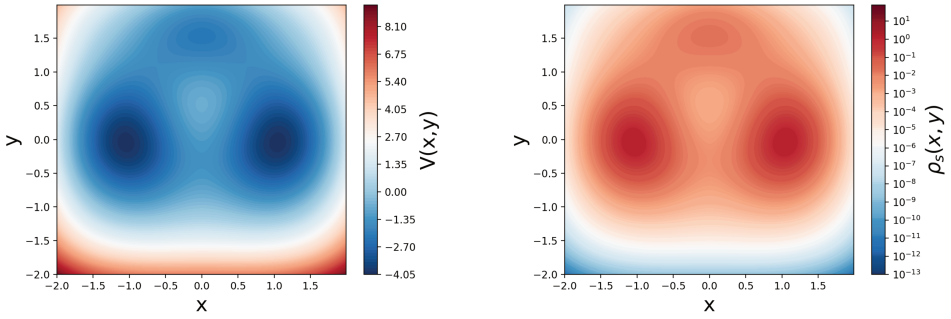
$$V(x,y) = 0.2x^4 + 0.2\left(y - \frac{1}{3}\right)^4 + 3e^{-x^2}\left(e^{-\left(y-\frac{1}{3}\right)^2} - e^{-\left(y-\frac{5}{3}\right)^2}\right) - 5e^{-y^2}\left(e^{(x+1)^2} + e^{(x-1)^2}\right). \tag{4.19}$$

The stationary distribution of the system is

$$\rho_s(\mathbf{x}) = Z^{-1} e^{-\frac{V(\mathbf{x})}{\epsilon}}, \tag{4.20}$$

where $Z = \int \mathrm{d}\mathbf{x}\, e^{-\frac{V(\mathbf{x})}{\epsilon}}$.

Figure 4.4 shows both the potential $V(\mathbf{x})$ (4.4a) and the stationary distribution $\rho_s(\mathbf{x})$ for $\epsilon = 0.5$ (4.4b). As can be seen in Fig. 4.4a, $V(\mathbf{x})$ has two global minima close to the points $\mathbf{x}_1 = (-1, 0)$ and $\mathbf{x}_2 = (1, 0)$, one local minimum close to the point $\mathbf{x}_m = (0, 1.5)$ and a saddle point close to $\mathbf{x}_s = (0, -0.5)$ — there are also two saddle points separating the global minima from the local minimum, approximately located at $(-0.6, 1.0)$ and $(0.6, 1.0)$. By comparing the panels 4.4a and 4.4b, it can be noted that small values of the invariant distribution correspond to large values of the potential and vice versa. In particular, Fig. 4.4b shows that $\rho_s(\mathbf{x})$ has global or local maxima at $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_m$.



(a) *Potential* $V(\mathbf{x})$. (b) *Stationary distribution* $\rho_s(\mathbf{x})$.

Figure 4.4: Color maps of the potential $V(\mathbf{x})$ (panel (a), left), defined by (4.19), and of the stationary distribution $\rho_s(\mathbf{x})$ (panel (b), right), defined by (4.20), for $\epsilon = 0.5$.

Let us consider the two sets $\mathcal{A} = \{\mathbf{x} : \sqrt{(\mathbf{x} - \mathbf{x}_1)^2} < 0.05\}$ and $\mathcal{B} = \{\mathbf{x} : \sqrt{(\mathbf{x} - \mathbf{x}_2)^2} < 0.05\}$. Note that these sets are defined to include the two maxima of the invariant distribution, where the dynamics spends most of the time. For $\epsilon = 0.5$, the relaxation time τ_r inside $\mathcal{A}$ or $\mathcal{B}$ is of order $O(1)$, while the average waiting time T_e to observe a transition between these two sets is of order $O(10^2)$.

We will now compute the committor function $q(\mathbf{x}) = \mathbb{P}[T_{\mathcal{B}}(\mathbf{x}) < T_{\mathcal{A}}(\mathbf{x})]$ for this system. First, we compute a reference committor function in the region $[-1, 1] \times [-1, 2]$ by direct sampling (as explained in Sec. 4.2.2), using a large amount of data: for each point on a 250×250 grid in this region, we sample 10 000 trajectories until they reach $\mathcal{A}$ or $\mathcal{B}$ and compute the value of the committor at that point using Eq. (4.5). This reference committor function is shown in Fig. 4.5. One can note that in a region around the set $\mathcal{A}$ the committor function is close to 0, while in the proximity of $\mathcal{B}$ it is mostly equal to 1; for $y \simeq -0.5$ and moving along the x direction $q(\mathbf{x})$ changes abruptly through the saddle point $\mathbf{x}_s$. On the contrary, around the relative minimum point $\mathbf{x}_m$ the committor function is mostly constant, with a value around 0.5, which corresponds to the probability to reach either of the two minima starting from this point.

Now we estimate the committor function using the method presented in Sec. 4.3.2. To do so, we need to generate some learning dataset and to choose

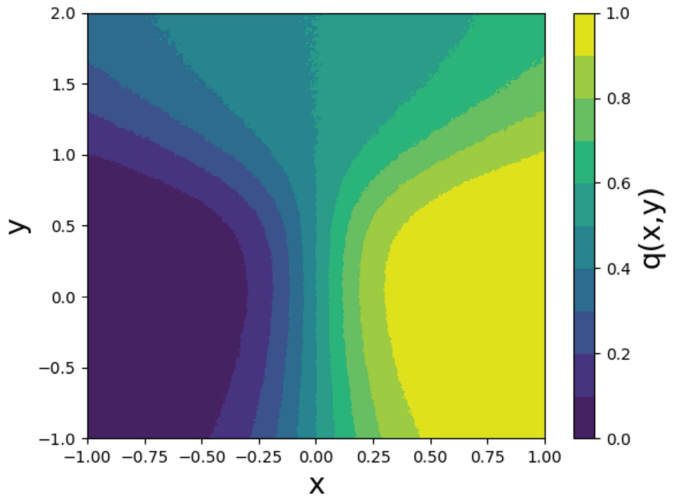
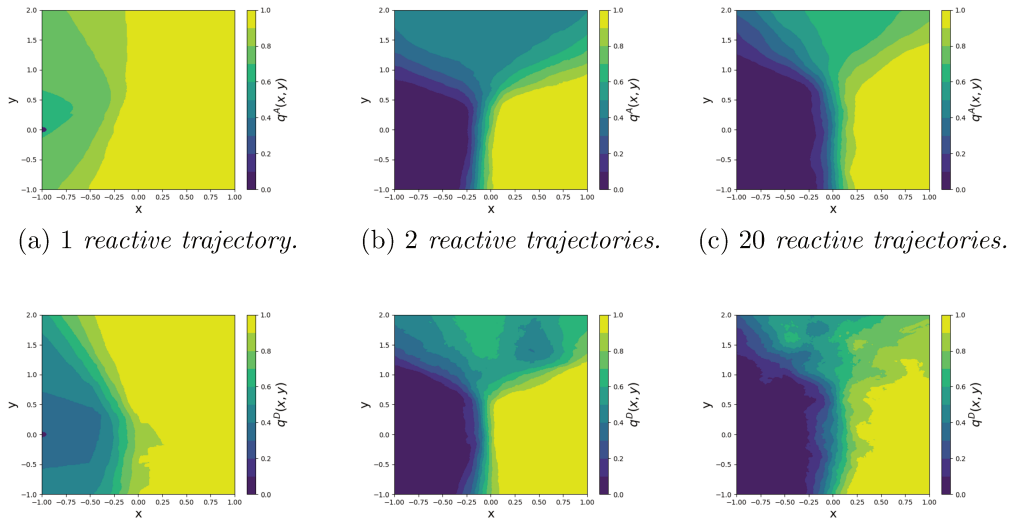


Figure 4.5: Committor function $q(\mathbf{x})$ for the 2D gradient system (4.18), computed using the real dynamics. The region $x \in [-1, 1]$, $y \in [-1, 2]$ is divided into $N_c = L \times L$ cells ($L = 250$) and, for each cell, $N = 10000$ Monte Carlo experiments are performed.

a distance and a number of nearest neighbors K. Because we want to measure the quality of our estimator $\hat{q}(\mathbf{x})$, by comparing it to the reference committor $q(\mathbf{x})$, as the quantity of available data varies, we generate three trajectories (using the real dynamics) of different length. Rather than fixing the length of the trajectory, we integrate each trajectory until a fixed number of transitions (1, 2 and 20) between sets $\mathcal{A}$ and $\mathcal{B}$ are observed. We then construct three analogue Markov chains using each of these trajectories as learning dataset and compute the corresponding committor function. For these computations, we have used the Euclidian distance and $K = 150$ analogues. The estimate of the committor function for the three choices of learning dataset are shown in Figs. 4.6a,4.6b,4.6c. Note that the method presented in Sec. 4.3.2 yields an estimate of the committor function only at the points included in the learning dataset. To represent the contour levels in Figs. 4.6a, 4.6b and 4.6c, we extend our estimate of the committor function to the whole region of interest by using a *k-nearest neighbor regression* method, as explained in Sec. 4.2.3. To avoid introducing additional parameters, we choose uniform weights $w_j = 1$ for all the nearest neighbor and we use the same number of neighbors as for constructing the analogue Markov chain $k = K = 150$.

In addition to the reference committor, we also want to compare the committor estimator based on the analogue method to a direct sampling estimate with the same amount of data. To do so, we also compute the committor function using Eq. (4.6) for the same three trajectories as above. In practice, because the exact same points are never visited twice, this amounts to assigning value 1 to a point in the trajectory if set $\mathcal{B}$ is visited before $\mathcal{A}$ in the rest of the trajectory, and value 0 otherwise. Again, this provides an estimate of the committor function only at points included in the learning dataset and we extend it to the region of interest with the same k-nearest neighbor method as above. This alternative estimator for the committor function, which we refer to as the *direct method*, is shown in Figs. 4.6d, 4.6e, and 4.6f.

Several conclusions can be drawn by comparing qualitatively the committor estimates shown in Fig. 4.6 with the reference committor shown in Fig. 4.5. First of all, note that a single reactive trajectory does not contain enough information to capture the structure of the committor function (Figs. 4.6a,4.6d). The committor estimates start to be qualitatively acceptable when two reactive trajectories are used (Figs. 4.6b,4.6e). This is due to the fact that our data set includes the two types of transition paths between $\mathcal{A}$ and $\mathcal{B}$ (the one that passes through the saddle point $\mathbf{x}_s$ and the one that



(a) 1 *reactive trajectory.* (b) 2 *reactive trajectories.* (c) 20 *reactive trajectories.*

(d) 1 *reactive trajectory.* (e) 2 *reactive trajectories.* (f) 20 *reactive trajectories.*

Figure 4.6: Estimates of the committor function $q(\mathbf{x})$ of the 2D gradient system (4.18) using the analogue method (top row, panels a–c) and the direct method (bottom row, panels d–f), for learning datasetsets of different length.

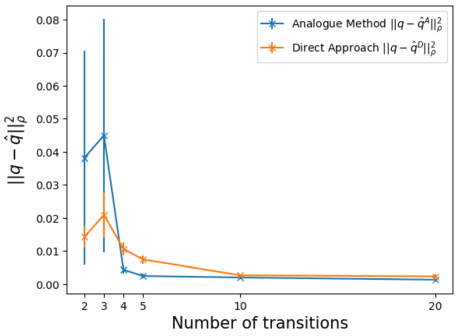


Figure 4.7: Error for the analogue and direct estimators of the committor function of the 2D gradient system (4.18) as function of the number of reactive trajectories in the learning dataset.

goes through the relative minimum $\mathbf{x}_m$). A comparison between Fig. 4.6b and Fig. 4.6e shows that the analogue method gives smoother results than the direct approach. However, note that both methods have a sharper transition region than that shown in Fig. 4.5. By increasing the number of reactive trajectories, a wider transition region is obtained (see Figs. 4.6c,4.6f) and the results appear more similar to the reference committor. Again, note that the result of Fig. 4.6c is smoother than that of Fig. 4.6f.

To quantify the error made in approximating $q(\mathbf{x})$ we consider the quantity

$$\|q-\hat{q}\|^2_{\rho_s} = \int \mathrm{d}\mathbf{x}[q(\mathbf{x})-\hat{q}(\mathbf{x})]^2\rho_s(\mathbf{x}) \approx \frac{1}{N_p}\sum_{i=1}^{N_p}(q(\mathbf{x}_i)-\hat{q}(\mathbf{x}_i))^2, \qquad (4.21)$$

where q is the true committor and $\hat{q}$ the estimate. The justification of using (4.21) as an error measurement has already been given in Sec. 4.2.4: $\|q-\hat{q}\|^2_{\rho_s}$ corresponds to the non-constant term in the Brier score (Eq. (4.11)). The errors computed from (4.21) for the two estimators of the committor function (the analogue method and the direct method) are represented in Fig. 4.7 as a function of the number of transitions in the . Each point in Fig. 4.7 corresponds to the average error computed over 10 independent realizations containing the same number of transitions while the error bar corresponds to the standard deviation. It can be noted that, for small datasets (2 or 3 reactive trajectories), the performances of the two methods are comparable within statistical errors but the direct approach seems to provide more

stable results. A simple interpretation is that when there is not enough data, the analogue Markov chain is not a good enough approximation of the real dynamics to provide any benefit to estimate the committor function. However, it becomes the case as the amount of data increases, and the analogue method outperforms the direct method as soon as the learning dataset contains at least 4 transitions. When the data contains at least 4 transitions, the error with the analogue method is two to three times smaller than the error with the direct method.

The Charney-DeVore model

We now apply the analogue method to compute a committor function for a more complex dynamics, the Charney–DeVore model [Charney and DeVore, 1979]. It is a simple toy model of atmospheric dynamics in the Northern Atlantic region, represented as a 2D channel with differential rotation. It is not intended to be realistic. Actually, the kind of multistability observed in this model is not observed in real atmosphere dynamics. The interest of this model is more methodological.

This model was introduced with the aim of proving that the combination of topography and barotropic instabilities can lead to different atmospheric flow regimes. The model is obtained by expanding the quasi-geostrophic stream function $\psi(z, y, t)$ (z corresponds to the longitude and y to the latitude) on the basis $\{\phi_{nm}(z, y)\}$ with

$$\begin{aligned} \phi_{0m} &= \sqrt{2}\cos\left(\frac{my}{b}\right), \\ \phi_{nm} &= \sqrt{2}\exp\left(inz\right)\sin\left(\frac{my}{b}\right), \end{aligned} \tag{4.22}$$

and truncating the series to retain only the first six terms. After the following change of variables [De Swart, 1989],

$$x_1 = \frac{1}{b}\psi_{01}, \qquad x_4 = \frac{1}{b}\psi_{02}, \tag{4.23}$$

$$x_2 = \frac{1}{\sqrt{2}b}(\psi_{11} + \psi_{-11}), \qquad x_5 = \frac{1}{\sqrt{2}b}(\psi_{12} + \psi_{-12}), \tag{4.24}$$

$$x_3 = \frac{i}{\sqrt{2}b}(\psi_{11} - \psi_{-11}), \qquad x_6 = \frac{i}{\sqrt{2}b}(\psi_{12} - \psi_{-12}), \tag{4.25}$$

the truncated equations of motion become

$$\begin{aligned}
\dot{x_1} &= \tilde{\gamma}_1 x_3 - C(x_1 - x_1^\star) + \sqrt{2\epsilon}\xi_1 \,, \\
\dot{x_2} &= -(\alpha_1 x_1 - \beta_1)x_3 - Cx_2 - \delta_1 x_4 x_6 + \sqrt{2\epsilon}\xi_2 \,, \\
\dot{x_3} &= (\alpha_1 x_1 - \beta_1)x_2 - \gamma_1 x_1 - Cx_3 + \delta_1 x_4 x_5 + \sqrt{2\epsilon}\xi_3 \,, \\
\dot{x_4} &= \tilde{\gamma}_2 x_6 - C(x_4 - x_4^\star) + \eta(x_2 x_6 - x_3 x_5) + \sqrt{2\epsilon}\xi_4 \,, \\
\dot{x_5} &= -(\alpha_2 x_1 - \beta_2)x_6 - Cx_5 - \delta_2 x_3 x_4 + \sqrt{2\epsilon}\xi_5 \,, \\
\dot{x_6} &= (\alpha_2 x_1 - \beta_2)x_5 - \gamma_2 x_4 - Cx_6 + \delta_2 x_2 x_4 + \sqrt{2\epsilon}\xi_6 \,,
\end{aligned} \tag{4.26}$$

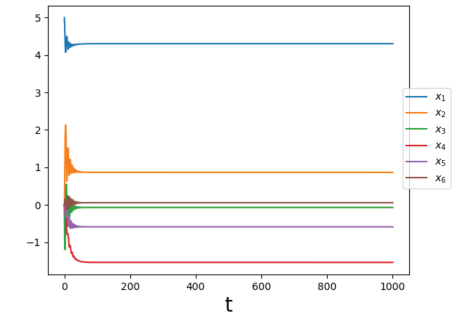
where a Gaussian white noise $\boldsymbol{\xi}(t)$ has been added with an arbitrary amplitude controlled by the parameter ϵ. All the components of the noise are independent and delta-correlated in time: $\langle \xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$. The parameters in (4.26) are defined as follows

$$\alpha_m = \frac{8\sqrt{2}}{\pi}\frac{m^2}{4m^2-1}\frac{b^2+m^2-1}{b^2+m^2}, \qquad \tilde{\gamma}_m = \gamma\frac{4m}{4m^2-1}\frac{\sqrt{2}b}{\pi}, \tag{4.27}$$

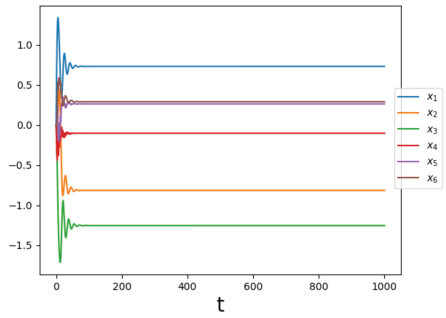
$$\beta_m = \frac{\beta b^2}{b^2+m^2}, \qquad \eta = \frac{16\sqrt{2}}{5\pi}, \tag{4.28}$$

$$\delta_m = \frac{64\sqrt{2}}{15\pi}\frac{b^2-m^2+1}{b^2+m^2}, \qquad \gamma_m = \gamma\frac{4m^3}{4m^2-1}\frac{\sqrt{2}b}{\pi(b^2+m^2)}. \tag{4.29}$$

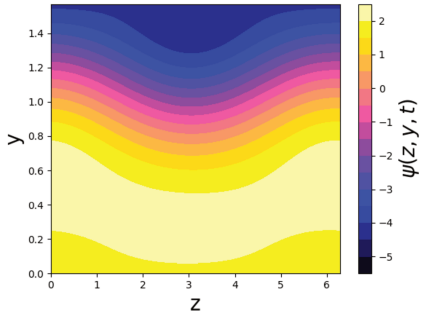
There are 7 free parameters in this model: $b, \gamma, \beta, C, x_1^\star, x_4^\star$, and the noise amplitude ϵ. For $\epsilon = 0$, the main feature of the system is the coexistence of multiple equilibrium states, in particular the existence of blocked flow and zonal flow regimes. The number and stability of these equilibrium states depend on the choice of the system parameters [De Swart, 1989, Crommelin et al., 2004]. We adopt the same choice made by T. Grafke *et al.* [Grafke et al., 2017, Grafke and Vanden-Eijnden, 2019], that is $\{b, \gamma, \beta, C, x_1^\star, x_4^\star\} = \{0.5, 1, 1.25, 0.1, 4.5, -1.8\}$. Crommelin *et al.* show that for these parameter values the system has two stable equilibrium points [Crommelin et al., 2004]: one corresponding to a zonal regime and the other to a blocked one. Figure 4.8 shows the convergence of the system towards the two equilibrium states as well as the corresponding stream function ψ for the deterministic model ($\epsilon = 0$). The panels 4.8a and 4.8b show clearly that, for this choice of parameters, the system exhibits multistability, and that the time it takes to reach the stationary regimes is of order $O(10)$. The two equilibria correspond to a *zonal* state, with almost horizontal streamlines



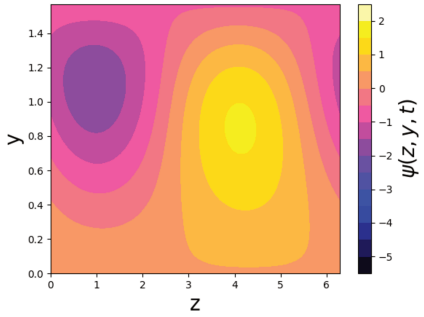
(a) *Convergence to the zonal regime.*



(b) *Convergence to the blocked regime.*



(c) *Stream function $\psi(z, y, t)$ in the zonal regime.*



(d) *Stream function $\psi(z, y, t)$ in the blocked regime.*

Figure 4.8: Time evolution of the six variables of the Charney-DeVore model for different initial conditions ((a) and (b)) showing the relaxation to two different states and the corresponding stream functions ((c) and (d)).

(Fig. 4.8c) and a *blocked* state, with strong cyclonic and anticyclonic structures (Fig. 4.8d). In the zonal regime the flow is characterized by a strong eastward jet $u_z \propto \partial_y \psi(z, y, t)$. Instead, in the blocked state there is no jet, the flow meanders strongly across the domain and it is characterized by the presence of vorticity.

For $\epsilon \neq 0$, the system can switch spontaneously from one regime to the other, under the influence of noise. To study the noise-induced transitions between the zonal and blocked states, we have to define the corresponding regions of the phase space. Let $\mathbf{x}_{eq}^Z$ and $\mathbf{x}_{eq}^B$ be the equilibrium points corresponding to zonal and blocked flow, respectively. Given two radii $r_B, r_Z > 0$, we define the sets

$$\begin{aligned} \mathcal{A} &= \{\mathbf{x} : \| \mathbf{x} - \mathbf{x}_{eq}^Z \| < r_Z\}, \\ \mathcal{B} &= \{\mathbf{x} : \| \mathbf{x} - \mathbf{x}_{eq}^B \| < r_B\}. \end{aligned} \tag{4.30}$$

In the rest of this section, we consider $r_Z = 0.8$, $r_B = 0.3$ and $\epsilon = 0.02$. For such parameters, the average time between two transitions is of order $O(10^3)$.

Let us now discuss the committor function $q(\mathbf{x}) = \mathbb{P}(T_{\mathcal{B}}(\mathbf{x}) < T_{\mathcal{A}}(\mathbf{x}))$ of the system. First of all, it should be noted that a direct computation of $q(\mathbf{x})$ in the whole phase space is not feasible. Indeed, such a calculation would require discretizing the six-dimensional phase space and to simulate a set of N trajectories for each point of the domain until they reach either $\mathcal{A}$ or $\mathcal{B}$. If 100 points along each direction were to be taken, then $N \times 10^{12}$ trajectories would have to be simulated. Considering a time of one millisecond to simulate N trajectories, the computation of $q(\mathbf{x})$ would still take $T_q = 10^9$ s ≈ 11574 days. Therefore, the reference committor $q(\mathbf{x})$ is computed on a limited number of points N_p distributed according to the invariant measure. Since the invariant distribution of the system is not known, the points N_p are sampled at regular time intervals over a very long trajectory. To be more specific, we consider a trajectory 10^7 time units long and we sample the N_p points at intervals $\delta t = 10^3$ time units. In this way, we ensure the statistical independence of the points and furthermore, by construction, their distribution will coincide with the invariant distribution of the system in the limit $N_p \to +\infty$. Then, the commitor function on those points can be computed by running N Montecarlo experiments for each of them.

After computing $q(\mathbf{x})$ along a trajectory in the six-dimensional space, it

is natural to ask how to represent such a function in a low-dimensional space. We will show below the empirical distribution of the value of the committor conditioned on the coordinate x_1, defined as:

$$\zeta(q|x_1) = \frac{\int \mathrm{d}\mathbf{y}\rho_s(\mathbf{y})\delta(q(\mathbf{y}) - q)\delta(y_1 - x_1)}{\int \mathrm{d}\mathbf{y}\rho_s(\mathbf{y})\delta(y_1 - x_1)}. \tag{4.31}$$

We chose to condition on the coordinate x_1 because the separation of the two attractors is larger in this direction than in the others. The distribution

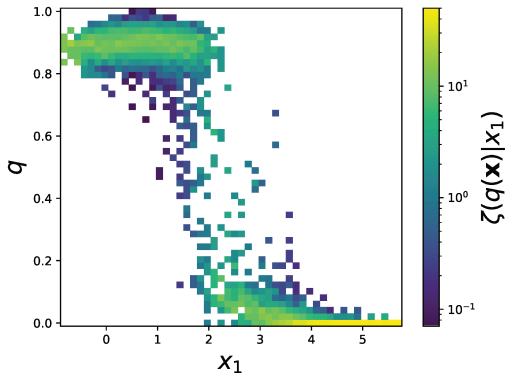


Figure 4.9: Color map of the conditional distribution $\zeta(q|x_1)$ for the reference committor of the Charney-DeVore model. The reference committor is computed using the real dynamics on $N_p = 10000$ points of the space space distributed according to the stationary distribution ρ_s. For each point, $N = 100$ Monte Carlo experiments are performed.

$\zeta(q|x_1)$ for the reference committor is represented on Fig. 4.9. We recognize some typical characteristics of a committor function in one dimension: in the neighborhood of the sets $\mathcal{A}$ (x_1 close to 4.3) and $\mathcal{B}$ (x_1 close to 0.7), $\zeta(q|x_1)$ is concentrated around 0 or 1 respectively, while in the proximity of the separatrix, although it is very dispersed, its average rapidly changes from values close to 0 to values close to 1, behaving roughly like a sigmoid function. The fact that the committor function exhibits some spread around 0 and 1 close to sets $\mathcal{A}$ and $\mathcal{B}$ can be explained by observing that many of the points are in fact located outside of the hyperballs defining these sets (4.30), although they lie in the basins of attraction of the zonal and blocked states.

We now estimate the committor function for the Charney-DeVore model using the analogue method. As in Sec. 4.3.3, we use several datasets of

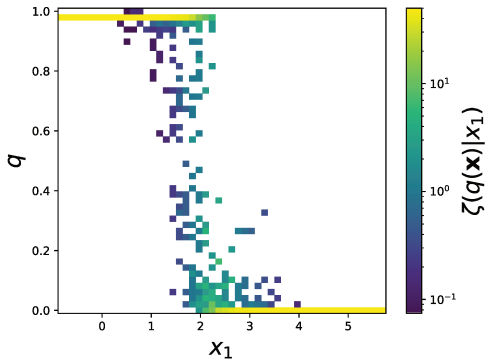
different size to build the analogue Markov chain used to estimate the committor. The size of these datasets is measured by the number $n = 2$, 5, 10, 15 of transitions between $\mathcal{A}$ and $\mathcal{B}$. As previously, we select $K = 150$ analogues using the Euclidean distance. We represent the conditional distributions $\zeta_A(q|x_1)$ for the estimates of the committor using 2 and 15 reactive trajectories in Fig. 4.10. By comparing Fig. 4.9 and Fig. 4.10, it can be noted that the conditional distributions $\zeta_A(q|x_1)$ provided by the analogue method have the same qualitative structure as the conditional distribution of the reference committor, with values concentrated close to 0 and 1 in the vicinity of sets $\mathcal{A}$ and $\mathcal{B}$, and a sharp transition region in between. However, the distributions are much more concentrated around the two set $\mathcal{A}$ and $\mathcal{B}$ than the reference one. This is probably because the phase space has not been explored sufficiently and therefore the analogues of points lying outside the hyperballs defining the sets are instead inside $\mathcal{A}$ and $\mathcal{B}$. Similarly, the transition region is narrower. The estimates obtained with the two datasets of different lengths are qualitatively very similar (see Figs. 4.10a and 4.10b), even if the distribution using 15 reactive trajectories (Fig. 4.10b) exhibits slightly more spread close to attractor $\mathcal{B}$ and a seemingly broader transition region.

We now compare the performances of the two data-based methods (the analogue method and the direct estimator) as the amount of data varies using the same procedure as in Sec. 4.3.3. The error associated to an estimate of the committor is given by the non-constant term of the Brier score (Eq. (4.11)), i.e.

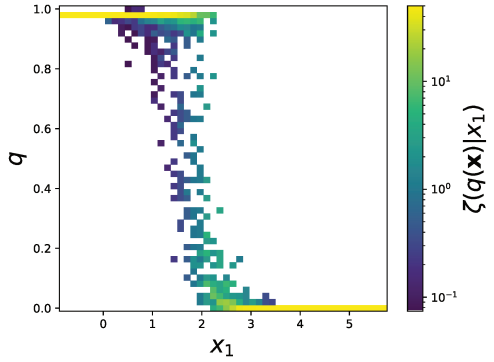
$$\|q - \hat{q}\|^2_{\rho_s} = \int \mathrm{d}\mathbf{x}(q(\mathbf{x}) - \hat{q}(\mathbf{x}))^2 \rho_s(\mathbf{x}) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} (q(\mathbf{x}_i) - \hat{q}(\mathbf{x}_i))^2, \qquad (4.32)$$

where q is the true committor, $\hat{q}$ its approximation and ρ_s is the invariant measure. Note that here, we are directly comparing the committor functions q and not the distributions $\zeta(q|x_1)$.

For each dataset size, we repeat the computation 10 times using different realizations of the trajectory. The best estimate of the error is computed as the empirical average over those realizations and the error bar corresponds to the standard deviation computed over the different experiments. These results are shown, as a function of the size of the dataset upon which the analogue Markov chain is built, in Fig. 4.11. The estimates of the committor function provided by the analogue method are more precise than those ob-



(a) 2 *reactive trajectory.*



(b) 15 *reactive trajectories.*

Figure 4.10: Color maps of the conditional distribution $\zeta_A(q|x_1)$ for the committor function of the Charney-DeVore model estimated using the analogue method $\hat{q}^A(\mathbf{x})$ with learning dataset containing 2 (a) and 15 (b) reactive trajectories.

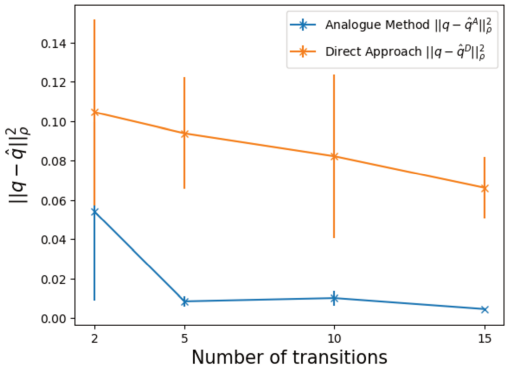


Figure 4.11: Error for the analogue and direct estimators of the committor function of the Charney-DeVore model as function of the number of reactive trajectories in the learning dataset.

tained by the direct approach, regardless of the length of the dataset. While the error associated to the direct approach decreases linearly with the number of transitions in the dataset, the error for the analogue method decreases faster for small datasets and reaches a plateau after about 5 reactive trajectories. In this latter regime, the error with the analogue method is roughly an order of magnitude smaller than with the direct approach. It is remarkable that although the system has a larger number of degrees of freedom than the 2D system studied in Sec. 4.3.3, the error associated to the analogue method is likewise small for datasets containing a relatively small number of trajectories (about 5). In other words, this suggests that the analogue method does not require larger datasets for estimating committor functions when the dimension of space increases. Here, we have only verified this statement for a moderate increase of the number of degrees of freedom, but we may hope that it remains true in higher dimension, as the relevant data should remain close to the transition paths where most of the information carried by the committor function is contained.

4.4 Using the learned committor function in Adaptive Multilevel Splitting

In Sec. 4.3, we estimated the committor function with the analogue method. We will now illustrate how this approximated committor can be used in a rare event simulation, using the Adaptive Multilevel Splitting (AMS) algorithm. This algorithm relies on a function used to select the trajectories leading to the rarest events, called the *score function*. The committor function is known to be the optimal score function, but it is generally not known exactly. We will show that using the estimated committor as a score function has two advantages. First, it provides a version of AMS where the user does not need to explicitly prescribe the score function. This is very useful in practice when we have little knowledge of the dynamics beyond the presence of the two attractors $\mathcal{A}$ and $\mathcal{B}$. In addition, it can improve the precision of the quantities computed with AMS, compared to user defined score functions. Indeed, it approximates the true committor, which leads to minimal errors on estimates. On other hand, user defined score functions, with analytical formulas, have no reason to be good approximations of the true committor in general.

4.4.1 The Adaptive Multilevel Splitting algorithm and the quality of score functions

Adaptive Multilevel Splitting is a *splitting* method designed to estimate the probability of rare events, inspired by the pioneering works of Kahn and Harris [Kahn and Harris, 1951] and Rosenbluth and Rosenbluth [Rosenbluth and Rosenbluth, 1955]. It has been proposed by Cérou & Guyader [Cérou and Guyader, 2007], as an improvement over Multilevel Splitting (see Ref. [Glasserman et al., 1998] for instance). Many variants have been developed since, and the algorithm has been applied in a variety of contexts [Rolland, 2018, Bouchet et al., 2019, Lopes and Lelièvre, 2019, Lestang et al., 2020]. The description of the algorithm given here follows the presentation of Lestang *et al.* [Lestang et al., 2018]. See the review article by Cérou, Guyader & Rousset [Cérou et al., 2019b] for a recent overview of the method and its applications.

For definiteness, we consider a continuous time Markov process X_t in the phase space $\mathcal{X}$. Let us define two regions $\mathcal{A}$ and $\mathcal{B}$ in phase space. We again

seek to estimate the probability $\alpha = \mathbb{P}[T_{\mathcal{B}} < T_{\mathcal{A}}]$, where $T_{\mathcal{D}} = \inf\{t > 0, X_t \in \mathcal{D} \text{ with } X_0 \in \mathcal{C}\}$ is the first hitting time of the set $\mathcal{D}$, starting from a set $\mathcal{C}$. The set $\mathcal{C}$ encloses the set $\mathcal{A}$. We also wish to compute the corresponding realizations of the dynamics.

The AMS algorithm computes these quantities iteratively. For this matter, the algorithm uses a *score function* ϕ, (sometimes termed *reaction coordinate*) a map from the phase space $\mathcal{X}$ to $\mathbb{R}$. Ideally, the score function is bounded from below by 0 and from above by 1, vanishes identically on $\mathcal{A}$ and is identically equal to 1 on $\mathcal{B}$. Our aim is to compare the efficiency of different score functions.

In order to run the algorithm, we first need to sample initial conditions according to the invariant measure restricted to the set $\mathcal{C}$. In practice, we sample these initial conditions on $\mathcal{C}$ by sampling long trajectories in the bassin of attraction of $\mathcal{A}$. Then the algorithm is initialized by sampling N independent trajectories, with initial conditions on the set $\mathcal{C}$ and run until they reach either the set $\mathcal{A}$ or the set $\mathcal{B}$. Let us denote by $\{\mathbf{x}_n^{(0)}(t)\}_{1\leq n\leq N}$ the initial ensemble of trajectories, where the subscript denotes the index of the trajectory in the ensemble and the superscript denotes the iteration of the algorithm. We associate a weight $w_0 = 1$ to those trajectories.

At each iteration $j \geq 1$, we apply the following *selection and mutation* steps, which are schematically illustrated in Fig. 4.12:

- We compute the score of each trajectory in the ensemble at iteration $j-1$: $\Phi_n^{(j)} = \sup_t \phi(t, \mathbf{x}_n^{(j-1)}(t))$.
- We determine the trajectories which have the lowest score: $\Phi_j^\star = \min_{1\leq n\leq N} \Phi_n^{(j)}$ and we set $n_{j,1}^\star, \ldots, n_{j,\ell_j}^\star$ the indices such that $\Phi_{n_{j,1}^\star}^{(j)} = \cdots = \Phi_{n_{j,\ell_j}^\star}^{(j)} = \Phi_j^\star$. One can have $\ell_j > 1$ in some iterations. If $\ell_j = N$ and not all the trajectories have reached $\mathcal{B}$, the algorithm stops: it leads to an *extinction*.
- We mutate each trajectory $\mathbf{x}_{n_{j,\ell}^\star}^{(j-1)}$ $(1 \leq \ell \leq \ell_j)$: for each of them, we choose a trajectory $\mathbf{x}_{n_\ell}^{(j-1)}$ $(n_\ell \neq n_{j,1}, \ldots n_{j,\ell_j})$ drawn randomly among the $N - \ell_j$ remaining trajectories. We determine the smallest time t such that $\phi(t, \mathbf{x}_{n_\ell}^{(j-1)}(t)) > \Phi_j^\star$, denoted by $t_{j,\ell}$. The new trajectory $\mathbf{x}_{n_{j,\ell}^\star}^{(j)}$ is set by copying the trajectory $\mathbf{x}_{n_\ell}^{(j-1)}$ from t_0 to $t_{j,\ell}$, and simulating

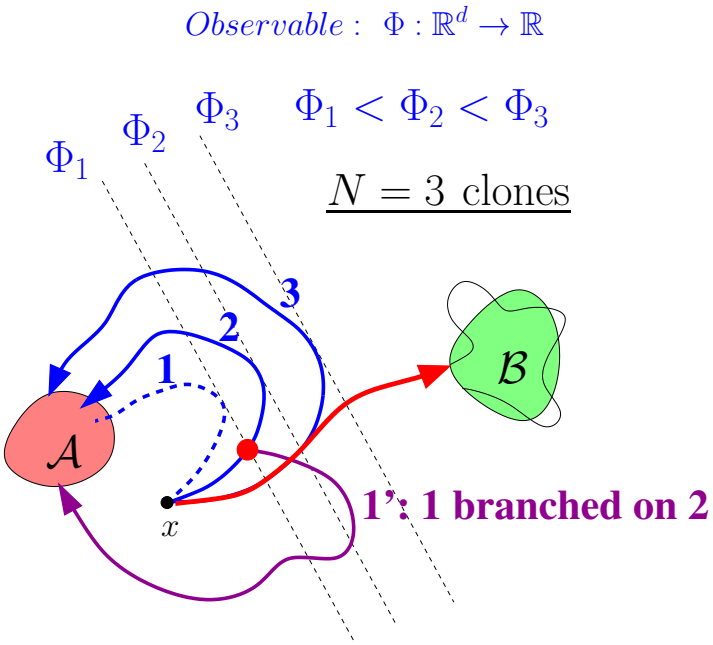


Figure 4.12: Sketch illustrating two iterations of AMS in a simplified example with 3 clones (Figure originally made for Ref. [Simonnet, 2016]), in order to compute trajectories going from set $\mathcal{A}$ to set $\mathcal{B}$. Trajectory 1 (dashed line) has the smallest excursion out of $\mathcal{A}$ as measured by the score function Φ. It is removed and branched on another trajectory (in that case trajectory 2, leading to the purple line). In the successive iteration, trajectory 2 has the smallest score function and is branched on trajectory 3 (leading to the red line).

the trajectory with a new independent realisation of the noise, starting from time $t_{j,\ell}$, until it hits either the set $\mathcal{A}$ or the set $\mathcal{B}$.

- Trajectories with higher scores are not modified at this step: $\mathbf{x}_n^{(j)} = \mathbf{x}_n^{(j-1)}$ for $n \neq n_{j,1}^\star, \dots, n_{j,\ell}^\star$.

- We compute the weight of iteration j: $w_j = \left(1 - \frac{\ell_j}{N}\right) w_{j-1}$.

The algorithm is iterated until all the trajectories reach the set $\mathcal{B}$. The number of iterations J is a random number. This leads to an estimator $\hat{\alpha}$ for the transition probability α:

$$\hat{\alpha} = w_J = \prod_{j=0}^{J} \left(1 - \frac{\ell_j}{N}\right) . \tag{4.33}$$

This estimator is a random variable, with one value obtained for each realization of the algorithm. We perform M independent realizations of the algorithm and compute the statistics of $\hat{\alpha}$: the empirical average and variance of $\hat{\alpha}$.

The mathematical properties of this estimator have been extensively studied [Cérou and Guyader, 2007, Guyader et al., 2011, Rolland and Simonnet, 2015, Bréhier, 2015, Bréhier et al., 2015, Bréhier et al., 2016a, Bréhier et al., 2016b, Simonnet, 2016]. The key property is that, for any N and score function ϕ, it is an unbiased estimator [Bréhier et al., 2016a, Bréhier et al., 2016b] with a finite variance. The variance $\sigma_\alpha(N)$ depends on N and on the score function. The optimal score function, with the lowest variance, is the committor function.

More precise results exist asymptotically for large N. It is then proven [Cérou et al., 2019a] that the variance scales like $\frac{1}{\sqrt{N}}$ asymptotically $\sigma_\alpha(N) \underset{N\to\infty}{\sim} \frac{G(\phi)}{\sqrt{N}}$. Moreover, when the score function is the committor function, G is minimal, and the variance scales like the ideal variance

$$\sigma_{\text{id}} = \frac{\alpha\sqrt{|\log(\alpha)|}}{\sqrt{N}} . \tag{4.34}$$

In many cases, an asymptotic scaling is observed in practice when the number of clones is larger than 100 (see for instance Ref. [Rolland, 2018], Fig. 14 (c)).

The computation of the empirical variance of α, given by $\sigma_\alpha(N,M) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\alpha}_m^2) - \left(\frac{1}{M}\sum_{m=1}^{M}\hat{\alpha}_m\right)^2}$, and its comparison to the ideal variance σ_{id} has often been used as an *a posteriori* test of the quality of the score function and how close it is to the committor [Rolland and Simonnet, 2015, Bréhier and Lelièvre, 2019, Rolland et al., 2016].

Although the estimator is actually unbiased ($\mathbb{E}[\hat{\alpha}] = \alpha$), in numerical uses of AMS, it is often observed that $\hat{\alpha}$ underestimates α in the large majority of the M realizations of the algorithm. These underestimates are such that the average $\langle\hat{\alpha}\rangle_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\alpha}_m$ over M realizations is most of the time strictly smaller than α although the average is α ($\mathbb{E}[\langle\hat{\alpha}\rangle_M] = \alpha$). This phenomenon is called an *apparent bias*. We note that a similar observation is made in the context of fixed Multilevel Splitting [Glasserman et al., 1998] and Importance Sampling [Devetsikiotis and Townsend, 1993]. In these contexts, it can be demonstrated that $\frac{1}{M}\sum_{m=1}^{M}\hat{\alpha}_m$ underestimates α with a probability that goes to 1 as parameters like ϵ, which control the rareness of the event, go to zero [Glasserman et al., 1998]. This happens if the score function yielding the levels of Multilevel Splitting is not adapted. As a consequence, the observed sample mean of $\hat{\alpha}$ will be strictly smaller than α unless an out of reach number of realizations of AMS is performed. It has been conjectured [Bréhier et al., 2016a] that the observed apparent bias phenomenon could be explained for the AMS by analogy with the studies for fixed multilevel splitting.

The apparent bias, measured through the difference $\alpha - \langle\hat{\alpha}\rangle_M$, decreases like $\frac{1}{N}$ as the number of clones N is increased. However,it has been observed that for some cases, the apparent bias seems to reach a plateau for extremely large values of N [Rolland and Simonnet, 2015]. We will see similar behavior in the following. In these situations, it is observed that this apparent bias is minimal when the score function is the committor function [Rolland and Simonnet, 2015]. As this apparent bias is a very important practical problem, we will use use the magnitude of this apparent bias as a measure of the quality of the score function.

We have seen that the committor function is the best score function for the AMS algorithm, and explained that the computation of the empirical variance and of the apparent bias are two ways to quantify the quality of a score function. We can also test the AMS computations by comparing the computation of other observables. For instance, we will compute the transition path duration, denoted τ. This physical quantity has proven to be a good indicator of whether AMS was correctly sampling transition paths [Rolland

and Simonnet, 2015].

To have an unbiased estimate of α and τ and validate the output of AMS computations, we perform a large number of Direct Numerical Simulations (DNS) of reactive trajectories. These DNS start like AMS computations with initial conditions on $\mathcal{C}$, we let them evolve until they reach either $\mathcal{A}$ or $\mathcal{B}$. The proportion of DNS that reach $\mathcal{B}$ before $\mathcal{A}$ yields a direct estimate of α. We also perform an estimate of τ by averaging the duration trajectories that reach $\mathcal{B}$ before $\mathcal{A}$.

The estimate of a quantity by AMS is deemed to be precise enough when the 95% confidence intervals of this estimate performed by AMS and by DNS overlap [Bréhier et al., 2016a]. These confidence intervals are constructed by noting that we look at the sum of independent random variables of finite variance. They therefore follow a central limit theorem and the sample mean of $\hat{\alpha}$ has a gaussian distribution. The confidence interval is then given by $\langle\alpha\rangle_M \pm 1.96\sigma_\alpha(N,M)$, with the empirical variance $\sigma_\alpha(N,M)$. Similar confidence intervals are constructed for α and τ for both AMS and DNS results.

4.4.2 The learned committor function

Our goal is to investigate the performance of a score function relying on a data-based estimate of the committor function, using the analogue method presented in Sec. 4.3. As mentioned above, this method only provides an estimate on the points initially present in the dataset. To extend the score function to the whole phase space, we proceed as explained in Sec. 4.2.3, with a nearest-neighbor method using an exponential kernel with width $\omega = 0.1$. Here, a small number of neighbors $\kappa = 10$ is used for efficient computations of the score function. Indeed, for each computation, a search through neighbors must be performed. For a given training dataset, this method defines a score function, which we shall denote ϕ_{dat}.

The use of a kernel is justified by the need to avoid regions of constant score function. Indeed, when $\mathbb{R}^D$ is divided in finite subvolumes, many points $\mathbf{y}$ have the same neighbours $\{\mathbf{x}_j\}_{1\leq j\leq\kappa}$ and thus would have the same score function if uniform weights were used. On the other hand, the kernel ensures a dependence on $\mathbf{y}$ even within such regions. In other words, the kernel ensures that sets of constant ϕ_{dat} are hypersurfaces and not hypervolumes and that much fewer $\mathbf{y}$ have the same values of Φ at each stage of the algorithm. Practice shows that this leads to more efficient branching by limiting the number of clones suppressed at each stage of the algorithm and the risk of

extinction.

We will test the use of the analogue based estimate of the committor as a score function for the AMS computations for the two systems presented in Sec. 4.3.3: the 2D three-well system (Sec. 4.4.3) and the Charney-DeVore model (Sec. 4.4.4). The test of the learned committor function will be twofold. First, we will consider a score function learned on a dataset displaying a large number of transitions and study the quality of the result as a function of the clone number N. This will allow us to discuss the phenomenon of apparent bias and how the learned committor function deals with it. The second aim will be to study the required size of the dataset to have good results with the AMS algorithm. This question is critical for complex systems for which data will be scarce because of computation costs. To address this question, we will then perform AMS computations with a fixed large number of clones $N = 1000$ and datasets of increasing size (measured in number of recorded transitions).

4.4.3 AMS study for the two dimensional three well model

In this subsection, we work on the dynamics of the two-dimensional three-well model presented in Sec. 4.3.3 (Eq. (4.18)). The sets $\mathcal{A}$ and $\mathcal{B}$ as well as the noise variance ϵ are defined as in Sec. 4.3.3.

Efficiency of the AMS algorithm with the learned committor function for large N for the three-well model

We first study the efficiency of the AMS algorithm when using the learned approximate committor funciton ϕ_{dat} as a score function, when the number of clones is increased with a fixed data set length.

The time series which is used to compute this score function has $N_p = 1.4 \cdot 10^5$ datapoints (effectively 1400 time units long) and displays 21 transitions. The results for the AMS algorithm with this score function will be compared to either DNS computations, or to AMS computations with two explicitly user defined score functions: $\phi_{\text{lin}}(\mathbf{x}) = \frac{x+1}{2}$ and $\phi_{\text{norm}}(\mathbf{x}) = \frac{\sqrt{(x+1)^2+\frac{1}{2}y^2}}{2}$. The performances of these score functions have been studied in detail in the literature [Rolland and Simonnet, 2015].

In Fig. 4.13 (a), we first show the transition probability $\langle \alpha \rangle_M$ as a function

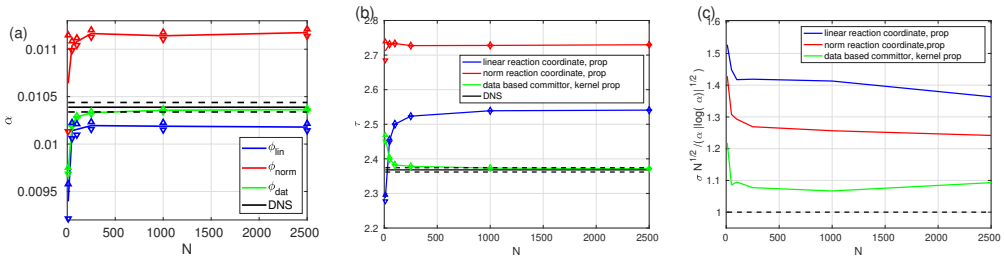


Figure 4.13: Efficiency of the AMS algorithm with the learned committor function for fixed and large dataset, for the three-well problem. Comparison of the estimated (a) Transition probability $\langle \alpha \rangle$, (b) Duration of reactive trajectories $\langle \tau \rangle$, and (c) Rescaled variance σ, as a function of the number of clones N. For each plot the black curve is the reference: either the DNS (a) and (b), or the optimal value 1 (c). The dashed black lines are the 95% confidence interval for the DNS. The color curves have been computed using the AMS, with respectively the learned committor function (green), the linear score function (blue) and the quadratic score function (red). The red and blue curves clearly illustrate the apparent bias phenomenon. The learned committor function gives excellent results, suppressing the apparent bias and giving smaller, close to optimal, empirical variance.

of the number of clones N used in AMS computations, using the three score functions ϕ_{dat}, ϕ_{lin} and ϕ_{norm}, and computed by means of DNS. Error bars show the 95% interval of confidence. One can first note that for the AMS computations, $\langle\alpha\rangle_M(N)$ is within 1% of its asymptotic value if the number of clones used is larger than $N = 100$. As noted in Sec 4.4.1 $\langle\alpha\rangle_M$ grows with N toward this asymptotic value. The confidence intervals of the probability, for AMS and DNS computations, do not overlap when we use the norm score function ϕ_{norm}: the asymptotic value of $\langle\alpha\rangle_M$ overestimates α. With the linear score function ϕ_{lin}, the asymptotic value of $\langle\alpha\rangle_M$ in turn underestimates α: this is a possible consequence of the apparent bias phenomenon [Bréhier et al., 2016a]. These results are in agreement with previous studies [Rolland and Simonnet, 2015], which have related these biases to errors in the relative sampling of transition paths. For instance, the linear score function selects preferentially trajectories going through the bottom channel (where paths remain around $y \simeq 0$ and where they cross the highest potential difference, see Fig. 4.4a and Ref. [Rolland and Simonnet, 2015], Fig. 7)a)), leading to the bias. By contrast, if the learned committor function ϕ_{dat} is used, the confidence intervals overlap as soon as $N \geq 250$, thus indicating that no bias can be detected in this estimate of α.

The results for τ, the average length of reactive trajectories, are qualitatively similar: Fig. 4.13 (b) shows $\langle\tau\rangle_M$ as a function of the number of clones used in AMS for the three score functions, compared to a reference DNS calculation. Error bars are again given by the 95% confidence interval. We first note that for the data-based and linear score functions $\langle\tau\rangle_M$ converges toward its asymptotic value to within 1% for $N \geq 250$; for the norm score function it is within that interval for all values of N. The 95% confidence interval of the AMS estimate with learned committor function and of DNS overlap if more than 1000 clones are used (a larger number than for the transition probability α; note however that the confidence intervals are narrower for τ than for α). On the other hand, both the linear and the norm score functions lead to overestimates of τ.

Finally, in Fig. 4.13 (c), we consider the rescaled variance $\sigma_\alpha(N, M)/\sigma_{\mathrm{id}}(N)$ of the estimator of α as a function of the clone number N. Here the absolute reference is the unit value, obtained for the optimal score function, the exact committor. We first note that for all score functions, the rescaled variance reaches a plateau if the number of clones is larger than $N = 100$. The value of this plateau is largest when we use the linear score function, with $\sigma_\alpha(N, M) = 1.4 \pm 0.02$. It is somewhat smaller for the norm score function,

with $\sigma_\alpha(N, M) = 1.25 \pm 0.02$. The best results are obtained for the learned committor function, with $\sigma_\alpha(N, M) = 1.12 \pm 0.2$. This again indicates that the computations performed using the learned committor function are the most precise, in that they come with the smallest statistical error, which is 10% larger than the smallest error possible.

All things considered, we conclude that if we use a large dataset to learn an estimate committor function with the analogue Markov chain, and use it as a score function for AMS, the estimates of transition properties show no apparent biases and converge to their true value when the number of clones is increased. The better precision of the results with the learned score function is also clearly visible for the lower statistical error measured by the empirical variance of the algorithm. This better precision is seen as soon as the number of clones N is of the order of 100 and the 95% confidence interval is reached for N of the order of a few hundreds.

Efficiency of the AMS algorithm with the learned committor function as a function of the dataset length for the three-well model

In Sec. 4.4.3, we used a large dataset with 21 transitions to accurately estimate the committor, before using it as a score function for the AMS. Compared to analytically defined score functions, this suppressed the apparent bias phenomenon and reduced the statistical error.

However, for many complex systems with very costly computations, it might not always be affordable to use a long dataset to learn the committor function. Moreover, in the initial stage of the study, one need to work with short datasets. Hence, we now study how the results of AMS computations using the learned committor function depend upon the size of the learning dataset, in the regime of short datasets, for the 2D three-well model.

For this matter, we sample trajectories of increasing length that contain an increasing number of transitions, from 1 to 21. For each number of transitions, we sample seven independent trajectories. For each of these datasets, we estimate the committor with the analogue method (Sec. 4.3) and use it as a score function in AMS computations with $N = 1000$ clones.

Fig. 4.14 (a,b) show the transition probability α and the average length of reactive trajectories τ as a function of the number of sampled transitions in the dataset. For each realization of the dataset which is used to learn the score function, we represent the best estimate and the 95% intervals of confidence, as different points. However, all the points are essentially

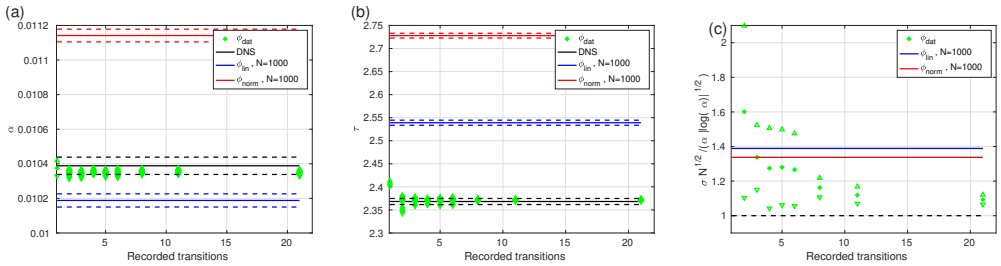


Figure 4.14: Efficiency of the AMS algorithm with the learned committor function, as a function of the dataset length, for the three-well problem. Comparison of the estimated (a) Transition probability $\langle\alpha\rangle$, (b) Duration of reactive trajectories $\langle\tau\rangle$, and (c) Rescaled variance σ, for each case averaged over independent realizations of the score function. For each plot the black curve is the reference one, either the DNS (a and b) or the optimal value 1 (c). The dashed black lines are the 95% confidence interval for the DNS. The color curves have been computed using the AMS, with respectively the learned committor function (green), the linear score function (blue) and the quadratic score function (red). The red and blue curves are constant values (they do not depend on the data set length) for comparison. The learned committor function gives much better results than the user defined score functions, even for very small datasets. With datasets containing only a few transitions, two to five, the results are already excellent. However, for such small datasets, the quality of the score function varies much from one realization to another.

superimposed: all the realizations of the score functions lead to the same results. We note that even if we use a short dataset to learn the committor function, the estimates are very precise: the intervals of confidence of the AMS algorithm and the DNS estimates overlap for all our datasets lengths, except for the shortest dataset (only one transition) for τ. In particular, they are always significantly more accurate than the estimates performed with the user defined score functions, as a consequence of the apparent bias phenomenon.

This is confirmed in Fig. 4.14 (c) by considering σ, the rescaled variance of the estimate of α. In this plot, for each dataset length, we have computed the empirical average and variance of the rescaled variances estimated with the different realizations of the score function. This first shows that the rescaled variance decreases as the number of transitions contained in the dataset increases, from 1.6 when the dataset contains only two transitions to almost 1.1 when the dataset contains 8 transitions or more. This indicates with datasets with 8 transitions or more, the statistical error is systematically reduced when using the learned score function is learned rather than user defined score functions.

We also note that the fluctuations of the variance between different dataset realizations decreases as the number of transitions contained in the dataset increases. If the dataset is short, no more than 6 transitions, one can obtain a score function that leads to better or worse results than analytically defined score functions with comparable probability. With a dataset with 3 transitions or more, the statistical error is most of time reduced when the score function is learned from datasets, compared to the case with user defined score functions.

Finally, we stress that for very short dataset, with only a few transitions, even if the variance on the estimate of α is of the same order for both user defined and learned score function, the systematic apparent bias is much smaller with the learned committor function.

4.4.4 Application to the Charney-DeVore model

We now perform the same tests for AMS computations using the learned committor function in the Charney-DeVore model (Eq. (4.26)). We use the same parameters and definition of the sets $\mathcal{A}$ and $\mathcal{B}$ as in Sec 4.3.3, describing transitions between zonal and blocked flows.

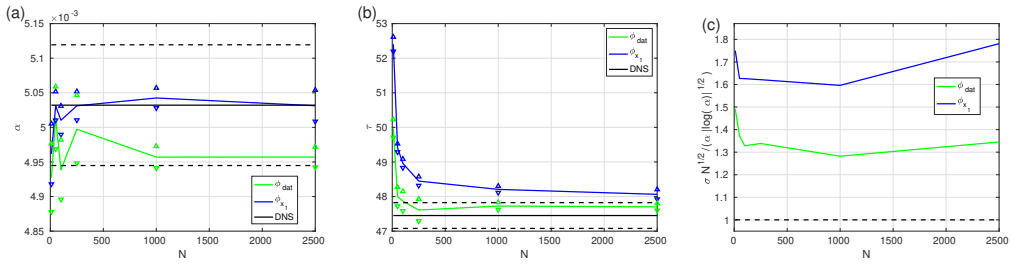


Figure 4.15: Efficiency of the AMS algorithm with the learned committor function for large dataset as a function of the clone number N, for the Charney-DeVore model. Comparison of the estimated (a) Transition probability $\langle\alpha\rangle$, (b) Duration of reactive trajectories $\langle\tau\rangle$, and (c) Rescaled variance σ. For each plot the black curve is the reference one, either the DNS (a and b) or the optimal value 1 (c). The dashed black lines are the 95% confidence interval for the DNS. The color curves have been computed using the AMS, with the learned committor function (green) and the linear score function (blue). The learned committor function gives excellent results, similar to the linear one for the weak apparent bias of the transition probability (a), and much better than the linear one for the variance and the length of reactive trajectories (b and c).

Efficiency of the AMS algorithm with the learned committor function for large N for the Charney-DeVore model

We proceed as in Sec. 4.4.3: we first learn the committor function ϕ_{dat} (Sec. 4.4.2) from a long trajectory, containing $3.4 \cdot 10^4$ data points and displaying 38 transitions. The perfomances will be compared to DNS results and to a simple linear score function $\phi_{x_1} = \frac{x_1 - x_{1,Z}}{x_{1,B} - x_{1,Z}}$ with $x_{1,Z} = 4.308$ and $x_{1,B} = 0.709$ (see Figs. 4.8a and 4.8b).

We first show the estimate of the transition probability α as a function of the number of clones used in AMS in Fig. 4.15 (a). For all three estimates, the 95% intervals of confidence are fairly large: 2% of the best estimate. All three intervals overlap if more than 100 clones are used in AMS computations. Based on this observable alone, both score functions give comparable results, and we cannot conclude on whether one is better than the other.

We then show the estimate of the average length of reactive trajectories as a function of the number of clones used in AMS computations in Fig. 4.15 (b). The two AMS estimates $\langle\tau\rangle_M$ decrease with N towards an asymptotic value.

With the learned committor function ϕ_{dat}, the 95% confidence intervals of the AMS and DNS estimates overlap if $N \geq 250$. This never happens for the linear score function ϕ_{x_1}.

Finally, Fig. 4.15 (c) shows the rescaled variance of the AMS estimator of α as a function of the number of clones. Both are compared to the reference value 1. The learned committor function significantly reduces the statistical error, compared to the linear score function.

We conclude that using the learned committor function computed from a long dataset leads to more precise results than using the user defined score function ϕ_{x_1}, especially for the statistical error and for the estimate of the duration of reactive trajectories. We note that AMS computations yield estimates close to the asymptotic value if $N \geq 1000$.

Efficiency of the AMS algorithm with the learned committor function as a function of the dataset length for the Charney-DeVore model

As we did with the 2D three-well model (Sec. 4.4.3), we now wish to determine the amount of data necessary to learn a committor function leading to good AMS estimates. Again, we sample longer and longer trajectories, containing from 1 to 99 transitions. From each of these datasets we learn a committor function and use it in AMS computations using $N = 1000$ clones. For each dataset length, we perform an average over independent realizations of the score function.

We first consider the transition probability α (Fig. 4.16 (a)) and the average length of the reactive trajectories τ (Fig. 4.16 (b)) as a function of the number of recorded transitions. We note that as soon as there are more than five recorded transitions in the dataset, using the AMS with the learned committor, the 95% intervals of confidence of α and τ overlap with the DNS estimate. This indicates that the learned committor function is relevant for much smaller datasets than used in Sec. 4.4.4. The results improve with the size of the dataset.

We now examine the rescaled variance of the AMS estimate of the transition probability α, using the learned committor function (Fig. 4.16 (c)). We note that if there are very few transitions recorded in the dataset, the variance can be larger than the one obtained using the linear score function. However, the statistical error quickly decreases for larger datasets: it is reduced by about 20% compared to the linear score function when the dataset

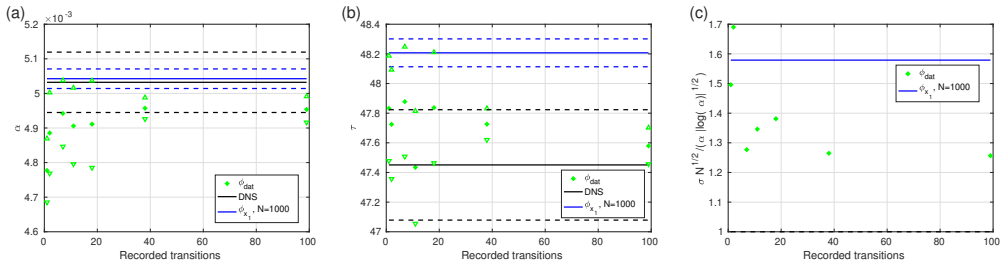


Figure 4.16: Efficiency of the AMS algorithm with the learned committor function as a function of the dataset length (measured in number of transitions), for the Charney-DeVore model. Comparison of the estimated (a) Transition probability $\langle\alpha\rangle$, (b) Duration of reactive trajectories $\langle\tau\rangle$, and (c) Rescaled variance σ. For each plot the black curve is the reference one, either the DNS (a and b) or the optimal value 1 (c). The dashed black lines are the 95% confidence interval for the DNS. The color curves have been computed using the AMS, with the learned committor function (green) and the linear score function (blue). For dataset as short as 5 transitions the AMS algorithm with the learned committor function leads to results as precise as the DNS, and more precise than the linear score function, for both the rescaled variance and trajectory duration. Having few transitions in the dataset leads to variability in the quality of the score function.

contains at least 38 transitions.

4.5 Conclusion

In this paper, we have proposed a data-driven approach for the computation of the committor function. This approach relies on the analogue method to define effective dynamics starting only from observations. We have shown that this defined a Markov chain on the observed states of the dynamics, which approximates the true propagator. This allows a spectral characterization of the committor function. Computing the committor function this way gives remarkably smooth and robust results for the committor function.

We have highlighted by means of two examples that it is possible to obtain fairly precise estimates of the committor function, even in cases where few observations are available. In addition, we have pointed out that these approximations are more precise than those provided by a naiver data-driven approach and that increasing the number of data results in a faster reduction of the error. These improvements are because that the analogue Markov chain is a dynamical approach, which uses all the information contained in the trajectories, while this is not the case for the direct approach, which treats all the points of the same reactive trajectory equally. We also stress that the analogue Markov chain approach can be used using any trajectories of any length, not necessarily distributed according to the invariant measure of the dynamics.

Finally, we provided evidence of the advantage of coupling the analogue method with a rare event algorithm. Indeed, learned committor with the analogue Markov chain can be used as a score function performing better than user defined score functions. This means that it is possible to develop an almost-fully automatic algorithm that requires very little knowledge and understanding of the system under consideration. The quality of the results suggest that better understanding can be obtained a-posteriori.

Although the learned committor function based on the analogue Markov chain, and its coupling with rare event algorithms, have revealed several very interesting advantages, some limitations might arise especially when one faces high-dimensional systems. We have tested the approach for a fairly complex dynamics with 6 degrees of freedom. It still has to be tested for more complex dynamics. For systems in high dimensions, the choice of the distance for the analogue method might be a critical issue.

Another interesting question would be to compare the quality of the estimation of the committor function using the analogue Markov chain, with other methods. It would be interesting to compare it other methods based using dynamical information, sometimes more complicated, for instance the direct Galerkin approximation [Thiede et al., 2019, Strahan et al., 2021] method. It would also be interesting to compare it to direct approaches using machine learning.

Chapter 5

Predicting extreme events using the analogue method: the heat-wave case

5.1 Introduction: heat waves and committor functions

Often rare events, for instance extreme heat waves or cold spells, have a huge impact on socio-economic systems [Ragone et al., 2018, Field et al., 2012, AghaKouchak et al., 2012, Herring et al., 2014, Coumou and Rahmstorf, 2012, Ragone and Bouchet, 2021]. Therefore, one of the major challenges nowadays is to accurately describe their dynamics and find effective ways to forecast the probability of occurrence of such events. The greatest difficulty in studying these events lies in the lack of observations. For example, consider the case of heat waves. These events happen every year in different parts of the world and have been observed in the past. The impact, amplitude and duration of these events vary widely depending on the region in which they occur, the period and many other factors. The most extreme events have a return time (the average time between two occurrences of the event) of the order of hundreds or thousands of years [Ragone et al., 2018, Ragone and Bouchet, 2021]. It is therefore not possible to rely on historical data because such events may have never been observed due to the many factors that contribute to their occurrence. Difficulties in analyzing these events also arise using climate models because of their huge computational costs.

5.1. INTRODUCTION: HEAT WAVES AND COMMITTOR FUNCTIONS

This study focuses on extremes of long lasting summer heat waves. Although various definitions can be given, roughly speaking a heatwave is an extended period of hot weather relative to the expected conditions of the area at that time of year. As an example the France region has been considered.

According to existing phenomenological theory[1] [Perkins, 2015, Horton et al., 2016], the principal causes of heat waves include large-scale atmospheric circulation patterns [Cassou et al., 2005, Della-Marta et al., 2007, Jézéquel et al., 2018] and lack of rainfall and soil moisture [Vautard et al., 2007, Zampieri et al., 2009, D'Andrea et al., 2016]. Although these are recognized as the main causes, it is extremely complex to establish causal relationships between them [Horton et al., 2016] and their influence varies in importance depending on the geographical region [Stefanon et al., 2012]. For instance, it is well established that long-lasting heat waves in extratropical regions are related to persistent weather regimes and blocking events [Lau and Kim, 2012, Hoskins and Woollings, 2015, Horton et al., 2016, Kornhuber et al., 2019, Ragone and Bouchet, 2021].

Therefore, studying extreme heat waves concerns the study of the nonlinear and turbulent dynamics of the atmosphere. Two key variables are the temperature and pressure fields, although it is important to note that relative humidity can also play a role. Surface temperature and pressure fields at a certain height of the vertical coordinate are usually considered. The most convenient vertical coordinate is the geopotential height. The geopotential $\Phi(\phi, \theta, z)$ is the gravitational potential energy per unit mass at latitude ϕ, longitude θ, and elevation z,

$$\Phi(\phi, \theta, z) = \int_0^z \mathrm{d}z'\, g(\phi, \theta, z), \tag{5.1}$$

where $g(\phi, \theta, z)$ is the gravity acceleration. The geopotential height $Z(\phi, \theta, z)$ is the geopotential normalized to the standard gravity at mean sea level $g_0 = 9.80665\ \mathrm{ms}^{-2}$, i.e.

$$Z(\phi, \theta, z) = \frac{\Phi(\phi, \theta, z)}{g_0}. \tag{5.2}$$

Instead of using the pressure fields at a certain value of the geopotential height, it is possible to look at the value of the geopotential height on a sur-

[1] The term phenomenological theory is used here to indicate a theory based on the observations of a certain phenomenon and which therefore does not derive from first principles.

face defined by a fixed pressure. Indeed, the geopotential height on a surface defined by a fixed pressure behaves as a streamfunction for the geostrophic wind vector and it is often used to visualize the state of the atmospheric circulation since it highlights regions of low pressure (cyclonic anomalies typically associated with bad weather) and high-pressure regions (anticyclonic anomalies typically associated with fair weather). The geopotential height is important in the study of heat waves as these are typically associated with persistent anticyclonic anomalies [Ragone et al., 2018]. As an example, consider the map of temperature at 850 hPa (colors) and geopotential height at 500 hPa (contours) in Fig. 5.1. As can be seen, the geopotential height gives indications on atmospheric circulation. Indeed, the two predominant features are the jet streams of the two hemispheres. Furthermore, it can be noted that the regions where the temperature is high (low) are associated with high (low) values of the geopotential height.

In this study, a specific criterion for selecting heat waves has been chosen. It consists of extremes of time averaged surface temperature fluctuations (anomalies) defined as

$$A(t) = \frac{1}{T}\int_t^{t+T} \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \left(T_s - \mathbb{E}_c\left[T_s\right]\right)(\mathbf{r}, t')\,\mathrm{d}\mathbf{r}\mathrm{d}t', \tag{5.3}$$

where $\mathcal{D}$ is a specific region of the globe, $|\mathcal{D}|$ represents the area of the region and $\mathbb{E}_c\left[\cdot\right]$ denotes the climatological average, i.e. an average over time made for any given location $\mathbf{r}$ and intra-year time t that preserves the intra-year seasonal effect. The integrand of Eq. (5.3) is the surface temperature anomaly field and will be indicated by $T_s^a(\mathbf{r}, t) = \left(T_s - \mathbb{E}_c\left[T_s\right]\right)(\mathbf{r}, t)$. Throughout the study, heat waves of duration $T = 15$ days will be analyzed. It should be underlined that extremes of temperature anomalies rather than extremes of absolute temperature are considered. This definition emphasizes dynamical characteristics, which are thought to be described reasonably well on a seasonal time scale, rather than physical impacts which are often related to the absolute temperature. However, it can be expected that those events have the same dynamical characteristics, at least for the most important aspects.

Considering that the predictability horizon of the atmosphere is about two weeks, it is easy to be convinced that the problem of predicting heat waves a few days in advance falls into the class of prediction problems at the predictability margin. Therefore, even in this case, the appropriate mathematical tool to deal with this problem is the committor function.

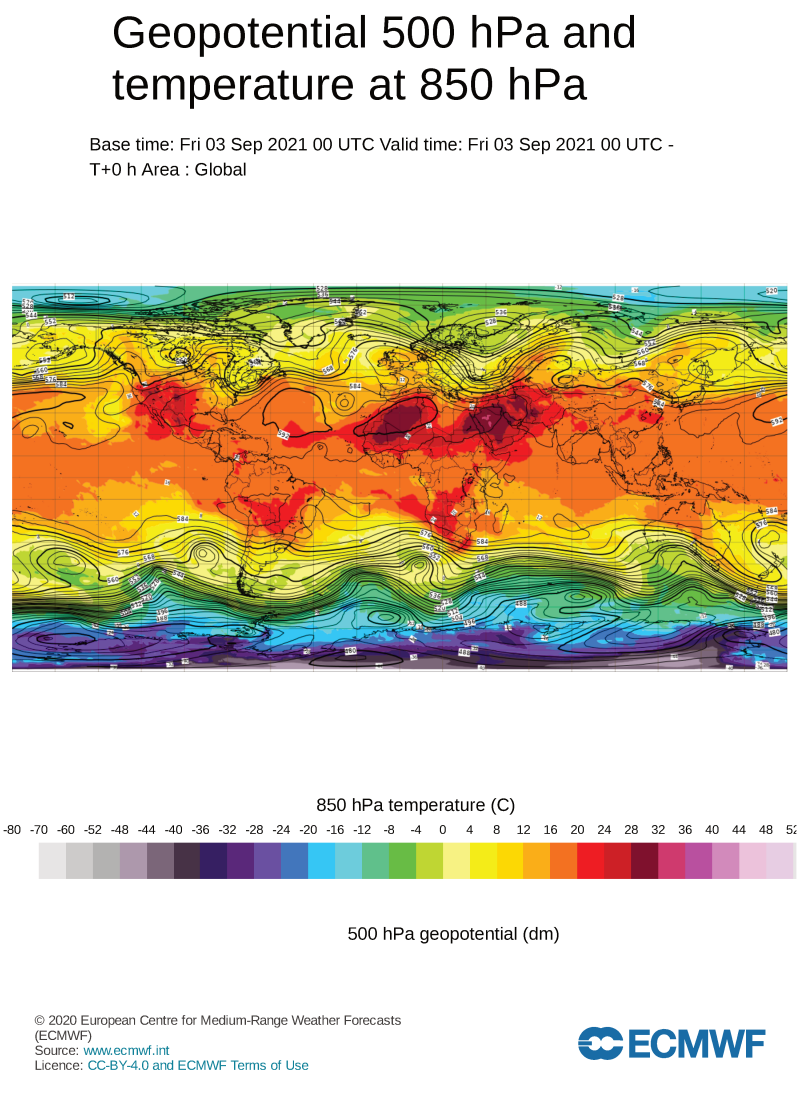


Figure 5.1: Map of 500 hPa geopotential height (contours) and temperature at 850 hPa (shading) on 03/09/2021 from the ECMWF HRES model. Height contours are labelled in tens of meters (dam) with an interval of 4 dam. The contours effectively show the main tropospheric waves that "control" weather - low heights indicate troughs and cyclones in the middle troposphere whilst high heights indicate ridges and anticyclones. Color shading indicates temperature at the 850 hPa level in degrees Celsius ($°C$), in 4 $°C$ color bands. This is the temperature approximately 1.5 km above sea level.

The committor function $q_\tau(x;a)$ is the probability that a heat waves of amplitude greater than a will occur in τ days as a function of the current state of the system x, i.e.

$$q_\tau(x;a) = \mathbb{P}(A(t_0+\tau) > a | X(t_0) = x). \tag{5.4}$$

The state variable x can be the set of all the variables that describe the system (temperature, pressure, soil moisture, wind velocity and so on) or a part of them (for instance temperature and geopotential height). Since at the predictability margin the system behaves like a stochastic process, the committor function $q_\tau(x;a)$ takes values in the interval $[0,1]$. In Sec. 2.5.1 it was shown that the probability defined in Eq. (5.4) is a committor function for the auxiliary process $Y(t) = [O(X(t),t),t]$, where $O(X(t),t)$ is the time averaged surface temperature fluctuation $A(t)$.

The computation of the committor starting from observations (being they real measurements or climate model outputs) is a complex task, especially for very extreme events (i.e. for large values of a), given the large dimensionality of the system and the scarcity of recorded events. However, it is crucial to access the information $q_\tau(x;a)$ contains. In fact, knowing how likely it is that, starting from the state x, a heat wave will occur in the next τ days could help to make informed decisions in order to mitigate the impact of such an event on societies. Furthermore, by studying the dependence of $q_\tau(x;a)$ on the different variables that define the state of the system x, it is possible, on the one hand, to identify more or less informative variables for the prediction of heat waves, and, on the other hand, to recognize recurrent dynamical patterns (such as teleconnection patterns). Finally, as explained in Chapter 4, $q_\tau(x;a)$ can be used to drive rare events algorithms, making available a larger collection of observations and helping the understanding of such events.

The aim of this study is to introduce the analogue Markov chain for the computation of extreme events probabilities, and in particular the probabilities of occurrence of heat waves.

The structure of the chapter is the following: Sec. 5.2 introduces the analogue method in the context of climate dynamics and explains how to build several types of analogue Markov chains. These different types of Markov chains are characterized by different properties. The first distinction is made between seasonal and annual Markov chains. The former are developed by considering dynamical observations related to a single season (e.g. summer)

and consequently are used to generate synthetic trajectories only in that season. The latter, instead, are developed using yearly trajectories. Within the class of annual Markov chains, a further distinction can be made between time-periodic or homogeneous Markov chains depending on whether the transition matrix depends on time or not. Then, Markov chains are used to generate synthetic trajectories. From these synthetic trajectories, some stationary properties of the chains, such as the distribution of the time averaged temperature A (Eq. (5.3)), the return times, and the probability distribution of the days visited by the trajectories, are computed. In addition, these quantities are compared to the ones computed from the real dynamics (where the term real refers here to climate model outputs). These consistency checks show that the type of Markov chain suitable for reproducing heat wave statistics depends on which variables are used as states of the chain. The section ends by showing that the analogue method allows the observation of events whose return time is much longer than the length of the trajectory used for building the analogue Markov chain. This result proves the ability of the analogue method to generate realistic trajectories useful for studying extreme events. Sec. 5.3 deals with the computation of the committor function using the analogue method. The committor function is computed by employing seasonal Markov chains. Indeed, as it will be better explained throughout the chapter, the committor function only depends on short trajectory statistics (since it is a function of the initial condition) and it is therefore not too much sensitive to the choice of the type of the analogue Markov chain. The predictive skills of the committor are evaluated by building a classifier (the analogue based classifier) that predicts the occurrence or not of a heat wave according to the current state of the system. Then, the classifier is used to assess the impact of some parameters, such as the coarse-graining time, the spatial domain on which analogues are computed and the combination of different physical quantities, on the predictability of extreme events. The section concludes by showing a comparison between the analogue based classifier and a classifier based on the committor function computed with a more naive approach (k-Nearest Neighbors regressor) highlighting that the analogue method provides more accurate predictions. Due to lack of time, a systematic study was not carried out in this section. Therefore most of the results obtained here must be thought of as insights rather than assertions.

5.2 The analogue method for heat waves

In Sec. 4.3, the analogue method has been introduced demonstrating its usefulness in computing committor functions. Given the success in low-dimensional stochastic dynamics, the purpose of this section is to illustrate its application to the study of heat waves. This is an extremely complex topic, especially regarding the occurrence of the more severe ones due to the difficulty in collecting sufficient statistics. Clearly, the statistics of the dynamics introduced by the analogue method converge to the real statistics when infinite data-set are considered, also keeping constant the number of analogues K. However, in the case under investigation, the data-set is too short and the convergence has to be studied empirically.

The use of analogues of atmospheric circulation for weather forecasting dates back to the second half of the 20th century with the pioneering works of Lorenz [Lorenz, 1969a, Lorenz, 1969c]. Over the years, forecasts based on analogues have been set aside due to the difficulty of finding good analogues within a limited number of observations [Van den Dool et al., 2007]. However, slightly modified versions of this method have sparked new interest in recent years, especially regarding the development of stochastic weather generators [Yiou et al., 2013, Yiou, 2014, Jézéquel et al., 2018, Yiou and Déandréis, 2019, Yiou and Jézéquel, 2020]. Following the ideas developed in these works, the analogue method can therefore be used to generate synthetic time series at low computational cost, thus allowing to overcome the lack of observations.

The dynamics that will be examined is deterministic as it is generated by a climate model (PLASIM). PLASIM is a general circulation model that aims at solving the fluid mechanics equations for the atmosphere [Fraedrich et al., 2005, Lunkeit et al., 2011]. At first glance, the use of a stochastic generator to learn deterministic dynamics may seem inappropriate but this choice is more reasonable than one might think. In fact, it is true that PLASIM is a deterministic climate model, but it is chaotic and, like all climate models, it features the evolution of thousands of variables which evolve on very different time scales. A well established result for dynamical systems with multiple time scales is that, if there is a gap between the time scales (i.e. slow variables evolve on much larger time scales than fast variables) and the dynamics of fast variables is mixing (i.e. the system looses the memory of its initial condition after a certain time), the dynamics of slow variables can be described by a system of Markovian stochastic equations [Kifer, 1992, Kifer, 2004, Givon et al., 2004, Melbourne and Stuart, 2011]. However, since heat waves originate

from the interplay of processes taking place on different time scales, the time scale gap does not exist for the climate model. In order to cope with this issue and to make the application of the analogue method meaningful, it could be useful to introduce a temporal coarse graining for the variables used to describe the system. To explain the reasoning behind it, it is better to first define a coarse grained variable. Let $\mathcal{F}(\mathbf{r}, t_0)$ be the value of a generic field at point $\mathbf{r} \in \mathbb{R}^2$ at time t_0. One can introduce the coarse grained field $\tilde{\mathcal{F}}(\mathbf{r}, t_0)$ where the coarse graining is done in time with a coarse graining duration equal to τ_c:

$$\tilde{\mathcal{F}}(\mathbf{r}, t_0) = \frac{1}{\tau_c} \int_{t_0 - \frac{\tau_c}{2}}^{t_0 + \frac{\tau_c}{2}} \mathrm{d}t \, \mathcal{F}(\mathbf{r}, t). \tag{5.5}$$

The typical correlation time τ_{corr} of the relevant variables for heat waves, such as temperature or geopotential height, is of the order of a few days in PLASIM. Considering a coarse graining time of the same order of the correlation time, i.e. $\tau_c \sim \tau_{corr}$, and looking at the system on time scales of the same order, one has that $\tilde{\mathcal{F}}(\mathbf{r}, t_0 + \tau_c)$ is correlated with $\tilde{\mathcal{F}}(\mathbf{r}, t_0)$ but much less with $\tilde{\mathcal{F}}(\mathbf{r}, t_0 - \tau_c)$. This justifies the choice of adopting a Markovian description, i.e. the analogue method, of the system.

In what follows it will be described how to adapt the analogue method to study heat waves. Since the development of the analogue Markov chain has already been explained in Sec. 4.3.1, the following discussion focuses solely on how to select analogues.

The state of the system will be defined by a vector $X_t = \left(\tilde{\mathcal{F}}_1(\mathbf{r}, t), \cdots, \tilde{\mathcal{F}}_I(\mathbf{r}, t)\right)$ where I is the number of variables considered. To identify the analogues of a point X, a distance $d(\cdot)$ must be defined. A good choice would be to use a Euclidean distance normalized to the local variance of the fields. The distance will be labeled with d^l where l stands for *local*. To be more precise, let $\sigma^2_{\tilde{\mathcal{F}}_i}(\mathbf{r})$ be the local variance of the field $\tilde{\mathcal{F}}_i(\mathbf{r}, t)$, i.e.

$$\sigma^2_{\tilde{\mathcal{F}}_i}(\mathbf{r}) = \mathbb{E}_t[\tilde{\mathcal{F}}_i^2(\mathbf{r}, t)] - \mathbb{E}_t[\tilde{\mathcal{F}}_i(\mathbf{r}, t)]^2. \tag{5.6}$$

where $\mathbb{E}_t[\cdot]$ is the expectation value over time (which will be approximated by time average over all time steps). Note that $\mathbb{E}_t[\cdot] \neq \mathbb{E}_c[\cdot]$ since $\mathbb{E}_t[\cdot]$ does not preserve intra-year seasonal effect.

Then, the distance d^l between two points $X_1 \equiv X_{t_1} = \left(\tilde{\mathcal{F}}_1(\mathbf{r}, t_1), \cdots, \tilde{\mathcal{F}}_I(\mathbf{r}, t_1)\right)$

and $X_2 \equiv X_{t_2} = \left(\tilde{\mathcal{F}}_1(\mathbf{r}, t_2), \cdots, \tilde{\mathcal{F}}_I(\mathbf{r}, t_2)\right)$ is defined as

$$d^l(X_1, X_2) = \left[\int \mathrm{d}\mathbf{r} \sum_{i=1}^{I} \frac{(\tilde{\mathcal{F}}_i(\mathbf{r}, t_1) - \tilde{\mathcal{F}}_i(\mathbf{r}, t_2))^2}{\sigma^2_{\tilde{\mathcal{F}}_i}(\mathbf{r})}\right]^{\frac{1}{2}}. \tag{5.7}$$

where the integral is to be intended on a region of space $\mathcal{D}$. Another possibility is to use a Euclidean distance normalized to the average variance, i.e.

$$d^g(X_1, X_2) = \left[\int \mathrm{d}\mathbf{r} \sum_{i=1}^{I} \frac{(\tilde{\mathcal{F}}_i(\mathbf{r}, t_1) - \tilde{\mathcal{F}}_i(\mathbf{r}, t_2))^2}{\Sigma^2_{\tilde{\mathcal{F}}_i}}\right]^{\frac{1}{2}}, \tag{5.8}$$

where

$$\Sigma^2_{\tilde{\mathcal{F}}_i} = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \mathrm{d}\mathbf{r} \sigma^2_{\tilde{\mathcal{F}}_i}(\mathbf{r}). \tag{5.9}$$

An alternative approach is to use the Mahalanobis distance $d^M(\cdot)$, which is

$$d^M(X_1, X_2) = \left[\int \mathrm{d}\mathbf{r_1}\mathrm{d}\mathbf{r_2} \sum_{i=1}^{I} (\tilde{\mathcal{F}}_i(\mathbf{r_1}, t_1) - \tilde{\mathcal{F}}_i(\mathbf{r_1}, t_2)) \mathrm{Cov}^{-1}_{\tilde{\mathcal{F}}_i}(\mathbf{r_1}, \mathbf{r_2}) (\tilde{\mathcal{F}}_i(\mathbf{r_2}, t_1) - \tilde{\mathcal{F}}_i(\mathbf{r_2}, t_2))\right]^{\frac{1}{2}}. \tag{5.10}$$

A priori, it is not obvious which is the most appropriate distance to use: the Mahalanobis distance is the most discriminative criterion but its computational cost is quite high because it requires to compute the product between large matrices. For this reason, the Euclidean distances have been preferred and the choice between the local and the global version depends on the spatial homogeneity of fields' fluctuations.

The last ingredient for the development of the analogue Markov chain is the choice of its time unit τ_M. Note that the unit time of the Markov chain τ_M and the coarse graining time τ_c are different objects. Indeed, the former represents the lag time between a state s_0 and its evolution s_1 while the latter is the duration of the time averages that define the coarse grained variables. Although they are different concepts, it this chapter, unless differently stated, $\tau_M = \tau_c$ will be considered.

In the following, several implementation of analogue Markov chains will be defined, analyzing the consistency of synthetic data with respect to the real ones.

5.2.1 Stationary vs time periodic analogue Markov chains

The main novelty between the systems studied in the previous chapter and the climate model analyzed here is the presence of a seasonal cycle in climate data. Hence, in the present case, several analogue Markov chains can be developed: they can be built on a seasonal (e.g. summer) or annual data-set. In the latter case, it is possible to make a further distinction between homogeneous analogue Markov chains (the transition matrix does not depend on time) or time-periodic analogue Markov chains (the transition matrix depends on time and has an annual periodicity).

Before illustrating how to develop the three Markov chains it might be convenient to briefly describe the data-set. It consists of 10 independent realizations of the climate dynamics produced by PLASIM. Each realization (also called batch) is 100 years long with outputs provided every 3 hours. Thus, each batch consists of time series (one for each model variable) of $N_b = 100 \times 360 \times 8 = 288000$ time steps (each month is 30 days long). Although the temporal coarse-graining procedure reduce the number of time steps to $\tilde{N}_b = N_b - \tau_c \times 8$, the search for analogues in a data-set made up of $\tilde{N}_b$ points is extremely slow. Therefore, it was chosen to use a sampling time $\tau_s = 1$ day to further reduce the size of the data-set. Note that three different times are relevant to define the analogue Markov chain, namely τ_c, τ_M and τ_s. The coarse graining time τ_c represents the duration of the time averages that define the coarse grained variables. The Markov chain time step τ_M is the lag time between two consecutive states of the Markov chain. Finally, the sampling time τ_s represents the lag time between two observations in the data-set, i.e. the sampling time of the climate model. Although these three times differ from each other, throughout the chapter $\tau_M = \tau_c$ is considered, unless otherwise stated.

By adopting $\tau_s = 1$ day, the number of points in each batch becomes $\hat{N}_b = 100 \times 360 - (\tau_c - 1)$. To be more explicit, using $\tau_c = 5$ days each batch will contain variables defined between January 3 of year 1 and December 28 of year 100. Thus, the total number of points in the data-set is $N_t = 10 \times \hat{N}_b \simeq 360000$.

Having defined the data organization, it can be explained how to build the different types of analogue Markov chains.

Seasonal analogue Markov chains

For the construction of this type of analogue Markov chains only the data corresponding to a given season (e.g. summer) are considered. This implies that instead of having 10 independent realizations each 100 years long, there will be 1000 independent realizations each 90 days long. Recalling the terminology of Sec. 4.3.1, the data-set will be indicated with $\{X_n\}_{1\leq n\leq N_t}$ with $N_t = 90000$. Unlike Sec. 4.3.1 where the data-set consisted of a single trajectory and the time step of the Markov chain coincided with the sampling time, here the data-set is composed of multiple trajectories and the time step of the Markov chain τ_M is not equal to the sampling time τ_c. Therefore, instead of excluding only the final point of the trajectory from the possible candidates for the analogues, the last τ_M points of each trajectory will be excluded as it is not possible to associate any transition from them. To be more precise, using $\tau_c = 5$, for each point X_n the analogues are searched between June 1 and August 25 of each year. Keeping these small changes in mind, the rest of the procedure is equivalent to that already described in Sec. 4.3.1. The K best analogues will be stored in a matrix $\mathcal{T}$ of size $N_t \times K$, where K is the analogue number and N_t the data-set length. The matrix $\mathcal{T}$ is a matrix of indices, i.e. the n-th row of $\mathcal{T}$ contains the indices of the K analogues of the point X_n. Note that the index n of each point X_n is nothing more than the time at which X_n was observed. As shown in Sec. 4.3.1, from the matrix $\mathcal{T}$ it is possible to derive the transition matrix of the analogue Markov chain G (Eq. (4.14)).

The synthetic trajectories can be built by applying the same procedure as in Sec. 4.3.1. Given an initial state of the Markov chain s_0, a random number k distributed uniformly in the interval $[1, K]$ is generated. Then s_1, the evolved state from s_0, will be $s_1 = \mathcal{T}_{s_0 k} + \tau_M$. Once again, note that τ_M can be added to the elements of the matrix $\mathcal{T}$ as the latter represent the times at which the corresponding states were observed. Therefore the state corresponding to s_1 (X_{s_1}) is the state that comes τ_M days after the state corresponding to $\mathcal{T}_{s_0 k}$ ($X_{\mathcal{T}_{s_0 k}}$). The entire synthetic trajectory is obtained by iterating the previous step.

Annual homogeneous analogue Markov chains

Since seasonal analogue Markov chains are learned using only the data corresponding to a single season, they can only be used to generate synthetic

trajectories in that season. Instead, the aim of this section is to introduce a type of analogue Markov chains that can be used to generate annual trajectories. In this way, generating a very long synthetic trajectory corresponds to simulating the dynamics of the system over several years.

This type of Markov chain is built using all the 10 batches. Therefore, the data-set is $\{X_n\}_{1\leq n\leq N_t}$ with $N_t = 10 \times (100 \times 360 - (\tau_c - 1))$. At this point, it should be noted that due to the strong seasonal cycle only fields in the same season share similar dynamics. Therefore, the analogues of a point X are searched not in the whole data-set but in a time window of two months centered on the day corresponding to the state X. To be clearer, the analogues of June 1st will be searched among the fields corresponding to the period May 1st - July 1st of all available years.

As in the previous case, attention must be paid to the end points of the trajectories. Therefore, considering that each batch loses the last $\frac{\tau_c-1}{2}$ days of December of year 100 due to temporal coarse-graining, the last $\tau_M + \frac{\tau_c-1}{2}$ December days of year 100 of each batch are excluded from the set in which analogues are searched. To clarify, with $\tau_c = \tau_M = 5$ days, the days following December 23 of year 100 of each batch can not be the analogues of any state.

Once the matrix $\mathcal{T}$ and the corresponding transition matrix G have been learned, they can be used to generate synthetic trajectories in the same manner as for the seasonal Markov chain.

Since the analogues of each point X depend exclusively on the calendar day of X, they do not evolve over time as the time of the Markov chain moves forward. Therefore, the transition matrix G is time independent. Recalling the terminology of Sec. 2.2, it means that this annual analogue Markov chain is homogeneous. This also implies that these Markov chains do not have seasonality or, more precisely, they have a random seasonality in the sense that the duration of the seasons and the duration of the year are random variables.

Annual time-periodic analogue Markov chains

This section aims to illustrate how to develop annual analogue Markov chains whose dynamics have the same periodicity as the climate model data.

As for the annual homogeneous Markov chains, the data-set considered for the development of the annual time-periodic Markov chains consists in 10 trajectories each long 100 years, i.e. $\{X_n\}_{1\leq n\leq N_t}$ with $N_t = 10 \times (100 \times 360 - (\tau_c - 1))$. The final points of the trajectories are also treated

as before, that is the last $\tau_M + \frac{\tau_c - 1}{2}$ December days of the last year of the trajectory are not considered as possible analogues.

The main difference with respect to how the annual homogeneous Markov chain is defined is that, in this case, a fundamental role is played by the calendar date of the Markov chain. The calendar date of the Markov chain, denoted by $\tilde{c}$, indicates the day of the year which corresponds to a given time $\tilde{t}$ of the Markov chain. Therefore, $\tilde{c}$ is a number which takes values between 1 and 360 corresponding to a day of the year (remember that each year in PLASIM is 360 days long). The time of the Markov chain $\tilde{t}$ keeps track of the time that flows in the synthetic dynamics and, at each time step, it increases by τ_M. The relation between $\tilde{t}$ and $\tilde{c}$ is $\tilde{c} = (\tilde{t} \mod 360)$, where mod indicates the modulo operation.

In order to develop a Markov chain whose transition matrix is time-periodic, the procedure for selecting the analogues differs from the one employed in the development of the annual homogeneous Markov chain. Indeed, in this case, the analogues of a state X are chosen in a time window of two months centered around the Markov chain calendar date $\tilde{c}$ and not on the date corresponding to the state itself.

An example can help to better understand the concept and it also illustrates how synthetic trajectories are generated in this case. Let c denotes the calendar date of a state X in the data-set, i.e. c represents the day at which X was observed. Consider $\tau_c = \tau_M = 5$ days and suppose one wants to generate a synthetic trajectory starting on January 1st. This means that the Markov chain calendar date is $\tilde{c}_0 = 1$ (since it is the first day of the year). Also the Markov chain time $\tilde{t}_0$ is equal to 1. Let s_0 be the initial state of the Markov chain corresponding to a point X and let c be the real calendar date of X. It has been explained that, for developing the annual analogue time-periodic Markov chain, the analogues have to be sought in a time window of two months centered around the Markov chain calendar date $\tilde{c}$. This means that, in this example, regardless of the value assumed by c (i.e. regardless of the day to which X corresponds), the analogues of X must be searched for in the period from December 1st to February 1st, i.e. in a time window of two months centered around $\tilde{c}_0 = 1$. Let $\mathcal{T}_{s_0}(\tilde{t}_0)$ indicates the set of K analogues of X at time $\tilde{t}_0 = 1$ (i.e. January 1st). To evolve s_0, an analogue k has to be selected from $\mathcal{T}_{s_0}(\tilde{t}_0)$. Then s_1, the evolved state from s_0, will be $s_1 = \mathcal{T}_{s_0 k}(\tilde{t}_0) + \tau_M$. At the same time, the Markov chain calendar date $\tilde{c}$ and the Markov chain time must also be updated, i.e. $\tilde{t}_1 = \tilde{t}_0 + \tau_M$ and $\tilde{c}_1 = (\tilde{t}_1 \mod 360)$. Then, the set of K analogues of s_1, denoted by $\mathcal{T}_{s_1}(\tilde{t}_1)$, is

computed by searching the best K analogues of s_1 in a period of two months centered around $\tilde{c}_1$. In this example, it corresponds to search analogues in the period between December 6th and February 6th. The Iteration of this scheme leads to the generation of a synthetic trajectory.

Applying this procedure, the set of K analogues of a state X changes as the Markov chain calendar date $\tilde{c}$ changes. Thus, analogues can be stored in $M = 360$ (one for each day of the year) $\{\mathcal{T}(m)\}_{1\leq m\leq M}$ matrices of size $N_t \times K$. To each of these M matrices it is possible to associate a transition matrix $G(m)$ making the Markov chain periodic in time.

5.2.2 Consistency of synthetic data: return time plot and time averaged temperature statistics

In this section the consistency of the statistics of the synthetic data compared to the real ones will be analyzed. Synthetic data refers to data produced using the dynamics of the Markov chain, while real data refers to the data generated by the climate model (PLASIM). Since the ultimate interest is to estimate committor functions for heat waves, it is useful to check how similar the distribution of the time averaged temperature $A(t)$ (Eq. (5.3)) computed with the analogue Markov chain is to the real one. For an heat wave duration $T = n\tau_c$, $A(t)$ can be expressed in term of the coarse-grained surface temperature anomalies field $\tilde{T}_s^a(\mathbf{r}, t)$, i.e.

$$A(t) = \frac{1}{T}\int_t^{t+T} \mathrm{d}t' \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \mathrm{d}\mathbf{r}(T_s - \mathbb{E}_c[T_s])(\mathbf{r}, t') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \mathrm{d}\mathbf{r}\tilde{T}_s^a(\mathbf{r}, t_i') \tag{5.11}$$

where $t_i' = t + (i - \frac{1}{2})\tau_c$. Eq. (5.11) means that $A(t)$ is obtained by adding the surface temperature anomalies averaged over the region $\mathcal{D}$ over n consecutive states of the Markov chain (remember $\tau_c = \tau_M$). In what follows a duration $T = 15$ days is considered (15−day heat waves).

The test can be formalized as follows. Let $\rho_A(A)$ be the real empirical distribution of $A(t)$ in summer, i.e. $\rho_A(A) = \frac{1}{t}\sum_t \delta(A - A(t))$, and let $\tilde{\rho}_A(A)$ be the distribution obtained with the analogue Markov chain. One has that

$$\tilde{\rho}_A(A) = \sum_{l\in\{\text{Summer}\}} \pi_l \mathbb{P}(A|s_0 = l) := \sum_{l\in\{\text{Summer}\}} \pi_l \tilde{\rho}_A^{(l)}(A), \tag{5.12}$$

where $\mathbb{P}(A|s_0 = l) \equiv \tilde{\rho}_A^{(l)}(A)$ represents the conditional probability of observing the value A knowing the initial state s_0, while π_l is the probability of

being in the state l (i.e. the stationary distribution of the Markov chain). Since A is defined using n consecutive states of a realization of the Markov chain (see Eq. (5.11)), for any initial condition $s_0 = l$, A can assume at most K^n different values, denoted by A_k^l, for $k = 1, \cdots, K^n$. Furthermore, each of the K^n realizations is equally probable. Therefore, the conditional probability $\mathbb{P}(A|s_0 = l)$ can be expressed as

$$\mathbb{P}(A|s_0 = l) = \tilde{\rho}_A^{(l)}(A) = \frac{1}{K^n} \sum_{k=1}^{K^n} \delta(A - A_k^l). \tag{5.13}$$

Combining (5.12) and (5.13), one has

$$\tilde{\rho}_A(A) = \sum_{l \in \{\text{Summer}\}} \frac{1}{K^n} \sum_{k=1}^{K^n} \pi_l \delta(A - A_k^l). \tag{5.14}$$

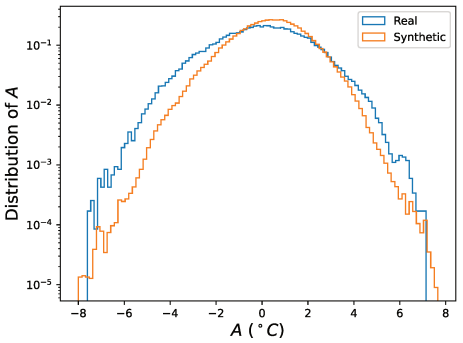
Hence, the first consistency check consists in comparing $\rho_A(A)$ and $\tilde{\rho}_A(A)$. From a practical point of view, $\tilde{\rho}_A(A)$ can be estimated on a very long synthetic trajectory.

In the following, the distributions resulting from different Markov chains will be considered. In each of the tests performed, a coarse-graining time $\tau_c = 5$ days was considered.

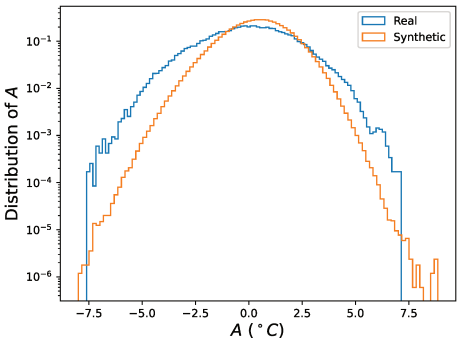
Consistency of time averaged temperature statistics: seasonal Markov chains

The first consistency test was performed on a seasonal summer (JJA) Markov chain built on the coarse-grained field of the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ defined in the North-Atlantic region (80°W − 30°E,30°N − 70°N), i.e. $X = \tilde{Z}$. The choice to use the geopotential height at 500 mb in the North-Atlantic region is due to the fact that, as explained in Sec. 5.1, the occurrence of heat waves is linked to persistent anticyclonic anomalies and the atmospheric circulation of this region is the one that most influences the occurrence of extreme heat waves over France.

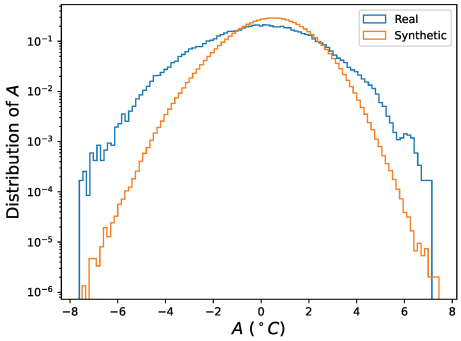
In this case, the analogues are computed by employing the Euclidean distance normalized to the average variance (Eq. (5.8)). The distributions of $A(t)$ for the Markov chains computed over very long trajectories are shown in Fig. 5.2 for different numbers of analogues K. It can be seen that, regardless of the number of analogues employed, the distribution obtained by using the



(a) $K = 5$ *analogues.*



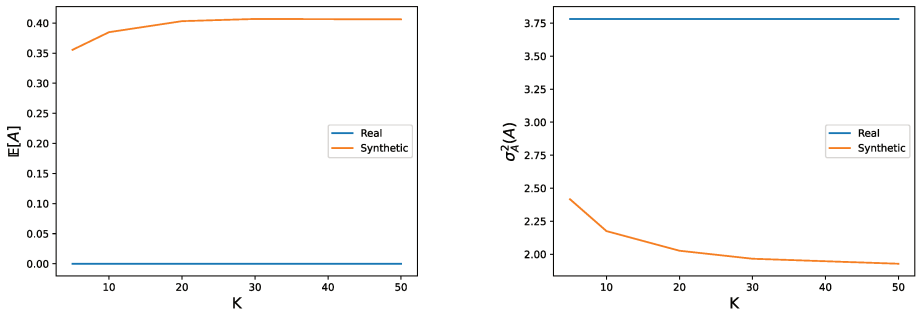
(b) $K = 20$ *analogues.*



(c) $K = 50$ *analogues.*

Figure 5.2: Comparison between $\rho_A(A)$ (real) and $\tilde{\rho}_A(A)$ (synthetic) for different numbers of analogues K. The analogue Markov chain has been defined on the geopotential height at 500 mb field $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.

analogue method does not match the real one. In particular, the differences are noticeable when comparing the tails of the distributions, that is, when comparing the occurrences of the most extreme events. Furthermore, a close inspection of Figs. 5.2a, 5.2b, 5.2c reveals that the differences are accentuated more and more when K increases. To make this analysis more quantitative, Fig. 5.3 shows the comparison between the first two moments (the mean $\mathbb{E}[A]$ and the variance $\sigma_A^2(A)$) of the distributions as the number of analogues K varies. It can be seen from Fig. 5.3a that the analogue method introduces a bias and therefore the mean of the distribution is slightly positive. It can also be noted that the bias slightly increases with K for small values of K and then remains almost constant when the number of analogues $K \gtrsim 20$. In itself this would not be a big problem since the study of heat waves involves the study of the right tail of the distribution (large values of A). On the other hand, Fig. 5.3b highlights a major problem. In the steady state of the Markov chain, the variance of the distribution is much lower than the real one. Furthermore, by increasing the number of analogues K, the variance drastically decreases passing from the value of 2.5 (which corresponds to a reduction of about 35% with respect to the real one) to a value of 2 (which corresponds to a reduction of about 50%). This places serious limits on the use of such a Markov chain for the study of extreme heat waves.

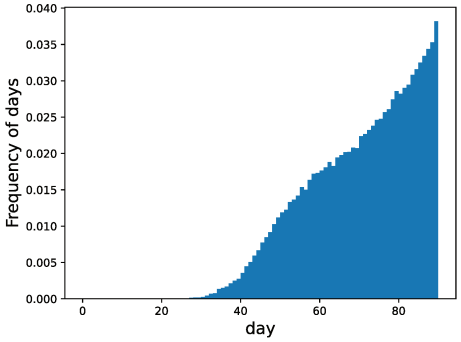


(a) *Mean.* (b) *Variance.*

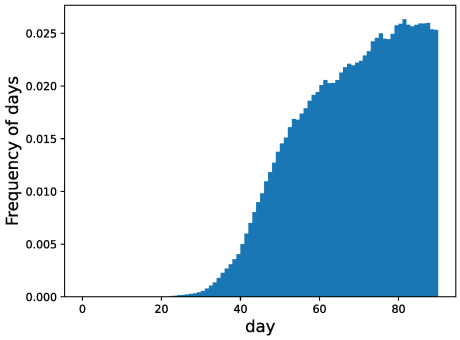
Figure 5.3: Mean $\mathbb{E}[A]$ and variance $\sigma_A^2(A)$ of the distribution of $A(t)$ as a function of the number of analogues K. The analogue Markov chain has been defined on the geopotential height at 500 mb field $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.

The reason for this behavior has to be sought in the stationary distribu-

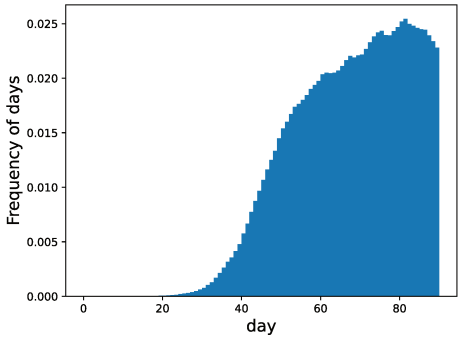
tion of calendar days c. Note that the actual calendar days c are considered (i.e. the days at which the states were observed) and not the Markov chain calendar days $\tilde{c}$. This distribution is shown in Fig. 5.4 for different numbers of analogues K. By looking at the three figures, it can be noted that there is



(a) $K = 5$ *analogues.*



(b) $K = 20$ *analogues.*



(c) $K = 50$ *analogues.*

Figure 5.4: Stationary distribution of days of the analogue Markov chain defined on the geopotential height at 500 mb field $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region, for different numbers of analogues K.

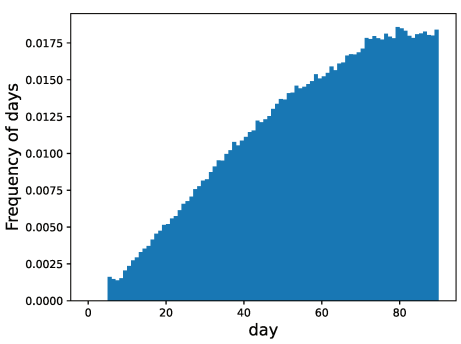
a systematic drift in the actual calendar day. This systematic temporal drift leads the chain to preferentially stay in states corresponding to the end of the summer. By increasing the number of analogues K this problem is partially mitigated but it nevertheless persists. Therefore, the systematic temporal drift prevents a good computation of the distribution of $A(t)$.

One source of the systematic temporal drift is that the seasonal analogue Markov chain always moves states forward in time. Indeed, a state s corre-

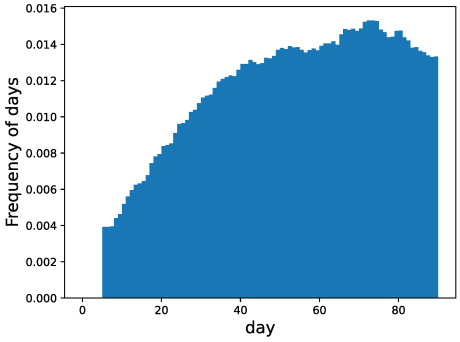
sponding to one of the first 5 days of June can only evolve towards states corresponding to days following June 5 (i.e. from June 6 onwards). Vice versa, a state s corresponding to the last 5 days of August cannot move forward more than 5 days and, on average, it evolves towards states corresponding to previous days.

One may wonder whether the systematic drift is exclusively due to the seasonal Markov chain approximation or if there are other sources such as the type of data used for the construction of the Markov chain. Therefore, an analogue Markov chain is built on the coarse-grained geopotential height at 500 mb *anomaly* field $\tilde{Z}^a(\mathbf{r},t) = \left(\tilde{Z} - \mathbb{E}_c\left[\tilde{Z}\right]\right)(\mathbf{r},t)$ in the North-Atlantic region (80°W − 30°E,30°N − 70°N). The analogues are searched by employing the Euclidean distance normalized to the average variance (Eq. (5.8)). Fig. 5.5 shows the invariant distribution of calendar days for this Markov chain. It can be noted that when only 5 analogues are considered (see Fig. 5.5a), the systematic drift is present, although it is less pronounced than in the previous case. However, as the number of analogues K increases, this problem tends to disappear and the distribution comes closer and closer to a uniform one (see Figs. 5.5b,5.5c). It means that the systematic drift is affected by the type of data used for the construction of the chain. Although the drift is present in both cases analyzed, it is much less pronounced when the Markov chain is built on geopotential height anomalies. Indeed, the ratio between the maximum and the minimum of the distribution is of order $O(1)$ when the Markov chain is built on geopotential height anomalies while it is of order $O(10)$ or more when the geopotential height (which has a strong seasonal cycle) is considered. Thus, it can be drown the conclusion that the seasonal Markov chain can be built only on variables whose seasonal cycle has been removed.

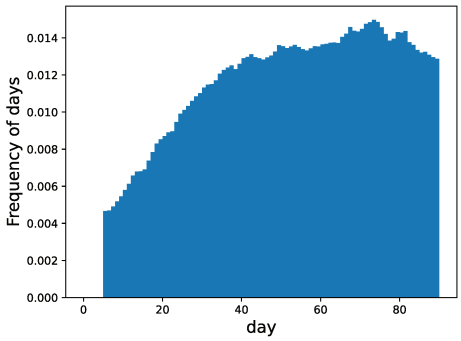
Having explained, at least partially, the reason for the systematic drift of the calendar day, one might ask whether the seasonal Markov chain built on geopotential height anomalies is able to reproduce the distribution of $A(t)$. The distributions of $A(t)$ for different numbers of analogues are shown in Fig. 5.6. It can be seen from the figure that although there is no perfect agreement, the discrepancies between the distributions are not too marked. However, by comparing Figs. 5.6a,5.6b,5.6c it can be noted that the distribution of $A(t)$ matches the real one less and less as the number of analogues K increases. Fig. 5.7 shows how the first two moments of the distribution vary as the number of analogues increases. It can be seen in Fig.5.7a that a



(a) $K = 5$ *analogues.*

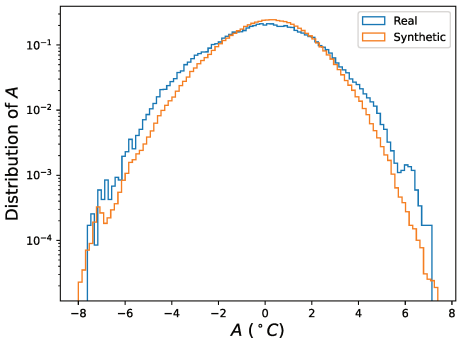


(b) $K = 20$ *analogues.*

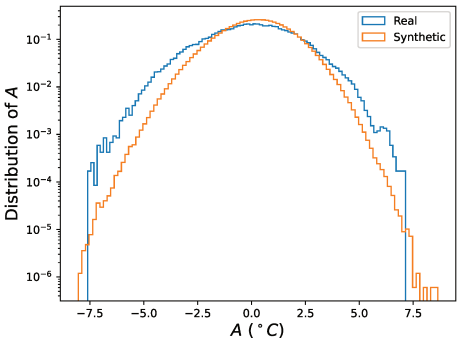


(c) $K = 50$ *analogues.*

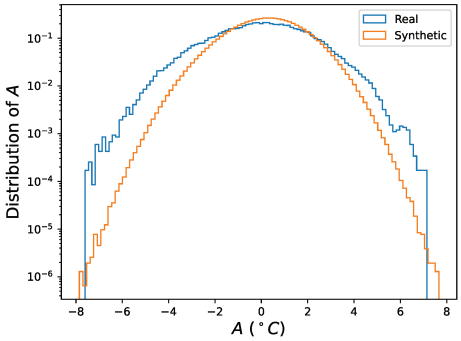
Figure 5.5: Stationary distribution of days of the analogue Markov chain defined on the geopotential height at 500 mb *anomaly* field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region, for different numbers of analogues K.



(a) $K = 5$ *analogues.*

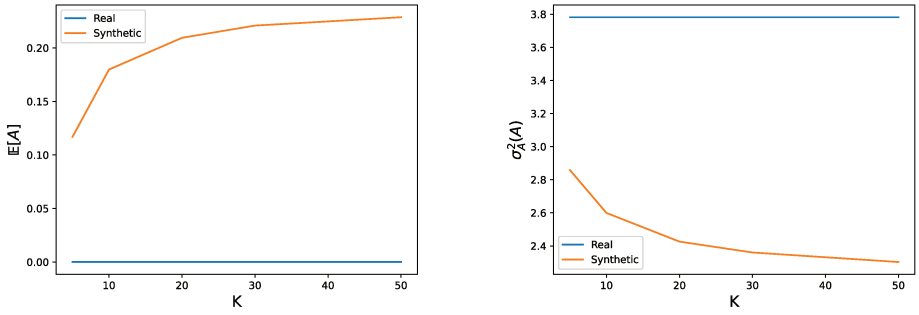


(b) $K = 20$ *analogues.*



(c) $K = 50$ *analogues.*

Figure 5.6: Comparison between $\rho_A(A)$ (real) and $\tilde{\rho}_A(A)$ (synthetic) for different numbers of analogues K. The analogue Markov chain has been defined on the geopotential height at 500 mb *anomaly* field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region.



(a) *Mean.* (b) *Variance.*

Figure 5.7: Mean $\mathbb{E}[A]$ and variance $\sigma_A^2(A)$ of the distribution of $A(t)$ as a function of the number of analogues K. The analogue Markov chain has been defined on the geopotential height at 500 mb anomaly field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region.

bias is also present in this case. However, the bias is about half of the one for the previous case, meaning that the bias is less important. The same consideration can be done for the variance (Fig. 5.7b). Indeed, there is still a reduction in variance of about 25%, which is not negligible. Furthermore, the increase in the number of analogues K produces further reductions in the variance, which goes from the value 2.8 (which corresponds to a reduction of about 25% with respect to the real one) to the value 2.5 (which corresponds to a reduction of about 35%). Therefore, although the problem of variance reduction also persists, it is much less pronounced than using fields with a strong seasonal cycle.

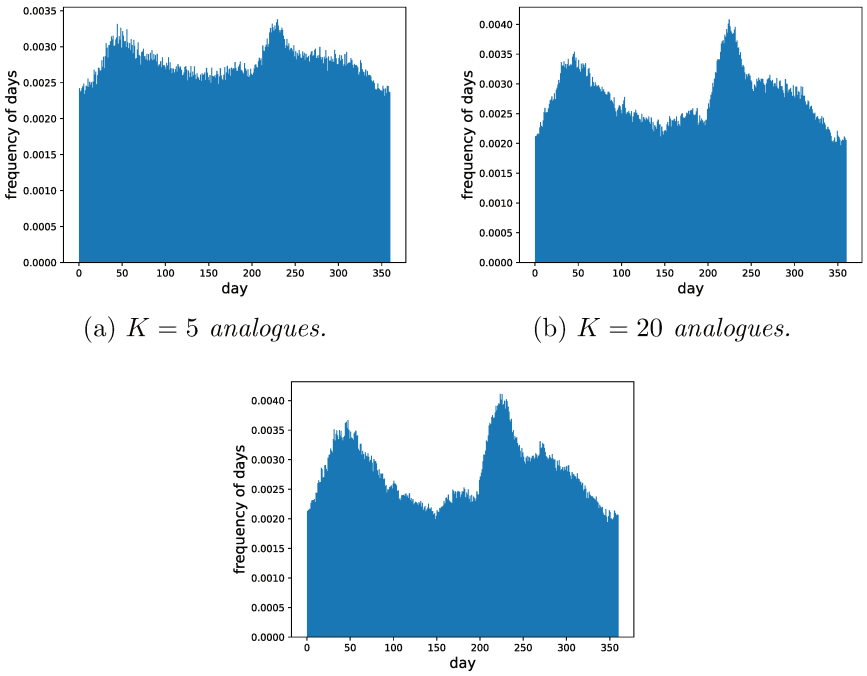
Note that, despite the problems mentioned above, these might not prevent a good computation of the committor function. In fact, the latter depends on the statistics of short trajectories while up to now the stationary properties of the chain have been analyzed. Thus, it can be concluded that seasonal analogue Markov chains built on variables that do not show a seasonal cycle can be employed in the study of extreme heat waves.

In the next subsections it will be explained how to construct Markov chains on variables that have a seasonal cycle. This will require to define annual Markov chains.

Consistency of time averaged temperature statistics: annual homogenous Markov chains

As mentioned earlier, seasonal Markov chains, because of the systematic drift, are not appropriate for dealing with the evolution of variables that have a seasonal cycle. This is the conceptual reason that leads to the introduction of annual Markov chains, which might not suffer from this problem. It is therefore important to understand if with such a chain it is possible to better reproduce the statistic of $A(t)$. Therefore, in this subsection the statistics of A on long synthetic trajectories generated by a homogeneous annual Markov chain are analyzed. The chain has been built on the coarse-grained geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region ($80°\mathrm{W} - 30°\mathrm{E}$,$30°\mathrm{N} - 70°\mathrm{N}$) and the analogues have been selected by employing the Euclidean distance normalized to the average variance (Eq. (5.8)).

Before analyzing the distributions of A as the number of analogues varies, it may be interesting to study the behavior of the stationary distributions of calendar days. They are shown in Fig. 5.8 for different numbers of analogues K. As can be seen from the figure, the distribution shows two peaks. The first one is in winter while the second one is in summer. Furthermore, by comparing Figs. 5.8a,5.8b,5.8c, it can be noted that the height of the peaks increases as the number of analogues increases. The explanation of these two peaks is linked to the presence of the seasonal cycle. In fact, winter and summer correspond respectively to the minimum and maximum of the seasonal cycle with a slowing down of the seasonal variations, and therefore it is easier to find analogues in these periods. This implies that a synthetic trajectory spends more time in states corresponding to these seasons. This also suggests that there may be problems in computing the distribution of A in the summer period. Indeed, the slowdown of the dynamics could make the duration of the year a random variable, preventing a good computation of the distribution of A. The distributions of A for different number of analogues are shown in Fig. 5.9. It can be seen from the figure that, regardless of the number of analogues considered, the Markov chain completely fails to reproduce the real distribution. Furthermore, by increasing the number of analogues the discrepancies are further accentuated. This is even more evident by looking at Fig. 5.10 which shows the first two moments of the distribution. Concerning the mean (Fig. 5.10a), this is not very far from the true value when only 5 analogues are considered. However, as the number of analogues increases, it grows further and further away from the real value.

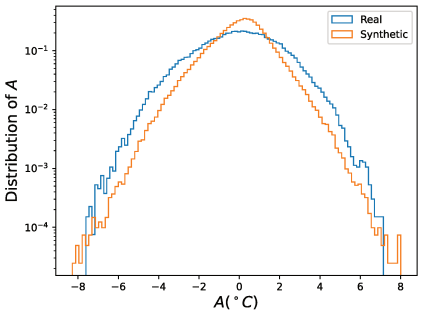


(a) $K = 5$ *analogues.*

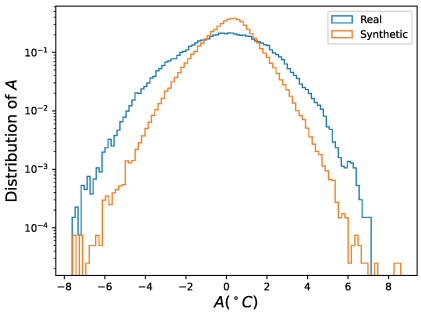
(b) $K = 20$ *analogues.*

(c) $K = 50$ *analogues.*

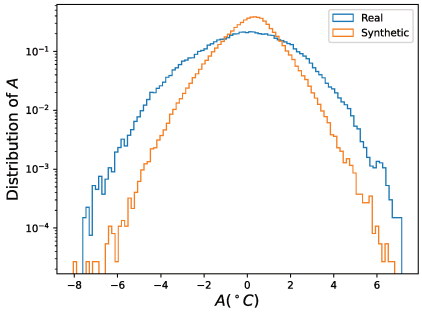
Figure 5.8: Stationary distribution of days of the annual homogeneous Markov chain defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region, for different numbers of analogues K.



(a) $K = 5$ *analogues.*

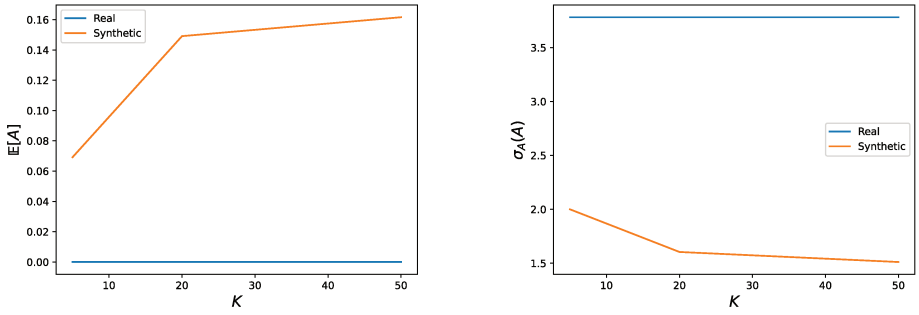


(b) $K = 20$ *analogues.*



(c) $K = 50$ *analogues.*

Figure 5.9: Comparison between $\rho_A(A)$ (real) and $\tilde{\rho}_A(A)$ (synthetic) for different numbers of analogues K. The homogeneous analogue Markov chain has been defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.



(a) *Mean.* (b) *Variance.*

Figure 5.10: Mean $\mathbb{E}[A]$ and variance $\sigma_A^2(A)$ of the distribution of $A(t)$ as a function of the number of analogues K. The homogeneous analogue Markov chain has been defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.

By looking at Fig. 5.10b, it should noted that even when considering only 5 analogues the variance is drastically reduced (it is about 50% smaller than the real value) and by increasing the number of analogues it continues to decrease further (up to a reduction of about 60%).

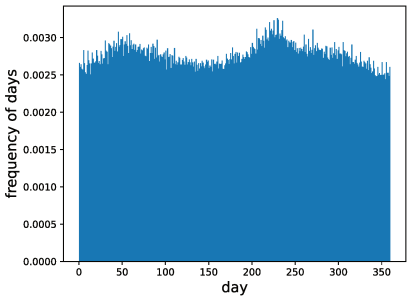
These results clearly show that the annual homogeneous Markov chain is not appropriate for the study of summer heat waves.

Consistency of time averaged temperature statistics: annual time-periodic Markov chains

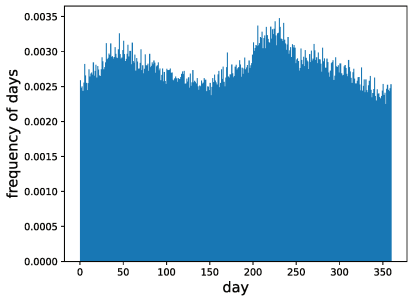
Given the failure of the annual homogeneous Markov chain to reproduce the statistic of A, the annual time-periodic Markov chain is now considered. Again, the Markov chain has been developed by considering the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region and by employing the Euclidean distance normalized to the average variance (Eq. (5.8)) for the analogue computation.

As in the previous case, the first analyzed quantity is the stationary distribution of calendar days, shown in Fig. 5.11 for different numbers of analogues K. Fig. 5.11 shows that the time periodic Markov chain does not completely remove the peaks in winter and summer, but in any case reduces the fluctuations of the distribution of calendar days. Furthermore, although increasing

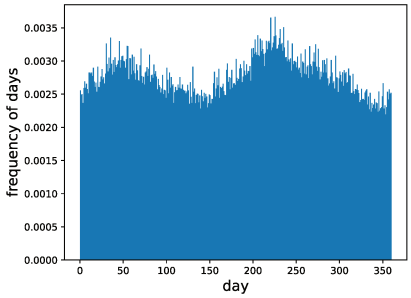
the number of analogues increases the height of the peaks, this is quite limited and certainly much smaller than the one observed with the homogeneous chain (the percentage differences between the peaks are about 15% for time periodic Markov chain and about 30% for the homogeneous one). This result is encouraging and suggests that time-periodic Markov chains do not suffer from the problems previously encountered in the computation of the distribution of A.



(a) $K = 5$ *analogues.*



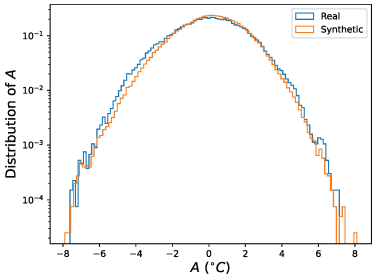
(b) $K = 10$ *analogues.*



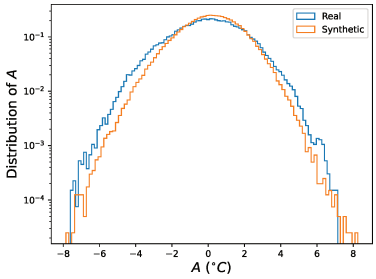
(c) $K = 20$ *analogues.*

Figure 5.11: Stationary distribution of days of the time-periodic analogue Markov chain defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region, for different numbers of analogues K.

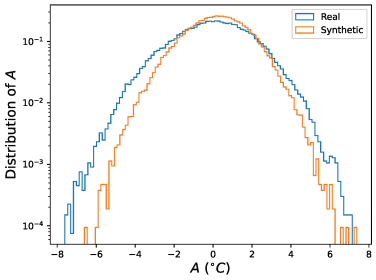
The distributions of A, for different number of analogues, are shown in Fig. 5.12. By looking at Fig. 5.12a, it can be noted that the distribution of A is very well reproduced when only 5 analogues are considered. Indeed, the synthetic distribution matches almost perfectly the real one. However, by looking at Figs. 5.12b,5.12c, it should be noted that the variance tends to decrease when the number of analogues increases. Fig. 5.13 clearly highlights



(a) $K = 5$ *analogues.*



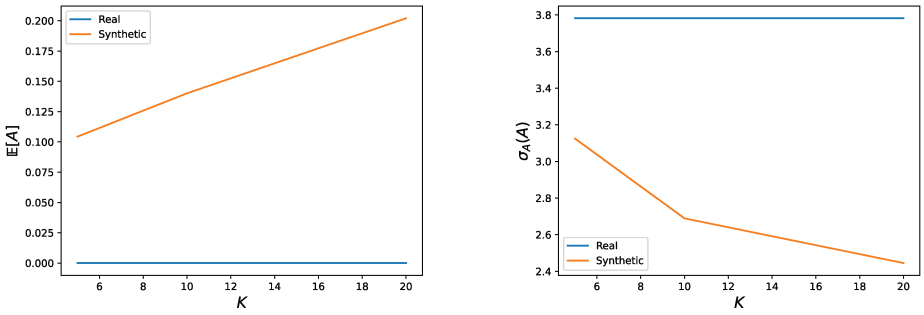
(b) $K = 10$ *analogues.*



(c) $K = 20$ *analogues.*

Figure 5.12: Comparison between $\rho_A(A)$ (real) and $\tilde{\rho}_A(A)$ (synthetic) for different numbers of analogues K. The time-periodic analogue Markov chain has been defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.

this behavior. Indeed, it can be seen (Fig. 5.13a) that the mean of the distri-



(a) *Mean.* (b) *Variance.*

Figure 5.13: Mean $\mathbb{E}[A]$ and variance $\sigma_A^2(A)$ of the distribution of $A(t)$ as a function of the number of analogues K. The time-periodic analogue Markov chain has been defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region.

bution increases when the number of analogues increases while the variance shows an opposite behavior, i.e. it is higher the fewer analogues are considered (see Fig. 5.13b). Despite the visual excellent agreement between the curves in Fig. 5.12, Fig. 5.13b shows that even for 5 analogues the variance of A is reduced by 15% with respect to the real one.

All things considered, it has been shown that by carefully selecting the number of analogues, the annual time-periodic Markov chain is able to reproduce the statistics of A. Indeed, the reduction of the variance when only 5 analogs are considered is about 15% which is considerably lower than that of the homogeneous Markov chain (reduction of about 50%) and also compared to that of the seasonal chain developed on the geopotential height anomaly field (reduction of about 25%). Furthermore, considering that even the distribution of calendar days is more uniform in this case (the percentage differences of the peaks are 15% for the annual time periodic chain and 30% for the annual homogeneous chain and the seasonal chain), it can be argued that the annual time periodic analogue Markov chain is the best tool to study the properties of the time averaged temperature A.

Consistency of return time plot

In the previous section, the consistency of typical fluctuations has been studied by analyzing the mean and the variance of the probability distribution of the time averaged temperature A computed by employing the analogue Markov chains. This section, instead, focuses on rare fluctuations. In particular, it aims to check the consistency between the return time plot computed from synthetic trajectories and the return time plot obtained from the real dynamics. Since the Markov chains that best reproduce the statistics of A are the seasonal analogue Markov chain built on the geopotential height anomaly field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region and the annual time-periodic Markov chain built on the geopotential height field $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region, these are the only chains considered in this section.

The return time of a heat wave of amplitude a is defined as the average time that elapses between the observation of two independent events with amplitude greater than a. It has been chosen to analyze the return times of summer maximum of the time averaged temperature $A(t)$. The summer maximum a_i of the i-th summer is defined as

$$a_i := \max_{t \in \text{summer } (i)} \{A(t)\}, \tag{5.15}$$

where $i = 1, \cdots, M$, and $M = 1000$ is the number of summers in the dataset. Note that with this definition one has 1 summer maximum per year. This choice is motivated by the fact that, in this way, independent events are analyzed, since the maximum of $A(t)$ in one summer is independent from that of another summer.

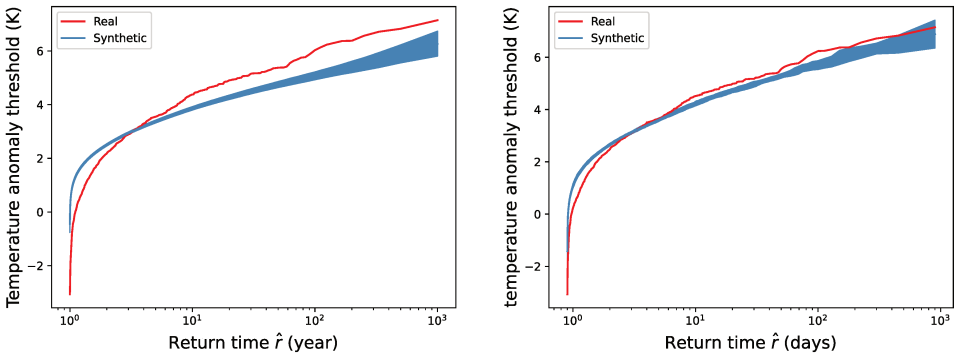
For estimating the return time plot, the sequence of summer maxima $\{a_i\}_{1 \leq i \leq M}$ has to be sorted in decreasing order. Let $\{\tilde{a}_i\}_{1 \leq i \leq M}$ be the ordered sequence. Finally, at each threshold $\tilde{a}_i$ it is possible to associate a return time

$$r(\tilde{a}_i) = \frac{M}{m} \text{ years}. \tag{5.16}$$

The quantity $r(a)$, known as return time or return period, is the average time between the occurrence of two heat waves with amplitude greater than a. At the same time, the r-year return level (or threshold) a^r can be defined as the value a^r such that a heat wave of amplitude greater than a^r is observed every r years on average.

Fig. 5.14 shows the comparison between the return time plot computed from the analogue Markov chains and the one computed using the real dy-

namics. The return time plots for the analogue Markov chain have been computed generating 10 trajectories, each 1000 years long, and the results shown in Figs. 5.14a,5.14b are the averages over the 10 realizations. The shaded areas show the empirical standard deviations computed using these 10 realizations. It can be seen from Fig. 5.14a that the Markov chain tends



(a) *Seasonal Markov chain.* (b) *Time-periodic Markov chain.*

Figure 5.14: Comparison between return time plots of 15-day heat waves computed from synthetic and real data. Left panel: return time plot computed using the seasonal summer Markov chain defined on the geopotential height at 500 mb anomaly field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region (20 analogues are used). Right panel: return time plot computed using the time-periodic Markov chain defined on the geopotential height at 500 mb field $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region (5 analogues are used).

to underestimate the return level for the rarer events (those with the longest return times). This result is consistent with the fact that the variance of the distribution of A is less than the one observed in the data, meaning that extreme events are slightly rarer in the Markov chain than in the real dynamics. Despite this, the discrepancies are not too large (the relative error between the return level estimated from the Markov chain dynamics and the return level estimated from the real one are about 10%) and the two curves are qualitatively in agreement. As regards Fig.5.14b, the same considerations can be made. In fact, it can be seen that also in this case the curve computed with the analogue method is slightly lower than the real one. However, unlike the previous case, the final part of the curves is in agreement within the statistical errors. This also could be expected as the variance of A for the

time-periodic chain is slightly greater than that of the seasonal chain.

These results show once again that the analogue method is capable of reproducing the statistics of A with a relatively good accuracy. Indeed, the return level can be estimated by using the analogue Markov chain with an error of about 10% for the seasonal analogue Markov chain or smaller (about 5%) for the annual time periodic Markov chain. It should be noted that the errors reported here refer to the errors in the estimation of the return levels for fixed return times. However, one might be interested in estimating the return times error for a given return level. In this case, it can be seen from Figs. 5.14a,5.14b that the errors in the estimation of the return times at certain return levels are much greater when the return times are estimated through the seasonal analogue Markov chain with respect to the case when the time-periodic analogue Markov chain is used. Therefore, it can be concluded that the time-periodic analogue Markov chain has to be preferred for the study of rare stationary fluctuations of the time averaged temperature A. Anyway, the analogue method is a promising tool for the study of extreme heat waves.

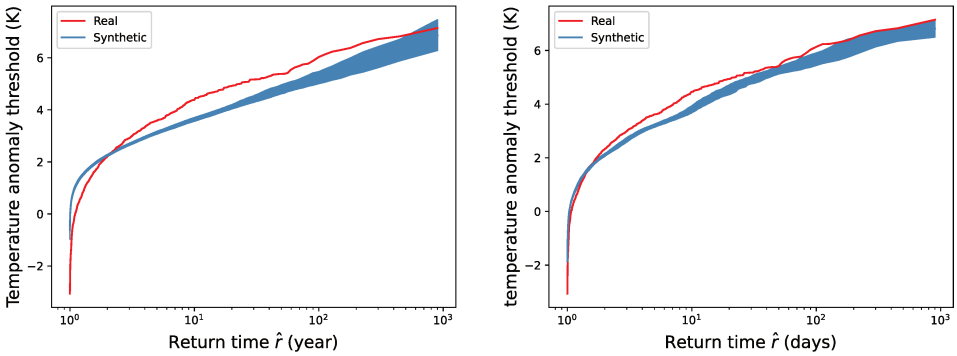
5.2.3 Extending return time plots

In the previous section it has been seen that the analogue method is able to reproduce some statistics of the original system (consistency checks). In particular, it has been shown that, under suitable conditions, it is possible to reproduce the probability distribution of A and the return time plot with an error of about 10% (when seasonal summer Markov chain was used) or about 5% (when the time-periodic analogue Markov chain was used).

The aim of this section is to understand whether it is possible to extend the return time plot for return times larger than the data-set length by using the analogue Markov chains beyond consistency. The idea is therefore to learn the dynamics of the Markov chain on a portion of the data and to use this analogue Markov chain to generate many synthetic summers. These synthetic summers are then used to estimate the return time plot. More precisely, the aim is to develop analogue Markov chains by using 100-year of the data for predicting return times plot up to 900 years. For this purpose, the Markov chains defined on the same variables used in the previous subsection have been considered. The only difference is that, in this case, instead of looking for analogues over all 1000 years available, they are sought in the first 100 years. Then, 10 synthetic trajectories of length 900 years have been

generated using the two different Markov chains and the return time plots have been drown.

The results are shown in Fig. 5.15, where the shaded areas represent the error bars equal to the standard deviations. It can be seen that, as for



(a) *Seasonal Markov chain.* (b) *Time-periodic Markov chain.*

Figure 5.15: Comparison between return time plots of 15-day heat waves computed from synthetic and real data. The Markov chains have been learned on 100 years only among the thousand availables. Left panel: return time plot computed using the seasonal Markov chain defined on the geopotential height at 500 mb anomaly field $\tilde{Z}^a(\mathbf{r}, t)$ in the North-Atlantic region (20 analogues are used). Right panel: return time plot computed using the time-periodic Markov chain defined on the geopotential height at 500 mb $\tilde{Z}(\mathbf{r}, t)$ in the North-Atlantic region (5 anlogues are used).

the consistency test, the seasonal Markov chain reproduces a curve which is qualitatively similar to the real one, but it is not accurate enough to be quantitatively in agreement (see Fig. 5.15a). Furthermore, the effect of the decrease in variance can still be seen as the curve is always below the real one. However, note that the two curves match for events with a return time of approximately 900 years.

The situation is different as regards the return time plot estimated with the time-periodic chain, shown in Fig. 5.15b. In this case the two curves are in agreement not only qualitatively but also quantitatively since there are several sections where they overlap within the errors. This is a remarkable result as it shows that the return time plot can be extended by a factor of 10. This opens up the possibility of studying return times for events that are

difficult to observe in climate models given their rarity. Moreover, given the low computational cost, it also allows to collect a statistics sufficient to make robust analysis, or in any case more robust than those possible only with direct simulations. It is worth stressing once again that to study the return times of heat waves it is preferable to use time-periodic analogue Markov chains.

In this section (Sec. 5.2) the statistically stationary statistics of 15-day heat waves were studied. The next section will instead focus on the transient statistics of these events, that is, the statistic computed on short trajectories.

5.3 Committor estimation through the analogue method

The tests performed in the previous section have proved the usefulness of the analogue method in the study of statistically stationary statistics of 15-day heat waves. Therefore, in this section the method is used in order to compute the committor function. According to the definition Eq. (5.4), the committor function $q_\tau(x; a)$ is the probability that a heat waves of duration T and amplitude greater than a will occur after τ days knowing that x is the state of the system at current time. In the context of analogue Markov chains, the state x corresponds to a state s of the chain. As in the previous section, 15-day heat waves are studied.

Before defining the committor function for Markov chains, it is important to note that this function only depends on the short trajectory statistic. Indeed, since $q_\tau(x; a)$ is a probability conditioned on the current state of the system, it only depends on the statistic of trajectory of length $T + \tau$. Therefore, there is not much difference between the committor computed by means of seasonal or time-periodic chains. Furthermore, the latter chains only add unnecessary complications as the committor function becomes a function not only of the initial state but also of the initial time. To avoid these further complications, exclusively the committor function estimated through the seasonal Markov chain will be discussed.

In Sec. 5.2.2 the quantity $\tilde{\rho}_A^{(s)}(A)$ has been introduced (Eq. (5.13)), which corresponds to the probability of having a heat waves of amplitude A and duration T knowing the initial state s of the system. Since for $\tau = 0$ the committor $q_0(s; a)$ is the probability of observing a heat wave with an amplitude

greater than a starting from the state $s_0 = s$, one has

$$q_0(s;a) = \int_a^{+\infty} \mathrm{d}A \tilde{\rho}_A^{(s)}(A) = \frac{1}{K^n} \int_a^{+\infty} \mathrm{d}A \sum_{k=1}^{K^n} \delta(A - A_k^s) = \frac{N_a(s)}{K^n}, \quad (5.17)$$

where $N_a(s)$ denotes the number of trajectories that start at $s_0 = 0$ and have $A > a$.

Once the committor function is computed at $\tau = 0$, it is easy to obtain the function at any time $\tau = m\tau_M$. Indeed, it is sufficient to multiply $q_0(s;a)$ by the power m of the transition matrix G, i.e.

$$q_\tau(s;a) = \sum_{s'} G_{ss'}^m q_0(s';a). \quad (5.18)$$

This formula is the discrete version of Eq. (2.105). It only says that the probability of observing a heat wave of duration T after τ days, knowing the state of the system s at the current time, is given by the sum over all states s' of the probability of having a heat wave knowing the state s' times the probability of going from s to s' in τ days.

In the following, three different thresholds will be considered $a_5 = 3.08\,\mathrm{K}$, $a_{2.5} = 3.7\,\mathrm{K}$, $a_{1.25} = 4.23\,\mathrm{K}$ which correspond to the 5%, 2.5% and 1.25% most extreme events of the time averaged temperature A or alternatively to events beyond the 95, 97.5 and 98.75 percentile.

5.3.1 Committor validation

To asses the quality of the estimates of the committor function, it is relevant to adopt a cross validation approach. Depending on the task one wants to focus on different score can be employed. In Sec. 2.4.5 two scores (the Brier score and the logarithmic score) for evaluating probabilistic predictability have been introduced. However, it has been explained that the Brier score is not appropriate for dealing with extreme events since it requires a huge number of test observations. Therefore, the Brier score can not be used in this context and the logarithmic score should be used. This applies to the task of carrying out probabilistic forecasts.

However, another interesting task is to build a quantifier, that is an object that, knowing the probability of occurrence of an event, predicts its occurrence or not. Such a quantifier can be very useful in decision-making

processes, where one needs to choose the most appropriate way to act. Furthermore, the performance of this object can be used to validate the estimate of the probability of occurrence of the event, that is, the estimate of the committor function.

Although both of these tasks are interesting, in what follows only the task of building a quantifier will be analyzed. Building a quantifier is equivalent to developing a classifier based on the committor function. It should be noted that predicting the occurrence of a heat wave through a classifier is equivalent to make a deterministic prediction. Although in this thesis it has been explained that it would be more appropriate to deal with probabilistic predictions, it has been chosen to use deterministic predictions mainly to compare the analogue method with a deep learning approach [Jacques-Dumas et al., 2021]. For lack of time it was not possible to study both deterministic and probabilistic predictions and therefore the validation of probabilistic predictions will be the subject of a future study. Thus, in the next subsection it will explained how to build a classifier based on the analogue method and how to test the performance of this classifier.

5.3.2 Classifier based on the analogue method

This section explains how to build a classifier based on the analogue method for predicting the occurrence of extreme heat waves. The advantage to develop a classifier based on the analogue method is that this classifier uses all the information contained in the dynamics.

A classifier can be thought as a deterministic function $C : \Omega \to \{0, 1\}$ where Ω is the phase-space of the system. Such a function takes a point $x \in \Omega$ as input and return a binary output $o \in \{0, 1\}$ that represents the class of the input (occurrence or not of the event). Over the years, different methods have been developed to accomplish this task, ranging from the simplest k-Nearest Neighbors and logistic regression, to more sophisticated methods such as random forests, vector support machines, and neural networks [MacKay and Mac Kay, 2003, Bishop, 2006, Theodoridis, 2015]. Unfortunately, all of these approaches face various difficulties with unbalanced data (i.e. the two classes are not equally populated) and therefore ad hoc modifications (such as undersampling or oversampling techniques) are required to obtain reasonable results [Krawczyk, 2016, Johnson and Khoshgoftaar, 2020, Jacques-Dumas et al., 2021].

In this section, an alternative approach is proposed which consists in

building a classifier based on the analogue method in order to extract the information contained in the committor. The scope of this classifier is to classify the system states according to whether in the next τ days a heat wave with an amplitude greater than a will occur or not. The basic idea is that the analogue method can be used to compute the committor function at points never seen before. Then a label 0 (no occurrence of the heat wave) or 1 (occurrence of the heat wave) can be associated to state x according to whether $q(x)$ is greater or less than a certain threshold value $\bar{q}$. A schematic illustration of such a procedure is shown in Fig. 5.16. Note that the step



Figure 5.16: A schematic illustration of the classifier based on the analogue method. The function takes a point of the phase space x as input and computes the committor function $q(x)$ using the analogue method. Then, it returns an output $o = 0$ if $q(x) < \bar{q}$ or $o = 1$ otherwise.

involving the analogue method is not strictly necessary for the construction of the classifier and can be replaced by any data-driven method that estimates the committor function.

The performance of the classifier is evaluated by employing the Matthews correlation coefficient (MCC) [Matthews, 1975]. The MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.19}$$

where TP is the number of positive events correctly classified, TN is the number of negative events correctly classified, while FP and FN represent the number of events in the two classes which are erroneously classified. The MCC takes values between -1 and 1 where $MCC = 1$ means that each event has been correctly classified, $MCC = -1$ that each event has been assigned to the wrong class, while $MCC = 0$ is associated with a random assignment. This score is considered very informative in the analysis of a binary classification, especially when the data are highly unbalanced, since it combines all the information contained in the confusion matrix (the

matrix formed by TN, FP, FN, TN) and is not sensitive to the choice of the class of positive events (as opposed to what happens with accuracy or F1 score) [Chicco, 2017, Chicco and Jurman, 2020]. Although the case where the MCC takes one of the values $\{-1, 0, 1\}$ is easily interpretable, the same is not valid for intermediate values. Hence, for a deeper understanding of the performance of the classifier it is useful to consider also the true positive rate (TPR) and the false detection rate (FDR)

$$TPR = \frac{TP}{TP + FN}, \quad FDR = \frac{FP}{TP + FP} \tag{5.20}$$

which measure the fraction of positive events correctly classified and the fraction of negative events erroneously classified as positive, respectively.

The analogue based classifier only depends on two hyperparameters (K and $\bar{q}$). The selection of these two parameters, which will be defined as the learning process, is based on the Monte Carlo cross-validation method [Dubitzky et al., 2007, Kuhn et al., 2013].

The learning process is performed as follows. Among the 1000 available trajectories, 100 are selected randomly. The set of 100 trajectories is denoted by $\mathcal{H}_{val}$ while the set of the remaining 900 trajectories is indicated by $\mathcal{H}_{learn}$. The analogue Markov chain is built on the set $\mathcal{H}_{learn}$. Then, by varying the number of analogues K, the committor function $q_K(x, \tau)$ for $x \in \mathcal{H}_{val}$ is computed. The subscript K is added to underline the dependence of q on the number of analogues considered. Finally, for each value of K and for each example x in the validation set $\mathcal{H}_{val}$, a set of binary output $o_K(x; \bar{q}(\tau))$ is computed. These outputs give rise to a set of MCC values denoted by $MCC(K, \bar{q}(\tau))$. By repeating this procedure $R = 20$ times, the average values $\langle MCC(K, \bar{q}(\tau))\rangle = \frac{1}{R}\sum_{r=1}^{R} MCC_r(K, \bar{q}(\tau))$ can be obtained and from these the optimal parameters $(K^\star, \bar{q}^\star(\tau))$ can be inferred, i.e.

$$(K^\star, \bar{q}^\star(\tau)) = \text{argmax}\{\langle MCC(K, \bar{q}(\tau))\rangle)\}. \tag{5.21}$$

Once the optimal hyperparameters have been learned, the performance of the classifier is tested by generating additional $R = 20$ test sets $\mathcal{H}_{test}$ and evaluating the average value of the MCC for $K = K^\star$ and $\bar{q}(\tau) = \bar{q}^\star(\tau)$.

In the following subsections the results of the predictions of the occurrence of heat waves, by increasing intensity of these events (5%, 2.5% and 1.25% most extreme events corresponding to events beyond the 95, 97.5 and 98.75 percentile), are illustrated. Furthermore, the impact of quantities such as the

coarse graining time, the size of the domain or the combination of different variables have on the quality of these predictions will be analyzed. Finally, a comparison between the analogue based classifier and one based on the committor function estimated with the k-Nearest Neighbors regressor will be performed.

Before starting to illustrate the results, it is convenient to briefly discuss the choice of the variables used for the predictions. Many of the analysis performed later will be done using the surface temperature anomaly field instead of the geopotential height anomaly field. This might seem contrary to what was stated in the introduction (Sec. 5.1) and with the analysis made in the previous section, where it was stated that the geopotential height is a key variable in the dynamics of heat waves. Although the importance of the geopotential height remains unquestionable, the ability of analogue Markov chains built on this variable to correctly predict the transient statistic of 15-day heat waves appeared limited, probably due to the lack of data. Therefore, it was decided to use the temperature anomaly field. From a physical point of view, this corresponds to trying to make predictions by exploiting the persistence of the temperature field.

Impact of temporal coarse-graining

The goal of this subsection is to answer the question: does the use of temporal coarse-graining improve the prediction skills of the classifier? For this purpose, two analogue Markov chains were constructed using two different variables, namely the surface temperature anomalies $\tilde{T}_s^a(\mathbf{r}, t)$ on the North Atlantic region (80°W − 30°E,30°N − 70°N), and the surface temperature anomalies averaged over France $\tilde{T}_F^a(t)$. For each of these variables, three coarse-graining times were used, namely $\tau_c = 1, 3, 5$ days.

It is important to note that, for prediction lag time $\tau = 0$, the coarse graining procedure always improves the quality of the forecasts. Indeed, the amounts of information about temperature at the beginning of the heat waves grows when the coarse graining time increases. However, this is only valid for a prediction lag time $\tau = 0$. For $\tau > 0$, there are no information about the temperature at the beginning of the heat wave and therefore it is not obvious to determine whether the coarse graining procedure improves or not the predictive skills of the analogue based classifier.

Fig. 5.17 shows the MCC obtained for the three classes of events as a function of the prediction lag time τ, for different coarse-graining times. The

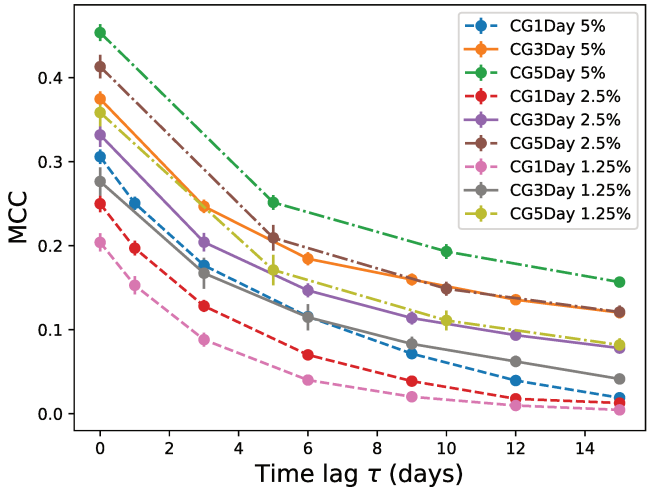


Figure 5.17: Comparison of MCC for different coarse graining times for the classifier built on the temperature anomaly $\tilde{T}_s^a(\mathbf{r}, t)$ in the North-Atlantic region.

dashed lines correspond to $\tau_c = 1$ day, the solid lines to $\tau_c = 3$ days and the dash-dotted lines corresponds to $\tau_c = 5$ days. Looking at the figure, several considerations can be made. The first is that the more extreme the event to be predicted, the worse the forecast is, regardless of the coarse-graining time. For $\tau_c = 1$ the classifier has modest predictive capabilities, which rapidly degrade increasing the prediction lag time. By increasing τ_c from 1 to 3 days, the predictive skills improve considerably (the MCC in the case of more extreme events for $\tau_c = 3$ days is almost equal to the case of less extreme events for $\tau_c = 1$ day). Furthermore, except for the 1.25% most extreme events, the classifier is able to predict events up to 15 days before their occurrence ($MCC > 0.1$), albeit with modest skill. Even more interesting is to look at the results obtained with $\tau_c = 5$. In this case, for all three classes of extreme events, the classifier is able to predict the occurrence of events 15 days in advance and, in the case of less extreme events, it is seen that the MCC is well above 0.1 ($MCC \sim 0.15$). These considerations lead to the conclusion that introducing a coarse-graining time greatly improves the predictive skills of the classifier. Note that the results are considered significant if the MCC is greater than 0.1. This threshold

was chosen arbitrarily and ensures that predictions are better than a random choice. Indeed, random classifications should have a $MCC = 0$ but the MCC depends on the data-set and therefore it fluctuates around 0 with typical fluctuations of the order 10^{-2}. Thus, a set of predictions is considered relevant if the MCC is at least an order of magnitude bigger than the fluctuations of the MCC for random predictions.

It seems natural to ask whether the behavior of the MCC when τ_c varies is linked or not to the fact that shorter trajectories were used to estimate the committor function (remember that the time step of the Markov chain τ_M is usually equal to the coarse graining time τ_c, i.e. $\tau_M = \tau_c$). Indeed, when $\tau_M = \tau_c$, as τ_c increases the length of the synthetic trajectories used to estimate the committor function decreases since the time distance τ_M between two successive states of the Markov chain increases as well.

To dispel this doubt, experiments were performed by building analogue Markov chains on the surface temperature anomalies averaged over France $\tilde{T}_F^a(t)$, varying the coarse-graining time τ_c but keeping $\tau_M = 1$ day. In this way the length of the synthetic trajectories used to estimate the committor function is constant for all three values of the coarse graining time τ_c. The results of these experiments are shown in Fig. 5.18. In this case, although the

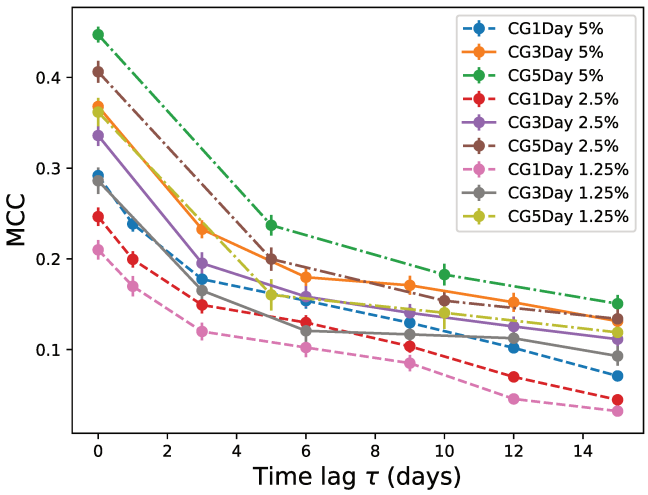


Figure 5.18: Comparison of MCC for different coarse graining times for the classifier built on $\tilde{T}_F^a(t)$.

analogue Markov chain is built on a simple scalar, the predictive abilities of the classifier are quite impressive. For all classes, it is possible to predict the occurrence of heat waves up to 15 days in advance (except for $\tau_c = 1$ where forecasts are possible up to 9 or 12 days in advance for classes corresponding to 1.25% and 2.5% more extreme events, respectively). The important thing to underline is that, once more, increasing the coarse-graining time leads to an increase in the predictive skills of the classifier.

Considering all the analysis shown it can be said that this property is valid in general, regardless of the variables used to construct the analogue Markov chain, although a clear motivation is missing. It could be hypothesized that this is due to a combination of two factors: the first is that the time averages eliminate statistical fluctuations from the variables while the second is that the evolution of these variables at time scales of the order of τ_c becomes more and more Markovian as τ_c grows, improving the agreement between the real dynamics and its approximation with a Markov chain. However, further verifications are necessary in order to confirm these hypotheses.

Impact of domain size: spatial coarse graining

The main question to be answered in this subsection is the following: does the domain on which the analogues are computed have an impact on the predictive capabilities of the classifier? To answer this question, it has been chosen to build 3 different analogue based classifiers using the surface temperature anomalies $\tilde{T}_s^a(\mathbf{r}, t)$ on 3 different regions, namely the Northern Hemisphere (180°W − 180°E,30°N − 90°N), the North-Atlantic region (80°W − 30°E,30°N − 70°N) and the France region (3°W − 5°E,42°N − 50°N). The coarse-graining time was taken equal to 3 days. Fig. 5.19 shows a snapshot of $\tilde{T}_s^a(\mathbf{r}, t)$ in the Northern-Hemisphere where the other two regions are highlighted by a black or red box respectively.

The results of the experiments performed using the three variables are shown in Fig. 5.20 where the dashed lines correspond to the Northern Hemisphere, the solid lines to the France region and the dash-dotted lines corresponds to the North-Atlantic region. Since the curves corresponding to the North-Atlantic region have been extensively discussed in the previous subsection, only those corresponding to the other two regions will be described.

Considering the Northern Hemisphere it can be seen that for the class of most extreme events the predictions are possible only at time $\tau = 0$ and

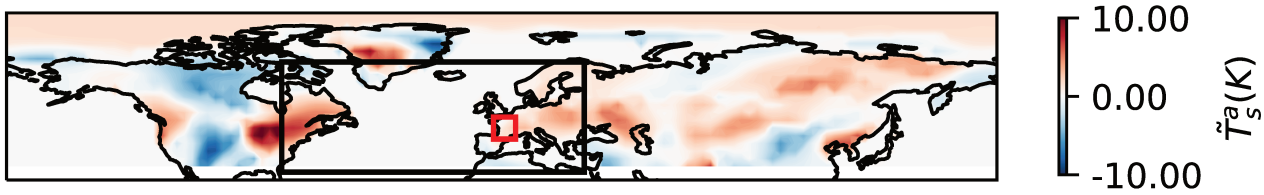


Figure 5.19: Snapshot of the surface temperature anomalies $\tilde{T}_s^a$ (i.e. fluctuations with respect to the climatological average) in the Northern Hemisphere (above $30°N$). The black box highlights the North Atlantic region while the France region is delimited by the red box.

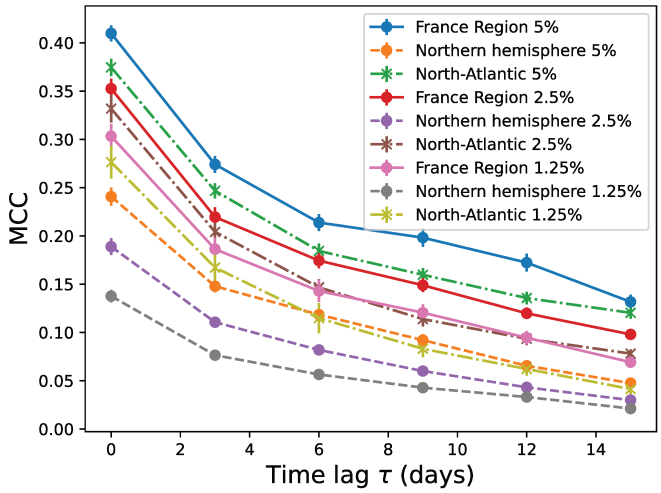


Figure 5.20: Comparison of MCC computed for the classifier built on the temperature anomaly $\tilde{T}_s^a(\mathbf{r}, t)$ in three different domains, i.e. the Northern Hemisphere, the North-Atlantic region and the France region. All the curves are realized using a 3-days coarse graining time.

for subsequent times the classifier does not perform better than a random choice (remember that a prediction is considered to be significant if the $MCC > 0.1$). The situation does not improve much for the intermediate class, where predictions are significant up to 6 days in advance. Regarding the 5% class of most extreme events, a significant fraction of heat waves can be predicted up to 12 days in advance. Despite this, considering the values assumed by the MCC, it can be stated that the analogue based classifier built using the temperature anomalies over such a large region does not provide accurate forecasts. The situation changes drastically when considering the small French region. In this case, for all three classes it is possible to make predictions up to 15 days in advance. In addition, it can be noted that for $\tau = 0$ the MCC assumes rather high values, thus highlighting how the persistence of heat waves can be used to make predictions. It is also important to note that the predictions provided by the classifier built on this small region are better than those obtained with the larger regions.

This analysis suggests that better predictions are obtained by reducing the region over which the analogues are computed. This is not surprising, as Euclidean distance is known to become a poor similarity criterion in high dimensions. It would therefore be tempting to say that the smaller the region, the better the prediction. However, comparing the solid blue curve of Fig. 5.20 and the solid orange curve of Fig. 5.18 it can be seen that the predictions made using the French region are better than those obtained based on the average temperature over France. It means that a compromise must be found between the information a variable carries with it and the dimension of the variable itself. To put it another way, a tradeoff between the complexity of the dynamics of the Markov chain and the information available is needed.

An interesting way to reduce the dimension of the system without losing too much information is to perform spatial coarse-graining. It is therefore important to know if this spatial coarse-graining can actually improve the predictive skills of the classifier. To achieve this results, it has been chosen to repeat the experiments by employing the Northern Hemisphere and its spatial coarse-grained counterpart. The spatial coarse-grained adopted here consists in an average over $6 = 3 \times 2$ grid points of the original field corresponding approximately to an average over 8.4° and 5.6° for longitude and latitude respectively. Two snapshots of the original field at the resolution provided by the climate model and its spatial coarse-grained counterpart are illustrated in Fig. 5.21.

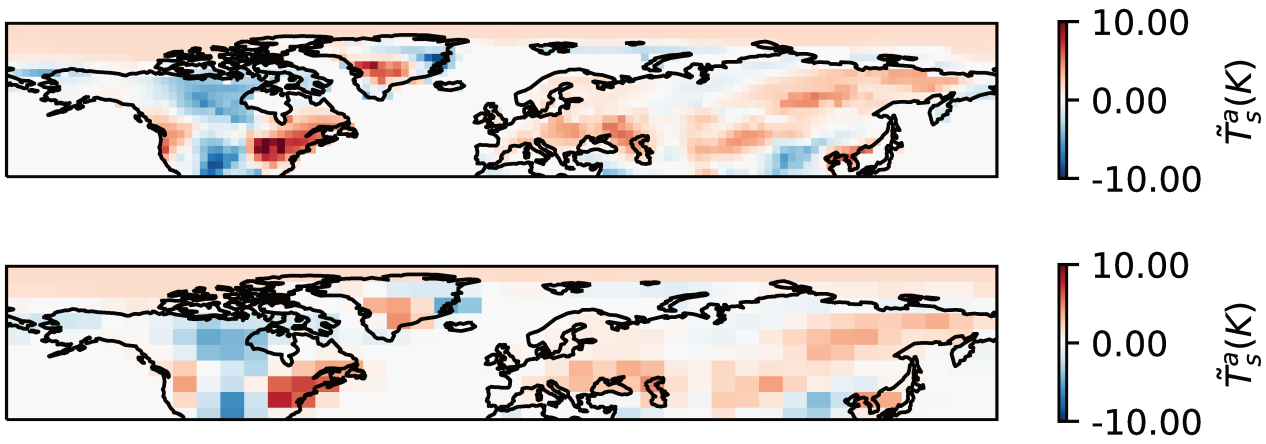


Figure 5.21: Snapshots of the surface temperature anomalies $\tilde{T}_s^a$ (i.e. fluctuations with respect to the climatological average) in the Northern Hemisphere (above $30°N$). The upper panel shows $\tilde{T}_s^a$ at the resolution provided by the climate model while the bottom panel shows its spatial coarse-grained counterpart.

Fig. 5.22 provides the results obtained using the surface temperature anomalies $\tilde{T}_s^a$ in the Northern Hemisphere with (solid lines) or without (dashed lines) spatial coarse-graining. It can be seen that, when the classifier is more accurate than a random one ($MCC > 0.1$), by reducing the dimension of the input variable results in higher predictive skills of the classifier. This clearly confirms the above statements and, moreover offers a method to improve the predictive skills of the classifier.

Variable selection

So far, the analogue Markov chain has been built using only one variable at a time. In this subsection, the predictive ability of the classifier is studied in the case where different variables are combined. Indeed, having in mind the discussion on the dimension of the system and the information contained in it, it is important to understand if using few variables which contain complementary information can improve heat waves forecasts. Therefore an analogue Markov chain built on a low-dimensional system is considered. In particular, the predictive skills of an analogue based classifier built on

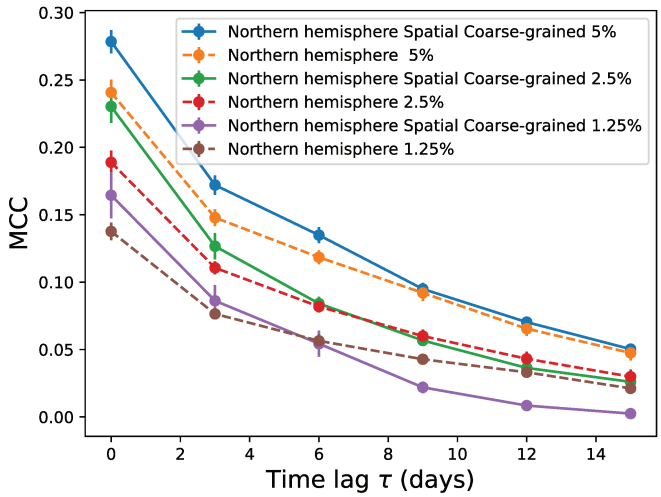


Figure 5.22: Comparison of the performance of the classifier based on the analogue Markov chain builded on the surface temperature anomalies of the northern hemisphere with or without spatial coarse-graining. A 3 days coarse-graining time is considered.

only three variables will be analyzed. The three variables considered are the surface temperature anomalies $\tilde{T}_F^a(t)$, the geopotential height anomalies $\tilde{Z}_F^a(t)$ and the soil moisture anomalies $\tilde{Sm}_F^a(t)$ averaged over France.

Fig. 5.23 shows the MCC obtained considering the analogue based classifier built on the low dimensional space. It can be seen that, for all the classes

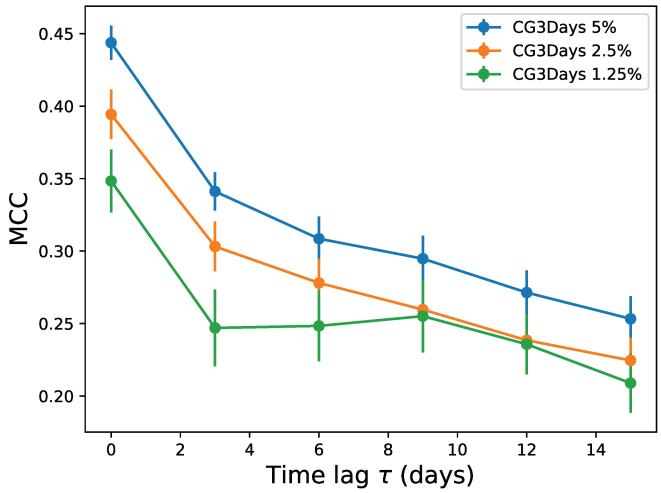


Figure 5.23: MCC computed from a Markov chain built on a low dimensional space. A 3-days coarse graining was used.

of extremes, the predictions are significant up to 15 days in advance. The values of the MCC at $\tau = 15$ days are quite high (ranging from 0.22 for the most extreme events up to 0.27 for the less extreme ones). In addition, with the exception of the transition from $\tau = 0$ to $\tau = 3$ days, the reduction of the MCC as τ varies is rather slow. This is probably related to the inclusion of the soil moisture, which evolves on slower time scales than the other two variables. However, it is surprising how the classifier built only on three variables provides such high MCC values, which compete if they do not beat the results obtained in [Jacques-Dumas et al., 2021] using a convolutional neural network (but different variables are employed there).

To better understand the performance of the classifier, however, it is useful to analyze not only the MCC but also the true positive rate (TPR) and the false detection rate (FDR). These quantities are shown in the left and right panel of Fig. 5.24, respectively. By looking at the left panel of Fig. 5.24,

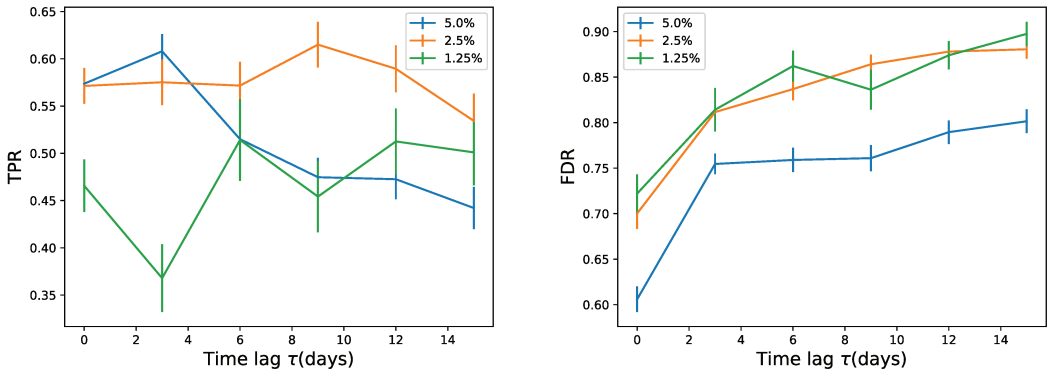


Figure 5.24: Rates of the analogue based classifier on a low dimensional space. Left panel: true positive rate (TPR). Right panel: false detection rate (FDR).

it can be inferred that the classifier correctly predicts the occurrence of about 50% of heat waves (being the TPR between 0.4 and 0.6). Furthermore, this number is roughly independent on the delay τ and on the heat wave class (although it is slightly lower for the most extreme class). Instead, the behavior of the FDR is different (right panel of Fig. 5.24) and helps to understand the behavior of the MCC. In fact, it should be noted that, for all three event classes, the FDR increases as τ increases. This indicates that as the prediction delay increases, more and more false alarms are recorded, explaining why the MCC value tends to decrease. It can also be noted that the more extreme the events, the more false alarms the classifier provides, thus explaining why the MCC is higher for less extreme event classes. For the 5% most extreme event class, the classifier predicts 4 correct events and 6 false alarms out of 10 predictions at time $\tau = 0$, while at $\tau = 15$ days out of four events it produces three false alarms and one correct detection. For the other two classes, at time $\tau = 0$ about 75 of the events are false detections and this number grows to 85 for $\tau = 15$ days.

These results highlight that although the MCC is quite high, the performances of the classifier are far from perfect. However, it should not be forgotten that the prediction of extreme events is an extremely complex task and therefore it is remarkable that about 50% of the events are correctly recorded 15 days in advance. Also consider that the predictive skills of the classifier on negative events (i.e. the non-occurrence of heat waves) are much

better (not shown) and contribute to the high value of the MCC.

All things considered, the analysis carried out in this subsection has shown that using a few very informative variables leads to good predictions. This further strengthens the statement made earlier on the need to find a compromise between the number of variables to use and the information these variables carry. Indeed, too many variables prevent a good computation of the analogues but, at the same time, as shown in this section, it is useful to combine the information coming from different variables.

Comparison between classifiers

To conclude the study of the analogue based classifier, it is shown that such a classifier performs better than one based on the committor function computed with a simpler approach, namely the k-Nearest Neighbors regressor. It was decided to perform this analysis using the low-dimensional system of the previous subsection. The reason for this choice lies on the one hand in the fact that it is one of the systems that provided the best forecasts, on the other hand because it is known that the k-Nearest Neighbors regressor performs better for low-dimensional systems.

The results obtained with the two classifiers are shown in Fig. 5.25, where the solid lines correspond to the analogue based classifier while the dash-dotted lines to the classifier based on the k-Nearest Neighbors regressor. From the figure it can be seen that the analogue based classifier outperforms the one based on the committor function estimated with the k-Nearest Neighbors method, although the latter shows good predictive abilities. Indeed, it should be noted that the curve of the MCC corresponding to the most extreme events class estimated from the analogue based classifier is almost always above the MCC corresponding to the less extreme events class estimated with the other method. This analysis is a further confirmation of the utility of the analogue method for estimating committor functions and predicting extreme events probabilities in climate systems.

5.4 Conclusions and perspectives

In this chapter the analogue method has been applied for studying seasonal extreme climate events, e.g. summer heat waves over France.

A heat wave of duration T and amplitude a has been defined as an event

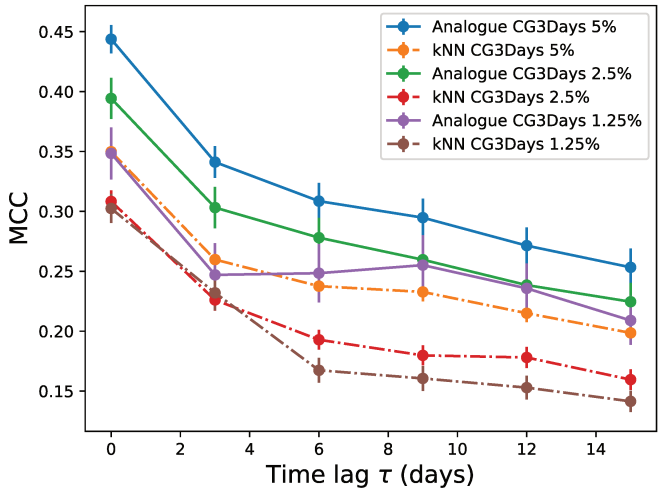


Figure 5.25: Comparison between the MCC computed using the analogue based classifier and the MCC computed using a classifier based on the estimation of the committor function provided by the k-Nearest Neighbors regressor.

such that the temporal average over period T of the surface temperature anomalies, averaged over the region of interest $\mathcal{D}$, is equal to a. In addition, the concept of a committor function for such events has been introduced. The committor encodes the probability that a heat wave of duration T and amplitude greater than a will occur after τ days, as a function of the current state of the system.

In order to study these phenomena, several analogue Markov chain have been developed. For each type of Markov chain, consistency checks were performed to ensure that the Markov chain approximation was able to correctly reproduce the stationary statistical properties of the events under investigation. These tests have shown that Markov chains learned on seasonal climate data provide meaningful results if the data does not exhibit a seasonal cycle while they fail in the other cases. It has also been shown that the use of data with a seasonal cycle is possible provided that annual time-periodic Markov chains are used. It can therefore be concluded that analogue Markov chains work quite well in the study of statistically stationary statistics of heat waves. Indeed, they reproduce well the distribution of the time averaged temperature A, although a reduction of the variance of about $25 - 35\%$ (seasonal analogue Markov chains) or about $15 - 35\%$ (annual time-period analogue Markov chains) is observed. Furthermore, they allows the computation of return times plots and estimate the return levels with an error of about 10% (seasonal analogue Markov chains) or about 5% (annual time-period analogue Markov chains).

The first remarkable result is to have shown that by using the learned dynamics, it is possible to extend the return times plot for return times larger than the data-set length. Indeed, it has been shown that by using 100-year of data it is possible to compute return times up to 900 years, i.e. return times 10 times bigger than the data-set length. This implies that the analogue method offers the possibility to compute return times at a much lower computational cost than traditional methods that solve a set of partial differential equations. In addition, it allows the collection of a much greater number of extreme events than those present in the data-set on which the dynamics is learned, opening up the possibility of carrying out significant statistical analysis.

After testing the usefulness of the analogue method, the analogue Markov chains have been used to compute the committor function. It has been explained that the choice of the type of analogue Markov chain does not affect much the computation of the committor and therefore it has been decided

to carry out the study using seasonal Markov chains. An analogue Markov chain based classifier has been built to assess the predictive informations encoded in the committor. The analogue based classifier is a tool that takes system variables as input and from these it asserts whether τ days later there will be a heat wave or not.

The analogue based classifier has been used to evaluate the impact of several parameters on the forecast of heat waves. It turns out that a temporal coarse-graining improves the performance of the classifier, likely because the coarse-graining procedure makes the temporal correlations less important and consequently the approximation of the dynamics with a Markov chain more realistic. Also the spatial domain on which analogues are computed affects the quality of the prediction. It has been shown that the smaller the domain the higher the performance of the classifier is. This is most likely due to the difficulty in finding good analogues in high-dimensional systems, especially using Euclidean distance as a similarity criterion, and therefore reducing the size of the system improves the quality of the analogues themselves. However, it has been noted that this is true as long as the fields in the smaller region contain sufficiently accurate information. For example, it has been shown that the analogue based classifier built on the surface temperature anomalies of France produces more accurate predictions than those obtained by building the classifier only on the average of these anomalies. This suggests that a compromise needs to be found between the dimension of the variables used as predictors and the information they hold. It has therefore been shown that spatial coarse-graining improves predictive abilities as it reduces the size of variables without discarding their information. To further strengthen the discussion between predictor dimensions and information in them it has been shown that an analogue based classifier built on only three very informative scalars, i.e. the surface temperature anomalies, the geopotential height anomalies and the soil moisture anomalies averaged over France, lead to excellent predictions. Finally, it has been shown that the analogue based classifier outperforms a more naive approach, that is a classifier based on the committor function estimated by a k-Nearest Neighbors regressor.

This work is conceived as a first step in understanding the analogue method applied to high-dimensional dynamical systems, such as the climate dynamics. Therefore, various modifications are possible to improve the coherence of the synthetic data produced by the analogue Markov chains, the quality of the estimation of the committor function and the predictive skills

of the analogue based classifier. For instance, following [Yiou, 2014, Jézéquel et al., 2018, Yiou and Déandréis, 2019, Yiou and Jézéquel, 2020], one could build analogue Markov chain where the where the probabilities of choosing one of the K analogues are not uniform as in this study but they are rather state dependent, for instance taking into account the distances of X_n to its analogues. Another improvement may consist in employing non-Euclidean distances as similarity criterion for the selection of analogues in order to take into account the spatial structure (i.e. spatial correlations) of the data. While both of these procedures can prove to be excellent ways to improve the performance of the analogue method by producing more accurate statistics and estimating the committor function more precisely, they can suffer from a lack of generality. Lack of generality is intended as a strong dependence on the data-set under investigation, thus making difficult to apply the same methodology to problems in different fields without having to perform long computations to recalibrate this tool. A more interesting modification of the algorithm proposed here would be to exploit the power of machine learning to perform a dimensional reduction of the system and to learn analogues in this reduced space. This can be done by employing autoencoders, as shown for example in [Kingma and Welling, 2013, Pulgar et al., 2018]. Furthermore, the dimensional reduction can be done keeping in mind that the ultimate goal is to estimate the probability of certain extreme events, and therefore extracting only the information relevant to the computation of such probabilities [Snoek et al., 2012, Banijamali and Ghodsi, 2016, Banijamali et al., 2018, Du et al., 2019]. This would allow to obtain a highly generalizable algorithm as it would be able to learn that contain most of the information useful for the computation of the committor function.

Chapter 6

Conclusions

This work was aimed at predicting the probability of climate extremes from dynamics and observations. More specifically, the work focused on the study of prediction problems at the predictability margin. These prediction problems are inherently probabilistic because the time scales on which they occur lie beyond the deterministic predictability time of the atmosphere, but it is nevertheless possible to make statistically significant predictions, depending on the current state of the system.

It has been explained that the mathematical object for dealing with prediction problems at the predictability margin is the committor function which encodes the probability that a given event occurs in the future as a function of the current state of the system.

In the first part of this thesis (Chapter 3), the committor function was studied in the context of a simple low-dimensional model proposed to explain the decadal amplitude changes of ENSO. Although this study was performed on a toy model for El-Niño, it allowed us to draw general conclusions valid for climate prediction problems. Based on the observation that the ability to predict the probability of occurrence of an event strongly differs depending on the initial state, a distinction can be traced between intrinsic probabilistic predictability (when the committor function is smooth and the probabilities do not depend sensitively from the initial conditions) and intrinsic probabilistic unpredictability (when the committor function depends sensitively on the initial conditions). This dichotomy between probabilistically predictable and unpredictable regions is expected to be a generic feature of prediction problems at the predictability margin. Therefore, it becomes crucial to develop a method that accurately estimates the committor function, especially in high

dimensional systems.

Estimating the committor function in high dimensional systems is an extremely complex task. Furthermore, when attention is paid to rare events, the problem of lack of observations also arises. To cope with the lack of data, rare event algorithms have been developed over the years, such as the family of splitting or cloning algorithms. To be efficient, those algorithms need to use a smart score functions during the selection stage and the optimal score functions are the committor functions. Since the committor function is precisely the object to be computed, it seems natural to consider an iterative approach where an algorithm is used to learn an estimate of the committor which is then used in the rare event algorithm to increase the number of samples improving the estimate of the committor function itself.

This iterative approach was developed in Chapter 4. A data-driven approach for estimating committor functions has been adopted. This approach relies on the analogue method used to define effective dynamics starting from dynamical observations only. It has been shown that this matches with the introduction of a Markov chain on the data so that its transition matrix approximates the true propagator of the system. Then, an estimate of the committor function has been obtained by employing classical methods for computing Markov chain committor functions, resulting in simple, robust, and efficient method. By means of two example, it has been established that the committor function can be accurately estimated from few observations. These approximations are more precise than those provided by a more naive data-driven approach and they converge faster to the exact committor when the number of observations increase. It was also pointed out that these improvements are due to the fact that the analogue method exploits all the dynamic information contained in a trajectory, while the direct approach treats the points of the same reactive trajectory equally. Finally, it has been proven that such learned approximate committor functions are extremely efficient score functions, when used with the Adaptive Multilevel Splitting algorithm. In this way, the rare events can be simulated with a minimal prior knowledge on the system and the results are much more precise than those obtained with a user-designed score function.

Finally, in Chapter 5 the analogue method has been applied to a complex climate data set in order to predict the probability of occurrence of heat waves. To this purpose, several types of analogue Markov chain have been introduced. It has been explained that it is possible to develop meaningful analogue Markov chains on seasonal data when the data do not exhibit a

strong seasonal cycle. In addition, it was explained that annual time-periodic analogue Markov chains should be considered when dealing with data that has a seasonal cycle. By using the dynamics introduced with the analogue Markov chains it was possible to extend the return time plot by an order of magnitude. Such a result is remarkable as it opens up the possibility of studying rare events by collecting a greater number of observations than traditional methods at lower computational cost. Then, the analogue Markov chain was used to compute the committor function. Starting from the latter, an analogue based classifier was developed with the aim of predicting the occurrence of heat waves several days ahead. The analogue based classifier was employed to assess the impact that different parameters, such as the coarse graining time and the spatial domain on which analogues are computed, have on the heat waves forecast. The main result is to have shown that there is a need for a compromise between the amount of information used for the prediction and the complexity of the effective dynamics introduced by the analogue Markov chain. Indeed, learning an analogue Markov chain on high dimensional data results in a poor estimate of the true propagator. At the same time, increasing the information available, especially if it comes from different variables, can greatly improve the performance of the classifier. Finally, it has been demonstrated that the analogue based classifier performs better than a classifier based on the committor function computed using a simpler approach.

Future developments are possible on several fronts. First of all, the study of the prediction of heat waves will be completed by analyzing and validating the probabilistic predictions that can be obtained by estimating the committor function through the analogue method. Then, regarding the analogue method, it is possible to develop analogue Markov chains where the probabilities of choosing one of the K analogues are not uniform but are rather state dependent, for instance taking into account the distances of X_n to its analogues. Furthermore, analogues could be selected by employing distances different from Euclidean ones or even by applying machine learning or other techniques to identify similarities between data. These modifications could be particularly relevant in the case of high dimensional systems because they could greatly improve the accuracy of the effective dynamics introduced by the analogue Markov chain. It would also be interesting to compare this method with other methods based on the use of dynamic information such as the Galerkin direct approximation method or with direct approaches that use machine learning. Regarding the coupling of the analogue method with

rare event algorithms, it would be useful to complete the study of the iterative procedure explained in Chapter 4. Indeed, it has been shown that the analogue method can provide an excellent score function for rare event algorithms, while it has not been tested whether the use of the data generated by these algorithms improves the approximation of the committor function itself or not. Finally, in light of the increasingly frequent catastrophic events occurring around the world, it would be very important to couple the analogue method and rare event algorithms to deepen the understanding of the dynamics and occurrence of rare or high impact climate events. Indeed, the committor function estimated through analogue Markov chains may prove to be a much more efficient score function for the rare event algorithms applied to climatic dynamics than the score functions used nowadays. This could therefore lead to the development of very efficient algorithms both as regards their computational cost and as regards the accuracy of the extreme trajectories produced by the rare event algorithms.

Bibliography

[Abbot et al., 2021] Abbot, D. S., Webber, R. J., Hadden, S., and Weare, J. (2021). Rare event sampling improves mercury instability statistics. *arXiv preprint arXiv:2106.09091*.

[Adams et al., 2008] Adams, D. A., Sander, L. M., and Ziff, R. M. (2008). Harmonic measure for percolation and ising clusters including rare events. *Physical review letters*, 101(14):144102.

[AghaKouchak et al., 2012] AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., and Sorooshian, S. (2012). *Extremes in a changing climate: detection, analysis and uncertainty*, volume 65. Springer Science & Business Media.

[Allen and Tildesley, 2017] Allen, M. P. and Tildesley, D. J. (2017). *Computer simulation of liquids*. Oxford university press.

[Altman, 1992] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[Arnold, 1974] Arnold, L. (1974). Stochastic differential equations. *New York*.

[Banijamali and Ghodsi, 2016] Banijamali, E. and Ghodsi, A. (2016). Semi-supervised representation learning based on probabilistic labeling. *arXiv preprint arXiv:1605.03072*.

[Banijamali et al., 2018] Banijamali, E., Karimi, A.-H., and Ghodsi, A. (2018). Deep variational sufficient dimensionality reduction. *arXiv preprint arXiv:1812.07641*.

[Bauer et al., 2015] Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47.

[Benedetti, 2010] Benedetti, R. (2010). Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211.

[Bishop, 2006] Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

[Bittracher et al., 2018] Bittracher, A., Koltai, P., Klus, S., Banisch, R., Dellnitz, M., and Schütte, C. (2018). Transition manifolds of complex metastable systems. *Journal of nonlinear science*, 28(2):471–512.

[Boffetta and Vulpiani, 2012] Boffetta, G. and Vulpiani, A. (2012). *Probabilità in Fisica: Un'introduzione*. Springer Science & Business Media.

[Bolhuis et al., 2002] Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002). Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, 53(1):291–318.

[Bouchet and Reygner, 2016] Bouchet, F. and Reygner, J. (2016). Generalisation of the Eyring-Kramers Transition Rate Formula to Irreversible Diffusion Processes. *ANNALES HENRI POINCARE*, 17(12):3499–3532.

[Bouchet et al., 2019] Bouchet, F., Rolland, J., and Simonnet, E. (2019). Rare event algorithm links transitions in turbulent flows with activated nucleations. *Physical review letters*, 122(7):074502.

[Bovier, 2006] Bovier, A. (2006). Lectures given at the 5th prague summer school on mathematical statistical physics, 2006.

[Bovier et al., 2004] Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M. (2004). Metastability in reversible diffusion processes i. sharp asymptotics for capcities and exit times.

[Bowman et al., 2013] Bowman, G. R., Pande, V. S., and Noé, F. (2013). *An introduction to Markov state models and their application to long timescale molecular simulation*, volume 797. Springer Science & Business Media.

[Bréhier, 2015] Bréhier, C.-E. (2015). Large deviations principle for the adaptive multilevel splitting algorithm in an idealized setting. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 12:717–742.

[Bréhier et al., 2016a] Bréhier, C.-E., Gazeau, M., Goudenège, L., Lelièvre, T., and Rousset, M. (2016a). Unbiasedness of some generalized adaptive multilevel splitting algorithms. *The Annals of Applied Probability*, 26(6):3559–3601.

[Bréhier et al., 2016b] Bréhier, C.-E., Goudenège, L., and Tudela, L. (2016b). Central limit theorem for adaptive multilevel splitting estimators in an idealized setting. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 245–260. Springer.

[Bréhier and Lelièvre, 2019] Bréhier, C.-E. and Lelièvre, T. (2019). On a new class of score functions to estimate tail probabilities of some stochastic processes with adaptive multilevel splitting. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(3):033126.

[Bréhier et al., 2015] Bréhier, C.-E., Lelièvre, T., and Rousset, M. (2015). Analysis of adaptive multilevel splitting algorithms in an idealized case. *ESAIM: Probability and Statistics*, 19:361–394.

[Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

[Bucklew, 2013] Bucklew, J. (2013). *Introduction to rare event simulation*. Springer Science & Business Media.

[Bugallo et al., 2017] Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79.

[Cassou et al., 2005] Cassou, C., Terray, L., and Phillips, A. S. (2005). Tropical atlantic influence on european heat waves. *Journal of climate*, 18(15):2805–2811.

[Castiglione et al., 2008] Castiglione, P., Falcioni, M., Lesne, A., and Vulpiani, A. (2008). *Chaos and coarse graining in statistical mechanics*. Cambridge University Press Cambridge.

[Cérou et al., 2019a] Cérou, F., Delyon, B., Guyader, A., and Rousset, M. (2019a). On the asymptotic normality of adaptive multilevel splitting. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):1–30.

[Cérou and Guyader, 2007] Cérou, F. and Guyader, A. (2007). Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443.

[Cérou et al., 2019b] Cérou, F., Guyader, A., and Rousset, M. (2019b). Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108.

[Charney and DeVore, 1979] 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.

[Chekroun et al., 2011] Chekroun, M. D., Kondrashov, D., and Ghil, M. (2011). Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *Proc. Natl. Acad. Sci. U.S.A.*, 108(29):11766.

[Chicco, 2017] Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1–17.

[Chicco and Jurman, 2020] Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

[Clarke, 2008] Clarke, A. J. (2008). *An introduction to the dynamics of El Niño & the Southern Oscillation*. Academic Press.

[Cobb et al., 2003] Cobb, K. M., Charles, C. D., Cheng, H., and Edwards, R. L. (2003). El Niño/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature*, 424(6946):271–276.

[Cobb et al., 2013] Cobb, K. M., Westphal, N., Sayani, H. R., Watson, J. T., Di Lorenzo, E., Cheng, H., Edwards, R., and Charles, C. D. (2013). Highly

variable El Niño–Southern Oscillation throughout the Holocene. *Science*, 339(6115):67–70.

[Coolidge, 1909] Coolidge, J. L. (1909). The gambler's ruin. *The Annals of Mathematics*, 10(4):181–192.

[Coumou and Rahmstorf, 2012] Coumou, D. and Rahmstorf, S. (2012). A decade of weather extremes. *Nat. Clim. Change*, 2(7):491.

[Crommelin et al., 2004] Crommelin, D. T., Opsteegh, J., and Verhulst, F. (2004). A mechanism for atmospheric regime behavior. *Journal of the atmospheric sciences*, 61(12):1406–1419.

[De Swart, 1989] De Swart, H. (1989). Analysis of a six-component atmospheric spectral model: Chaos, predictability and vacillation. *Physica D: Nonlinear Phenomena*, 36(3):222–234.

[Del Moral, 2012] Del Moral, P. (2012). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer Science & Business Media.

[Della-Marta et al., 2007] Della-Marta, P. M., Luterbacher, J., von Weissenfluh, H., Xoplaki, E., Brunet, M., and Wanner, H. (2007). Summer heat waves over western europe 1880–2003, their relationship to large-scale forcings and predictability. *Climate Dynamics*, 29(2):251–275.

[Dematteis et al., 2018] Dematteis, G., Grafke, T., and Vanden-Eijnden, E. (2018). Rogue waves and large deviations in deep sea. *Proceedings of the National Academy of Sciences*, 115(5):855–860.

[Devetsikiotis and Townsend, 1993] Devetsikiotis, M. and Townsend, J. K. (1993). Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM transactions on networking*, 1(3):293–305.

[Dijkstra, 2013] Dijkstra, H. A. (2013). *Nonlinear climate dynamics*. Cambridge University Press.

[Donovan et al., 2016] Donovan, R. M., Tapia, J.-J., Sullivan, D. P., Faeder, J. R., Murphy, R. F., Dittrich, M., and Zuckerman, D. M. (2016). Unbiased rare event sampling in spatial stochastic systems biology models

using a weighted ensemble of trajectories. *PLoS computational biology*, 12(2):e1004611.

[Du et al., 2019] Du, F., Zhang, J., Ji, N., Hu, J., and Zhang, C. (2019). Discriminative representation learning with supervised auto-encoder. *Neural Processing Letters*, 49(2):507–520.

[Dubitzky et al., 2007] Dubitzky, W., Granzow, M., and Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

[Dykman and Ryvkine, 2005] Dykman, M. I. and Ryvkine, D. (2005). Activated escape of periodically modulated systems. *Phys. Rev. Lett.*, 94(7):070602.

[D'Andrea et al., 2016] D'Andrea, F., Drobinski, P., and Stéfanon, M. (2016). European heat waves: the effect of soil moisture, vegetation, and land use. *Dynamics and predictability of large-scale, high-impact weather and climate events*, 2:185.

[E et al., 2005] E, W., Ren, W., and Vanden-Eijnden, E. (2005). Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chemical Physics Letters*, 413(1-3):242–247.

[Ebener et al., 2019] Ebener, L., Margazoglou, G., Friedrich, J., Biferale, L., and Grauer, R. (2019). Instanton based importance sampling for rare events in stochastic pdes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063102.

[Eyring, 1935] Eyring, H. (1935). The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115.

[Farkas, 1927] Farkas, L. (1927). The velocity of nucleus formation in supersaturated vapors. *Zeitschrift für Physikalische Chemie*, 125:236–242.

[Fedorov et al., 2006] Fedorov, A., Dekens, P., McCarthy, M., Ravelo, A., DeMenocal, P., Barreiro, M., Pacanowski, R., and Philander, S. (2006). The Pliocene paradox (mechanisms for a permanent El Niño). *Science*, 312(5779):1485–1489.

[Feng and Dijkstra, 2017] Feng, Q. Y. and Dijkstra, H. A. (2017). Climate network stability measures of El Niño variability. *Chaos*, 27(3):035801.

[Field et al., 2012] Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

[Finkel et al., 2020] 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 et al., 2021] Finkel, J., Webber, R. J., Abbot, D. S., Gerber, E. P., and Weare, J. (2021). Learning forecasts of rare stratospheric transitions from short simulations. *arXiv preprint arXiv:2102.07760*.

[Fraedrich et al., 2005] Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., and Lunkeit, F. (2005). The planet simulator: Towards a user friendly model. *Meteorologische Zeitschrift*, 14(3):299–304.

[Freidlin and Wentzell, 2012] Freidlin, M. I. and Wentzell, A. D. (2012). *Random Perturbations of Dynamical Systems*. Springer-Verlag, 3dr ed. New York.

[Froyland et al., 2014] Froyland, G., Gottwald, G. A., and Hammerlindl, A. (2014). A computational method to extract macroscopic variables and their dynamics in multiscale systems. *SIAM Journal on Applied Dynamical Systems*, 13(4):1816–1846.

[Galfi and Lucarini, 2021] Galfi, 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.

[Gardiner et al., 1985] Gardiner, C. W. et al. (1985). *Handbook of stochastic methods*, volume 3. springer Berlin.

[Garnett and Campbell, 1882] Garnett, W. and Campbell, L. (1882). *The Life of James Clerk Maxwell, with a Selection from His Correspondence*

and Occasional Writings and a Sketch of His Contributions to Science. Macmillan.

[Giannakis, 2019] Giannakis, D. (2019). Data-driven spectral decomposition and forecasting of ergodic dynamical systems. *Applied and Computational Harmonic Analysis*, 47(2):338–396.

[Giannakis et al., 2015] Giannakis, D., Slawinska, J., and Zhao, Z. (2015). Spatiotemporal feature extraction with data-driven koopman operators. In *Feature Extraction: Modern Questions and Challenges*, pages 103–115. PMLR.

[Givon et al., 2004] Givon, D., Kupferman, R., and Stuart, A. (2004). Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, 17(6):R55.

[Glasserman et al., 1998] Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. (1998). A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control*, 43(12):1666–1679.

[Grafke et al., 2013] Grafke, T., Grauer, R., and Schäfer, T. (2013). Instanton filtering for the stochastic burgers equation. *Journal of Physics A: Mathematical and Theoretical*, 46(6):062002.

[Grafke et al., 2015] Grafke, T., Grauer, R., and Schindel, S. (2015). Efficient computation of instantons for multi-dimensional turbulent flows with large scale forcing. *Communications in Computational Physics*, 18(3):577–592.

[Grafke et al., 2017] Grafke, T., Schäfer, T., and Vanden-Eijnden, E. (2017). Long term effects of small random perturbations on dynamical systems: Theoretical and computational tools. In *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, pages 17–55. Springer.

[Grafke and Vanden-Eijnden, 2019] Grafke, T. and Vanden-Eijnden, E. (2019). Numerical computation of rare events via large deviation theory. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063118.

[Graham, 1987] Graham, R. (1987). Macroscopic potentials, bifurcations and noise in dissipative systems. In *Fluctuations and Stochastic Phenomena in Condensed Matter*, pages 1–34. Springer.

[Guckenheimer et al., 2017] Guckenheimer, J., Timmermann, A., Dijkstra, H., and Roberts, A. (2017). (Un) predictability of strong El Niño events. *Dynamics and Statistics of the Climate System*, 2(1):dzx004.

[Guyader et al., 2011] Guyader, A., Hengartner, N., and Matzner-Løber, E. (2011). Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics & Optimization*, 64(2):171–196.

[Hartmann et al., 2014] Hartmann, C., Banisch, R., Sarich, M., Badowski, T., and Schütte, C. (2014). Characterization of rare events in molecular dynamics. *Entropy*, 16(1):350–376.

[Herbert and Bouchet, 2017] Herbert, C. and Bouchet, F. (2017). Predictability of escape for a stochastic saddle-node bifurcation: when rare events are typical. *Phys. Rev. E*, 96:030201(R).

[Herring et al., 2014] Herring, S. C., Hoerling, M. P., Peterson, T. C., and Stott, P. A. (2014). Explaining extreme events of 2013 from a climate perspective. *Bull. Am. Meteorol. Soc.*, 95(9):S1–S104.

[Horton et al., 2016] Horton, R. M., Mankin, J. S., Lesk, C., Coffel, E., and Raymond, C. (2016). A review of recent advances in research on extreme heat events. *Current Climate Change Reports*, 2(4):242–259.

[Hoskins and Woollings, 2015] Hoskins, B. and Woollings, T. (2015). Persistent extratropical regimes and climate extremes. *Current Climate Change Reports*, 1(3):115–124.

[Jacques-Dumas et al., 2021] Jacques-Dumas, V., Ragone, F., Bouchet, F., Borgnat, P., and Abry, P. (2021). Deep learning based extreme heatwave forecast. *arXiv preprint arXiv:2103.09743*.

[Jézéquel et al., 2018] Jézéquel, A., Yiou, P., and Radanovics, S. (2018). Role of circulation in european heatwaves using flow analogues. *Climate dynamics*, 50(3):1145–1159.

[Jin, 1997a] Jin, F.-F. (1997a). An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model. *J. Atmos. Sci.*, 54(7):811–829.

[Jin, 1997b] Jin, F.-F. (1997b). An equatorial ocean recharge paradigm for ENSO. Part II: A stripped-down coupled model. *J. Atmos. Sci.*, 54(7):830–847.

[Johnson and Khoshgoftaar, 2020] Johnson, J. M. and Khoshgoftaar, T. M. (2020). The effects of data sampling with deep learning and highly imbalanced big data. *Information Systems Frontiers*, 22(5):1113–1131.

[Kahn and Harris, 1951] Kahn, H. and Harris, T. E. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30.

[Kalnay, 2003] Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press.

[Khider et al., 2011] Khider, D., Stott, L., Emile-Geay, J., Thunell, R., and Hammond, D. (2011). Assessing el niño southern oscillation variability during the past millennium. *Paleoceanography*, 26(3).

[Khoo et al., 2019] Khoo, Y., Lu, J., and Ying, L. (2019). Solving for high-dimensional committor functions using artificial neural networks. *Research in the Mathematical Sciences*, 6(1):1–13.

[Kifer, 1992] Kifer, Y. (1992). Averaging in dynamical systems and large deviations. *Inventiones Mathematicae*, 110(1):337–370.

[Kifer, 2004] Kifer, Y. (2004). Averaging principle for fully coupled dynamical systems and large deviations. *Ergodic Theory and Dynamical Systems*, 24(03):847–871.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kornhuber et al., 2019] Kornhuber, K., Osprey, S., Coumou, D., Petri, S., Petoukhov, V., Rahmstorf, S., and Gray, L. (2019). Extreme weather events in early summer 2018 connected by a recurrent hemispheric wave-7 pattern. *Environmental Research Letters*, 14(5):054002.

[Kramers, 1940] Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304.

[Krawczyk, 2016] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

[Kuhn et al., 2013] Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

[Kullback, 1997] Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[Latif et al., 1994] Latif, M., Barnett, T. P., Cane, M. A., Flügel, M., Graham, N. E., Von Storch, H., Xu, J.-S., and Zebiak, S. E. (1994). A review of ENSO prediction studies. *Clim. Dyn.*, 9(4-5):167–179.

[Lau and Kim, 2012] Lau, W. K. and Kim, K.-M. (2012). The 2010 pakistan flood and russian heat wave: Teleconnection of hydrometeorological extremes. *Journal of Hydrometeorology*, 13(1):392–403.

[Laurie and Bouchet, 2015] Laurie, J. and Bouchet, F. (2015). Computation of rare transitions in the barotropic quasi-geostrophic equations. *New Journal of Physics*, 17(1):015009.

[Lestang et al., 2020] Lestang, T., Bouchet, F., and Lévêque, E. (2020). Numerical study of extreme mechanical force exerted by a turbulent flow on a bluff body by direct and rare-event sampling techniques. *Journal of Fluid Mechanics*, 895.

[Lestang et al., 2018] 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.

[Lguensat et al., 2017] Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., and Fablet, R. (2017). The analog data assimilation. *Monthly Weather Review*, 145(10):4093–4107.

[Li et al., 2019] 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.

[Lopes and Lelièvre, 2019] Lopes, L. J. and Lelièvre, T. (2019). Analysis of the adaptive multilevel splitting method on the isomerization of alanine dipeptide. *Journal of computational chemistry*, 40(11):1198–1208.

[Lorenz, 1972] Lorenz, E. (1972). *Predictability: does the flap of a butterfly's wing in Brazil set off a tornado in Texas?* na.

[Lorenz, 1963] Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141.

[Lorenz, 1969a] Lorenz, E. N. (1969a). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences*, 26(4):636–646.

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

[Lorenz, 1969c] Lorenz, E. N. (1969c). Three approaches to atmospheric predictability. *Bulletin of the American Meteorological Society*, 50(3454):349.

[Lucarini and Bódai, 2019] Lucarini, V. and Bódai, T. (2019). Transitions across melancholia states in a climate model: reconciling the deterministic and stochastic points of view. *Phys. Rev. Lett.*, 122(15):158701.

[Lucente et al., 2019] Lucente, D., Duffner, S., Herbert, C., Rolland, J., and Bouchet, F. (2019). Machine learning of committor functions for predicting high impact climate events. In Brajard, J., Charantonis, A., Chen, C., and Runge, J., editors, *Proceedings of the 9th International Workshop on Climate Informatics: CI 2019*. NCAR.

[Lucente et al., 2021] Lucente, D., Herbert, C., and Bouchet, F. (2021). Committor functions for climate phenomena at the predictability margin: The example of el niño southern oscillation in the jin and timmerman model. *arXiv preprint arXiv:2106.14990*.

[Ludescher et al., 2014] Ludescher, J., Gozolchiani, A., Bogachev, M. I., Bunde, A., Havlin, S., and Schellnhuber, H. J. (2014). Very early warning of next El Niño. *Proc. Natl Acad. Sci. USA*, 111(6):2064–2066.

[Lunkeit et al., 2011] Lunkeit, F., Blessing, S., Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., and Sielmann, F. (2011). Planet simulator. Technical report, Reference Manual. Technical Report, University of Hamburg. https://www. mi ?

[MacKay and Mac Kay, 2003] MacKay, D. J. and Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Mardt et al., 2018] Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):1–11.

[Massey et al., 2015] Massey, N., Jones, R., Otto, F., Aina, T., Wilson, S., Murphy, J., Hassell, D., Yamazaki, Y., and Allen, M. (2015). weather@ home?development and validation of a very large ensemble modelling system for probabilistic event attribution. *Quarterly Journal of the Royal Meteorological Society*, 141(690):1528–1545.

[Matthews, 1975] 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.

[McGregor et al., 2013] McGregor, S., Timmermann, A., England, M. H., Timm, O. E., and Wittenberg, A. T. (2013). Inferred changes in El Niño-Southern Oscillation variance over the past six centuries. *Clim. Past*, 9(5).

[McPhaden et al., 2015] McPhaden, M. J., Timmermann, A., Widlansky, M. J., Balmaseda, M. A., and Stockdale, T. N. (2015). The curious case of the el niño that never happened: a perspective from 40 years of progress in climate research and forecasting. *Bull. Amer. Meteor. Soc.*, 96(10):1647–1665.

[Melbourne and Stuart, 2011] Melbourne, I. and Stuart, A. (2011). A note on diffusion limits of chaotic skew-product flows. *Nonlinearity*, 24(4):1361.

[Metzner, 2008] Metzner, P. (2008). *Transition path theory for Markov processes*. PhD thesis, Doctoral thesis, Department of Mathematics and Computer Science, Freie Universität Berlin.

[Metzner et al., 2006] Metzner, P., Schütte, C., and Vanden-Eijnden, E. (2006). Illustration of transition path theory on a collection of simple examples. *The Journal of chemical physics*, 125(8):084110.

[Metzner et al., 2009] Metzner, P., Schütte, C., and Vanden-Eijnden, E. (2009). Transition path theory for markov jump processes. *Multiscale Modeling & Simulation*, 7(3):1192–1219.

[Miron et al., 2021] Miron, P., Beron-Vera, F., 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.

[Nemoto and Alexakis, 2018] Nemoto, T. and Alexakis, A. (2018). Method to measure efficiently rare fluctuations of turbulence intensity for turbulent-laminar transitions in pipe flows. *Physical Review E*, 97(2):022207.

[Nemoto et al., 2016] 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.

[Noé and Rosta, 2019] Noé, F. and Rosta, E. (2019). Markov models of molecular kinetics.

[Noé et al., 2009] 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.

[Nooteboom et al., 2018] Nooteboom, P. D., Feng, Q. Y., López, C., Hernández-García, E., and Dijkstra, H. A. (2018). Using Network Theory and Machine Learning to predict El Niño. *Earth Syst. Dynam.*, 9:969–983.

[Norris et al., 1998] Norris, J. R., Norris, J. R., and Norris, J. R. (1998). *Markov chains*. Number 2. Cambridge university press.

[Novikov, 1959] Novikov, E. (1959). On the problem of predictability of synoptic processes. *Izv. Acad. Sci. USSR, Geophys. Ser*, 11:1209–1211.

[Onsager, 1938] Onsager, L. (1938). Initial recombination of ions. *Physical Review*, 54(8):554.

[Pérez-Espigares and Hurtado, 2019] Pérez-Espigares, C. and Hurtado, P. I. (2019). Sampling rare events across dynamical phase transitions. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8):083106.

[Perkins, 2015] Perkins, S. E. (2015). A review on the scientific understanding of heatwaves—their measurement, driving mechanisms, and changes at the global scale. *Atmospheric Research*, 164:242–267.

[Philander, 1990] Philander, S. G. (1990). *El Niño, La Niña, and the Southern Oscillation*, volume 46 of *International geophysics series*. Academic Press.

[Platzer et al., 2021a] Platzer, P., Yiou, P., Naveau, P., Filipot, J.-F., Thiébaut, M., and Tandeo, P. (2021a). Probability distributions for analog-to-target distances. *arXiv preprint arXiv:2101.10640*.

[Platzer et al., 2021b] Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Zhen, Y., Ailliot, P., and Filipot, J.-F. (2021b). Using local dynamics to explain analog forecasting of chaotic systems. *Journal of the Atmospheric Sciences*.

[Plotkin et al., 2019] Plotkin, D. A., Webber, R. J., O'Neill, M. E., Weare, J., and Abbot, D. S. (2019). Maximizing simulated tropical cyclone intensity with action minimization. *Journal of Advances in Modeling Earth Systems*, 11(4):863–891.

[Poincaré, 1905] Poincaré, H. (1905). *Science and hypothesis*. Science Press.

[Pozun et al., 2012] Pozun, Z. D., Hansen, K., Sheppard, D., Rupp, M., Müller, K.-R., and Henkelman, G. (2012). Optimizing transition states via kernel-based machine learning. *The Journal of chemical physics*, 136(17):174101.

[Prinz et al., 2011] Prinz, J.-H., Held, M., Smith, J. C., and Noé, F. (2011). Efficient computation, sensitivity, and error analysis of committor probabilities for complex dynamical processes. *Multiscale Modeling & Simulation*, 9(2):545–567.

[Pulgar et al., 2018] Pulgar, F. J., Charte, F., Rivera, A. J., and Del Jesus, M. J. (2018). Aeknn: An autoencoder knn-based classifier with built-in dimensionality reduction. *arXiv preprint arXiv:1802.08465*.

[Ragone and Bouchet, 2020] 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 and Bouchet, 2021] 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 et al., 2018] 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.

[Redner, 2001] Redner, S. (2001). *A guide to first-passage processes*. Cambridge university press.

[Rickaby and Halloran, 2005] Rickaby, R. E. M. and Halloran, P. (2005). Cool La Niña during the warmth of the Pliocene? *Science*, 307(5717):1948–1952.

[Roberts, 2012] Roberts, A. (2012). Modify the improved euler scheme to integrate stochastic differential equations. *arXiv preprint arXiv:1210.0933*.

[Roberts et al., 2016] Roberts, A., Guckenheimer, J., Widiasih, E., Timmermann, A., and Jones, C. K. (2016). Mixed-mode oscillations of El Niño-Southern Oscillation. *J. Atmos. Sci.*, 73(4):1755–1766.

[Rolland, 2018] Rolland, J. (2018). Extremely rare collapse and build-up of turbulence in stochastic models of transitional wall flows. *Physical Review E*, 97(2):023109.

[Rolland, 2021] Rolland, J. (2021). Collapse of transitional wall turbulence captured using a rare events algorithm. *arXiv preprint arXiv:2103.16460*.

[Rolland et al., 2016] 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.

[Rolland and Simonnet, 2015] Rolland, J. and Simonnet, E. (2015). Statistical behaviour of adaptive multilevel splitting algorithms in simple models. *Journal of Computational Physics*, 283:541–558.

[Rosenbluth and Rosenbluth, 1955] Rosenbluth, M. N. and Rosenbluth, A. W. (1955). Monte carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23(2):356–359.

[Sarachik and Cane, 2010] Sarachik, E. S. and Cane, M. A. (2010). *The El Nino-Southern Oscillation phenomenon*. Cambridge University Press.

[Schütte et al., 1999] Schütte, C., Fischer, A., Huisinga, W., and Deuflhard, P. (1999). A direct approach to conformational dynamics based on hybrid monte carlo. *Journal of Computational Physics*, 151(1):146–168.

[Schütte et al., 2011] Schütte, C., Noé, F., Lu, J., Sarich, M., and Vanden-Eijnden, E. (2011). Markov state models based on milestoning. *J. Chem. Phys.*, 134(20):05B609.

[Schütte and Sarich, 2015] Schütte, C. and Sarich, M. (2015). A critical appraisal of markov state models. *Eur. Phys. J. Spec. Top.*, 224(12):2445–2462.

[Seneviratne et al., 2021] Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S. M., Wehner, M., Zhou, B., Chan, J., Sorteberg, A., and Vera, C. (2021). Weather and climate extreme events in a changing climate. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 11–1–11–345. IPCC, Cambridge University Press.

[Shiryaev, 2004] Shiryaev, A. (2004). Probability-1, probability-2. *MCCME, Moscow*.

[Shukla et al., 2016] Shukla, V., Fauve, S., and Brachet, M. (2016). Statistical theory of reversals in two-dimensional confined turbulent flows. *Phys. Rev. E*, 94(6):061101.

[Simonnet, 2016] Simonnet, E. (2016). Combinatorial analysis of the adaptive last particle method. *Statistics and Computing*, 26(1-2):211–230.

[Simonnet et al., 2021] 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.

[Slade, 2014] Slade, G. W. (2014). Markov chains for fun and profit: From gambler's ruin to phase locked loops.

[Smith et al., 1985] Smith, G. D., Smith, G. D., and Smith, G. D. S. (1985). *Numerical solution of partial differential equations: finite difference methods*. Oxford university press.

[Snoek et al., 2012] Snoek, J., Adams, R. P., and Larochelle, H. (2012). Nonparametric guidance of autoencoder representations using label information. *Journal of Machine Learning Research*.

[Stefanon et al., 2012] Stefanon, M., D'Andrea, F., and Drobinski, P. (2012). Heatwave classification over europe and the mediterranean region. *Environmental Research Letters*, 7(1):014023.

[Stocker, 2014] Stocker, T. (2014). *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge university press.

[Strahan et al., 2021] 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.

[Tailleur and Kurchan, 2007] Tailleur, J. and Kurchan, J. (2007). Probing rare physical trajectories with lyapunov weighted dynamics. *Nature Physics*, 3(3):203–207.

[Tantet et al., 2015] Tantet, A., van der Burgt, F. R., and Dijkstra, H. A. (2015). An early warning indicator for atmospheric blocking events using transfer operators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):036406.

[Theodoridis, 2015] Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic press.

[Thiede et al., 2019] Thiede, E. H., Giannakis, D., Dinner, A. R., and Weare, J. (2019). Galerkin approximation of dynamical quantities using trajectory data. *J. Chem. Phys.*, 150(24):244111.

[Thompson, 1957] Thompson, P. D. (1957). Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, 9(3):275–295.

[Timmermann and Jin, 2002] Timmermann, A. and Jin, F.-F. (2002). A nonlinear mechanism for decadal El Niño amplitude changes. *Geophys. Res. Lett.*, 29(1):3–1.

[Timmermann et al., 2003] Timmermann, A., Jin, F.-F., and Abshagen, J. (2003). A nonlinear theory for El Niño bursting. *J. Atmos. Sci.*, 60(1):152–165.

[Ulam, 2004] Ulam, S. M. (2004). *Problems in modern mathematics*. Courier Corporation.

[Van den Dool et al., 2007] Van den Dool, H., Cpc, P. S., and Van Den Dool, H. (2007). *Empirical methods in short-term climate prediction*. Oxford University Press on Demand.

[Van Kampen, 1992] Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*, volume 1. Elsevier.

[Vanden-Eijnden, 2006] Vanden-Eijnden, E. (2006). Transition path theory. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, pages 453–493. Springer.

[Vanden-Eijnden et al., 2010] Vanden-Eijnden, E. et al. (2010). Transition-path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, 61:391–420.

[Vanden-Eijnden and Venturoli, 2009] Vanden-Eijnden, E. and Venturoli, M. (2009). Markovian milestoning with voronoi tessellations. *The Journal of chemical physics*, 130(19):194101.

[Vautard et al., 2007] 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).

[Vulpiani et al., 2014] Vulpiani, A., Cecconi, F., Cencini, M., Puglisi, A., and Vergni, D. (2014). Large deviations in physics. *The Legacy of the Law of Large Numbers (Springer, 2014)*.

[Wang and Landau, 2001] Wang, F. and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050.

[Webber et al., 2019] 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.

[Weinan and Vanden-Eijnden, 2006] Weinan, E. and Vanden-Eijnden, E. (2006). Towards a theory of transition paths. *Journal of statistical physics*, 123(3):503–523.

[Wilks, 2011] Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

[Williams et al., 2015] Williams, M. O., Kevrekidis, I. G., and Rowley, C. W. (2015). A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346.

[Woillez and Bouchet, 2020] Woillez, E. and Bouchet, F. (2020). Instantons for the destabilization of the inner solar system. *Physical Review Letters*, 125(2):021101.

[Wouters and Bouchet, 2016] 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.

[Yiou, 2014] Yiou, P. (2014). Anawege: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7(2):531–543.

[Yiou and Déandréis, 2019] Yiou, P. and Déandréis, C. (2019). Stochastic ensemble climate forecast with an analogue model. *Geoscientific Model Development*, 12(2):723–734.

[Yiou and Jézéquel, 2020] Yiou, P. and Jézéquel, A. (2020). Simulation of extreme heat waves with empirical importance sampling. *Geoscientific Model Development*, 13(2):763–781.

[Yiou et al., 2013] 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(5-6):1333–1344.

[Zampieri et al., 2009] Zampieri, M., D'andrea, F., Vautard, R., Ciais, P., de Noblet-Ducoudré, N., and Yiou, P. (2009). Hot european summers and the role of soil moisture in the propagation of mediterranean drought. *Journal of Climate*, 22(18):4747–4758.