University of Reading

Climate on the Edge

Metastability, Melancholia States and Critical Transitions of the Ocean Circulation

Reyk Börner

Thesis submitted for the degree
Doctor of Philosophy in Mathematics

Department of Mathematics and Statistics
School of Mathematical, Physical and Computational Sciences

University of Reading

March 2025

Supervisors
Valerio Lucarini
Jeroen Wouters

"So, how close are we?"

Everyone I tell I'm studying
climate tipping points

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

I further confirm that I have not used any artificial intelligence tools such as large language models to assist with the writing of the text. I declare the use of ChatGPT as a coding and debugging assistant.

Reyk Börner

Final version with minor corrections and updated references (July 2025).

Online resources, data and software:
`https://reykboerner.github.io/projects/phd-thesis`

Abstract

Earth's climate is a nonlinear, multiscale complex system out of equilibrium. The variability of paleoclimate records reveals metastable dynamics, commonly associated with climate tipping elements – components of the Earth system susceptible to abrupt, self-perpetuating, and potentially irreversible shifts between distinct dynamical states. Under anthropogenic climate change, there is growing concern that the Atlantic Meridional Overturning Circulation (AMOC), a major proposed tipping element and vital climate component, could undergo a critical transition to a weakened or collapsed flow regime in the near future. Despite severe projected impacts for humanity and nature, the risk of an AMOC transition remains highly uncertain. Data and current methods are too limited to reliably estimate the proximity to potential tipping points.

This thesis seeks to improve our understanding of the processes and predictability of climate transitions based on dynamical systems theory. From conceptual models to an earth system model of intermediate complexity, we present a global view on the quasipotential landscape of the AMOC. Our approach combines concepts from stochastic and multiscale dynamics, large deviation theory, and chaos theory to understand the role of Melancholia (M) states – saddles located on the boundary between basins of competing attracting states (also called edge states). We explicitly compute M states in the climate models considered. From two to thousands of dimensions, we show how M states can give rise to complex dynamics and long chaotic transients. Connecting these findings with simulations under time-dependent forcing, we build a dynamical and physical understanding of metastability observed in state-of-the-art earth system models under future scenarios. We highlight fundamental limits to climate predictability and the standard use of asymptotic theories, proposing ways forward. Overall, the thesis contributes a step towards mapping out the stability landscape of the Earth system, which is essential for assessing the resilience of our planet and a safe operating space for humanity.

Contents

1 Introduction **6**
1.1 Metastability in the Earth system . . . 9
1.2 The Atlantic Meridional Overturning Circulation . . . 15
1.3 Climate predictability . . . 20
1.4 Research aims & thesis outline . . . 22

2 Theory, concepts & methods **25**
2.1 Elements of dynamical systems theory . . . 26
2.2 Numerical methods . . . 50

3 Stability landscape and tipping in an AMOC box model
Three coupled bath tubs **60**
3.1 Rooth's interhemispheric box model . . . 62
3.2 Stability and hysteresis under freshwater forcing . . . 64
3.3 Noise-induced transitions . . . 71
3.4 Discussion & conclusion . . . 78

4 Saddle avoidance in multiscale stochastic systems
Defying the weak-noise limit **80**
4.1 Multiscale stochastic systems . . . 82
4.2 Additive noise: FitzHugh-Nagumo model . . . 83
4.3 Multiplicative noise: Two competing species . . . 94
4.4 Reconciling the fast-slow and weak-noise limits . . . 99
4.5 Discussion & conclusion . . . 102

5 Predictability and chaos in a conceptual climate model
A study in five dimensions **104**
5.1 Stommel-L84: A conceptual ocean-atmosphere model . . . 106
5.2 Stability landscape of the AMOC . . . 109
5.3 Limits to predictability . . . 113
5.4 Overshoot tipping: Safe operating space? . . . 118
5.5 Discussion & conclusion . . . 120

6 Metastability of the AMOC in an intermediate-complexity climate model
A study in 10^5 dimensions **122**
6.1 PlaSim-LSG: Intermediate-complexity climate model . . . 124
6.2 AMOC stability and response . . . 126
6.3 Edge tracking: Constructing the Melancholia state . . . 130
6.4 Climate of the Melancholia state . . . 134
6.5 Boundary crisis: From Melancholia to ghost state . . . 141
6.6 Trajectories under future nonautonomous forcing . . . 145
6.7 Discussion & conclusion . . . 148

7 Synthesis & outlook **151**
7.1 Summary . . . 152
7.2 Research outcomes . . . 154
7.3 Concluding remarks . . . 158

A Supplemental figures **160**

Acknowledgments & data availablility **164**

References **167**

Terminology

2.1 Phase space . . . 27
2.2 Frozen system . . . 28
2.3 Attractor . . . 28
2.4 Basin of attraction . . . 29
2.5 Lyapunov exponent, Lyapunov spectrum . . . 30
2.6 Multistability . . . 31
2.7 Basin boundary . . . 32
2.8 Gradient system, Nongradient system . . . 32
2.9 Quasipotential I . . . 33
2.10 Saddle . . . 34
2.11 Melancholia state (Edge state) . . . 34
2.12 Invariant measure . . . 36
2.13 Quasipotential II . . . 36
2.14 Pathspace . . . 37
2.15 Freidlin-Wentzell action . . . 37
2.16 Instanton . . . 38
2.17 Freidlin-Wentzell quasipotential . . . 38
2.18 Arrhenius' law: Mean first-exit time . . . 39
2.19 Metastability . . . 42
2.20 Tipping point . . . 43
2.21 Bifurcation-induced tipping (B-tipping) . . . 44
2.22 Rate-induced tipping (R-tipping) . . . 45
2.23 Noise-induced tipping (N-tipping) . . . 45
2.24 Chaotic transient . . . 45
2.25 Boundary crisis . . . 46
2.26 Ghost state . . . 46
2.27 Onsager-Machlup action . . . 55
2.28 Geometric action . . . 57
3.1 Bifurcation: Saddle-node, Hopf, Homoclinic . . . 70
4.1 Transition path . . . 82

Chapter 1

Introduction

By the end of this century, will Reading be warmer on average than it is today? Yes, you might predict, given that the UK is currently warming at a rate of around 0.25°C per decade (Kendon et al. (2024), Fig. 1.1), consistent with the observed increase in global mean surface temperature since the 1900s (Masson-Delmotte et al. 2021). As greenhouse gas concentrations in the atmosphere keep rising due to human-caused emissions, global warming is projected to continue. But against the global trend, the UK and northern Europe could potentially experience a substantial cooling by the year 2100 (Liu et al. 2020). Why is that?

The answer lies in the ocean. In the North Atlantic, warm and salty water arrives from the south with a surface current that extends the Gulf Stream. As the salty water cools, it becomes denser than the water below, sinking to great depth (Kuhlbrodt et al. 2007). This deep-water formation of several million cubic meters of water every second acts like a pump that drives a large-scale ocean current called the *Atlantic Meridional Overturning Circulation* (AMOC). The AMOC transports vast amounts of heat northwards, corresponding to about fifty times the human energy consumption (Rahmstorf 2024), which is subsequently released to the atmosphere (Weijer et al. 2019). This heat maintains Europe's relatively warm climate compared to other places at the same latitude. Without the AMOC, average temperatures in Europe and the entire northern hemisphere would be significantly colder, regionally decreasing by up to 7 °C or more (Jackson et al. 2015; Liu et al. 2017; van Westen et al. 2024).

The AMOC is one of the proposed *climate tipping elements* – components of the Earth system believed susceptible to large-scale, abrupt and potentially irreversible shifts (Lenton et al. 2008). Such nonlinear changes are associated with *tipping points*, critical thresholds beyond which the system destabilizes and may transition to a qualitatively different state. Climate tipping elements are characterized by a combination of negative (stabilizing) and positive (destabilizing) feedbacks that give rise to the coexistence of multiple equilibrium states, or *multistability* (Feudel 2008). In the case of the AMOC, the main positive feedback is the *salt-advection feedback*, which makes the AMOC sensitive to salinity[1] changes in the North Atlantic controlled by freshwater transport and surface freshwater fluxes. Climate models robustly show that if the freshwater input becomes too large, the AMOC can tip into a state with a much weaker or collapsed flow (Manabe and Stouffer 1988; Marotzke and Willebrand 1991; Rahmstorf 1996; Rahmstorf et al. 2005; van Westen and Dijkstra 2023). This competing state coexists with today's vigorous state across a range of freshwater forcing strengths, establishing the AMOC as a multistable system (Stommel 1961; Broecker et al. 1985).

Paleoclimate records evidence that Earth's climate has experienced abrupt shifts in the

[1] Salinity measures the saltiness of water (typically in units of grams of salt per kilogram of water). Freshwater is unsalty water with a salinity close to zero.

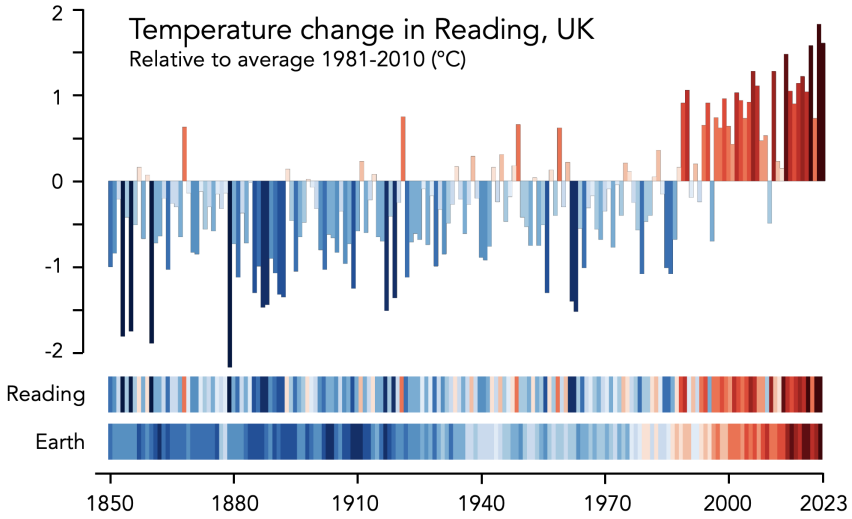

Figure 1.1 | **Yearly temperature anomaly in Reading, UK** from 1850 to 2023, relative to the period 1981-2010. Two lower bar graphs visualize the change as *climate stripes* for Reading and the globe, respectively. Adapted from Ed Hawkins, National Centre for Atmospheric Science, University of Reading, under CC-BY4.0 license. Data from the UK Met Office (`showyourstripes.info`).

past (Alley et al. 2003; Maslin 2009; Rousseau et al. 2023). For example, ice cores drilled to the base of the Greenland ice sheet reveal a fascinating history of climate variability spanning the past 250 000 years (Dansgaard et al. 1993). Oxygen isotopes in the ice of the last glacial period encode a series of sudden regime switches between warmer and colder episodes – a discovery that has profoundly changed our understanding of Earth's stability (Johnsen et al. 1992). It became clear that the climate system is *metastable*, meaning that the climate history features multiple long-lived yet ultimately transient dynamical regimes connected by critical transitions (Rousseau et al. 2023; Rossi et al. 2025). While the detailed mechanisms remain debated, it is widely accepted that the Atlantic ocean circulation played a major role in causing the climatic shifts recorded in the Greenland ice cores (Clark et al. 2002; Rahmstorf 2002; Weijer et al. 2019). Combining paleoclimate proxy data with model simulations strongly suggests that the AMOC is multistable and has acted as a tipping element in the past (Lynch-Stieglitz 2017), which raises the question whether the circulation is likely to "tip" again.

As a result of strong Arctic warming over the past years (Rantanen et al. 2022), the air-sea fluxes into the North Atlantic are currently changing. Warmer air temperatures imply a stronger hydrological cycle, leading to more precipitation and river runoff (Bintanja and Selten 2014). Arctic sea ice is in rapid decline (Stroeve et al. 2007), and the Greenland ice sheet is discharging an Olympic-size swimming pool into the ocean every three seconds (357 ± 58 Gt yr^{-1} meltwater runoff (Slater et al. 2021) plus 126 ± 9 Gt yr^{-1} ice discharge in the 2010s, Mouginot et al. (2019)). In addition to the increased freshwater flux, rising surface temperatures lead to an increased warming of the upper ocean, further stratifying the water column. These changes in salinity and temperature act to reduce the surface density, which counteracts North Atlantic deep-water formation. Climate model simulations consistently predict a decline of the AMOC as a response to global warming[2] (Weijer et al. 2020; Masson-Delmotte et al. 2021). Does this imply that the AMOC is approaching a tipping point beyond which it would abruptly and irreversibly collapse?

A substantial weakening or shutdown of the AMOC would have dramatic impacts (Jackson et al. 2015; Liu et al. 2020; Bellomo et al. 2023; Bellomo and Mehling 2024). Northern Europe could cool by several degrees Celsius (van Westen et al. 2024) while

[2]Many simulations under future scenarios show a weakening of the AMOC despite neglecting the contribution from Greenland melting (Swingedouw et al. 2022).

getting drier (Ritchie et al. 2020) and stormier (Bellomo et al. 2023). The tropical rain belt (Intertropical Convergence Zone, ITCZ) is expected to shift southward, strongly impacting the local rainfall distribution (Kang et al. 2008) and monsoon systems (Ben-Yami, Good, et al. 2024) that millions of people depend on. The circulation changes would further affect marine ecosystems and the oceanic uptake of carbon (Boot et al. 2024; Boot et al. 2025). These changes could play out within decades, challenging the capacity of society and nature to adapt to them (Loriani et al. 2023).

Thus, there is an urgent need to understand the evolution of the AMOC under future climate change, including the risk of a critical transition. The prospect of anticipating a forthcoming tipping event has spurred various research efforts to detect early warning signs in observational data (Feng et al. 2014; Boers 2021; Michel et al. 2022; Ditlevsen and Ditlevsen 2023; van Westen et al. 2024). At the same time, observations of the AMOC are sparse: continuous direct observations of the overturning circulation have only begun twenty years ago (Frajka-Williams et al. 2019), and the validity of fingerprints (proxy observables) reaching further back in time is debated (Ben-Yami, Morr, et al. 2024). The variability of the AMOC on annual to multidecadal timescales adds to the difficulty of detecting lasting trends, resulting in different assessments of whether the AMOC has recently weakened (Rahmstorf et al. 2015; Caesar et al. 2018; Jackson et al. 2022; Terhaar et al. 2025).

Due to limited observations, much of our knowledge of the ocean circulation relies on modeling studies. While multistability of the AMOC proves a consistent feature across models, large inter-model differences and model biases remain, making it difficult to quantify transition probabilities and the proximity to potential tipping points (Liu et al. 2017; Weijer et al. 2020; van Westen and Dijkstra 2024). Furthermore, AMOC studies in high-dimensional earth system models are limited by computational cost, requiring a trade-off between fidelity and simulation length. This calls for a concerted effort to draw from the entire climate model hierarchy – bridging the gap between simulation and understanding by linking simple and complex models (Held 2005).

Given the challenges of observing and modeling the AMOC, a mathematical approach towards metastability and tipping phenomena can help overcome key knowledge gaps (Ashwin et al. 2012; Feudel et al. 2018; Ghil and Lucarini 2020). The complexities of the Earth system bring standard mathematical methods to their limit, since many underlying assumptions do not necessarily hold in reality. Therefore, a synthesis between advanced applied mathematics and earth system science holds great potential to deliver answers to the pressing problems of the climate crisis (Lucarini and Chekroun 2023; Dijkstra 2013; Lucarini et al. 2014).

This thesis addresses the metastability and transition behavior of the climate system from a dynamical systems perspective, taking a global view on the phase space of climate models – from conceptual box models to a state-of-the-art earth system model. Our approach emphasizes dynamical and physical understanding to clarify what we can or cannot predict about our future. The unifying concept behind our investigations is that of an edge state, or *Melancholia state*: an unstable saddle state that tends to mediate metastable dynamics in the midst of competing climate equilibria. While we will focus on the AMOC as a tipping element, we will discuss how the concepts and findings are transferable to other metastable complex systems in the Earth system and beyond.

In this chapter, we provide a background in selected topics of Earth system science, with a particular focus on the AMOC. We illustrate the concepts of multistability,

metastability and tipping, which we define mathematically in chapter 2.2.4. Emphasizing the knowledge gaps that we seek to address leads us to formulating the main aims and research questions of this work (section 1.4.3). An introduction to the theory and mathematical concepts underlying this thesis follows in chapter 2.2.4.

1.1 Metastability in the Earth system

The Earth system comprises the atmosphere, hydrosphere (oceans, lakes, liquid water), cryosphere (ice), lithosphere (rocks, land masses) and biosphere (living organisms). Of course, there is also the human sphere with its infrastructure, technology, economy and social dynamics. In Earth system modeling, however, human activities are usually taken as an input and not modeled explicitly[3].

The components of the Earth system interact with each other, forming a nonlinear, chaotic, forced and dissipative complex dynamical system out of equilibrium (Peixóto and Oort 1984; Ghil and Lucarini 2020). For example, the turbulent flows in the atmosphere and ocean continuously try to balance the inhomogeneous distribution of incoming solar radiation between the equator and the poles, which varies over time (Milanković 1941). This redistribution of solar energy drives Earth's weather and climate, modulated by a plethora of processes from continental drift, vulcanism and the carbon cycle to, recently, human influences.

The spatiotemporal variability of weather and climate spans several orders of magnitude – from the nucleation of a cloud droplet to global ocean currents and from an afternoon cloud burst to cycles of ice ages and warm periods (von der Heydt et al. 2021). Altogether, the climate signal can be summarized in terms of a power spectrum, displaying local peaks corresponding to the periodicities of Earth's orbit as well as modes of internal variability (see Fig. 2 of von der Heydt et al. (2021)). Metastability produces a defining signal in the low-frequency range of the spectrum, shaped by critical transitions between competing climatic regimes. However, due to their irregular nature, they generally do not show up as peaks in the spectrum.

1.1.1 Multistability and tipping elements

In a simple calculation, Budyko (1969) and Sellers (1969) showed that the global ice cover can have multiple stable extents for a given magnitude of solar forcing. The reason for this multistability lies in the ice-albedo feedback: a larger ice cover reflects more solar radiation, thereby increasing Earth's reflectivity (albedo). The reduced heat absorption leads to a lower global mean temperature, favoring further ice growth. In the reverse direction, a reduction in ice cover self-amplifies by reducing the albedo, which increases temperatures, causing further ice retreat, in a positive feedback loop.

The positive ice-albedo feedback acts together with a negative thermal radiative feedback (Planck response): a warmer planet emits more thermal radiation to space than a colder planet, which stabilizes global temperatures. Based on these two processes, we can write a zero-dimensional energy balance equation for Earth, stating that at equilibrium (in a statistical sense) the incoming energy must balance the outgoing energy (North et al.

[3] If human activities such as the economy are included, this is known as integrated assessment modeling (Nordhaus 2013).

1981):

$$\frac{S_0}{4}(1-\alpha(T)) = \varepsilon\sigma_{\mathrm{SB}}T^4\,. \tag{1.1}$$

Here S_0 denotes the solar constant (the incident shortwave radiative flux), $0 < \alpha < 1$ the planetary albedo, T the global mean surface temperature, $0 < \varepsilon \leq 1$ the emissivity of Earth's surface, and σ_{SB} the Stefan-Boltzmann constant. The right-hand side describes the heat loss due to the Planck response, which depends nonlinearly on T. The left-hand side represents the incident net radiation, where the albedo depends on T owing to the ice-albedo feedback. If the dependence is nonlinear, e.g. $\alpha(T) \propto \tanh(T)$ to interpolate between fully ice-free and ice-covered conditions, there may exist a parameter range in which two temperature values yield a stable solution of Eq. (1.1) for given S_0.

Snowball Earth. The conclusion of this calculation is that two possible climates coexist for a range of solar forcing: a warm, largely ice-free world and a cold, frozen world. This has led to the "Snowball Earth" hypothesis, stating that our planet may have been completely ice-covered one or more times in the past (Harland 1964; Hoffman and Schrag 2002). Indeed, carbon isotope records dating back hundreds of millions of years hint at abrupt climate changes that would be explained by a global glaciation (Hoffman et al. 1998). Transitions between the Snowball and warm states could have been triggered by various processes, such as continental drift and volcanic activity. The bistability of Earth's global climate associated with these two states has been found and investigated in energy balance models (Bódai et al. 2015; Lucarini et al. 2022) as well as climate models of intermediate complexity (Lucarini and Bódai 2017; Kaszás et al. 2019; Margazoglou et al. 2021).

An important parameter in Eq. (1.1) is the emissivity ε, which depends on the chemical composition of the atmosphere. A smaller emissivity corresponds to a larger atmospheric attenuation coefficient m. Half a century ago, Sellers (1969) wrote, "A 3% decrease in m should be sufficient to put the globe on the brink of an ice age. Fortunately, because of the increasing carbon dioxide content of the atmosphere, m is more likely to increase than decrease. Hence, the global mean temperature should slowly rise due to this factor." Today, this anticipated temperature increase (Manabe and Wetherald 1967) is the major reason for concern of crossing potential tipping points (Armstrong McKay et al. 2022).

Climate tipping elements. Multistability has not only been proposed for the global energy budget, but also for various subcomponents of the Earth system, collectively termed *climate tipping elements* (Lenton et al. 2008). Due to the coexistence of different states, these subcomponents could undergo abrupt and possibly irreversible changes induced by small external changes or random internal variability (see section 2.1.7).

Besides the AMOC, potential Earth system tipping elements include the Greenland ice sheet, the West and East Antarctic ice sheets, the Amazon rainforest, the African summer monsoon and tropical coral reefs, among others (Lenton et al. 2023) (Fig. 1.2). Closely connected with the AMOC, oceanic deep convection in the Labrador and Irminger Seas has been identified as a possible separate tipping element (Sgubin et al. 2017).

While these systems differ greatly in their dynamical processes and characteristic timescales, they share the common property of positive feedbacks that may drive self-perpetuating change once a critical threshold is crossed. The assessments by Armstrong

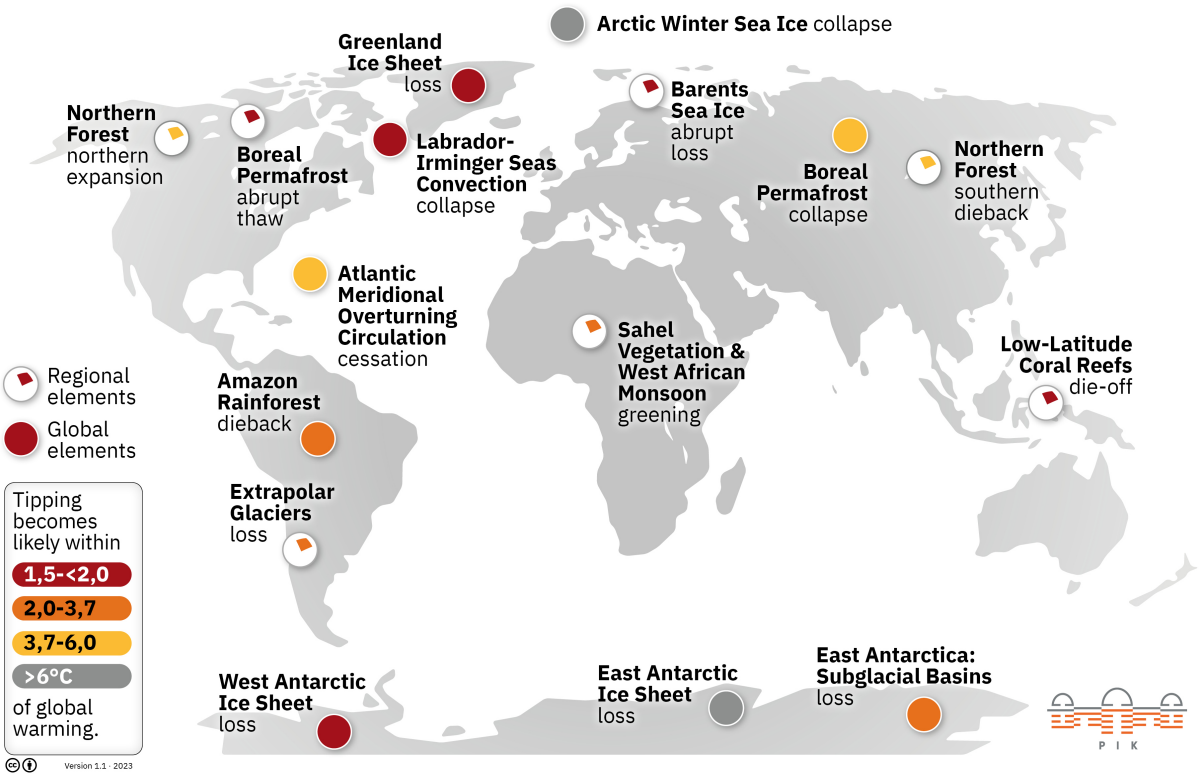

Figure 1.2 | **Map of proposed climate tipping elements** and their estimated critical temperature threshold ranges based on the assessment by Armstrong McKay et al. (2022). Figure designed at the Potsdam Institute for Climate Impact Research (PIK) under CC-BY license.

McKay et al. (2022) and Lenton et al. (2023) estimate such thresholds in terms of the magnitude of global warming, as indicated in Fig. 1.2. However, the temperature ranges are highly uncertain, and other variables than the global mean surface temperature may be more appropriate for determining potential thresholds of individual tipping elements.

Crossing a tipping point in any of the tipping elements would likely have severe regional and global impacts (Armstrong McKay et al. 2022). Additionally, the tipping elements are interconnected and could potentially trigger each other in a tipping cascade (Wunderling et al. 2024). For example, the AMOC interacts with the Greenland and Antarctic ice sheets (Sinet et al. 2023; Klose et al. 2024), as well as the Amazon rainforest (Akabane et al. 2024; Nian et al. 2023) and monsoon systems (Ben-Yami, Good, et al. 2024).

1.1.2 Metastable climate of the past

To understand our possible futures, it is informative to look into the past. Regionally, nature has revealed metastable behavior in recent decades, such as shifts in lake ecosystems (Scheffer et al. 1993). Large-scale Earth system components, on the other hand, typically evolve on longer timescales. Paleoclimate records from sediment cores, ice cores, speleothems or fossilized pollen tell stories of the prehistoric past, taking us up to millions of years back in time.

Based on isotope measurements of deep-sea benthic foraminifera (microfossils of single-cell organisms), the CENOGRID record gives insight into global temperatures over the past 66 million years (Westerhold et al. 2020). Four distinct climate states have been

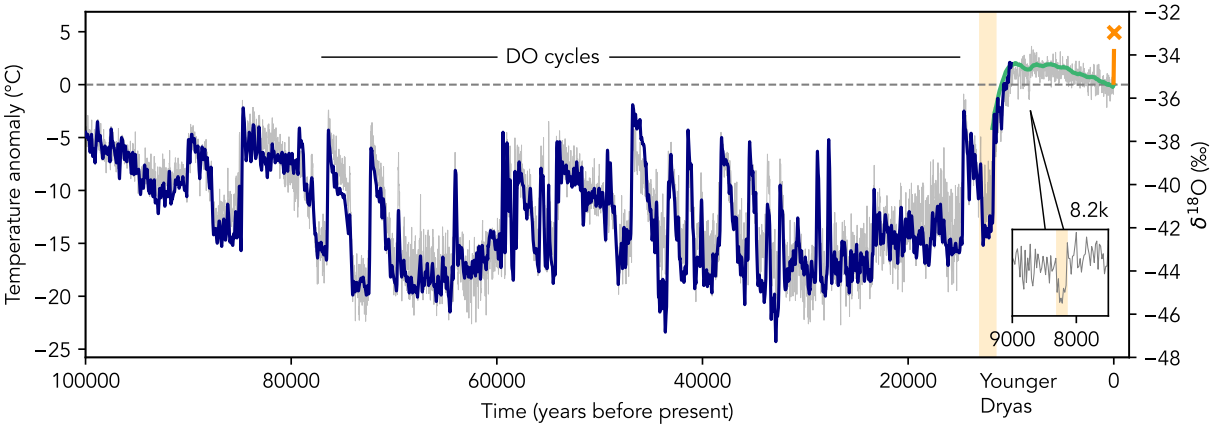

Figure 1.3 | **Arctic temperature record of the past 100 000 years**, displaying the Dansgaard-Oeschger (DO) cycles, the Younger Dryas events (yellow shading, main figure) and the 8k event (yellow shading, inset figure). The temperature scale shows anomalies relative to the 20th century C.E. Line colors represent different temperature reconstructions, based on δ^{18}O data of the NGRIP Greenland ice core (dark blue; Kindler et al. (2014)), a multi-proxy analysis by Kaufman et al. (2020) (green, taking the average of their General Additive Model (GAM) and Composite Plus Scale (CPS) methods for the northern-most latitude band), and a multi-product analysis by Rantanen et al. (2022) (orange; see also Masson-Delmotte et al. (2012)). The estimated Arctic temperature in the year 2050, given current trends, is marked (orange cross). For comparison, the δ^{18}O data of the NGRIP ice core is plotted in the background (gray) using the GICC05modelext dating (Wolff et al. 2010).

identified in the record, as well as five relatively abrupt transitions between them. While the climate variability in each state is paced by quasi-periodic orbital forcing, the nonlinear and state-dependent climate response to the forcing highlights that strong internal feedback processes drive the nonlinearities of the climate system.

As we move closer to the present, proxy data with increasing temporal resolution become available. In the aforementioned ice cores extracted from the Greenland ice sheet, oxygen isotopes in tiny frozen air bubbles encode the Arctic temperature conditions of the past 120,000 years (Kindler et al. 2014). Throughout the last ice age (around 115,000 to 11,000 years ago), Greenland temperatures underwent numerous abrupt changes, as captured in the dated δ^{18}O record of the NGRIP ice core (Wolff et al. 2010) and other records (Bond et al. 1993) (Fig. 1.3). The most famous of these transitions are the *Dansgaard-Oeschger* (DO) events – sudden warming events where local temperatures rose by up to 15 °C within around a century (Dansgaard et al. 1993). Each DO event is followed by a more gradual cooling, forming a so-called DO cycle.

Around 19,000 years ago, the summer insolation at northern latitudes began to increase owing to the orbital configuration, initiating a slow warming trend (Shakun and Carlson 2010). This warming was interrupted by a sudden cooling and subsequent final DO event (Younger Dryas period, around 12,900 to 11,000 ago), which terminated the last glacial period and led to a new climate epoch: the Holocene (Dansgaard et al. 1989; Clement and Peterson 2008; Mayewski et al. 2004). The climate of the Holocene has been remarkably stable compared to the last ice age, providing the conditions that supported human development until today (Rockström et al. 2009). Yet, also the Holocene has seen abrupt climate change, as exemplified by the so-called 8k event (Alley and Ágústsdóttir 2005) (see inset of Fig. 1.3).

The mechanisms of abrupt transitions in the Arctic climate of the last glacial period

remain difficult to decipher, both physically and dynamically. While many factors come into play, it is widely accepted that changes in the AMOC played a central role (Clark et al. 2002; Rahmstorf 2002; Lynch-Stieglitz 2017). Mathematically, DO events have been associated with various dynamical concepts, including bistability (Broecker et al. 1985) and noise-induced transitions (Ditlevsen and Johnsen 2010), relaxation oscillations (Broecker et al. 1990; Peltier and Vettoretti 2014; Vettoretti and Peltier 2018; Romé et al. 2025), excitability (Riechers et al. 2024; Timmermann et al. 2003), stochastic resonance (Alley et al. 2001; Ganopolski and Rahmstorf 2002; Cimatoribus et al. 2013), and intermittency (Gottwald 2021).

In summary, there is an abundant evidence of competing states and critical transitions in the past, which cannot be explained by external forcing alone but points towards nonlinear internal feedbacks giving rise to metastability.

1.1.3 Earth's stability landscape

Mathematically, we strive for a formal framework that can explain the complex dynamics of the Earth system. Dynamical systems theory provides a geometric description of the dynamics in terms of the *phase space*, i.e. the vector space containing all possible system states. To reason about the system's evolution in the phase space, the concept of a *stability landscape* has proven a powerful tool.

We give a mathematical introduction to stability landscapes in chapter 2.2.4. In nontechnical terms, the basic idea can be explained with the picture of a ball rolling in a hilly landscape[4]. Left alone, the ball will always roll downhill until it reaches the bottom of a valley, where it will stay indefinitely unless it is perturbed or the landscape changes. Local minima of the landscape thus correspond to stable states of the system, and "mountain crests" separate the valleys which represent different *basins of attraction*. The deeper a valley is, the more stable (or resilient to perturbations) it is, since we would need to push the ball harder (i.e. provide more energy) to make it escape from the valley. In this analogy, critical transitions occur when the ball jumps from one valley to another. The relative height of "mountain passes" between valleys thereby represents the energy barrier that must be overcome.

Stability landscapes are a popular concept in various scientific areas, from chemical reactions (Wales 2015; Ensing et al. 2006) to Waddington's epigenetic landscape in cell biology (Waddington 1957; Ferrell 2012). The idea links closely to energy potentials and the notion of free energy in statistical mechanics (Touchette 2009). Besides being intuitive, stability landscapes can be formally defined in terms of a *quasipotential* (see chapter 2.2.4), even for complex systems out of equilibrium. If the quasipotential is known, many global properties of the system follow directly, such as the most likely pathways of transitions. Stability landscapes thus offer a quantitative tool for studying metastability in complex systems (Zhou et al. 2012).

It is intriguing to think about what Earth's stability landscape looked like in the past, how it looks today, and how it might change in the future. Rousseau et al. (2023) recently proposed to understand the climate shifts in the CENOGRID record as motions in Earth's quasipotential landscape. Analyzing ice core data, Livina et al. (2010) used the idea of a changing potential landscape to infer global stability changes. In simplified climate models, the quasipotential has been explored for the multistability in global mean

[4]For this analogy to work, the ball must be an overdamped particle, i.e. subjected to high friction.

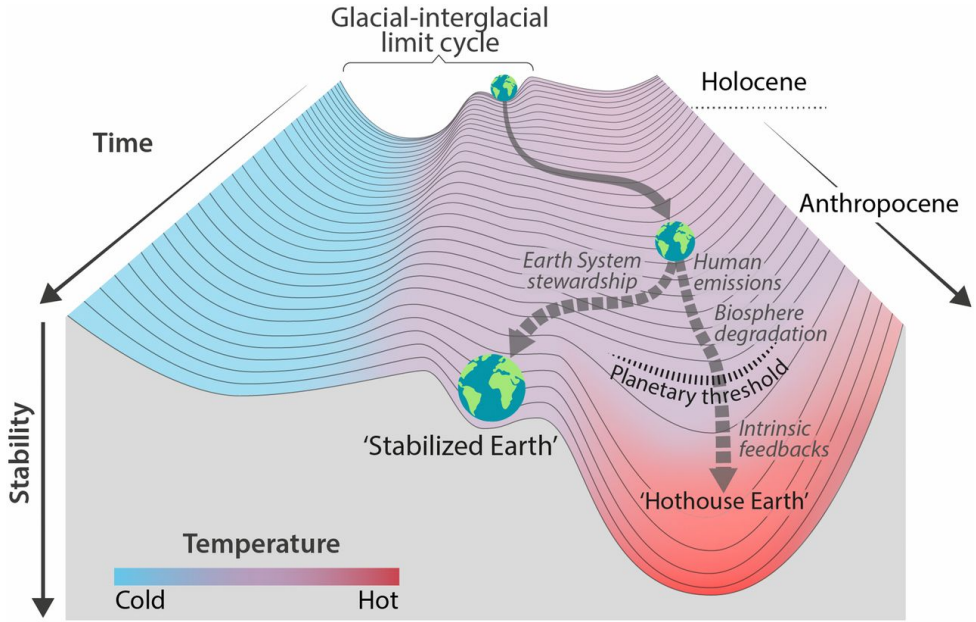

Figure 1.4 | **Cartoon of the stability landscape of the Earth system**. The state of the Earth is imagined as a ball rolling in the landscape, which morphs over time as the external conditions change. At each time instance, multiple stable states are possible, represented by different valleys. Here Steffen et al. (2018) propose the emergence of a "Hothouse Earth" state at the transition from the Holocene to the Anthropocene; we are instead concerned with the stability landscape associated with multiple competing AMOC states. Reproduced from Steffen et al. (2018), under BY-CC 4.0 license.

surface temperature associated with Snowball Earth (Lucarini and Bódai 2017; Lucarini and Bódai 2020; Margazoglou et al. 2021). In fact, given Earth's multistability on various scales, we may imagine a hierarchical landscape where hills and valleys are nested within each other as one zooms in to smaller scales (Margazoglou et al. 2021). For example, Lohmann, Dijkstra, et al. (2024) presented a qualitative illustration of the complex stability landscape found for the AMOC in a global ocean model. A proposition for how the stability landscape of global climate could be morphing over time has been sketched by Steffen et al. (2018) in a widely discussed perspective paper (Fig. 1.4).

In summary, the idea of a stability landscape has been established also in climate science. However, attempts to map out the landscape in detail remain rare. Without question, this is a challenging task for a high-dimensional system like the Earth system, and precisely quantifying the full landscape is currently not feasible. Nonetheless, we argue that valuable insight can be gained by exploring the stability landscape as far as possible. For example, even if we cannot compute the full landscape, we may focus on special regions of interest, such as the "mountain passes" on the edge between valleys that connect different climatic states. Such objects, which we call *Melancholia states* (Lucarini and Bódai 2017), are a core theme of this thesis.

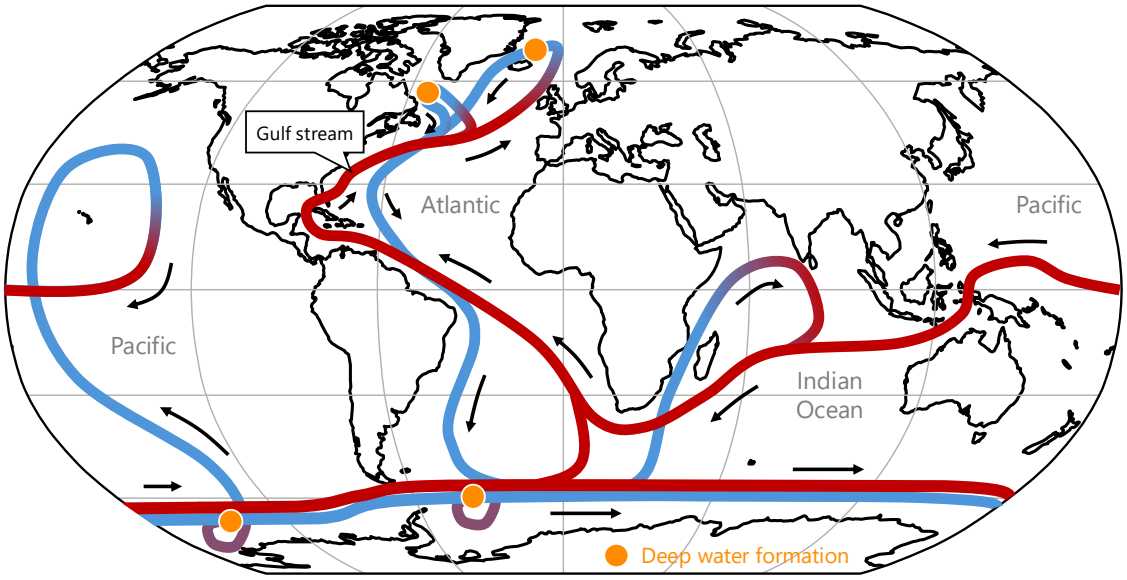

Figure 1.5 | **Idealized representation of the global thermohaline ocean circulation.** Warm surface currents (red) and cold deep currents (blue) form an oceanic "conveyer belt," with major deep water formation sites marked (orange circles). The Gulf stream contributes to a part of this circulation off the North American east coast. Figure reproduced based on Rahmstorf (2002), Kuhlbrodt et al. (2007) and Robert Simmon, NASA.

1.2 The Atlantic Meridional Overturning Circulation

1.2.1 A global ocean conveyor belt

The Atlantic Meridional Overturning Circulation (AMOC) is a system of large-scale ocean currents that extends throughout the Atlantic Ocean and is connected to a worldwide network of oceanic flows (Kuhlbrodt et al. 2007; Buckley and Marshall 2016). Schematically, the global ocean circulation has been described as a conveyor belt that transports water and tracers around the globe, with warm currents at the surface and cold currents in the deep ocean (Fig. 1.5). Only in a few localized regions at high latitudes, surface water becomes so dense that it sinks to great depth, connecting the surface and deep currents (Rahmstorf 2002). Of course, the schematic depiction of a conveyer belt is only a simplification. The ocean is highly turbulent, characterized by highly complex flow patterns as visualized in Fig. 1.6.

Starting from the tropics, the AMOC follows the Gulf stream, the western boundary current off the American east coast. While the wind-driven component of the Gulf stream recirculates in a subtropical gyre, the AMOC continues northward along the North Atlantic current until reaching deep water formation sites in the Labrador, Irminger and Nordic Seas. The North Atlantic Deep Water then recirculates southward as a deep western boundary current. Upwelling in the south closes the loop, forming a meridional (i.e. north-south) overturning circulation.

1.2.2 Processes and feedbacks

The large-scale ocean currents are driven by surface winds and density differences. The component due to density gradients is known as the *thermohaline circulation*, since density

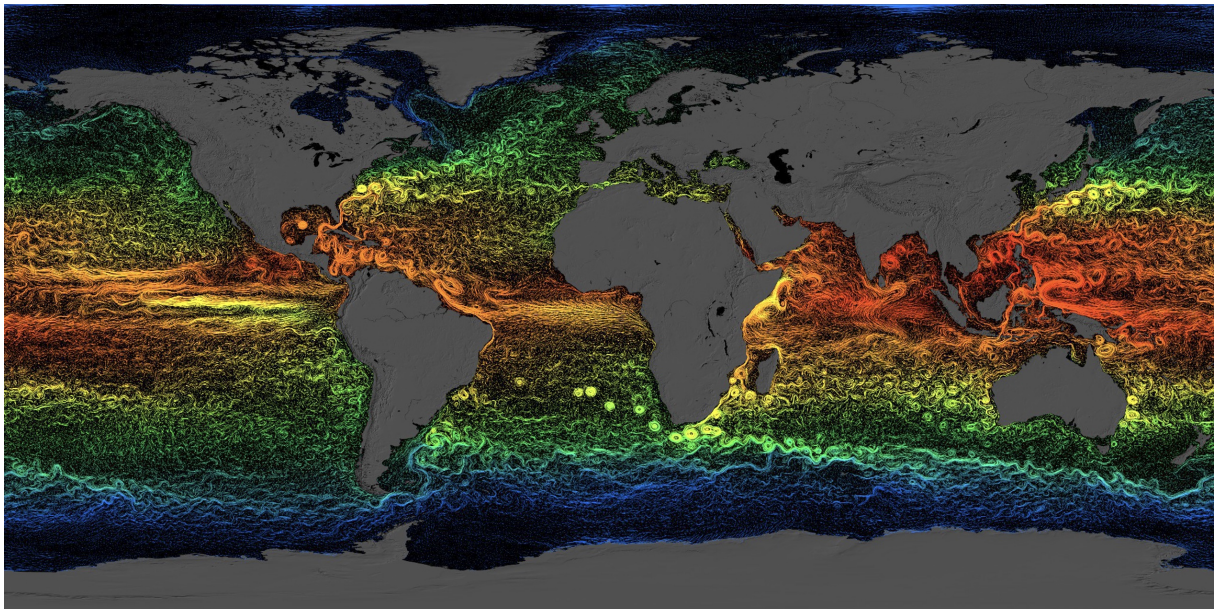

Figure 1.6 | **World map of sea surface currents** showing a high degree of turbulence, colored by sea surface temperature (SST) where the color intensity represents the flow speed. Data produced by the ECCO2 reanalysis combining the MITgcm model with satellite and in-situ observations. Figure produced by Greg Shira, NASA/Goddard Space Flight Center Scientific Visualization Studio (https://svs.gsfc.nasa.gov/3912/).

differences are induced by fluxes of heat ("thermo") and salt ("haline"). The relationship between density ρ, temperature T and salinity S is expressed in the equation of state, which to first order may be approximated as a linear equation,

$$\rho(T, S) = \rho_0 - \alpha T + \beta S \,, \tag{1.2}$$

where ρ_0 is a reference density and α and β are expansion coefficients. As the signs indicate, density increases with increasing salinity and decreasing temperature. When comparing densities between different water parcels, it is useful to consider the potential temperature Θ instead of absolute temperature T, since the former corrects for pressure effects on density (Isachsen and Christensen 2017).

A more accurate approximation of the equation of state that also accounts for nonlinearities and the depth dependence is given by the S-EOS 10 formula (Vallis 2017),

$$\rho(\Theta, S, z) = \frac{1}{\rho_0}\left(-a_0\left(1 + \frac{\lambda_1 T_a}{2} + \mu_1 z\right)T_a + b_0\left(1 - \frac{\lambda_2 S_a}{2} - \mu_2 z\right)S_a - \nu T_a S_a\right) . \tag{1.3}$$

Here $T_a = \Theta - 10°\mathrm{C}$, $S_a = S - 35\,\mathrm{g\,kg^{-1}}$ and $\rho_0 = 1026\,\mathrm{kg\,m^{-3}}$. Furthermore, $a_0 = 0.166$, $b_0 = 0.766$, $\lambda_1 = 5.95 \times 10^{-2}$, $\lambda_2 = 5.49 \times 10^{-4}$, $\mu_1 = 1.50 \times 10^{-4}$, $\mu_2 = 1.11 \times 10^{-5}$ and $\nu = 2.43 \times 10^{-3}$.

Salt-advection feedback. In the subtropics, surface waters are warm and salty due to enhanced evaporation. The AMOC transports vast amounts of this heat and salt northwards, where the heat is lost to the atmosphere, such that the water cools. The cold, salty water attains a high density, triggering deep convection and sinking. A stronger AMOC

supplies more salt to the subpolar North Atlantic, which favors deep water formation. In turn, enhanced deep water formation leads to a strengthening of the overturning circulation. This creates a positive feedback loop known as the *salt-advection feedback*: a stronger AMOC advects more salt northward, supporting deep convection; this strengthens the AMOC, which advects more salt, and so on (Weijer et al. 2019). In reverse, a weakening of the AMOC would cause itself to weaken further, implying the possibility of a self-perpetuating collapse.

Since the AMOC is a basin-wide circulation, the salt-advection feedback depends on whether the AMOC sustains a net import or export of salt into the Atlantic. Therefore, Rahmstorf (1996) proposed the net freshwater transport at the southern border of the Atlantic, $F_{\mathrm{ovS}} = F_{\mathrm{ov}}(\phi_S)$, as a stability indicator:

$$F_{\mathrm{ov}}(\phi) = -\frac{1}{S_0}\int_{z_0}^{0}\left(\int_{\varphi_W}^{\varphi_E}\big(v(\phi,\varphi,z)-\bar{v}(\phi)\big)r_\circ|\cos\phi|\mathrm{d}\varphi\right)\big(\langle S(\phi,z)\rangle - S_0\big)\,\mathrm{d}z, \tag{1.4}$$

where $\phi_S = -34°\mathrm{S}$ is the latitude at the Atlantic southern border, φ is the longitude (integrated from the western boundary φ_W to the eastern boundary φ_E of the Atlantic basin), z denotes depth, S_0 is a reference salinity and $r_\circ$ is Earth's radius. From the meridional velocity field v, the zonal section average $\bar{v}$ is subtracted to yield velocity anomalies. Furthermore, $\langle S\rangle$ denotes the zonally averaged salinity field in the Atlantic.

A negative F_{ovS} supports the positive salt-advection feedback as described above, indicating that the AMOC is in a bistable regime since a collapsed state becomes possible (Rahmstorf 1996; Hawkins et al. 2011; Weijer et al. 2019; van Westen et al. 2024). Conversely, if $F_{\mathrm{ovS}} > 0$, the AMOC exports net salinity.

Other drivers and feedbacks. While the salt-advection feedback is considered the key mechanism of bistability, many other processes influence the AMOC. Highlighting the competing effects of salinity and temperature, there is a stabilizing thermal advective feedback: a stronger AMOC not only transports more salt northwards but also more heat, leading to warmer North Atlantic sea surface temperatures. This increases buoyancy by reducing the surface density, resulting in a negative feedback.

On the other hand, an increase in air temperature in the subpolar North Atlantic (due to global warming and Arctic amplification (Rantanen et al. 2022)) leads to higher precipitation amounts and accelerated melting of glaciers and sea ice. Both the heat and freshwater fluxes into the North Atlantic at the air-sea interface act to weaken the AMOC.

The response of the North Atlantic circulation can be highly nonlinear due to critical behavior in deep convection and sea ice formation. For instance, convection kicks in at a relatively sharp stratification threshold when the water column becomes unstable (Kuhlbrodt et al. 2001; Levermann and Born 2007). This instability has been associated with multistability in the SPG, which has been proposed as a separate tipping element but is closely interlinked with the AMOC (Sgubin et al. 2017). Sea ice locally adds a strong positive feedback: the growth of sea ice inhibits ocean heat loss to the atmosphere, thereby increasing buoyancy and shutting down convection. This, in turn, reduces the advection of warm water in to the region, amplifying cooling and thus sea ice growth. These feedbacks act on relatively fast timescales and make the AMOC highly sensitive to changes in the subpolar North Atlantic.

Finally, it is important to recognize that the AMOC is not purely density-driven (Kuhlbrodt et al. 2007). Wind-driven upper ocean mixing is essential to maintain an

overturning circulation, and upwelling in the Southern Ocean due to Ekman suction likely plays a crucial role in closing the circulation loop (Buizert and Schmittner 2015; Baker et al. 2025). To summarize, important drivers (among others; see Weijer et al. (2019)) impacting AMOC stability include the salt-advection feedback (positive), the thermal advective feedback (negative), surface heat and freshwater exchange, convection thresholds, ocean-sea ice feedbacks (positive) and Southern Ocean winds.

1.2.3 Measuring the AMOC

The ocean covers 70% of Earth's surface, and nowadays satellites monitor every square kilometer of it. Yet, a saying goes that we may know more about outer space than we know about the deep sea (Schätzing 2005). In fact, three quarters of the sea floor remain unexplored (NOAA 2025), since satellite sensors cannot penetrate the sea surface and expeditions mostly sample the upper ocean. Fortunately, new observation systems and the increasing deployment of autonomous profiling floats are unveiling more of the three-dimensional ocean day by day (Jayne et al. 2017; Zilberman et al. 2023). This could help understand the uncertain future of global ocean currents.

Continuous, direct measurements of the AMOC only begun in 2004 with the RAPID array, a collection of measurement systems transecting the Atlantic at 26°N (McCarthy et al. 2012). More recently, arrays have been installed further south (SAMBA) and in the Subpolar Gyre (SPG) region (OSNAP) (Frajka-Williams et al. 2019).

It is well-known that the AMOC features decadal and centennial variability, including the North Atlantic Oscillation (Jiang et al. 2021; Latif et al. 2022; Mehling et al. 2022). Therefore, the current time period of direct observations is too short to robustly assess whether the AMOC is declining on a more long-term trend (Kilbourne et al. 2022). To overcome this limitation, several studies have turned to so-called AMOC fingerprints – proxy observables that are thought to correlate with the AMOC overturning strength (Caesar et al. 2018; Michel et al. 2022). For example, satellite observations show that a region in the SPG (called the "cold blob") has warmed less compared to other regions over the past years, while coastal waters off the U.S. east coast have warmed more (Caesar et al. 2018). This has been interpreted as a signal of AMOC weakening (Rahmstorf et al. 2015) but its validity has been questioned (Kilbourne et al. 2022; Terhaar et al. 2025). Overall, whether the AMOC has weakened or not during recent years remains a topic of scientific debate.

The AMOC strength typically refers to the meridional volume transport due to overturning. This quantity can be measured via the meridional streamfunction Ψ,

$$\Psi(\phi, z) = -\int_{z_0}^{z}\int_{\varphi_W}^{\varphi_E} v(\varphi, \phi, z') \cdot r_\circ |\cos\phi| \mathrm{d}\varphi \mathrm{d}z' \,. \tag{1.5}$$

Here ϕ is the latitude and the meridional velocity field v is integrated zonally across the Atlantic basin. Commonly, the AMOC strength is then defined as the maximum of Ψ over depth at a given latitude, often chosen to be 26.5°N.

1.2.4 The AMOC in a hierarchy of climate models

In light of limited observations, much of our understanding of the AMOC stems from climate models. The complexity of AMOC models covers the entire climate model hierarchy,

from conceptual box models to state-of-the-art earth system models. With a celebrated two-box model, Stommel (1961) showed that the thermohaline flow between two asymmetrically forced sea basins may feature two coexisting stable flow states. While Stommel originally did not have the AMOC in mind, his work laid the foundation for identifying the AMOC as a multistable system (Broecker et al. 1985; Manabe and Stouffer 1988; Manabe and Stouffer 1999).

Rahmstorf (1995) investigated the multistability of the AMOC via a hysteresis experiment, i.e. by slowly (quasi-adiabatically) ramping the freshwater flux into the North Atlantic up and down. This experiment showed a bistable region where a strong and weak AMOC state coexist. The hysteresis experiment has since been performed in intermediate-complexity climate models (Rahmstorf et al. 2005; Hawkins et al. 2011; Willeit and Ganopolski 2024) and a state-of-the-art earth system model (van Westen and Dijkstra 2023). Across the model hierarchy, the models exhibit a remarkable qualitative agreement with the bifurcation diagram found for the conceptual Stommel model.

Quantitatively, climate models differ considerably in simulating the AMOC. The earth system models participating in the Coupled Model Intercomparison Project phase 6 (CMIP6) may be considered the current state of the art. Still, these models typically have a horizontal resolution of around 1°, such that turbulent eddies and other processes must be parameterized. While the multimodel mean of the AMOC strength is close to the observed value, individual CMIP6 models range from 8 to 32 Sv in AMOC strength (Weijer et al. 2020). Many models have known biases in the salinity field and surface freshwater fluxes, which can strongly impact the location of bifurcation points (Mecking et al. 2017). Particularly, most CMIP6 models produce a positive F_{ovS}, whereas observations point to a negative F_{ovS} (van Westen and Dijkstra 2024). The AMOC response to freshwater hosing has also been shown to vary across models (Jackson et al. 2023). Intermodel spread in the magnitude of AMOC weakening simulated over the 21st century can partly be attributed to biases in ocean stratification and overturning depth, which affects the amount of weakening more than the choice of greenhouse gas emissions scenario (Bonan et al. 2025).

Qualitatively, CMIP5 and CMIP6 models are consistent in predicting a decline of the AMOC (Weaver et al. 2012; Cheng et al. 2013; Schleussner et al. 2014; Weijer et al. 2020) until the year 2100, but none of the models simulate an abrupt collapse over that time period (although a substantial weakening within a century could be considered abrupt on oceanic timescales). The 6th Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), published in 2021, thus concludes: "The AMOC will very likely weaken over the 21st century (high confidence), although a collapse is very unlikely (medium confidence)" (Masson-Delmotte et al. 2021). The "medium confidence" classification highlights the uncertainties involved in this estimate.

In fact, the likelihood of AMOC tipping could be underestimated for technical reasons (Gent 2018). Due to model biases and surface flux adjustments to correct biases artificially, climate models might be "too stable" compared to reality (Mecking et al. 2017; Liu et al. 2017). Another possible reason lies in the high computational cost of high-resolution earth system models: AMOC transitions might be observed more frequently in simplified models simply because they can be simulated for more model years. Viewing AMOC transitions as rare events, the relatively slow oceanic timescales from decades to centuries imply that state-of-the-art models are necessarily undersampled.

To increase our confidence in climate model projections, we need a robust physical

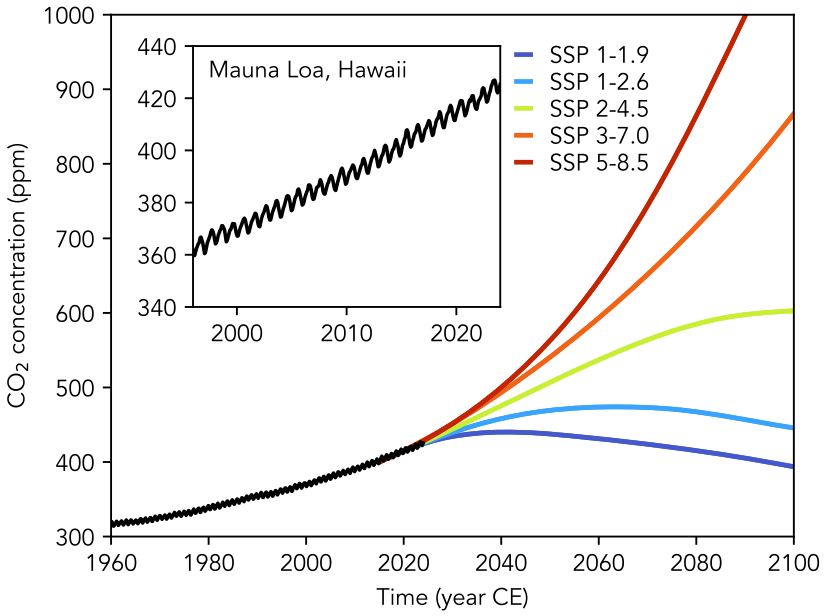

Figure 1.7 | **Historical and projected future atmospheric CO_2 concentrations.** The historical record (black line) was recorded at Mauna Loa, Hawaii (Thoning et al. 1989), known as the Keeling curve (Keeling et al. 1976). The observations show an increase by 18% since the author of this thesis was born (inset). Colored curves show possible future scenarios until 2100 based on Shared Socioeconomic Pathways (SSPs) (Meinshausen et al. 2020).

understanding of their behavior. Arguably, this can only be achieved by connecting the full model hierarchy – from box models to state-of-the-art earth system models. Conceptual climate models provide invaluable insight into the underlying dynamics of metastable earth system components, but their properties must be linked to high-dimensional models to gain confidence that they enable understanding of the real climate system. For this purpose, earth system models of intermediate complexity (EMICs) are very useful tools. They permit extensive investigations due to computational speed while incorporating many of the relevant processes in a simplified manner.

1.3 Climate predictability

A transition of the AMOC before the year 2100 cannot be ruled out. Human activities, such as the burning of fossil fuels and land use, are rapidly and profoundly changing our planet (IPCC 2023). Mathematically, this raises the question of whether we can solve the boundary value problem: how will the Earth system respond to the external time-dependent forcing? Since this response will be state-dependent in general, we must equally consider the initial value problem: where will we evolve from our current state?

When looking to the future, one challenge to climate prediction arises from the uncertainty in the boundary condition: how will anthropogenic climate forcing evolve? Since this is undecided, it is common practice in climate modeling to study the response to different forcing scenarios that cover the plausible range of future human development (Fig. 1.7). A standardized portfolio of greenhouse gas emissions scenarios is provided by the Shared Socioeconomic Pathways (SSPs; Meinshausen et al. (2020)). The wide span of projected CO_2 concentrations across SSPs emphasizes both the uncertainty and agency of human action.

In principle, uncertainties due to data quality, model limitations and boundary condition constraints are reducible. By contrast, the chaotic nature of the climate system implies fundamental, irreducible limits to predictability, as we describe below.

1.3.1 Predictability of chaotic systems

Lorenz (1975) proposed two kinds of predictability that play a fundamental role in earth system science. The first kind is deeply linked with chaos: given an initial condition with finite precision, for how long can we predict the system's evolution before the chaotic divergence of nearby trajectories prevents this? In chaotic systems, the limited predictability of the first kind implies a finite time horizon at which the inital value problem can be solved before uncertainties grow exponentially. This explains why weather forecasts lose skill after a maximum of a few weeks.

Predictability of the second kind asks about the long-term behavior of the system: which asymptotic state will the system end up in, provided an external forcing with some uncertainty? In this context, we are concerned with predicting probability measures – instead of individual trajectories – that describe the mean and higher moments of variability of the climate state. For an ergodic measure, the inital condition does not matter after a while, since trajectories eventually populate the whole measure (Ghil and Sciamarella (2023) and Tantet (2016); see Term 2.3). Thus, long-term *climate* predictions can be made despite the unpredictability of long-term *weather*.

In metastable systems, the situation becomes more complicated. Near an instability, the asymptotic state of the system can depend sensitively on the initial condition (Knutti and Stocker 2002). Even if an invariant probability measure exists, it may be multimodal, giving rise to different outcomes for the long-term average evolution of the system depending on the initial condition, whereby the mixing property is lost. Lohmann, Wuyts, et al. (2024) have recently termed this the "predictability of storylines": even if both the first and second kind of predictability are highly restricted, it may still be possible to predict a finite number of possible outcomes, or storylines.

Overall, both the transient dynamics as well as the final state of a metastable system can become virtually uncertain near a critical threshold, posing a fundamental limit to the predictability of tipping points. For the AMOC, Romanou et al. (2023) demonstrated a striking example of this in a CMIP6 model under an intermediate SSP scenario (Fig. 1.8). There, an ensemble of ten initial conditions splits, with some members undergoing a sustained AMOC weakening while others show an AMOC recovery. We will return to these simulations at the end of the thesis (chapter 6).

1.3.2 Anticipating transitions: Early warning signs

In light of the risk associated with an AMOC transition and other tipping events, there is a strong mathematical interest in developing robust *early warning signs* of critical transitions. Since the release of the IPCC report, a series of articles have applied statistical early warning methods to observed time series of AMOC fingerprints to investigate whether the AMOC is currently losing stability (Boers 2021; Michel et al. 2022; Ditlevsen and Ditlevsen 2023). The general idea relies on the concept of *critical slowing down*: as a system approaches a tipping point, the stabilizing restoring forces of the system diminish, which can be measured as an increase in variance and auto-correlation over a moving time window (Dakos et al. 2008; Carpenter and Brock 2006; Boers et al. 2022). All studies indeed detect an increase, suggesting that the AMOC is currently nearing a tipping point. Proposing an early warning indicator based on the F_{ovS} index, van Westen et al. (2024) came to similar conclusions. Ditlevsen and Ditlevsen (2023) extrapolated the signal of critical slowing down to deduce a timing of collapse between 2025 and 2095 with 95%

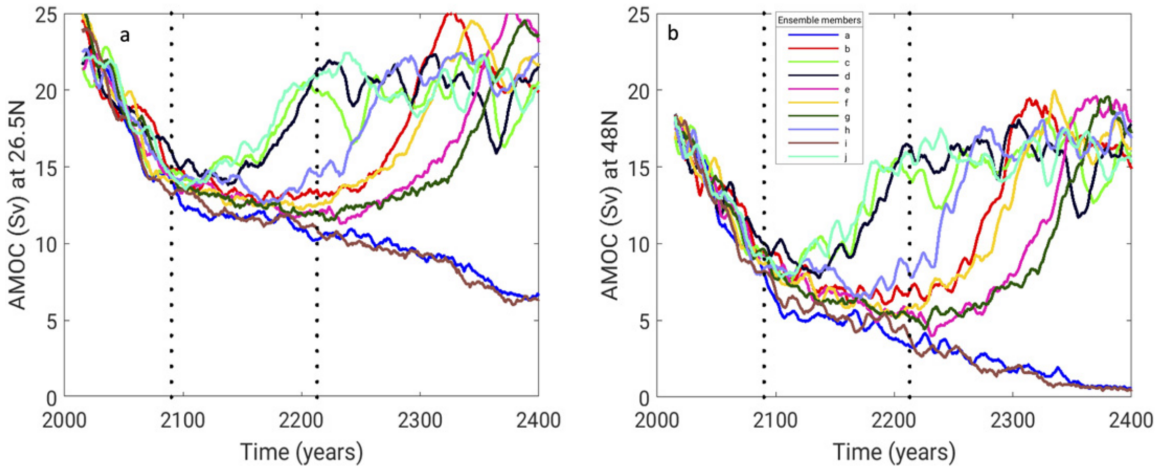

Figure 1.8 | **Ensemble splitting of the AMOC** in a CMIP6 model (NASA GISS E2-1-G) under an intermediate emissions scenario (SSP2-4.5). Each panel shows 10 ensemble members under the same forcing, plotting the meridional stream function maximum at 26.5°N (left) and 48°N (right). Figure reproduced from Romanou et al. (2023), with permission. Published 2023 by the American Meteorological Society.

confidence. Together, these studies suggest that an AMOC collapse by 2100 could be more likely than estimated by the IPCC AR6.

However, the early warning signs applied in these studies have strong limitations (Ben-Yami, Morr, et al. 2024). One issue relates to the source data and appropriateness of the fingerprints. Additionally, the method requires several restrictive assumptions on the dynamical properties of the underlying system which are not necessarily fulfilled by the real climate system. These restrictions pertain to the type of bifurcation causing the tipping point and the properties of stochastic fluctuations. Importantly, as a local stability concept, critical slowing down assumes the system is near a statistical equilibrium. Applying statistical early warning signs has revealed inconsistent results (Hummel et al. 2024) and false outcomes (Zimmerman et al. 2025). At the same time, the relevance of predicting potential climate tipping events has sparked a scientific effort to tackle the shortcomings of early warning signs (e.g. Morr and Boers (2024), Ritchie and Sieber (2016), Bernuzzi and Kuehn (2023), Moinat et al. (2024), Lohmann et al. (2025) and Santos Gutiérrez and Lucarini (2022)). These advances may pave the way towards more reliable methods, but inherent limitations due to the chaotic nature of the Earth system will likely remain.

1.4 Research aims & thesis outline

In this thesis, we seek to gain a deeper understanding of climate metastability by exploring the stability landscape of the AMOC in a range of climate models across the model hierarchy. Taking a phase space view on the dynamics, we aim to contribute a *global* perspective that complements the *local* view of near-equilibrium dynamics and early warning signs. Since computing the full quasipotential landscape is prohibitive in most models, we will focus on transition paths in the phase space. Transition paths give insight into the transient dynamics of metastable systems and encode the physical information needed to deduce transition mechanisms. Specifically, we will investigate the role of Melancholia states for transition paths under stochastic or chaotic nonautonomous forcing. With a

numerical method called the edge tracking algorithm, Melancholia states can be computed also in high-dimensional complex systems, and deploying this tool will allow us to connect our findings across the model hierarchy. Lastly, we will emphasize the relevance of transient chaos and limited predictability for making sense of climate tipping points.

1.4.1 Knowledge gaps

Our work is motivated by the objective to tackle the following broader knowledge gaps:

- The phase space and stability landscape of high-dimensional climate models remains underexplored.
- Determining most probable transition paths in non-gradient systems under stochastic forcing is an open mathematical problem beyond the weak-noise limit, particularly for state-dependent noise.
- The role of unstable states and the relevance of long transients for determining the stability and evolution of dynamical systems, such as the AMOC, is not fully understood.
- A Melancholia state of the AMOC has not previously been calculated in a coupled general circulation model.
- Studies of the long-term response of the AMOC to forcing scenarios beyond the year 2100 remain sparse. There has been a strong focus on freshwater forcing, while the stability with respect to CO_2 changes has been less studied.
- Conceptually, tipping is often regarded as a binary problem, and there is a need to deepen our understanding of what we can or cannot predict.

1.4.2 Research questions

The thesis addresses the following research questions.

Research questions

Q1. What does the stability landscape of the AMOC look like? Is the picture obtained from conceptual climate models consistent with the dynamics of comprehensive earth system models?

Q2. Can we determine the Melancholia state(s) separating competing AMOC states in a high-dimensional climate model? If so, what are the dynamical, physical and climatological properties of such state(s)?

Q3. What is the role of Melancholia states for metastable dynamics, transition paths and long transients under stochastic and chaotic forcing?

Q4. What can we learn from Melancholia states about the global stability of the Earth system, transition mechanisms and the predictability of our future climate?

1.4.3 Thesis outline

The thesis brings together four studies that explore the role of Melancholia states for metastable behavior in models of varying complexity, from two to 10^5 degrees of freedom. To connect these studies, a common mathematical basis is essential. In chapter 2.2.4, we review the concepts, theory and numerical methods that provide the underlying framework for this work.

To illustrate key ideas in an AMOC context, we begin with a simple box model of the interhemispheric thermohaline circulation (chapter 3, based on Börner et al. (2022)). We focus on the dynamics near critical thresholds and the transition behavior under stochastic freshwater forcing. This investigation leads to the unexpected observation of saddle avoidance of noise-induced transition paths, a phenomenon that we study in detail in chapter 4 (based on Börner et al. (2024b)). Specifically, we clarify the link between saddle avoidance and timescale separation in multiscale dynamical systems driven by weak yet finite noise. Based on large deviation theory, we present a finite-noise variational approach that allows to aptly predict the transition behavior.

In chapter 5 (based on Mehling et al. (2024)), we return to the AMOC and explore its metastability in a conceptual ocean-atmosphere model, where now, instead of a stochastic representation, the fast variability of the atmosphere is explicitly modeled as a chaotic system. The ocean component consists of the seminal Stommel model, for which we explicitly compute the global quasipotential. The ocean-atmosphere coupling gives rise to complex dynamics that highlight fundamental limits to the predictability of the climate system's long-term evolution, particularly under overshoot forcing scenarios. Our conceptual study outlines a recipe that we apply to a high-dimensional climate model in chapter 6 (based on Börner et al. (2025a)). In an intermediate-complexity earth system model, we explicitly compute the Melancholia state for two different levels of CO_2 forcing. We characterize the climate dynamics of the Melancholia state and leverage the results to understand the metastable AMOC behavior observed in simulations of future climate change, featuring an ensemble splitting and long chaotic transients.

Chapter 7 starts with a concise chapter-by-chapter summary of the thesis. Finally, we review the research outcomes of this work and reflect on the implications for future research as well as for our understanding of climate tipping points.

This PhD work also initiated the open-source software package CriticalTransitions.jl, available[5] in the Julia language (see sections 2.2.4 and 4.2.5).

[5]Source code and documentation: `https://github.com/juliaDynamics/criticalTransitions.jl`.

Chapter 2

Theory, concepts & methods

2.1 Elements of dynamical systems theory 26

2.1.1 Nonautonomous dynamical systems 26

2.1.2 The deterministic view 27

2.1.3 Multistability and the stability landscape 31

2.1.4 Saddles, edge states and Melancholia states 34

2.1.5 The stochastic view 35

2.1.6 Large deviation theory 37

2.1.7 Metastability, tipping and transient chaos 42

2.1.8 The role of Melancholia states for metastable dynamics 47

2.1.9 Multiscale dynamics and homogenization 48

2.2 Numerical methods 50

2.2.1 Finding Melancholia states: Edge tracking algorithm 50

2.2.2 Sampling noise-induced transitions 53

2.2.3 Predicting transition paths 56

2.2.4 Computing quasipotentials 58

This thesis seeks to capitalize on an interdisciplinary approach towards earth system science. To be successful, interdisciplinarity needs a common language. Following the geophysical introduction in chapter 1.4.3, we here outline the mathematical concepts and methodology that form the basis of this work.

Over the past decades, the complexities of the Earth system have often required pushing scientific frontiers, leading to breakthroughs also beyond earth system science, for example in chaos theory (Lorenz 1963), data assimilation (Ghil and Malanotte-Rizzoli 1991) and machine learning (Geer 2021), early-warning signs (Scheffer et al. 2009), nonlinear dynamics (Ashwin et al. 2012), stochastic dynamics (Hasselmann 1976), and response theory (Lucarini and Chekroun 2024). The iconic Lorenz system, the paradigm of chaos theory, was originally developed in the context of atmospheric convection (Lorenz 1963). The 2021 Nobel prize in physics, shared between Parisi and the climate scientists Hasselmann and Manabe, recognized their "groundbreaking contributions to our understanding of complex physical systems"[1] in general (Ravishankara et al. 2022).

2.1 Elements of dynamical systems theory

Dynamical systems theory provides a mathematical framework for studying systems that change over time. It is concerned with systems for which the time evolution can be expressed by a dynamical rule that iteratively leads from one time instance to the next. The theory finds application in almost every field of science, from protein folding and epidemiology to power grids and the stock market. The theory is also fundamental for our understanding of Earth system dynamics and our past, present, and future climate (Caraballo and Han 2016; Dijkstra 2013).

2.1.1 Nonautonomous dynamical systems

Throughout this thesis, we treat a climate model (or any other dynamical model of nature) as a nonlinear, nonautonomous dynamical system, which we write in its general form as a system of stochastic differential equations,

$$\mathrm{d}\boldsymbol{x} = \underbrace{\boldsymbol{b}\big(\boldsymbol{x},\, \Lambda(t)\big)\mathrm{d}t}_{\text{deterministic}} + \underbrace{\sigma \boldsymbol{Q}^{1/2}(\boldsymbol{x})\mathrm{d}\mathcal{N}_t}_{\text{stochastic}}\,, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0\,. \tag{2.1}$$

Here the climate state $\boldsymbol{x}(t) \in \mathbb{R}^D$ at time $t \geq 0$ is described by a D-dimensional state vector, initialized from the deterministic *initial condition* $\boldsymbol{x}_0$ (in other words, we take the initial probability distribution of the random variable $\boldsymbol{x}(0)$ to be $P(\boldsymbol{x}(0) = \boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}_0)$, where $\delta(\boldsymbol{x})$ denotes the Dirac δ-function). On the right hand side of Eq. (2.1), the first term describes the deterministic evolution of the system governed by the continuous *drift* field $\boldsymbol{b} : \mathbb{R}^D \times \mathbb{R}^N \to \mathbb{R}^D$, subject to a time-dependent forcing given by the parameter drift $\Lambda(t) : \mathbb{R} \to \mathbb{R}^N$. Unless stated otherwise, we will assume that $\boldsymbol{b}$ is continuously differentiable. The second term represents stochastic forcing of Itô type, where $\sigma \geq 0$ is the noise strength, $\boldsymbol{Q}(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^{D\times D}$ is the noise covariance matrix, and $\mathcal{N}_t$ is a D-dimensional stochastic process[2] (Pavliotis 2014). The matrix square root of $\boldsymbol{Q}$ can

[1] Source: https://www.nobelprize.org/prizes/lists/all-nobel-prizes-in-physics/, visited on 8 January 2025.

[2] Besides the formulation here, different conventions exist for defining the stochastic process (Pavliotis 2014).

be obtained, for example, via the Cholesky decomposition $\boldsymbol{Q}(\boldsymbol{x}) = \boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{\Sigma}^\top(\boldsymbol{x})$, where $\boldsymbol{\Sigma} = \boldsymbol{Q}^{1/2}$ is sometimes called the diffusion matrix (Dahiya and Cameron 2018a). We assume the covariance matrix to be a continuous function of the state vector $\boldsymbol{x}$, permitting state-dependent noise. To ensure uniqueness of the term $\sigma\boldsymbol{Q}^{1/2}$, we may require $\boldsymbol{Q}$ to be normalized according to some criterion[3] and absorb the normalization constant in σ.

A *nonautonomous* system is a dynamical system that depends explicitly on time. This is the case if the external forcing $\Lambda(t)$ varies over time but also if the noise properties are explicitly time-dependent, e.g. $\boldsymbol{Q} \equiv \boldsymbol{Q}(\boldsymbol{x}, t)$.

In a deterministic setting, the stochastic term is absent ($\sigma = 0$) and the focus lies on the properties of the drift field $\boldsymbol{b}$ as well as the system's response to the forcing $\Lambda(t)$. In the presence of noise ($\sigma > 0$), the situation changes fundamentally, and we are interested in the interplay of the deterministic dynamics and stochastic forcing.

To see how the dynamical system (2.1) evolves in time, we can simply plot the time-series of the components of $\boldsymbol{x}(t)$. However, a deeper understanding of the dynamics is offered by the geometric view of the state variables co-varying in time. This leads to the concept of the *phase space*, which is crucial for this work.

Term 2.1. **Phase space**

The phase space $\Gamma \subseteq \mathbb{R}^D$ of a dynamical system is the subset of the D-dimensional Euclidean vector space, spanned by all state components of the state vector $\boldsymbol{x}$, where the vector fields $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ are defined.

A point $\boldsymbol{x}$ in phase space corresponds to a state of the system, and over time the system evolves along a *trajectory* forming a continuous curve in phase space. Formally, we can also think of the evolution of the system in terms of the *flow* $\Phi_t^\Lambda(\boldsymbol{x}) : \boldsymbol{x}(0) \mapsto \boldsymbol{x}(t)$ that maps the initial condition $\boldsymbol{x}_0 = \boldsymbol{x}(0)$ to its (future or past) solution[4] at time t (given a forcing protocol $\Lambda(t)$).

For systems with more than three dimensions ($D > 3$), the phase space becomes difficult to visualize. Additionally, the phase space volume grows rapidly with increasing D, known as the "curse of dimensionality" (Bellman 2010). Often, however, the dynamics of interest do not occupy the full phase space but are largely restricted to a lower-dimensional subset of Γ. It is therefore common to consider a *reduced phase space* of variables $\boldsymbol{y} = \xi(\boldsymbol{x})$ (sometimes called *latent space*), where $\xi : \mathbb{R}^D \to \mathbb{R}^k$ is a suitable mapping and $2 \leq k < D$. Note that while we consider Markovian dynamics in Eq. (2.1), the dynamics may become non-Markovian under the transformation ξ.

2.1.2 The deterministic view

Let us first discuss Eq. (2.1) in the deterministic case,

$$\dot{\boldsymbol{x}} = \boldsymbol{b}\big(\boldsymbol{x},\, \Lambda(t)\big)\,, \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0\,, \tag{2.2}$$

[3] For homogeneous (state-independent) noise, a suitable normalization factor is $D/L_1(\boldsymbol{Q})$, where L_1 denotes the L1 matrix norm.

[4] We implicitly assume the existence and uniqueness of solutions to Eq. (2.1).

where we introduced the common notation $\dot{\boldsymbol{x}} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$. Generally, climate dynamics must be viewed in a nonautonomous context, since the boundary conditions change continuously. For example, Earth's orbit around the sun causes the solar energy reaching Earth to vary over the course of millenia (Milanković 1941; Berger 1988); its axial tilt causes the incident energy distribution to vary annualy, while the Earth's rotation around itself causes a daily cycle. Since the 1900s, the rapid increase in greenhouse gas concentrations in the atmosphere constitutes another strongly time-dependent radiative forcing (Masson-Delmotte et al. 2021).

In principle, a time-varying forcing parameter could be considered as an additional dynamical variable, turning the non-autonomous system into a higher-dimensional autonomous one. However, this would require that the evolution equation of the forcing is known; furthermore, it can be useful to view certain parameters like CO_2 concentrations as external control parameters, as we will see.

Autonomous dynamics. Still, we can learn much about the nonautonomous system by considering the *frozen system*, i.e. a snapshot of the system at fixed external forcing (Wieczorek et al. 2023). After all, whenever a "non" precedes an adjective like linear, equilibrium, or autonomous, it seems common practice to take the theory applying to the case without the "non" and use or generalize it as far as possible.

Term 2.2. **Frozen system**

The frozen system described by

$$\dot{\boldsymbol{x}} = \boldsymbol{b}(\boldsymbol{x}, \lambda)\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{2.3}$$

is an autonomous version of the nonautonomous system (2.2) where the forcing is fixed to $\Lambda(t) = \lambda$.

The frozen system is amenable to the general framework of autonomous dynamical systems, which we summarize in the following (see also Caraballo and Han (2016)).

A trajectory of the frozen system initialized at a state $\boldsymbol{x}_0$ will evolve under the drift $\boldsymbol{b}(\boldsymbol{x}, \lambda)$. After waiting sufficiently long, it will converge to its *asymptotic state*, i.e. the invariant set in phase space reached after infinite time ($t \to +\infty$). An *invariant set* X is a subset of Γ that fulfills $\Phi_t(X) = X$ for all $t \in \mathbb{R}$. The asymptotic behavior of the system brings us to the notion of attractors. Intuitively, an attractor is a region in phase space that trajectories are attracted to and asymptotically tend towards over time.

Term 2.3. **Attractor**

An *attractor* $\boldsymbol{A} \subset \Gamma$ is a set of points in phase space satisfying the following properties:

(i) $\boldsymbol{A}$ is an invariant set, i.e.

$$\boldsymbol{x}_0 \in \boldsymbol{A} \Rightarrow \Phi_t(\boldsymbol{x}_0) \in \boldsymbol{A} \quad \forall t \geq 0\,;$$

(ii) $\boldsymbol{A}$ is asymptotically stable for points in an open neighborhood $\mathcal{B}(\boldsymbol{A}) \supset \boldsymbol{A}$ with nonzero measure,

$$\mathcal{B}(\boldsymbol{A}) = \left\{ \boldsymbol{y} \in \Gamma : \lim_{t \to +\infty} d(\Phi_t(\boldsymbol{y}),\, \boldsymbol{A}) = 0 \right\}\,;$$

(iii) any two open (non-empty) subsets $X, Y \subset \boldsymbol{A}$ are connected by transitivity, i.e. there exists $t > 0$ such that

$$Y \cap \Phi_t(X) \neq \emptyset\,.$$

The transitivity property ensures that if the attractor is an extended set (not a fixed point), trajectories initialized on the attractor will eventually explore the entire set, visiting any open subset of the attractor infinitely often as $t \to \infty$.

Term 2.4. **Basin of attraction**

The *basin of attraction* $\mathcal{B}(\boldsymbol{A}) \subseteq \Gamma$ of an attractor $\boldsymbol{A}$ is the set of initial conditions that converge to $\boldsymbol{A}$ as $t \to \infty$, corresponding to the open neighborhood $\mathcal{B}(\boldsymbol{A})$ defined in Term 2.3.

There are different types of attractors:

- Stable fixed point. In the simplest case, an attractor consists of a single point $\boldsymbol{x}^*$, a *stable fixed point* or *stable node* at which $\boldsymbol{b}(\boldsymbol{x}^*, \lambda) = 0$. For example, a pendulum swinging in air without external perturbations will come to rest at the center point.
- Stable limit cycle. Imagine another pendulum swinging in vacuum, where friction is absent. This pendulum will keep swinging back and forth forever at its natural frequency, so its attractor is a *stable limit cycle*, a closed curve in the phase space spanned by the phase angle and angular velocity.
- Chaotic attractor. Attractors can also be *chaotic* invariant sets, as famously exemplified by the Lorenz attractor (Lorenz 1963). In this case, the attractor is an extended invariant set consisting of infinitely many unstable periodic orbits (UPOs) (Maiocchi et al. 2022). While a trajectory initialized on the Lorenz attractor will never leave the attractor, it will keep swirling around from the neighborhood of one orbit to another.

Attractors are stable in the sense that the system tends back to the attractor after small perturbations away from it. This is quantified by *linear stability analysis*, which gives a local measure of stability. Consider two infinitesimally separated points $\boldsymbol{x}_0$ and $\boldsymbol{x}' = \boldsymbol{x}_0 + \boldsymbol{\delta}$. Locally around $\boldsymbol{x}_0$, the dynamics are approximated by the linearized dynamics $\dot{\boldsymbol{\delta}} \approx \boldsymbol{J}(\boldsymbol{x}_0)\boldsymbol{\delta}$, where and $\boldsymbol{J}(\boldsymbol{x})$ is the Jacobian of $\boldsymbol{b}$ at $\boldsymbol{x}$ (assuming that $\boldsymbol{b}$ is continuously differentiable).

A fixed point $\boldsymbol{x}^*$ with $\boldsymbol{b}(\boldsymbol{x}^*, \lambda) = 0$ is stable if and only if all eigenvalues of $\boldsymbol{J}(\boldsymbol{x}^*)$ have negative real parts. Then, the dynamics are contracting towards $\boldsymbol{x}^*$. By contrast, a saddle point is a fixed point at which the real parts of the Jacobian eigenvalues are both positive and negative, corresponding to unstable and stable directions, respectively.

Lyapunov spectrum. Chaotic systems are characterized by a sensitive dependence on the initial condition: small perturbations in the initial condition tend to grow exponentially over time (Alligood et al. 1996). Suppose we initialize two trajectories with an infinitesimal distance vector $\boldsymbol{\delta}_0$ between them. At time t (for t not too large), the magnitude of the distance roughly follows

$$||\boldsymbol{\delta}(t)|| \approx ||\boldsymbol{\delta}_0|| e^{\gamma_1 t} \,, \tag{2.4}$$

where γ_1 is a local estimate of the *maximum Lyapunov exponent* (MLE) (Pikovsky and Politi 2016; Datseris and Parlitz 2022). A positive MLE is a signature of chaos, since then infinitesimal perturbations lead to diverging trajectories.

Lyapunov exponents quantify the expansion and contraction of the flow in phase space. To define them globally for an invariant set[5], consider the fundamental matrix $\boldsymbol{M}_t \in \mathbb{R}^{D\times D}$ that governs the linearized evolution of $\boldsymbol{\delta}$,

$$\boldsymbol{\delta}(t) = \boldsymbol{M}_t(\boldsymbol{x}_0)\boldsymbol{\delta}_0 \,. \tag{2.5}$$

As shown by Oseledet (1968), there exits a matrix $\boldsymbol{\Omega}$,

$$\boldsymbol{\Omega}(\boldsymbol{x}) = \lim_{t\to\infty} (\boldsymbol{M}_t^\top(\boldsymbol{x})\boldsymbol{M}_t(\boldsymbol{x}))^{1/(2t)} \tag{2.6}$$

that has positive real eigenvalues ω_i, $i = 1, \ldots, D$ and that is almost independent of $\boldsymbol{x}$ (Pikovsky and Politi 2016; Maiocchi et al. 2024).

Term 2.5. **Lyapunov exponent, Lyapunov spectrum**

A *Lyapunov exponent* measures the rate of divergence of nearby trajectories on an invariant set along a direction in phase space.

For a D-dimensional dynamical system, there are at most D Lyapunov exponents γ_i, given by $\gamma_i = \ln \omega_i$ for $i = 1, \ldots, D$, where ω_i is an eigenvalue of $\boldsymbol{\Omega}$ in Eq. (2.6).

The set of Lyapunov exponents forms the *Lyapunov spectrum*,

$$\{\gamma_1, \gamma_2, \ldots, \gamma_D\} \quad \text{with} \quad \gamma_1 \geq \gamma_2 \cdots \geq \gamma_D \,.$$

A positive MLE, $\lambda_1 > 0$, indicates chaos.

[5] This derivation assumes that the system is ergodic (Pikovsky and Politi 2016).

Since the matrix $\mathbf{\Omega}$ is an asymptotic quantity, the associated Lyapunov exponents characterize the averaged dynamics across the invariant set. In practice, the Lyapunov spectrum can be computed by selecting a set of orthonormal perturbation vectors (spanning a hypersphere) and evolving them according to the linearized dynamics until some time Δt (Datseris and Parlitz 2022). This deforms the hypersphere into a hyper-ellipsoid, and the stretching and squeezing factors in each direction are recorded. By iteratively re-orthonormalizing and evolving the set of perturbation vectors (Benettin et al. 1980), we obtain a series of values for each Lyapunov exponent. The theorem by Oseledet (1968) justifies averaging over each series to estimate the Lyapunov spectrum of the invariant set.

At a stable fixed point, all Lyapunov exponents are negative. On a limit cycle, the maximum Lyapunov exponent is zero, accounting for the steady-state flow along the limit cycle: tangential to the limit cycle, nearby trajectories neither converge nor diverge. On a chaotic attractor, at least the maximum Lyapunov exponent is positive, characterizing the chaotic instability on the attractor, one or more exponents are zero, describing the motion along the attractor, and negative exponents account for the convergence of initial conditions to the attractor.

2.1.3 Multistability and the stability landscape

Many systems in nature possess more than one attractor, giving rise to the ubiquitous phenomenon of *multistability* (Feudel 2008; Pisarchik and Feudel 2014).

Term 2.6. **Multistability**

A dynamical system is *multistable* at a given forcing λ if and only if it possesses more than one attractor.

Specifically, a system with two attractors is called a *bistable* system.

As an example, consider a bottle of sparkling water[6] standing on a table. For simplicity, let the state of the bottle be fully described by its tilt angle, giving a one-dimensional system. If you tilt it only slightly and let go, the bottle will sway back and return to the upright state (tilt angle 0°), where it will stay. The upright state is thus an attractor. If you tilt the bottle more, there will be a critical tilt angle beyond which, once you let go, the bottle will tip over and end up lying horizontally (tilt angle 90°). The tipped state is another attractor. The bottle thus has two attractors; it is bistable. Whether the bottle ends up upright or tipped after you tilt it depends on the tilt angle at which you let go.

Definiton 2.6 implies that multistable systems feature more than one basin of attraction, and therefore the asymptotic state depends on the initial condition. To understand the asymptotic behavior of multistable systems, considering the *local* stability in the vicinity of each attractor is insufficient. Instead, a *global* stability perspective is needed that investigates the phase space beyond the attractors.

[6] This bottle featured in the Fairbrother Lecture 2024, in which I had the opportunity to present my PhD work to a wider audience; see `https://reykboerner.github.io/fairbrother-lecture` for a recording.

In the example of the bottle, the basins of attraction of the upright and the tipped state are separated by the point marking the critical tilt angle. In higher dimensions, the basins of attraction are volumes in the phase space, separated from each other by hyper-surfaces. These hyper-surfaces are called separatrices or *basin boundaries*.

Term 2.7. **Basin boundary**

A basin boundary $\partial\mathcal{B} \subset \Gamma$ is a subset of the phase space with dimension D_b satisfying $D-1 \leq D_b < D$ that separates a basin of attraction $\mathcal{B}$ from the rest of the phase space.

Topologically, the basin boundary is the boundary of the set $\mathcal{B}$.

In the simple case, the basin boundary is a smooth manifold with dimension $D_b = D-1$. For example, you need a 2D surface to divide a 3D space (think of a curtain separating a room) and a 1D line to divide a 2D plane (draw a line across a sheet of paper). However, particularly in chaotic systems, basin boundaries can be *fractal* with a non-integer *fractal dimension* (Grebogi et al. 1983b; Lucarini and Bódai 2017; Mehling et al. 2024). As we will see, this has important implications for the predictability of the system.

Stability landscape. When studying the global stability of multistable systems, common questions include: how large is the basin of attraction of each attractor? How stable is each basin relative to the others? How resilient is the system to perturbations, and how will it respond? These questions can be elegantly addressed via the concept of a stability landscape, which we introduced in section 1.1.3.

Generally, a stability landscape is defined as a scalar function $\mathcal{V} : \Gamma \to \mathbb{R}$ defined over the phase space, just like a topological map would map the position $(x, y) \in \mathbb{R}^2$ to the elevation $h \in \mathbb{R}$ at that position. However, the picture of the system evolving downhill in the direction of steepest descent (recall the analogy of a rolling ball in section 1.1.3) is only strictly true if the dynamics $\boldsymbol{b}(\boldsymbol{x}, \lambda)$ can be derived from a *potential* $V : \mathbb{R}^D \to \mathbb{R}$, such that the system is a *gradient* system.

Term 2.8. **Gradient system, Nongradient system**

The frozen system $\dot{\boldsymbol{x}} = \boldsymbol{b}(\boldsymbol{x}, \lambda)$ is *gradient* if and only if a potential $V : \mathbb{R}^D \to \mathbb{R}$ exists such that

$$\boldsymbol{b}(\boldsymbol{x}, \lambda) = -\boldsymbol{\nabla} V(\boldsymbol{x}, \lambda)\,. \tag{2.7}$$

Else, the system is *nongradient*.

If such a potential exists, V is the natural choice for the stability landscape $\mathcal{V}$, which is also called *energy landscape* since the elevation of the landscape represents the system's potential or free energy at the given state. The system thus tends to minimize its energy, and the height of the mountain crests between valleys quantifies the energy barrier that prevents the system to move from one valley to another.

One-dimensional systems are necessarily gradient. Also higher-dimensional physical systems may be gradient in some cases, e.g. in thermodynamic equilibrium, but most real-world systems are nongradient. The climate system is a chaotic system out of equilibrium and thus certainly nongradient. Can we still find a scalar function $\mathcal{V}$ that quantifies the energetic landscape and global stability of nongradient systems?

The answer is yes, in principle. Instead of Eq. (2.7), the dynamics are more generally given by

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\boldsymbol{\nabla}\mathcal{V}(\boldsymbol{x},\lambda) + \mathcal{R}(\boldsymbol{x},\lambda)\,, \tag{2.8}$$

where $\mathcal{V}$ is the *quasipotential* and the rest term $\mathcal{R}$ accounts for nongradient effects. Note that if $\mathcal{R} = 0$, then $\mathcal{V} = V$.

Term 2.9. **Quasipotential I**

The *quasipotential* $\mathcal{V} : \Gamma \to \mathbb{R}$ of a deterministic dynamical system can be defined, provided that it exists, via a decomposition of the drift field $\boldsymbol{b}$ as

$$\boldsymbol{\nabla}\mathcal{V}(\boldsymbol{x},\lambda) = \mathcal{R}(\boldsymbol{x},\lambda) - \boldsymbol{b}(\boldsymbol{x},\lambda)\,, \tag{2.9}$$

satisfying the orthogonality condition,

$$\langle \boldsymbol{\nabla}\mathcal{V}, \mathcal{R}\rangle = 0\,. \tag{2.10}$$

Here $\langle\cdot,\cdot\rangle$ denotes the scalar product. See also definitions 2.13 and 2.17.

Then, it can be shown that $\mathcal{V}$ is a *Lyapunov function* (Zhou et al. 2012), i.e., along a trajectory of the system in forward time,

$$\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}t} \leq 0\,, \tag{2.11}$$

where the equality is only attained at attractors. Lyapunov functions have key properties that ensure the desired behavior of a stability landscape: Eq. (2.11) implies that unforced trajectories will always travel "downhill" in $\mathcal{V}$ and that the attractors of $\boldsymbol{b}$ correspond to local minima of $\mathcal{V}$. The quasipotential thus offers a scalar function in phase space that generalizes the notion of a potential to nongradient systems. The term $\mathcal{R}$ describes motions along isolines of $\mathcal{V}$ and hence does not contribute energetically.

Mathematically, the quasipotential is the solution of the Hamilton-Jacobi equation,

$$\langle \boldsymbol{\nabla}\mathcal{V}, \boldsymbol{b} + \boldsymbol{\nabla}\mathcal{V}\rangle = 0\,. \tag{2.12}$$

This can be seen by combining Eqs. (2.10) and (2.9).

Numerous numerical algorithms exist to solve the Hamilton-Jacobi equation in simple systems (Cameron 2012; Dahiya and Cameron 2018a; Grafke and Vanden-Eijnden 2019). In high dimensions the calculation is usually prohibitive and a solution is not guaranteed to exist. Yet, even if the quasipotential cannot be explicitly computed for a given system, it may still provide a useful mental picture for reasoning about the global stability of the system.

2.1.4 Saddles, edge states and Melancholia states

In a multistable system, the basin boundaries are crucial objects in phase space, since they mark thresholds where the asymptotic behavior of the system changes dramatically. Two initial conditions on either side of the basin boundary will converge to different attractors, even if the initial conditions are very close to each other. What happens when you initialize the system precisely *on* the basin boundary?

On the basin boundary $\partial\mathcal{B} \subset \Gamma$, the flow of the system is governed by the drift field $\boldsymbol{b}$ like in the rest of the phase space. When restricting the dynamics to $\partial\mathcal{B}$, it is common that there are also attractors on $\partial\mathcal{B}$. Such sets are called *edge states*, or *Melancholia states*, or *saddles* – pick your favorite name. They attract initial conditions within the basin boundary but are globally unstable, since any finite perturbation away from the boundary would cause the system to converge to one of the attractors of the full phase space.

Term 2.10. **Saddle**

A *saddle point* $\boldsymbol{x}_M \in \mathbb{R}^{\mathbb{D}}$ of the frozen system (Eq. (2.3)) is an equilibrium state satisfying $\boldsymbol{b}(\boldsymbol{x}_M, \lambda) = 0$ where Jacobian of $\boldsymbol{b}$ evaluated at $\boldsymbol{x}_M$ has eigenvalues with negative and positive real parts, respectively.

A *chaotic saddle* $\boldsymbol{M} \subset \Gamma$ is a non-attracting chaotic invariant set in phase space that has both stable and unstable directions, i.e. it is neither an attractor nor a repeller.

Saddles thus have both a *stable set*, i.e. a set of states attracted by the saddle, and an *unstable set*, a set of states repelled away from the saddle. If these sets are manifolds, they are also called stable and unstable manifold, respectively.

We are especially interested in saddles that have only one unstable direction, since their stable sets are able to divide the phase space.

Term 2.11. **Melancholia state (Edge state)**

An *edge state* $\boldsymbol{M} \subset \partial\mathcal{B}$ of a multistable system of the form (2.3) is a non-attracting invariant set embedded in the basin boundary $\partial\mathcal{B}$ that acts as an attractor of the dynamics restricted to $\partial\mathcal{B}$.

An edge state is thus a saddle whose stable set divides the phase space.

The name *Melancholia state* (M state) is used as a synonym for edge state in the context of chaotic dynamics of the Earth system.

In other words, the stable set of the edge state is a subset of the basin boundary, and its unstable set traces the relaxation paths the system would take to the attractors of each of the adjacent basins of attraction. A system may possess multiple edge states, in which case the dynamics restricted on the basin boundary are multistable themselves.

Even though saddles are unstable, they can dominate the transient dynamics of the

system. Particularly, chaotic saddles are associated with transient chaos (Lai and Tél 2011; Szabó and Tél 1994; Morozov et al. 2023; Mehling et al. 2024), as we explore in detail in this thesis.[7]

2.1.5 The stochastic view

Adding noise changes things fundamentally. In the previous sections, we have focused on deterministic, autonomous systems to introduce key concepts of dynamical systems theory for multistable systems. In the absence of external forcing and random perturbations, a multistable system will never leave the basin of attraction it has been initialized in. But as soon as we add stochastic forcing, transitions from one basin to another become possible. The meaning behind adding noise in climate models is discussed in section 2.1.9.

Let us now turn on the noise in Eq. (2.1), i.e. we consider noise strengths $\sigma > 0$ with otherwise fixed external forcing ($\Lambda(t) = \lambda$), yielding the stochastic differential equation (SDE) of the Itô type,[8]

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{x})\mathrm{d}t + \sigma\boldsymbol{\Sigma}(\boldsymbol{x})\mathrm{d}\boldsymbol{W}_t\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0\,. \tag{2.13}$$

Here the noise process $\boldsymbol{\mathcal{N}}_t$ is a D-dimensional standard Wiener process $\boldsymbol{W}_t$ (i.e. Gaussian white noise with mean zero and variance one). Correlations between variables and a possible state-dependence of the noise are encapsulated in the covariance matrix $\boldsymbol{Q}(\boldsymbol{x}) = \boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{\Sigma}(\boldsymbol{x})^\top$. In this work, we will restrict to non-degenerate noise, which means that the noise is non-vanishing in all D degrees of freedom of the system[9]. In Eq. (2.13) the noise is non-degenerate if and only if the covariance matrix is positive definite. This also implies that $\boldsymbol{Q}$ is invertible. For a discussion on the relevance of choosing the Îto or Stratonovich interpretation of SDEs in a climate context, see Kypke and Ditlevsen (2024).

Suppose that the deterministic dynamics $\dot{\boldsymbol{x}} = \boldsymbol{b}(\boldsymbol{x})$ are multistable (see definition 2.6). The deterministic drift will pull a trajectory towards the attractor of the basin it was initialized in. By contrast, the stochastic perturbations can push the trajectory in different directions, according to the covariance matrix. Assuming unbounded noise, we will eventually observe a realization of the Wiener process that pushes the system across the basin boundary. As quantified below, the time we have to wait depends on the height of the quasipotential along the basin boundary.

Over time, a trajectory of Eq. (2.13) forced by unbounded noise will thus explore the entire phase space, visiting all competing basins of attraction. This is true for any finite noise $\sigma > 0$, but of course exploring remote regions of the phase space takes longer if the noise is weak. The probability distribution $P_\sigma(\boldsymbol{x}, t)$ of the system being in state $\boldsymbol{x}$ at time t is given by the *Fokker-Planck equation* (Risken 1996),

$$\frac{\partial P_\sigma(\boldsymbol{x},t)}{\partial t} = -\frac{\partial}{\partial x_i}\big(b_i(\boldsymbol{x})P_\sigma(\boldsymbol{x},t)\big) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i \partial x_j}\big(Q_{ij}(\boldsymbol{x})P_\sigma(x,t)\big)\,, \tag{2.14}$$

[7] The term *Melancholia state* was coined by Valerio Lucarini in reference to the film "Melancholia" by Lars von Trier, which he was forced to watch (Lucarini and Bódai 2017). The film masterfully portrays how the fate of a dynamical system (Earth and a rogue planet) becomes uncertain as it lingers near an edge state. While the need for an additional name has been questioned in the community, it certainly is a beautiful name. Hopefully, definitions 2.10 and 2.11 add clarity to the terminology. Throughout the thesis, we will use the terms interchangeably, depending on how melancholic we feel.

[8] In this section we drop the argument λ in the notation.

[9] Even if the noise is degenerate, one might assume that the random forcing propagates onto all degrees of freedom if the system is sufficiently mixing.

where $\boldsymbol{b} = (b_1, \ldots, b_D)$ and $\boldsymbol{Q} = (Q_{ij})$, and we have used Einstein's summation convention. Let us denote by $\boldsymbol{\rho}_\sigma = P_\sigma(\boldsymbol{x}, t \to \infty)$ the steady-state (time-independent) solution of the Fokker-Planck equation.

Term 2.12. **Invariant measure**

The *invariant measure* $\boldsymbol{\rho}_\sigma : \Gamma \to \mathbb{R}$ of a stochastic dynamical system is its steady-state probability density in phase space for given noise strength σ.

Assuming that the system is *ergodic*, it makes no difference whether we calculate the expectation of an observable from a (sufficiently long) time average along a trajectory or a space average over the invariant measure, according to Birkhoff's ergodic theorem (Birkhoff 1931; Tantet 2016).

If the noise is large compared to the drift, the system will move around almost like a particle under free diffusion. The weaker the noise, the more the structure of the deterministic drift field becomes apparent. For multistable systems driven by weak noise, the invariant measure is characterized by multiple peaks at the locations of the (deterministic) attractors, since the system tends to spend longer times near them due to the attracting deterministic forces. This motivates using the invariant measure to define a quasipotential for stochastic dynamical systems.

Term 2.13. **Quasipotential II**

The *quasipotential* $\tilde{\mathcal{V}} : \Gamma \to \mathbb{R}$, or *nonequilibrium potential*, of a system governed by Eq. (2.13) is obtained in the weak-noise limit ($\sigma \to 0$) by (Graham and Tél 1986; Graham and Tél 1984)

$$\tilde{\mathcal{V}}(\boldsymbol{x}) = -\lim_{\sigma \to 0} \sigma^2 \ln \boldsymbol{\rho}_\sigma(\boldsymbol{x}) \,. \tag{2.15}$$

See also definitions 2.9 and 2.17.

This quasipotential solves the Hamilton-Jacobi equation,

$$Q_{ij}(\boldsymbol{x}) \frac{\partial \tilde{\mathcal{V}}}{\partial x_i} \frac{\partial \tilde{\mathcal{V}}}{\partial x_j} + b_i(\boldsymbol{x}) \frac{\partial \tilde{\mathcal{V}}}{\partial x_i} = 0 \,, \tag{2.16}$$

which is equivalent to the previously derived Hamilton-Jacobi equation (2.12) if the covariance matrix is proportional to the unit matrix, i.e. $\boldsymbol{Q} \propto \mathbb{I}_D$ (isotropic noise). For positive definite $\boldsymbol{Q}$ (sometimes even positive semidefinite $\boldsymbol{Q}$, see Graham and Tél (1986)), it can be shown that the quasipotential defined in Eq. (2.15) is also a Lyapunov function (see Eq. (2.11)).

The quasipotential $\tilde{\mathcal{V}}$ has minima at the attractors of $\boldsymbol{b}$ and saddle properties at saddles of $\boldsymbol{b}$ (Graham and Tél 1986). The quasipotantial heights at the attractors quantify how stable the attractors are relative to each other, and the quasipotential height difference between an attractor and the lowest region on its basin boundary – typically a saddle or

unstable limit cycle – represents the energy barrier that must be overcome to exit the basin of attraction (see Kramers' law, definition 2.18).

In the following, we make the discussion of stochastic dynamical systems driven by weak Gaussian noise more mathematically precise. Under weak noise, excursions of the system from an attractor qualify as *rare events*, which brings us into the realm of *large deviation theory*.

2.1.6 Large deviation theory

How can you predict the behavior of a random dynamical system? This seems impossible, given that randomness is by definition unpredictable. On the other hand, rare events become increasingly predictable the more rare they become. This is because they tend to happen in the least unlikely way, an idea that is rigorously formulated by large deviation theory (Freidlin and Wentzell 1998; Touchette 2009).

Consider a stochastic dynamical system described by Eq. (2.13). We are interested in the probability of observing a trajectory $\boldsymbol{\phi}_t^{IF}$ that starts at the state $\boldsymbol{x}_I$ at time $t = 0$ and ends at $\boldsymbol{x}_F$ at $t = T$,

$$\boldsymbol{\phi}_t^{IF} := \{x(t) : 0 \leq t \leq T,\, \boldsymbol{x}(t_i) = \boldsymbol{x}_I,\, \boldsymbol{x}(t_f) = \boldsymbol{x}_F\}\,.$$

To this end, let us define the set $\mathcal{C}_{ij}^T$ of all (twice differentiable) paths in phase space leading from $\boldsymbol{x}_I$ to $\boldsymbol{x}_F$ in time T:

$$\mathcal{C}_{IF}^T := \{\boldsymbol{\varphi}_t : 0 \leq t \leq T,\, \boldsymbol{\varphi}_0 = \boldsymbol{x}_I,\, \boldsymbol{\varphi}_T = \boldsymbol{x}_F\}\,.$$

It will further become useful to define the notion of pathspace.

Term 2.14. **Pathspace**

The *pathspace* $\Pi := \mathcal{C}([I, F], \Gamma)$ associated with a stochastic dynamical system is the function space of continuous paths that live in its phase space Γ, parameterized over an interval $[I, F]$.

In the weak-noise limit ($\sigma \to 0$), the probability that the trajectory $\boldsymbol{\phi}_t^{IF}$ remains inside a tube of radius δ around the path $\boldsymbol{\varphi}_t \in \mathcal{C}_{IF}^T$ (in phase space) follows

$$\mathbb{P}_\sigma\left(\sup_{0\leq t\leq T} ||\boldsymbol{\phi}_t - \boldsymbol{\varphi}_t|| < \delta\right) \asymp \exp\left(-\frac{S_T[\boldsymbol{\varphi}_t]}{\sigma^2}\right)\,, \tag{2.17}$$

where $||\cdot||$ denotes a suitable norm in phase space (e.g. the Euclidean norm) and the symbol $\asymp$ denotes asymptotic logarithmic equivalence, i.e., $a \asymp b(\sigma) \Leftrightarrow \lim_{\sigma\to 0} \ln(a)/\ln(b(\sigma)) = 1$ (Freidlin and Wentzell 1998).

Eq. (2.17) has the form of a so-called *large deviation principle*. It states that some paths in $\mathcal{C}_{IF}^T$, namely those with a relatively small value of the (non-negative) functional S_T, become exponentially more likely than other paths as the noise goes to zero. This functional is the *Freidlin-Wentzell action*.

Term 2.15. **Freidlin-Wentzell action**

The *Freidlin-Wentzell (FW) action* functional $S_T : \Pi \to \mathbb{R}_{>0}$ is given by

$$S_T[\boldsymbol{\varphi}_t] = \frac{1}{2}\int_0^T ||\dot{\boldsymbol{\varphi}}_t - \boldsymbol{b}(\boldsymbol{\varphi}_t)||^2_{\boldsymbol{Q}(\boldsymbol{\varphi}_t)}\mathrm{d}t\,, \tag{2.18}$$

where the $\boldsymbol{Q}$-metric is defined as $||\boldsymbol{v}||_{\boldsymbol{Q}} := \sqrt{\langle \boldsymbol{v}, \boldsymbol{Q}^{-1}\boldsymbol{v}\rangle}$.

Here $\dot{\boldsymbol{\varphi}}_t$ denotes the time derivative along the path $\boldsymbol{\varphi}_t$ and $\boldsymbol{b}(\boldsymbol{\varphi}_t)$ is the drift along the path. The FW action thus measures how much the path deviates from the deterministic flow of the system, weighted by the covariance matrix. In other words, the action quantifies the amount of "work" done against the drift. Along deterministic trajectories, the FW action vanishes.

The large deviation principle (Eq. (2.17)) tells us that, in the weak-noise limit, the by far most probable way to get from $\boldsymbol{x}_I$ to $\boldsymbol{x}_F$ is via the path that minimizes the FW action. It is the least costly path in the sense of the noise performing minimal "work" against the deterministic drift. Deviations from this path become exponentially unlikely. (For finite noise, the appropriate variational formulation is given in terms of the Onsager-Machlup action functional; see section 2.2.2 and chapter 4.)

Term 2.16. **Instanton**

The *instanton* $\boldsymbol{\varphi}^*_{IF}$ is the *minimum action path*,

$$\boldsymbol{\varphi}^*_{IF} := \underset{\boldsymbol{\varphi}_t \in \mathcal{C}^T_{IF},\, T>0}{\arg\min}\; S_T[\boldsymbol{\varphi}_t]\,, \tag{2.19}$$

leading from $\boldsymbol{x}_I$ to $\boldsymbol{x}_F$ (Freidlin and Wentzell 1998; Grafke and Vanden-Eijnden 2019).

In the weak-noise limit, the instanton is the *most probable transition path* from $\boldsymbol{x}_I$ to $\boldsymbol{x}_F$.

Note that the minimization in Eq. (2.19) is performed over all paths in $\mathcal{C}^T_{IF}$ and over all times $T > 0$.

Freidlin-Wentzell (FW) theory thus provides a variational principle that allows to predict transition paths in the weak-noise limit based on the drift $\boldsymbol{b}$ and the covariance matrix $\boldsymbol{Q}$. The action value of the instanton quantifies how "difficult" it is to reach $\boldsymbol{x}_F$ from $\boldsymbol{x}_I$. In principle, if we compute the instantons from a reference point to all points in phase space and map out their action values, we obtain an energetic landscape of the system. This motivates defining a quasipotential in terms of the FW action.

Term 2.17. **Freidlin-Wentzell quasipotential**

The *FW quasipotential* $V_A : \Gamma \to \mathbb{R}$ with respect to an attractor $\boldsymbol{A}$ is defined as

$$V_{\boldsymbol{A}}(\boldsymbol{x}_F) = \inf_{T>0} \inf_{\boldsymbol{\varphi}_t \in \mathcal{C}^T_{AF}} S_T[\boldsymbol{\varphi}_t]\,, \tag{2.20}$$

where $\mathcal{C}^T_{AF}$ is the set of all paths leading from a point $\boldsymbol{x}_A \in \boldsymbol{A}$ to $\boldsymbol{x}_F$ in time T.

On the reference attractor, we have $V_{\boldsymbol{A}}(\boldsymbol{x}_A) = 0$, which is the only strict local minimum and the strict global minimum in the phase space. Although the FW quasipotential is usually defined with respect to an attractor, the reference point could generally be any point.

As it turns out, the FW quasipotential allows an orthogonal decomposition in analogy to Eq. (2.9), which for isotropic noise ($\boldsymbol{Q} = \mathbb{I}_D$) reads

$$\boldsymbol{b}(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{\nabla} V_{\boldsymbol{A}}(\boldsymbol{x}) + \mathcal{R}_{\boldsymbol{A}}(\boldsymbol{x})\,, \qquad \langle \boldsymbol{\nabla} V_{\boldsymbol{A}},\, \mathcal{R}_{\boldsymbol{A}} \rangle = 0\,. \tag{2.21}$$

The rest term $\mathcal{R}_{\boldsymbol{A}}$ embodies the nongradient nature of $\boldsymbol{b}$. This decomposition shows that deterministic trajectories evolve "downhill" in the FW quasipotential landscape whilst also moving along isolines of $V_{\boldsymbol{A}}$ due to $\mathcal{R}_{\boldsymbol{A}}$ (a bit like geostrophic "wind", for friends of geophysical fluid dynamics). In gradient systems, the instanton from $\boldsymbol{A}$ to a point $\boldsymbol{y}$ in its basin of attraction leads directly "uphill" in the potential V, i.e. it solves the equation $\dot{\boldsymbol{x}} = \boldsymbol{\nabla} V(\boldsymbol{x})$. In nongradient systems (with $\boldsymbol{Q} = \mathbb{I}_D$), the corresponding instanton equation is

$$\dot{\boldsymbol{x}} = \frac{1}{2}\boldsymbol{\nabla} V_{\boldsymbol{A}}(\boldsymbol{x}) + \mathcal{R}_{\boldsymbol{A}}(\boldsymbol{x})\,, \tag{2.22}$$

so the instanton does not climb directly "uphill" due to the orthogonal "wind" $\mathcal{R}_{\boldsymbol{A}}$ which diverts the path in the same direction as for deterministic relaxations (Cameron 2012). As a result, the instanton from $\boldsymbol{A}$ to $\boldsymbol{y}$ is generally not the reverse of the relaxation path from $\boldsymbol{y}$ to $\boldsymbol{A}$. The property $\mathcal{R}_{\boldsymbol{A}} \neq 0$ often – though not necessarily – breaks *detailed balance*, meaning that a nonzero probability current J_{eq} remains also in a statistical steady-state. (Here J_{eq} follows from the steady-state solution of the Fokker-Planck equation (2.14), $\partial P(x,t)/\partial t = -\boldsymbol{\nabla} \cdot J(x,t)$ and detailed balance requires $J_{\mathrm{eq}}(x) = 0\ \forall x$.) However, detailed balance can also be broken by the noise term in gradient systems, and in special cases nongradient systems may satisfy detailed balance.

How long does it take on average until a trajectory initialized on an attractor escapes from its basin of attraction? Arrhenius (1889) found that the duration depends on $\exp(\Delta/\sigma^2)$, where Δ is the height of the energy barrier that must be overcome. For gradient systems, the *mean first-exit time* $\langle \tau_\sigma \rangle$ is given by the Eyring-Kramers' law (Eyring 1935; Kramers 1940; Berglund 2013),

$$\langle \tau_\sigma \rangle \asymp C(V) \cdot \exp\left(\frac{\Delta V}{\sigma^2}\right)\,, \tag{2.23}$$

where C is a pre-exponential factor and $\Delta V = V(\boldsymbol{x}_M) - V(\boldsymbol{x}_A)$ is the potential difference between the attractor $\boldsymbol{A}$ and the saddle $\boldsymbol{x}_M$ with the lowest potential height on the basin boundary $\partial\mathcal{B}(\boldsymbol{A})$. The exponential scaling is known as *Arrhenius' law*. The quasipotential allows to generalize this result to nongradient systems.

Term 2.18. **Arrhenius' law: Mean first-exit time**

In nongradient systems described by Eq. (2.13), the *mean first-exit time*, i.e. the expected time for a trajectory initialized at $\boldsymbol{x}_A \in \boldsymbol{A}$ to reach the basin boundary $\partial\mathcal{B}(\boldsymbol{A})$ scales like

$$\langle \tau_\sigma \rangle \asymp \exp\left(\frac{\min_{\boldsymbol{x}\in\partial\mathcal{B}(\boldsymbol{A})} V_A(\boldsymbol{x})}{\sigma^2} \right) . \tag{2.24}$$

The minimum of V_A on $\partial\mathcal{B}$ is typically attained at an edge state (a saddle of the deterministic drift $\boldsymbol{b}$). The inverse of $\langle \tau_\sigma \rangle$ quantifies the *escape rate* from the basin of attraction of $\boldsymbol{A}$.

In the case where the attractors and edge states of the system are fixed points, the pre-exponential factor in Eq. (2.24) has been derived by Bouchet and Reygner (2016) and Bouchet and Reygner (2022); see also Maier and Stein (1997) and Berglund (2013).

If we interpret the quasipotential as an energetic landscape, we obtain a very intuitive picture of the noise-induced escape. Imagine you are a hiker, standing at the bottom of the quasipotential valley at the attractor $\boldsymbol{A}$. How do you get out of the valley most efficiently? Clearly not by climbing the highest peak, but by taking the *mountain pass* – the lowest saddle point along the mountain crest surrounding the valley. This mountain pass is the minimum of V_A in Eq. 2.24, motivating the idea of edge states as gateways for noise-induced transitions out of a basin of attraction (Kraut and Feudel 2003; Margazoglou et al. 2021). The fact that this picture can be misleading is the subject of chapter 4.

Local vs. global quasipotential. We have introduced three different definitions of a quasipotential, each presenting an energy-like scalar field that generalizes the notion of a potential landscape to nongradient systems (definitions 2.9, 2.13 and 2.17). Are they all the same?

In a deterministic context, the quasipotential $\mathcal{V}$ in Eq. (2.9) was derived via the composition of the drift into a gradient term and an orthogonal rest term (known as the normal decomposition). Under stochastic forcing with Gaussian noise, Graham and Tél (1986) defined a quasipotential $\tilde{\mathcal{V}}$ in terms of the invariant measure by taking the weak-noise limit (Eq. (2.15)). As shown, these quasipotentials both solve the same Hamilton-Jacobi equation under the restriction to isotropic noise (identity covariance matrix $\boldsymbol{Q}$). The Hamilton-Jacobi equation offers a general framework for constructing quasipotential functions for nongradient systems that satisfy the properties of a Lyapunov function (Zhou et al. 2012). However, the existence and uniqueness of a solution is not guaranteed (Cameron 2012). Note that the quasipotential $\tilde{\mathcal{V}}$ derived via the invariant measure may also be written as a decomposition into a gradient term plus a term that involves the steady-state probability current (Zhou et al. 2012). This establishes a link between the nongradient property and the lack of *detailed balance* characterized by a non-vanishing steady-state probability current.

The quasipotentials $\mathcal{V}$ and $\tilde{\mathcal{V}}$ are *global* quasipotentials characterizing the stability landscape in the whole phase space, with local minima at each attractor. By contrast, the FW quasipotential V_A (Eq. (2.20)) is a *local* quasipotential: it depends on the reference attractor. Within the basin of attraction of the reference attractor $\boldsymbol{A}$, it can be verified that

$$V_A(\boldsymbol{x}) = 2\mathcal{V}(\boldsymbol{x}) + c \text{ for all } \boldsymbol{x} \in \mathcal{B}(\boldsymbol{A}) ,$$

where c is a constant[10] (Zhou et al. 2012). Outside of this basin, however, V_A no longer reflects the global stability landscape. This is because the FW quasipotential is constant along deterministic trajectories, rather than decreasing (since the FW action is then zero). Starting from $\boldsymbol{A}$, the FW quasipotential only captures the "uphill" parts of instantonic paths in the stability landscape. Thus, after crossing the basin boundary $\partial\mathcal{B}(\boldsymbol{A})$, the FW quasipotential is either constant at points to which a deterministic path exists or increases further. This emphasizes that $V_A(\boldsymbol{x})$ represents the "difficulty" to reach $\boldsymbol{x}$ from $\boldsymbol{A}$. The FW quasipotential still solves the Hamilton-Jacobi equation, but it is not globally a Lyapunov function since $\mathrm{d}V_A/\mathrm{d}t$ can be zero outside of $\mathcal{B}(\boldsymbol{A})$ along paths which do not evolve on an attractor.

In some cases, a global quasipotential may be constructed by stitching together local FW quasipotentials computed with respect to each attractor (Zhou and Li 2016). For example, in a bistable system with attractors $\boldsymbol{A}$ and $\boldsymbol{B}$ and one saddle point $\boldsymbol{x}_M$, we can calculate the quasipotentials V_A and V_B and align them along the basin boundary such that $V_A(\boldsymbol{x}_M) = V_B(\boldsymbol{x}_M)$. However, the resulting quasipotential will likely be discontinuous on the basin boundary. To avoid this, a meaningful global landscape may be constructed as

$$V_{\text{global}}(\boldsymbol{x}) = \min_k \big(V_k(\boldsymbol{x}) - V_k(\boldsymbol{x}_M)\big) + \max_k V_k(\boldsymbol{x}_M) \ , \tag{2.25}$$

where the index $k \in \{\boldsymbol{A}, \boldsymbol{B}\}$. The addition of the second term on the right hand side ensures that V_{global} is zero at the attractor with the deeper valley. If the system has more than two attractors and/or more than one edge state, it becomes less clear how to align local quasipotentials relative to each other.

For further theoretical background on quasipotentials, see the works of Freidlin and Wentzell (1998), Cameron (2012), Graham and Tél (1986), Zhou et al. (2012), Zhou and Li (2016), and Nolting and Abbott (2016).

Non-Gaussian noise. While classical Freidlin-Wentzell theory is derived under the assumption of Gaussian white noise, many random processes in nature are not well described by a normal distribution. Particularly, the observed distribution is often heavy-tailed compared to a normal distribution (Hurst 1951), as Ditlevsen (1999) showed for ice core data that records the climate of the past. Developing a large deviation theory for non-Gaussian noise and studying the behavior of climate models under non-Gaussian stochastic forcing are thus important research problems (Lucarini et al. 2022; Zheng et al. 2020).

A generalization to non-Gaussian noise processes is offered by the class of symmetric α-stable Lévy processes, where $\alpha \in (0, 2]$ is a parameter that characterizes the tails of the distribution decaying like $x^{-\alpha-1}$ (Imkeller and Pavlyukevich 2006b). Gaussian noise presents the limiting case $\alpha = 2$. While a Wiener process is continuous, α-stable processes with $\alpha < 2$ are continuous in probability but feature discontinuous jumps. As α decreases, the jumps become increasingly rare and extreme in magnitude.

The jumpy behavior of α-stable Lévy noise impacts the mechanism of escaping a basin of attraction in stochastic dynamical systems. If the process $\boldsymbol{\mathcal{N}}_t$ in Eq. (2.1) is an α-stable process, the mean first-exit time $\langle \tau_\sigma \rangle$ scales like (Imkeller and Pavlyukevich 2006a; Gao

[10] Note that generally potential functions are only defined up to an additive constant (which vanishes when taking the gradient), and we are only interested in the potential *differences* between points.

et al. 2014)

$$\langle \tau_\sigma \rangle \sim \sigma^{-\alpha}\,, \tag{2.26}$$

which is fundamentally different from the result in Eq. (2.24) obtained for Gaussian noise (Lucarini et al. 2022). Intuitively, the escape by means of a Wiener process is characterized by a sequence of small uphill steps in the quasipotential, whereas for Lévy noise with small α the escape occurs predominantly via single jumps across the basin boundary. For Gaussian noise it is thus the quasipotential height difference between attractor and edge state that matters, whereas for Lévy noise it is rather the distance between the attractor and the basin boundary in the direction of the jumps (Del Amo and Ditlevsen 2025). For a derivation of the minimum action path under Lévy flights, see Hu and Duan (2020).

2.1.7 Metastability, tipping and transient chaos

The picture of a stability landscape offers an intuitive view on the multistability of complex dynamical systems. Valleys of the landscape correspond to the basins of attraction of different attractors, each of which often represents a qualitatively different physical state. Transitions between attractors are thus typically associated with large-scale regime shifts – so-called *critical transitions*. But how does the system get from one valley to another? In this section, we discuss the mechanisms and phenomenology of critical transitions.

Multistability characterizes the global stability of a system possessing multiple attractors (see definition 2.6). In the absence of nonautonomous forcing, the system is bound to a single basin of attraction determined by the initial condition. By contrast, as soon as the system is subjected to Gaussian noise with finite amplitude, it will eventually hop from one basin to another. This implies that technically none of the attractors are asymptotically stable anymore; instead, the statistical steady state of the system is given by the invariant measure (see section 2.1.5). In short: multistability plus noise gives rise to *metastability*.

While the term metastability has been used differently depending on the context, we follow Rossi et al. (2025) in proposing a definition rooted in dynamical systems theory:

Term 2.19. **Metastability**

Metastability describes dynamical behavior characterized by

- multiple distinct dynamical regimes in which they spend long but ultimately transient episodes
- abrupt switches (critical transitions) between the different persistent regimes.

Here abruptness is understood in the sense that the transitions occur within a short time period compared to the long-lived transient episodes.

A *metastable* system is a dynamical system that supports metastability, i.e. it features one or more long-lived states that are not asymptotically stable. Here a state refers to a region in the phase space with distinct properties.

Metastability can arise in different ways. Besides stochastic forcing combined with multistability, also time-dependent deterministic forcing can induce critical transitions between regimes, as we discuss below. Even without external forcing, chaos (particularly chaotic saddles) and oscillatory dynamics with multiple timescales often cause metastable behavior (Rossi et al. 2025). All of these mechanisms play a role for climate dynamics.

Over the past years, the term *tipping point* has become popular in the context of metastable dynamics and transition behavior (Gladwell 2000; Lenton et al. 2008; Ashwin et al. 2012). In this work, we use a relatively broad definition.

Term 2.20. **Tipping point**

A *tipping point* is a threshold beyond which a dynamical system may undergo an abrupt, self-perpetuating and potentially irreversible change.

Here the threshold typically refers to a critical value of a control parameter of the system, such as the CO_2 concentration in the atmosphere. Let us break down the adjectives involved in the definition:

- *Abrupt* means that the change occurs within a short time period compared to the characteristic time scale of the system.
- The change is *self-perpetuating* if no further external forces are needed to drive the change once the threshold is crossed.
- If the change is *irreversible*, it cannot be undone by simply restoring the control parameter to its original value before crossing the threshold.

To ensure a precise language, we distinguish between a *tipping point* referring to the threshold and a *tipping event* (or just *tipping*) referring to the change itself.

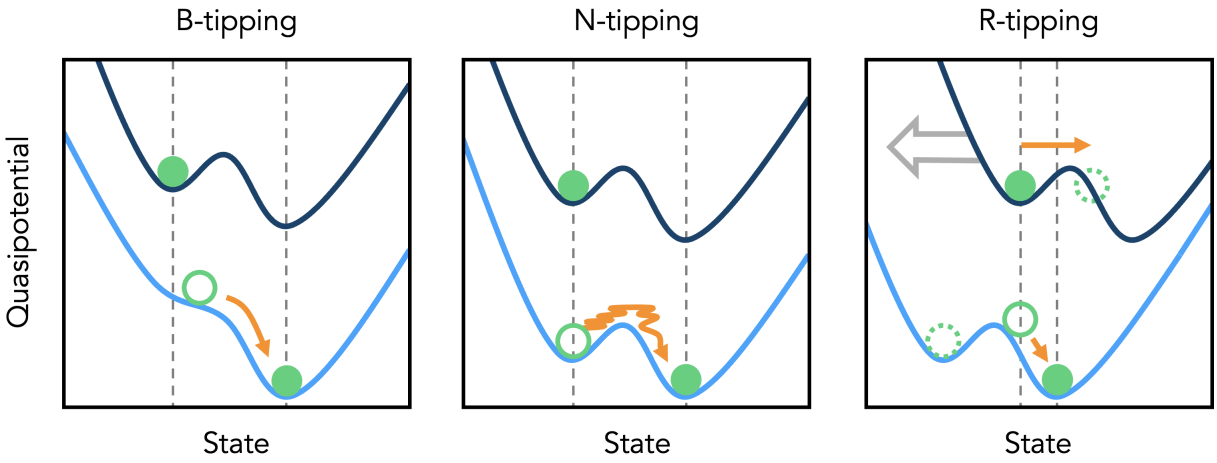

Figure 2.1 | **Bifurcation (B), noise (N) and rate-induced (R) tipping** in a schematic double-well quasipotential of a bistable system. The dark (light) blue outline illustrates the quasipotential before (after) the transition; the green ball represents the system's state. B-tipping: The quasipotential changes from bistable to monostable. N-tipping: The quasipotential remains unchanged but noise kicks the system over. R-tipping: The rapid change of a control parameter shifts the landscape, and the system fails to track its original attractor under the given rate of change.

B, R, and N-tipping. Tipping events are commonly classified in three categories, following Ashwin et al. (2012). Each category defines a different mechanism of tipping, but all are based on an underlying dynamical system with multistability. While the classification is useful for the mathematical treatment of tipping behavior, in reality there is no clear distinction between these mechanisms and they may work in concert (Slyman and Jones 2023).

Term 2.21. **Bifurcation-induced tipping (B-tipping)**

A *bifurcation-induced transition* occurs when an external control parameter is varied quasi-adiabatically (i.e. very slowly) across a bifurcation point where the current attractor supporting the system loses stability.

A *bifurcation* is a qualitative change in the global stability properties of the dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{b}(\boldsymbol{x}, \lambda)$ occurring at a critical value $\lambda = \lambda_c$ of the control parameter.

B-tipping presents the most classical scenario that can be understood in terms of equilibrium dynamics, since the quasi-adiabatic drift of the control parameter ensures that the system continuously tracks an equilibrium branch, i.e. it follows the attractor up to the bifurcation point. As illustrated in Fig. 2.1 for the case of a saddle-node bifurcation, the quasipotential around that attractor becomes shallower as the bifurcation is approached, reflecting that the attractor gradually loses stability. This leads to *critical slowing-down*: as the restoring forces of the system weaken, the response to perturbations becomes more sluggish, which can be measured as an increase in variance and autocorrelation along the timeseries of the trajectory (Carpenter and Brock 2006). Under certain assumptions on the dynamical properties of the system, this feature can be exploited to detect *early warning signals* of approaching a tipping point (Dakos et al. 2008). However, critical

slowing-down indicators do not work for all types of bifurcations.

Term 2.22. **Rate-induced tipping (R-tipping)**

A *rate-induced transition* occurs when an external control parameter is varied beyond a critical rate such that the trajectory of the nonautonomous system fails to track the attractor it was initialized in.

Unlike B-tipping, R-tipping is a truly nonautonomous phenomenon (Wieczorek et al. 2023; Feudel 2023). Due to the time-dependent forcing, the properties of the attractors and their location in phase space generally changes over time. Instead of invariant attractors of the frozen system, we must consider time-dependent *snapshot attractors* (Drótos et al. 2015), and the system may be driven far from equilibrium. This complicates the detection of early warning signs (Lohmann et al. 2021), though several methods have been proposed (Ritchie and Sieber 2016; Huang et al. 2024).

Term 2.23. **Noise-induced tipping (N-tipping)**

A *noise-induced transition* occurs in noise-driven dynamical systems when a realization of the stochastic forcing kicks the trajectory from one basin of attraction (of the deterministic dynamics) to another, leading to a critical transition.

To understand and predict N-tipping, large deviation theory (section 2.1.6) provides a suitable framework provided that the noise is sufficiently weak (Horsthemke and Lefever 1984; Börner et al. 2024b). Of course, predictions can only be made in a probabilistic sense (Ditlevsen and Johnsen 2010).

In addition to the three types of tipping introduced above, *shock-induced tipping* (S-tipping) describes the case where a single perturbation kicks the system from an attractor outside of its basin of attraction, leading to a critical transition (Halekotte and Feudel 2020). S-tipping relates to a common notion of *resilience* of a system: what is the minimal perturbation that causes the system to fail to return to the initial state after the perturbation (Holling 1973; Krakovská et al. 2023)? This problem becomes especially relevant in the presence of non-Gaussian noise with a heavy-tailed distribution, such as Lévy noise (see sections 2.1.6 and 3.3.2).

Chaotic transients. Suppose you observe a critical transition in a timeseries. After spending a long time in a certain dynamical regime, the timeseries suddenly switches to a different regime. Did the system cross a bifurcation? Did it tip due to high rates of forcing changes? Or did noise induce the transition? None of these might be the case. Instead, the transition could be due to *transient chaos*.

Chaotic systems can be metastable even under fixed external forcing in a fully deterministic setting. This is because the system's structure in the phase space is often highly complex, exhibiting regions resembling a maze where trajectories linger for extended periods of time but eventually escape (Lai and Tél 2011; Morozov et al. 2023).

Term 2.24. **Chaotic transient**

A *chaotic transient* is a trajectory exhibiting chaotic behavior of a finite lifetime due to non-attracting chaotic sets in the phase space of the system.

This definition based on Lai and Tél (2011) highlights the prominent role of chaotic saddles, which together with chaotic repellers make up the group of non-attracting chaotic sets. Even though chaotic saddles are asymptotically unstable, they possess a stable set that attracts initial conditions in a subset of the phase space (see section 2.1.4). Trajectories initialized near the stable set may thus stay in the vicinity of the saddle for ultralong yet finite times. In principle, since we can only observe systems for a finite time, we will never be able to distinguish whether a long-lived state is transient or asymptotically stable.

When studying chaotic transients, a basic question is: how long is the transient? Since this typically depends sensitively on the initial condition, we are concerned with the lifetime distribution of an initial condition ensemble that covers a phase space volume Σ enclosing the non-attracting chaotic set (e.g. the saddle). To this end, we measure the number N of ensemble members remaining in Σ at time t. As it turns out, N decays exponentially,

$$N(t) = N_0 e^{-\kappa t} \ , \tag{2.27}$$

where N_0 is the ensemble size and κ denotes the *escape rate* of the saddle (Yorke and Yorke 1979). Hence, lifetime τ near the chaotic saddle is exponentially distributed with a *mean lifetime* $\langle \tau \rangle = 1/\kappa$ (Ott 2002; Lai and Tél 2011).

Non-attracting chaotic sets also appear in the context of *crises* (Grebogi et al. 1983a; Szabó and Tél 1994).

Term 2.25. **Boundary crisis**

A *boundary crisis* marks the destruction of a chaotic attractor and its basin of attraction as a control parameter λ is varied across a bifurcation point marked by a critical value λ_c.

Typically, the crisis occurs when the attractor collides with an edge state on its basin boundary.

Even though the attractor is annihilated at the boundary crisis, its geometric properties in the phase space are often reflected in the dynamics for parameter values greater than but close to λ_c (without loss of generality, assume the attractor exists for $\lambda < \lambda_c$).

Term 2.26. **Ghost state**

Within a parameter range $[\lambda_c, \lambda_c + w]$, for some $w > 0$, the non-attracting (almost-invariant) chaotic set remaining in place of the former attractor that was annihilated at the boundary crisis is called a *ghost state*.

Though ghost states (Yeung and Strogatz 1998; Deco and Jirsa 2012; Morozov et al. 2023; Koch et al. 2024) are not invariant sets, the slow dynamics near a ghost state can induce long transients. For λ greater than but close to λ_c, the lifetimes of the transients are exponentially distributed, where the mean lifetime is related to the distance to the bifurcation via a power law (Grebogi et al. 1986):

$$\langle \tau \rangle \propto |\lambda - \lambda_c|^{-\eta} . \tag{2.28}$$

Here η is the critical exponent. This shows that chaotic transients due to the ghost state become infinitely long as the bifurcation point is approached "from behind". Consequently, it is practically impossible to precisely determine λ_c from simulations of finite duration (Mehling et al. 2024).

As proposed by Kantz and Grassberger (1985), the lifetime of a non-attracting chaotic set can be combined with its Lyapunov spectrum (definition 2.5) to gain further insight into the phase space structure, particularly the dimension of the chaotic set (Hunt et al. 1996; Sweet and Ott 2000). Particularly, if the escape rate κ is significantly smaller than the maximum Lyapunov exponent γ_1, the *fractal dimension* of the nonattracting chaotic set is given by (Bódai and Lucarini 2020),

$$D_b = D - \frac{\kappa}{\gamma_1} . \tag{2.29}$$

The smaller the ratio κ/γ_1 is, the more fractal the geometry of the set becomes, up to almost full phase space dimension D (Mehling et al. 2024; Lucarini and Bódai 2017).

2.1.8 The role of Melancholia states for metastable dynamics

As clarified in section 2.1.4, an edge state or Melancholia state is a saddle embedded in a basin boundary. Thus, what we know about chaotic saddles directly applies to chaotic Melancholia states[11]. Melancholia states thus play a central role in governing transient dynamics, both in parameter regimes before a crisis and – in the form of a remnant ghost state – beyond a crisis (Lai and Tél 2011).

Melancholia states are directly involved in the mechanisms of B- and N-tipping. In the case of B-tipping, the attractor that loses stability collides with a Melancholia state (e.g. a saddle point or unstable limit cycle) at the bifurcation point, such that the trajectory tips once it reaches the Melancholia state – provided that the forcing varies sufficiently slowly.

From large deviation theory, we also know that Melancholia states act as N-tipping gateways in the weak-noise limit (Freidlin and Wentzell 1998; Lucarini and Bódai 2019; Lucarini and Bódai 2020). For finite noise, the picture is more complicated, as we discuss in chapter 4 (Börner et al. 2024b). Nonetheless, noise-induced transitions tend to cross between competing basins of attraction over a region of relatively low quasipotential associated with a Melancholia state whose stable set locally forms the basin boundary (Feudel 2023).

For R-tipping, where the full nonautonomous dynamics must be considered, so-called R-tipping edge states may be introduced to extend the discussion beyond the autonomous case (Wieczorek et al. 2023). Assuming that the nonautonomous system limits to an autonomous future limit system, Melancholia states of the future limit system and their

[11] T. Tél, personal communication.

stable manifolds allow to define critical rates that separate tracking from tipping trajectories (Wieczorek et al. 2010; Wieczorek et al. 2023; Feudel 2023).

In the design of early warning signals, Melancholia states could help identify the relevant observables expected to most prominently and robustly display critical slowing-down (Lohmann et al. 2025). The idea is that fluctuations increase mainly in degrees of freedom directed towards the Melancholia state as a bifurcation is approached. In short, we argue that to understand metastability and tipping from a dynamical systems perspective, Melancholia states are essential.

2.1.9 Multiscale dynamics and homogenization

In the previous sections, we have introduced two viewpoints on metastable complex systems: the deterministic view with possibly chaotic dynamics as well as the stochastic view. Due to the mathematical differences, each approach comes with a distinct set of tools and way of thinking. Yet, the system we seek to describe – the Earth system – is of course the same in either case. Here we outline a way of reconciling the two viewpoints.

Climate variability occurs on a continuum of temporal scales spanning several orders of magnitude (see section 1.2.1). Ideally, we would like to resolve the dynamics on all scales, but this is not feasible and does not necessarily facilitate understanding (Held 2005). At the same time, climate variables are often expressed as time averages over several years, thus averaging out the high-frequency variability from diurnal to interannual timescales. How can we effectively reduce the number of degrees of freedom?

Following Hasselmann (1976), we may conceptually divide the spectrum of climate variability into a slow component (the "climate") and a fast component (the "weather"). While the slow dynamics are associated with inertial earth system components such as oceans and ice sheets, the chaotic dynamics of the atmosphere dominate the fast dynamics across days and seasons.

When building a reduced-order climate model, a basic way of incorporating unresolved processes is to parameterize their average effect on the resolved dynamics (Pavliotis and Stuart 2008). However, this neglects possible interactions across scales and implies that variability can only arise from time-dependent external forcing. The groundbreaking work by (Hasselmann 1976) instead proposes to model the unresolved fast component as stochastic forcing. The slow climate component then evolves under the integrated effect of fast random perturbations. Hasselmann's stochastic climate model presents a tool to effectively reduce the dimensionality of the system while still explaining key features of observed climate dynamics, such as the red variability spectrum (Arnold 2001).

Mathematically, a link between the stochastic and deterministic formulations of multiscale systems is given by homogenization theory (Gottwald et al. 2017; Kelly and Melbourne 2017). To understand the main idea, let us first introduce the framework of *fast-slow systems*, which offers an asymptotic theory of systems with multiple timescales (Kuehn 2015).

Fast-slow systems. We are concerned with dynamical systems where the time derivatives of some variables are much larger in magnitude than those of the others (Kuehn 2015).

In the simplest case, we have one *fast* variable x and one *slow* variable y,

$$\begin{aligned} \varepsilon \dot{x} &= f(x, y, \varepsilon) \\ \dot{y} &= g(x, y, \varepsilon) \,, \end{aligned} \tag{2.30}$$

where $\dot{u} = \mathrm{d}u/\mathrm{d}t$ and the timescale parameter $0 < \varepsilon \ll 1$ is a small parameter. Equation (2.30) is written in terms of the slow timescale t on which the slow dynamics take place. As $\varepsilon \to 0$, the timescale separation diverges and in the singular limit ($\varepsilon = 0$) the system reduces to the *slow subsystem*,

$$\begin{aligned} 0 &= f(x, y, 0) \\ \dot{y} &= g(x, y, 0) \,. \end{aligned} \tag{2.31}$$

Thus, near equilibria of the fast variables where $f = 0$, the system is largely governed by the slow dynamics $\dot{y} = g(x, y)$ which evolve on the *critical manifold* C_0,

$$C_0 = \{(x, y) \in \mathbb{R}^2 : f(x, y, 0) = 0\} \,. \tag{2.32}$$

Away from the critical manifold, we expect that the fast variable dominates and rapidly brings trajectories close to the critical manifold, as described by the *fast subsystem* ($\dot{x} = f(x, y, 0)$; $\dot{y} = 0$). This means that as the timescale separation becomes very large ($\varepsilon \to 0$), the system spends most of its time near the critical manifold. In the fast subsystem, the slow variable y takes the role of an external parameter; in the slow subsystem, the fast variable x is effectively enslaved to its equilibrium as a function of y.

Homogenization theory. Returning to the idea of a fast-slow climate model, let us now consider a fast-slow system of the following form (Kelly and Melbourne 2017; Ashwin et al. 2025),

$$\begin{aligned} \dot{\boldsymbol{y}} &= \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{\sqrt{\varepsilon}} \boldsymbol{b}_0(\boldsymbol{x}, \boldsymbol{y}) \\ \dot{\boldsymbol{x}} &= \frac{1}{\varepsilon} \boldsymbol{g}(\boldsymbol{x}) \,, \end{aligned} \tag{2.33}$$

where $\boldsymbol{y} \in \mathbb{R}^D$ describes the slow climate state we aim to model and $\boldsymbol{x} \in \mathbb{R}^d$ are a collection of fast, chaotic variables describing the weather. The function $\boldsymbol{b}_0$ represents the fluctuations of $\boldsymbol{y}$ due to the coupling to the fast variables, which are assumed to average to 0 over the fast dynamics generated by $\boldsymbol{g}$. Note that here the slow dynamics do not couple back to the fast dynamics.

Under fairly general conditions, Kelly and Melbourne (2017) have shown that, in the limit of infinite timescale separation between the fast and slow dynamics ($\varepsilon \to 0$), solutions of the slow dynamics of Eq. (2.33) converge weakly to solutions of an Itô stochastic differential equation,

$$\mathrm{d}\boldsymbol{y} = \tilde{\boldsymbol{b}}(\boldsymbol{y})\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{y})\mathrm{d}\boldsymbol{W}_t \,, \tag{2.34}$$

where $\boldsymbol{W}_t$ denotes a D-dimensional standard Wiener process (see also Gottwald et al. (2017), Pavliotis and Stuart (2008)). This reduction procedure is known as *homogenization*.

In reality, it appears reasonable to assume a two-way coupling, where the fast dynamics also depend on the slow dynamics such that $\boldsymbol{g} \equiv \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$. This more complicated case has been addressed by Engel et al. (2021).

In summary, homogenization theory provides a mathematical justification for stochastic climate modeling. Under the assumption of a large timescale separation between the fast and slow dynamics, Eqs. (2.33) and (2.34) establish a connection between stochastic and fast chaotic forcing. We consider stochastic forcing in chapters 3 and 4. However, it is clear that the limit of infinite timescale separation does not apply to climate variability. It may thus also be relevant to explicitly resolve the fast chaos, which we consider in chapter 5. An alternative approach to decouple multiscale systems devoid of a clear timescale separation has been proposed by Wouters and Lucarini (2012).

2.2 Numerical methods

Now that we have gone through the theory, how do we compute things in practice? This section reviews the numerical methods that we will apply to specific problems in the following chapters.

Software package. Alongside this project, we have developed `CriticalTransitions.jl`[12], a package written in the Julia language[13]. The code aims to provide a toolbox for sampling and analyzing critical transitions under stochastic and time-dependent forcing. Several of the algorithms discussed in the following are implemented and documented[14] there for general dynamical systems.

2.2.1 Finding Melancholia states: Edge tracking algorithm

In simple low-dimensional systems, equilibrium states of the (deterministic) system are found by solving $\boldsymbol{b}(\boldsymbol{x}) = 0$ (see Eq. (2.3)). As soon as the system becomes chaotic or high-dimensional, this approach typically becomes infeasible. Still, attractors can often be found by integrating the system in time: we start with some initial condition and see what the trajectory converges to after a sufficiently long time. To find multiple attractors in multistable systems, we can run an initial condition ensemble distributed over the competing basins of attraction. Of course, this is challenging in high-dimensional systems and limited by computational constraints but, in principle, works.

This recipe generally does not work for saddles because they are unstable. Unless we get very lucky and choose an initial condition from the stable set of the saddle, any trajectory will diverge from the saddle. However, there is a way to still compute saddles using only forward-time integration of the system. The *edge tracking algorithm* is such a method that allows finding saddles with a single unstable direction, which applies to edge states (or Melancholia states).[15]

[12] Link to the package: `https://github.com/JuliaDynamics/CriticalTransitions.jl`. Developed in collaboration with Orjan Ameye, George Datseris, Ryan Deeley, Calvin Nesbitt, Raphael Römer and Jeroen Wouters.

[13] Homepage of the Julia language: `https://julialang.org/`, visited on 10 January 2025.

[14] Package documentation: `https://juliadynamics.github.io/CriticalTransitions.jl/`.

[15] Parts of the text in this subsection are based on Börner et al. (2025a).

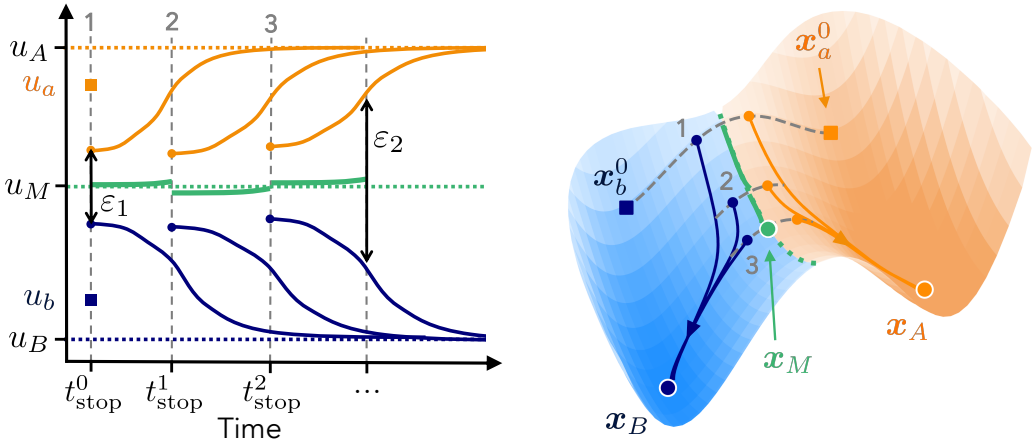

Figure 2.2 | **Edge tracking algorithm** illustrated as the resulting timeseries projected onto the coordinate u (left) and in the corresponding quasipotential landscape of a bistable system (right). The landscape shows the basins of attraction of the attractors $\boldsymbol{x}_A$ (orange shading) and $\boldsymbol{x}_B$ (blue shading), separated by the basin boundary (green dashed). Starting from $\boldsymbol{x}^0_a$ and $\boldsymbol{x}^0_b$, iterations 1-3 yield a pseudotrajectory (green solid line) that leads close to the Melancholia state $\boldsymbol{x}_M$. Gray dashed lines indicate the bisections. Figure adapted from Börner et al. (2025a), with modifications.

The edge tracking algorithm goes back to the method proposed by Battelino et al. (1988) and Skufca et al. (2006). It is also known as the "saddle-straddle method" and has originally been applied to identify the edge between laminar and turbulent flow in planar shear flows (Skufca et al. 2006; Schneider et al. 2008). It was introduced to climate science through explorations of the "Snowball Earth" transition (Lucarini and Bódai 2017; Lucarini and Bódai 2019; Lucarini and Bódai 2020; Lucarini et al. 2022), and recently applied to the AMOC in conceptual models (Mehling et al. 2024; Axelsen et al. 2024) as well as a global ocean model (Lohmann and Lucarini 2024).

For simplicity, let us consider a bistable deterministic system with attractors A and B located at $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$, respectively. Our goal is to find the Melancholia state $\boldsymbol{x}_M$ located on the basin boundary that separates the basins of attraction $\mathcal{B}(A)$ and $\mathcal{B}(B)$ (Fig. 2.2). The general idea is to counteract the global instability of the saddle by running two simulations in parallel and repeatedly "pulling" them together.

Starting from two initial conditions $\boldsymbol{x}^{(0)}_a \in \mathcal{B}(A)$ and $\boldsymbol{x}^{(0)}_b \in \mathcal{B}(B)$, the algorithm consists of an iterative loop with two steps. For the i-th iteration ($i \geq 1$):

1. *Bisection.* Between the two initial conditions $\boldsymbol{x}^{(i-1)}_a$ and $\boldsymbol{x}^{(i-1)}_b$, bisect repeatedly along a straight line in phase space to obtain two new initial conditions $\boldsymbol{y}^{(i)}_a$ and $\boldsymbol{y}^{(i)}_b$ that are less than a distance $\varepsilon_1 > ||\boldsymbol{y}^{(i)}_a - \boldsymbol{y}^{(i)}_b||$ apart while still converging to different attractors (one to A and the other to B).

2. *Tracking.* From each of the two new initial conditions, run a simulation, $\boldsymbol{y}^{(i)}_a(t)$ and $\boldsymbol{y}^{(i)}_b(t)$. Stop the simulations at $t^{(i)}_{\text{stop}}$ when the two trajectories diverge by more than a distance ε_2 from each other:

$$t^{(i)}_{\text{stop}} = \inf_t \left\| \boldsymbol{y}^{(i)}_a(t) - \boldsymbol{y}^{(i)}_b(t) \right\| \geq \varepsilon_2 \,.$$

   Use the end points of these simulations as initial conditions for the next iteration: $\boldsymbol{x}^{(i)}_a = \boldsymbol{y}^{(i)}_a(t^{(i)}_{\text{stop}})$ and $\boldsymbol{x}^{(i)}_b = \boldsymbol{y}^{(i)}_b(t^{(i)}_{\text{stop}})$. Repeat 1.

Here the norm $||\cdot||$ could be the Euclidean distance in the (normalized) phase space or any other appropriate metric that reflects the separation between the two states. Often the difference in a one-dimensional observable u is chosen, which is fine as long as the bisections are performed in the full phase space. The bisection step can be formulated in a pseudo-code as follows:

Bisection step

Let $\boldsymbol{a} = \boldsymbol{x}_a^{(i-1)}$ and $\boldsymbol{b} = \boldsymbol{x}_b^{(i-1)}$

`while` $||\boldsymbol{b} - \boldsymbol{a}|| \geq \varepsilon_1$:

    Set $\boldsymbol{c} = \frac{1}{2}(\boldsymbol{a} + \boldsymbol{b})$
    `if` $\boldsymbol{c} \in \mathcal{B}(A)$:
        set $\boldsymbol{a} = \boldsymbol{c}$
    `else`:
        set $\boldsymbol{b} = \boldsymbol{c}$

Set $\boldsymbol{y}_a^{(i)} = \boldsymbol{a}$ and $\boldsymbol{y}_b^{(i)} = \boldsymbol{b}$.

This pseudo-code neglects the possibility that $\boldsymbol{c}$ lands precisely on the basin boundary or in a possible third basin of attraction. In such cases, one workaround is to slightly perturb the candidate $\boldsymbol{c}$ along the direction $\boldsymbol{b} - \boldsymbol{a}$ until $\boldsymbol{c} \in \mathcal{B}(A) \cup \mathcal{B}(B)$.

Running the algorithm for N iterations yields two strings of trajectory segments that trace the basin boundary on either side of it. By concatenating each of these strings in time and taking their average at each time point, we obtain a pseudotrajectory $\boldsymbol{x}_{\text{edge}}(t)$ that approximates a trajectory on the basin boundary:

$$\boldsymbol{x}_{\text{edge}}(t) = \left\{ \frac{1}{2}\left(\boldsymbol{y}_a^{(i)}(t) + \boldsymbol{y}_b^{(i)}(t)\right) : i \in \{1, \ldots, N\} \text{ and } t \in \left[t_{\text{stop}}^{(i-1)},\ t_{\text{stop}}^{(i)}\right) \right\} .$$

The repeated rebisection of initial conditions counteracts the instability that causes any trajectory initialized near the boundary to eventually diverge from it. Based on the property of Melancholia states as attracting sets when restricted to the boundary, the pseudotrajectory is expected to converge to a Melancholia state. If the Melancholia state is simply a saddle point, the algorithm will converge to that point. In the case of an unstable limit cycle or chaotic set, the algorithm converges to this invariant set and successively populates its invariant measure. Thus, after removing the initial transient period t_{trans}, the pseudotrajectory $\boldsymbol{x}_{\text{edge}}(t)$ for $t \geq t_{\text{trans}}$ is an approximate construction of the Melancholia state.

The precise course of the pseudotrajectory $\boldsymbol{x}_{\text{edge}}(t)$ depends on the choice of the starting initial conditions $\boldsymbol{x}_a^{(0)}$ and $\boldsymbol{x}_b^{(0)}$ as well as the hyperparameters ε_1 and ε_2, which affect both the accuracy and the speed of the method. In principle, decreasing ε_1 and ε_2 improves the accuracy of the result but can also lead to slow convergence. Setting the ratio $\varepsilon_2/\varepsilon_1 \geq 1$ requires a trade-off: a small ratio leads to frequent rebisections, giving a fragmented pseudotrajectory of many short segments that can distort the dynamics,

especially in the high-frequency variability[16]. A large ratio can reduce the accuracy and slow down the algorithm because more bisections are needed per rebisection step. In the cases studied here, choosing the ratio $\varepsilon_2/\varepsilon_1$ between 2 to 10 proved suitable.

It is not obvious that the edge tracking algorithm can handle systems with a more complicated phase space structure than the smooth double-well potential in Fig. 2.2. However, the algorithm has worked robustly even in high-dimensional chaotic systems with a fractal basin boundary. In that case, there may be many filaments of the boundary intersecting with a straight line between points $\boldsymbol{a}$ and $\boldsymbol{b}$. By only updating either $\boldsymbol{a}$ or $\boldsymbol{b}$ with $\boldsymbol{c}$ during any bisection step (where $\boldsymbol{c}$ belongs to the same basin of attraction as the updated point), the algorithm guarantees to always keep two initial conditions on either side of the boundary. Despite the possibly complex geometry of the basin boundary, the iterative tracking steps tend to follow, over time, a coarse-grained flow along the boundary leading to a Melancholia state. Thus, even though repeating the algorithm with slightly different settings can lead to rather different pseudotrajectories in chaotic systems, each realization will eventually approximate the dynamics of the same invariant set (provided they lead to the same Melancholia state).

Other techniques to compute saddles include the PIM-Triple method (Nusse and Yorke 1989) and its generalization by Moresco and Ponce Dawson (1999), as well as control-based continuation (Sieber et al. 2014; Panagiotopoulos et al. 2023). They are not explored in this thesis.

Software implementation. An efficient and tested code for the edge tracking algorithm is publicly available in the Julia package `Attractors.jl` (since version 1.15), a subpackage of `DynamicalSystems.jl`. For more information, see the documentation[17] and code example[18].

2.2.2 Sampling noise-induced transitions

When subjecting a multistable dynamical system to noise, critical transitions between the competing attractors become possible (see section 2.1.5). However, if the noise is weak, transitions will be rare, and long simulations will be required to generate noise-induced transitions.

Monte Carlo simulation. One option is to sample transitions by brute force, using direct simulation. The simplest numerical scheme for solving an Itô SDE of the form in Eq. (2.13) with Gaussian noise is the *Euler-Maruyama* scheme (Kloeden and Platen 1992). Discretizing time in steps of duration Δt, the solution $\boldsymbol{x}_i$ at the i-th step is given in terms of the solution at the previous step plus a random perturbation,

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \boldsymbol{b}(\boldsymbol{x}_{i-1})\Delta t + \sigma\sqrt{\Delta t}\boldsymbol{\Sigma}(\boldsymbol{x}_{i-1}) \cdot \boldsymbol{\xi}_{i-1}\,, \quad i \geq 1\,. \tag{2.35}$$

Here $\boldsymbol{x}_0$ is the initial condition and $\boldsymbol{\xi}_i$ is a vector of increments of independent Wiener processes.

Many improved numerical schemes with higher accuracy exist, but the Euler-Maruyama scheme remains commonly used due to its simplicity and speed. For our purposes, we

[16] Jan Sieber, personal communication.
[17] https://juliadynamics.github.io/Attractors.jl/stable/api/#Attractors.edgetracking
[18] https://juliadynamics.github.io/Attractors.jl/stable/examples/#Edge-tracking

deem the method appropriate as long as we choose a suitably small time step Δt to contain the discretization error. Note that the stochastic term scales with the square root of Δt for Gaussian noise. In chapter 3, we will also consider α-stable Lévy noise, in which case $\boldsymbol{\xi}_i$ denotes increments of the Lévy process and the stochastic term scales like $(\Delta t)^{1/\alpha}$ instead of $(\Delta t)^{1/2}$.

Simulating Lévy noise

Chambers et al. (1976) developed a method to generate α-stable random numbers, which for symmetric distributions (skewness parameter $\beta = 0$) simplifies to the following steps:

1. Draw a uniformly distributed random number $U \in (-\frac{\pi}{2}, \frac{\pi}{2})$
2. Draw another independent uniformly distributed random number $V \in (0, 1)$ and let $W = -\ln(V)$
3. Compute $r_\alpha = \frac{\sin(\alpha U)}{(\cos(U))^{1/\alpha}} \cdot \left(\frac{\cos((1-\alpha)U)}{W}\right)^{\frac{1-\alpha}{\alpha}}$.

Samples r_α can be used as stationary, independent noise increments to construct Lévy processes in Monte Carlo simulations.

Suppose we wish to sample a transition from $\boldsymbol{x}_A$ to $\boldsymbol{x}_B$. If we solve Eq. (2.35) for a sufficiently long time, the random trajectory will eventually reach a neighborhood B around $\boldsymbol{x}_B$. Monte Carlo *rejection sampling* consists of launching an ensemble of trajectories at $\boldsymbol{x}_0 = \boldsymbol{x}_a$, running each simulation until it either reaches B or a time limit $t_{\max}$, and collecting all trajectories that made it to B. This technique is implemented in CriticalTransitions.jl as the `transitions` function[19] with multi-threading support.

The efficiency of sampling transitions via Monte Carlo simulation can be greatly improved by employing *rare event algorithms* based on *importance sampling* (Bouchet, Rolland, and Wouters 2019; Grafke and Vanden-Eijnden 2019). These methods are increasingly and successfully applied in climate science (Gálfi et al. 2021; Ragone et al. 2018; Cini et al. 2024; Jacques-Dumas et al. 2024). In this thesis, we explore yet a different sampling technique called *pathspace sampling*, which is based on large deviation theory and has much potential especially for low-dimensional systems.

Pathspace transition path sampling. A clear disadvantage of Monte Carlo sampling in phase space is that we can only prescribe the starting point $\boldsymbol{x}_A$ and must wait for simulations to reach the vicinity of the desired end point $\boldsymbol{x}_B$. In pathspace (definition 2.14), we can specify both start and end points and sample random paths as *stochastic bridges* between them (Aguilar et al. 2022; Grafke and Laio 2024).

The general idea is the following (Stuart et al. 2004): instead of solving the SDE (2.13) in physical time t, we solve a stochastic partial differential equation (SPDE) that describes how a path from $\boldsymbol{x}_A$ to $\boldsymbol{x}_B$ (parameterized in physical time) evolves in *virtual time* z under

[19] Documentation: `https://juliadynamics.github.io/CriticalTransitions.jl/stable/man/sampling/`.

different realizations of the noise. Hairer et al. (2007) provide the mathematical basis for this. They showed that for systems of the form of Eq. (2.13) with gradient drift and identity covariance matrix, the distribution of sample paths conditioned on $\boldsymbol{x}(0) = \boldsymbol{x}_A$ and $\boldsymbol{x}(T) = \boldsymbol{x}_B$ with a fixed travel time T is the same as the distribution of the solution $\boldsymbol{x}(t, z)$ to the boundary value problem defined by the SPDE,

$$\frac{\partial \boldsymbol{x}}{\partial z} = \frac{\delta}{\delta \boldsymbol{x}} \ln(q_{\text{path}}(\boldsymbol{x})) + \sqrt{2}\boldsymbol{\eta}(t, z)\,, \quad \boldsymbol{x}(t, 0) = \boldsymbol{x}_0(t)\,,\ t \in (0, T)\,,\ z \geq 0\,, \tag{2.36}$$

with boundary conditions $\boldsymbol{x}(0, z) = \boldsymbol{x}_A$ and $\boldsymbol{x}(T, z) = \boldsymbol{x}_B$ for all $z \geq 0$. Here $\boldsymbol{\eta}(t, z)$ denotes D-dimensional spatio-temporal white noise satisfying

$$\langle \eta_i(t, z)\eta_j(t', z')\rangle = \delta(t - t')\delta(z - z')\delta_{ij}\,, \quad \langle \eta_i(t, z)\rangle = 0 \quad \text{for } i, j \in \{1, \ldots, D\}\,,$$

and the path potential $q_{\text{path}}(\boldsymbol{x})$ is given by

$$q_{\text{path}}(\boldsymbol{x}) = \exp\left\{-\frac{1}{2}\int_0^T \left(\left\|\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} - \boldsymbol{b}(\boldsymbol{x})\right\|^2_{(\sigma\boldsymbol{\Sigma})} + \boldsymbol{\nabla}\cdot\boldsymbol{b}(\boldsymbol{x})\right) \mathrm{d}t\right\}\,, \tag{2.37}$$

where the norm $||\cdot||_{(\sigma\boldsymbol{\Sigma})}$ is defined via the scalar product,

$$\begin{aligned}\langle a, b\rangle_{(\sigma\boldsymbol{\Sigma})} &= \boldsymbol{a}^\top \left(|\sigma|^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\right)^{-1}\boldsymbol{b}\,,\\ \|a\|^2_{(\sigma\boldsymbol{\Sigma})} &= \langle a, a\rangle_{(\sigma\boldsymbol{\Sigma})}\,.\end{aligned}$$

The path potential (Eq. (2.37)) is the negative exponential of the *Onsager-Machlup action* (Onsager and Machlup 1953; Horsthemke and Bach 1975; Börner et al. 2024b).

Term 2.27. **Onsager-Machlup action**

The *Onsager-Machlup (OM) action* functional $\tilde{S}^\sigma_T : \Pi \to \mathbb{R}$ is given by

$$\tilde{S}^\sigma_T[\boldsymbol{\varphi}_t] = \frac{1}{2}\int_0^T \left(\|\dot{\boldsymbol{\varphi}}_t - \boldsymbol{b}(\boldsymbol{\varphi}_t)\|^2_{(\sigma\boldsymbol{\Sigma})} + \boldsymbol{\nabla}\cdot\boldsymbol{b}(\boldsymbol{\varphi}_t)\right) \mathrm{d}t\,, \tag{2.38}$$

where the $(\sigma\boldsymbol{\Sigma})$-norm is defined below Eq. (2.37).

The OM action is related to the FW action S_T (Eq. (2.18)) via

$$\tilde{S}^\sigma_T[\boldsymbol{\varphi}_t] = S_T[\boldsymbol{\varphi}_t] + \frac{1}{2}\int_0^T \boldsymbol{\nabla}\cdot\boldsymbol{b}(\boldsymbol{\varphi}_t)\,\mathrm{d}t\,. \tag{2.39}$$

Unlike the FW action, the OM action depends on the noise strength σ. It has been a debated question in the literature whether the FW action or the OM action is the correct action functional in the variational problem of finding the most probable path (Dürr and Bach 1978; Pinski and Stuart 2010; Bröcker 2019; Du et al. 2021; Gladrow et al. 2021). The ambiguity stems from the fact that stochastic paths are not differentiable, requiring a choice when performing stochastic calculus (Itô vs. Stratonovich; van Kampen (1981)). When deriving a discretized form of the action functional, the additional divergence term

in the OM action can either arise or not depending on the choice of discretization and the order of limits taken (Adib 2008; Cugliandolo and Lecomte 2017). Backed by experimental evidence, it is increasingly appreciated that the OM action is the appropriate functional for finite noise (Bröcker 2019; Gladrow et al. 2021; Fleurantin et al. 2023; Börner et al. 2024b). A physical motivation for the divergence term is often lacking in the literature. We will further discuss the relevance and justification of the OM action in chapter 4.

Calculating the functional derivative $\delta/\delta\boldsymbol{x}$ in Eq. (2.36) and performing a rescaling of the virtual time $z \to \sigma^{-2}z$ leads to the explicit form

$$\frac{\partial \boldsymbol{x}}{\partial z} = \boldsymbol{Q}^{-1}\partial_{tt}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\nabla}\langle \boldsymbol{b}, \boldsymbol{Q}^{-1}\boldsymbol{b}\rangle - \Big(\boldsymbol{Q}^{-1}(\boldsymbol{\nabla}\boldsymbol{b}) - (\boldsymbol{\nabla}\boldsymbol{b})^{\top}\boldsymbol{Q}^{-1}\Big)\partial_t\boldsymbol{x} - \frac{\sigma^2}{2}\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{b}) + \sqrt{2}\sigma\boldsymbol{\eta}(t,z)\,, \tag{2.40}$$

where we have introduced shorthand notations for partial derivatives in physical time ($\partial_t = \partial/\partial t$ and $\partial_{tt} = \partial^2/\partial t^2$), $\boldsymbol{\nabla}\boldsymbol{b}$ for the Jacobian of $\boldsymbol{b}$, and $\boldsymbol{\nabla}\langle \boldsymbol{b}, \boldsymbol{Q}^{-1}\boldsymbol{b}\rangle$ for the gradient of the scalar product. Here $\boldsymbol{Q} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}$ is the noise covariance matrix as in Eq. (2.13), which is symmetric and invertible by assumption.

Intuitively, solving the SPDE may be interpreted as stochastically perturbed gradient descent in pathspace (Pinski and Stuart 2010). Thereby the gradient term is given by the functional derivative of the OM action, $-\delta/\delta\boldsymbol{\varphi}_t(\tilde{S}_T^{\sigma}[\boldsymbol{\varphi}_t])$. We thus expect the solution of Eq. (2.36) to converge to a local minimum of this gradient and, after a transient convergence period, reach a statistical steady state. Once the statistical steady state is reached, any solution $\boldsymbol{x}(t,z)$ at given virtual time z corresponds to a sample transition path from $\boldsymbol{x}_A$ to $\boldsymbol{x}_B$ in physical time T. The theorem by Hairer et al. (2007) states that this steady-state distribution in pathspace is equivalent to the density distribution of the solution of the SDE (2.13) conditioned on the start and end points. While this result was shown for gradient systems only, Hairer et al. (2007) conjecture still holds for non-gradient systems where $\boldsymbol{b} \neq -\boldsymbol{\nabla}V$. We will test this conjecture in chapter 3.

Note that even if the drift $\boldsymbol{b}$ of the SDE is non-gradient, the corresponding SPDE describes gradient dynamics. Solving the SPDE efficiently generates sample transition paths after a convergence period. Since the SPDE takes the form of a Langevin equation in pathspace, the method is also called *Langevin MCMC* (Stuart et al. 2004; Pinski and Stuart 2010).

To summarize, in this thesis we will sample noise-induced transition paths either by Monte Carlo rejection sampling or pathspace transition path sampling. As we propose in section 4.2.4, pathspace sampling can also be turned into a method to compute most probable transition paths in finite-noise and finite-time contexts.

2.2.3 Predicting transition paths

As discussed in section 2.1.6, FW theory can be used to compute the most probable transition path in the weak-noise limit – the instanton, minimizer of the FW action (definition 2.16). Different numerical algorithms exist to solve the action minimization problem (Grafke and Vanden-Eijnden 2019; E et al. 2004; Heymann and Vanden-Eijnden 2008b; Grafke et al. 2014; Simonnet 2023).

Equation (2.19) reveals that we are dealing with a double-minimization problem: the instanton is found over all paths $\boldsymbol{\varphi}_t$ with travel time T, over all $T > 0$. The true minimizer is generally attained at $T = \infty$, which is of course impractical. However, if we compute

the FW action minimizer for a given T, we often see a convergence of the minimum action path for $T > T_{\text{cutoff}}$, so that the true instanton is approximated by the action minimizer at T_{cutoff}. Performing the minimization for fixed time is referred to as the *minimum action method* (MAM) (E et al. 2004).

Alternatively, there is a clever way to circumvent the minimization over all $T > 0$, known as the *geometric minimum action method* (gMAM) (Freidlin and Wentzell 1998; Heymann and Vanden-Eijnden 2008a; Heymann and Vanden-Eijnden 2008b). The method exploits a correspondence between the minimization of the FW action (Eq. (2.18)) and the minimization of a different action functional, the *geometric action*.

Term 2.28. **Geometric action**

The *geometric action* $\hat{S} : C([0, L], \mathbb{R}^D) \to \mathbb{R}_{\geq 0}$ along a curve $\boldsymbol{\psi}_s$ in the phase space of Eq. (2.13), parameterized by $s \in [0, L]$, is given by

$$\hat{S}[\boldsymbol{\psi}_s] = \int_0^L \left(||\boldsymbol{\psi}_s'||_{\boldsymbol{Q}(\boldsymbol{\psi}_s)} ||\boldsymbol{b}(\boldsymbol{\psi}_s)||_{\boldsymbol{Q}(\boldsymbol{\psi}_s)} - \langle \boldsymbol{\psi}_s', \boldsymbol{b}(\boldsymbol{\psi}_s) \rangle_{\boldsymbol{Q}(\boldsymbol{\psi}_s)} \right) \mathrm{d}s \,. \tag{2.41}$$

Here L is the arc length of $\boldsymbol{\psi}_s$, the prime denotes the derivative with respect to s, and the norm and dot product are defined via $\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\boldsymbol{Q}} := \boldsymbol{a}^\top \boldsymbol{Q}^{-1} \boldsymbol{b}$.

Suppose a path $\boldsymbol{\varphi}_t$, parameterized in time, minimizes the FW action. By rescaling time locally along the path such that $||\dot{\boldsymbol{\varphi}}_t||_{\boldsymbol{Q}(\boldsymbol{\varphi}_t)} = ||\boldsymbol{b}(\boldsymbol{\varphi}_t)||_{\boldsymbol{Q}(\boldsymbol{\varphi}_t)}$, we can obtain a geometric representation $\boldsymbol{\psi}_s$ of the minimum action path, then re-parameterized in arc length s along the path, that minimizes the geometric action:

$$\min_{\boldsymbol{\varphi}_t, T} S_T[\boldsymbol{\varphi}_t] = \min_{\boldsymbol{\psi}_s} \hat{S}[\boldsymbol{\psi}_s] \,. \tag{2.42}$$

For a proof and further explanation, see Heymann and Vanden-Eijnden (2008a) and Cameron (2012). The double minimization problem thus reduces to a single minimization problem on the space of curves.

Recently, a deep learning version of gMAM has been introduced by Simonnet (2023). Further methods to compute instantons include the *augmented Lagrangian method*, which also works for degenerate noise covariance matrices (Schorlepp et al. 2022; Soons et al. 2024; Bernuzzi and Grafke 2024). If the FW quasipotential V_A with respect to $\boldsymbol{x}_A$ is known for the system, the instanton from $\boldsymbol{x}_A$ to a point $\boldsymbol{x}$ follows immediately. Numerically, it can be computed from the quasipotential using a *shooting method* by solving the (normalized) instanton equation (2.22) backwards from $\boldsymbol{x}$ to $\boldsymbol{x}_A$ (Dahiya and Cameron 2018a; Cameron 2012).

Software implementation. The MAM and gMAM algorithms are implemented in the `CriticalTransitions.jl` package[20]. The path is discretized into N points in the D-dimensional phase space, resulting in an $N \times D$-dimensional optimization problem. The

[20] Documentation: https://juliadynamics.github.io/CriticalTransitions.jl/stable/man/largedeviations/

convergence rate can depend strongly on the gradient of the action in pathspace. In the examples we studied (see chapter 3), we found gMAM to be more robust than MAM.

The definition of the SDE (2.13) leaves an ambiguity in the combination of the noise strength value σ and covariance matrix $\boldsymbol{Q}$, since any scalar prefactor in $\boldsymbol{Q}$ can be absorbed in σ^2. Since the FW action is independent of σ, multiplying $\boldsymbol{Q}$ by a scalar also scales the FW action value accordingly. To ensure consistency in the action value, we propose to normalize the covariance matrix when defining the optimization problem by dividing it by the factor $c = L_1(\boldsymbol{Q})/D$, i.e. the quotient of the L_1 matrix norm and the dimensionality D of the system. This solution is implemented in CriticalTransitions.jl (as of version 0.4).

2.2.4 Computing quasipotentials

Knowing the quasipotential of a dynamical system is highly desirable, as it offers a quantitative view of the global stability of the system. In the case of the FW quasipotential, we obtain the stability landscape with respect to a reference point and can readily compute most probable transition paths and transition rates. However, solving the Hamilton-Jacobi equation (see Eqs. (2.12) and (2.16)) is difficult in practice, and in general a solution is not even guaranteed to exist.

The *ordered upwind method* (Sethian and Vladimirsky 2003) presents a technique based on the dynamic programming principle to solve Hamilton-Jacobi equations numerically. It has been adapted by Cameron (2012) to calculate the quasipotential in two-dimensional systems. A more accurate and faster way to compute the quasipotential on a mesh is the *ordered line integral method* (OLIM) introduced by Dahiya and Cameron (2018b), which has been generalized to anisotropic and state-dependent noise (Dahiya and Cameron 2018a) and extended to $D = 3$ dimensions (Yang et al. 2019).

Ordered Line Integral Method. Instead of solving the Hamilton-Jacobi equation, the OLIM algorithms[21] are based on the minimization of the geometric action (Eq. (2.41)). Starting from the reference point $\boldsymbol{x}_A$ (taken to be an attractor), tentative quasipotential values are computed at neighboring points within a given update radius Kh, where h is the grid step size and $K \in \mathbb{N}_{>0}$ a hyperparameter. At each step, the tentative values are computed by locally minimizing the action along straight lines of lengths up to the update radius, whereby the action integral is solved numerically using one of several possible quadrature rules (Dahiya and Cameron 2018b). Grid points are then updated based on the status and tentative value of their neighbors. Step by step, final values of the quasipotential are determined across an expanding region around $\boldsymbol{x}_A$.

Beyond 3D. Extending the OLIM procedure to higher dimensions is hindered by the *curse of dimensionality*: the computational cost grows exponentially with D. To our knowledge, there are currently no quasipotential solvers for more than three dimensions. If we are interested in the quasipotential within a contained region of the phase space, we can select points in that region and compute instantons to these points using the methods outlined in the previous section. The action value along each of these instantons then corresponds to the quasipotential value at the end point.

[21] Software implementations of various OLIM algorithms are available in the C language on M. Cameron's homepage: `https://www.math.umd.edu/~mariakc/OLIM.html`, visited on 20 January 2025.

If simulation data of the system are available that cover the phase space region of interest, machine learning methods can offer a promising alternative also in higher dimensions. Inspired by the normal decomposition of the drift (Eq. (2.21)), Lin et al. (2021) proposed training two neural networks in parallel, one corresponding to the quasipotential and the other to the nongradient rest term, under the constraint of orthogonality imposed as a cost function. Lastly, the quasipotential in definition 2.13 (Graham and Tél 1986) may be constructed by empirically estimating the invariant measure $\boldsymbol{\rho}_\sigma$, as done e.g. by Margazoglou et al. (2021) and Rousseau et al. (2023) for climate models. This likewise requires data from simulations or observations covering the relevant phase space region. If the phase space is sampled by running a long stochastic simulation, it is important to balance the choice of the noise strength σ to be sufficiently large to explore the phase space yet as close as possible to the weak-noise limit (Zhou and Li 2016).

Chapter 3

Stability landscape and tipping in an AMOC box model

Three coupled bath tubs

3.1 Rooth's interhemispheric box model . . . 62
3.1.1 Model description . . . 62
3.1.2 Modifications to the original model . . . 64
3.2 Stability and hysteresis under freshwater forcing . . . 64
3.2.1 Equilibria under symmetric forcing . . . 64
3.2.2 Melancholia state . . . 65
3.2.3 Resilience against freshwater bursts . . . 67
3.2.4 Bifurcation diagram for freshwater forcing . . . 69
3.3 Noise-induced transitions . . . 71
3.3.1 Gaussian noise . . . 72
3.3.2 Lévy noise . . . 74
3.4 Discussion & conclusion . . . 78

This chapter presents unpublished work that has partly been presented by Börner et al. (2022).

Introduction

Let us begin at the bottom of the climate model hierarchy. The famous ocean box model of Stommel (1961) and its one-dimensional stochastic formulation (Cessi 1994) probably present the simplest conceptual models of a bistable thermohaline circulation. These models are used to describe an Atlantic overturning circulation in the northern hemisphere, based on temperature and salinity differences between a subtropical box and a subpolar box.

A key characteristic of the AMOC is that it transports heat northward on *both* hemispheres. The Atlantic Ocean is therefore the only ocean basin with a net northward oceanic heat transport across the equator. Moreover, it has been proposed that processes in the Southern Ocean such as freshwater fluxes and Antarctic Bottom Water (AABW) formation play an important role for the AMOC and its stability (Kuhlbrodt et al. 2007; Swingedouw et al. 2009; Buizert and Schmittner 2015; Cimatoribus et al. 2012). This motivates incorporating the southern hemisphere in an AMOC box model to mimic a pole-to-pole overturning circulation with interhemispheric heat exchange.

Rooth (1982) proposed an interhemispheric model consisting of three boxes representing the northern, tropical, and southern Atlantic, respectively. Like the Stommel model, it is purely deterministic and features a bistable circulation. A key difference is that the Rooth model parameterizes the overturning strength in terms of the density difference between the high-latitude boxes, not the equator-to-pole density difference. This choice is supported by ocean general circulation models (Rahmstorf 1996; Klinger and Marotzke 1999; Hughes and Weaver 1994) and leads to distinct stability properties (Scott et al. 1999; Lucarini and Stone 2005), as we will see. Scott et al. (1999) argue that the interhemispheric description offers a better analogue of the Atlantic thermohaline circulation compared to hemispheric models.

Here we investigate the stability landscape and transition behavior of the Rooth model, exemplifying how the dynamical systems approach outlined in chapter 2.2.4 can be applied to the climate system. First, we focus on the stability properties of the AMOC as a function of the freshwater flux in the North Atlantic. We then add stochastic forcing to examine noise-induced transitions, both for Gaussian noise and Lévy noise. Rather than attempting to provide a full account of the dynamics of the Rooth model, we will highlight a few qualitative features that motivate and guide the analysis in the subsequent chapters.

Mathematically, the Rooth model is a non-smooth, nonlinear and multiscale five-dimensional dynamical system. It is computationally cheap and amenable to limited analytical treatment, yet already too high-dimensional to visualize the full phase space. In five dimensions, the curse of dimensionality becomes noticeable, and we cannot compute the full quasipotential with the mesh-based algorithms introduced in section 2.2.4 anymore. Nonetheless, by extending our stability analysis beyond stable equilibria, exploring the phase space by adding noise, and selecting suitable low-dimensional projections of the phase space, we will gain an understanding of the model's global stability landscape.

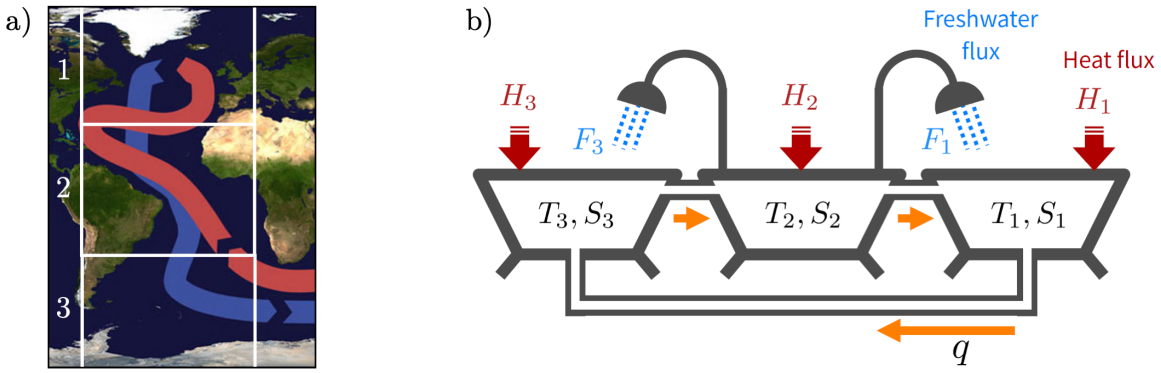

Figure 3.1 | **Rooth's interhemispheric box model.** (a) Map of the Atlantic sector showing the approximate geographical region represented by each box, superimposed on a sketch of the large-scale AMOC currents. Adapted from Brisbane (Wikimedia Commons) under CC BY-SA 3.0 license. (b) Schematic of the model setup. Orange arrows indicate the flow direction for $q > 0$.

3.1 Rooth's interhemispheric box model

3.1.1 Model description

The Rooth model (Rooth 1982) partitions the Atlantic Ocean into three interconnected boxes (Fig. 3.1):

(1) The northern box, representing the North Atlantic north of 30°N,

(2) The tropical box, representing the tropical Atlantic (30°S to 30°N),

(3) The southern box, representing the Atlantic sector south of 30°S.

Each box is 5 km deep and the high-latitude boxes have the same volume ($V_0 = 1.08 \times 10^{17}\,\mathrm{m}^3$), which is half the volume of the tropical box (Scott et al. 1999). It is assumed that the water in each box is well-mixed. The high-latitude boxes are connected to the tropical box via a passage of negligible volume, allowing for a flow between them mimicking surface currents. Additionally, the high-latitude boxes are connected to each other, permitting "deep" currents that bypass the tropical box (Fig. 3.1).

The state of each box ($i \in \{1, 2, 3\}$) is determined by its mean temperature T_i (in units °C) and salinity S_i (in dimensionless units $\mathrm{g\,kg}^{-1}$). Assuming a linear equation of state, the density of each box is given by

$$\rho_i = \rho_0(1 - \alpha T_i + \beta S_i) \tag{3.1}$$

where α and β are the thermal and haline expansion coefficients, respectively (Table 3.1), and $\rho_0 = 1000\,\mathrm{kg\,m}^{-3}$ is the reference density. The flow strength q is assumed proportional to the density difference between box 1 and box 3,

$$q = \kappa\big(\alpha(T_3 - T_1) + \beta(S_1 - S_3)\big)\ , \tag{3.2}$$

where k is the hydraulic constant, which is tuned to yield a reasonable equilibrium AMOC overturning strength (Scott et al. 1999). The total salinity is conserved, such that we

may express S_2 as a function of S_1 and S_3. This leaves a state vector of five independent variables,

$$\boldsymbol{x}(t) = \big(T_1(t),\ T_2(t),\ T_3(t),\ S_1(t),\ S_3(t)\big)\,.$$

Their dynamics is governed by a system of ordinary differential equations (Rooth 1982; Scott et al. 1999; Lucarini and Stone 2005),

$$\dot{T}_1 = \begin{cases} q(T_2 - T_1) + H_1 & q \geq 0 \\ q(T_1 - T_3) + H_1 & q < 0 \end{cases} \tag{3.3a}$$

$$\dot{T}_2 = \begin{cases} \frac{q}{V}(T_3 - T_2) + H_2 & q \geq 0 \\ \frac{q}{V}(T_2 - T_1) + H_2 & q < 0 \end{cases} \tag{3.3b}$$

$$\dot{T}_3 = \begin{cases} q(T_1 - T_3) + H_3 & q \geq 0 \\ q(T_3 - T_2) + H_3 & q < 0 \end{cases} \tag{3.3c}$$

$$\dot{S}_1 = \begin{cases} q(S_2 - S_1) - F_1 & q \geq 0 \\ q(S_1 - S_3) - F_1 & q < 0 \end{cases} \tag{3.3d}$$

$$\dot{S}_3 = \begin{cases} q(S_1 - S_3) - F_3 & q > 0 \\ q(S_3 - S_2) - F_3 & q \leq 0\,. \end{cases} \tag{3.3e}$$

Here V denotes the volume ratio between the equatorial and polar boxes and $S_2 = (S_0 - S_1 + S_3)/V$, where S_0 is the total salinity. Furthermore, H_i and F_i denote heat and freshwater fluxes into the boxes, respectively. The heat fluxes are modeled as a Newtonian relaxation to a reference temperature Θ_i representing a typical climatological value for each box,

$$H_i = \lambda_T(\Theta_i - T_i) \qquad i = 1, 2, 3\,, \tag{3.4}$$

where $\lambda_T = 1.29 \times 10^{-9}\,\mathrm{s}^{-1}$ is the relaxation rate corresponding to a timescale of around 25 years. This relaxation crudely accounts for both radiative surface fluxes as well as meridional oceanic heat transport. The freshwater fluxes are parameterized as

$$F_i = \bar{F}_i - \lambda_S(\tilde{S}_i - S_i) \qquad i = 1, 3\,, \tag{3.5}$$

where $\bar{F}_i$ is a constant mimicking the surface freshwater flux and the additional term accounts for salt diffusion between the boxes, approximated as a Newtonian relaxation[1] at rate λ_S to a reference salinity $\tilde{S}_i$. Since the total salinity is conserved, the sum of surface freshwater fluxes entering the high-latitude boxes must equal the surface freshwater flux out of the tropical box,

$$F_2 = -\frac{1}{V}(F_1 + F_3)\,. \tag{3.6}$$

This mimics enhanced evaporation in the tropics and precipitation at high-latitude in accordance with observations (differing from the box model by Rahmstorf (1996) where the surface freshwater flux is always northward). The values of the model parameters are listed in Table 3.1.

[1] To model salt diffusion more physically, the relaxation should depend on the salinity difference *between* boxes and not the salinity difference between each box and a reference value $\bar{S}_i$. This would introduce an additional linear coupling.

Parameter	Value	Unit	Description
F_1	is varied	$\mathrm{g\,kg^{-1}\,s^{-1}}$	Northern freshwater flux
$\bar{F}_3$	9×10^{-11}	$\mathrm{g\,kg^{-1}\,s^{-1}}$	Southern freshwater flux ($\approx 0.27\,\mathrm{Sv}$)
κ	1.5×10^{-6}	$\mathrm{s^{-1}}$	Hydraulic constant
α	1.5×10^{-4}	$\mathrm{K^{-1}}$	Thermal expansion coeff.
β	8×10^{-4}	–	Haline expansion coeff.
λ_T	1.29×10^{-9}	$\mathrm{s^{-1}}$	Temperature restoring rate (≈ 25 years)
λ_S	3×10^{-11}	$\mathrm{s^{-1}}$	Salinity restoring rate (≈ 1000 years)
Θ_1, Θ_2, Θ_3	0.0, 30.0, 0.0	°C	Reference temperatures
$\tilde{S}_1$, $\tilde{S}_2$, $\tilde{S}_3$	35.0, 35.0, 35.0	$\mathrm{g\,kg^{-1}}$	Reference salinities
S_0	140.0	–	Total salinity
V	2	–	Box volume ratio

Table 3.1 | Parameters of the Rooth model (settings based on Scott et al. (1999) and Lucarini and Stone (2005)).

3.1.2 Modifications to the original model

Salt diffusion. In the original formulation of the model (Rooth 1982), diffusive fluxes of salinity between the boxes are neglected, i.e. $\lambda_S = 0$. As we will see, this leads to unphysical divergent behavior near an instability threshold. We argue that turbulent eddies in the ocean will cause diffusive freshwater exchange across the Atlantic basin (Treguier et al. 2012; Treguier et al. 2014), in addition to wind-driven gyre transport (Wood et al. 2019), and that therefore λ_S should be nonzero. We set $\lambda_S = 3 \times 10^{-11}\,\mathrm{s^{-1}}$, corresponding to a diffusive timescale of around 1000 years. The temperature relaxation is thus significantly faster than the salinity relaxation: $\lambda_T/\lambda_S = 43$.

Smooth dynamics. A mathematical challenge of the Rooth model (as well as other AMOC box models (Stommel 1961; Cimatoribus et al. 2014; Wood et al. 2019)) is the non-smooth behavior at $q = 0$, where the drift

$$\boldsymbol{b}_{\mathrm{Rooth}}(\boldsymbol{x}) = (\dot{T}_1,\ \dot{T}_2,\ \dot{T}_3,\ \dot{S}_1,\ \dot{S}_3)$$

is not differentiable. Where necessary, we may approximate the dynamics in Eq. (3.3) by a smooth version,

$$\dot{T}_1 = \frac{q}{2}\big(c_+(T_2 - T_1) + c_-(T_1 - T_3)\big) + H_1\ , \tag{3.7}$$

and similarly for $\dot{T}_2$, $\dot{T}_3$, $\dot{S}_1$, $\dot{S}_3$, where $c_\pm = 1 \pm \tanh(\xi q)$. The smooth version limits to the original ODE as $\xi \to \infty$. For $\xi \sim 10^{12}\,\mathrm{s}$ (note that $q \sim 10^{-10}\,\mathrm{s^{-1}}$), we find that differences between the solutions of Eqs. (3.7) and (3.3) become negligible.

3.2 Stability and hysteresis under freshwater forcing

3.2.1 Equilibria under symmetric forcing

With the parameter settings in Table 3.1, the Rooth model is symmetric with respect to both hemispheres if the freshwater fluxes $\bar{F}_1$ and $\bar{F}_3$ are identical. Even under symmetric forcing, two stable equilibria coexist: a solution with $q > 0$, corresponding to a northward surface current and sinking in the northern box, and a solution with $q < 0$ representing

Eigenvalue	ON/OFF ($\bar{F}_1 = \bar{F}_3$)	ON ($\bar{F}_1 = 1.5\bar{F}_3$)	OFF ($\bar{F}_1 = 1.5\bar{F}_3$)	M ($\bar{F}_1 = \bar{F}_3$)	M ($\bar{F}_1 = 1.5\bar{F}_3$)
γ_1	-0.215	-0.205	-0.215	-0.110	-0.111
γ_2	-0.041	-0.041	-0.041	-0.041	-0.041
γ_3	-0.040	-0.040	-0.040	-0.041	-0.041
γ_4	$-0.006 - 0.007i$	$-0.004 - 0.008i$	$-0.009 - 0.008i$	-0.001	-0.001
γ_5	$-0.006 + 0.007i$	$-0.004 + 0.008i$	$-0.009 + 0.008i$	$+0.081$	$+0.079$

Table 3.2 | Eigenvalues of the Jacobian of $\boldsymbol{b}_{\text{Rooth}}$ at each equilibrium for $\bar{F}_1 = \bar{F}_3$ and $\bar{F}_1 = 1.5\bar{F}_3$. For symmetric forcing, the eigenvalues of the ON and OFF states are identical due to the symmetry.

a reversed circulation. The $q > 0$ solution resembles today's overturning circulation, and we will call it the "ON" state in reference to the literature on AMOC multistability. Correspondingly, we will refer to the $q < 0$ solution as the "OFF" state (even though in this box model the circulation is not collapsed but reversed, like in the Stommel model).

With $\bar{F}_1 = \bar{F}_3 = 9 \times 10^{-11}\,\mathrm{g\,kg^{-1}\,s^{-1}} \approx 0.27\,\mathrm{Sv}$, which roughly corresponds to the hydrology of the southern hemisphere (Baumgartner and Reichel 1975), the ON and OFF states (for the settings in Table 3.1) are

$$\boldsymbol{x}_{\text{ON}} \approx (2.54°\text{C},\ 28.62°\text{C},\ 0.23°\text{C},\ 34.80\,\mathrm{g\,kg^{-1}},\ 34.26\,\mathrm{g\,kg^{-1}})\,,$$
$$\boldsymbol{x}_{\text{OFF}} \approx (0.23°\text{C},\ 28.62°\text{C},\ 2.54°\text{C},\ 34.26\,\mathrm{g\,kg^{-1}},\ 34.80\,\mathrm{g\,kg^{-1}})\,.$$

This yields an AMOC strength of $q_{\text{ON}} \approx 1.26 \times 10^{-10}\,\mathrm{s^{-1}}$ corresponding to around 13.6 Sv, which differs from the value of $q \approx 15.5\,\mathrm{Sv}$ reported in Scott et al. (1999). The difference originates from the choice of the salinity restoring rate λ_S, which we set to $3 \times 10^{-11}\,\mathrm{s^{-1}}$ instead of zero (see section 3.1.2). Due to the nonzero salt diffusion, the high-latitude boxes are slightly saltier at the stable equilibria, which also affects the equilibrium temperatures and the flow strength q.

The Jacobian of $\boldsymbol{b}_{\text{Rooth}}$ at the ON and OFF states each has a pair of complex conjugate eigenvalues, indicating damped oscillatory behavior as trajectories converge to these equilibria (Table 3.2). The real parts of all eigenvalues are negative as required for a stable fixed point.

In addition to the ON and OFF attractors, the model has an unstable equilibrium at

$$\boldsymbol{x}_M \approx (0.0°\text{C},\ 30.0°\text{C},\ 0.0°\text{C},\ 32.0\,\mathrm{g\,kg^{-1}},\ 32.0\,\mathrm{g\,kg^{-1}})\,,$$

where there is no flow ($q = 0$) and the freshwater forcing is entirely compensated by salt diffusion from the tropical box to the high-latitude boxes. This can only be achieved if box 2 is saltier than boxes 1 and 3 ($S_2 = 38\,\mathrm{g\,kg^{-1}}$ at $\boldsymbol{x}_M$). As the diffusivity decreases, i.e. $\lambda_S \to 0$, the equator-to-pole salinity difference required to sustain the unstable equilibrium diverges, and the solution $\boldsymbol{x}_M(\lambda_S)$ becomes increasingly unphysical. Rooth (1982) and Scott et al. (1999) comment on the "no-flow solution" for $\lambda_S = 0$ as being a transient solution where $S_1 = S_3 = S_0/V - S_2 = -\bar{F}_1 t$, which they conclude is physically unrealistic but relevant for the transient response to large perturbations. Indeed, this solution is interesting from a global stability perspective, as it describes the dynamics on the basin boundary (see below).

3.2.2 Melancholia state

In section 2.1.4, we defined a Melancholia (M) state as an unstable state on the basin boundary that is an attractor of the dynamics restricted to the basin boundary (definition

2.11). The unstable equilibrium $\boldsymbol{x}_M$ is in fact an M state – a saddle point whose stable manifold is the four-dimensional basin boundary $\partial\mathcal{B}$. We can verify this by calculating the Jacobian eigenvalues at $\boldsymbol{x}_M$, giving four negative and one positive eigenvalue. The basin boundary contains the line

$$\mathcal{M} = \{\boldsymbol{x} : T_i = \Theta_i \ \forall i \text{ and } S_1 = S_3\} \subset \partial\mathcal{B}\,,$$

along which the boxes decouple and the surface heat fluxes vanish ($q = 0$, $\dot{q} = 0$, $H_i = 0$), but we emphasize that the entire basin boundary is a hypersurface of dimension $D_b = D - 1 = 4$ (if $D_b < D - 1$, it would not separate the basins of attraction of $\boldsymbol{x}_{\mathrm{ON}}$ and $\boldsymbol{x}_{\mathrm{OFF}}$).

For $\lambda_S = 3 \times 10^{-11}\,\mathrm{s}^{-1}$, the dynamics on the basin boundary are attracted to $\boldsymbol{x}_M$ with $S_M = S_1 = S_3 \approx 32\,\mathrm{g\,kg}^{-1}$. As mentioned in the previous section, $S_M \to -\infty$ as $\lambda_S \to 0$. In other words, the original model formulation with $\lambda_S = 0$ possesses a *singular* M state at $\boldsymbol{x}_M = (\Theta_1, \Theta_2, \Theta_3, -\infty, -\infty)$. For practical purposes, the singularity of the M state may not be an issue, since any infinitesimal perturbation away from the basin boundary $\partial\mathcal{B}$ will cause the model to converge either to the ON state or the OFF state. For the mathematical aim of constructing a stability landscape, however, this presents a pathological situation.

Recall that candidates for a scalar function Φ representing a stability landscape should have the properties of a Lyapunov function (Eq. (2.11)), specifically: $\mathrm{d}\Phi/\mathrm{d}t < 0$ along trajectories out of equilibrium. Since $\dot{S}_i = -\bar{F}_1$ for $i \in \{1, 3\}$ along $\mathcal{M}$, the function Φ must decrease monotonically along $\mathcal{M}$ for decreasing S_i. This implies that no global minimum of Φ exists at finite salinity, and there is no physically reasonable "mountain pass" in the stability landscape that would act as the gateway for noise-induced transitions in the weak-noise limit (see section 2.1.6). Contrarily, for $\lambda_S > 0$ an M state $\boldsymbol{x}_M$ exists at finite salinity, and in principle it should be possible to construct a Lyapunov function that assumes a local minimum on the basin boundary at $\boldsymbol{x}_M$, as well as local minima at the ON and OFF states.

Asymmetric forcing. Let us now consider the case of asymmetric freshwater forcing ($\bar{F}_1 \neq \bar{F}_3$). A ratio of $\bar{F} := \bar{F}_1/\bar{F}_3 = 1.5$ may be considered more representative of Earth's hydrology (Lucarini and Stone 2005). This changes the stability landscape of the model. The forcing asymmetry breaks the symmetry of the ON and OFF states, and the basin boundary no longer contains the set $\mathcal{M}$. However, an M state $\boldsymbol{x}_M$ still exists with $q \neq 0$. We can find it either by calculating the zeroes of $\boldsymbol{b}_{\mathrm{Rooth}} = 0$ numerically or by using the edge tracking algorithm (see section 2.2.1). Since the edge tracking algorithm is an essential tool of this thesis, let us also apply it here.

Using our implementation in the `Attractors.jl` software package[2] (Datseris et al. 2023), we initialize the algorithm from the ON and OFF states (at $\bar{F} = 1.5$), setting $\varepsilon_1 = 10^{-4}$ and $\varepsilon_2 = 5\varepsilon_1$ measured in terms of the Euclidean distance in the five-dimensional phase space. We integrate the model using a 5th-order Runge-Kutta solver (Tsitouras 2011; Rackauckas and Nie 2017). The algorithm tracks along the basin boundary until it converges to a state $\boldsymbol{x}_M$ after 113 iterations[3] (Fig. 3.2). By plotting the edge track in

[2] Documentation: `https://juliadynamics.github.io/DynamicalSystemsDocs.jl/attractors/stable/api/#Attractors.edgetracking`

[3] The number of iterations naturally depends on the initial conditions and hyperparameter settings of the algorithm.

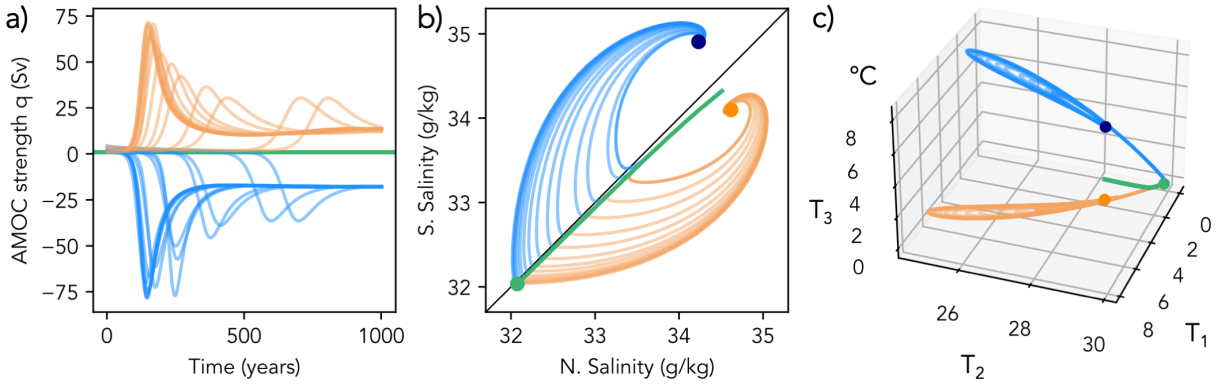

Figure 3.2 | **Edge tracking in the Rooth model** for $\bar{F}_1/\bar{F}_3 = 1.5$. The algorithm is initialized from the ON (orange point) and OFF (blue point) states. Orange (blue) lines show the relaxation paths of the pair of trajectories used at every 20-th iteration of the algorithm. The algorithm tracks along the basin boundary (green line) until it converges to the M state (green point). (a) AMOC timeseries of the relaxation paths, with the AMOC strength of the M state indicated (green line). (b) Phase space view of the edge tracking procedure in salinity space. (c) as (b) but in temperature space.

the reduced phase space spanned by the salinity variables (salinity space), we see that it lies close to the $S_1 = S_3$ line but is closer to the ON state than to the OFF state. We also observe the rotational behavior in the relaxation paths anticipated from the complex Jacobian eigenvalues of the attractors.

For reference, the equilibria at $\bar{F} = 1.5$ are

$$\begin{aligned}
\boldsymbol{x}_{\mathrm{ON}} &\approx (2.45°\mathrm{C},\ 28.67°\mathrm{C},\ 0.21°\mathrm{C},\ 34.62\,\mathrm{g\,kg^{-1}},\ 34.10\,\mathrm{g\,kg^{-1}})\,,\\
\boldsymbol{x}_{\mathrm{OFF}} &\approx (0.37°\mathrm{C},\ 28.21°\mathrm{C},\ 3.22°\mathrm{C},\ 34.23\,\mathrm{g\,kg^{-1}},\ 34.91\,\mathrm{g\,kg^{-1}})\,,\\
\boldsymbol{x}_{M} &\approx (0.18°\mathrm{C},\ 29.91°\mathrm{C},\ 0.00°\mathrm{C},\ 32.08\,\mathrm{g\,kg^{-1}},\ 32.03\,\mathrm{g\,kg^{-1}})\,,
\end{aligned}$$

with $q(\boldsymbol{x}_{\mathrm{ON}}) \approx 13.04\,\mathrm{Sv}$, $q(\boldsymbol{x}_{\mathrm{OFF}}) \approx -17.96\,\mathrm{Sv}$ and $q(\boldsymbol{x}_M) \approx 0.85\,\mathrm{Sv}$.

3.2.3 Resilience against freshwater bursts

Abrupt climate changes during the last glacial period are often linked to AMOC transitions, some of which could have been triggered by large, sudden freshwater discharges in the North Atlantic (see section 2.1.7). In the context of our box model, we may ask: how much additional freshwater discharge into the northern box can the ON circulation withstand without tipping to the OFF state?

We can quantify this in terms of the *minimal fatal shock*, i.e. the minimal instantaneous perturbation needed to kick a multistable system from an attractor out of its basin of attraction (Halekotte and Feudel 2020). This provides a simple measure of global stability and resilience in the sense of the system's ability to return to the original state following a perturbation (see section 2.1.7). Generally, the minimal fatal shock is a vector in the phase space. Since we are interested in the minimal fatal North Atlantic freshwater perturbation, we only consider perturbations of $\bar{F}_1$, which point in the direction of S_1.

Adding an instantaneous freshwater burst corresponds to reducing S_1 in the initial condition. Via bisection, we find that for $\bar{F} = 1.5$ the critical northern freshwater shock lies within

$$0.299\,338\,\mathrm{g\,kg^{-1}} < \Delta\bar{F}_{1,\mathrm{crit}}\Delta t < 0.299\,339\,\mathrm{g\,kg^{-1}}\,,$$

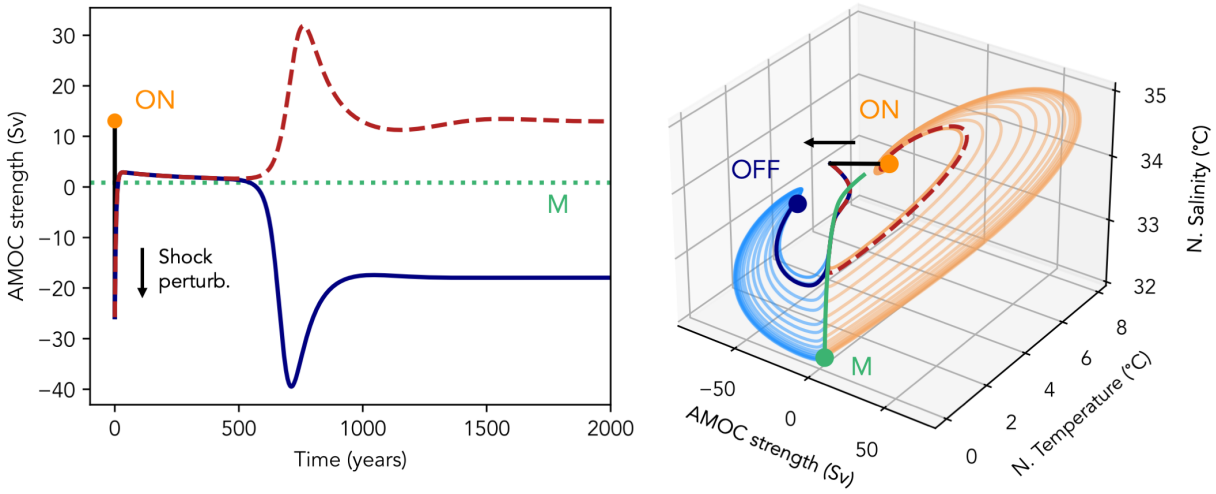

Figure 3.3 | **Minimal fatal freshwater shock** in the northern box. Left: Subcritical (red dashed) and supercritical (blue) trajectory after a shock perturbation (black line) applied to the ON state (orange). The AMOC strength q of the M state is indicated (green). Right: Reduced phase space view spanned by q, T_1 (N. temperature) and S_1 (N. salinity), showing also the edge tracking relaxation trajectories from Fig. 3.2.

where Δt is the simulation time step. This value is about 25% smaller than the value found by Scott et al. (1999) for a corresponding experiment with the original model ($0.387\,\mathrm{g\,kg^{-1}}$ for $\lambda_S = 0$). The salt relaxation term has thus made the system more vulnerable to shock-induced tipping.

Forcing the model with the lower and upper bound of this estimate yields a subcritical and supercritical trajectory, respectively (Fig. 3.3, see also Scott et al. (1999)). Since the AMOC strength depends instantly on the high-latitude salinity difference, the perturbation leads to an immediate reversal and strengthening of the circulation for both trajectories. This quickly brings warm water from the tropics into the South Atlantic, counteracting the reversal. Then, both trajectories undergo a long transient where they track the basin boundary towards the M state. Due to the weak flow, the boxes are weakly coupled and the fast ocean-atmosphere heat exchange stabilizes the transient. In the meantime, however, the equator-pole salinity gradient grows steadily, strengthening a destabilizing salt-advection feedback. After around 500 years the super- and subcritical trajectories diverge from each other and relax to the OFF and ON state, respectively. A phase space projection onto q, S_1 and T_1 shows that the relaxation paths resemble those found during the edge tracking procedure (Fig. 3.3). This is not surprising, since the minimal fatal shock kicks the system directly to the basin boundary.

The minimal fatal shock experiment highlights that the asymptotic state of the system following a critical perturbation can be effectively unpredictable for a long time. Without precisely knowing the location of the basin boundary, the two trajectories in Fig. 3.3 are virtually indistinguishable for more than 500 years. In other words, imagine we live in the Rooth model and experience a near-critical freshwater burst. Even hundreds of years after the burst, we might not know whether the AMOC has "survived" the shock or not. Of course, this situation appears because we are very close to the basin boundary, which arguably has a low probability of occurrence under random shocks. In more complex systems with complicated basin boundary geometries, the chance of landing near the

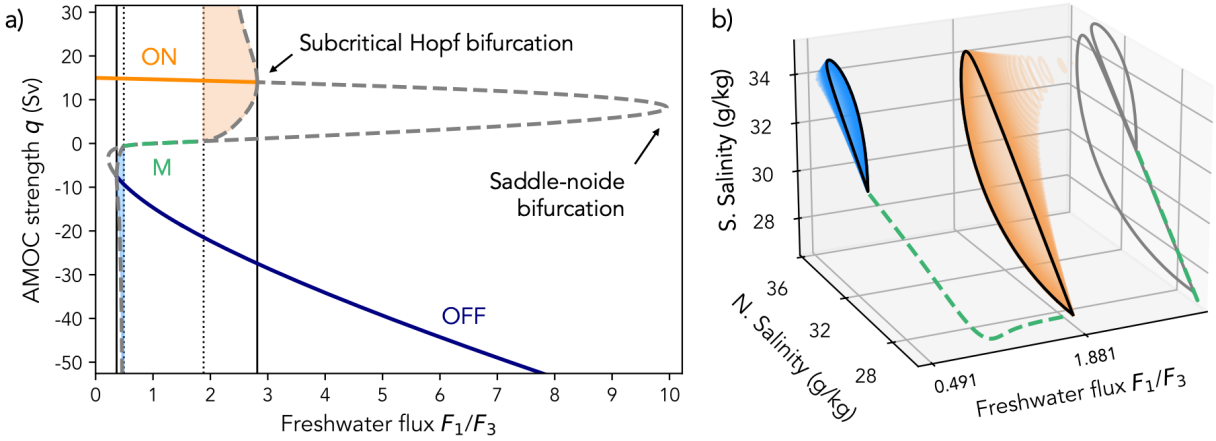

Figure 3.4 | **Bifurcation diagram** of the Rooth model with respect to $\bar{F}_1$, with $\bar{F}_3 = 9 \times 10^{-11}\,\mathrm{s}^{-1}$ and $\lambda_S = 1.1 \times 10^{-11}\,\mathrm{s}^{-1}$. (a) Continuation of equilibria as a function of the freshwater forcing ratio F_1/F_3, projected onto the AMOC strength q. Shaded areas indicate unstable limit cycles, bounded by subcritical Hopf (solid vertical lines) and homoclinic (dotted vertical lines) bifurcations. In a part of the bistable regime, the Melancholia state is a saddle point (green dashed). (b) Phase space projections (salinity space) of the unstable limit cycles (blue, orange) and saddle M state (green) as a function of the bifurcation parameter F_1/F_3.

boundary becomes more relevant, as we discuss in chapter 5.

3.2.4 Bifurcation diagram for freshwater forcing

The shock-induced AMOC tipping in the previous section may be seen as the limiting case of rate-induced tipping where the parameter $\bar{F}_1$ is changed at an infinitely fast rate (see section 2.1.7). In the supercritical case, the system is catapulted into the competing basin of attraction and thus fails to track the ON attractor. Conversely, varying the control parameter infinitely slowly (adiabatically) across a range of interest yields the *bifurcation diagram* of the system with respect to $\bar{F}_1$. The adiabatic rate of change allows the system to track the attractor in phase space until it undergoes a qualitative change in its stability – a bifurcation.

Bifurcation diagrams under North Atlantic freshwater forcing have been computed for various climate models across the model hierarchy (see section 1.2.4) (Manabe and Stouffer 1995; Rahmstorf et al. 2005; Hawkins et al. 2011; Jackson et al. 2023; van Westen and Dijkstra 2023). In complex models, this is commonly achieved via hosing experiments where the freshwater input is ramped up and down as slowly as possible within computational constraints. Using numerical continuation methods, both stable and unstable equilibrium branches have been computed for a global ocean model (Dijkstra and Weijer 2005).

Given the simplicity of the Rooth model, we can study its stability across parameter space using standard continuation algorithms. Here we use the Matlab-based `MATCONT` software (Dhooge et al. 2003). For this purpose, we use the smoothened version of the drift $\boldsymbol{b}_{\mathrm{Rooth}}$ (see Eq. (3.7)) with $\xi = 10^{13}\,\mathrm{s}$. After rescaling time in the model equations from seconds to years, the method worked robustly.

The stability properties of the Rooth model with respect to $\bar{F}_1$ have been studied by Scott et al. (1999), who showed that the model has a bistable regime limited by two

subcritical Hopf bifurcations. In our modified version of the model, where a nonsingular M state exists, we can complement their results by exactly calculating the unstable equilibrium branch connecting the limit cycles emanating from the Hopf bifurcations. Additionally, we explicitly compute these unstable limit cycles.

As a short recap in bifurcation theory: there are different types of bifurcations (Crawford 1991). At a local bifurcation, a fixed point loses stability as the real part of an eigenvalue of its Jacobian crosses the imaginary axis. For higher-dimensional invariant sets, more complicated bifurcations arise.

Term 3.1. **Bifurcation: Saddle-node, Hopf, Homoclinic**

At a *saddle-node bifurcation*, or *fold* bifurcation, two fixed points collide and annihilate each other.

At a *Hopf bifurcation*, a fixed point loses stability and turns into a limit cycle. If the bifurcation is *supercritical*, the resulting limit cycle is stable; in the *subcritical* case it is unstable.

A *homoclinic bifurcation* is a global bifurcation where a saddle point collides with a limit cycle. At the bifurcation, the limit cycle becomes a homoclinic orbit with infinite period.

Bifurcation diagram. In the region of near-symmetric freshwater forcing, the modified Rooth model is bistable, and the M state is the saddle point on the basin boundary (Fig. 3.4). As we increase the ratio $\bar{F} = \bar{F}_1/\bar{F}_3$, the M state collides with an unstable limit cycle at a homoclinic bifurcation (Titz et al. 2002). This limit cycle contracts as $\bar{F}$ is further increased. In this regime, the limit cycle becomes the Melancholia state, located on the basin boundary (as a reminder, the basin boundary is a 4D hypersurface). The saddle point still exists in the phase space but no longer lies on the basin boundary and is thus not an edge state. Increasing $\bar{F}_1$ further, the contracting limit cycle eventually collides with the stable ON state at a subcritical Hopf bifurcation. Beyond the Hopf bifurcation, the ON state is unstable. The unstable portion of the ON branch can still be continued up to a saddle-node bifurcation, where it meets with the saddle point. Note that the saddle-node bifurcation is located at $\bar{F} \approx 10$, far away from the bistable regime. Here, the OFF state is the only attractor, with a strong reversed circulation of more than 50 Sv.

If we follow the stable OFF branch back towards smaller values of $\bar{F}$, we first enter the bistable regime, then the M state switches from limit cycle to saddle, only to become a limit cycle again at another homoclinic bifurcation; this limit cycle contracts with decreasing $\bar{F}$ until it collides with the OFF state at the second subcritical Hopf bifurcation, marking the lower end of the bistable regime. Slightly outside the bistable regime, there is another saddle-node bifurcation at which the unstable OFF state collides with the saddle point.

Qualitatively, we thus recover the same bifurcation diagram for $\lambda_T > 0$ as for the original Rooth model (Scott et al. 1999). In the original version, the stable equilibrium depends only on the freshwater flux of the high-latitude box in which there is upwelling.

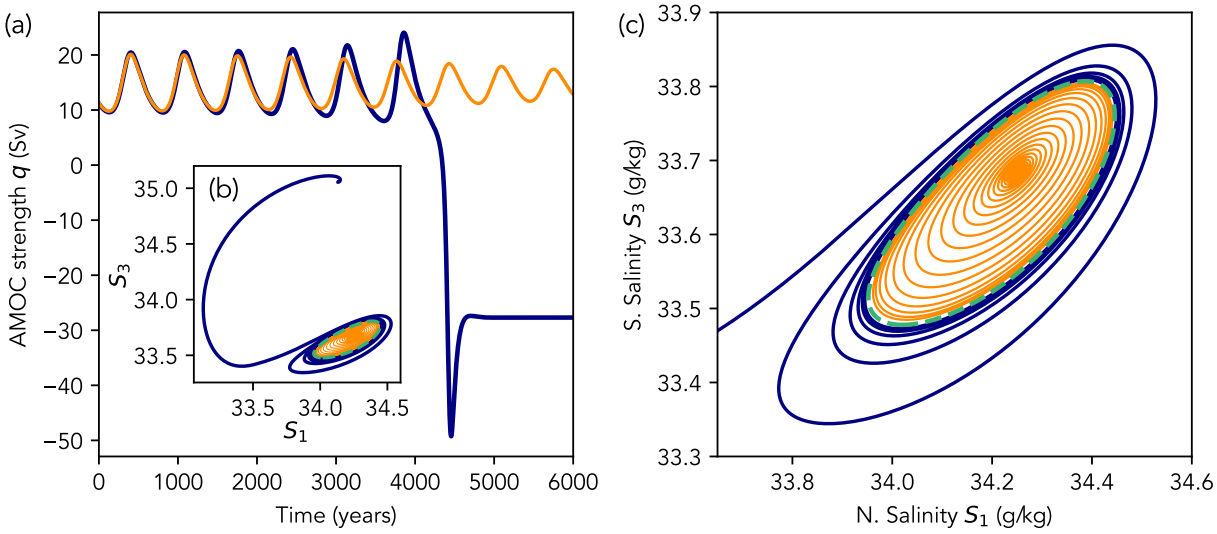

Figure 3.5 | **Transient dynamics near the unstable limit cycle** for $\bar{F} = 2.74$. (a) Two trajectories from nearby initial conditions, one within (orange) and one outside (blue) the limit cycle. (b) Phase space projection onto salinity space, showing also the limit cycle (green dashed). (c) Zoom-in of (b). We recommend thinking of this figure when watching the film *Melancholia*.

This implies that the flow strength of the ON state is invariant under changes in $\bar{F}_1$. By contrast, with $\lambda_T > 0$ the strength of the ON branch depends weakly on $\bar{F}_1$ (Fig. 3.4).

Unstable limit cycle. At each end of the bistable regime, between the subcritical Hopf bifurcation and the homoclinic bifurcation, the unstable limit cycle constitutes the M state. To investigate the dynamics near the M state, we initialize one trajectory slightly inside the limit cycle (in salinity space) and another slightly outside (Fig. 3.5).

Initially, both trajectories undergo similar multi-centennial oscillations as they track the limit cycle. As illuminated by (Scott et al. 1999), the oscillations are related to salt advection in the model: given the flushing timescale of the boxes and how they are connected to each other, salinity anomalies in the tropical box must be advected through the northern box to reach the southern box for $q > 0$. This introduces different timescales in the salinity response, leading to oscillations. The inner trajectory exhibits damped oscillations and spirals towards the ON state, to which it converges. By contrast, the outer trajectory increases in amplitude and sharply transitions to the OFF state after six cycles. Notably, the transition occurs only after a transient period of 4000 years. Melancholia states can give rise to long transients in their vicinity, which challenges the predictability of the system's long-term behavior.

3.3 Noise-induced transitions

So far, we have studied the Rooth model deterministically, focusing on the dynamics near the Melancholia state in the bistable regime. To conclude this introductory chapter, let us now apply stochastic forcing to explore the transition behavior between the ON and OFF states. We will contrast Gaussian and non-Gaussian noise in the form of α-stable Lévy noise.

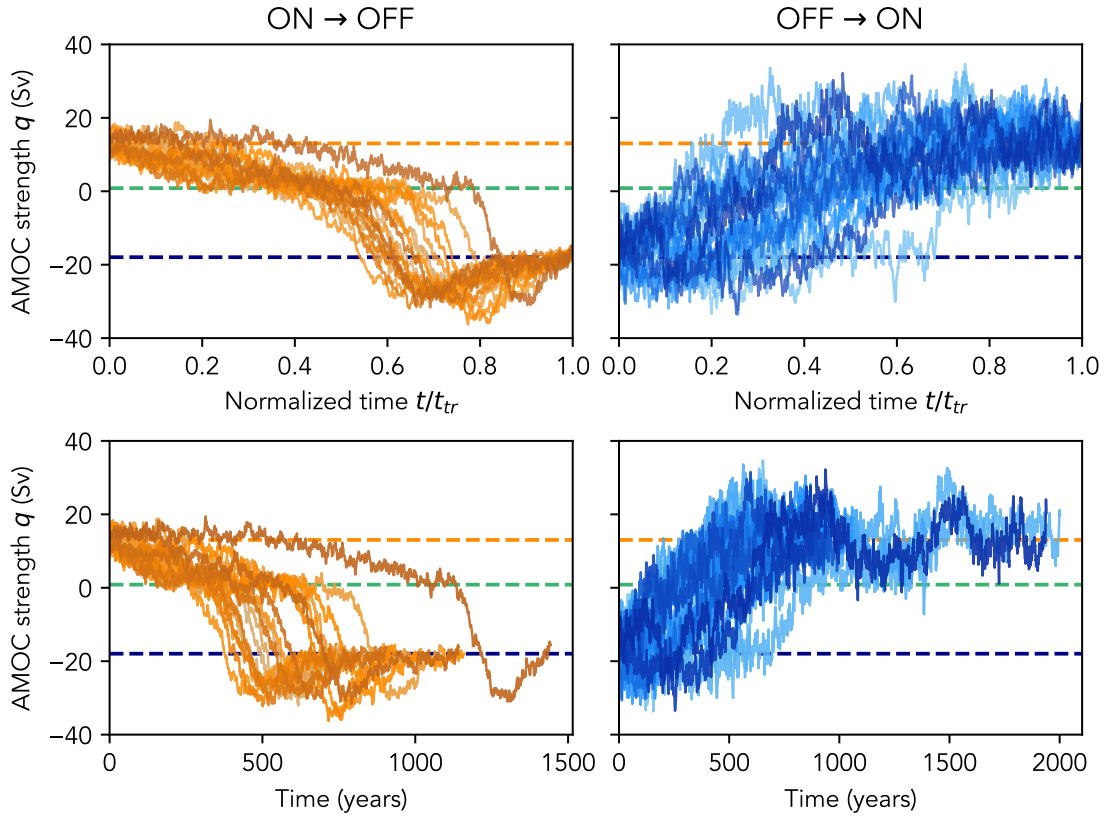

Figure 3.6 | **Sample transition paths** under isotropic Gaussian noise for the scenarios ON → OFF (left, $\sigma = 0.0045$) and OFF → ON (right, $\sigma = 0.0112$). Top panels show the same as the bottom panels but in units of the transition time t_{tr} for each individual path. Dashed lines indicate the AMOC strength of the ON (orange), OFF (blue) and M (green) state, respectively.

3.3.1 Gaussian noise

First, we consider an SDE of the form in Eq. (2.13),

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{b}_{\text{Rooth}}(\boldsymbol{x})\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{W}_t\,, \tag{3.8}$$

where $\sigma > 0$ is the noise strength and $\boldsymbol{W}_t$ denotes a 5-dimensional vector of independent standard Wiener processes. Thus, we consider isotropic Gaussian noise with the noise covariance matrix $\boldsymbol{Q} = \mathbb{I}_5$.

In reality, the stochastic variability will generally be correlated between state variables; the variances may differ significantly among the variables and also be state-dependent (Chapman et al. 2024). Here, we regard the noise rather as a means to explore the phase space and generate noise-induced transitions. We select relatively weak noise such that transitions occur as rare events, in which case we might expect large deviation theory to describe the statistics of the tipping behavior (see section 2.1.6). Physically, the noise may be viewed as an idealized way of accounting for high-frequency variability in the atmospheric heat and freshwater fluxes.

To sample noise-induced transition paths of Eq. (3.8), we employ Monte Carlo rejection sampling as described in section 2.2.2 (see also section 4.2.2), using the `transitions` method of the CriticalTransitions.jl package[4]. We set $\bar{F} = 1.5$ and consider both transition

[4]Open-source code: `https://github.com/JuliaDynamics/CriticalTransitions.jl`. Here we used version v0.5 but recommend using the latest version.

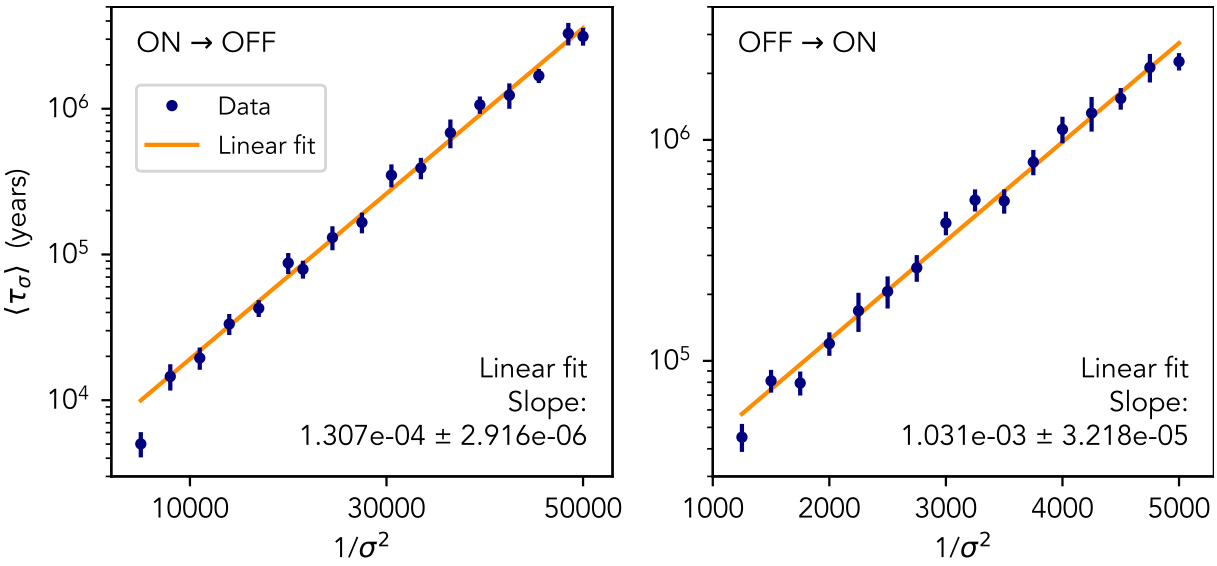

Figure 3.7 | **Kramers' scaling (Arrhenius law) in the Rooth model** for the scenarios ON → OFF (left) and OFF → ON (right). $\langle \tau_\sigma \rangle$ is plotted on a logarithmic scale. For the linear least-squares fit, the three data points with the largest noise strengths were omitted to obtain a more accurate result.

scenarios ON → OFF as well as OFF → ON. The SDE is solved with the Euler-Maruyama scheme (Eq. (2.35)) using a time step of $\Delta t = 0.01$ years (3.65 days).

For the ON → OFF scenario, we sample 64 transitions each at 17 different noise strengths in the interval $\sigma \in [0.0044, 0.0142]$. At the upper end of this interval, the mean waiting time to observe a transition is on the order of 10 000 years, increasing to around 3 million years at the lower end. We obtain a similar range of waiting times in the OFF → ON scenario when selecting $\sigma \in [0.0141, 0.0283]$. Here we collect 64 transitions each at 16 different noise levels.

At the lowest noise strength considered in each scenario (waiting time of around 3 million years), the transition time series of the flow strength q show a similar behavior in each scenario, but the timing of the transition differs depending on how quickly the trajectories reach the basin boundary after leaving a neighborhood (ball of radius 0.1 in phase space) around the initial attractor (Fig. 3.6). Clearly, the transition paths from OFF → ON are noisier.

Kramers' scaling of transition rates. The fact that the ON → OFF and OFF → ON scenarios require different noise strengths to yield similar waiting times indicates an asymmetry in the transition rates. According to Arrhenius' law for nongradient systems (see Eq. (2.24)), the scaling relationship between the noise strength σ and mean first-exit time $\langle \tau_\sigma \rangle$ is given by,

$$\left\langle \tau_\sigma^A \right\rangle \asymp \exp\left(\sigma^{-2} V_A(\boldsymbol{x}_M)\right) \tag{3.9}$$

where $A \in \{\text{ON}, \text{OFF}\}$ indicates the reference attractor and we assume that the minimum of the quasipotential V_A along the basin boundary is attained at the M state $\boldsymbol{x}_M$. Our simulations verify the exponential relationship within the noise ranges considered (Fig. 3.7). By performing a linear least-squares fit of $\ln \left\langle \tau_\sigma^A \right\rangle$ against $1/\sigma^2$, we obtain an estimate of the quasipotential value at $\boldsymbol{x}_M$ with respect to each attractor. Intuitively, this value

represents the nonequilibrium "energetic" barrier that must be overcome by the stochastic trajectory to transition to the competing attractor.

We find that the barrier from OFF $\rightarrow$ ON is around 7.9 times higher than the barrier from ON $\rightarrow$ OFF. Thus, by analyzing the statistics of noise-induced transitions, we gain insight into the stability landscape of the underlying deterministic system: we learned that at $\bar{F} = 1.5$ the ON state is significantly less stable – and thus more sensitive to perturbations – than the OFF state. This strong asymmetry in the quasipotential is not directly deducible from the bifurcation diagram in Fig. 3.4.

Transition paths. Analyzing the transition paths in phase space offers insight into the mechanism and typicality of transitions. To do this, we plot the sample transition paths in the reduced phase spaces of salinity and temperature (Fig. 3.8). We consider two cases: (1) similar rareness of transitions between the scenarios ON $\rightarrow$ OFF and OFF $\rightarrow$ ON, and (2) equal noise strengths for both scenarios.

We observe that the density of transition paths concentrates around some typical path for each scenario (Lucarini et al. 2023). A rough estimate of a mean sample transition path (MTP) is obtained by interpolating each path in normalized time (with respect to its transition time) and then averaging over the transition path ensemble at each time point. This procedure is not ideal because the abrupt part of the transitions occurs at different normalized times (see Fig. 3.6). For our purposes, however, the resulting MTP captures the essence of the behavior. We discuss a different way of estimating the MTP in section 4.2.2.

In temperature space, all sample transitions follow a similar path that clearly leads via the M state, as we would expect in the weak-noise limit. However, in salinity space we get a more nuanced picture (Figs 3.8a and c). First of all, the MTP from ON to OFF is not the same as the reverse of OFF to ON. This is a well-known feature of nongradient systems that highlights the lack of detailed balance (see section 2.1.6). Transition paths from OFF to ON climb up to the basin boundary in a clear channel but distribute over a larger region after crossing the basin boundary, owing to the rotational drift around the ON state. In the ON $\rightarrow$ OFF scenario, we see that weaker noise leads to a more distinct typical transition path that undergoes an excursion to lower high-latitude salinities (Figs 3.8a and c). We can understand this by recalling that the flow along the basin boundary leads towards the M state located at lower high-latitude salinity. For weaker noise, transition paths tend to follow the stable manifold of the M state for longer before the noise kicks them over to one of the attractors.

As an unexpected result, none of the sampled transition paths get close to the M state in salinity space, which is located at $(S_1, S_3) \approx (32, 32)\,\mathrm{g\,kg^{-1}}$ (see also Fig. 3.11). This is surprising because the transitions represent one-in-three-million-year events, an thus we would expect them to be accurately described by Freidlin-Wentzell theory (see section 2.1.6). What we encounter here is a phenomenon termed *saddle avoidance*, which we study in detail in chapter 4.

3.3.2 Lévy noise

Finally, let us take a brief look at the transition behavior under degenerate non-Gaussian noise (see page 41, section 2.1.6), i.e. we consider the SDE,

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{b}_{\mathrm{Rooth}}(\boldsymbol{x})\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{L}_t^{\alpha}\,, \tag{3.10}$$

where $\boldsymbol{L}_t^\alpha = (0, 0, 0, L_t^\alpha, 0)$ and L_t^α is the increment of an α-stable Lévy process. Thus, we add stochastic forcing only to the freshwater flux in the northern box. We simulate the symmetric Lévy process using the method by Chambers et al. (1976) implemented in Matlab (`stblrnd` function). As is well known, the heavy-tailed distribution of Lévy noise fundamentally changes the mechanism and most probable path of transitions (Lucarini et al. 2022). Physically, this choice is motivated by the idea that non-Gaussian freshwater fluctuations could arise from rare, sudden bursts of meltwater discharge into the North Atlantic, e.g. from the Greenland ice sheet or sea ice.

Here we consider symmetric freshwater forcing ($\bar{F}_1 = \bar{F}_3 = 10.5 \times 10^{-11}\,\mathrm{g\,kg^{-1}\,s^{-1}}$) and a noise process with[5] $\alpha = 1$, skewness parameter $\beta = 0$ (symmetric distribution) and $\sigma = 5 \times 10^{-12}$. At $\boldsymbol{\alpha = 1}$ the forcing is highly non-Gaussian and the variability of the noise-driven system is characterized by large jumps in the North Atlantic salinity S_1 (Fig. 3.9).

The transition paths look qualitatively different compared to the Gaussian case (Fig. 3.10). While Gaussian noise-induced transitions are realized by a concatenation of incremental kicks in the right direction, the extremes of the Lévy distribution make it possible to jump to the competing basin of attraction at a single time step. Since we are only forcing the northern box, we see that these jumps displace the system in the S_1 direction, just across the basin boundary. Consequently, we do not expect that Lévy noise-induced transition paths would travel via the saddle. In temperature space, however, we see that transition paths still undergo an excursion towards the M state after they have been kicked across the basin boundary.

[5]Note that this α refers to the α-stable Lévy process and is different from the thermal expansion coefficient α in Table 3.1.

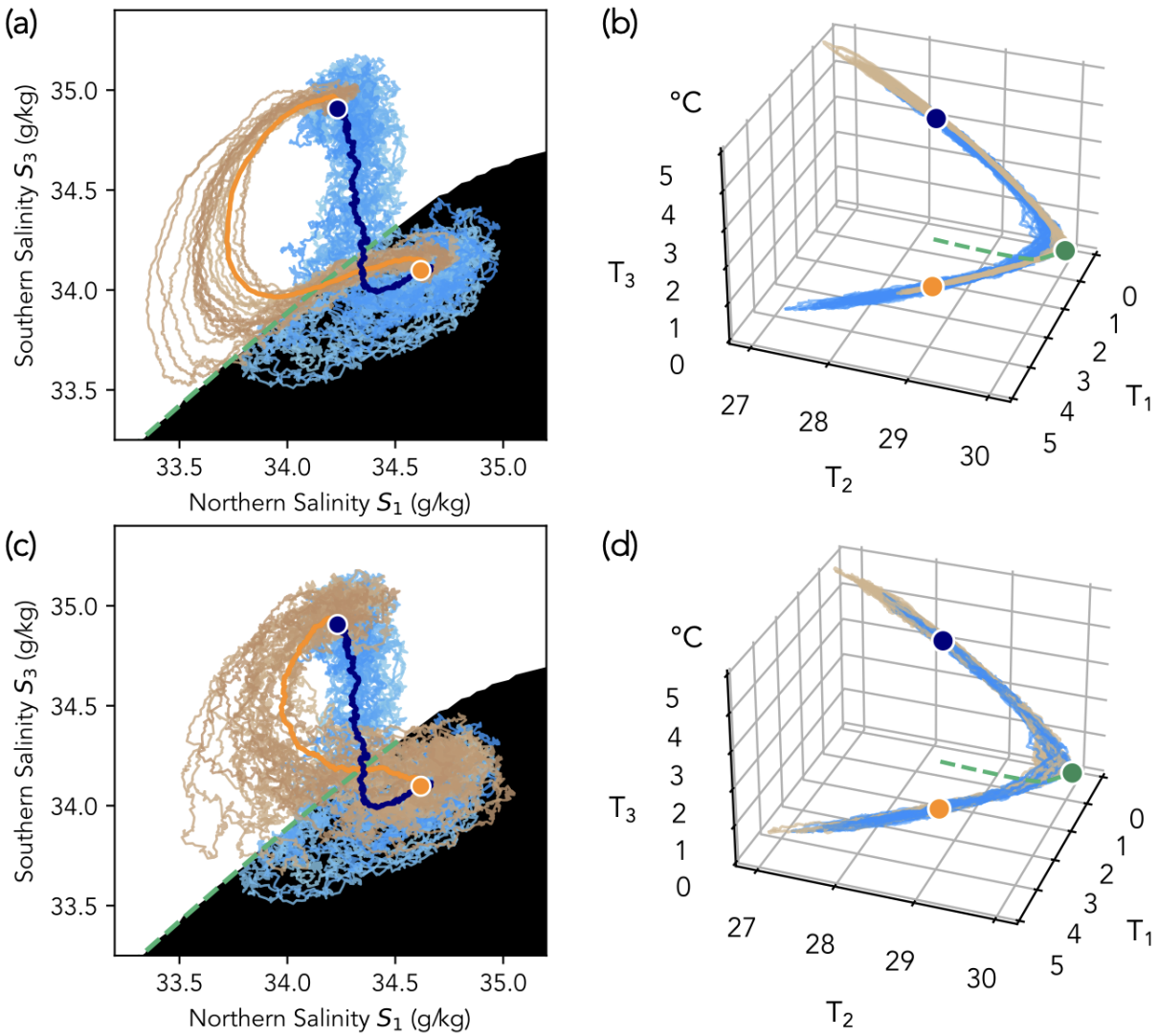

Figure 3.8 | **Sample transition paths** under isotropic Gaussian noise for the scenarios ON $\rightarrow$ OFF (orange) and OFF $\rightarrow$ ON (blue), projected onto salinity space (left) and temperature space (right). (a)-(b): Noise strengths $\sigma_{ON\rightarrow OFF} = 0.0045$ and $\sigma_{OFF\rightarrow ON} = 0.0112$ are chosen to yield a similar rareness in each scenario. (c)-(d): Both scenarios are sampled with $\sigma = 0.0112$. The ON (orange), OFF (blue), and M (green) states are shown as points, along with the edge trajectory along the basin boundary (green). In (a) and (c), the basins of attraction (black/white) are projected onto the plane containing $\boldsymbol{x}_{\text{ON}}, \boldsymbol{x}_{\text{OFF}}$ and $\boldsymbol{x}_M$.

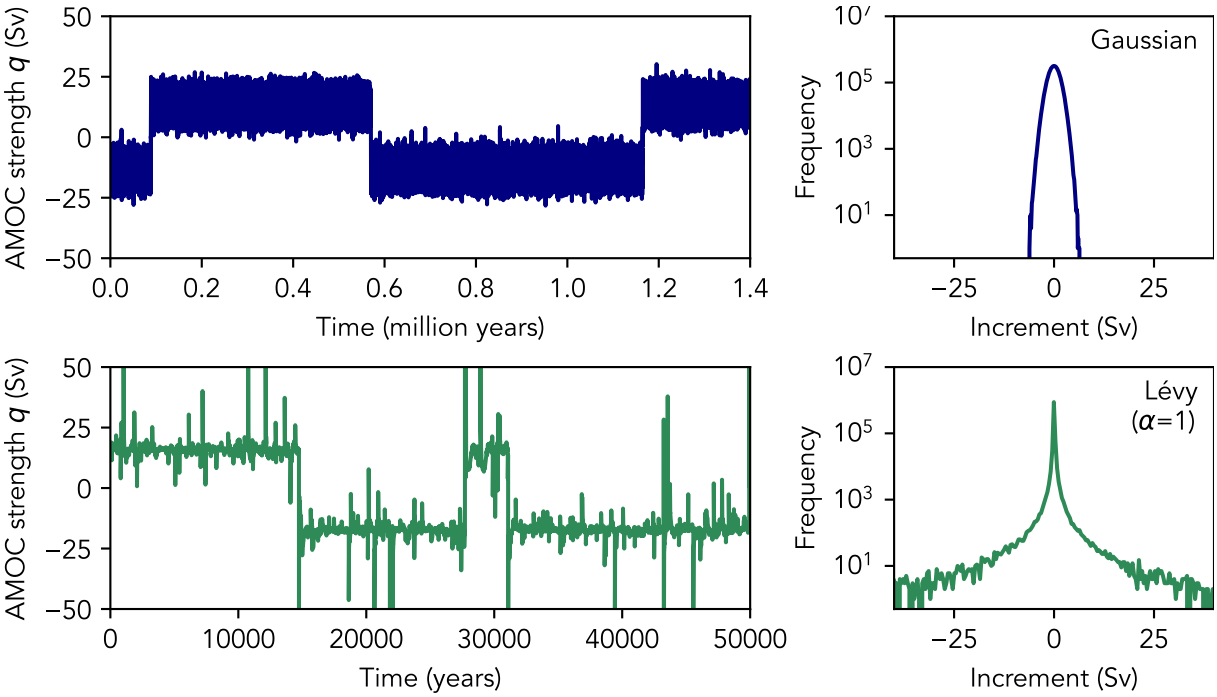

Figure 3.9 | **Noise-induced transitions under Gaussian vs. Lévy noise.** Top: Isotropic Gaussian noise, showing the AMOC time series (left) and distribution of increments across the time series (right). Bottom: Degenerate symmetric α-stable Lévy noise with $\alpha = 1$.

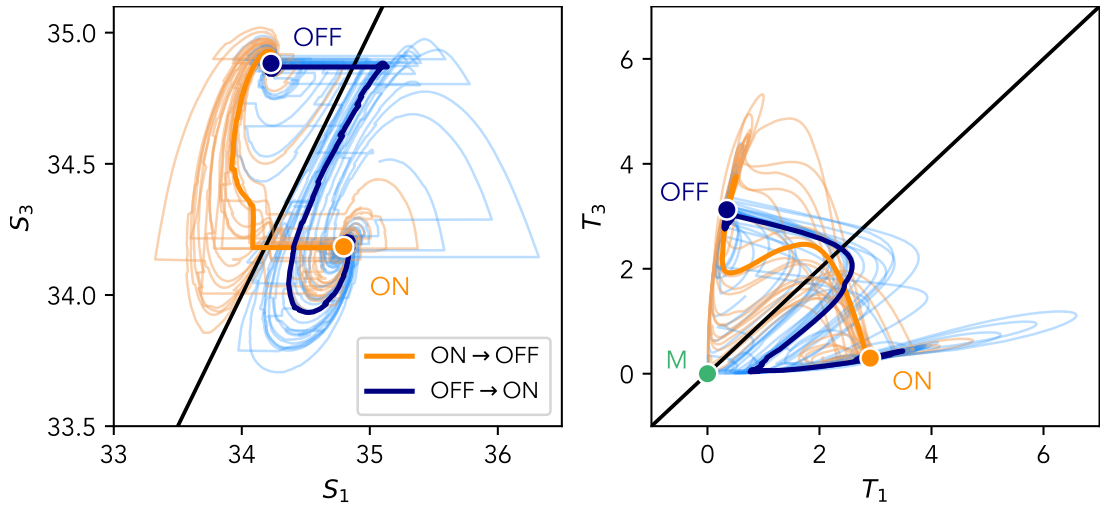

Figure 3.10 | **Lévy noise-induced transition paths** for the scenarios ON → OFF (orange) and OFF → ON (blue), projected onto salinity space (left) and the temperature space of the high-latitude boxes (right). Thick solid lines indicate the approximate MTP. The basin boundary is marked (black line).

3.4 Discussion & conclusion

In this chapter, we used Rooth's interhemispheric box model (Rooth 1982) as a toy model of the thermohaline circulation to exemplify some of the key concepts and methods underlying this thesis. The stability of the model has been studied in detail by Rooth (1982), Scott et al. (1999) and Lucarini and Stone (2005), so much of the dynamics described here were already known. However, less attention has been paid to the unstable equilibria of the model. As we showed, the only saddle point of the system is located at infinite salinity in the bistable regime without limit cycles. Based on the properties of Lyapunov functions, we argued that this implies the lack of a finite quasipotential minimum on the basin boundary that would act as a "mountain pass" for noise-induced transitions in the weak-noise limit.

To resolve this issue, we modified the model by adding a salinity relaxation term that may crudely be interpreted as horizontal turbulent diffusion of salt between the boxes. The bifurcation diagram of this modified model is qualitatively identical to that of the uncoupled model (Scott et al. 1999), but the phase space location of the Melancholia state depends strongly on the salinity restoration timescale. This highlights that one-dimensional projections of the system state can mask significant changes in the stability landscape of the system.

In the Rooth model, the bistable regime of the AMOC is bounded by the two subcritical Hopf bifurcations. This contrasts with the case of other box models, e.g. those by Stommel (1961) and Cimatoribus et al. (2014), where saddle-node bifurcations restrict the bistable regime. On the other hand, a Hopf bifurcation is found in the box model studied by Alkhayuon et al. (2019). There, the Hopf bifurcation is very close to the subsequent saddle-node bifurcation, whereas in the Rooth model the saddle-node bifurcation (at which the ON state disappears) lies at an unphysically high asymmetry in the high-latitude freshwater flux. Overall, the discussion about saddle-node vs. Hopf bifurcation is relevant also for the interpretation of AMOC behavior in general circulation models, as well as the applicability of early-warning indicators (Ditlevsen and Ditlevsen 2023; Ben-Yami, Morr, et al. 2024). It may thus be helpful to clarify how the design of conceptual ocean models impacts their bifurcation structure, and whether differences can be physically constrained. While the Stommel model (Stommel 1961) serves as a two-dimensional

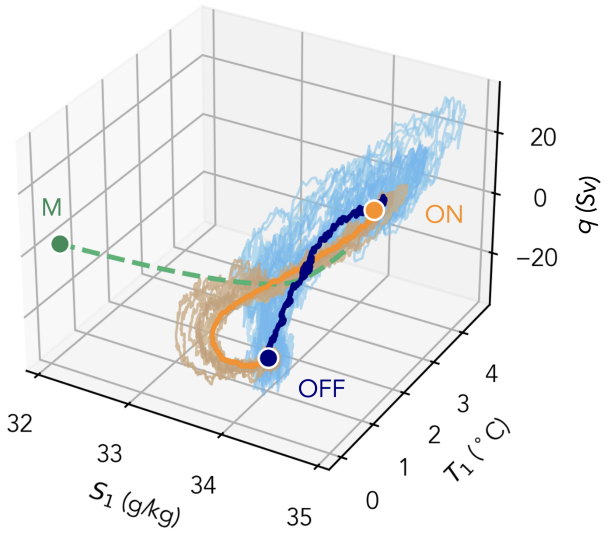

Figure 3.11 | **Saddle avoidance in the Rooth model** under weak isotropic Gaussian noise, depictted in the reduced phase space of northen temperature T_1, northern salinity S_3, and AMOC strength q. The green dashed line indicates the edge track from the ON and OFF states to the M state. Orange (blue) sample transition paths lead from ON → OFF (OFF → ON).

paradigm of a bistable ocean circulation, the more elaborate box models by Wood et al. (2019) and Cimatoribus et al. (2014) are arguably more realistic and thus preferable over the Rooth model.

Adding weak Gaussian noise to the model, we observed that noise-induced transition paths do not travel via the saddle in all degrees of freedom. This result was surprising because the chosen noise strength was so weak that transitions occurred only once every three million years, on average, qualifying them as rare events. Based on Freidlin-Wentzell theory (see section 2.1.6, we would naïvely expect that saddles act as gateways of noise-induced transitions, which holds in the weak-noise limit. Here, while transition paths travel via the M state in temperature space (fast relaxation time), they avoid it in salinity space (slow relaxation time). Chapter 4 is dedicated to explaining this phenomenon, which we refer to as *saddle avoidance*.

The appearance of saddle avoidance has important implications for the interpretation of the dynamics in higher-dimensional systems where the relevant Melancholia state has not been (or cannot be) computed. For example, Margazoglou et al. (2021), studying noise-induced transitions between "Snowball Earth" and the coexisting warm state, found that the MTP in one transition direction does not intersect with the MTP in the opposite direction. Based on this, they concluded that multiple M states must coexist on the basin boundary. Knowing about saddle avoidance, however, it is also possible that only one M state exists but the transitions avoid the M state. The converse inference can also be problematic: consider Fig. 3.11. In the projection shown, it appears that both MTPs intersect near the basin boundary. Without knowledge of the M state, we might conclude that the M state is located at the point where the MTPs meet in between the two attractors. Here, this would be a false conclusion.

Lastly, we touched only briefly upon the interesting case of α-stable Lévy noise. As argued here and discussed by Lucarini et al. (2022) and Del Amo and Ditlevsen (2025), critical transitions induced by Lévy noise (with α sufficiently different from 2) are typically realized by a single extreme event in the stochastic process, kicking the system across the basin boundary. The case of Lévy noise is thus closely related to the minimal fatal shock problem. Consequently, the transition probability under Lévy noise will depend crucially on the direction of the noise in phase space, as the distance to the basin boundary in that direction determines the escape rate. In the regime where the unstable limit cycle constitutes the M state, this would impliy that, for symmetric Lévy noise, the transition statistics will depend on two possibly asymmetric escape rates corresponding to the sign of the direction of the perturbation. This hypothesis could be explored in future work.

To conclude, focusing on the unstable invariant sets of the Rooth model showed that transient behavior with a high sensitivity to the initial condition can arise near the basin boundary. For example, the minimal fatal shock experiment highlights that the asymptotic state of the system following a critical perturbation can be effectively unpredictable for a long time. In this simple non-chaotic system, one needs to get very close to the basin boundary to observe such behavior. But as we will see (chapters 5 and 6), the phase space region with complex transient behavior can be much larger in more complex systems.

Chapter 4

Saddle avoidance in multiscale stochastic systems

Defying the weak-noise limit

4.1 Multiscale stochastic systems . . . 82
4.1.1 Two competing limits . . . 82
4.1.2 Defining transitions and weak noise . . . 82
4.2 Additive noise: FitzHugh-Nagumo model . . . 83
4.2.1 Model description . . . 83
4.2.2 Sample transition paths and instantons . . . 84
4.2.3 Quasipotentials . . . 86
4.2.4 Finite-noise prediction of transition paths . . . 88
4.2.5 First-exit statistics . . . 90
4.3 Multiplicative noise: Two competing species . . . 94
4.3.1 Model description . . . 94
4.3.2 Global stability and quasipotential landscape . . . 95
4.3.3 Sample transition paths and instantons . . . 96
4.3.4 Onsager-Machlup action minimizer . . . 98
4.4 Reconciling the fast-slow and weak-noise limits . . . 99
4.5 Discussion & conclusion . . . 102

This chapter is based on the study published by Börner et al. (2024b).

Introduction

A surprising finding of the previous chapter was that noise-induced transition paths do not travel via the saddle in the ocean box model we studied – even though the noise was so weak that transitions occurred as one-in-a-million-years events. We speculated that this could be related to the temperature and salinity dynamics evolving on different timescales. To understand the cause of *saddle avoidance*, let us land ashore and leave the ocean for this chapter. Our aim is to replicate saddle avoidance in simple multistable systems featuring what we hypothesize to be key ingredients: nongradient dynamics and multiple timescales. We need at least two dimensions for this. Choosing 2D systems allows us to get a full picture of their phase plane and explicitly compute quasipotentials. Nonetheless, the toy models we study here may be viewed as minimal models that illustrate a more general phenomenon of potential relevance in various fields.

Noise-induced tipping (N-tipping, see section 2.1.7) occurs in many real-world systems (Horsthemke and Lefever 1984), from brain activity (Brinkman et al. 2022), gene regulation (Bressloff 2017) and human culture (Klemm et al. 2003) to lasers (Masoller 2002), planetary atmospheres (Bouchet, Rolland, and Simonnet 2019) and the earth system (Boers et al. 2022; Ditlevsen and Johnsen 2010). Transitions often represent high-impact low-probability events, such as financial crashes (Bouchaud and Cont 1998) and ecosystem collapse (Bashkirtseva and Ryashko 2011), besides climate tipping points (Castellana et al. 2019; Cini et al. 2024). Understanding noise-induced transitions is thus crucial to assess a system's stability and resilience (Holling 1973). Particularly, predicting the *path* of transitions in phase space can offer insights into the transition mechanism.

The common belief that saddles act as gateways of noise-induced transitions is supported by the large deviation principle formulated by Freidlin and Wentzell (1998), as outlined in section 2.1.6. Since the FW quasipotential offers a nongradient analogue to the potential in gradient systems, it is tempting to interpret the quasipotential as a stability landscape with the common properties of a potential. However, the analogy is limited, since the deterministic dynamics of nongradient systems is governed not only by the gradient of the quasipotential but also the additional orthogonal term (see Eq. (2.21)). This can give rise to surprising behavior that clashes with the expectation derived in the weak-noise limit, even for weak finite noise.

In reality, the weak-noise limit is never attained. Various counter-examples attest that if the noise is weak yet non-infinitesimal, noise-induced transitions do not necessarily pass near a saddle (Maier and Stein 1993; Luchinsky et al. 1999; Maier and Stein 1997; Berezhkovskii et al. 1989; Northrup and McCammon 1983; Agmon and Kosloff 1987; Schäfer et al. 2017). The question is: *how* weak must the noise be for results from FW theory to give an adequate approximation? Under which conditions do finite-noise effects become important, and how can we predict the behavior in that case?

This chapter is dedicated to answering these questions. We focus on the role of multiscale dynamics, combining the theory of fast-slow systems with the intuition of the quasipotential landscape to understand saddle avoidance. Furthermore, we extend the least-action variational principle of FW theory to aptly predict transition paths under finite noise. Illustrated by two simple examples, our results offer a deeper understanding of saddle avoidance from a large deviation theoretical perspective.

4.1 Multiscale stochastic systems

The general form of dynamical systems we consider in this chapter is the Itô stochastic differential equation introduced in Eq. (2.13),

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{x})\,\mathrm{d}t + \sigma\boldsymbol{\Sigma}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{W}_t\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0\,,\ t \geq 0\ . \tag{4.1}$$

We are interested in the case where the deterministic drift $\boldsymbol{b}$ is nongradient, multistable and *multiscale*. By multiscale we mean that the degrees of freedom of the system vary on different timescales, where the degree of timescale separation is quantified by a parameter ε, such that $\boldsymbol{b} \equiv \boldsymbol{b}(\boldsymbol{x}, \varepsilon)$. We will restrict our discussion to systems where the attractors and unstable equilibria are fixed points.

The noise strength σ shall be small in the sense that noise-induced transitions between the attractors qualify as rare events. Recall that we assume the covariance matrix $\boldsymbol{Q}(\boldsymbol{x}) = \boldsymbol{\Sigma}(\boldsymbol{x})\boldsymbol{\Sigma}(\boldsymbol{x})^\top$ to be invertible (non-degenerate noise). We will study additive noise in section 4.2.5 and multiplicative noise in section 4.3.4.

4.1.1 Two competing limits

If we now add weak stochastic forcing to the fast-slow system, the resulting stochastic system is controlled by two small parameters: the timescale parameter ε and the noise strength σ. The asymptotic theory of fast-slow systems deals with the limit $\varepsilon \to 0$, whereas FW theory applies to the weak-noise limit $\sigma \to 0$. We are thus facing a double-limit problem (Kuehn et al. 2022), and it is not a priori clear which limiting behavior dominates for finite ε and σ.

4.1.2 Defining transitions and weak noise

Before we start playing with our toy models, we should specify under which conditions we classify a trajectory as a transition, and what we mean by weak noise. Suppose the system (Eq. (4.1)) has two stable equilibria $\boldsymbol{x}_R$, $\boldsymbol{x}_L$ with basins of attraction $\mathcal{B}_R$, $\mathcal{B}_L \subset \mathbb{R}^D$, separated by the basin boundary $\partial\mathcal{B}$.

Term 4.1. **Transition path**

A *transition path* $\boldsymbol{\phi}_t^{RL}$ is a trajectory that, after exiting a small neighborhood R around $\boldsymbol{x}_R$, enters a small neighborhood L around $\boldsymbol{x}_L$ without re-entering R.

Two distinct timescales characterize the transition $\boldsymbol{\phi}_t^{RL}$. The *mean first-exit time* $\langle \tau_\sigma^{RL} \rangle$, i.e. the expected waiting time until a trajectory initialized at $\boldsymbol{x}_R$ leaves $\mathcal{B}_R$, and the *mean transition time* $\langle t_\sigma^{RL} \rangle$, i.e. the average time $\boldsymbol{\phi}_t^{RL}$ takes to travel to L after last leaving R. Large values of the *rareness*,

$$r_\sigma^{RL} := \frac{\langle \tau_\sigma^{RL} \rangle}{\langle t_\sigma^{RL} \rangle}\,, \tag{4.2}$$

indicate that individual transitions are *rare* and therefore occur as a memory-less Poisson process.

In what follows we choose σ sufficiently small such that $r_\sigma^{RL} \gg 1$. Hence, by *weak* noise we mean that the noise is of a magnitude where transitions qualify as rare events and one might therefore expect FW theory to apply.

4.2 Additive noise: FitzHugh-Nagumo model

As a paradigmatic example, let us consider the FitzHugh-Nagumo (FHN) model. It was originally conceived to describe a spiking neuron (FitzHugh 1961; Nagumo et al. 1962) and has been widely studied as a conceptual multiscale model in theoretical neuroscience and beyond.

4.2.1 Model description

The FHN model can be written as

$$\dot{\boldsymbol{x}} = \boldsymbol{b}_{\mathrm{FHN}}(u, v) = \begin{pmatrix} \varepsilon^{-1}(-u^3 + u - v) \\ -\beta v + u \end{pmatrix}, \tag{4.3}$$

where $\boldsymbol{x} = (u, v) \in \mathbb{R}^2$ is the two-dimensional state vector and $\varepsilon > 0$ denotes the timescale parameter. For $\varepsilon \ll 1$, the variable u is fast, while v is slow. We set $\beta = 3$, in which case the FHN model is bistable. The two stable fixed points are located at $\boldsymbol{x}_{R,L} = \pm(\sqrt{2/3}, \sqrt{2/3^3})$ and the model has a saddle point at $\boldsymbol{x}_M = (0, 0)$. Due to the nonlinearity in the u equation, the model is nongradient.

In the limit $\varepsilon \to 0$, the critical manifold is given by

$$C_0 := \left\{(u, v) \in \mathbb{R}^2 : v = -u^3 + u\right\},$$

corresponding to the u-nullcline. The cubic form of C_0 yields a fast subsystem with two saddle-node bifurcations with respect to v. Two stable branches of C_0 are separated by an unstable branch between the fold points $v_f^{\pm} = \pm\sqrt{4/3^3}$. We will also investigate the inverse timescale separation ($1/\varepsilon \to 0$) where u is slow and v is fast. In that case, the critical manifold is given by the v-nullcline,

$$C_0' = \{(u, v) \in \mathbb{R}^2 : \beta v = u\}.$$

Along this line the fast dynamics are stable everywhere.

We wish to study the transition behavior of the FHN model driven by additive Gaussian white noise. Thus, we consider Eq. (4.1) with $\boldsymbol{b} = \boldsymbol{b}_{\mathrm{FHN}}$ and identity covariance matrix $\boldsymbol{Q} = \boldsymbol{I}_2$, giving the stochastic FHN model,

$$\begin{aligned} \mathrm{d}u &= \frac{1}{\varepsilon}\big(-u^3 + u - v\big)\mathrm{d}t + \sigma \mathrm{d}W_{u,t} \\ \mathrm{d}v &= \big(-\beta v + u\big)\mathrm{d}t + \sigma \mathrm{d}W_{v,t} \end{aligned} \tag{4.4}$$

with $\beta = 3$ and free parameters ε and σ.

Problem setup. We study Eq. (4.4) for four different values of timescale separation, namely $\varepsilon \in \{0.01, 0.1, 1, 10\}$. For each ε-case, we fix $\sigma(\varepsilon)$ to maintain a constant rareness $r_\sigma^{RL} \approx 10^5$.

Note that the FHN model is point symmetric with respect to the origin, i.e. $\boldsymbol{b}_{\mathrm{FHN}}(\boldsymbol{x}) = -\boldsymbol{b}_{\mathrm{FHN}}(-\boldsymbol{x})$. This implies that the transition behavior from $\boldsymbol{x}_L$ to $\boldsymbol{x}_R$ will be a mirror image of the transition behavior from $\boldsymbol{x}_R$ to $\boldsymbol{x}_L$, and it suffices to study one direction[1].

4.2.2 Sample transition paths and instantons

Methodology. The first step is to generate ensembles of sample transition paths from $\boldsymbol{x}_R$ to $\boldsymbol{x}_L$ by numerical simulation of Eq. (4.4) for each ε-case. We do this via Monte Carlo rejection sampling: we initialize a large number of trajectories at $\boldsymbol{x}_R$ and collect those that reached the neighborhood L within time $t_{\max} = 10^8$ until we have at least 100 samples. The neighborhoods are defined as circles centered at R and L with radii as given in Table 4.1. We then save the portion of each trajectory between last leaving the neighborhood around the start point and first entering the neighborhood around the end point. Equation (4.4) is integrated using the Euler-Maruyama scheme (Eq. (2.35)) as in the previous chapter; the integration time step is detailed in Table 4.1 along with the noise strength σ chosen for each ε. The rareness r_σ is estimated empirically from the sampled ensemble. Our efficient code to perform rejection sampling as described is publicly available as the `transitions` function of the CriticalTransitions.jl package[2].

Using the geometric minimum action method (gMAM, see section 2.2.3 and Heymann and Vanden-Eijnden (2008b)), we compute the instanton $\boldsymbol{\varphi}^{*RL}$, the minimizer of the FW action functional (Eq. (2.18)), for each ε-case. This method is available in CriticalTransitions.jl as the `geometric_min_action_method` function. The path $\boldsymbol{\varphi}$ over which the geometric action is minimized is initialized as a straight line with 265 equally spaced points between the fixed start point $\boldsymbol{x}_R$ and the end point. As end points, we choose the final attractors $\boldsymbol{x}_L$ except for $\varepsilon = 0.01$, where convergence to the instanton is very slow; in this case, we construct the instanton by concatenating the FW minimizer from $\boldsymbol{x}_R$ to the saddle with the heteroclinic orbit leading from the saddle to $\boldsymbol{x}_L$. As we confirm in the next section, the saddle indeed marks the global minimum of the quasipotential along the basin boundary, which justifies this procedure.

Results. Phase space plots of the sample transition paths and corresponding instantons are shown in Fig. 4.1 for the different cases $\varepsilon \in \{0.01, 0.1, 1, 10\}$. The ensembles of transition paths concentrate within a tube around a *mean sample transition path* (MTP), which we compute by spatial averaging over the ensemble (Fig. 4.1a). Specifically, we interpolate the sample transition paths in the u direction and define the MTP as the ensemble mean in the v direction as a function of u.

The computed instantons always pass through the saddle. They approach it at an angle relative to $\partial\mathcal{B}$ that narrows as ε decreases, becoming tangential to $\partial\mathcal{B}$ when $\varepsilon < 1/\beta$ (Maier and Stein 1997) (Fig. 4.1a). For $\varepsilon \in \{1, 10\}$, the MTP closely matches the

[1] This does not contradict the fact that the FHN model is nongradient, i.e. detailed balance is violated, and instantons from a point with lower quasipotential to one with higher quasipotential generally do not coincide with the relaxation path in the opposite direction.

[2] `https://github.com/JuliaDynamics/CriticalTransitions.jl`. We used version `v0.2` but recommend the latest version which features numerous improvements.

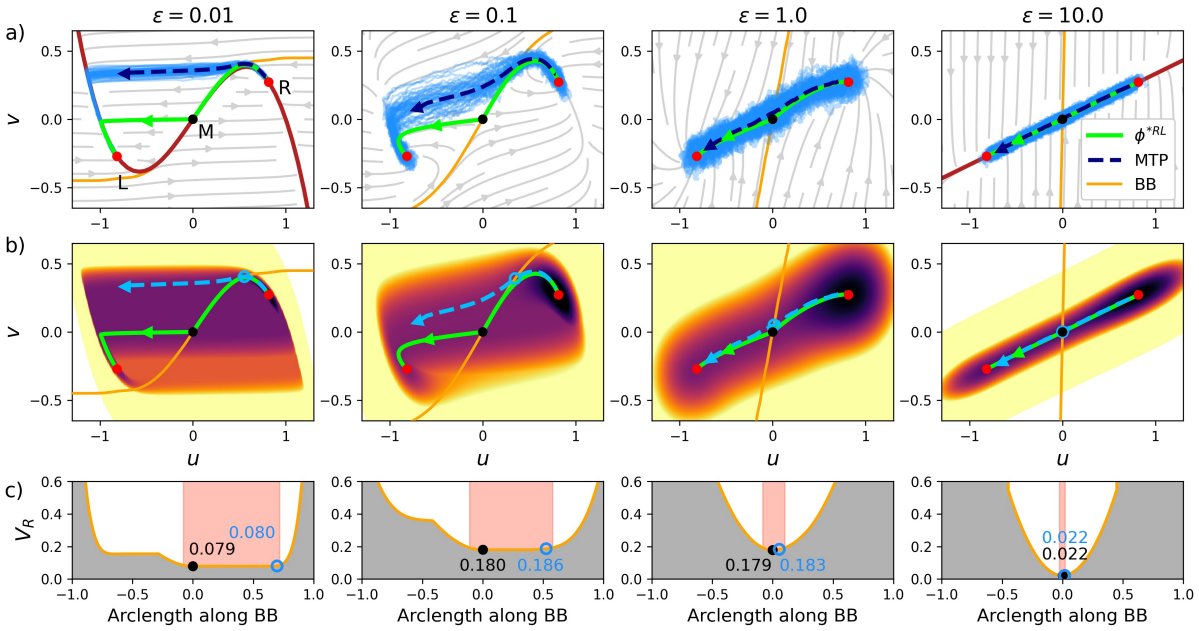

Figure 4.1 | **Transition paths and quasipotential landscape of the FHN model** for $\varepsilon = (0.01, 0.1, 1, 10)$ at $\sigma(\varepsilon) \approx (0.08, 0.12, 0.12, 0.04)$. (a) Phase space with equilibria $\boldsymbol{x}_{L,R}$ (red), the saddle $\boldsymbol{x}_M$ (black), the instanton $\boldsymbol{\varphi}^{*RL}$ (green), 50 sample transition paths (light blue), and the MTP (blue dashed). The drift field $\boldsymbol{b}$ (gray flow lines) and basin boundary (BB, orange) are indicated, along with the critical manifold C_0 (red, left-most panel) and C_0' (right-most panel). (b) Like (a) but with the FW quasipotential $V_R(\boldsymbol{x})$ plotted on a logarithmic colormap (black: $V_R = 0$; brighter means larger V_R), showing $\boldsymbol{x}_c$ (blue circle) where the MTP crosses the basin boundary. (c) Quasipotential $V_R(\boldsymbol{x})$ along the basin boundary, indicating values at $\boldsymbol{x}_M$ (black) and $\boldsymbol{x}_c$ (blue), and the region $\mathcal{Z}$ (red shading). Figure reproduced from (Börner et al. 2024b).

instanton and all sample paths cross the basin boundary within $\sim \sigma$ distance on either side of the saddle. Contrarily, for $\varepsilon \in \{0.01, 0.1\}$ sample transitions avoid the saddle: the MTP diverges from the instanton after getting close to the basin boundary, which happens before reaching the saddle. Once the noise kicks the trajectory into the competing basin, it is repelled from the basin boundary stronger than it is attracted towards the saddle.

This multiple timescale effect manifests itself in the ratio of the negative and positive eigenvalues $\lambda_\mp$ of the Jacobian of $\boldsymbol{b}$ at the saddle, $\mu := |\lambda_-|/\lambda_+$. If $\mu < 1$, Maier and Stein (1997) showed that sample transition paths avoid the saddle on an extended lengthscale $\mathcal{O}(\sigma^\mu)$ (Bobrovsky and Zeitouni 1992; Day 1994): their exit locations (where they cross the basin boundary) follow a one-sided Weibull distribution (Coles 2001), whose mode approaches the saddle only logarithmically as $\sigma \to 0$ (Maier and Stein 1997; Luchinsky et al. 1999). This contrasts with the case $\mu > 1$ where the distribution of exit locations is centered around the saddle on a lengthscale $\mathcal{O}(\sigma)$ as $\sigma \to 0$ (Maier and Stein 1997). In the FHN model, $\mu < 1$ if and only if $0 < \varepsilon < 1/\beta$, and we have $\mu = (\beta - 1)\varepsilon + \mathcal{O}(\varepsilon^2)$ as $\varepsilon \to 0$, which highlights the relationship between saddle avoidance and timescale separation.

However, there is no abrupt change in behavior at $\mu = 1$, as we found by performing additional simulations (via pathspace sampling, see section 4.2.4) for various ε-values at fixed r_σ (Fig. 4.2). The amount of saddle avoidance as a function of ε increases smoothly with decreasing ε in an exponential fashion. Here we measure saddle avoidance as the arclength along the basin boundary between the saddle and the mode of the exit point

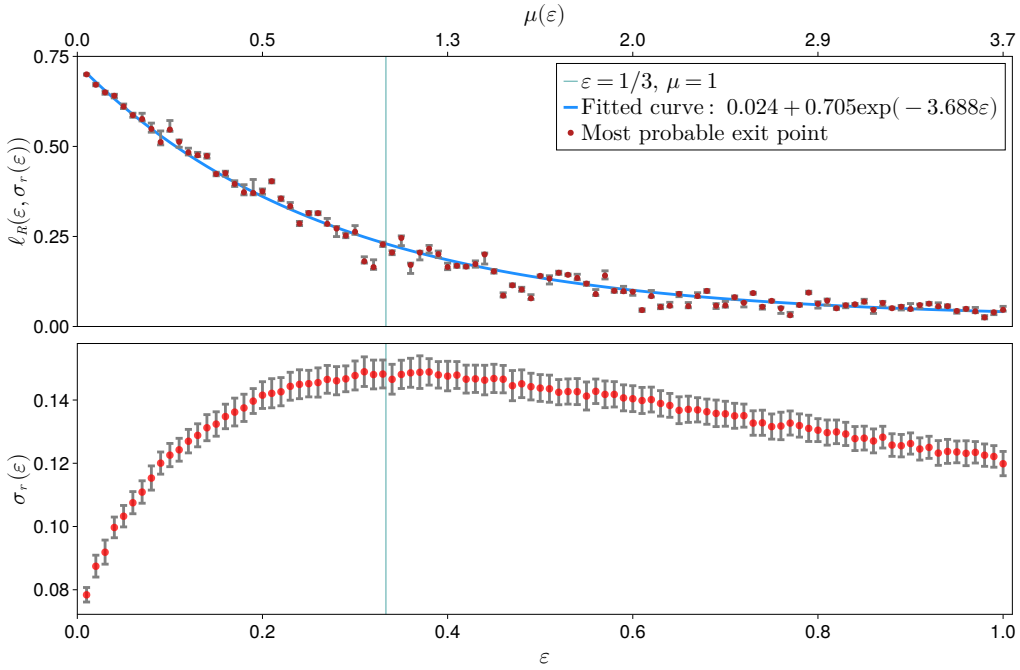

Figure 4.2 | **Saddle avoidance and rareness as a function of ε in the FHN model.** Top: Empirical estimates (error bars indicate the 0.05-0.95 interquantile range from bootstrapping) of the distance ℓ_R (in arclength along the basin boundary) between the saddle $\boldsymbol{x}_M$ and the most probable exit point of transition paths, given by the mode of the exit point distribution for rareness $r_\sigma = 10^5$. The vertical line indicates $\varepsilon = 1/3$ where $\mu = 1$. Bottom: Noise strength σ_r at which transitions occur with rareness $r_\sigma = 10^5$ as a function of ε, estimated by extrapolating a weighted least-squares fit of (σ, r_σ) computed from Monte Carlo rejection sampling ensembles (error bars show ∼90% confidence intervals). Note the non-monotonic dependence of σ_r on ε. Figure taken from Börner et al. (2024b), Supplemental Material, produced by Ryan Deeley.

distribution of transition paths (i.e. the distribution of where transition paths cross the basin boundary).

The cases $\varepsilon = 0.01$ and $\varepsilon = 10$ exemplify the behavior we would anticipate based on fast-slow systems theory for $\varepsilon \to 0$ and $1/\varepsilon \to 0$, respectively. Recall that for $\varepsilon = 0.01$, the critical manifold C_0 has a bistable regime with two saddle-node bifurcations with respect to v. The MTP escapes from the basin boundary near the bifurcation point v_f^+ where C_0 becomes unstable (Fig. 4.1a). After leaving the right branch of C_0, the sample transition paths rapidly shoot to the left branch of C_0, where they are attracted and slowly move towards $\boldsymbol{x}_L$ due to deterministic relaxation. In the inverse limit $1/\varepsilon \to 0$, where $\mu \gg 1$ and the corresponding critical manifold is given by C_0', the limiting behavior of both σ and $1/\varepsilon$ aligns by forcing the instanton and sample transition paths onto C_0', as we observe for $\varepsilon = 10$.

4.2.3 Quasipotentials

Can we understand the observed transition behavior by means of the FW quasipotential V_R (Eq. (2.20)), which measures the "difficulty" of reaching a point $\boldsymbol{x}$ from $\boldsymbol{x}_R$? As $\sigma \to 0$, the probability of passing through the global minimum $\boldsymbol{x}^*$ of V_R along the basin boundary $\partial\mathcal{B}$ approaches 1 (Freidlin and Wentzell 1998). For finite σ, however, the large

	FHN				COMP ($AB \to A$)		COMP ($AB \to B$)	
ε	0.01	0.1	1	10	0.01	1	0.01	1
σ	0.07842	0.11907	0.12000	0.04189	0.006	0.0425	0.006	0.0425
r_σ	$\approx 10^5$	$\approx 10^5$	$\approx 10^5$	$\approx 10^5$	≈ 100	≈ 16000	≈ 1000	≈ 15000
N	100	306	100	100	100	100	100	100
Δt	0.00218	0.00218	0.00218	0.00218	10^{-2}	10^{-2}	10^{-2}	10^{-2}
radius	0.01	0.01	0.01	0.01	0.025	0.025	0.05	0.025

Table 4.1 | Specifications for transition paths sampled via Monte Carlo simulation in the FHN and COMP models. Here N is the ensemble size, Δt the integration time step, and radius refers to the radii of the circular neighborhoods around each attractor.

deviation principle underlying FW theory (Eq. (2.17)) suggests that transitions may cross with similar probability in regions $\mathcal{Z}(\sigma) = \{\boldsymbol{z} \in \partial\mathcal{B} : V_R(\boldsymbol{z}) \leq V_R(\boldsymbol{x}^*) + \sigma^2\}$.

Methodology. We compute V_R for the four ε-cases considered using the OLIM4VAD algorithm, an ordered line integral method for variable and anisotropic diffusion which thus also supports state-dependent covariance matrices (see section 2.2.4 and Dahiya and Cameron (2018a)). We use the C script `OLIM4VAD.c` available on M. Cameron's homepage[3] with the settings `Nx=1024, Ny=1024`, and `TOL=1.0e-12`. For the FHN model, we set `K=22` and `XMIN=-2.0, XMAX=2.0, YMIN=-1.5, YMAX=1.5`, with $\boldsymbol{x}_R = (\sqrt{2/3}, \sqrt{2/27})$ as the reference point.

Results. As per definition, the quasipotentials are zero at $\boldsymbol{x}_R$ and increase monotonically along the instanton to $\boldsymbol{x}_L$. For $\varepsilon = 10$, a steep and narrow trench of low quasipotential connects $\boldsymbol{x}_R$ with $\boldsymbol{x}_L$ (Fig. 4.1b). As ε decreases, however, V_R increasingly flattens along the v-direction, leading to an extended *quasipotential plateau* around the saddle $\boldsymbol{x}_M$ (Fig. 4.3). Along $\partial\mathcal{B}$, V_R indeed assumes a global minimum at $\boldsymbol{x}_M$ for all ε, but its curvature around $\boldsymbol{x}_M$ becomes small for $\varepsilon \ll 1$ (Fig. 4.1c). The low curvature implies the lack of a clear "mountain pass" and a widened "danger zone" $\mathcal{Z}(\sigma)$ around $\boldsymbol{x}_M$ (red shading in Fig. 4.1c), whose arclength converges more slowly to 0 as $\sigma \to 0$. This large deviation theoretical argument aligns with the occurrence of saddle avoidance on the extended lengthscale $\mathcal{O}(\sigma^\mu)$ (see above and Maier and Stein (1997)). Indeed, the MTP crosses the basin boundary at the outer end of the danger zone. This is likely related to the instanton approaching the basin boundary tangentially (see section 4.2.5).

Note that the quasipotential height at the saddle is higher in the intermediate ε-cases $\varepsilon \in \{0.1, 1.0\}$ compared to the extreme ones ($\varepsilon \in \{0.01, 10\}$). This aligns with the finding that a larger noise strength σ_r is needed at intermediate ε-values to maintain a constant rareness r_σ of transitions (Fig. 4.2).

Another interesting feature is the asymmetry of the quasipotential along the basin boundary with respect to the saddle, which highlights the nongradient nature of the FHN model especially for $\varepsilon < 1/\beta$. In the cases $\varepsilon \in \{0.01, 0.1\}$, a discontinuity in the slope of the quasipotential is apparent (Fig. 4.1c). This kink occurs at a point $\boldsymbol{x}_K$ where the quasipotential is twice the height at the saddle, $V_R(\boldsymbol{x}_K) = 2V_R(\boldsymbol{x}_M)$. Between $\boldsymbol{x}_M$ and $\boldsymbol{x}_K$, the most efficient (and thus most probable) way to reach the basin boundary in the weak-noise limit follows an instantonic path that climbs up the quasipotential within the basin of attraction $\mathcal{B}_R$ of $\boldsymbol{x}_R$. By contrast, points on the basin boundary beyond $\boldsymbol{x}_K$

[3] https://www.math.umd.edu/~mariakc/OLIM.html, visited on 19 March 2024.

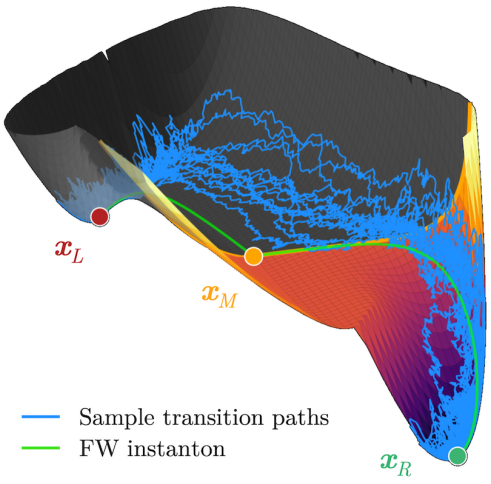

Figure 4.3 | **3D visualization of saddle avoidance in the quasipotential landscape** of the FHN model for $\varepsilon = 0.1$. Sample transition paths (blue), the instanton ϕ^{*RL} (green), and the fixed points are shown. The quasipotential is colored as in Fig. 4.1b) up to the basin boundary (orange line).

are most efficiently reached by first transitioning from $\boldsymbol{x}_R$ to $\boldsymbol{x}_L$ and then climbing up to the basin boundary within the basin of attraction $\mathcal{B}_L$ of $\boldsymbol{x}_L$ (again, in the weak-noise limit). Hereby the "cost" of the transition from $\boldsymbol{x}_R$ to $\boldsymbol{x}_L$ is given by $V_R(\boldsymbol{x}_M)$. This highlights that the most probable transition path from an attractor to a point in its basin of attraction can sometimes lead through competing basins of attraction.

4.2.4 Finite-noise prediction of transition paths

While the quasipotential can explain the occurrence of saddle avoidance, the question remains how to predict the location distribution of transition paths for finite noise. If the noise is additive, there is increasing confidence that the appropriate variational problem entails minimizing the Onsager-Machlup (OM) action (Eq. (2.38)), which adds a σ-dependent correction term to the FW action (see section 2.2.2) (Stratonovich 1971; Horsthemke and Bach 1975; Dürr and Bach 1978; Pinski and Stuart 2010; Gladrow et al. 2021; Thorneywork et al. 2024).

Based on this, we propose that minimizing the OM action allows us to derive a candidate for the most probable transition path, also for cases with saddle avoidance. Although the minimization over all $T > 0$ is ill-defined in the OM case (Li and Li 2021), one can often select a characteristic path travel time T. Here we set T to the mean transition time, $T = \langle t_\sigma^{RL} \rangle$, empirically estimated from the transition path ensemble for each ε-case.

Methodology. As described in section 2.2.2, the variational problem of minimizing the OM action can be formulated as an SPDE in pathspace. For the FHN model (Eq. (4.4)), the SPDE in Eq. (2.40) (with $\boldsymbol{x} = (u, v)^\top$ and $\boldsymbol{\eta}(t, z) = (\eta_u(t, z), \eta_v(t, z))^\top$) is specifically given by

$$\begin{aligned}
\frac{\partial u}{\partial z} &= \frac{\partial^2 u}{\partial t^2} - \frac{1}{\varepsilon^2}(-u^3 + u - v)(-3u^2 + 1) + (\beta v - u) + \left(1 + \frac{1}{\varepsilon}\right)\frac{\partial v}{\partial t} + \frac{3\sigma^2}{\varepsilon}u + \sqrt{2}\sigma\eta_u(t, z) \\
\frac{\partial v}{\partial z} &= \frac{\partial^2 v}{\partial t^2} + \frac{1}{\varepsilon^2}(-u^3 + u - v) + (-\beta^2 v + \beta u) - \left(1 + \frac{1}{\varepsilon}\right)\frac{\partial u}{\partial t} + \sqrt{2}\sigma\eta_v(t, z)\,.
\end{aligned} \tag{4.5}$$

Solving this SPDE in virtual time z corresponds to a gradient descent in the OM action plus spatio-temporal white noise. If we fix the boundary conditions

$$\boldsymbol{x}(0, z) = \boldsymbol{x}_R\,, \quad \boldsymbol{x}(T, z) = \boldsymbol{x}_L \quad \text{for all } z \geq 0\,,$$

this allows us to sample transition paths in pathspace once the SPDE has converged to a steady state. We use pathspace sampling in different ways:

Method 1) To compute **transition path densities** by integrating Eq. (4.5) for a time period $z \in [z_i, z_f]$ with $T = \langle t_\sigma^{RL} \rangle$, following a spin-up period.

Method 2) To compute **action minimizers** by solving Eq. (4.5) for a time period $z \in [z_i, z_f]$ with $T = \langle t_\sigma^{RL} \rangle$, while linearly decreasing the magnitude of the additive noise processes η_u, η_v to zero. The noise processes $\eta_{u,v}$ are thus replaced by $\eta'_{u,v}(z,t) = h(z)\eta_{u,v}(x,t)$ where $h(z) = 1 - (z - z_i)/(z_f - z_i)$, such that the primed processes now have decreasing variance over z. This can be thought of as gradient descent with simulated annealing.

For each method, we implement two versions:

Version a) We integrate Eq. (4.5) as is, which corresponds to gradient descent in the the OM action.

Version b) We omit the term proportional to σ^2 in Eq. (4.5), which is the correction term introduced by the OM action. Omitting this term leads to gradient descent in the FW action (Eq. (2.18)). Here we choose T sufficiently large to converge to the minimization problem over all $T > 0$.

Settings used for method 1 are summarized in Table 4.2. The initial path is taken as a linear line between the attractors; to account for the spin-up period, we only save paths across the second half of the virtual time steps. Method 2b is expected to yield the FW instanton and thus presents an alternative to the gMAM method used above.

Results. We focus on the FHN model for $\varepsilon = 0.1$ and $\sigma \approx 0.119$. Using method 1a, we obtain a transition path density that agrees with the observed distribution of the sample transition path ensemble generated with Monte Carlo rejection sampling (Fig. 4.4a). This supports the conjecture by (Hairer et al. 2007) that the theory underlying pathspace sampling (see section 2.2.2) indeed also holds for nongradient systems.

Furthermore, using methods 2a and 2b, respectively, we compute the minimizers of the OM and FW minimizers over the set $\mathcal{C}_{RL}^T$ of all paths leading from $\boldsymbol{x}_R$ to $\boldsymbol{x}_L$ in time T (see Table 4.2). As expected, the FW minimizer corresponds to the instanton obtained with gMAM and passes through the saddle. By contrast, the OM minimizer avoids the saddle and yields an appropriate approximation of the MTP (Fig. 4.4a).

Let us now focus on the distribution of *crossing points*, i.e. the points where sample transition paths cross the basin boundary. Comparing the crossing point distributions

Model	Functional	ε	σ	Δt	T	Δz	N_z
FHN	OM	0.1	0.11907	2.18×10^{-3}	2.79807	$\Delta t/1500$	5×10^7
FHN	FW	0.1	0.11907	2.18×10^{-3}	15.0	$\Delta t/1500$	5×10^7
COMP ($AB \to B$)	OM	0.01	0.006	3.07	390.0	10^{-2}	5×10^7

Table 4.2 | Specifications for pathspace sampling of the FHN and COMP models. The column Functional indicates which action functional was used for the minimization (OM - method 2; FW - method 3). N_z denotes the number of virtual time steps.

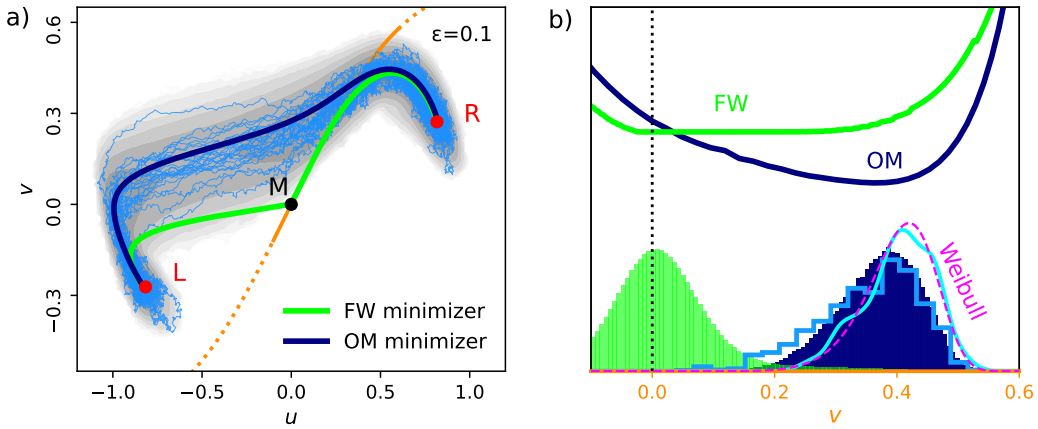

Figure 4.4 | **Predicting transition paths in the FHN model.** (a) FW (green) and OM (dark blue) minimizer compared with sample paths (light blue) for the transition ϕ_t^{RL}, showing the pathspace sampling density (gray contours) and basin boundary (orange). (b) Cross section along basin the boundary (solid orange line in a), projected onto v). Top lines: quasipotential as obtained by minimizing the FW action (green) and OM action (blue). Histograms show the crossing point distributions obtained from direct sampling (blue outline) compared with those from pathspace sampling using the FW (green) and OM (dark blue) action functionals. Additionally, the sampled first-exit point distribution (cyan) is compared with the rescaled Weibull distribution (magenta, see section 4.2.5). Figure adapted from (Börner et al. 2024b), with modifications.

obtained with method 1a (OM pathspace sampling) and method 1b (FW pathspace sampling) with that of the ensemble from Monte Carlo rejection sampling, we confirm the agreement between OM pathspace sampling and Monte Carlo sampling (Fig. 4.4b). By contrast, the distribution obtained from FW pathspace sampling is centered around the saddle point and very different from the observed crossing point distribution. This highlights that the FW action is not the appropriate functional for this variational problem under finite noise.

A finite-noise, finite-time quasipotential. Our results show that the OM minimizer gives better predictions of sample transition paths. This motivates constructing a finite-noise, finite-time quasipotential $\tilde{V}_R(\boldsymbol{x};\, \sigma, T)$ in the spirit of Eq. (2.20) but based on the OM action $\tilde{S}_T^\sigma$ (Eq. (2.38)) with fixed T,

$$\tilde{V}_R(\boldsymbol{x}_F;\, \sigma, T) = \inf_{\boldsymbol{\varphi}_t \in \mathcal{C}_{RF}} \tilde{S}_T^\sigma[\boldsymbol{\varphi}_t]\,, \tag{4.6}$$

where $\mathcal{C}_{RF}^T$ is the set of all paths leading from $\boldsymbol{x}_R$ to $\boldsymbol{x}_F$ in time T.

We compute this quantity along the basin boundary in two steps. First, we use method 2a to compute the OM action minimizer to different points on the basin boundary. Then, the OM action value along that minimizer gives the value of $\tilde{V}_R$. This quantity indeed exhibits a minimum on the basin boundary where sample transition paths are observed to cross (Fig. 4.4b).

4.2.5 First-exit statistics

To deepen the connection between our findings and previous studies, we conclude our analysis of the FHN model through the lens of the classical *first-exit problem*: in our case,

the escape from the basin of attraction of $\boldsymbol{x}_R$. This is an old problem in the literature (Hänggi et al. 1990; Maier and Stein 1993; Schuss and Spivak 1998; Kraut and Feudel 2003; Berglund 2013), given its relevance in various applications (e.g. chemical reactions (Arrhenius 1889), information technology (Bobrovsky and Schuss 1982) and power grids (Schäfer et al. 2017)).

Mean first-exit time. As discussed in section 2.1.6, the mean-first exit time $\langle \tau_\sigma^{RL} \rangle$ is given by Eyring-Kramers' law for gradients systems (Eyring 1935; Kramers 1940) and its generalization to the non-gradient case (Freidlin and Wentzell 1998; Bouchet and Reygner 2016; Bouchet and Reygner 2022). Since the FHN model attains the quasipotential minimum on the basin boundary at the saddle point $\boldsymbol{x}_M$, we expect the exponential scaling (Arrhenius' law; see Eq. (2.24)),

$$\langle \tau_\sigma^{RL} \rangle \asymp \exp\left(\frac{V_R(\boldsymbol{x}_M)}{\sigma^2}\right) \quad \text{as } \sigma \to 0\,. \tag{4.7}$$

More generally, based on Freidlin and Wentzell (1998) we would expect that the rate (the inverse of $\langle \tau_\sigma^{RL} \rangle$) at which trajectories reach the vicinity of a point $\boldsymbol{x}_c$ on the basin boundary scales like $\exp(-V_R(\boldsymbol{x}_c)/\sigma^2)$ in the weak-noise limit. However, we showed that for $\varepsilon \in \{0.01, 0.1\}$ sample transition paths do not distribute proportionally to $\exp(-V_R/\sigma^2)$ across the flat quasipotential region but are skewed away from the saddle (Figs. 4.1c and 4.4b): the boundary-crossing location $\boldsymbol{x}_c(\varepsilon, \sigma)$ of the MTP lies at the end of the danger zone $\mathcal{Z}$, far from $\boldsymbol{x}_M$. This is related to the instanton approaching the basin boundary tangentially (Maier and Stein 1997).

Given these discrepancies between FW theory and the finite-noise transition behavior, does Arrhenius' law still apply when there is saddle avoidance? To test this, we empirically estimate the mean first-exit time for a range of σ-values for each ε-case. We then plot $\ln \langle \tau_\sigma^{RL} \rangle$ as a function of $1/\sigma^2$ and perform a linear weighted least-squares fit of the data points. An exponential scaling according to Eq. (4.7) would require that the relationship between $\ln \langle \tau_\sigma^{RL} \rangle$ and $1/\sigma^2$ is approximately linear and that the slope corresponds to the quasipotential height at the saddle $V_R(\boldsymbol{x}_M)$.

Indeed, we observe Arrhenius' law in the transition paths ensembles for all ε-cases considered (Fig. 4.5). The p-values for each fit obtained from Chi-squared tests are

$$(0.213,\ 0.915,\ 0.981\,0.997) \text{ for } \varepsilon = (0.01, 0.1, 1, 10)\,,$$

supporting a linear relationship. The value of $V_R(\boldsymbol{x}_M)$ lies within the 90% confidence interval for all ε-values with the exception of $\varepsilon = 0.01$, where the values lies slightly outside.

In summary, despite saddle avoidance, the exponential scaling in Eyring-Kramers' law matches the quasipotential height at the saddle. An energetic argument supporting this finding is that the quasipotential at the mean crossing point $\boldsymbol{x}_c(\varepsilon, \sigma)$ is $V_R(\boldsymbol{x}_c) \approx V_R(\boldsymbol{x}_M) + \sigma^2$, yielding a comparable exponent as long as $\sigma^2 \ll V_R(\boldsymbol{x}_M)$. Our empirical results indicate that the pre-exponential factor is approximately constant in the examined range of σ-values (not shown) and hence does not break the exponential scaling. These findings corroborate the results of Bouchet and Reygner (2022) and Maier and Stein (1997) who derived that the pre-exponential factor is independent of σ under certain assumptions. The saddle thus remains relevant for predicting transition rates – even if

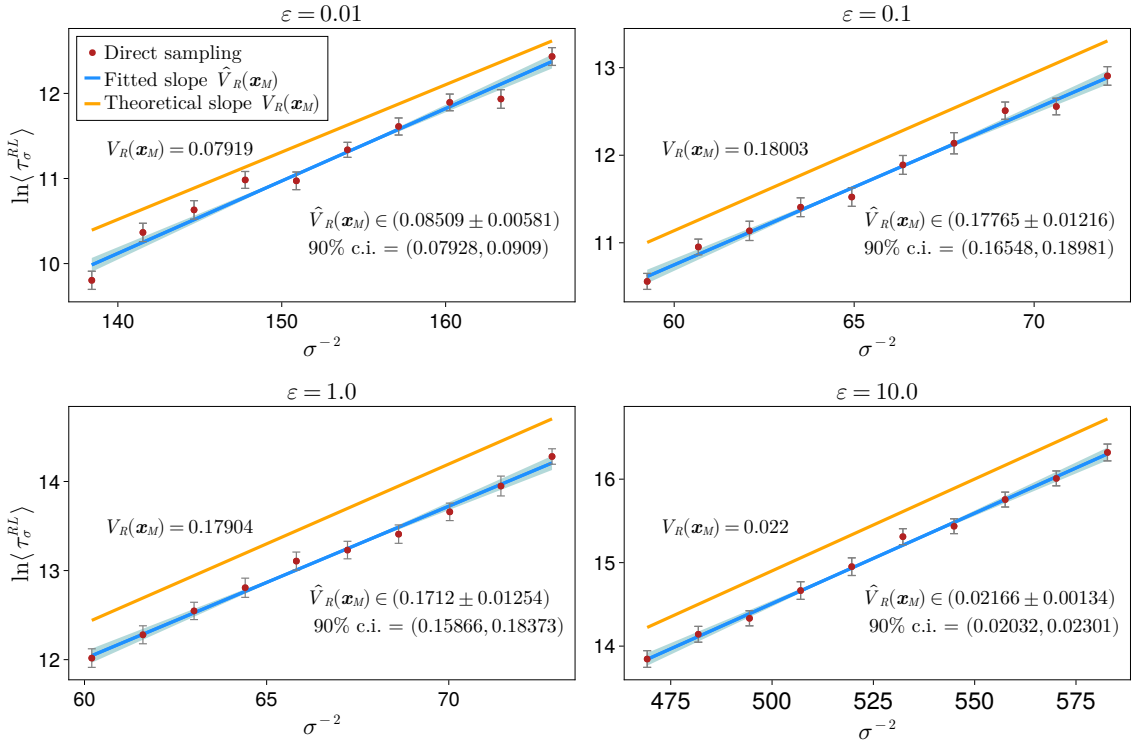

Figure 4.5 | **Arrhenius' law in the FHN model.** Panels show the dependence of $\langle \tau_\sigma^{RL} \rangle$ on the noise strength σ for each $\varepsilon \in \{0.01, 0.1, 1, 10\}$. Empirical estimates from Monte Carlo rejection sampling (red) are fitted via a linear weighted least-squares fit (blue), with an estimated slope $\hat{V}_R(\boldsymbol{x}_M)$ and its 90% confidence interval displayed. For comparison, the theoretical slope (orange) corresponding to the quasipotential at the saddle, $V_R(\boldsymbol{x}_M)$, is shown. Figure reproduced from Börner et al. (2024b), Supplemental Material, produced by Ryan Deeley.

the corresponding transition paths avoid the saddle. Conversely, the height of the energy barrier at the saddle can be approximately estimated from transition samples even if they avoid the saddle. Generally, however, saddle avoidance can strongly impact the pre-exponential factor in Eyring-Kramers' law and thereby affect the transition rate.

First-exit point distribution. Let $\boldsymbol{x}(t)$ be a trajectory of Eq. (4.4) initialized at $\boldsymbol{x}(0) = \boldsymbol{x}_R$. Its first-exit point is given by $\boldsymbol{x}_c = \boldsymbol{x}(t')$, where $t' = \inf_{t>0}\{t : \boldsymbol{x}(t) \in \partial\mathcal{B}_R\}$. We are interested in the distribution p_σ of first-exit points of an ensemble of transition paths.

Maier and Stein (1997) studied the first-exit point distribution of two-dimensional nongradient systems of the form in Eq. (4.1), assuming that the system has a saddle point $\boldsymbol{x}_M$ at which the Jacobian of $\boldsymbol{b}$ can be written as

$$\nabla \boldsymbol{b}(\boldsymbol{x}_M) = \begin{pmatrix} 1 & 0 \\ c & -\mu \end{pmatrix}. \tag{4.8}$$

Here $\boldsymbol{c} \in \mathbb{R}$ is a constant and $\mu = |\lambda_-|/\lambda_+$ is the ratio of the stable and unstable eigenvalues of $\nabla \boldsymbol{b}$ at $\boldsymbol{x}_M$ as before (see section 4.2.2).

In gradient systems, p_σ converges to a normal distribution centered at the saddle with standard deviation $\mathcal{O}(\sigma)$ as $\sigma \to 0$ (Maier and Stein 1997). For non-gradient systems, the authors showed that this is markedly different and the distribution tends to be non-Gaussian, skewed away from the saddle. Particularly, if $\mu < 1$ and the covariance matrix

$\boldsymbol{Q}(\boldsymbol{x}_M)$ at the saddle is orthogonal, the first-exit point distribution converges asymptotically to a *Weibull distribution*,

$$p_\sigma(s;\,\mu,A) \sim \begin{cases} \dfrac{2s^{(2/\mu)-1}}{\mu(A\sigma^\mu)^{2/\mu}} \exp\left(-\left(\dfrac{s}{A\sigma^\mu}\right)^{2/\mu}\right) & s>0 \\ 0 & s\le 0 \end{cases}, \tag{4.9}$$

controlled by the shape parameter $2/\mu$ and scale parameter $A\sigma^\mu$ (see Eq. (6.16) of Maier and Stein (1997)). Here $s \coloneqq q - q_M$ is the distance to $\boldsymbol{x}_M$ along the vertical coordinate q. Note that p_σ is entirely distributed on one side of the saddle. Since the mode of p_σ equals $A((2-\mu)/2)^{2/\mu}\sigma^\mu$, the first-exit points are distributed on a length scale $\mathcal{O}(\sigma^\mu)$ away from the saddle, and p_σ converges only logarithmically to the saddle as $\sigma \to 0$.

We can cast the FHN model (Eq. (4.3)) into a form satisfying Eq. (4.8) and the orthogonality of $\boldsymbol{Q}(\boldsymbol{x}_M)$ via a coordinate transformation and a rescaling of time. Defining the rotation matrix $\boldsymbol{H}$,

$$\boldsymbol{H} = \frac{1}{\sqrt{m^2+1}}\begin{pmatrix} 1 & m \\ -m & 1 \end{pmatrix}, \tag{4.10}$$

where $m = \left(\sqrt{(1+3\varepsilon)^2 - 4\varepsilon} - (1+3\varepsilon)\right)/(2\varepsilon)$, we first transform from $\boldsymbol{x} = (u,v)$ to the coordinates $(p,q) = \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$. In the transformed coordinates, the drift field (Eq. (4.3)) reads

$$\tilde{\boldsymbol{b}}_{\mathrm{FHN}}(\boldsymbol{y}) = \frac{1}{\varepsilon(m^2+1)}\begin{pmatrix} -\dfrac{(p-mq)^3}{m^2+1} + \zeta p \\ \dfrac{m(p-mq)^3}{m^2+1} + \nu p + \chi q \end{pmatrix} = \boldsymbol{H}\boldsymbol{b}_{\mathrm{FHN}}(\boldsymbol{H}^{-1}\boldsymbol{y}), \tag{4.11}$$

where

$$\zeta = -3\varepsilon m^2 + (\varepsilon - 1)m + 1, \quad \nu = m^2 - (1+3\varepsilon)m + \varepsilon, \quad \chi = m^2 + (1-\varepsilon)m - 3\varepsilon.$$

Then, in rescaled time $\tau = t/\omega$, with $\omega = \varepsilon(m^2+1)/\zeta > 0$, we obtain the SDE,

$$\boldsymbol{y} = \omega\tilde{\boldsymbol{b}}_{\mathrm{FHN}}(\boldsymbol{y})\,\mathrm{d}\tau + \tilde{\sigma}\boldsymbol{H}\;\mathrm{d}\boldsymbol{W}_\tau\,, \tag{4.12}$$

where $\tilde{\sigma} = \sqrt{\omega}\sigma$ and $\boldsymbol{W}_\tau$ is a Wiener process in the rescaled time. Equation (4.12) is equivalent to the original FHN system at $\varepsilon = 0.1$ if we set $m = -0.82109$, $\zeta = 1.53673$, $\nu = 1.84161$, $\chi = -0.36479$, $\omega = 0.10895$, and $\mu = 0.23738$. The noise strength $\sigma = 0.11907$ corresponds to $\tilde{\sigma} = 0.03930$ in rescaled time. The stable fixed points of the transformed FHN model are located at $\boldsymbol{y}_{R,L} = \boldsymbol{H}\boldsymbol{x}_{R,L}$ while the saddle $\boldsymbol{y}_M = (0,0)$ remains at the origin.

According to Maier and Stein (1997), the first-exit point distribution of the transformed FHN model should follow the Weibull distribution in Eq. (4.9) as a function of the vertical coordinate q. To verify this, we take the first-exit points $\{x_c\}$ of the ensemble of 306 sample transition paths for $\varepsilon = 0.1$ and transform their coordinates to $\boldsymbol{y}_c = \boldsymbol{H}\boldsymbol{x}_c$. Indeed, their distribution along the q coordinate is accurately described by the Weibull distribution $p_{\tilde{\sigma}}(q;\,\mu,A)$ with $\tilde{\sigma} = 0.03930$, $\mu = 0.23738$ and $A = 1.23$, as shown in Fig. 4.4b (where the distributions are plotted as a function of the coordinate v after transforming back from q). We determined the value of A via a least-squares fit since evaluating the

ratio v/u^{μ} along the numerically computed instantons did not yield a consistent limiting value.

While the observed first-exit point distribution (of the transition path ensemble at $\varepsilon = 0.1$) is well represented by a Weibull distribution, there is a clear difference between the *first-exit* point distribution and the *crossing* point distribution, where the latter considers all crossings of the basin boundary, not only the first one (Fig. 4.4b). We explain this with the fact that the sample transition paths often cross the basin boundary multiple times before reaching the competing attractor $\boldsymbol{x}_L$. As they cross multiple times, they are attracted towards the saddle along the basin boundary, such that the first-exit point distribution is more heavily skewed away from the saddle than the crossing point distribution.

4.3 Multiplicative noise: Two competing species

Our exploration of the FHN model has shown how timescale separation causes saddle avoidance in a simple bistable system with additive noise. To illustrate that saddle avoidance is a more general phenomenon in multiscale stochastic systems, we study an additional two-dimensional example featuring more than two attractors and multiplicative noise.

4.3.1 Model description

We consider a model inspired by population dynamics that describes the competition between two species A and B. The model, which we call the COMP model, is based on Bazykin (1998) with modifications by Ulrike Feudel (Börner et al. 2024b). The growth of each population is modeled with an Allee effect (Stephens et al. 1999) according to

$$\boldsymbol{b}_{\text{comp}}(x, y) = \begin{pmatrix} x(x-\alpha_A)(1-x) - \beta_A xy \\ \varepsilon[y(y-\alpha_B)(1-y) - \beta_B xy] \end{pmatrix} . \tag{4.13}$$

Here x and y denote the population densities of species A and B, respectively, which each go extinct below their critical density $\alpha_A = 0.1$, $\alpha_B = 0.3$. The parameter ε represents the ratio of net growth rates of the two species. The competition term is controlled by $\beta_A = 0.18$ and $\beta_B = 0.1$.

Population dynamics in ecological systems are typically subjected to fluctuations due to internal variability (*demographic noise*) and due to changes in the environmental conditions (*environmental noise*) (Meng et al. 2020). Demographic noise depends on the population size, and commonly a square-root dependence on the population size is proposed (Weissmann et al. 2018). This motivates studying the COMP model with multiplicative, or state-dependent, noise.

Specifically, we study the SDE,

$$\begin{aligned} \mathrm{d}x &= \big(x(x-\alpha_A)(1-x) - \beta_A xy\big)\mathrm{d}t + \sigma\sqrt{x}\mathrm{d}W_{x,t} \\ \mathrm{d}y &= \big(\varepsilon[y(y-\alpha_B)(1-y) - \beta_B xy]\big)\mathrm{d}t + \sigma\sqrt{y}\mathrm{d}W_{y,t}\,, \end{aligned} \tag{4.14}$$

where $W_{x,t}$ and $W_{y,t}$ denote independent standard Wiener processes as before but the noise term now depends on the population. This system corresponds to Eq. (4.1) with $\boldsymbol{b} = \boldsymbol{b}_{\text{comp}}$ and a state-dependent covariance matrix $\boldsymbol{Q} = \mathrm{diag}(x, y)$.

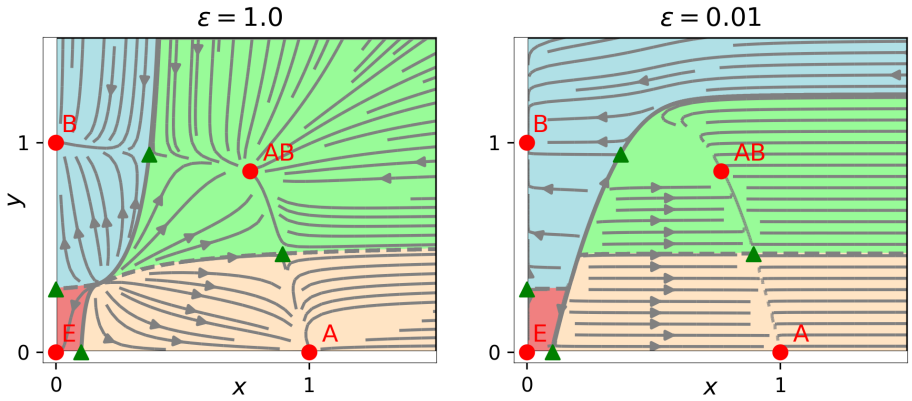

Figure 4.6 | **Global stability of the COMP model** for $\varepsilon = 1$ (left) and $\varepsilon = 0.01$ (right). The stable fixed points (saddle points) are marked as red circles (green triangles). Flow lines of the drift $\boldsymbol{b}_{\text{comp}}$ are indicated as gray lines and colored shading visualizes the competing basins of attraction.

4.3.2 Global stability and quasipotential landscape

Multistability. With the chosen parameter settings, the COMP model possesses four stable equilibrium points: a state $\boldsymbol{x}_{AB}$ where both species coexist, two states $\boldsymbol{x}_A, \boldsymbol{x}_B$ where only one species survives, and a full extinction state $\boldsymbol{x}_E$:

$$\boldsymbol{x}_{AB} \approx (0.767, 0.864)\,, \quad \boldsymbol{x}_A = (1, 0)\,, \quad \boldsymbol{x}_B = (0, 1)\,, \quad \boldsymbol{x}_E = (0, 0)\,.$$

Additionally, four saddle points and one repeller exist in the non-negative quadrant (Fig. 4.6). They are embedded in the basin boundaries that separate the four basins of attraction. The coordinates of the saddles read

$$\boldsymbol{x}_{AB\leftrightarrow A} \approx (0.894, 0.468)\,, \quad \boldsymbol{x}_{AB\leftrightarrow B} \approx (0.369, 0.943)\,, \quad \boldsymbol{x}_{E\leftrightarrow A} = (0.1, 0)\,, \quad \boldsymbol{x}_{E\leftrightarrow B} = (0, 0.3)\,,$$

where the subscript indicates which basins of attraction each saddle separates. Due to the multistability, several transition scenarios are possible, and each saddle mediates a transition scenario.

Let us focus on the extinction of one species, realized by the two scenarios $AB \to A$ and $AB \to B$. We will contrast two cases:

- $\varepsilon = 1$: Both species have the same net growth rate.
- $\varepsilon = 0.01$: The population x of species A grows much faster than the population y of species B.

Quasipotentials. As for the FHN model, we compute the FW quasipotential for each ε-case to understand the transition behavior in terms of the system's stability landscape. The `OLIM4VAD` algorithm by Dahiya and Cameron (2018a) supports variable anisotropic diffusion, so we can use it also in the case of multiplicative noise. Here we set `K=30` and `XMIN=-1.0, XMAX=3.0, YMIN=-1.0, YMAX=3.0`, with $\boldsymbol{x}_{AB} = (0.767, 0.864)$ as the reference point.

If both species have similar net growth rates ($\varepsilon = 1$), the quasipotential V_{AB} exhibits a clear channel of low quasipotential from $AB \to A$ and $AB \to B$, respectively (Fig. 4.7a). Along the relevant basin boundary, a distinct quasipotential minimum is located

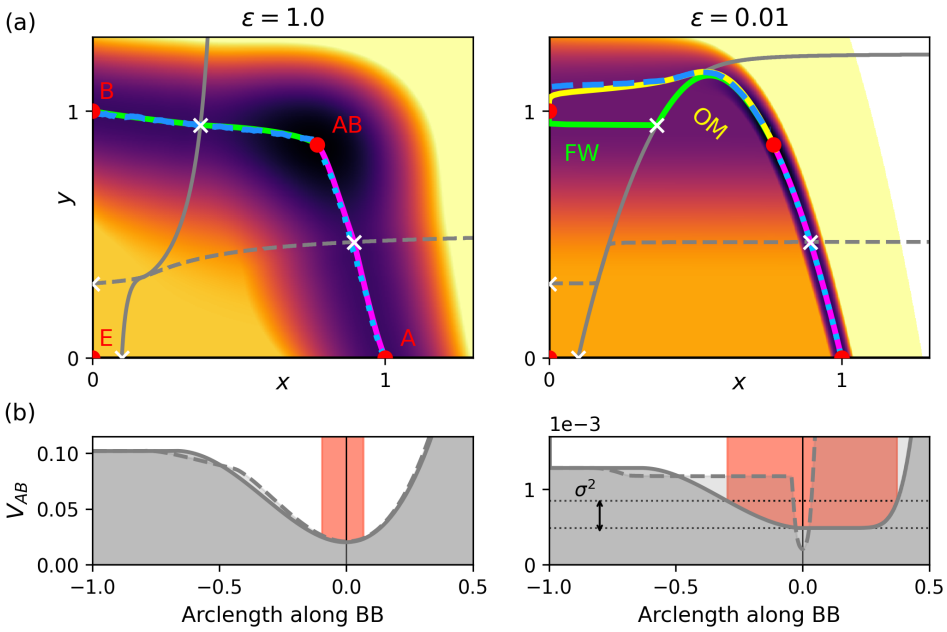

Figure 4.7 | **Quasipotential landscape of the COMP model.** (a) V_{AB} with respect to $\boldsymbol{x}_{AB}$ (black: $V_{AB} = 0$; brighter color indicates larger V_{AB}) for $\varepsilon \in \{1, 0.01\}$. The stable fixes points (red points), saddles (white crosses) and basin boundaries (gray solid/dashed lines) are indicated. Additionally, the FW instantons from $\boldsymbol{x}_{AB}$ to $\boldsymbol{x}_A$ (magenta) and $\boldsymbol{x}_B$ (green) are compared with the corresponding MTPs (dashed blue). For the saddle-avoidant scenario $AB \to B$ ($\varepsilon = 0.01$), the estimated OM minimizer (yellow) is shown. (b) V_{AB} along the solid/dashed basin boundary displayed in (a), as a function of arclength from the respective saddle, highlighting the danger zone $\mathcal{Z}$ (red shading) for the solid basin boundary. Figure reproduced from Börner et al. (2024b).

at the saddle $\boldsymbol{x}_{AB\leftrightarrow A}$ and $\boldsymbol{x}_{AB\leftrightarrow B}$, respectively (Fig. 4.7b). The MTPs track the FW instanton and cross near the corresponding saddle. Contrarily, when A grows faster than B ($\varepsilon = 0.01$), an asymmetry emerges in the quasipotential landscape: along the route $AB \to A$, the quasipotential channel becomes very steep and narrow, yielding a well-defined "mountain pass" at the saddle $x_{AB\leftrightarrow A}$. Between $\boldsymbol{x}_{AB}$ and $\boldsymbol{x}_B$, however, the region of relatively low quasipotential widens and becomes flat around the saddle $x_{AB\leftrightarrow B}$. Similar to the FHN model with $\varepsilon < 1$, this leads to an extended danger zone $\mathcal{Z}$ across the quasipotential plateau where the "energy" provided by the stochastic fluctuations ($\sim \sigma^2$) exceeds the quasipotential height difference relative to the saddle.

Based on our insights from the FHN model analysis, we expect that FW theory will adequately describe the transition behavior for $\varepsilon = 1$ and for the scenario $AB \to A$ in the case $\varepsilon = 0.01$, but that it will fail to describe transition scenario $AB \to B$ if $\varepsilon = 0.01$ due to the flat quasipotential. We verify this in the next section by sampling transition paths and computing instantons.

4.3.3 Sample transition paths and instantons

Methodology. Analogous to the FHN model, we generate ensembles of transition paths via Monte Carlo rejection sampling, using the settings listed in Table 4.1. As before, we fix the noise strength such that the rareness $r_\sigma \gg 1$ (see Eq. (4.2)). It turns out that noise-induced transitions in the COMP model can have very long mean transition times

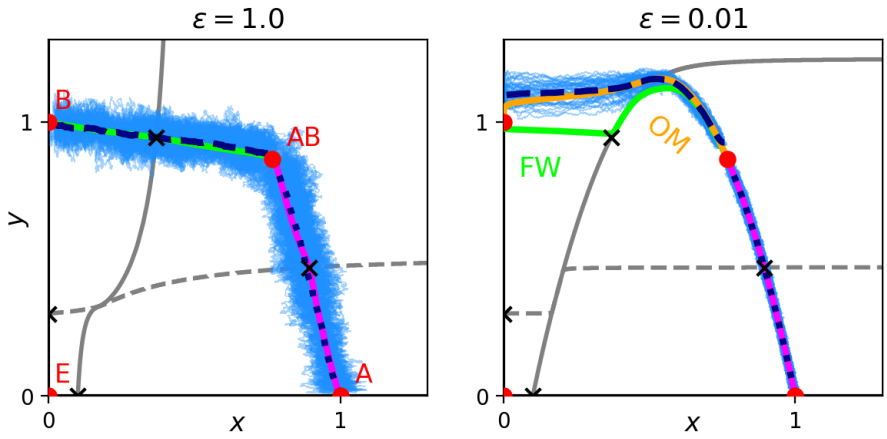

Figure 4.8 | **Noise-induced transition path ensembles in the COMP model** for $\varepsilon = 1$ (left) and $\varepsilon = 0.01$ (right). As Fig. 4.7(a) but showing the ensembles of sample transition paths instead of the quasipotential. Simulation settings for the ensembles are summarized in Table 4.1. Figure reproduced from Börner et al. (2024b), Supplementary Material.

$\langle t_\sigma \rangle$, which makes it computationally expensive to achieve a large r_σ. Nonetheless, we choose σ to ensure that $r_\sigma \geq 100$ for all transition scenarios (see Table 4.1).

Additionally, we compute the FW instanton for each scenario as before, using the gMAM algorithm (Heymann and Vanden-Eijnden 2008b). Here me must take into account that the covariance matrix, and thus the $\boldsymbol{Q}$-norm in the FW action functional (see Eq. (2.18)), is now state-dependent ($\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{x})$). We initialize the instantons as straight lines in the phase space discretized by 250 points. In the case $\varepsilon = 0.01$, due to slow convergence of gMAM, we construct the instanton from $\boldsymbol{x}_{AB}$ to $\boldsymbol{x}_B$ by stitching together the gMAM instanton from $\boldsymbol{x}_{AB}$ to $\boldsymbol{x}_{AB\leftrightarrow B}$ and the heteroclinic orbit from $\boldsymbol{x}_{AB\leftrightarrow B}$ to $\boldsymbol{x}_B$. We verified in the previous section that the quasipotential attains its global minimum along the relevant basin boundary at $\boldsymbol{x}_{AB\leftrightarrow B}$, such that this approach is valid.

Results. The ensembles of sample transition paths concentrate within a tube around a typical path, confirming that we are in a weak-noise regime (Fig. 4.8). Therefore, we can describe their mean behavior in terms of the mean sample transition path (MTP) as for the FHN model. As anticipated, the MTP agrees with the FW instanton in all cases except the case $AB \to B$ at $\varepsilon = 0.01$, where the quasipotential is flat. In this case, the instanton tangentially approaches the basin boundary and passes through the saddle but all sample transition paths avoid the saddle.

In summary, the COMP model exhibits saddle avoidance due to timescale separation just like the FHN model. Since the COMP model has more than two attractors, FW theory may accurately predict the path of one transition scenario while failing for another – in the same model under the same settings. The flatness of the quasipotential and grazing nature of the instanton approaching the basin boundary are key indicators of saddle avoidance.

4.3.4 Onsager-Machlup action minimizer

Once again, we are left with the question of how to accurately predict the transition paths in the case where the ensemble diverges from the instanton and avoids the saddle. Does the minimization of the OM action provide the answer, as in the FHN model? Generally, the OM action is not the suitable functional for multiplicative noise (Dürr and Bach 1978), and the prediction of the most probable transition path under finite multiplicative noise is an topic of ongoing research (Thorneywork et al. 2024; Kappler et al. 2024).

However, the demographic noise considered here presents a special case where we can transform the SDE (4.14) into an SDE with additive noise[4]. We thus propose that a finite-noise most probable transition path may be computed for the COMP model by transforming the model to a version with additive noise, computing the OM minimizer in that model, and then transforming back to the original model.

We are looking for a coordinate transform under which the noise term in Eq. (4.14) becomes independent of x and y. The general formula for SDEs is given by Itô's formula (Gardiner 2009). Specifically, by choosing the change of variables $\boldsymbol{l} := (m, n) = h(\boldsymbol{x})$, where $h : (x, y) \mapsto (2\sqrt{x}, 2\sqrt{y})$, we arrive at an SDE of the form in Eq. (4.1) with $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{l}) = \mathbb{I}_2$ and the transformed drift $\tilde{\boldsymbol{b}}_{\text{comp}}$,

$$\tilde{\boldsymbol{b}}_{\text{comp}}(\boldsymbol{l}) := \begin{pmatrix} \tilde{b}^{(1)}_{\text{comp}}(\boldsymbol{l}) \\ \tilde{b}^{(2)}_{\text{comp}}(\boldsymbol{l}) \end{pmatrix} = \begin{pmatrix} \frac{m}{2}\left(\frac{m^2}{4} - \alpha_A\right)\left(1 - \frac{m^2}{4}\right) - \beta_A \frac{mn^2}{8} - \frac{\sigma^2}{2m} \\ \varepsilon\left[\frac{n}{2}\left(\frac{n^2}{4} - \alpha_B\right)\left(1 - \frac{n^2}{4}\right) - \beta_B \frac{m^2 n}{8}\right] - \frac{\sigma^2}{2n} \end{pmatrix} . \tag{4.15}$$

Thus, the transformation leads from multiplicative noise ($\boldsymbol{Q}(\boldsymbol{x}) = \text{diag}(x, y)$) to additive noise ($\tilde{\boldsymbol{Q}}(\boldsymbol{l}) = \mathbb{I}_2$). Note that the transformed drift depends on the noise strength σ (but is purely deterministic).

Pathspace SPDE. To obtain the OM minimizer in the transformed system, we perform pathspace sampling with simulated annealing, as described in section 4.2.4 (method 2, version a). The SPDE for pathspace sampling is given by Eq. (2.40), where the individual terms are specifically

$$\boldsymbol{\nabla}\left\langle \tilde{\boldsymbol{b}}_{\text{comp}}(\boldsymbol{l}), \tilde{\boldsymbol{b}}_{\text{comp}}(\boldsymbol{l}) \right\rangle = \begin{pmatrix} \frac{\partial\left(\tilde{b}^{(1)}_{\text{comp}}\right)^2}{\partial m} + \frac{\partial\left(\tilde{b}^{(2)}_{\text{comp}}\right)^2}{\partial m} \\ \frac{\partial\left(\tilde{b}^{(1)}_{\text{comp}}\right)^2}{\partial n} + \frac{\partial\left(\tilde{b}^{(2)}_{\text{comp}}\right)^2}{\partial n} \end{pmatrix} , \tag{4.16}$$

[4]We would like to thank Moshir Harsh for suggesting this transformation.

with

$$\frac{\partial\left(\tilde{b}^{(1)}_{\mathrm{comp}}(\boldsymbol{l})\right)^2}{\partial m} z = \frac{\big(m(m^2-4\alpha_A)(-m^2+4)-4\beta_A n^2 m\big)\big(-5m^4+12\alpha_A m^2+12m^2-16\alpha_A-4\beta_A n^2\big)}{512} - \frac{\sigma^2 m(-m^2+2\alpha_A+2)}{8} - \frac{\sigma^4}{2m^3},$$

$$\frac{\partial\left(\tilde{b}^{(2)}_{\mathrm{comp}}(\boldsymbol{l})\right)^2}{\partial n} = \frac{\varepsilon^2\big(n(n^2-4\alpha_B)(-n^2+4)-4\beta_B m^2 n\big)\big(-5n^4+12\alpha_B n^2+12n^2-16\alpha_B-4\beta_B m^2\big)}{512} - \frac{\varepsilon\sigma^2 n(-n^2+2\alpha_B+2)}{8} - \frac{\sigma^4}{2n^3},$$

$$\frac{\partial\left(\tilde{b}^{(1)}_{\mathrm{comp}}(\boldsymbol{l})\right)^2}{\partial n} = \frac{\beta_A mn\big(4\beta_A mn^2 - m(-m^2+4)(m^2-4\alpha_A)\big)}{64} + \frac{\beta_A\sigma^2 n}{4},$$

$$\frac{\partial\left(\tilde{b}^{(2)}_{\mathrm{comp}}(\boldsymbol{l})\right)^2}{\partial m} = \frac{\varepsilon^2\beta_B nm\big(4\beta_B nm^2 - n(-n^2+4)(n^2-4\alpha_B)\big)}{64} + \frac{\varepsilon\beta_B\sigma^2 m}{4},$$

as well as

$$\big(\boldsymbol{\nabla}\tilde{\boldsymbol{b}}_{\mathrm{comp}}(\boldsymbol{l})\big) - \big(\boldsymbol{\nabla}\tilde{\boldsymbol{b}}_{\mathrm{comp}}(\boldsymbol{l})\big)^\top = \begin{pmatrix} 0 & -\frac{\beta_A mn}{4} + \frac{\varepsilon\beta_B mn}{4} \\ -\frac{\varepsilon\beta_B mn}{4} + \frac{\beta_A mn}{4} & 0 \end{pmatrix}, \tag{4.17}$$

and

$$\boldsymbol{\nabla}\Big(\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{b}}_{\mathrm{comp}}(\boldsymbol{l})\Big) = \begin{pmatrix} -\frac{5m^3}{8} + \frac{3(1+\alpha_A)m}{4} - \frac{\sigma^2}{m^3} - \frac{\varepsilon\beta_B m}{4} \\ \varepsilon\left[-\frac{5n^3}{8} + \frac{3(1+\alpha_B)n}{4}\right] - \frac{\sigma^2}{n^3} - \frac{\beta_A n}{4} \end{pmatrix}. \tag{4.18}$$

We integrate the SPDE (2.40) with boundary conditions $\boldsymbol{x}(0,z) = h(\boldsymbol{x}_{AB})$ and $\boldsymbol{x}(T,z) = h(\boldsymbol{x}_B)$ for all $z \geq 0$. For the value of T we choose the mean transition time $T = 390$ of the transition path ensemble for $AB \to B$ at $\varepsilon = 0.01$. Further settings are provided in Table 4.2.

Results. To convert the resulting OM minimizer path back to the original coordinate frame, we apply the inverse transformation $h^{-1} : (m,n) = (m^2/4, n^2/4)$. This yields what we consider the OM minimizer of the COMP model (yellow line in Figs. 4.7 and 4.8). The OM minimizer avoids the saddle and yields a reasonable approximation of the MTP, although the agreement diminishes towards the end of the path, close to the absorbing boundary $x = 0$. This could be related to the fact that the transformed drift $\tilde{\boldsymbol{b}}_{\mathrm{comp}}$ contains a term proportional to σ^2 that diverges as $m \to 0$ (and thus $x \to 0$).

4.4 Reconciling the fast-slow and weak-noise limits

To conclude our study of saddle avoidance, let us find out why the OM minimizer captures saddle avoidance in agreement with sample transition paths, whereas the FW minimizer (the instanton) does not.

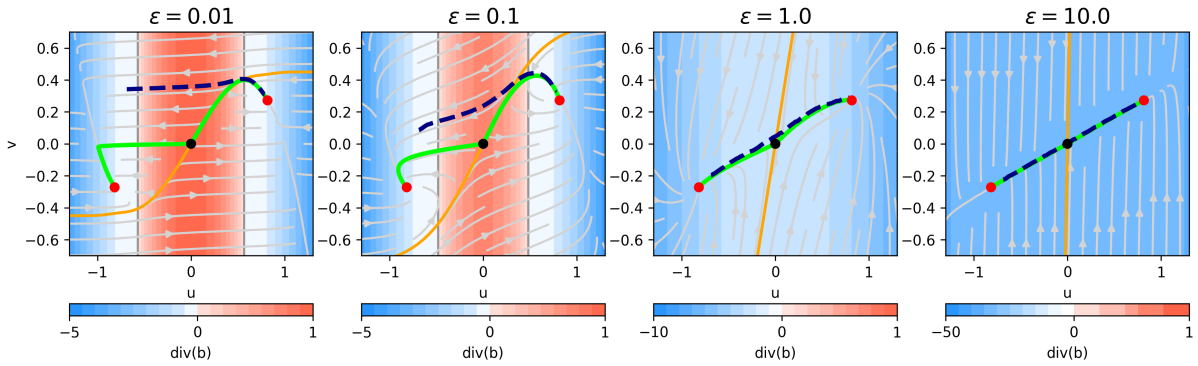

Figure 4.9 | **Divergence of the drift field** in the FHN model for $\varepsilon \in \{0.01, 0.1, 1, 10\}$. Colored shading displays the magnitude of $\boldsymbol{\nabla} \cdot \boldsymbol{b}_{\text{FHN}}$, where positive (red) and negative (blue) values are plotted on separate linear color scales. Additional elements in each panel are the same as in Fig. 4.1a. Figure reproduced from Börner et al. (2024b), Supplemental Material.

The difference between the FW action (Eq. (2.18)) and the OM action (Eq. (2.38)) is the additional divergence term in the OM action,

$$\tilde{S}_T^{\sigma}[\boldsymbol{\varphi}_t] - S_T[\boldsymbol{\varphi}_t] = \frac{\sigma^2}{2} \int_0^T \boldsymbol{\nabla} \cdot \boldsymbol{b}(\boldsymbol{\varphi}_t)\, \mathrm{d}t \ , \tag{4.19}$$

which depends on σ and adds a penalty to the FW action where the divergence of the drift, $\boldsymbol{\nabla} \cdot \boldsymbol{b}$, is positive. The OM action limits to the FW action as $\sigma \to 0$ and the divergence term may be viewed as a finite-noise correction to the FW action (Fleurantin et al. 2023). Our results in the FHN and COMP models suggest that this correction becomes more important as the timescale parameter $\varepsilon \to 0$. What is the connection between the divergence of the drift, saddle avoidance, and timescale separation in fast-slow systems?

As mentioned in section 4.2.2, there is a bijective relationship between the timescale parameter ε and the parameter $\mu = |\lambda_-|/\lambda_+$, the ratio of the stable and unstable Jacobian eigenvalues of the drift $\boldsymbol{b}$ at the saddle. The eigenvalues $\lambda_\pm$ are the roots of the characteristic equation, $\det(\boldsymbol{\nabla b} - \lambda \mathbb{I}_2) = 0$, which in the two-dimensional case can be expanded as

$$0 = \lambda^2 - (\boldsymbol{\nabla} \cdot \boldsymbol{b})\lambda + \det(\boldsymbol{\nabla b}) \ . \tag{4.20}$$

Thus, the eigenvalues take the form,

$$\lambda_\pm = \theta \pm \psi \quad \text{with } \theta = \frac{\boldsymbol{\nabla} \cdot \boldsymbol{b}}{2} \, , \ \psi = \sqrt{\frac{(\boldsymbol{\nabla} \cdot \boldsymbol{b})^2}{4} - \det(\boldsymbol{\nabla b})} \, . \tag{4.21}$$

Note that θ is directly proportional to the divergence, $\boldsymbol{\nabla} \cdot \boldsymbol{b}$. Now, we use the fact that we are considering the eigenvalues at the saddle, where by definition $\lambda_+ > 0$ and $\lambda_- < 0$. Then, $|\lambda_-| = -\lambda_- = \psi - \theta$, such that

$$\mu = \frac{|\lambda_-|}{\lambda_+} = \frac{\psi - \theta}{\theta + \psi} \, .$$

Maier and Stein (1997) showed that saddle avoidance occurs if $\mu < 1$. This implies

$$\mu < 1 \iff \psi - \theta < \psi + \theta \iff 2\theta > 0 \iff \boldsymbol{\nabla} \cdot \boldsymbol{b} > 0 \, . \tag{4.22}$$

We have thus shown that the condition for saddle avoidance $\mu < 1$ is equivalent to having a positive divergence of the drift at the saddle. Due to continuity, the divergence is then also positive in a neighborhood around the saddle.

Let us consider a general two-dimensional fast-slow system of the form in Eq. (2.30). Via a rescaling of time $t' = t/\varepsilon$, we arrive at the equivalent form

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t'} &= f(x, y) \\ \frac{\mathrm{d}y}{\mathrm{d}t'} &= \varepsilon g(x, y) \ , \end{aligned} \tag{4.23}$$

where we assume that f and g are independent of ε. The divergence of the drift $\boldsymbol{b} = (f, g)$ of this system is given by

$$\boldsymbol{\nabla} \cdot \boldsymbol{b}(x, y) = \frac{\partial f}{\partial x} + \varepsilon \frac{\partial g}{\partial y} \, ,$$

so the timescale parameter could determine the sign of the divergence depending on the values of $\partial f/\partial x$ and $\partial g/\partial y$. Recall that as $\varepsilon \to 0$, the dynamics are confined to the critical manifold C_0 where $f = 0$. The stability of the critical manifold is given by the sign of $\partial f/\partial x$; particularly, the critical manifold is unstable where $\partial f/\partial x > 0$.

From these considerations, we conclude that if the saddle lies on an unstable branch of the critical manifold, there always exists an $\varepsilon_c > 0$ below which $\boldsymbol{\nabla} \cdot \boldsymbol{b} > 0$ at the saddle. Consequently, for $\varepsilon < \varepsilon_c$ we get an added cost in the OM action functional for paths passing through the saddle, which increases as $\varepsilon \to 0$ but decreases proportionally to σ^2 as $\sigma \to 0$.

In the FHN model, we found that $\varepsilon_c = 1/\beta$, corresponding to $\mu = 1$. For $\varepsilon > \varepsilon_c$, the divergence is negative everywhere (Fig. 4.9). By contrast, for $\varepsilon < \varepsilon_c$ there exists a region $\mathcal{D}$ of positive divergence around the saddle,

$$\mathcal{D} = \left\{ (u, v) : |u| < \sqrt{\frac{1 - \beta\varepsilon}{3}} \right\} \, ,$$

where $\beta = 3$ in the cases we studied.

Since the integral in the correction term of the OM action (Eq. (4.19)) is taken over the entire path $\boldsymbol{\varphi}_t$, it is the divergence along the entire path that matters in the minimization problem. The OM minimizer will thus find the optimal balance between a) following a low action path in the FW quasipotential and b) staying in regions of low divergence of $\boldsymbol{b}$ for as much as possible. This balance is weighted by σ^2, so the stronger the noise, the more the divergence becomes important.

In fast-slow systems, one way to minimize the added cost of the divergence term is by passing through any inevitable regions of positive divergence as quickly as possible – following the fast variable. In the cases $\varepsilon \in \{0.01, 0.1\}$ in the FHN model, following the FW instanton up to the saddle is a slow process, since the instanton becomes tangential to the critical manifold. Thus, it is much "cheaper" for transition paths to cross the basin boundary near the fold point of the critical manifold where they are quickly repelled across the region of positive divergence (Fig. 4.9).

Based on these considerations, we argue that the OM action is the suitable finite-noise action functional in stochastic fast-slow systems because it captures the competition between the limits of weak noise ($\sigma \to 0$) and infinite timescale separation ($\varepsilon \to 0$).

4.5 Discussion & conclusion

The key result of this chapter is that multiscale dynamics can cause noise-induced transitions to bypass the saddle point between competing attracting states. Since physical systems are typically non-gradient, noisy and multiscale, this phenomenon can appear in various applications, even where transitions classify as rare events. Transition path ensembles may deviate strongly from the minimizer of the FW action for weak yet finite noise, while Kramers' law remains valid. These properties are possible due to a flat quasipotential along the basin boundary, which may occur due to timescale separation in the drift term for additive and multiplicative noise alike.

Despite avoiding the instanton, sample transition paths still tend to bundle within a tube around a typical transition path, manifesting the general notion of dynamical typicality (Galfi and Lucarini 2021; Lucarini et al. 2023). Transition paths can be efficiently sampled in pathspace under stochastic minimization of the appropriate action functional. For additive noise of finite amplitude, minimizing the OM action (for an appropriate path travel time T) yields an apt prediction of this most probable transition path and the transition path distribution, in contrast to minimizing the FW action. For multiplicative noise, computing finite-noise most probable transition paths is a topic of ongoing research (Thorneywork et al. 2024; Kappler et al. 2024; Dürr and Bach 1978). Interestingly, estimating the FW minimizer with machine learning using *deep gMAM* (Simonnet 2023) yields a path that more closely resembles the OM minimizer instead of the true FW instanton[5]. Addressing the mathematical challenges of multiplicative and degenerate noise are of great interest for future work. Furthermore, it would be interesting to study how saddle avoidance depends on the anisotropy of the noise, which is known to affect the course of transition paths (Dannenberg et al. 2014).

The studies by Maier and Stein (1997) and Luchinsky et al. (1999) presented a path-geometric study of saddle avoidance in two dimensions based on the local stability at the saddle. Our main contribution lies in linking saddle avoidance to timescale separation, which – to the best of our knowledge – had not been done previously. Furthermore, our work complements the local asymptotic approach by (a) providing a global stability viewpoint based on the quasipotential, (b) establishing a link to fast-slow systems theory by considering the stability of the critical manifold and (c) clarifying the relationship between the first-exit point distribution, positive divergence at the saddle and timescale separation. These concepts may enable anticipating saddle avoidance also in non-gradient fast-slow systems of higher dimension: if the saddle exists on an unstable branch of the critical manifold, we conjecture that repulsion away from the saddle outweighing attraction towards it will generally cause the instanton to approach the basin boundary tangentially, resulting in a quasipotential plateau along the saddle's stable set. Other physical interpretations of saddle avoidance, such as anisotropic friction (Berezhkovskii et al. 1989), may be interpreted as a multiscale feature.

Even though saddle avoidance has been observed already 40 years ago, we found that its possibility and implications have not been widely appreciated in our scientific field, including the dynamical systems community. One reason for this could be that the theory and intuition about stochastic dynamics are often guided by the gradient case (Term 2.8). As clarified by Maier and Stein (1997), a crucial difference between gradient and nongradient systems is that the length scale of saddle avoidance converges linearly

[5] Eric Simonnet, personal communication.

as $\sigma \to 0$ in gradient systems but only logarithmically in nongradient systems. In our terms, unexpected behavior occurs where the quasipotential becomes flat, since there the nongradient term $\mathcal{R}_R$ (Eq. (2.21)) becomes important. In the latter case, the weak-noise limit might present an inadequate approximation for any noise strength at which transitions have a chance of being observed within numerical constraints or on timescales of interest.

While transitions driven by Lévy noise are known to avoid the saddle (Lucarini et al. 2022), our results challenge the generic role of saddles as gateways of noise-induced transitions also under Gaussian noise. Saddle avoidance limits the classical applicability of FW theory for predicting most probable transition paths in multiscale systems, since the regime of weak but finite noise will often apply to rare events of relevance. Our results in chapter 3 showed that this can become relevant also for climate applications, specifically for the multiscale and stochastically forced thermohaline circulation. In the context of high-impact low-likelihood events, understanding where in state space transitions will likely occur is crucial to assess a system's resilience to random fluctuations.

Chapter 5

Predictability and chaos in a conceptual climate model

A study in five dimensions

5.1 Stommel-L84: A conceptual ocean-atmosphere model 106

5.1.1 Stommel's box model 106

5.1.2 The Lorenz '84 model 107

5.1.3 Coupled model 107

5.2 Stability landscape of the AMOC 109

5.2.1 Quasipotential of the ocean 109

5.2.2 Bifurcation diagram and moving equilibria of the coupled model 111

5.3 Limits to predictability 113

5.3.1 Bistable regime: Chaotic saddle and fractal basin boundary . . . 113

5.3.2 Monostable regime: Long chaotic transients 116

5.4 Overshoot tipping: Safe operating space? 118

5.5 Discussion & conclusion 120

Parts of this chapter are based on Mehling et al. (2024).

Introduction

In the box model study conducted in chapter 3, we investigated the response of the thermohaline circulation to stochastic freshwater perturbations. We may interpret the noise as a parameterization of fast, unresolved processes, such as chaotic weather variability and variable meltwater input driving the net freshwater flux at the sea surface. As discussed in section 2.1.9, this idea underlies the stochastic climate model proposed by Hasselmann (1976) and is mathematically justified by homogenization theory in the limit of infinite timescale separation (Kelly and Melbourne 2017).

However, in reality the timescales of different earth system components and modes of variability are not perfectly separated. Outside the limit of infinite timescale separation, it can be more appropriate to model the fast component as bounded chaotic forcing (Ashwin et al. 2025). Moreover, the ocean and atmosphere interact in both directions. To address these points, we now consider a conceptual ocean-atmosphere model in which we explicitly resolve the chaotic variability. This brings us from stochastic dynamics to the realm of deterministic chaos, opening an alternative mathematical viewpoint on the problem of AMOC metastability (see chapter 2.2.4).

The model consists of a slow, bistable ocean component given by Stommel's seminal box model (Stommel 1961) and a fast chaotic atmosphere represented by the reduced-order model of the atmospheric circulation proposed by Lorenz (1984) (hereafter L84). Gottwald (2021) combined the Stommel and L84 models with a stochastic sea ice model to simulate Dansgaard-Oeschger events. Other recent studies (Ashwin and Newman 2021; Axelsen et al. 2024; Ashwin et al. 2025; De Melo Viríssimo et al. 2024) have coupled the Stommel model to L84 or the famous Lorenz system (Lorenz 1963), which is structurally similar to the L84 model. These setups offer a very useful conceptualization of key ingredients of the real Earth system: multistability, chaos, interacting components, and multiscale dynamics.

We are primarily interested in how the bidirectional coupling between the ocean and atmosphere influences the stability landscape in which the slow dynamics evolve – and what this means for the system's transient behavior as well as the predictability of the asymptotic state. As a point of reference, we first quantify the stability landscape of the uncoupled Stommel model by explicitly calculating its global quasipotential. In the coupled system, we use the edge tracking algorithm to compute chaotic Melancholia states for different background freshwater forcing, which allows us to construct the bifurcation diagram of the full system. Quantifying the dynamical properties of the Melancholia state reveals important insights into the structure of the basin boundary. Furthermore, we investigate the transient behavior beyond a boundary crisis. Lastly, we consider the problem of forcing overshoots, where a bifurcation point is crossed temporarily.

The findings in this chapter illustrate two points: fundamental limits to climate predictability on the one hand, and a route to comprehending complex dynamics on the other. While they pertain to a conceptual model, they lay a foundation for understanding chapter 6, where we apply the same approach to a fully coupled, intermediate-complexity climate model.

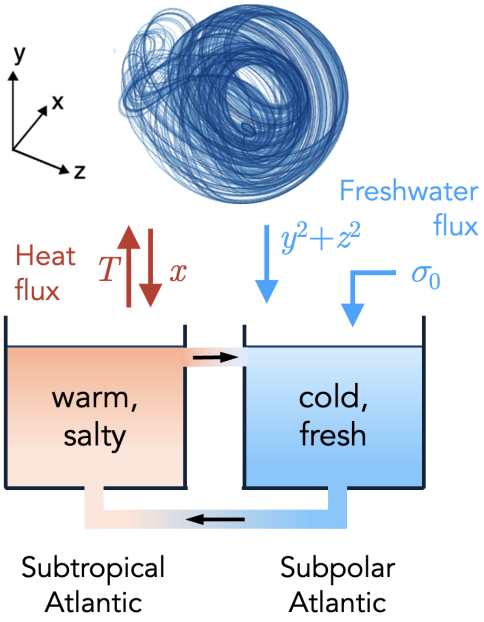

Figure 5.1 | **Schematic of the coupled Stommel-L84 model.** The two boxes of the Stommel model are coupled to the chaotic L84 model via heat and freshwater fluxes. The black arrows between the ocean boxes indicate a flow with $\Psi > 0$. Adapted from Mehling et al. (2024), with modifications.

5.1 Stommel-L84: A conceptual ocean-atmosphere model

5.1.1 Stommel's box model

The Stommel model consists of two well-mixed boxes representing the subtropical Atlantic and subpolar North Atlantic, respectively (Fig. 5.1). The system is driven by atmospheric heat and freshwater fluxes, and the dynamical state is described by the temperature and salinity in each box. The evolution equations can be formulated in terms of the meridional temperature difference $T := T_e - T_p$ and the meridional salinity difference $S := S_e - S_p$, where the subscripts e and p refer to the equatorial and polar box, respectively. As in the case of the Rooth model, the flow strength Ψ of the overturning circulation is assumed proportional to the density gradient (see Eq. (3.1)). In the dimensionless units chosen here, this simplifies to $\Psi = T - S$.

Following the version derived by Cessi (1994) and the notation of Gottwald (2021), we write the Stommel model as

$$\begin{aligned} \dot{T} &= -\frac{1}{\varepsilon_a}(T - \theta) - T - \mu|\Psi|T \\ \dot{S} &= \sigma_0 - S - \mu|\Psi|S, \end{aligned} \tag{5.1}$$

where $\varepsilon_a = t_r/t_d$ is the ratio between the timescale t_r of Newtonian temperature relaxation and the diffusive timescale t_d of the ocean. The ratio between the advective (t_a) and diffusive timescales of the ocean is defined as $\mu = t_a/t_d$. The meridional temperature gradient relaxes linearly to the reference value θ, and σ_0 denotes the freshwater flux from the subtropical into the subpolar box. The advective terms are proportional to the absolute value of the AMOC strength Ψ, where $\Psi > 0$ represents a thermally driven circulation like in today's climate, and $\Psi < 0$ represents a salinity-driven AMOC with a reversed circulation.

The freshwater flux σ_0 is usually considered the main control parameter that determines the stability landscape of the model. The bifurcation diagram with respect to σ_0 exhibits a bistable regime bounded by two saddle-node bifurcations (Fig. 5.2). At lower

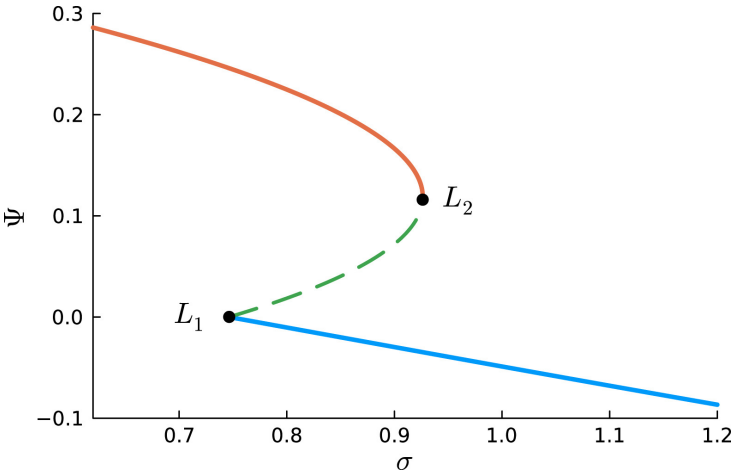

Figure 5.2 | **Bifurcation diagram of the Stommel model** as a function of the freshwater flux σ_0, showing the stable ON branch (orange), stable OFF branch (blue), and unstable saddle branch (green). The two saddle-node bifurcation points are marked by L_1 and L_2. Reproduced from Mehling et al. (2024).

northern freshwater forcing, only a stable fixed point with $\Psi > 0$ (ON) exists. As σ_0 is increased, a saddle point and a second stable fixed point (OFF) with $\Psi < 0$ emerges at $\sigma_{0,c1}$. At $\sigma_{0,c2} > \sigma_{0,c1}$, the saddle point collides with the ON state, leaving the OFF state as the only attractor.

5.1.2 The Lorenz '84 model

The conceptual model of Lorenz (1984) for the atmospheric general circulation (hereafter L84) consists of three coupled ordinary differential equations,

$$\begin{aligned} \dot{x} &= -(y^2 + z^2) - a\,(x - F) \\ \dot{y} &= xy - bxz - (y - G) \\ \dot{z} &= bxy + xz - z\,. \end{aligned} \tag{5.2}$$

Here x represents the strength of the mid-latitude westerly winds, assumed proportional to the meridional density gradient. The variables y and z mimic the phases of large-scale eddies which transport heat polewards at a rate $\sim y^2 + z^2$. The time unit in these equations is the damping timescale of the eddies, taken to be 5 days (Lorenz 1984). The total energy of the system is given by $E = (x^2 + y^2 + z^2)/2$. The model is externally controlled by the meridional and zonal temperature gradients F and G, respectively.

Depending on the choice of parameter values, the L84 model exhibits rich dynamical behavior with a complex bifurcation structure (Wang et al. 2014; Broer et al. 2002). Here we consider the original values $a = 0.25$ and $b = 4$ in Lorenz (1984) and set $G = 1$, yielding a system with a single chaotic attractor (Fig. 5.3).

5.1.3 Coupled model

We can couple the two models via the heat and freshwater fluxes between ocean and atmosphere. Following Gottwald (2021) and Roebber (1995), we make the oceanic relaxation temperature θ dependent upon the anomalies in the westerly winds x. Additionally, we couple the freshwater flux to the eddy transport, $\sigma_0 \propto \Delta := y^2 + z^2$. In the other direction, it is reasonable that the magnitude of the westerly winds depends on the oceanic meridional temperature gradient T.

While the timescale of the L84 model is determined by the eddy damping timescale of around five days, the timescales of the Stommel model are set by the parameters t_d, t_a and

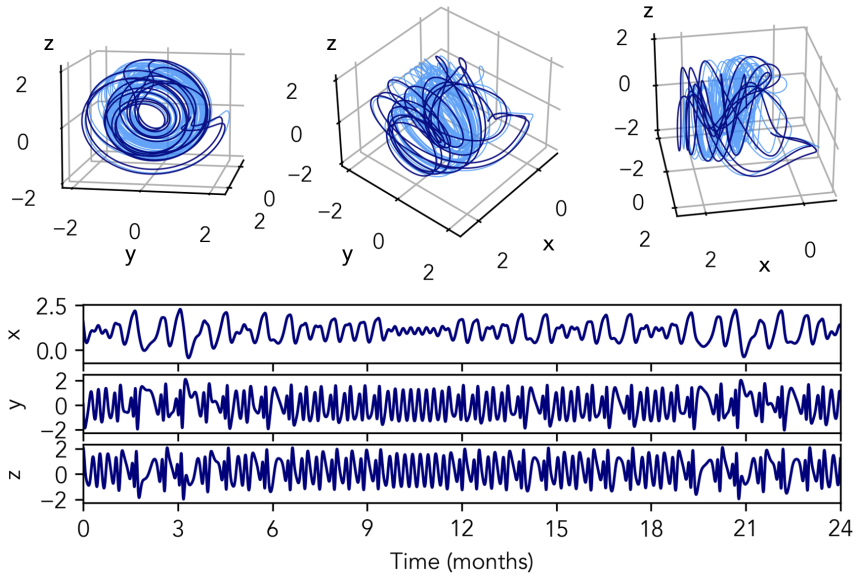

Figure 5.3 | **Chaotic attractor of the Lorenz '84 model.** Top: Phase space view of trajectories on the attractor from different viewing angles. Bottom: Time series of each component over a 2-year interval.

t_r. To reflect the timescale separation between the oceanic and atmospheric dynamics, we introduce a timescale parameter ε_f. Altogether, the coupled equations read

$$\begin{aligned}
\varepsilon_f \dot{x} &= -\Delta - a\,(x - F_0 - F_1 T) \\
\varepsilon_f \dot{y} &= xy - bxz - (y - G_0) \\
\varepsilon_f \dot{z} &= bxy + xz - z \\
\dot{T} &= -\frac{1}{\varepsilon_a}(T - T_{\text{surf}}) - T - \mu|S - T|T \\
\dot{S} &= S_{\text{surf}} - S - \mu|S - T|S\,,
\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}
T_{\text{surf}} &= \theta_0 + \theta_1 \frac{x - \bar{x}}{\sqrt{\varepsilon_f}}\,, \\
S_{\text{surf}} &= \sigma_0 + \sigma_1 \frac{\Delta - \bar{\Delta}}{\sqrt{\varepsilon_f}}\,.
\end{aligned}$$

Equation (5.3) presents a fast-slow system formulated in slow time, i.e. on the timescale of the slow component (see section 2.1.9). With $\varepsilon_f = 3 \times 10^{-4}$, we get an oceanic timescale of around 100 years and an atmospheric timescale on the order of 10 days (hence one time unit corresponds to about one century). The standard parameter values of the coupled model are summarized in Table 5.1.

Even though the Stommel-L84 model consists "only" of five ordinary differential equations, the large timescale separation adds computational cost because we must integrate the system on the fast timescale of the atmosphere. This requires a small time step relative to the oceanic dynamics of interest. Here we use a fifth-order Runge-Kutta scheme with a fixed time step of $\Delta t = 7.5 \times 10^{-6}$, which ensures that the chaotic dynamics are adequately resolved.

Parameter	Value	Physical interpretation
a	0.25	–
b	4	–
F_0	8	Atmospheric meridional temperature gradient
G_0	1	Atmospheric zonal temperature gradient
ε_a	0.34	Ratio between relaxation and diffusive timescales
μ	7.5	Ratio between advective and diffusive timescales
θ_0	1	Atmospheric temperature gradient
σ_0	is varied	Freshwater flux
θ_1	0.0195	Ocean-to-air temperature coupling strength
σ_1	0.00934	Ocean-to-air freshwater coupling strength
$\bar{x}$	1.0147	Time-mean of x (uncoupled L84 model)
$\bar{\Delta}$	1.7463	Time-mean of Δ (uncoupled L84 model)
F_1	0.1	Air-to-ocean temperature coupling strength
ε_f	3×10^{-4}	Ratio between atmospheric and oceanic timescale

Table 5.1 | Parameters of the coupled Stommel-L84 model.

Note that the bi-directional coupling generally complicates the derivation of a homogenized stochastic differential equation for the ocean component (Engel et al. 2021). Here, the additivity of the coupling between the Stommel and L84 models yields a relatively straightforward result, as shown in Appendix B of Mehling et al. (2024).

5.2 Stability landscape of the AMOC

Due to the large timescale separation and weak coupling between the ocean and atmosphere, we expect that the bistability of the Stommel model is retained also in the coupled model. In terms of a stability landscape, we might imagine that the coarse-grained structure resembles a double-well quasipotential, with a fine structure imposed by the L84 dynamics.

5.2.1 Quasipotential of the ocean

First, let us consider the stability landscape of the uncoupled Stommel model. Even though the stability properties of the model are well known (Stommel 1961), to our knowledge an explicit calculation of its quasipotential has not been published.

Let us consider a freshwater flux of $\sigma_0 = 0.9$ located in the bistable regime (see Fig. 5.2). As in chapter 4, we use the `OLIM4VAD` algorithm (see section 2.2.4) to compute the Freidlin-Wentzell quasipotentials with respect to the ON and OFF attractors of the AMOC. We assume an identity covariance matrix (i.e. isotropic noise).

The resulting quasipotentials are shown in Fig. 5.4. They are continuous as required but not everywhere differentiable (Cameron 2012). As per definition, each quasipotential is zero at the corresponding reference attractor, marking the global minimum. In both cases (V_{ON} and V_{OFF}), the attractors are connected via a channel of low quasipotential that leads across the saddle M. Comparing V_{OFF} with the drift field, we clearly see the nongradient nature of the system: the flow lines do not lead perpendicular to the

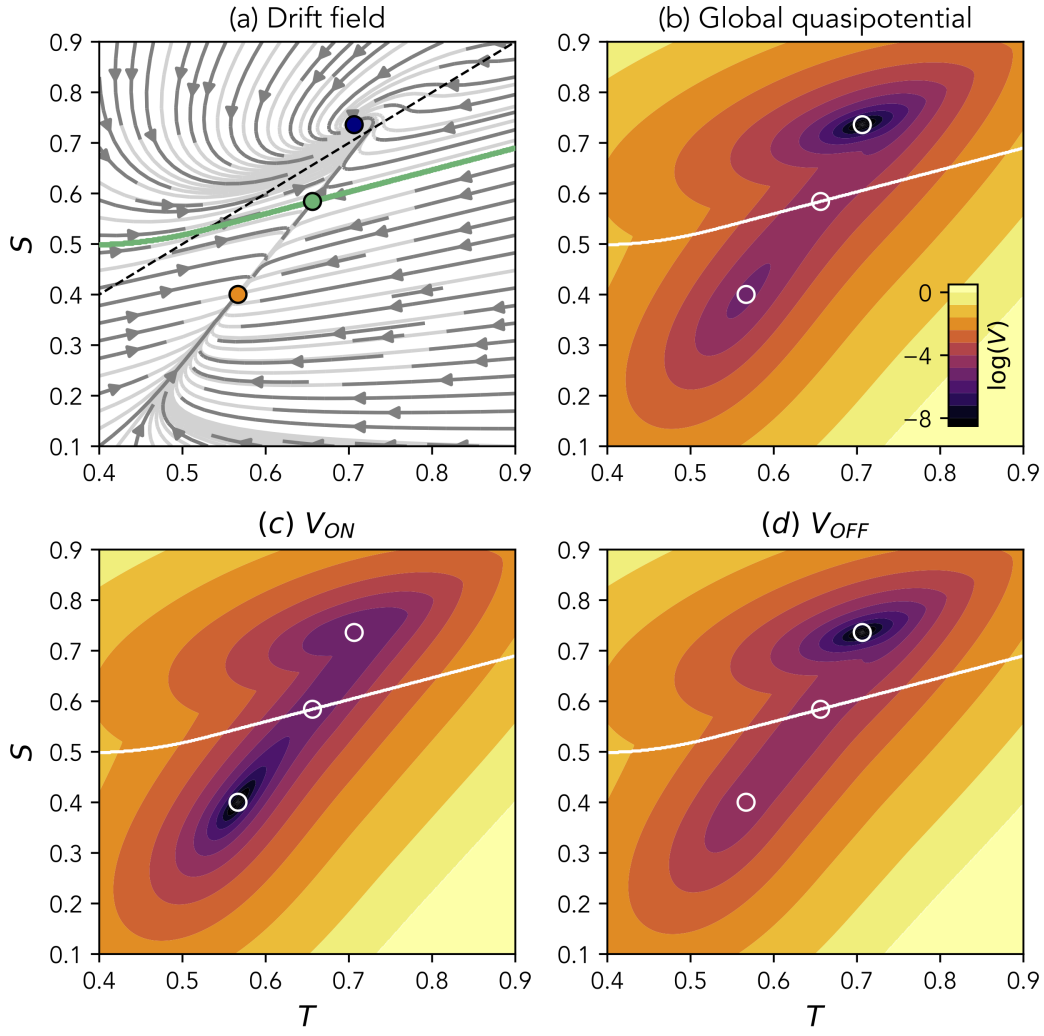

Figure 5.4 | **Quasipotentials of the Stommel model** for $\sigma_0 = 0.9$ and isotropic noise. (a) Drift field, indicating the locations of the ON (orange), OFF (blue) and M states (green). The basin boundary (green line) and the $q = 0$ line (black dashed) are shown. (b) Global quasipotential V_{global}, constructed from the local quasipotentials (c) V_{ON} and (d) V_{OFF}. Quasipotentials are plotted on a logarithmic color scale where black corresponds to $V = 0$ and brighter colors indicate higher values.

isolines of the quasipotential but spiral inwards towards the OFF state, indiating a nonzero rotational component (see Eq. (2.21)).

Since there is only one saddle point in the system, we can stitch the local quasipotentials V_{ON} and V_{OFF} together to obtain the global quasipotential (see section 2.1.6). To do this, we shift V_{ON} by a positive constant such that its value at M is identical to the value of V_{OFF} at M. Then, at each point $\boldsymbol{x} = (T, S)$ in phase space, we take the minimum, $\min(V_{\text{ON}}(\boldsymbol{x}), V_{\text{OFF}}(\boldsymbol{x}))$, as the global quasipotential value (Fig. 5.5). The global quasipotential V_{global} has the doublewell shape that characterizes the bistability of the system (Figs. 5.4 and 5.5). We can immediately deduce interesting dynamical features from looking at the quasipotential. For example, since $V_{\text{global}}(\boldsymbol{x}_{\text{ON}}) > V_{\text{global}}(\boldsymbol{x}_{\text{OFF}})$, we find that at $\sigma_0 = 0.9$ the ON state is less stable than the OFF state. We further observe that the global quasipotential has a relatively low curvature around the ON state compared with the OFF state, which relates to the restoring rate near each attractor. Most probable transition paths and transition rates can be directly derived from the quasipotential (as

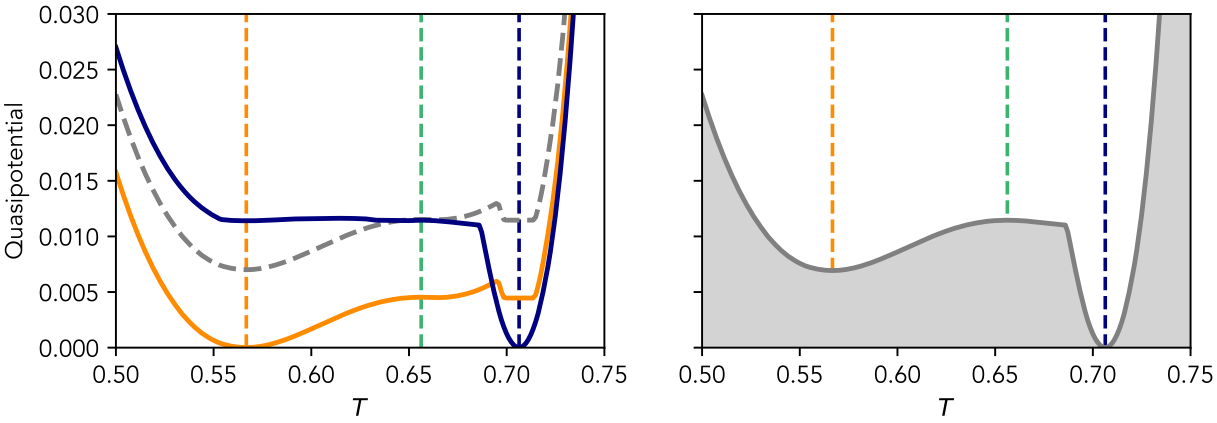

Figure 5.5 | **Construction of the global quasipotential** from the local quasipotentials V_{ON} (orange) and V_{OFF} (blue). Here the quasipotentials are plotted along the quadratic curve that intersects with the ON, OFF, and M states, whose location is indicated (vertical dashed lines). Left: V_{ON} is shifted upwards (gray curve) to align with V_{OFF} at the saddle. Right: Cross section of the resulting global quasipotential.

outlined in sections 2.1.6 and 2.2.3; we do not pursue this here).

5.2.2 Bifurcation diagram and moving equilibria of the coupled model

Now that we have quantified the stability landscape of the Stommel model, let us turn to the coupled Stommel-L84 model. Due to the coupling, the attractors of the system are not fixed points anymore but chaotic sets. Similarly, we expect that the saddle point turns into a chaotic saddle. Since the saddle point of the Stommel model is located on the basin boundary, we may anticipate that this chaotic saddle is a Melancholia state, in which case it should be possible to compute it using the edge tracking algorithm (see section 2.2.1).

In the edge tracking procedure, we must define a metric to quantify the divergence of the two parallel trajectories in the phase space. While we simply used the Euclidean distance in chapter 3, this choice is not appropriate here due to the different magnitudes and timescales of variability between the atmosphere and ocean. Rather, the metric should capture the divergence in the oceanic state without being dominated by the divergence of trajectories on the chaotic attractor of the atmospheric dynamics. To ensure this, we take the Euclidean distance after rescaling the atmospheric variables by a scaling factor $\xi < 1$ (we set $\xi = 1/5000$). It can be viewed as a hyperparameter of the edge tracking method and does not alter the dynamics of the system. With the rebisection parameters set to $\varepsilon_1 = 0.0025$ and $\varepsilon_2 = 0.004$, we run the edge tracking algorithm for 2000 iterations, generating a pseudo-trajectory of around 50 time units (≈ 5000 years). As anticipated, the pseudo-trajectory converges to a localized region in phase space after a transient period. Removing the initial transient, the pseudo-trajectory yields an estimate of the dynamics on the Melancholia state.

Applying the edge tracking algorithm for various values of the imposed freshwater flux σ_0, we can compute the unstable branch of the bifurcation diagram for the coupled model (Fig. 5.6; see also Fig. 8 by Mehling et al. (2024)). The stable branches are obtained from long simulations run for each σ_0 value, starting in either basin of attraction.

We find that the bifurcation diagram projected onto Ψ largely resembles that of the

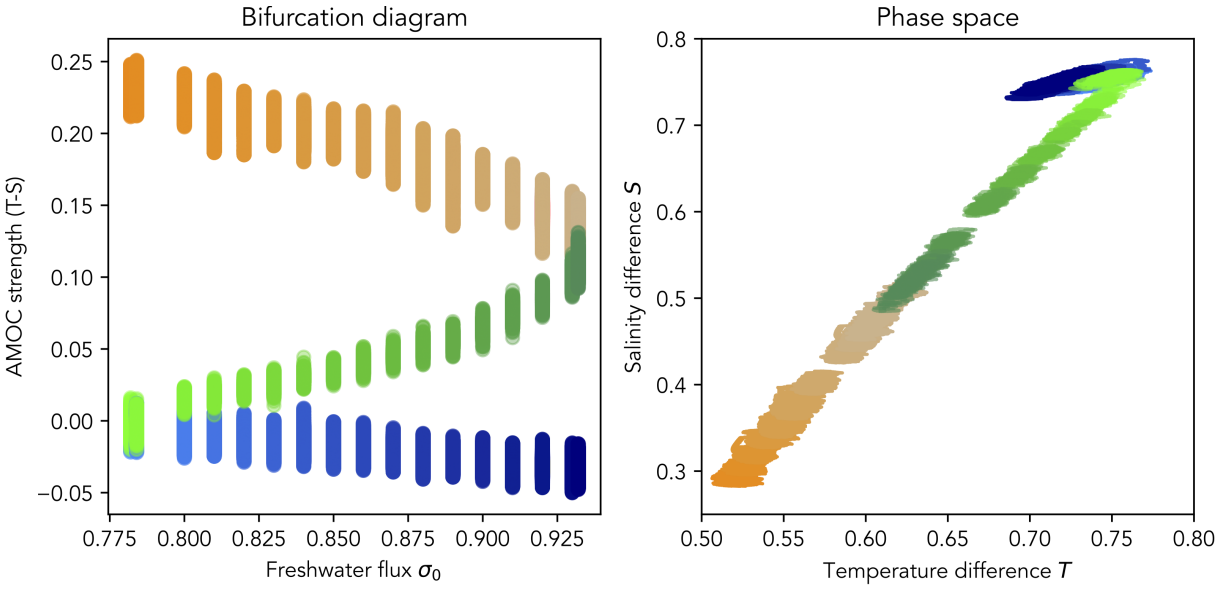

Figure 5.6 | **Chaotic attractors and saddle** of the Stommel-L84 model as a function of the freshwater flux σ_0, shown as a bifurcation diagram (left) and in phase space (right). The colors represent the ON (orange) and OFF (blue) attractors as well as the saddle (green) as they move in the parameter/phase space.

uncoupled Stommel model (Fig. 5.2). Since instead of fixed points we now have chaotic invariant sets, the branches have a finite width in Ψ representing the size of the invariant set. As a consequence, the bistable regime is effectively narrower compared to the uncoupled case. The bifurcation points correspond to boundary crises (Term 2.25) where the Melancholia state collides with one of the attractors. In the case of the OFF state, this happens at a significantly larger value of σ_0 compared to the uncoupled saddle-node bifurcation point L_1 (as we estimate below).

Taking a phase space perspective, we can trace how the equilibria move in the T–S plane as a function of the freshwater forcing (Figs. 5.6 and 5.7). While the location of the OFF state is relatively independent of σ_0, the ON and M states move towards each other as σ_0 is increased. This illustrates the idea of a *tipping window* as proposed by (Ashwin et al. 2025): a window of σ_0 values where the unstable M state "overlaps" with the former attractor at each of the boundary crises. While the tipping window strictly belongs to the monostable regime, the ghost state formed by the merged chaotic sets supports long chaotic transients, such that the dynamics may appear bistable for finite simulation times (see sections 2.1.7, 5.3.2 and 6.5.3). Hence, edge tracking still works despite being in the monostable regime. This case relates to the original application of the edge tracking algorithm in fluid dynamics, where the edge state between a stable laminar flow and an ultralong transient of turbulent flow is computed (Skufca et al. 2006).

Inspecting a three-dimensional projection of the phase space (Fig. 5.7), we observe that the chaotic ON, OFF, and M states take up a similar volume in the reduced phase space. The M state features spikes of higher atmospheric energy E compared to the attractors, which could be a signature of the global instability.

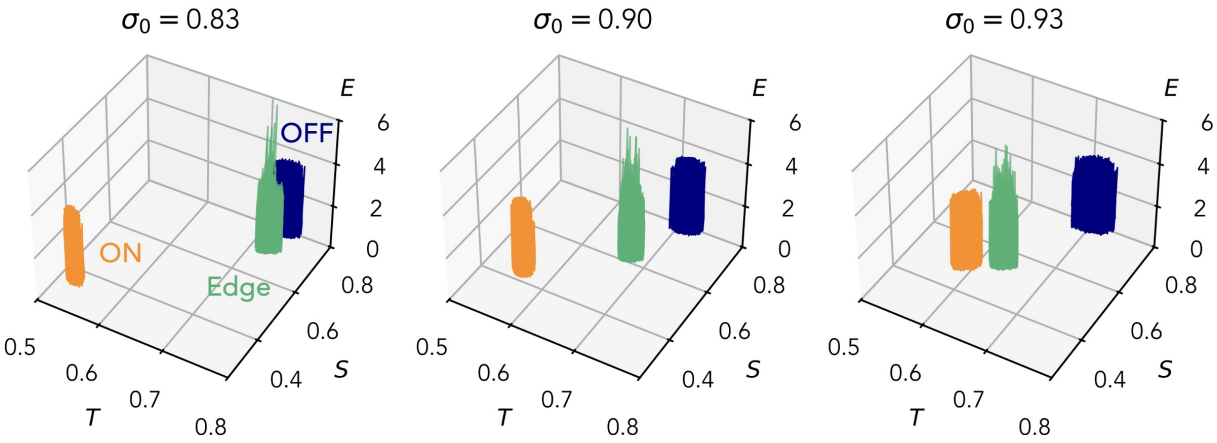

Figure 5.7 | **Phase space view of the chaotic invariant sets** (ON, OFF attractors and Melancholia state) for selected magnitudes of freshwater forcing. The vertical axis shows the energy E of the atmosphere.

5.3 Limits to predictability

5.3.1 Bistable regime: Chaotic saddle and fractal basin boundary

Let us now focus on the bistable regime, where the M state separates the ON and OFF states. Since the M state is a chaotic saddle, we can characterize its dynamics in terms of its mean lifetime and Lyapunov spectrum (Ott (2002); see sections 2.1.2 and 2.1.7). Both can be obtained by sampling a large number of initial conditions from the pseudo-trajectory. Since a state initialized on the pseudo-trajectory is very close to but not precisely on the actual saddle, its trajectory will remain in the vicinity of the saddle for some time before converging to one of the attractors.

Lifetime of the Melancholia state. Let Ω be the bounding box (in five dimensions) containing the pseudo-trajectory approximating the M state. Following Eq. (2.27) (Ott 2002), we expect the number of ensemble members remaining within Ω to decay exponentially. We take $N_0 = 600$ initial conditions in Ω and record the times at which they escape Ω. The exponential scaling is not expected to hold for small and large t due to the non-uniform initial distribution in Ω and the finite size of the ensemble, respectively. However, we can obtain the mean lifetime $\langle\tau\rangle$ via a least-squares fit of the slope of $\log N(t)/N_0$ (see Eq. (2.27)) against t for intermediate values of t.

We perform this analysis for several values of σ_0 in the bistable regime. In close agreement with the theory, the number of ensemble members remaining in Ω shows a clear exponential decay (Fig. 5.8a). The mean lifetime $\langle\tau\rangle$ ranges from roughly 60 to 250 years depending on σ_0 (Fig. 5.8b). Approaching the boundary crisis where the ON state is annihilated, trajectories tend to spend longer times near the M state. This asymmetry of the mean lifetime with respect to the distance to each bifurcation point might be relatable to the asymmetry in the double-well global quasipotential landscape.

Lyapunov spectrum. The mean lifetime characterizes the global instability of the Melancholia state associated with the bistability of the AMOC, i.e. the slow "climate" component (Bódai and Lucarini 2020). Let us now turn to the local instability of the chaotic

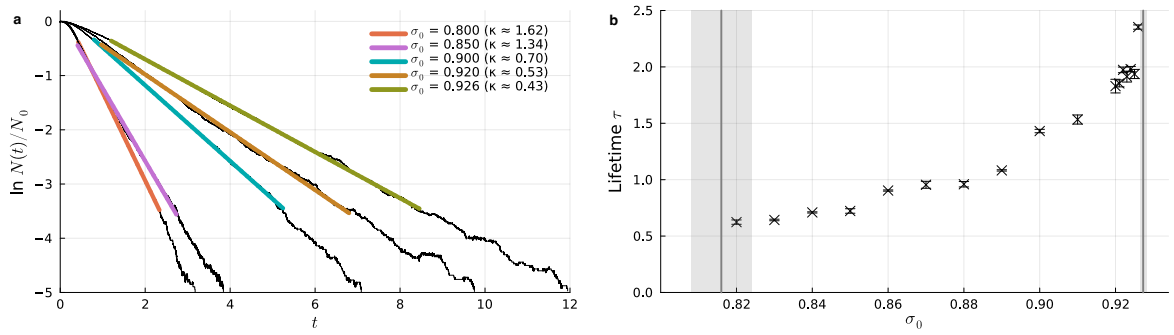

Figure 5.8 | **Lifetime of the chaotic saddle.** (a) Fraction of trajectories $N(t)/N_0$ within the saddle bounding box Ω as a function of time for different values of σ_0. Colored lines indicate linear fits in lin-log space, whose slope is $-\kappa = -1/\langle\tau\rangle$, in the range $N/N_0 \in [0.03, 0.7]$. (b) Mean lifetime $\langle\tau\rangle$ obtained from the fits in panel (a). Error bars reflect the uncertainty in $\langle\tau\rangle$ with respect to reasonable changes in the fitting range in panel (a). The vertical lines and shaded areas indicate estimates of the critical values $\sigma_{0,c}$ plus/minus one standard deviation. Reproduced from Mehling et al. (2024).

"weather" dynamics on the M state, captured by the Lyapunov exponents of the M state. To compute the Lyapunov spectrum, we apply the method of evolving an orthonormal perturbation matrix with iterative Gram-Schmidt orthonormalization (Benettin et al. (1980); see section 2.1.2) as implemented by Datseris (2018) in the `DynamicalSystems.jl` software library. We choose 2000 initial conditions on the pseudo-trajectory that have the longest lifetimes in Ω and then average over the Lyapunov spectra of a subset of these individual trajectories while they remain in Ω. For consistency, we tested the same method for the attractors, where averaging over 2000 finite-time Lyapunov exponents calculated over a similar duration ($t = 7.5$) yielded a comparable result to letting the algorithm converge along one long trajectory.

The Lyapunov spectra $\{\gamma_i\}$, $i = 1, \ldots, 5$ for the saddle and each of the two attractors are given in Tab. 5.2 for $\sigma_0 = 0.9$. For the attractors, we expect one positive Lyapunov exponent characterizing the chaotic instability on the attracting set, a zero exponent associated with the motion along the attractor, and otherwise negative exponents encapsulating the convergence of initial conditions to the attractor. Indeed, we find a large, positive maximum Lyapunov exponent (MLE) γ_1, while γ_2 is very close to zero and all remaining Lyapunov exponents are negative. Similarly, the saddle possesses one positive exponent (γ_1), one exponent close to zero (γ_2) and three negative exponents. Note that obtaining a good numerical estimate for the vanishing Lyapunov exponent of a saddle is more difficult than for an attractor. This is because the trajectories considered for the

	ON	OFF	M state
γ_1	535 ± 2	531 ± 2	538 ± 2
γ_2	-0.04 ± 0.02	-0.015 ± 0.017	0.13 ± 0.07
γ_3	-0.67 ± 0.02	-2.65 ± 0.03	-0.56 ± 0.14
γ_4	-8.11 ± 0.13	-2.91 ± 0.04	-6.0 ± 0.3
γ_5	-1230 ± 2	-1155 ± 2	-1182 ± 4

Table 5.2 | Lyapunov exponents for the two attractors and the M state for $\sigma_0 = 0.9$. Uncertainty estimates are 95% confidence intervals generated from bootstrap resampling of 100 trajectories each from the 2000-member ensemble.

evaluation of the Lyapunov exponents have a finite lifetime, as they unavoidably end up veering towards one of the competing attractors, realizing a stretching along the direction of the flow. We have verified that the estimate of γ_2 converges towards zero when we compute it based on longer-lived trajectories, whereby the stretching along the flow is reduced (see Fig. S1 in the Supplementary Information of Mehling et al. (2024)).

The magnitude of the Lyapunov exponents varies greatly for each state, with $|\gamma_1|$ and $|\gamma_5|$ being much larger than the other exponents. We can link this to the timescale separation ε_f between the ocean and atmosphere. Recomputing the Lyapunov spectra for varying ε_f shows that $|\gamma_1|$ and $|\gamma_5|$ scale with $\gamma_{i,\mathrm{L84}}/\varepsilon_f$, where $\gamma_{i,\mathrm{L84}}$ is the corresponding Lyapunov vector of the uncoupled L84 model, while $|\gamma_2|$, $|\gamma_3|$ and $|\gamma_4|$ are largely independent of ε_f (see Mehling et al. (2024)). Thus, owing to the weak air-sea coupling, we can associate the large Lyapunov exponents (in absolute value) with the atmospheric dynamics; particularly, the MLE is related to the predictability horizon of the "weather" limited by the sensitive dependence on the initial condition (predictability of the first kind; see section 1.3).

The results shown here exemplarily for $\sigma_0 = 0.9$ are qualitatively representative for all values of σ_0 which we have tested within the bistable regime, and there is no clear dependence of γ_1 or γ_2 on the freshwater flux. The MLE γ_1 is very similar across different σ_0 (mean: 536 ± 4) and the second Lyapunov exponent is very small ($\gamma_2 < 0.3 \ll \gamma_1$) for all σ_0.

Fractal dimension. Now that we have quantified the lifetime and maximum Lyapunov exponent of the M state, we can use the Kantz-Grassberger formula (Kantz and Grassberger 1985) to calculate the dimension of the stable set of the M state, i.e. the basin boundary (Hunt et al. 1996; Sweet and Ott 2000; Hsu et al. 1988). In fact, since the escape rate from the saddle $\kappa = 1/\langle\tau\rangle$ is smaller than the MLE γ_1, we can use Eq. (2.29) derived by Bódai and Lucarini (2020).

The basin boundary dimension D_b is between 4.997 and 4.999 for different values of σ_0. In other words, the basin boundary is an almost space-filling set: its dimension is less than 0.1% smaller than the dimensionality of the phase space ($D = 5$). As Eq. (2.29) tells us, this situation occurs if the fast chaotic instability on the chaotic saddle is large compared to the escape rate associated with the global instability.

What does this result mean for the predictability of the asymptotic state of the AMOC? A fractal basin boundary implies the existence of an extended region in phase space – a "gray zone" – where initial conditions belonging to opposite basins of attraction lie arbitrarily close to each other. In this region, the slightest perturbation can fundamentally alter the fate of the AMOC, since it can change whether the AMOC will converge to the ON or OFF state, respectively.

We can express the limited predictability in terms of the *final state sensitivity*, based on the uncertainty exponent $\alpha = D - D_b \approx 0.002$ (McDonald et al. 1985). According to McDonald et al. (1985), the fraction f of a bounded phase space volume Ω in which the asymptotic state is uncertain scales like

$$f \sim \varepsilon^{\alpha}\,, \tag{5.4}$$

where $0 < \varepsilon$ is the precision at which an initial condition is known. Hence, if we halve the error in the initial condition, we improve the predictability by $1 - 0.5^{\alpha}$, which in this case is less than 1%. This means that the asymptotic state depends sensitively on the

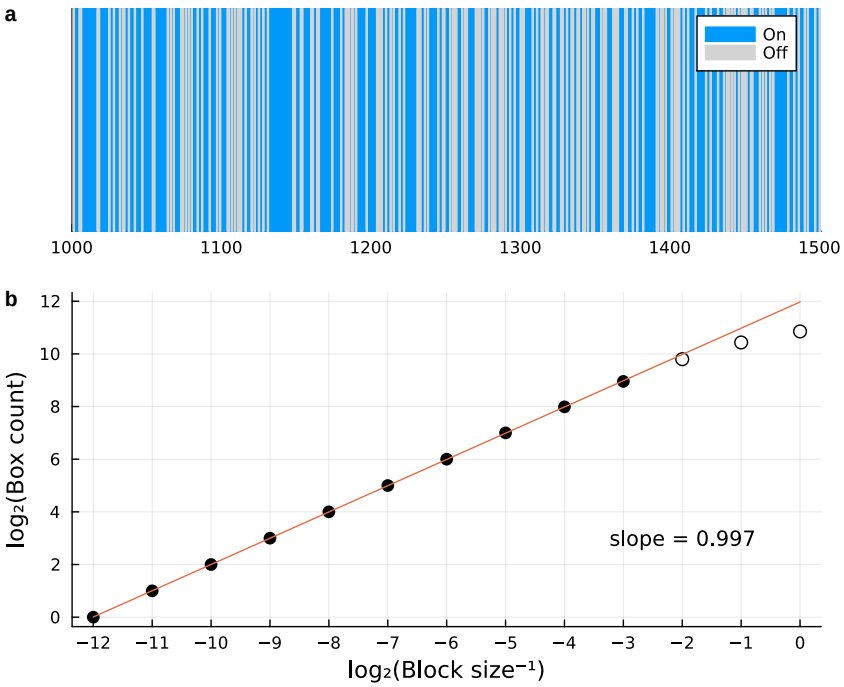

Figure 5.9 | **Fractal basin boundary** along a sample cross-boundary section for $\sigma_0 = 0.9$. (a) Final AMOC state (color-coded) of the forward integration of 512 evenly spaced initial conditions along a one-dimensional phase space transect across the basin boundary. (b) Computation of the box-counting dimension from 2^{13} initial conditions for the same boundary section. Unfilled markers are not taken into account for the linear fit. Reproduced from Mehling et al. (2024).

initial condition near the basin boundary, and the fate of the AMOC becomes essentially unpredictable there.

We can alternatively estimate the fractal dimension of the basin boundary via the box counting dimension, as performed by Lucarini and Bódai (2017). To this end, we sample evenly spaced initial conditions along a line that intersects with the basin boundary, and determine the attractor to which each initial condition eventually converges. Fig. 5.9a shows a sample plot of the outcomes of 2^{13} initial conditions, where the endpoints of the sampling interval are within $\delta_1 = 10^{-3}$ of the basin boundary (measured in terms of the rescaled Euclidean distance defined above) but away from the M state. The result looks like a fractal bar code, indicating that the asymptotic state varies nonmonotonically along the transect. Repeating this procedure in different regions of phase space and then for each value of σ_0 considered previously, we find that the box-counting dimension is also very close to 1 (between 0.988 and 0.999) for all σ_0, in agreement with D_b.

5.3.2 Monostable regime: Long chaotic transients

Suppose we are on the ON state at $\sigma_0 = 0.9$ (Fig. 5.6). As σ_0 increases, the ON state and M state approach each other until they collide in a boundary crisis, marked by the critical threshold $\sigma_{0,c2}$. (At the other end of the bistable region, the OFF and M states collide at $\sigma_{0,c1}$.) What happens if we increase σ_0 beyond $\sigma_{0,c2}$?

As an example, we select a set of $N = 200$ initial conditions by sampling the ON state for $\sigma_0 = 0.926$ at timesteps every $\Delta t = 7.5$. This sampling interval is much larger

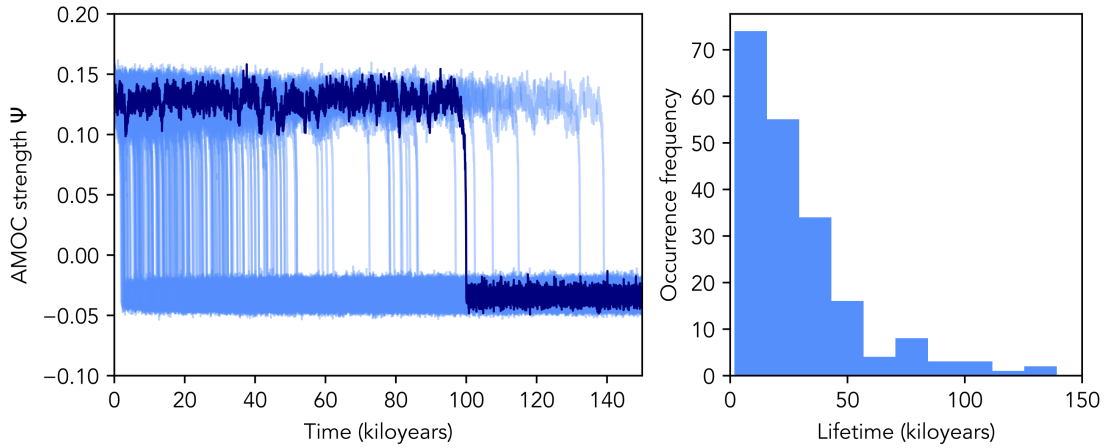

Figure 5.10 | **Long chaotic transients** for $\sigma_0 = 0.932$. Initial conditions were sampled from the ON attractor for $\sigma_0 = 0.926$ and were integrated forward after the parameter change to $\sigma_0 = 0.932$. Left: AMOC timeseries for all 200 ensemble members, highlighting one exemplary member. Right: Lifetime distribution of the transients.

than the Lyapunov timescale of the fast system (≈ 0.002 time units) and therefore we can expect the initial conditions to be uncorrelated. At the given value of σ_0 the ON state is stable. Then, we integrate trajectories from these initial conditions but with the freshwater parameter instantaneously set to $\sigma_0 = 0.932$, a value slightly larger than the bifurcation point L_2 of the uncoupled Stommel model, $\sigma(L_2) = 0.9263$. Due to the weak coupling, we may assume that this value also lies close to the bifurcation point of the coupled model.

The 200 trajectories exhibit transient chaos (Fig. 5.10). Eventually, all trajectories collapse to the OFF state, confirming that at $\sigma_0 = 0.932$ the system has passed a boundary crisis and entered a monostable regime. At the same time, we observe the signature of the "ghost" of the ON state, which causes trajectories to spend an extended time near the former ON state before abruptly transitioning to the OFF state (see also Axelsen et al. (2024)). The lifetimes of the transients are roughly exponentially distributed (as expected; see section 2.1.9 and Yorke and Yorke (1979)) with a mean lifetime of $\langle\tau\rangle \approx 20\,000$ years and reach up to 140,000 years (Fig. 5.10).

As demonstrated by Grebogi et al. (1986) (see Eq. (2.28) in section 2.1.7), the mean lifetime typically follows a power law as the system approaches the critical value of the control parameter $\sigma_{0,c}$. To verify this, we run initial condition ensembles for different values of σ_0, sampled from the ON state near the saddle-node bifurcation point L_2 of the uncoupled Stommel model and from the OFF state near L_1 (see Figs. 5.2 and 5.6).

Because $\langle\tau\rangle \to \infty$ as $\sigma_0 \to \sigma_{0,c}$, we need to choose a threshold time $T_{\max}$ after which we deem a trajectory "stable," even though it may still be a very long chaotic transient. This means that $\sigma_{0,c}$ would be biased if determined by direct simulation, as it would depend on the choice of $T_{\max}$. Therefore, we look to find a range of values for σ_0 with $\langle\tau\rangle$ sufficiently large for the values to be exponentially distributed, but smaller than $T_{\max}$. Using Eq. 2.28, we then determine $\sigma_{0,c}$ and γ within this region via a weighted least-squares fit.

Near both crisis points, we fing a clear power-law relationship between the distance to the threshold $|\sigma_0 - \sigma_{0,c}|$ and the mean lifetime of transients tracking the ghost state, as evidenced by the linear dependence in the log-log plots (Fig. 5.11). Here we only evaluate the range of σ_0 in which the lifetimes of an ensemble of $N = 100$ initial conditions

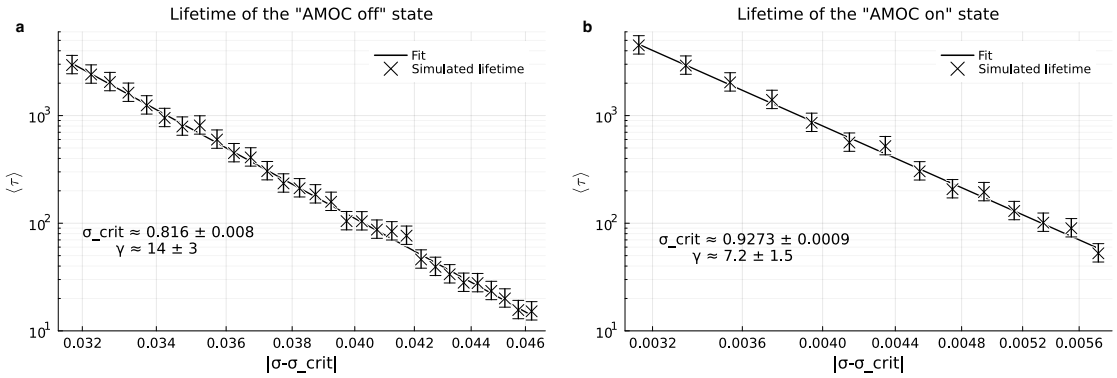

Figure 5.11 | **Power-law scaling of the mean lifetime** $\langle\tau\rangle$ of chaotic transients, corresponding 95% confidence intervals and power-law fit for different values of σ_0 for (a) the transient OFF state and (b) the transient ON state outside the regime of bistability. Note that all axes are logarithmic and that the subplots use different x axes. Reproduced from Mehling et al. (2024).

approximately follow an exponential distribution, and in which $\tau < T_{max} = 10^4$ for all trajectories. From the corresponding fits, we estimate the bifurcation points,

$$\sigma_{0,\mathrm{c1}} = 0.816 \pm 0.008 \text{ and } \sigma_{0,\mathrm{c2}} = 0.9273 \pm 0.0009\,,$$

where the subscripts c1 (c2) refer to the boundary crisis where the OFF (ON) state collides with the Melancholia state.

The critical exponents obtained from the fit are $\gamma_1 = 14 \pm 3$ below and $\gamma_2 = 7.2 \pm 1.5$ above the bistable region of σ_0. This implies that long transients can be observed over a wider range of parameter values near the ghost of the OFF state compared to that of the ON state (Ott 2002). For instance, the parameter range $|\sigma_0 - \sigma_{0,\mathrm{c}}|$ for which $\langle\tau\rangle > 100$ is 0.04 at $\sigma_{0,\mathrm{c1}}$ and 0.005 at $\sigma_{0,\mathrm{c2}}$. Both critical exponents are larger than those of many typical dynamical systems (Ott 2002).

5.4 Overshoot tipping: Safe operating space?

As the greenhouse gas concentrations keep rising and we are close to reaching 1.5 °C of global warming, the question of "safe overshoots" is gaining increasing attention (Ritchie et al. 2019; Ritchie et al. 2021; Schleussner et al. 2024): is it possible to temporarily exceed a critical threshold without triggering a critical transition?

Using the scaling law in Eq. (2.28) and the quantile function of the exponential distribution, we can calculate how far the "safe operating space" of the AMOC is extended beyond the genuinely bistable regime through the existence of long transients. In other words, up to which value of σ_0 will the desirable AMOC state (in our case, the ON state) continue to act like a stable attractor, even though the boundary crisis has been surpassed? The exponential distribution underlying the transient lifetime means that this safe operating space is inherently linked to a timescale of interest and an acceptable probability of survival on the desired AMOC state (or, equivalently, a tolerable probability of tipping to the undesired state).

The survival probability on the ON state as a function of $|\sigma_0 - \sigma_{0,\mathrm{c}}|$ and the chosen timescale is portrayed in Fig. 5.12. For example, overshooting to $\sigma_0 \approx 0.934$ would yield a 5% probability of obtaining a transient lifetime of less than 100 years (red marker in Fig.

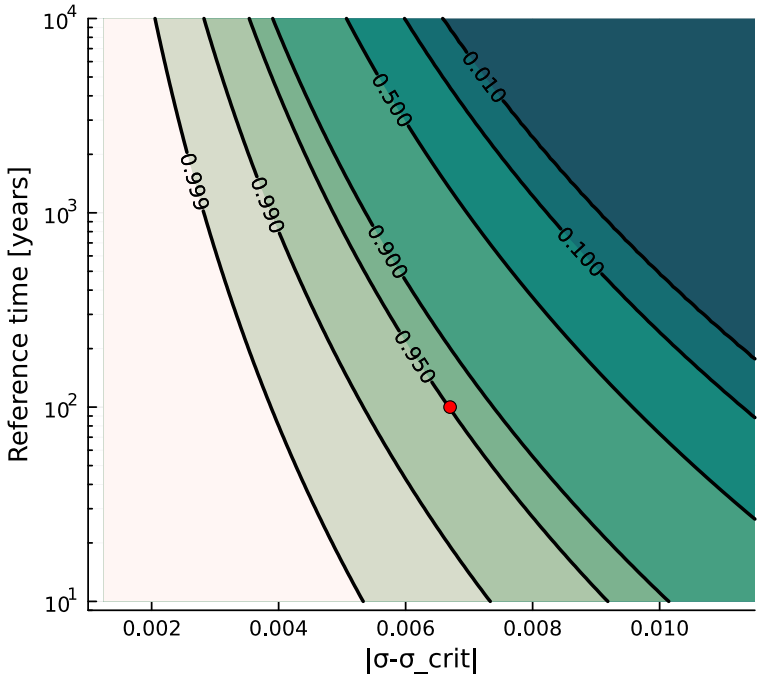

Figure 5.12 | **Survival probability** in the region of chaotic transients beyond the critical value $\sigma_{0,c2} \approx 0.927$ as a function of the distance to the critical value $|\sigma_0 - \sigma_{0,c}|$ and the reference time horizon. The red marker indicates $\sigma_0 \approx 0.934$ for which the transient tipping probability is limited to 5% on a timescale of 100 years. Reproduced from Mehling et al. (2024).

5.12). On the other hand, if the probability of a collapse should be limited to 0.1% within the same timescale, the freshwater forcing range is limited to no more than $\sigma_0 \approx 0.931$. In this way, if the control parameter is returned sufficiently quickly back to the bistable regime, a transition might be avoided despite overshooting the critical value (Ritchie et al. 2021).

The fact that there is a forcing range of nontrivial tipping probability implies that the safe operating space (a term coined by Rockström et al. (2009)) must be understood in a probabilistic sense. In this range, the asymptotic state can depend sensitively on the initial condition, which can give rise to ensemble splitting.

To illustrate this, we consider a step overshoot forcing where σ_0 is instantaneously increased to a value exceeding $\sigma_{0,c2}$ by $\Delta\sigma$ for a duration t_{over}, after which the forcing is instantaneously restored to its original value (within the bistable regime). We run this experiment for $t_{over} = 35$ time units and different values of $\Delta\sigma$, generating ensembles of 200 members each. The initial conditions of the ensemble members are sampled from the ON state at $\sigma_0 = 0.926$ as before.

For small values of $\Delta\sigma$, all ensemble members remain on the ON state, despite temporarily having entered the monostable regime where the ON state is annihilated (Fig. 5.13). For sufficiently large values of $\Delta\sigma$, all ensemble members collapse during the overshoot. However, the timing of tipping can differ by thousands of years among ensemble members. In between these two cases, we observe a parameter regime that features a splitting of the ensemble. This regime has recently been termed "tipping window" by Ashwin et al. (2025). The fraction of tipping trajectories depends on the manitude of overshoot, in agreement with Fig. 5.12. Furthermore, the range of tipping times widens among those trajectories that transition. Thus, in the intermediate regime, tipping becomes highly unpredictable.

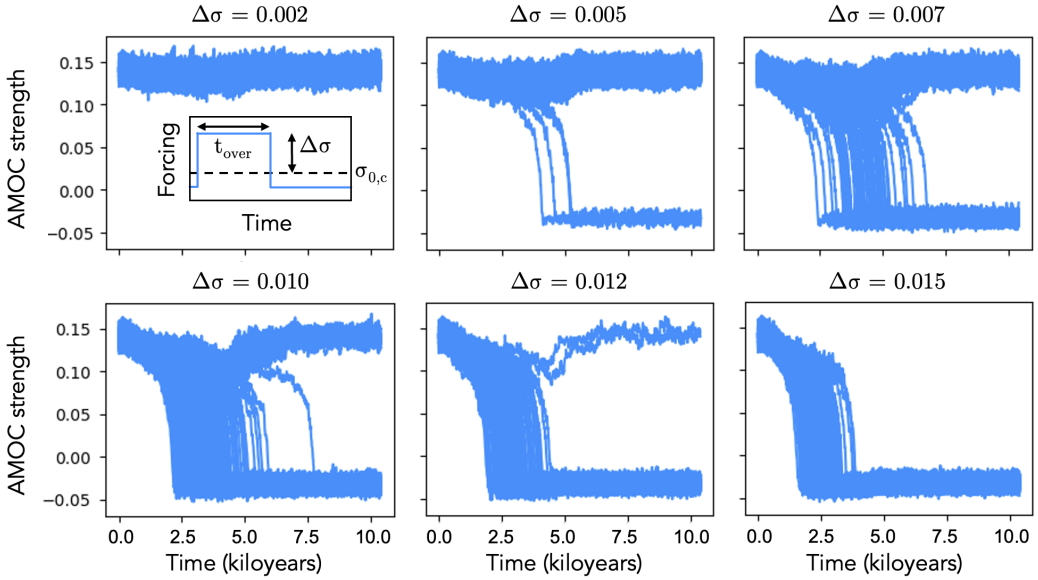

Figure 5.13 | **Ensemble splitting under freshwater overshoot** for a step forcing protocol with overshoot magnitude $\Delta\sigma$ and duration $t_{\text{step}} = 35$ time units ($t \in [5, 40]$), as depicted in the inset figure (top left). Each panel shows 200 ensemble members.

5.5 Discussion & conclusion

This chapter presented a conceptual study of metastability arising from chaotic, multiscale dynamics. Instead of following the Hasselmann (1976) approach of modeling the influence of the atmosphere on the ocean as noise, we explicitly resolved the fast, chaotic atmospheric dynamics. Our model brought together the seminal models of Stommel (1961) and Lorenz (1984) with a two-way coupling. For the ocean component, we explicitly computed the global quasipotential landscape of a bistable AMOC, before exploring how the ocean-atmosphere coupling affects the geometric fine structure of the landscape.

We investigated two fundamental limits to predictability of the asymptotic AMOC state: a fractal basin boundary in the bistable AMOC regime and chaotic transients in the monostable regime. In the bistable regime, we computed the chaotic Melancholia state between the two competing AMOC attractors using the edge tracking algorithm and found a large timescale separation between the fast chaotic motion on the saddle and the slow escape rate from the saddle. This timescale separation implies a fractal basin boundary with close to full phase space dimension, which we verified by explicitly computing the box-counting dimension of the boundary. In the monostable regime, chaotic transients with exponentially distributed lifetimes arise close to the bifurcation points over a relatively wide range of the freshwater parameter σ_0.

This complex behavior complicates the assessment of a system's resilience in two distinct ways. In dynamical systems theory, one common definition of resilience relies on measuring the minimal "kick" perturbation that causes the system to transition into a competing (undesired) attractor (Holling 1973; Krakovská et al. 2023; Halekotte and Feudel 2020) (see section 3.2.3). A fractal basin boundary thus implies a complete loss of resilience in an extended region in state space, a fuzzy "gray zone" where an arbitrarily small perturbation may cause a critical transition. In parameter space, resilience is often defined via the distance to critical thresholds delimiting a safe operating space (Rockström

et al. 2009; Krakovská et al. 2023). The presence of long chaotic transients near ghost states renders it essentially impossible to determine the exact position of such bifurcation thresholds on finite timescales by observing or simulating only a single time series. Hence, defining a safe operating space requires a probabilistic definition due to the exponential distribution of transient lifetimes, and predictability depends crucially on the timescale of interest.

The limits of predictability discussed here arise from the chaotic multiscale nature of the coupled atmosphere–ocean system. We derived these findings from a conceptual climate model that consisted of four key ingredients for this behavior – bistability, chaotic motion, timescale separation and weak coupling. These properties are also characteristic of more complex climate models, and in chapter 6 we will show that our conclusions indeed generalize to higher-fidelity AMOC models.

Though we have not discussed the properties of noise-induced transitions in the Stommel model here, we can deduce from the quasipotentials in Fig. 5.4 that saddle avoidance is likely not an issue in this model. This is because the saddle point of the Stommel model marks a well-defined quasipotential minimum for the transition scenarios in both directions. In fact, considering the drift field, the Stommel model may be viewed as a multiscale system that corresponds more closely to the FitzHugh-Nagumo model with $\varepsilon = 10$ (see chapter 4), whereas the Rooth model seems more related to the case $\varepsilon = 0.1$. Along these lines, it is interesting to think about how the design of AMOC box models impacts their transition properties and stability landscape.

Chapter 6

Metastability of the AMOC in an intermediate-complexity climate model

A study in 10^5 dimensions

6.1 PlaSim-LSG: Intermediate-complexity climate model . . . 124
6.1.1 Model description . . . 124
6.1.2 Geography and diagnostics . . . 125
6.2 AMOC stability and response . . . 126
6.2.1 Two attractors: Strong and weak AMOC . . . 126
6.2.2 Quasi-equilibrium simulations: Bistable CO_2 regime . . . 127
6.2.3 Ensemble splitting: Response to future emissions scenarios . . . 128
6.3 Edge tracking: Constructing the Melancholia state . . . 130
6.3.1 Locating the basin boundary . . . 130
6.3.2 Tracking towards an edge state . . . 130
6.3.3 Reduced phase space . . . 132
6.4 Climate of the Melancholia state . . . 134
6.4.1 Energetics . . . 135
6.4.2 Eccentric observables . . . 136
6.4.3 Drivers of the unstable oscillations . . . 137
6.5 Boundary crisis: From Melancholia to ghost state . . . 141
6.5.1 AMOC stability at increased CO_2 . . . 142
6.5.2 Collision of ON and edge states . . . 142
6.5.3 Chaotic transients . . . 143
6.6 Trajectories under future nonautonomous forcing . . . 145
6.6.1 PlaSim-LSG: Response to SSP scenarios . . . 145
6.6.2 CMIP6 model: Interpreting the "stochastic bifurcation" . . . 146
6.7 Discussion & conclusion . . . 148

This chapter is based on the manuscript by Börner et al. (2025a), submitted to Philosophical Transactions of the Royal Society A.

Introduction

So far, we have studied low-dimensional, conceptual models – the "spherical cows" of climate science (Harte 1988) (chapters 3 and 5). While these simple models permit detailed numerical analyses and enable process understanding, it is clear that they oversimplify or neglect many real-world processes of likely relevance. On the other end of the model hierarchy, state-of-the-art earth system models are expensive to run, such that mapping out their phase space and stability landscape becomes prohibitive. Earth system models of intermediate complexity (EMICs; see section 1.2.4) offer a bridge between simulation and understanding (Held 2005; Claussen et al. 2002).

In this chapter, we assess the metastability of the ocean circulation in PlaSim-LSG, a fully coupled EMIC with around 10^5 degrees of freedom. Our key objective is to gain a global view on the dynamics by computing Melancholia states of the AMOC, assuming they exist in the model. Using the edge tracking algorithm (see section 2.2.1), Lohmann and Lucarini (2024) recently computed a Melancholia state of the AMOC in a global ocean circulation model. Here we seek to add a significant step of realism by using a model that couples the ocean dynamics to a dynamic atmosphere, a hydrological cycle, and an interactive sea ice component.

Our strategy builds on the findings of chapter 5, where we studied transient chaos and the splitting of an initial condition ensemble following a boundary crisis with respect to freshwater forcing. However, as argued in section 1.2.4, it is equally relevant to study the stability of the AMOC with respect to CO_2 (Willeit and Ganopolski 2024), given high continued greenhouse gas emissions. With a coupled model like PlaSim-LSG, we can directly investigate the earth system response to changes in CO_2 concentrations.

To give a graphical summary of the findings in this chapter, we propose the problem setting schematically depicted in Fig. 6.1. It resembles the bifurcation diagram in Fig. 5.6 but now refers to CO_2 as the control parameter on the horizontal axis and considers chaotic dynamics. As we will see, the AMOC in PlaSim-LSG seems to undergo a bifurcation from

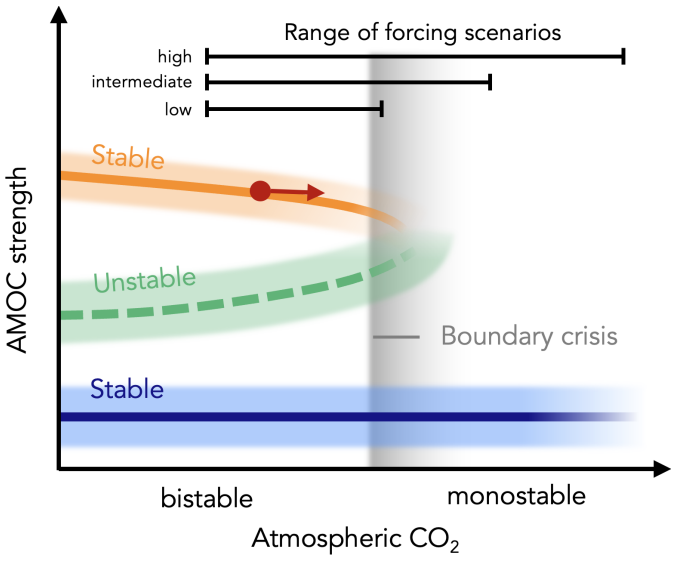

Figure 6.1 | **Setting considered in this study.** We investigate the global stability of the AMOC at two CO_2 levels, one in a bistable regime and one in a monostable regime, near a boundary crisis. We then use the results to understand the AMOC behavior under time-dependent CO_2 forcing scenarios. The gray region indicates the tipping window in which long transients and ensemble splitting may occur. Reproduced from Börner et al. (2025a).

a bistable regime at lower CO_2 values to a monostable regime at higher CO_2 values.

Our study combines simulations at fixed CO_2, which yield snapshots of the stability landscape for a given frozen system (see section 2.1.2), with runs under time-dependent CO_2 forcing. We use the results to interpret the behavior observed in CMIP6 (Coupled Model Intercomparison Project phase 6) models under future greenhouse gas emission scenarios (Romanou et al. 2023; Gu et al. 2024). Specifically, we offer a dynamical systems-based explanation for the ensemble splitting observed in the NASA-GISS-E2-1G model (hereafter GISS model), referred to as a "stochastic bifurcation" by Romanou et al. (2023).

6.1 PlaSim-LSG: Intermediate-complexity climate model

PlaSim-LSG[1] is a coupled general circulation model of intermediate complexity, comprising a dynamic atmosphere, ocean, sea ice component and hydrological cycle (Fraedrich et al. 2005; Maier-Reimer et al. 1993; Angeloni et al. 2020). Ice sheets and vegetation are prescribed in our setup. With $D \approx 10^5$ degrees of freedom, it balances detail with computational affordability, producing around 700 model simulation years per day running on a single core.

Various versions of the model (sometimes with a slab ocean instead of a dynamical ocean) have previously been used to study its climate variability (Angeloni 2022), optimal fingerprinting of climate change (Lucarini and Chekroun 2024), the Snowball Earth transition (Kaszás et al. 2019; Margazoglou et al. 2021) and extremes (D'Errico et al. 2022), applying rare event algorithms (Ragone et al. 2018; Wouters et al. 2023; Sauer et al. 2024; Cini et al. 2024). Specifically regarding the AMOC, PlaSim-LSG has been used to study mechanisms of multicentennial variability (Mehling et al. 2024) as well as spontaneous AMOC tipping (Cini et al. 2024).

6.1.1 Model description

The atmosphere component of the Planet Simulator (PlaSim) (Fraedrich et al. 2005) solves the moist primitive equations, describing the conservation of mass and momentum as well as basic thermodynamics, using simplified parameterizations of radiation, convection, precipitation and cloud processes. The prognostic equations are formulated in a spectral represenation truncated at T21 resolution horizontally (roughly corresponding to a $5.6° \times 5.6°$ grid) with 10 vertical levels.

The atmosphere is coupled to the Large Scale Geostrophic (LSG) ocean model (Maier-Reimer et al. 1993) via a 50 m thick slab ocean, which serves as the surface layer of LSG and regulates the heat, freshwater and momentum fluxes between the ocean and atmosphere. Under the assumption that the nonlinear terms of the Navier-Stokes equations can be neglected for large-scale ocean flows (Hasselmann 1982), the model solves the momentum equations and evolution equations for temperature and salt based on hydrostatic balance and the Boussinesq approximation. Convection is not explicitly resolved but accounted for via a convective adjustment scheme. At each time step, the scheme mixes vertically adjacent grid boxes whenever they are unstably stratified, starting from the top and iterating through the water column. Discretized on an E grid (Arakawa and Lamb

[1]The PlaSim-LSG model is available as open-source code at `https://github.com/jhardenberg/PLASIM`.

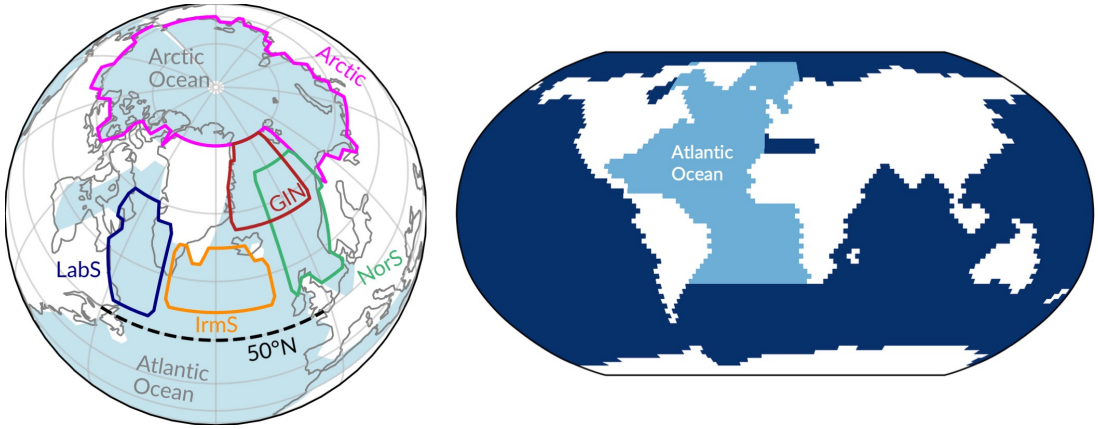

Figure 6.2 | **Geography of PlaSim-LSG**. Left: Map of geographical regions defined in this chapter. Right: Land-sea mask at the ocean surface, showing the region we define as the Atlantic Ocean basin. Figure reproduced from Börner et al. (2025a), Supplemental Information.

1977), LSG has an effective horizontal resolution of $3.5° \times 3.5°$ and 22 vertical layers on a stretched grid with thicknesses ranging from 50 m at the surface to 1000 m in the deep ocean.

The thermodynamic sea ice module of PlaSim-LSG is based on a zero-layer model (Semtner 1976) that computes the ice thickness from the thermodynamic balance at the ice-air and ice-ocean interface. Ice is melted from above and frozen from below when the sea temperature drops below freezing, ensuring a closed heat budget.

The model is configured to roughly reflect present-day climatic conditions (with orbital parameters corresponding to around 2000 CE). Its climate sensitivity is slightly above 4 °C, which lies well within the CMIP6 range (Angeloni et al. 2020). At the baseline CO_2 concentration of 360 ppm (a level recorded in the year 1995), the default initialization of the model produces a vigorous AMOC with an average strength of around 16 Sv ($1\,\text{Sv} = 10^6\,\text{m}^3\,\text{s}^{-1}$) at 26°N. This is close to today's observed value of 16.9 ± 1.2 Sv (Johns et al. 2023).

6.1.2 Geography and diagnostics

Figure 6.2 gives an overview of the oceanic spatial resolution and geography of PlaSim-LSG. Geographical regions and ocean basins are defined with the aim of matching the region masks used for the GISS model and specified by Romanou et al. (2023).

Unless stated otherwise, throughout this chapter we define the *AMOC strength* as the maximum of the meridional overturning streamfunction (see Eq. (1.5)) between 46°N and 66°N and below 100 m depth. This choice follows previous AMOC studies with PlaSim-LSG (Mehling et al. 2022; Cini et al. 2024) but differs from the standard definition used in many studies (e.g. Hawkins et al. (2011) and van Westen et al. (2024)), which refers to the streamfunction maximum at 26.5°N. The density is computed from salinity and potential temperature using the simplified EOS in Eq. (1.3).

Throughout this chapter, we consider annual mean data of the model output. Thus, when referring to the *sea ice* extent, we define the sea ice border as the boundary of the region where the annual mean sea ice thickness is at least 5 cm.

Since convection is parameterized via a convective adjustment scheme, we must choose a way of estimating the *convection depth*. For each horizontal grid point, we start at the sea

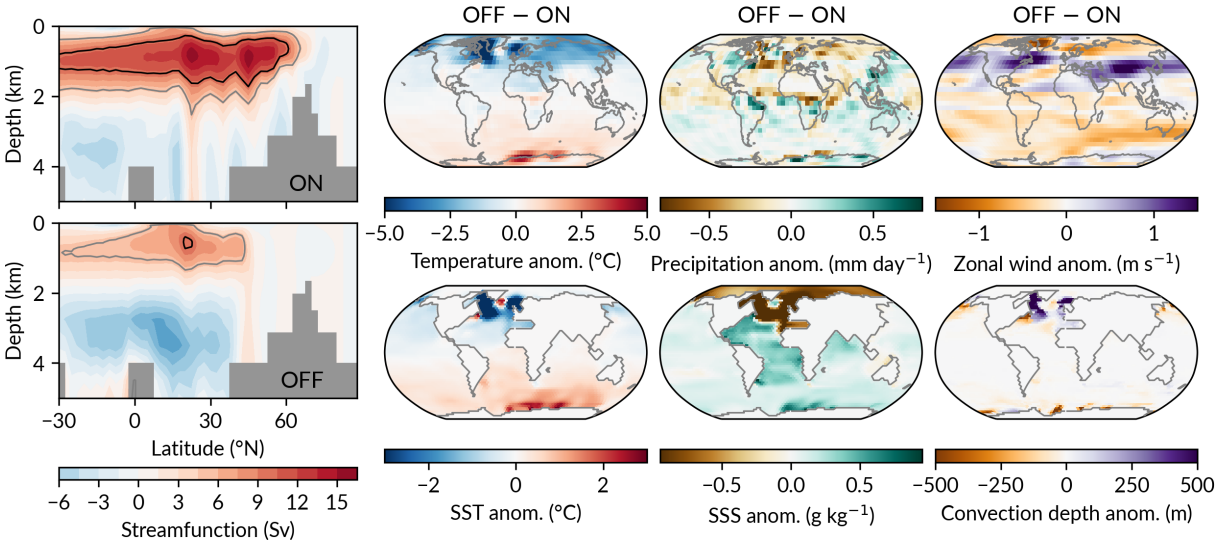

Figure 6.3 | **AMOC bistability in PlaSim-LSG at 360 ppm CO_2.** (a) and (b) show the Atlantic meridional streamfunction of ON and OFF, respectively. (c)-(h) show anomalies of OFF relative to ON for (c) surface air temperature, (d) precipitation, (e) zonal wind speed in the mid-troposphere (around 300-800 hPa), (f) sea surface temperature, (g) sea surface salinity, and (h) oceanic convection depth. All panels are computed from 1000-year time averages. Reproduced from Börner et al. (2025a).

surface and descend until reaching a vertical level for which the annual mean of convective adjustment events is zero. The depth of the previous level (where annual mean convection is nonzero) is taken as the convection depth.

6.2 AMOC stability and response

6.2.1 Two attractors: Strong and weak AMOC

At 360 ppm CO_2, the model features (at least) two distinct stable AMOC states: a strong interhemispheric overturning cell with an average maximum strength of 16 Sv and a much weaker and shallower overturning circulation that shuts down to less than about 2 Sv north of 46°N (Fig. 6.3a). As customary, we refer to these equilibria as the ON and OFF state, respectively. Their stability has been verified via 4000-year long unforced simulations.

The ON state resembles today's climatology and the large-scale ocean circulation pattern currently observed in the Atlantic Ocean (Frajka-Williams et al. 2019). In the OFF state, the Atlantic meridional streamfunction collapses in the Subpolar Gyre region, while a weakened meridional volume transport remains at lower latitudes ($\approx$ 8 Sv at 26°N, see Fig. 6.3b). Thus, the OFF state in PlaSim-LSG represents a weak AMOC rather than an entirely collapsed overturning circulation. A weak stable AMOC state is found in some models (Jackson et al. 2023; Romanou et al. 2023), while in other models the OFF state corresponds to a full AMOC collapse (e.g., van Westen and Dijkstra (2023) and Lohmann, Dijkstra, et al. (2024)). Still, the OFF state is characterized by the typical climate signal associated with an AMOC collapse, including a reduction of mean surface air temperature in the Northern Hemisphere that can locally exceed 10 °C, a drying of North Atlantic regions including northern Europe, and a southward shift of the tropical

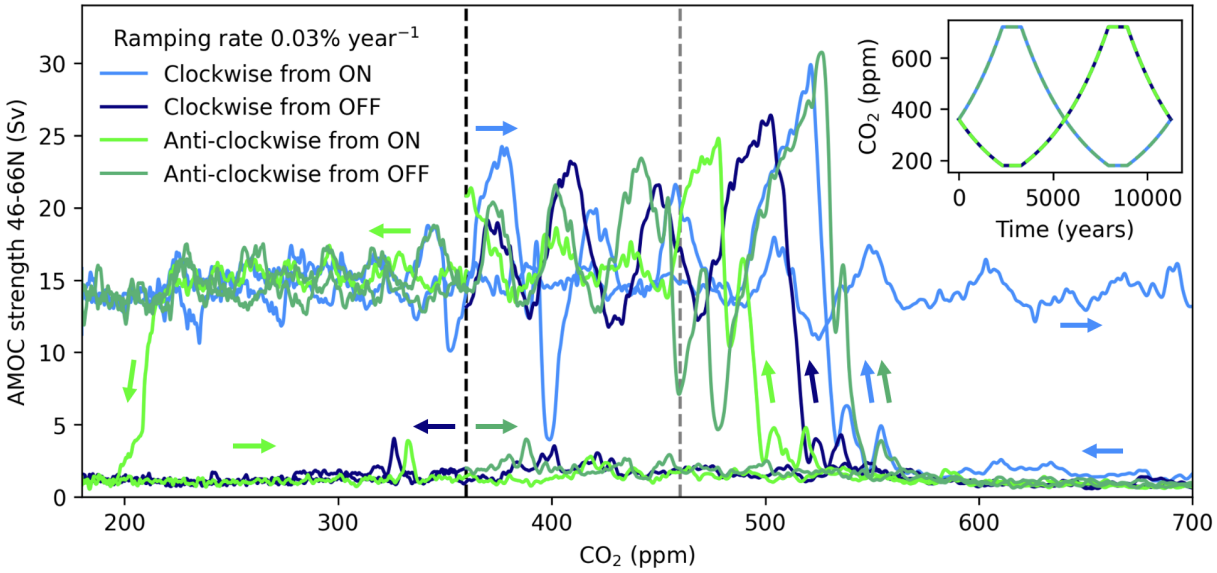

Figure 6.4 | **AMOC hysteresis experiments with respect to CO_2.** The CO_2 ramping protocols are shown in the inset figure. All runs start at 360 ppm, either from the ON state (light blue, light green) or from the OFF state (dark blue, dark green), with arrows indicating the direction.

rain belt (Intertropical Convergence Zone, ITCZ) (Fig. 6.3). We also find a strengthened polar jet stream in the northern hemisphere, combined with a large-scale reduction of zonal winds in other regions. Shifts in the tropical circulation induced by an AMOC decline can have major impacts on the stability of the Amazon rainforest, which exemplifies the risk of tipping cascades among Earth system tipping elements (Akabane et al. 2024).

In the OFF state, the time-averaged sea surface temperature (SST) is more than 2°C (up to 9°C) colder than in the ON state in large parts of the North Atlantic, while the Southern Ocean is up to 3°C warmer. The Atlantic subtropical gyre region, southern Atlantic, Indian Ocean and Southern Ocean are saltier in the OFF state, whereas the North Atlantic and Arctic Ocean are substantially fresher (with the exception of the Irminger Sea). This is a clear signature of the salt-advection feedback and meridional ocean heat transport: the weakened AMOC transports less salt and heat from the tropics to the north.

It is well established that the AMOC is closely connected with sites of deep oceanic convection in the North Atlantic, where dense water sinks, acting as a "pump" of the overturning circulation. In models and observations, major deep convection sites are located in the Labrador Sea (LabS), Irminger Sea (IrmS), and Norwegian Sea (NorS; see Fig. 6.2 for a map of these regions). In PlaSim-LSG, the transition from the ON to the OFF state is characterized by a shutdown of deep convection in the LabS and NorS (Fig. 6.3f), while the convection depth increases in several other locations.

In summary, the AMOC ON and OFF states have a qualitatively different climate on a global scale. In order to understand the transition behavior between these states, we now investigate the global stability of the AMOC beyond the stable equilibria.

6.2.2 Quasi-equilibrium simulations: Bistable CO_2 regime

To understand how AMOC multistability in PlaSim-LSG depends on the CO_2 concentration, we perform four hysteresis experiments where we slowly ramp the CO_2 level at a

rate of 0.03% per year. The exponential change in CO_2 corresponds to a linear change in radiative forcing.

Specifically, we start at 360 ppm and apply two protocols (see inset of Fig. 6.4):

(a) Ramping up to $2\times CO_2$ (720 ppm), equilibrating for 1000 years, ramping down to $0.5\times CO_2$ (180 ppm), equilibrating again for 1000 years, and ramping back up to 360 ppm.

(b) The time-reversal of (a).

We run each of these protocols starting from an initial condition on the ON and the OFF state, respectively.

The resulting AMOC trajectories show a rather surprising behavior (Fig. 6.4). All four runs exhibit a rapid recovery of the AMOC from OFF→ON at around 500-550 ppm. The protocol (a) run initialized from the ON state retains a strong AMOC until reaching the 1000-year plateau at 720 ppm, during which the AMOC collapses to the OFF state. Overall, the results suggest that the AMOC may be bistable in the entire range of CO_2 values covered ([180,720] ppm). However, as we will see, this is not the case.

Based on the four simulations, the true bifurcation diagram of the AMOC with respect to CO_2 in PlaSim-LSG does not become fully clear. Possibly, the ramping rate is still too high to track equilibrium conditions quasi-adiabatically. An additional difficulty arises from the chaotic behavior of the model, which we focus on in the following sections. While the limits of the bistable regime with respect to CO_2 remain elusive, we know that at 360 ppm the AMOC is bistable in the model (on a timescale of at least 4000 years).

In addition to the hysteresis experiments, we conducted further CO_2 ramping experiments where we varied the ramping rate or branched off at different CO_2 levels (landing climates), continuing under constant forcing. The forcing protocols and responses in AMOC strength are summarized in the Appendix (Figs. A.6 and A.7).

6.2.3 Ensemble splitting: Response to future emissions scenarios

After looking at the quasi-equilibrium response of the AMOC, let us now turn to the transient response under future climate change. We force the model with CO_2 projections of Shared Socioeconomic Pathways (SSPs) – standardized climate change scenarios for greenhouse gas concentrations until 2500 (Meinshausen et al. 2020) (Fig. 6.5a). For each SSP, we launch an ensemble of runs starting in the forcing year 1995, with initial conditions branched off from a 2000-year control run at 360 ppm. We focus on SSP1-2.6 (strong emissions reduction), SSP2-4.5 (intermediate emissions) and SSP4-6.0 (high emissions). Beyond 2500, we assume that the CO_2 concentration continues to decay exponentially to 330 ppm at the rate of decay reached towards the end of the SSP scenario.

The AMOC shows qualitatively different behavior under these three climate change scenarios (Fig 6.5b). For SSP1-2.6, the vigorous AMOC state is maintained over the 1000-year simulation, as shown for three ensemble members. In the SSP4-6.0 scenario, the AMOC collapses north of 46 °N for all ensemble members. The collapse starts around the year 2200 and most of the decline happens within 100 years. Strikingly, in the intermediate SSP2-4.5 scenario, the selected ensemble members exhibit ensemble splitting, with the North Atlantic AMOC sometimes persisting and sometimes collapsing after strongly varying transients. While one ensemble member rapidly declines after 2200, another undergoes an initial decline, followed by a recovery and overshoot, only to abruptly collapse

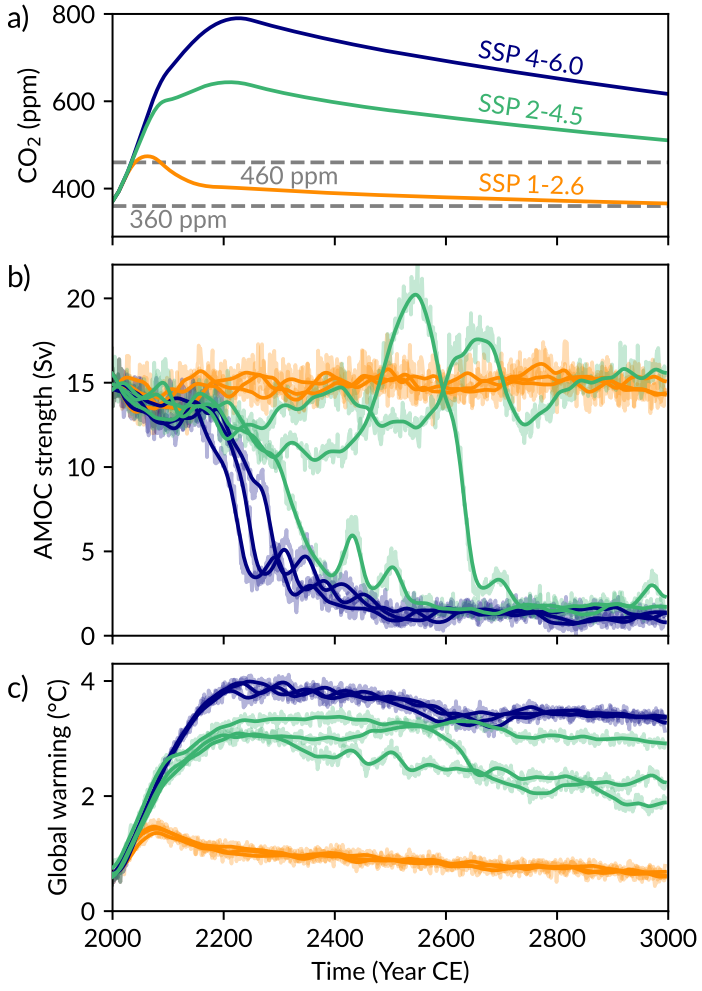

Figure 6.5 | **Simulated evolution of the AMOC and global warming** in PlaSim-LSG under three extended SSP scenarios from 2000 to 3000 CE. a) Atmospheric CO_2 concentration for each scenario, indicating 360 and 460 ppm as dashed lines. b) AMOC strength (10-year smoothed, with annual variability shown as faint lines) for simulations (three ensemble members each) forced by the corresponding SSP scenario as color-coded. c) as b) but showing global mean surface temperature change relative to the 1850-1900 reference. Reproduced from Börner et al. (2025a).

a few hundred years later. By contrast, the third ensemble member remains close to the initial AMOC strength, albeit with increased centennial variability.

The differing AMOC behavior imprints itself on the global climate, including the global mean surface temperature (Fig. 6.5c). As expected, the amount of global warming increases from the low to the high emissions scenario, but global warming in the SSP2-4.5 scenario can differ by up to 1 °C depending on the state of the AMOC. Generally, AMOC weakening reduces the global mean surface temperature, in line with expectations.

The results shown in Fig. 6.5 should not be taken as reliable future climate projections, given the reduced complexity of the model and the fact that we neglect other greenhouse gas emissions besides CO_2, such as methane, and other climate-altering forcings such as those associated with land use change and injection of aerosol into the troposphre as a result of pollution. Nonetheless, the AMOC behavior under SSP2-4.5 resembles the ensemble splitting found in the GISS model (Romanou et al. 2023), as we discuss in section 6.6.2. To better understand this behavior, we now explore the global stability properties of the AMOC in detail.

6.3 Edge tracking: Constructing the Melancholia state

As in chapter 5, our plan is to implement the edge tracking algorithm to construct a Melancholia state of the model (see section 2.2.1). However, since we now have 10^5 instead of five dimensions, its realization is technically more challenging and computationally involved.

6.3.1 Locating the basin boundary

We take two model restart files from previous PlaSim-LSG simulations by Mehling et al. (2022) as the pair of initial conditions $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ for the edge tracking procedure. The simulations were performed at 285 ppm CO_2 with differing vertical diffusivity profiles, which strongly affected the AMOC strength (Mehling et al. 2022). In our model configuration (at 360 ppm), these two initial conditions are located near the ON and OFF state, respectively, and converge to these.

The first step is to locate the basin boundary that necessarily lies somewhere along a straight line between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. Interpolating along all variables x_i with $i = 1, \ldots, D$ in the model restart files ($D \approx 10^5$), we compute new initial conditions $x_{j,i} = x_{a,i} + 0.1j(x_{b,i} - x_{a,i})$ with $j = 1, 2, \ldots, 9$. This required writing Fortran subroutines that read, interpolate, and write restart files, which for PlaSim-LSG are stored in binary format.

Computing the AMOC strengths for these initial conditions shows that the states $\boldsymbol{x}_j$ are monotonically decreasing in AMOC strength with increasing j from around 15 to 3 Sv (Fig. 6.6a). However, the differences in AMOC strength between adjacent states are not equidistant, reflecting that the AMOC strength is a nonlinear mapping of the full phase space.

From the initial conditions $\boldsymbol{x}_j$, we run parallel simulations for 200 years each. While the trajectory initialized at $\boldsymbol{x}_1$ remains close to the ON state in AMOC strength, all other trajectories converge to the OFF state. This implies that a part of the basin boundary is located between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in phase space. This pair of initial conditions $\boldsymbol{x}_a^{(0)}$ and $\boldsymbol{x}_b^{(0)}$ constitutes the starting points for the first edge tracking iteration (see below).

6.3.2 Tracking towards an edge state

The computational cost of the edge tracking algorithm depends on how slowly trajectories near the basin boundary converge to the attractors. This is because the asymptotic state of each new initial condition must be determined by simulation, which can take hundreds of model years for the AMOC. With PlaSim-LSG, however, we can exploit the fact that multiple simulations can be run in parallel. Thus, instead of successive bisections as described in section 2.2.1 (and implemented in chapter 5 and by Lohmann and Lucarini (2024)), we compute nine equidistant initial conditions at once by linear interpolation and run parallel simulations from them. That way, we can reduce the distance ε between initial conditions by a factor 10 in one interpolation step (instead of 2).

As a distance measure between trajectories, we simply use the difference in the 10-year smoothed AMOC strength (between 46-66°N), i.e. the scalar variable u in section 2.2.1 is here the AMOC strength. Then, the k-th iteration of the edge tracking implementation in PlaSim-LSG consists of the following steps:

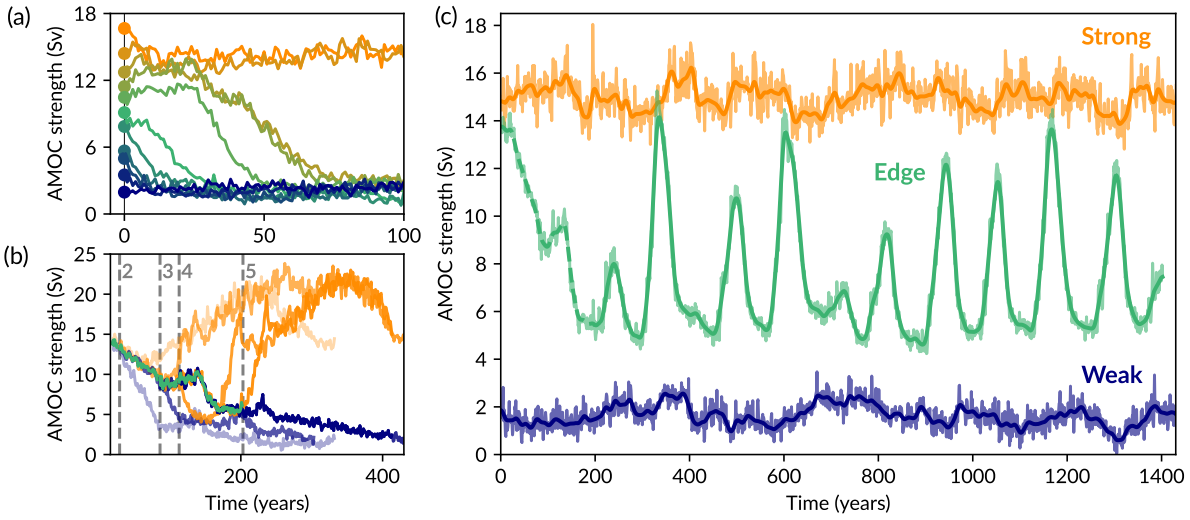

Figure 6.6 | **Edge tracking and AMOC states at 360ppm CO_2.** (a) Interpolating initial conditions between the Strong (orange) and Weak (blue) AMOC state allows locating the basin boundary. (b) Iterations 2-5 of the edge tracking algorithm, showing the trajectories that converge to the Strong (orange) and Weak (blue) state, respectively. The edge pseudotrajectory (green) is constructed from segments of these trajectories. (c) Edge trajectory (green) and trajectories on the Strong (orange) and Weak (blue) attractors. The AMOC strength is measured between 46-66N. Reproduced from Börner et al. (2025a).

1. Compute interpolated initial conditions $\boldsymbol{x}_j^{(k)} = \boldsymbol{x}_a^{(k-1)} + 0.1j\left(\boldsymbol{x}_b^{(k-1)} - \boldsymbol{x}_a^{(k-1)}\right)$ for $j = 0, \ldots, 10$.

2. Run parallel simulations from $\boldsymbol{x}_j^{(k)}$ for 350 model years each (or longer if necessary).

3. Find new indices $m, n \in \{0, \ldots, 10\}$ such that the trajectories from $\boldsymbol{x}_m^{(k)}$ and $\boldsymbol{x}_n^{(k)}$ stay close to each other ($< 1\,\mathrm{Sv}$ difference in smoothed AMOC strength) for as long as possible but converge to different attractors.

4. Select the time t_k at which the trajectories from $\boldsymbol{x}_m^{(k)}$ and $\boldsymbol{x}_n^{(k)}$ first diverge by 1 Sv in AMOC strength and use their states at t_k as new initial conditions $\boldsymbol{x}_a^{(k)}$ and $\boldsymbol{x}_b^{(k)}$.

5. Increase k by 1 and repeat step 1.

For the first few iterations of edge tracking, spanning about 200 years, the resulting pseudotrajectory (hereafter called *edge trajectory*) decreases in AMOC strength from 14 Sv to about 5 Sv. Subsequently, the AMOC increases again, and the edge trajectory begins a series of large AMOC oscillations (Fig. 6.6c). The quasiperiodic oscillations vary in amplitude from 3 to 10 Sv, with a mean period of 118 ± 7 years (estimated from 10 peaks). This behavior persists until the edge tracking was stopped after around 1400 years (39 iterations).

The recurrent pattern of centennial AMOC cycles suggests that the edge trajectory has converged to an edge state and thereafter evolves on this unstable set. This claim is supported by the fact that the specific potential energy of the global ocean is relatively constant after convergence (Fig. 6.9b) and that the salinity in the deep Pacific, Indian, and Southern Oceans has equilibrated (not shown). Since the oscillations are neither perfectly

periodic nor constant in amplitude, the edge state appears to be a chaotic saddle with a more complex geometry than that of an unstable limit cycle. This nonattracting invariant set is approximated by the edge trajectory after removing the initial transient of 200 years.

The variations in oscillation amplitude and frequency could be explained in terms of unstable period orbits (UPOs): assuming that the chaotic edge state supports an infinite set of UPOs, the edge tracking algorithm will likely track different UPOs across its iterations (Grebogi et al. 1988; Cvitanović 1991; Ashwin et al. 2025).

Owing to chaos, the precise course of the edge trajectory would differ if we would repeat the edge tracking procedure from slightly different initial conditions or with a slightly different criterion for choosing t_k. Recall (see section 2.2.1) that the edge trajectory is stitched together from averaged segments of trajectory pairs shadowing the edge state. Ideally, the interpolation steps only counteract the divergence transversal to the stable set of the edge state, but for a chaotic saddle this is not necessarily the case[2]. This might distort the dynamics along the edge trajectory. However, the oscillations found here are a true feature, not artifacts: during some edge tracking iterations, the pair of trajectories completed one or more oscillatory cycles in free run, without re-interpolation, before diverging to different attractors. We have also re-run the edge tracking algorithm from different initial conditions and obtained the same oscillations in terms of mean period and amplitude (not shown). Moreover, the oscillations become characteristic for freely evolving trajectories near criticality, as shown in section 6.5.3 (see also Cini et al. (2025)).

We emphasize that the edge trajectory varying in time does not mean the edge state itself is time-dependent: since we fix the external forcing, the edge state is invariant in time, and the the edge trajectory reflects the dynamics *on* the edge state.

6.3.3 Reduced phase space

Looking at the one-dimensional AMOC timeseries (Fig. 6.6c) gives the impression that the edge trajectory oscillates back and forth between the ON and OFF states. However, visualizing the dynamics in a reduced phase space clarifies that the Edge state is separated from the attractors in phase space (Fig. 6.8).

Determining a suitable low-dimensional phase space projection of the 10^5-dimensional dynamics is challenging due to the countless possible combinations of variables. Our physical understanding of the AMOC suggests that the salinity and temperature in the Atlantic basin will be key variables. Using a data-driven reduction approach, we perform an empirical orthogonal function (EOF) analysis, a reduction technique that determines independent spatial patterns that explain most of the temporal variability in the data (Navarra and Simoncini 2010).

Specifically, we combine around 20 000 years of edge tracking simulations into one long trajectory (Fig. A.1). The data includes trajectories attracted by the edge state as well as relaxation paths that encapsulate the dynamics along the unstable set of the edge state. We then take the timeseries of the zonally averaged salinity field in the Atlantic as input data[3], using the `eofs` package in Python (Dawson 2016). We find that this choice contains sufficient information to disentangle the dynamics (Fig. 6.7). The leading three EOFs of the salinity field explain around 90% of the variance and show distinct patterns.

[2] Jan Sieber, personal communication.

[3] We performed the same analysis for the meridional streamfunction, temperature and density, but the salinity appeared most informative.

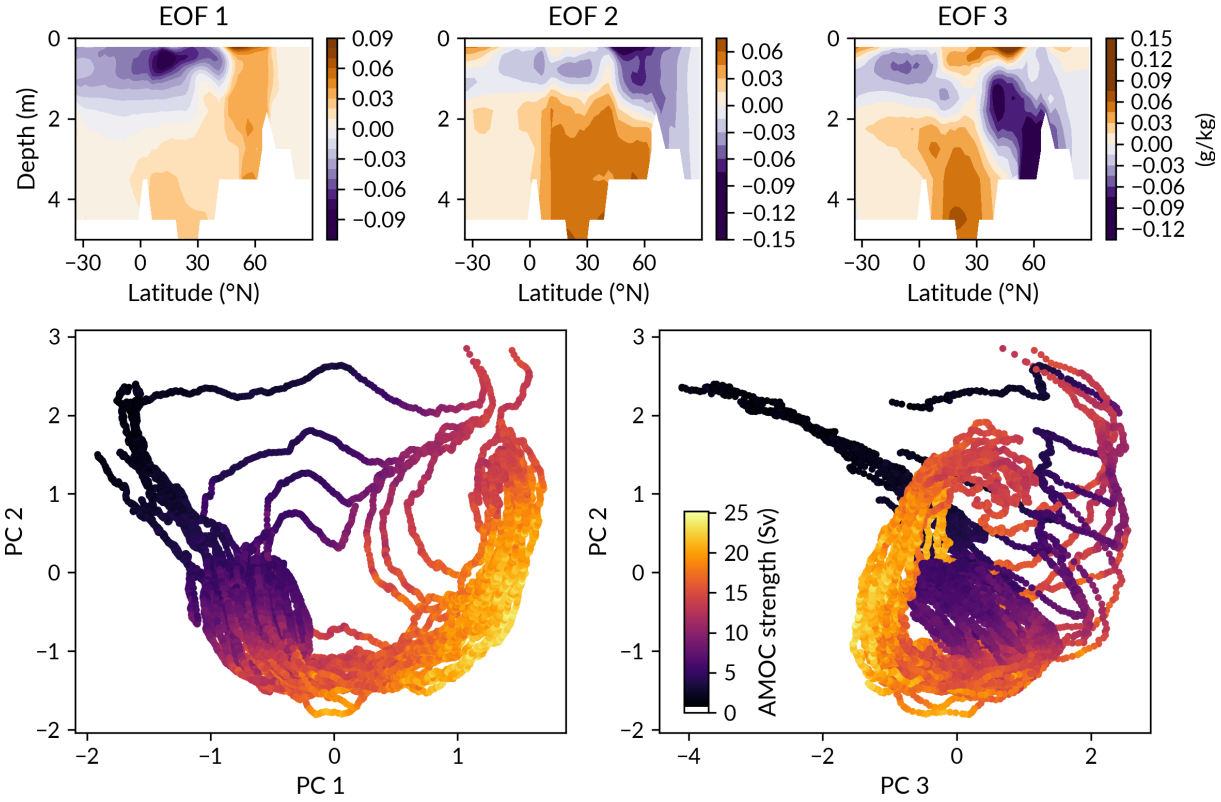

Figure 6.7 | **EOF analysis of the edge tracking simulations.** Top panels show the first three EOFs of the zonally averaged Atlantic salinity field for the simulation data summarized in Fig. A.1. Bottom panels show different PCs plotted against each other, colored by AMOC strength along the trajectories. Reproduced from Börner et al. (2025a), Supplemental Information.

Moreover, plotting the principal components (PCs) corresponding to the EOFs against each other, we obtain a reduced phase space in which the dynamical regimes of the edge, ON and OFF states are well separated. In this space, we can see how the trajectories of the first few edge tracking iterations trace a path on the basin boundary before diverging from it, but have not yet converged. The bulk of trajectories, however, is concentrated in a subspace that reveals the edge state and its unstable set.

From the results of the EOF analysis, we derive an operational definition of a reduced phase space is spanned by the following three variables:

- The *meridional* salinity gradient (SG) in the Atlantic, measured as the mean salinity difference between 0-20 N and 40-80 N in the top 1000 m (omitting the top 100 m),
- The *vertical* SG in the North Atlantic, defined as the mean salinity difference between the depths 100-1000 m and 1000-3000 m at 46-66°N,
- The *deep North Atlantic salinity anomaly*, defined as the mean salinity anomaly relative to $35\,\mathrm{g\,kg^{-1}}$ in the Atlantic basin north of 50 °N and below 1000 m depth.

This choice is inspired by the meridional salinity dipole in the upper 1000 m of EOF1 and the vertical gradient in the North Atlantic in EOF2. The third variable covers the negative anomaly seen in EOF3.

The benefit of using these derived variables, instead of directly using the principal components of the EOFs, is that they can easily be computed for any spatially resolved

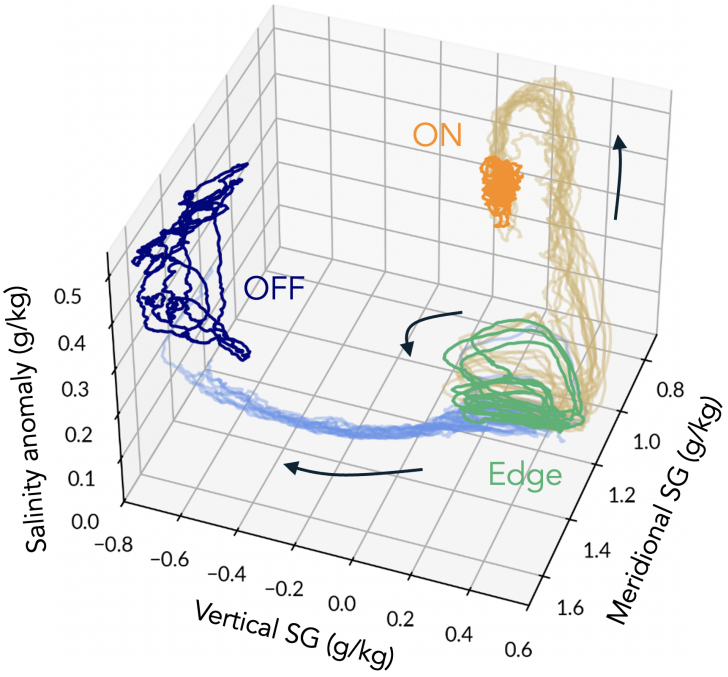

Figure 6.8 | **Phase space projection** onto the meridional SG, vertical SG, and salinity anomaly in the deep North Atlantic (below 1000 m, north of 50°S). Faint orange (blue) lines show trajectories relaxing from near the Edge state to the ON (OFF) state. Arrows indicate the time direction. Reproduced from Börner et al. (2025a).

ocean model, permitting inter-model comparisons. The meridional SG is negatively correlated with the AMOC strength, since a stronger AMOC transports more salt to the North Atlantic, reducing the salinity difference between low and high latitudes. The vertical SG and deep salinity anomaly are related to deep convection and the stability of the water column in the North Atlantic.

Viewing the trajectories of Fig. 6.6 in the reduced phase space, we see that each of the ON, OFF, and Edge states occupies a distinct region (Fig. 6.8). The Edge state has a higher vertical SG and fresher deep North Atlantic than both the ON and OFF states. The OFF state has the saltiest deep North Atlantic and largest meridional SG. While the ON state covers a relatively small volume of the reduced phase space, the AMOC oscillations of the Edge state are clearly seen as loops in the meridional-vertical SG plane. Also the OFF state exhibits relatively large internal variability that is captured in this projection but not in the AMOC strength. This low-frequency variability on multi-centennial timescales is caused by global inter-basin salt exchanges (not shown).

The simulations used to perform edge tracking also provide insight into the transition pathways from the Edge state to each of the the ON and OFF states. The trajectories of the final 20 edge tracking iterations reveal clear characteristic pathways to either attractor (Fig. 6.8), tracing the unstable set of the edge state.

6.4 Climate of the Melancholia state

The pseudotrajectory on the edge state is constructed from segments of actual model trajectories, meaning that we can explore its weather and climate as with any other model simulation. This provides insight into what the world looks like near the edge state and into the processes involved in the instability of the AMOC.

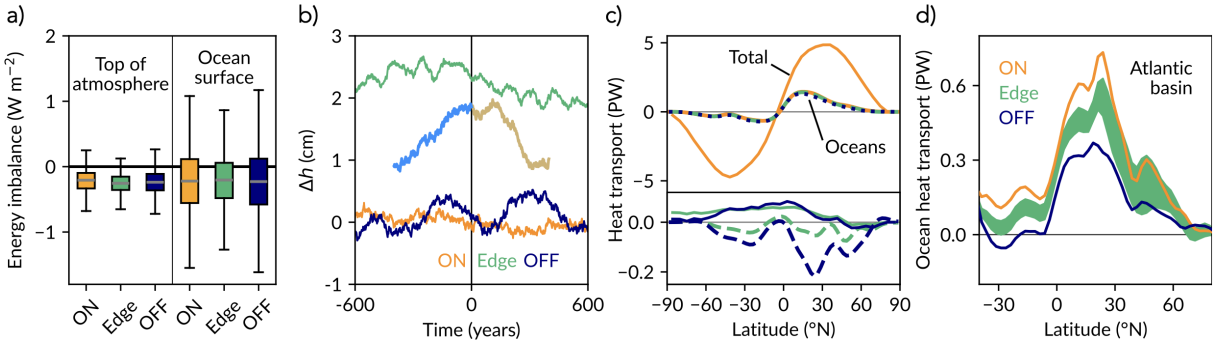

Figure 6.9 | **Energetics of the climate states.** (a) Imbalance of the top of atmosphere radiation (left) and heat flux at the sea surface (right), integrated over the globe for ON, OFF and Edge (negative imbalance means the Earth/ocean is losing energy). (b) Oceanic enter of mass anomaly Δh (relative to $1970.3126\,\mathrm{m}$ below sea level) for years 200-1400 of the edge trajectory and corresponding time intervals for ON and OFF. The 400-year long relaxation paths from Edge $\rightarrow$ ON (beige) and Edge $\rightarrow$ OFF (light blue, plotted in reverse time) are shown for one of the edge tracking iterations. (c) Northward meridional heat transport, showing the total from atmosphere and oceans (ON, orange) and the oceanic contribution (all states, dotted). Bottom inset: Difference in total (solid) and oceanic (dashed) heat transport of OFF (blue) and Edge (green) relative to ON. (d) Oceanic heat transport in the Atlantic basin only, showing the variability of the oscillations on the Edge state (green band). Reproduced from Börner et al. (2025a).

6.4.1 Energetics

The first question is whether the edge state is energetically aligned with steady state conditions, which require an approximately vanishing global energy budget for the climate as a whole as well as for its sub-components (Peixóto and Oort 1984; Lucarini and Ragone 2011). Indeed, both the radiative balance at the top of the atmosphere as well as the globally integrated net surface heat flux between the ocean and atmosphere are close to zero (comparable to the ON and OFF states; Fig. 6.9a). At the edge state, the global ocean as well as the planet as a whole are thus in steady state conditions in terms of the energy budget.

The meridional heat transport of the ocean and atmosphere combined is nearly identical for the ON, OFF and edge states, despite differences in the ocean circulation (6.9b). This means that the atmosphere largely compensates for changes in oceanic heat transport (Knietzsch et al. 2015), as reported in previous studies on the AMOC collapse (Povea-Pérez et al. 2024; Orbe et al. 2023). This is a manifestation of the Bjerknes compensation (Bjerknes 1964) (see also (Stone 1978)). It is well known that, because of the AMOC, the Atlantic Ocean is the only ocean basin with a northward oceanic heat transport on both hemispheres, causing an asymmetry of the oceanic meridional heat transport. A reduced AMOC thus decreases this asymmetry, as we observe for the OFF state (Fig. 6.9c, lower panel). Interestingly, the change in the atmospheric transport slightly overcompensates the reduction in the oceanic transport (Fig. 6.9c, upper panel). The time-averaged oceanic meridional heat transport of the edge state lies in between that of the ON and OFF states, though the AMOC oscillations cause temporal variations of more than 100 TW especially in the northern midlatitudes (Fig. 6.9d).

Based on the picture of a double-well stability landscape of the AMOC (see Fig. 2.2), we expect that the edge state has a higher potential energy than the ON and OFF

states. Lohmann and Lucarini (2024), using a global ocean model, identified the dynamic enthalpy as the appropriate thermodynamic potential (Young 2010). In a coupled model like PlaSim-LSG, also the potential energy of the atmosphere as well as the energy fluxes associated with sea ice thermodynamics and vegetation must be considered for a full account of potential energy. For simplicity, we use the elevation h of the oceanic center of mass,

$$h = H - \frac{\int_0^H z\bar{\rho}(z)\mathrm{d}z}{\int_0^H \bar{\rho}(z)\mathrm{d}z}\,, \tag{6.1}$$

as an energy measure to compare the oceanic specific potential energy among the different AMOC states. Here H is the maximum depth of the sea floor, z is the depth coordinate (positive downwards), and $\bar{\rho}$ is the horizontally integrated density across the ocean.

We find that the edge state has a significantly higher center of mass – and thus specific potential energy – compared to the two attractors. The ON and OFF states have a comparable center of mass, with the OFF state exhibiting multi-centennial variability due to global salt exchanges, as also observed in Fig. 6.8.

To understand which geographical regions contribute most to the higher center of mass, we calculate the time average of h for the water column at each horizontal grid point. Mapping out the difference Δh between the Edge state and each of the attractors shows that the Edge state has a higher specific potential energy in most of the global ocean (Figs. 6.10h-i). Compared to the ON state, the largest difference is located in the Northern Atlantic, though the Edge state has a lower center of mass in some of the deep convection zones around Greenland, specifically the Labrador Sea and Norwegian Sea. Compared to the OFF state, the area with a strongly elevated center of mass extends from the North Atlantic into the Arctic Ocean, while the center of mass is substantially lower in the Gulf Stream region.

Even though we have not explicitly calculated the relevant thermodynamic potential, the center of mass analysis offers a crude means to speculate about the quasipotential landscape of the model. Our findings match the theoretical expectation that the edge state should have a higher quasipotential value than the attractors (see sections 2.1.6 and 5.2.1).

6.4.2 Eccentric observables

Since the Edge state lies on the basin boundary between the ON and OFF states, one might expect that its climate lies somewhere in between that of the ON and OFF states, too. For example, the AMOC strength on the edge state oscillates between the strong and weak AMOC states, as does the meridional ocean heat transport. However, we have also seen that the Edge state features a fresher deep North Atlantic than both attractors, and a higher center of mass. In a high-dimensional system like our climate model, there may be many directions in which the Edge state lies outside of the ON and OFF states. These directions in phase space could be particularly relevant for detecting early warning signs, as proposed by Lohmann et al. (2025).We call such phase space variables *eccentric observables*, since transition paths would undergo non-monotonic excursions in these observables provided that they pass via the vicinity of the Edge state.

In most ocean regions, the time-averaged sea surface salinity (SSS) and sea surface temperature (SST) of the Edge state lie in between that of the ON and OFF states (Fig.

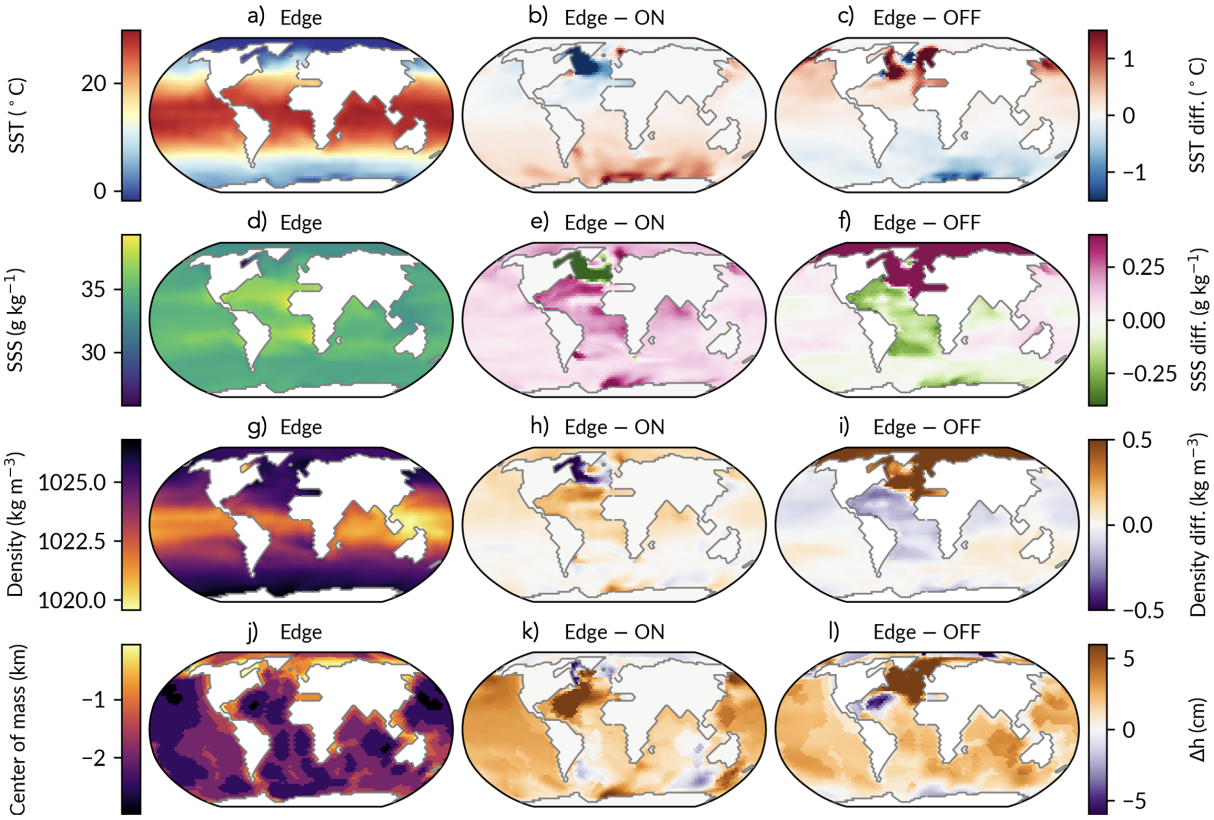

Figure 6.10 | **Ocean properties of the Edge state**, displayed as time averages over the final 640 years of the edge trajectory in absolute values (first row) and as differences relative to the ON (second) and OFF (third row) states: (a)-(c) sea surface temperature (upper 100 m), (d)-(f) sea surface salinity (upper 100 m), (g)-(i) surface density (upper 100 m), and (j)-(l) water column center of mass. Reproduced from Börner et al. (2025a).

6.10). However, almost the entire Arctic Ocean is saltier and denser in the upper ocean relative to the attractors. Parts of the Norwegian Sea are warmer than both attractors, and the northwestern Pacific Ocean is warmer and saltier on the Edge state. Other eccentric observables include the sea ice cover in the Irminger Sea and the surface freshwater flux in the Norwegian Sea.

6.4.3 Drivers of the unstable oscillations

The most prominent dynamical feature of the Edge state are the large AMOC oscillations with a period of around 120 years and an amplitude of up to 10 Sv between 46-66°N (Fig. 6.6c; see also Fig. A.2). At 26°N, the AMOC oscillations are qualitatively similar but have a smaller amplitude (Fig. 6.11a). Together with the overturning strength, many other climate observables oscillate at this frequency. What drives the unstable oscillations?

Key processes known to interact with the AMOC are North Atlantic deep convection, sea ice formation, and high-latitude surface freshwater fluxes. As shown in Fig. 6.3, the transition from ON to OFF in our model is characterized by a shutdown of deep convection sites in the Labrador and Norwegian Seas. On the Edge state, deep convection persists in the Norwegian Sea, though the mean convection depth oscillates slightly with the AMOC strength (Fig. 6.11b). Convection is inactive in the Irminger Sea. In the Labrador Sea,

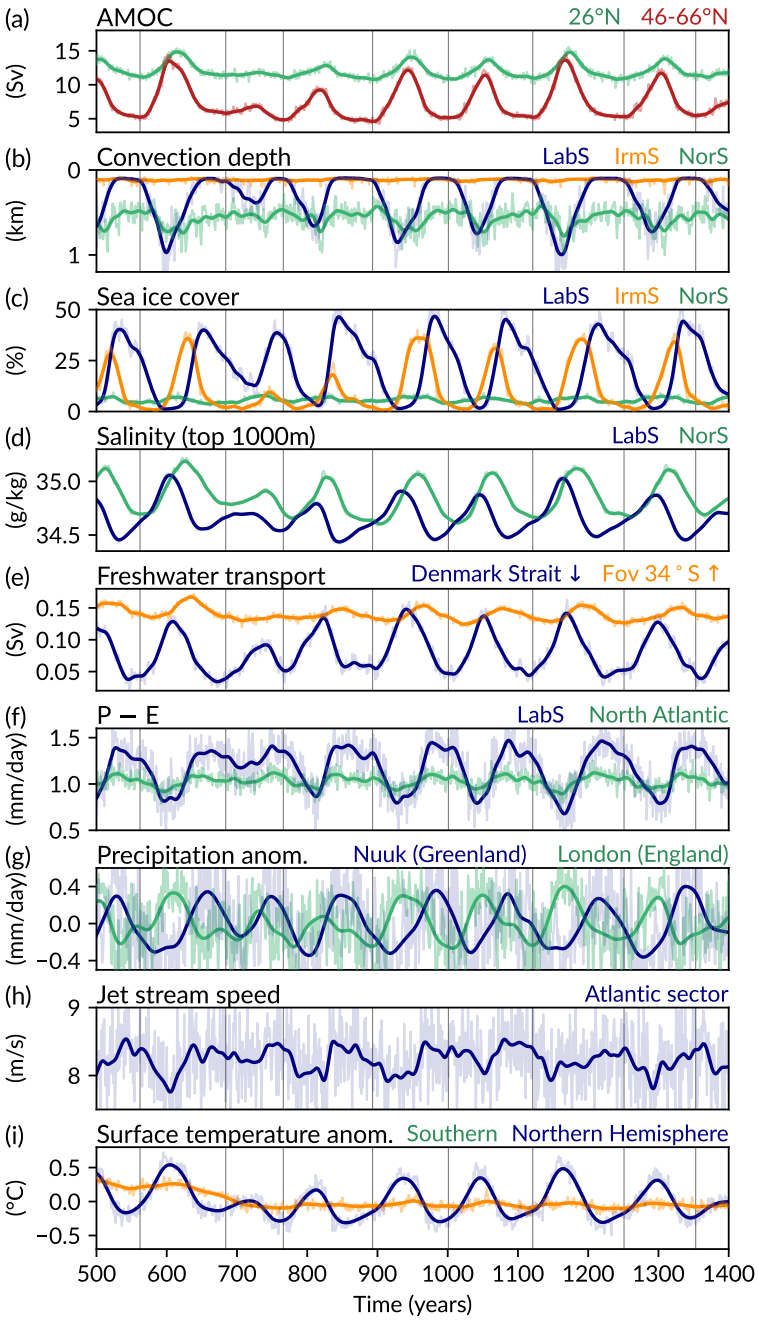

Figure 6.11 | **Oscillations on the edge state** captured in (a) the AMOC strength at 46-66°N (red) and 26°N (green), (b) the annual maximum convection depth in the LS, (c) sea ice cover in the LS (blue) and Irminger Sea (green), (d) mean salinity (averaged over the top $1000\,\mathrm{m}$) in the LS (blue) and across the Atlantic north of 50°N (green), (e) LS precipitation minus evaporation, (f) precipitation anomaly over the LS (blue) and Northern Europe (green), (g) jet stream index in the North Atlantic sector, and (h) mean surface temperature anomalies in the Northern (blue) and Southern (green) Hemispheres. Thick lines are smoothed with a 5-year Gaussian filter (10 years for precipitation). Reproduced from Börner et al. (2025a).

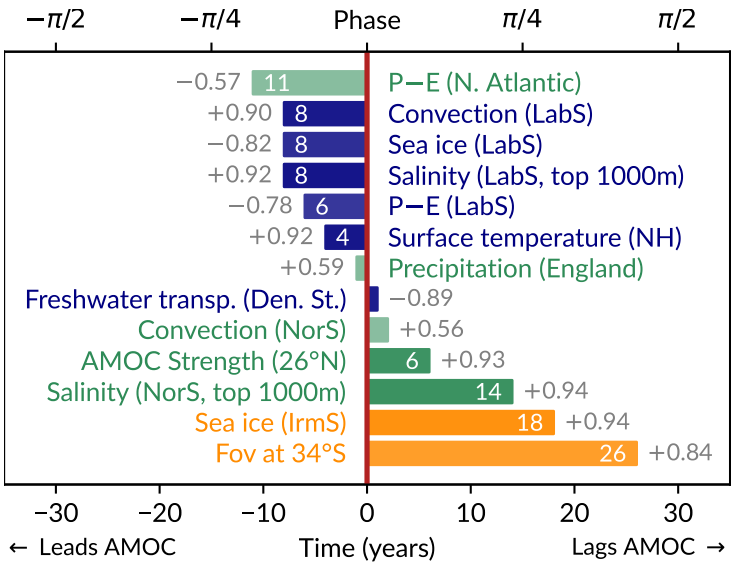

Figure 6.12 | **Lag correlations with AMOC strength** for the timeseries shown in Fig. 6.11. A negative (positive) lag time means the signal is leading (following) the AMOC at 46-66°N. White numbers inside the bars indicate the lag time; gray numbers give the correlation with the AMOC timeseries at that lag (no smoothing, annual resolution). Reproduced from Börner et al. (2025a).

deep convection undergoes large variations, switching on and off in close correspondence with the AMOC strength.

Large oscillations also appear in the sea ice cover fraction of the Labrador and Irminger Seas (Fig. 6.11c and Fig. A.4). In both regions, the sea ice retreats almost entirely during the minimum of the cycle. However, there is a clear phase shift between the two regions, as we quantify below. In the Norwegian Sea, there is little sea ice on the Edge state at all times.

The salinity in the top 1000 m of the Labrador Sea exhibits oscillations that resemble a relaxation oscillation, with a slow build-up of salt succeeded by a rapid re-freshening (Fig. 6.11d). This asymmetry in salt oscillations has also been found in other contexts where AMOC oscillations occur (Romé et al. 2025). The upper ocean salinity also oscillates in the Norwegian Sea but with a significant phase lag. Generally, the salinity changes could be caused by horizontal advection, convection, or surface freshwater fluxes. Indeed, both the meridional freshwater transport into the Atlantic basin as well as the surface freshwater fluxes in the North Atlantic mirror the AMOC oscillations. We find a particularly strongly oscillating southward freshwater transport through the Denmark Strait between Greenland and Iceland (Fig. 6.11e) and a large amplitude in precipitation minus evaporation (P−E) over the Labrador Sea (Fig. 6.11f).

In the atmosphere, many variables likewise display variability on the 120-year timescale. For example, mean precipitation (smoothed over a 20-year period) oscillates over southwest Greenland and England, yet nearly in opposite phases (Fig. 6.11g). We further find that the North Atlantic mid-latitude jet stream strength, calculated according to Woollings et al. (2010), undergoes similar oscillations (Fig. 6.11h). Overall, however, due to the much higher interannual variability in the atmosphere compared to the ocean, correlations with the AMOC strength are lower. The hemispheric mean surface temperature oscillates with an amplitude of up to 1°C on the northern hemisphere but remains relatively constant on the southern hemisphere (Fig. 6.11i).

To systematically investigate how the various oscillating variables are related to each other in time, we compute lag correlations between the AMOC strength at 46-66°N and all other variables, considering time lags between −120 and 120 years. We select all variables whose 3-year smoothed timeseries has a maximum lag correlation above 0.8 in absolute

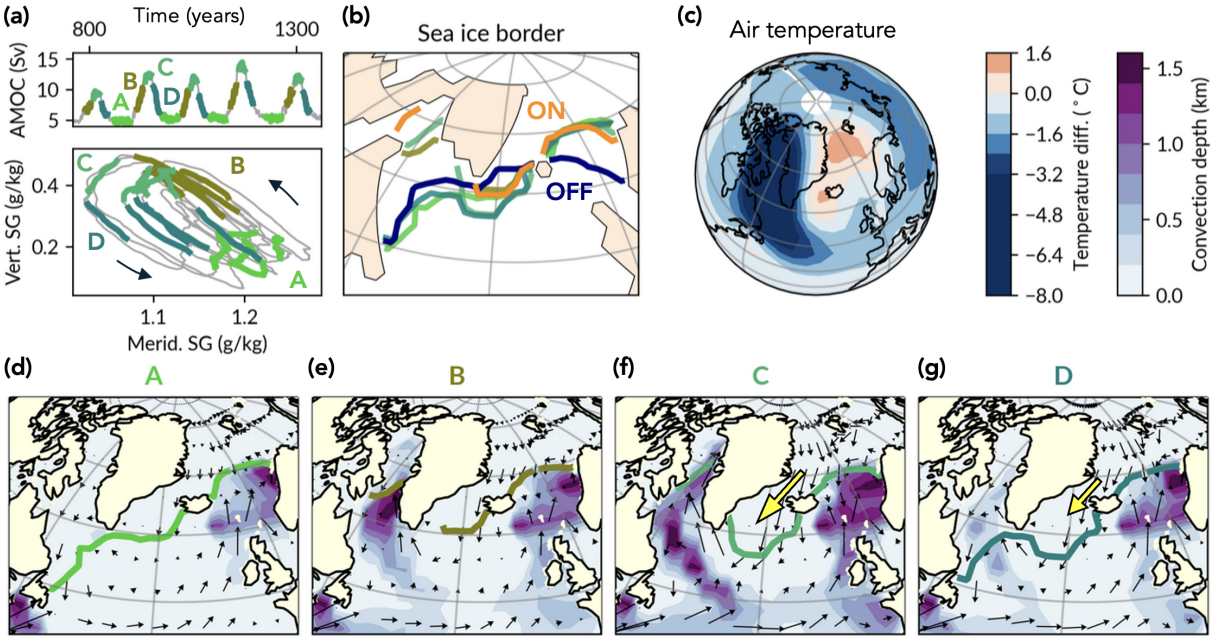

Figure 6.13 | **Phases of the AMOC oscillations.** (a) Segmentation of the final 5 oscillations into phases A-D as labeled, shown as a timeseries (top panel) and projected onto the reduced phase space of meridional and vertical SG (bottom panel). (b) Annual mean sea ice border for all phases compared with ON and OFF. (c) Surface air temperature difference for A minus C. (d)-(g) Maps of convection depth (shading), surface currents (black arrows) and the sea ice border (thick line) in the North Atlantic for phases A-D, respectively. The yellow arrow represents the strength of the freshwater flux through the Denmark Strait. Reproduced from Börner et al. (2025a).

value. Correlation values reported in the following and in Fig. 6.12 instead refer to the unfiltered timeseries, which have annual resolution.

Interestingly, the upper ocean salinity, deep convection, sea ice, and P−E in the Labrador Sea all lead the AMOC by 6-8 years, with correlations ranging between 0.78 and 0.92 (in absolute value). P−E averaged across the entire North Atlantic (between 50-80°N) has an even larger lead time of 11 years, though the correlation is less strong. Furthermore, the southward freshwater transport through the Denmark Strait is strongly negatively correlated with the AMOC strength at a lag time of one year. The AMOC strength measured at 26°N follows the AMOC at 46-66°N by 6 years, and the freshwater transport at the Atlantic southern border (34°S) lags behind by 26 years. Likewise, the upper ocean salinity in the Norwegian Sea and the sea ice cover in the Irminger Sea follow the AMOC.

Our analysis shows that the Labrador Sea is a key region in driving the AMOC oscillations on the Edge state. The fact that the salt and volume transport in the tropical Atlantic lags the AMOC strength further north suggests that the salt-advection feedback does not initiate the oscillations, though it likely plays an important role in amplifying them. Rather, ocean-cryosphere-atmosphere interactions appear crucial for triggering the AMOC cycles.

To gain further insight into the processes at play, we now divide each AMOC cycle into four phases (A–minimum, B–rise, C–maximum, D–decline; see Fig. 6.13) and consider time averages for each phase over the final five oscillations of the edge trajectory. In phase A, the Labrador Sea is ice-covered, preventing deep convection and thus maintaining a

weak AMOC. However, deep water formation in the Norwegian Sea ensures that the AMOC is not as weak as in the OFF state. In phase B, sea ice retreats in the Labrador Sea, which allows the ocean to release heat to the atmosphere and consequently deep convection to be activated. The salt-advection feedback kicks in, supplying warm and salty water to the Labrador Sea, enhancing sea ice retreat and convection up to the AMOC maximum in phase C. Then, however, the salinity and convection in the Labrador start to decrease, along with sea ice expansion. One possible reason for this reversal could be the strong freshwater influx from the Arctic Ocean through the Denmark Strait, which peaks in phase C and reaches to the Labrador Sea. The freshwater flow is concentrated in the upper ocean, implying that it can disrupt convection by freshening the upper water column. Another explanation could involve surface fluxes of heat and freshwater. In phase D, sea ice rapidly expands to cover the entire Labrador Sea and convection shuts down, causing the AMOC decline.

Fully deciphering the oscillation mechanism goes beyond the scope of this study. Nonetheless, we can identify multiple competing processes that could produce cyclic behavior: a competition between sea ice and convection in the Labrador Sea, a competition between salt advection by the AMOC and freshwater advection from the Arctic Ocean, as well as a competition between deep water formation sites in the Labrador and Norwegian Seas.

The competition between deep convection sites could explain the anti-phase pattern observed in precipitation between Greenland and England (Fig. 6.11g) as well as in surface air temperatures between the Greenland-Iceland-Norwegian (GIN) Seas and the rest of the high northern latitudes (Fig. 6.13c). During phase A, at the AMOC minimum, air temperatures are warmer over the GIN Seas than during phase C, in contrast with the surrounding areas, especially the Labrador Sea. We hypothesize that this is because the Norwegian Sea is the only deep convection zone in the North Atlantic during phase A, directing the meridional heat transport to that region and reducing cold inflows from the Arctic Ocean.

Exploring the dynamics on the Edge state thus reveals distinct modes of climate variability that are absent in the ON and OFF attractors. The edge tracking method allowed to capture the centennial climate oscillations even though they are asymptotically unstable. Despite being unstable, however, such oscillations can persist for hundreds of years on Earth system trajectories evolving near Edge state, as we show in the following section.

6.5 Boundary crisis: From Melancholia to ghost state

So far, we have investigated the global stability of the AMOC in PlaSim-LSG at constant external forcing $\boldsymbol{\lambda}$, with the CO_2 concentration set to 360 ppm. However, the radiative forcing of the Earth is currently undergoing rapid change as CO_2 concentrations are increasing at a rate of around 0.56% per year. Consequently, the stability landscape of the Earth system is continuously evolving as a function of $\boldsymbol{\Lambda}(t)$ (see Eq. (2.2)). In this nonautonomous context, attractors and edge states must be viewed in a pullback or snapshot sense as they are moving in phase space subject to the change of the control $\boldsymbol{\Lambda}$.

As we know from bifurcation theory and chapter 5, there may be critical forcing levels $\boldsymbol{\lambda}_c$ at which the global stability landscape changes qualitatively. For example, new attractors may emerge, existing ones may disappear, or attractors may switch between

periodic and non-periodic behavior. An important case are *boundary crises* where an attractor is annihilated by colliding with an edge state embedded in a basin boundary.

6.5.1 AMOC stability at increased CO_2

To gain insight into how the stability landscape of the AMOC changes under a CO_2 increase, we now consider the frozen system (i.e., fixed external forcing) at a CO_2 level of 460 ppm. Analogous to our investigation at 360 ppm, we run long simulations (4000 years), initialized from the ON and OFF state obtained at 360 ppm, respectively. Additionally, we run the edge tracking algorithm (sec. 6.3.3), likewise initialized from the ON and OFF states at 360 ppm.

We find that at 460 ppm the AMOC in our model is no longer in the bistable regime (Fig. 6.14). The OFF state at 460 ppm persists for the 4000 years of simulation and resembles the OFF state at 360 ppm in terms of AMOC strength (Fig. 6.14d). By contrast, the simulation initialized from the former ON state eventually collapses to the OFF state after a 2700-year long transient. For the first 1300 years, this trajectory (beige line in Fig. 6.14d) maintains a relatively strong AMOC, though the AMOC strength begins multi-centennial oscillations that grow in amplitude reaching up to 10 Sv. Then, the AMOC strength abruptly declines to less than 5 Sv and enters a period of large AMOC oscillations that resemble those of the Edge state in period and amplitude. After six cycles, the AMOC suddenly recovers and overshoots to 22 Sv, thereafter steeply declines again, undergoes two intermediate AMOC oscillations, and eventually falls to the OFF state where the trajectory remains for the final 1200 years of simulation.

We conclude that the AMOC is monostable at 460 ppm, with the OFF state being the only asymptotically stable attractor. What happened to the ON and Edge states?

6.5.2 Collision of ON and edge states

Despite the fact that the ON state has lost its stability at 460 ppm, edge tracking between the ON and OFF states is still possible for a while. This is because the former ON state is transiently stable for a few hundred years. We can thus find pairs of initial conditions that converge to a weak and, temporarily, a strong AMOC state, respectively. Running the edge tracking algorithm at 460 ppm produces several large AMOC oscillations that resemble the Edge state dynamics at 360 ppm, though the AMOC minimum is initially lower and the period of around 100 years is slightly shorter than at 360 ppm.

After about 750 years of edge tracking, the edge trajectory interrupts its oscillatory behavior and follows a course that is characteristic for relaxation paths from the Edge state to the former ON state. Seemingly, the edge tracking algorithm loses track of the Edge state and instead approaches the ON state. However, we know that the ON state is not an attractor anymore. Rather, the former ON and Edge states are now an intertwined chaotic object – a ghost state, as found also in chapter 5.

To see this, let us project the dynamics onto the reduced phase space spanned by the meridional and vertical SG (see section 6.3.3). At 460 ppm, the OFF state has moved to a slightly larger mean meridional SG and displays higher variability in the vertical SG compared to 360 ppm, but qualitatively it is located in the same region of the reduced phase space (Fig. 6.14b). The former ON state and Edge state now take up a larger volume in the projected phase space and are not separated anymore. We see how the

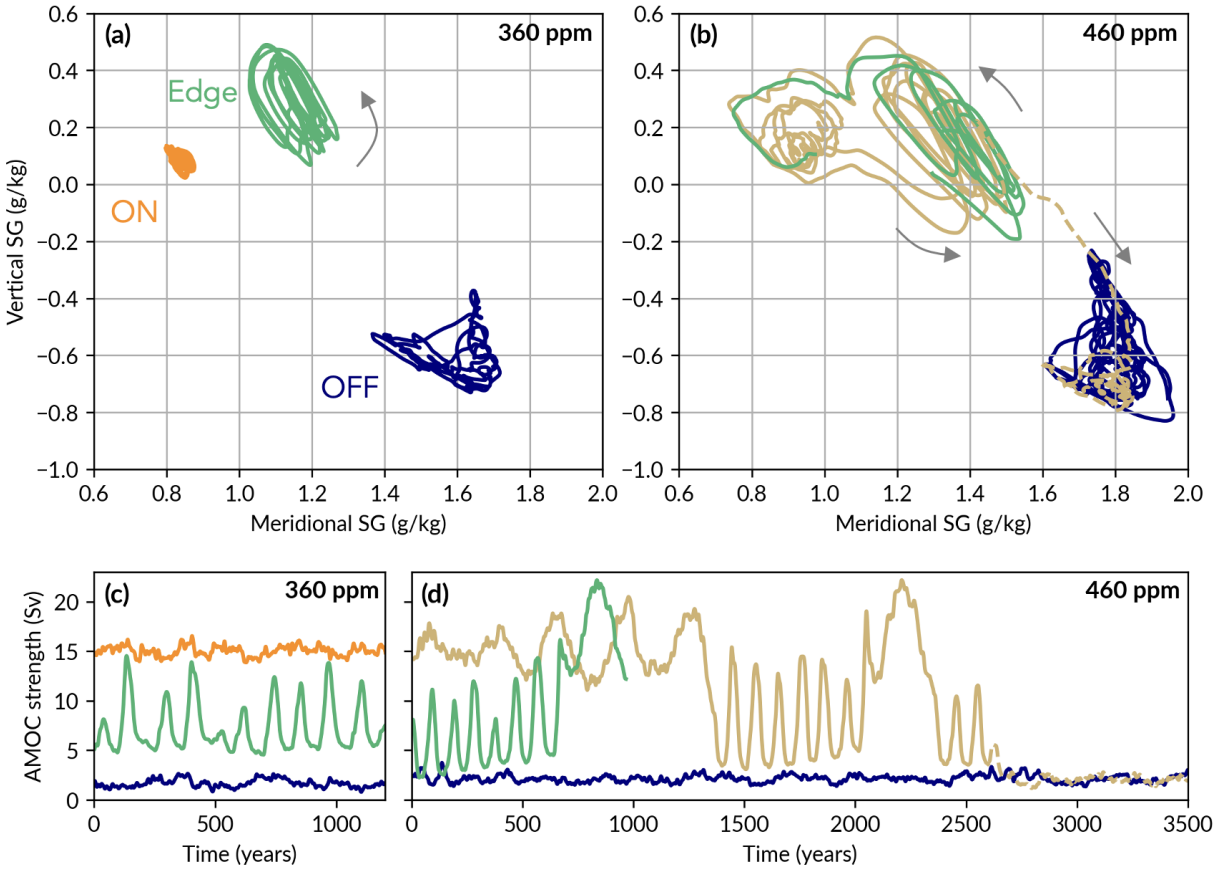

Figure 6.14 | Phase space projections and AMOC timeseries of the ON, OFF and edge states for 360 ppm (a,c) and 460 ppm (b,d). The time direction is shown by arrows. In panels (b) and (d), the trajectory initialized from the ON state is shown in a beige color to emphasize that the ON state is not an attractor anymore at 460 ppm. Reproduced from Börner et al. (2025a).

edge trajectory oscillates in a region that overlaps with the Edge state at 360 ppm and extends further to higher meridional and lower vertical SG values. Then, it transitions to the region of the former ON state. Conversely, the trajectory initialized from the former ON state circles around the ON state region and then transitions to the Edge state region, where it undergoes the same oscillations as the edge trajectory before moving to the OFF state.

Based on the phase space view taken here, we propose that the ghost state embodies the union of two interconnected phase space regions with rotational dynamics: multi-centennial oscillations near the the former ON state and centennial oscillations near the former Edge state. Since both regions are not separated in phase space, trajectories can chaotically switch back and forth between both oscillatory modes until they necessarily escape the ghost state and converge to the OFF state. Indeed, we observe this behavior in simulations initiated near the ghost state, as described below.

6.5.3 Chaotic transients

The unforced trajectory evolving under 460 ppm from the ON state existing at 360 ppm tells a tale of a simulated AMOC collapse that fundamentally contrasts the conventional picture of tipping events as abrupt, monotonic shifts (Fig. 6.15). If we would live on this

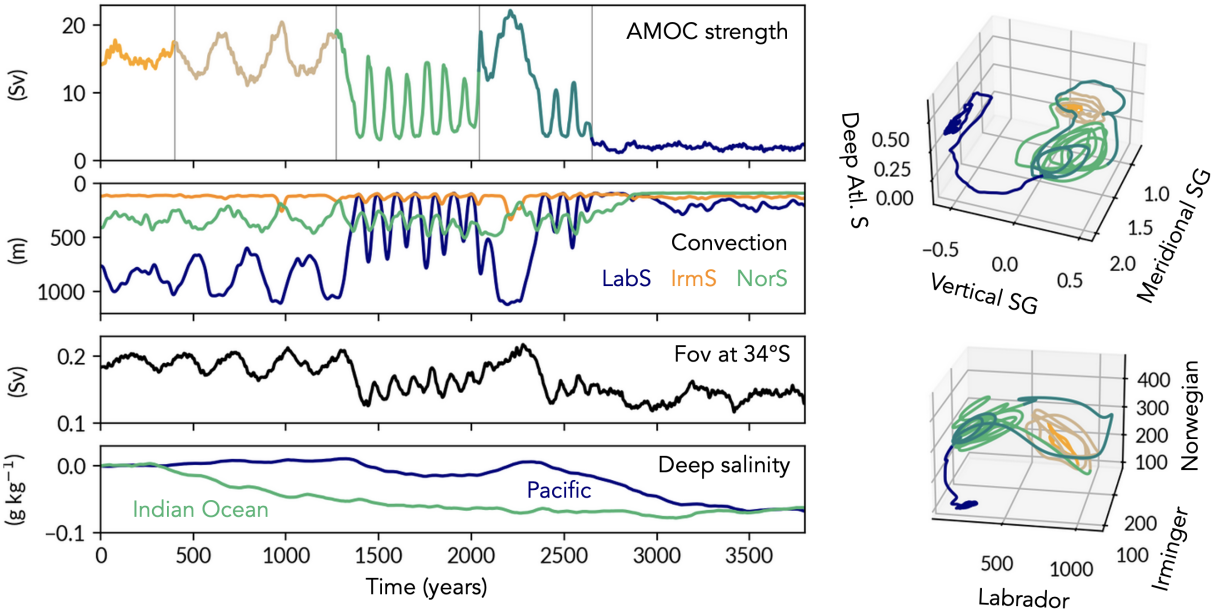

Figure 6.15 | **A pathway of transient AMOC tipping**, following an unforced trajectory at 460 ppm CO_2, initialized from the former ON state (the same as the beige trajectory in Fig. 6.14). The right panels show two different phase space projections: the proposed reduced salinity space (top, same as Fig. 6.8) and a projection onto convection depth (in meters) in different North Atlantic regions (bottom). Here Fov refers to the F_{ovS} index (see Eq. (1.4)). The deep salinity is defined as the mean salinity below 1000 m depth in the respective basin.

trajectory, it would be extremely difficult to predict the future climate. We would repeatedly anticipate a collapse, but the AMOC would keep recovering many times. Throughout the transient, the climate would undergo continued, drastic changes. Perhaps the most reliable indicators found here are the mean deep ocean salinity (below 1000 m depth) in the Indian and Pacific Oceans. Especially the Indian Ocean decreases roughly monotonically during the chaotic transient.

To further explore the transient dynamics in the monostable regime, we exploit simulations produced as part of the edge tracking procedure at 460 ppm. Specifically, we consider the ensemble of 11 simulations used for the last iteration of edge tracking before the edge trajectory jumps to the ON state region (iteration 13). These simulations are run from nearby initial conditions interpolated between one that collapsed (member 0) and another that maintained a strong AMOC (member 10) within 500 years during the previous edge tracking iteration.

The ensemble reveals a rich, non-monotonic behavior over a period of 2000 years (Fig. 6.16). Initially, all trajectories undergo a spike in AMOC strength corresponding to an excursion to the former ON state region. Thereafter, the AMOC evolution varies greatly between ensemble members. While members 0 and 9 exhibit the multi-centennial mode of variability associated with the former ON state, member 3 shows repeated episodes of Edge state-like oscillations. Member 2 sustains a relatively constant, strong AMOC strength for 2000 years, whereas members 8 and 9 feature more erratic behavior with occasional spikes of high and low AMOC strength. Only four of eleven ensemble members collapse to the OFF state within 2000 years.

This demonstrates that the transient dynamics near the ghost state is essentially unpredictable and can last for thousands of years. The long lifetime of the ghost state

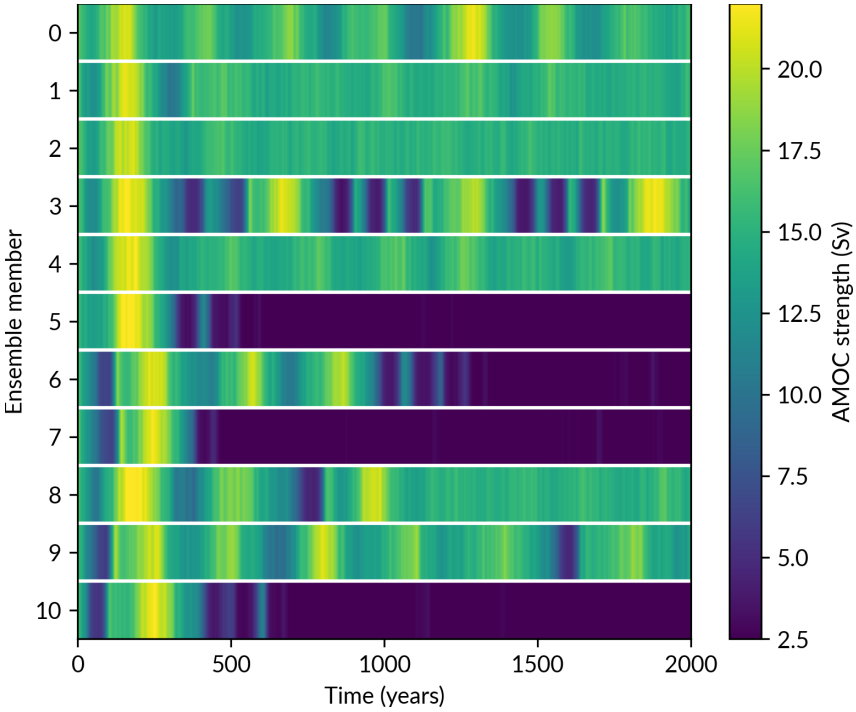

Figure 6.16 | **Transient chaos** in an 11-member ensemble of simulations initialized near the ghost state at a fixed CO_2 concentration of 460 ppm. Each horizontal panel represents one 2000-year simulation colored by the AMOC strength. Reproduced from Börner et al. (2025a).

suggests that at 460 ppm our model is close to the boundary crisis where the ON and Edge state merged, which occurs somewhere between 360 and 460 ppm. Further away from the critical CO_2 value, the ghost state is expected to have a shorter life, see discussion in (Mehling et al. 2024). Indeed, attempting to perform edge tracking at 500 and 540 ppm proved unsuccessful because the model quickly diverged from the ON state, as a result of the enhanced instability of the system.

6.6 Trajectories under future nonautonomous forcing

Our study has been focusing on the stability landscape of PlaSim-LSG at fixed external forcing, i.e., we investigated the model as an autonomous dynamical system at two different CO_2 forcing levels. However, constant forcing is a poor representation of current and projected future greenhouse gas emissions. What can our results tell us about the transition behavior of the AMOC in a non-autonomous context?

6.6.1 PlaSim-LSG: Response to SSP scenarios

To address this question, we return to the CO_2 forcing experiments introduced at the beginning of this chapter (Fig. 6.5), where we forced PlaSim-LSG with the CO_2 projections of low, intermediate, and high emission SSP scenarios. Recall that the AMOC persists under the low emissions scenario, transitions to the OFF state at high emissions, and exhibits a splitting of the simulation ensemble at intermediate emissions.

We can now inspect these simulations in the reduced phase space projection to see how their trajectories in phase space relate to the model's stability landscape, particularly the Edge state and ghost state (Fig. 6.17). For SSP1-2.6, the ensemble members virtually do not move in the phase space and remain in the region of the ON state. In the SSP2-4.5 scenario, we observe that the ensemble trajectories initially experience an increase in

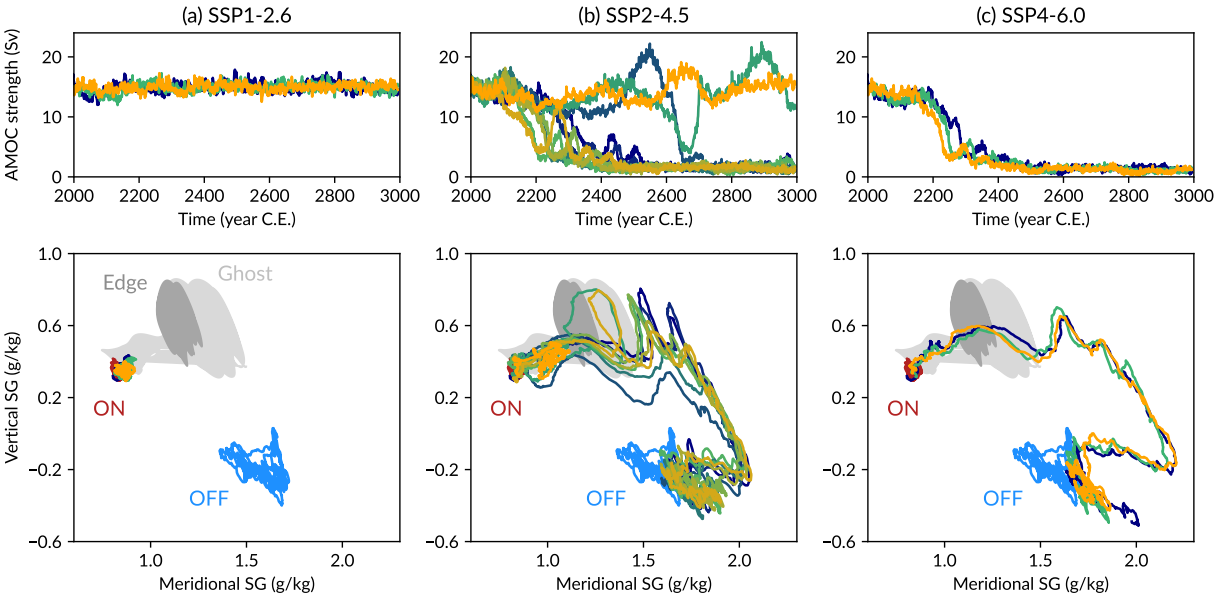

Figure 6.17 | **Trajectories of PlaSim-LSG under future CO_2 emission scenarios** for ensemble members run under the (a) SSP1-2.6, (b) SSP2-4.5 and (c) SSP4-6.0 scenario. Top panels show the AMOC timeseries, bottom panels show their projection onto the reduced phase space spanned by the meridional and vertical SG. Dark and light gray shaded areas indicate the region of the Edge state at 360 ppm and ghost state at 460 ppm, respectively. Reproduced from Börner et al. (2025a).

meridional SG, moving towards the region of the Edge state at 360 ppm. Over the 1000-year simulation period, one ensemble member remains to the left of the Edge state region, maintaining a strong AMOC. Other trajectories travel through the lower part of the Edge state region (where the AMOC is weakest, see Fig. 6.13a) and collapse to the OFF state. Yet again, some ensemble members perform one or more cycles of an oscillatory motion before converging to the OFF state. Strikingly, these oscillations occur in the region of the Edge and ghost states or to the right of it. Since the ghost state at 460 ppm extends to higher meridional SG values than the edge state at 360 ppm, it is likely that the ghost state at CO_2 levels above 460 ppm expands to even higher meridional SG values. In the SSP4-6.0 scenario, the trajectories pass straight through the ghost state region, as if the oscillatory regime of the ghost state would be "invisible" to them.

These results indicate that the Edge state and, beyond the boundary crisis, the corresponding ghost state play a key role in the ensemble splitting with respect to the AMOC strength, observed under the intermediate CO_2 forcing. Under the low emissions scenario, the trajectories do not travel to the Edge state region, while under the high emissions scenario the forcing rate is so high that the dynamical structure of the frozen system is masked. Indeed, the SSP1-2.6 scenario remains below 460 ppm (besides a short overshoot, see Fig. 6.5a), such that the ON state continues to exist. By contrast, the SSP2-4.5 and SSP4-6.0 scenarios stay above 460 ppm after the year 2050, such that we assume the model is in the monostable regime from then onwards.

6.6.2 CMIP6 model: Interpreting the "stochastic bifurcation"

The diverging AMOC behavior of ensemble members observed in PlaSim-LSG under SSP2-4.5 is intriguingly reminiscent of the so-called "stochastic bifurcation" found under

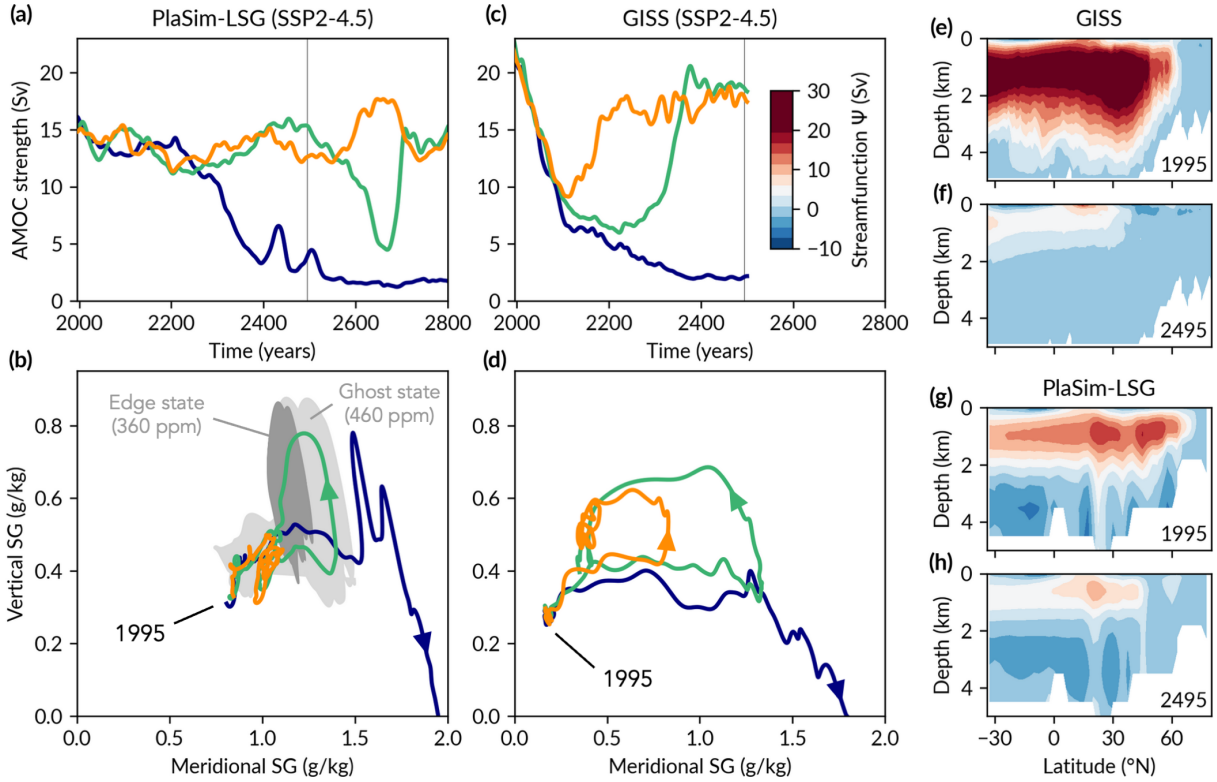

Figure 6.18 | **Comparison between GISS and PlaSim-LSG simulations** under the SSP2-4.5 scenario. (a) AMOC strength at 46-66°N and (b) reduced phase space trajectories for the three selected ensemble members in PlaSim-LSG. Arrows indicate the time direction. Dark and light gray shaded areas indicate the region of the Edge state (at 360 ppm) and ghost state (at 460 ppm), respectively. (c)-(d) like (a)-(b) but for selected ensemble members of the GISS model, colored by qualitative similarity. (e)-(h) 5-year averages of the Atlantic meridional streamfunctions along the dark blue trajectories of each model, starting in the years 1995 and 2495 as labeled. Reproduced from Börner et al. (2025a).

the same forcing scenario in the GISS model (Romanou et al. 2023). We are thus interested in whether the global stability perspective presented here can help explain the dynamics in GISS.

For this purpose, we select three ensemble members of the GISS simulations under SSP2-4.5 (members 1, 7, and 10). The simulations extend until the year 2500 and show divergent AMOC behavior (Fig. 6.18a). Following an initial AMOC weakening in all members, the first member starts to recover around the year 2100, the second recovers after 2200, whereas the third remains in a weak AMOC state until 2500 (but eventually recovers, see the Supplementary Material of Romanou et al. (2023)). Similar to PlaSim-LSG, the weak AMOC state is characterized by a collapse of the overturning cell north of 45°N while a weak overturning circulation is maintained south of 45°N (Fig. 6.18h). In the year 1995, when CO_2 levels are at 360 ppm and the AMOC is in the ON state, the AMOC is around 30% stronger in GISS than in PlaSim-LSG, though the meridional streamfunctions have a qualitatively similar shape (Fig. 6.18e, g).

To relate these simulations to our results, we likewise select three ensemble members from the SSP2-4.5 simulations with PlaSim-LSG, based on their qualitative similarity with each of the GISS ensemble member (Fig. 6.18c). The first member maintains a strong AMOC, the second undergoes a weakening to about 5 Sv followed by a recovery, and the third collapses to the OFF state within a time window until the year 2800. Note

that the evolution of the trajectories is delayed in PlaSim-LSG compared to GISS, and the initial AMOC weakening is less pronounced.

We now compare the reduced phase space trajectories of these simulations between the two models. Analogously to our PlaSim-LSG analysis, we can compute the meridional and vertical SG in GISS based on the zonally averaged salinity field in the Atlantic basin. Due to the complexity of the GISS model, its AMOC stability landscape with respect to CO_2 and the properties of potential edge states or ghost states are not known. However, we can study how the trajectories relate to the Edge state and ghost state found in PlaSim-LSG.

The reduced phase space dynamics are qualitatively very similar between the two models (Fig. 6.18b, d). The GISS trajectories start off from a significantly lower meridional SG than the PlaSim-LSG trajectories, in line with the fact that the AMOC in GISS is significantly stronger at that time. The vertical SG values are in good agreement between the models. As CO_2 forcing increases, all trajectories initially move towards larger meridional and slightly larger vertical SG values. The AMOC recovery in GISS is characterized by a counter-clockwise rotation, where the loop performs a larger excursion for the trajectory that recovers from lower AMOC values. Interestingly, the reversal of the GISS trajectory with a late recovery occurs directly in the phase space region where the Edge state is located in PlaSim-LSG, following a path that resembles that of the recovering PlaSim-LSG trajectory.

The collapsing trajectory in GISS skims the bottom end of the ghost state region before traveling to high meridional and low vertical SG values. This path is qualitatively similar to the collapsing PlaSim-LSG trajectory. Although the collapsing GISS simulation does not display any AMOC oscillations seen in the collapsing PlaSim-LSG simulation, there are still similar upward spikes in the phase space trajectory that might hint at similar, yet dampened dynamics.

To summarize, we find that the splitting of the GISS ensemble occurs in the same region of the projected phase space in which the Edge state is located in PlaSim-LSG. This supports the proposition that the "stochastic bifurcation", as the ensemble splitting is termed by Romanou et al. (2023), could indeed be mediated by an AMOC edge state in that model.

6.7 Discussion & conclusion

Our study was conducted with a climate model of intermediate complexity that inevitably relies on simplifications and neglects numerous processes of potential relevance. Therefore, the quantitative results of our investigation may be highly model-dependent and not representative of reality. On the other hand, we believe that PlaSim-LSG is to date the most comprehensive earth system model in which an edge state has been explicitly computed. Our results thus add a significant step towards realism to recent studies investigating edge states of the AMOC in a conceptual climate model (Mehling et al. 2024) and a global ocean circulation model (Lohmann and Lucarini 2024). Furthermore, the similarity in the dynamics between PlaSim-LSG and the more realistic GISS model suggests that the global stability view established here could provide key insights into the behavior of state-of-the-art earth system models. As is increasingly clear, AMOC metastability and tipping behavior is not restricted to simple climate models but occurs across the model hierarchy (Romanou et al. 2023; van Westen and Dijkstra 2023; Cini

et al. 2024; Willeit and Ganopolski 2024). The phase space of comprehensive models is just more challenging to explore.

Our work presents a proof of concept that analyzing edge states using the edge tracking algorithm is a powerful method of global stability analysis that works robustly from two to 10^5 degrees of freedom. The key limitation for its application in even higher-dimensional systems is the computational cost of running long simulations. In order to produce 1400 years of edge trajectory in PlaSim-LSG at 360 ppm, around 70 000 years of simulation (roughly 3000 CPU hours) were required in the procedure, giving an efficiency of 2%. However, this number depends strongly on the system, particularly how quickly trajectories diverge from the basin boundary. Improving the efficiency of edge tracking could be a valuable objective of future work.

From a dynamical systems perspective, it is remarkable that the edge tracking algorithm converges to an edge state despite the complex geometry of the basin boundary, which is typically fractal (Bódai and Lucarini 2020; Lucarini and Bódai 2017; Mehling et al. 2024). In our study, the assumption of a fractal basin boundary is supported by the non-monotonicity of the asymptotic state of initial conditions interpolated across the basin boundary. Lucarini and Bódai (2017) discovered a fractal basin boundary in a model where the dynamical core of PlaSim, PUMA, was coupled to an energy balance model. Moreover, it is interesting that edge tracking continues to work for multiple iterations at 460 ppm where the model is in the monostable regime and a basin boundary does not exist. This connects our results with the original application of the algorithm, tracking the edge to turbulence in planar shear flow (Skufca et al. 2006), where the turbulent state is an asymptotically unstable, ultralong transient and the laminar state is the sole true attractor. From an earth system modeling perspective, it is not obvious that the interpolation between initial conditions in all dynamical variables yields new initial conditions that generate numerically stable and physical trajectories. We argue that even if a new initial condition obtained by interpolation is unphysical, the structure (convexity) of the physical equations governing the climate dynamics ensures that the trajectory quickly relaxes to a physical state.

The AMOC edge state investigated by Lohmann and Lucarini (2024) was constructed in the global ocean circulation model Veros (Häfner et al. 2018), an ocean-only model where interactions with the atmosphere and cryosphere were not dynamically represented. Here, including atmosphere-ocean-sea ice feedbacks revealed a much richer dynamics on the edge state. Nonetheless, some consistent properties emerge from comparing the two studies. Lohmann and Lucarini (2024) emphasized that the edge state features a less "spicy" (i.e., colder and fresher) deep North Atlantic than the attractors. This aligns with our results (see Fig. A.3). Based on their results, Lohmann and Lucarini (2024) concluded that the most relevant regions for anticipating AMOC transitions are located in the deep sea, which is impractical for observations. Our results instead suggest that many eccentric observables are found also in the surface ocean when interactions with the atmosphere and cryosphere are considered. This could potentially be exploited for improved early warning systems of AMOC changes (Lohmann et al. 2025; Lenton et al. 2024).

AMOC oscillations have received wide interest in the paleoclimate community due to their potential for explaining past abrupt climate change in the northern hemisphere (see section 1.2.1). In the paleoclimate context, studies often focus on determining "sweet spots" for oscillations in parameter space (Malmierca-Vallet et al. 2024). Here, we demon-

strate a sweet spot for oscillations in phase space: while the ON and OFF states do not exhibit oscillations at 360 ppm, strong unstable oscillations occur near the edge state. The drivers of the unstable oscillations on the edge state in PlaSim-LSG involve many processes that have previously identified in oscillation mechanisms, such as deep convection at different locations, sea ice interactions, and transports between the Arctic and Atlantic Oceans (Romé et al. 2025). As a perhaps novel variant of these mechanisms, our results indicate a teleconnection between the deep convection sites of the Labrador and Greenland Seas, possibly mediated by the Denmark Strait throughflow. Dynamically, the presence of unstable oscillations near an instability might hint at the existence of a subcritical Hopf bifurcation, similar to the situation in the Rooth model (chapter 3, see Fig. 3.4).

The AMOC overturning strength is the most common quantity considered in the study of AMOC metastability. For instance, Cini et al. (2024) used the AMOC strength as their score function when applying a rare event algorithm to simulate spontaneous AMOC transitions. However, our results suggest that the AMOC strength may be a poor indicator of AMOC stability due to the oscillations and long transients. To better understand the AMOC and develop efficient rare event algorithms, it is thus important to consider additional variables. Our results may help inform the choice of suitable indicator variables.

The limited predictability of the AMOC near an instability has already been suggested by Knutti and Stocker (2002). The studies by Lohmann, Wuyts, et al. (2024) and Romanou et al. (2023) have recently reiterated this idea by demonstrating an ensemble splitting. In the context of the Snowball Earth transition, Kaszás et al. (2019) found an ensemble splitting in PlaSim (without the dynamic LSG ocean). Our findings allow to link this behavior the presence of an edge state and, beyond the boundary crisis, a ghost state. We can thus directly link the dynamics of earth system models to fundamental concepts of dynamical systems theory that are often only explored in low-dimensional systems.

Even though the AMOC edge state found in PlaSim-LSG is a physically reasonable steady-state, the physical processes described in this chapter may differ from reality. First, the thermodynamic sea ice module is rather simplified and may unrealistically amplify the nonlinear behavior near the edge state. Second, the convective adjustment scheme can only crudely approximate true convection processes. Moreover, PlaSim-LSG does not resolve ocean eddies due to the coarse resolution, which may play an important role for AMOC stability. Despite the limitations of the model, we note that we find a critical threshold of the AMOC with respect to atmospheric CO_2 concetrations at less than 460 ppm. This level is expected to be reached before the year 2050, supporting the possibility that the AMOC loses stability within this century – even if the subsequent transition could take longer to play out.

Chapter 7

Synthesis & outlook

7.1 Summary **152**

7.2 Research outcomes **154**

7.2.1 Assessment of research questions 154

7.2.2 Main contributions 157

7.2.3 Open challenges and future directions 157

7.3 Concluding remarks **158**

7.1 Summary

This thesis is concerned with the metastable behavior of the Atlantic overturning circulation, a key component of Earth's climate system. Taking a dynamical systems perspective, our aim was to explore the stability landscape and transition behavior of the AMOC in different climate models. As a unifying theme, the analysis focuses on the role of edge states, or Melancholia (M) states, for critical transitions between competing AMOC states.

Chapters 1 and 2 provide the context and mathematical background that motivate the work. In light of ongoing climate change, it cannot be ruled out that the AMOC will substantially weaken or even collapse within the coming decades, posing a severe climate risk. Despite intensive research on the AMOC, the uncertainties about the future fate of the AMOC remain large. To address the challenges in detecting and predicting critical AMOC behavior, we propose a research strategy that takes a global perspective on the system's dynamics in the phase space.

The concepts and methods underlying our approach stem from dynamical systems theory. In chapter 2.2.4, we review multistability and metastability, multiscale chaotic and stochastic dynamics, as well as large deviation theory. We further introduce the edge tracking algorithm and numerical techniques for sampling transition paths and computing quasipotentials.

Chapter 3 presents a motivating example of how the dynamical systems approach introduced in chapter 2.2.4 can be put to practice in a box model of the AMOC. We investigate Rooth's interhemispheric 3-box model, a five-dimensional non-smooth system of ordinary differential equations. The original formulation of the model features a singular edge state, which we fix by introducing an additional salt restoration term. Due to the different timescales of temperature and salinity relaxation as well as advection, this is a multiscale system.

Using numerical continuation, we explicitly compute the bifurcation diagram of the model with respect to the freshwater flux into the North Atlantic box. Subcritical Hopf bifurcations bound a bistable region in which the edge state is either a saddle point or an unstable limit cycle. We probe the global stability of the model with freshwater shock experiments and demonstrate that long transients can emerge close to the M state.

Additionally, we examine the transition behavior of the model under stochastic forcing. Noise-induced transitions driven by isotropic Gaussian noise display the expected Kramers scaling of the mean first-exit time. Surprisingly, however, the transition paths bypass the M state in the salinity variables, despite weak noise. Under freshwater forcing subjected to α-stable Lévy noise, we observe a fundamentally different transition behavior dominated by single extreme events, as expected from the probability distribution of the noise increments.

Chapter 4 takes a closer look at noise-induced transition paths avoiding the saddle, a phenomenon discovered unexpectedly in the previous chaper. In the zero-noise limit, Freidlin-Wentzell theory states that transition paths between competing attractors must cross the relevant basin boundary at the quasipotential minimum on the boundary, typically attained at an M state. However, we show that timescale separation can cause saddle avoidance, even under weak yet finite noise. To this end, we study the transition

behavior in two simple two-dimensional fast-slow systems: the bistable FitzHugh-Nagumo model with additive noise and a population dynamical system of two competing species with multiplicative noise.

Computing most probable transition paths (instantons) and quasipotentials, we quantify and explain differences between Freidlin-Wentzell theory and the statistics of sample transition paths. To resolve the disagreement, we present a pathspace sampling technique based on the Onsager-Machlup action that allows to accurately compute most probable transition paths and the transition path distribution under finite noise. While this solution is limited to additive noise, we can transform the competing species model to a stochastic differential equation with additive noise, allowing to obtain a minimizer of the Onsager-Machlup action also in that case.

Chapter 5 brings us back to the AMOC with an investigation of a conceptual climate model. This time, instead of applying stochastic forcing to the ocean, we explicitly model the ocean-atmosphere interactions as chaotic forcing. Specifically, we consider the Stommel box model bi-directionally coupled to the Lorenz '84 model. This yields a five-dimensional system of ordinary differential equations mimicking a slow, bistable ocean and a fast, chaotic atmosphere.

First, we calculate the global quasipotential landscape of the uncoupled Stommel model by a stitching technique joining the local quasipotentials with respect to each of the two attractors of the model. Then, we focus on the chaotic dynamics of the coupled system as a function of additional freshwater input. Using the edge tracking algorithm, we explicitly compute the M state, a chaotic saddle separating the two chaotic attractors, for a range of freshwater forcing values. This allows us to construct the bifurcation diagram of the coupled model.

In the bistable regime, we find that the M state is embedded in a fractal basin boundary. The fractal dimension is computed from the mean lifetime of the M state and its maximum Lyapunov exponent, showing that the basin boundary has almost the full phase space dimension. This indicates that the predictability of the asymptotic state of the AMOC is virtually lost in an extended region around the basin boundary. Increasing the freshwater input beyond a critical threshold, we study long chaotic transients and quantify their exponentially distributed lifetime as a function of the distance to criticality. Based on this, we propose an approximate method to estimate the transition probability under temporary forcing overshoot scenarios. The results underscore the need for a probabilistic notion of a safe operating space.

Chapter 6 closely follows the problem setting of chapter 5, now using a much more complex climate model[1]. Specifically, we investigate the metastability of the AMOC in PlaSim-LSG, an intermediate-complexity general circulation model featuring dynamic ocean, atmosphere, sea ice and hydrological cycle components.

For two different atmospheric carbon dioxide (CO_2) levels, we employ edge tracking to construct the Melancholia (M) state that separates the strong and weak AMOC attractors found in the model. While being unstable, the M state can govern the transient climate for centuries. We find that the M state exhibits strong AMOC oscillations on centennial timescales driven by oceanic convection, sea ice, and salt advection in the North Atlantic Subpolar Gyre.

[1]This text is partly taken from (Börner et al. 2025a)

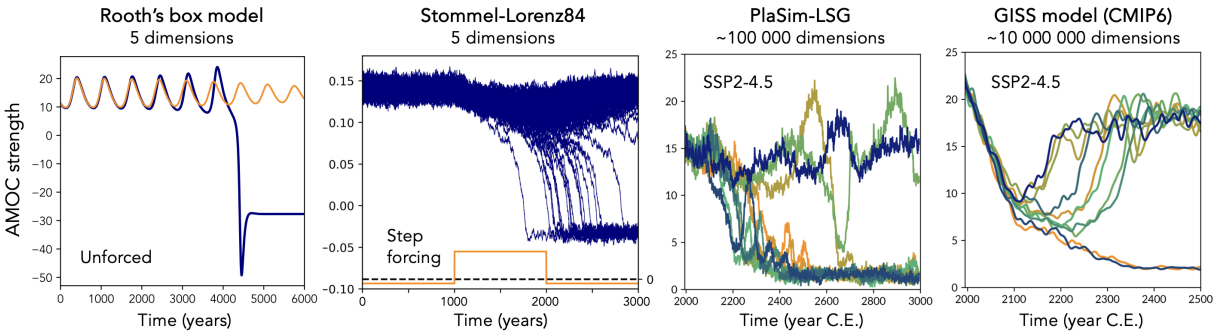

Figure 7.1 | **AMOC on the edge across a hierarchy of climate models.** From five to millions of dimensions, the possibility of an ensemble splitting near a critical thresholds highlights a fundamental limit to future climate predictability.

Combining simulations with fixed and time-dependent CO_2 forcing, we demonstrate that in this model the AMOC undergoes a boundary crisis at CO_2 levels projected to be reached in the next decade. Under an intermediate CO_2 emissions scenario, long chaotic transients and a splitting of ensemble trajectories occur, which we explain in terms of the "ghost" of the M state. We link this phenomenology to the chaotic AMOC behavior recently observed in state-of-the-art earth system models.

7.2 Research outcomes

7.2.1 Assessment of research questions

Q1 What does the stability landscape of the AMOC look like? Is the picture obtained from conceptual climate models consistent with the dynamics of comprehensive earth system models?

The work in this thesis was driven by the goal of quantifying the stability landscape of the climate system, particularly of the AMOC, and its dependence on forcing changes. While constructing the full landscape still remains out of reach for high-dimensional complex systems, there are different ways of how we can nonetheless gain an insight into its structure. By pursuing these approaches, we obtained a more nuanced picture of the global stability of the AMOC beyond the paradigm of a double-well potential.

One way to start is by analyzing the stability landscape in low-order models where computing the quasipotential is feasible. In Stommel's box model, we obtained the global quasipotential, offering a complete and quantitative view of the stability landscape for that model. Our quasipotential analysis of the two toy models in chapter 4 suggests that the presence of multiple timescales can lead to extended quasipotential plateaus – flat regions that can be relevant also for the multiscale dynamics of the AMOC, e.g. leading to saddle avoidance in Rooth's box model.

If the full quasipotential cannot be computed, we can still target phase space regions of interest, particularly Melancholia states. By analyzing the dynamical properties of M states in different climate models, we gained an understanding of the structure of the associated basin boundary. Interactions between the slow, multistable ocean dynamics and the fast, chaotic variability of the atmosphere have an intricate effect on the "fine

structure" of the stability landscape. For example, we found indications of a fractal basin boundary both in the Stommel-L84 and PlaSim-LSG models. Furtermore, we showed how the stability landscape near the edge state changes during a boundary crisis, giving rise to a long-lived yet globally unstable ghost state.

Despite the greatly different dimensionality and level of detail between the Stommel-L84 and PlaSim-LSG models, their metastable dynamics exhibited a striking (qualitative) similarity. Combined with the similarities shared between PlaSim-LSG and the more sophisticated GISS model, this supports the hypothesis that reduced-order models can help understand the stability landscape of state-of-the-art earth system models. At the same time, the Rooth model (chapter 3) exemplified that even the qualitative features of the stability landscape can depend strongly on the parameter choices of the model, highlighting the problem of model biases.

Q2 Can we determine the Melancholia state(s) separating competing AMOC states in a high-dimensional climate model? If so, what are the dynamical, physical and climatological properties of such state(s)?

Using the edge tracking algorithm, we were able to construct a Melancholia state separating the two stable AMOC states in the intermediate-complexity climate model PlaSim-LSG. To the best of our knowledge, this is to date the most complex and high-dimensional climate model in which an M state, and more generally a chaotic saddle, has been explicitly computed.

Repeating the edge tracking procedure from similar but different initial conditions led to the same M state. However, in principle it is possible that additional M states exist which could be found by further varying the initial states of the edge tracking procedure.

Using up to 1200 years of a pseudotrajectory approximating the M state, we obtained a detailed picture of its climate. The M state is in a non-equilibrium steady state in terms of the global energy budget but has a significantly higher oceanic center of mass than the two stable attractors, indicating a higher potential energy. The dynamics on the M state are characterized by strong ocean-sea ice interactions involving convective and advective processes in the North Atlantic and Arctic Oceans, leading to strong unstable AMOC oscillations with a period of roughly 120 years.

Q3 What is the role of Melancholia states for metastable dynamics, transition paths and long transients under stochastic and chaotic forcing?

Melancholia states play a fundamental role for the metastability and transient behavior of complex multistable systems. In many cases, M states mediate critical transitions, acting as gateways between competing attractors. Yet, our work on saddle avoidance (chapter 4) also demonstrated that transitioning via an M state is not guaranteed. We conjecture that transition paths may generally bypass the relevant M state if (a) the noise is not sufficiently weak and/or, in the case of nonautonomous deterministic forcing, (b) the forcing rate is not sufficiently small. In multiscale non-gradient systems with a multistable slow manifold, it is possible that any realistic amplitude of stochastic perturbations leads to saddle avoidance. This challenges the classical picture derived from gradient systems and emphasizes the need to go beyond asymptotic theories. Still, our results suggest that transition paths generally cross

basin boundaries via a quasipotential plateau associated with an M state. In higher-dimensional multiscale systems, it is possible that transitions pass near an M state along the fast degrees of freedom while avoiding it along the slow ones.

In section 2.1.7, we summarized the role of M states in each of the three tipping mechanisms of B-, N-, and R-tipping. While this classification is conceptually useful, real tipping scenarios arguably result from a combination of these mechanisms. In the case of pure bifurcation-induced tipping due to a saddle-node or subcritical Hopf bifurcation, an attractor disappears by colliding with an M state. Noise-induced transitions become increasingly likely near such a bifurcation, since the M state then tends to be close to the attractor in phase space. Rate-induced tipping in the sense of Wieczorek et al. (2023) depends on which side of the moving basin boundary the trajectory ends up on, and close to a bifurcation this moving basin boundary likely corresponds to the stable set of the moving M state. Thus, considering that critical transitions in forced systems often occur due to joint effects of noise, rate, and crossing bifurcations of the underlying frozen system, the dynamics of Melancholia states can become relevant for nonautonomous trajectories in multiple ways.

The finding that chaotic saddles are associated with chaotic transients is well-known in the literature (Lai and Tél 2011) but its relevance for climate science has so far not been widely recognized. We have shown that long transients appear also in ocean-atmosphere dynamics due to the presence of chaotic Melancholia states and potential boundary crises. Notably, in PlaSim-LSG long transients occurred under a plausible CO_2 emissions scenario, even though this overshoot scenario temporarily exceeds the critical CO_2 level by more than 100 ppm. This suggests that tipping windows (Römer and Ashwin 2025) can be relatively wide regimes in parameter space where the ghost of an M state governs the transient dynamics.

Q4 What can we learn from Melancholia states about the global stability of the Earth system, transition mechanisms and the predictability of our future climate?

Ultimately, we would like to predict the phase space trajectory of the Earth system under future climate forcing. In the context of multistability, we are interested in the likely transition paths leading from our current climate state to an alternative competing state. Knowing the transition path allows us to infer transition mechanisms, understand the transient climate conditions during the transition, and possibly anticipate the transition in advance. This would provide valuable information for mitigation and adaptation strategies.

Since calculating most likely transition paths is often unfeasible, analyzing M states marks an important step towards this goal. Locating M states in phase space can yield indications of the rough direction of transition paths, and their stable sets divide the phase space. Understanding the dynamics on the edge state and the structure of the basin boundary offer an interpretation for the chaotic and unpredictable behavior encountered in state-of-the-art climate simulations. Indeed, in all climate models we studied, the edge state mediated an ensemble splitting of AMOC trajectories (Fig. 7.1). Our results in chapter 6 suggest that the dynamical properties of the AMOC edge state are, to some extent, conserved under varied external parameters: the unstable AMOC oscillations in the present-day climate of PlaSim-LSG "foreshadowed" the oscillatory transient dynamics of the AMOC at higher CO_2 concentrations.

7.2.2 Main contributions

In summary, the present work makes the following scientific contributions:

- Demonstrating saddle avoidance of noise-induced transitions in a geoscientific context; linking saddle avoidance to timescale separation in multiscale systems and elucidating its occurrence via the quasipotential of Freidlin-Wentzell theory.
- Proposing a pathspace sampling approach to compute transition path distributions and most probable transition paths under finite noise, based on a variational formulation in terms of the Onsager-Machlup action.
- Highlighting fundamental limits to predictability of the future AMOC evolution in a conceptual climate model, by showing that the model features a fractal basin boundary and a ghost state
- Explicitly computing a Melancholia state of the AMOC in a fully coupled climate model with 10^5 degrees of freedom; characterizing its dynamical and physical properties.
- Providing evidence of a boundary crisis of the AMOC with respect to CO_2 in an intermediate-complexity climate model; illustrating the resulting ghost state and using this knowledge to interpret the transient, noncoherent behavior observed in earth system models.
- Advancing the application of dynamical systems theory and methods like edge tracking in Earth system science, deepening our understanding of the Earth system's stability landscape.

7.2.3 Open challenges and future directions

The probability of a future AMOC transition and our proximity to potential tipping points in the real Earth system remain unknown. To better understand the climate models we rely on, we argue that there is a value in extending the global perspective presented here to other models and climate subsystems. The edge tracking algorithm presents a versatile method to compute Melancholia states. Knowledge of the edge state could inform the score function of rare event simulations, ideally allowing to sample transitions in large climate models more efficiently. Along the lines of Lohmann et al. (2025), Melancholia states could further help improve early-warning indicators.

In principe, the edge tracking algorithm could also be run under nonautonomous forcing, opening up the possibility of constructing a pullback Melancholia state. The minimization problem of the Onsager-Machlup action could be exploited to construct finite-noise, finite-time quasipotentials. Such a stability landscape that refers to a specific timescale of interest could be more relevant for applications than the Freidlin-Wentzell quasipotential derived in the weak-noise limit.

To overcome the barriers of calculating quasipotentials in higher dimensions, machine learning approaches could help that learn the dynamics under the constraint of the orthogonal decomposition of the flow (Lin et al. 2021). There is also a need to better understand transition pathways under realistic boundary conditions. To work towards

this goal, it seems promising to further develop a comprehensive theory and toolbox of large deviations for degenerate, anisotropic and multiplicative noise.

Overall, our results show that results of dynamical systems theory hold also in highly complex climate models consisting of millions of lines of code rather than a closed-form equation. The apparent gap between the mathematical treatment of low-dimensional climate models and analyses of state-of-the-art models can thus be overcome. At the same time, the complexities of the Earth system mean that asymptotic theories (e.g. in the limits of weak noise, infinite timescale separation, and adiabatically slow forcing) only give approximations that may fail in practice. In many cases, it has been possible to develop a theory beyond these asymptotic limits, and future research in this direction is likely to yield improvements of high relevance for the real world.

7.3 Concluding remarks

So, how close are we? — When telling others about my research topic, I get this question a lot. Many people have heard about climate tipping points or read catastrophic news stories about an impending AMOC collapse. After all, it is the year 2025, and according to a headline[2] of The Guardian, a misleading account of Ditlevsen and Ditlevsen (2023), the shutdown could happen this year.

In the simplified climate model used in chapter 6, we found that the current AMOC state becomes unstable at CO_2 values exceeding 460 ppm. This level is projected to be crossed during the next 10 to 20 years. The results are consistent with other recent results, e.g. by Ditlevsen and Ditlevsen (2023) and van Westen et al. (2024). Despite the strong assumptions and limitations underlying these estimates, ignoring the risk of an AMOC transition would be irresponsible.

Given the serious risk, there is an understandable yearning for certainty. Tipping points are often discussed as a binary problem: either we cross a tipping point, with immediate catastrophic consequences, or we avoid tipping and remain safe. However, already today the climate crisis poses one of the gravest threats to human wellbeing, whether the AMOC will tip or not. The existence of tipping points aggravates the threat, since the abruptness and potential irreversibility of tipping events makes mitigation and adaptation more difficult.

Ultimately, the timescales of climatic change are crucial for human impacts. A collapse of the Greenland ice sheet, playing out over millennia, may be defined an abrupt event on glacial timescales but appears slow compared to a human lifetime. The fact that the AMOC could shut down within decades to a century makes it a particularly relevant tipping element. However, the existence of long transients requires to distinguish between the trigger of a transition, such as a loss of stability due to crossing a critical threshold, versus the time of the actual abrupt change. These two timings may lie centuries apart. In fact, even if the AMOC would lose stability this year, it is likely that we would not realize it for a while.

For Reading and the UK, the question of an AMOC transition could imply opposite directions of adaptation: if the AMOC is sustained, the UK is expected to warm and become wetter. Should the AMOC substantially decline, the trend over the coming

[2] Source: https://www.theguardian.com/environment/2023/jul/25/gulf-stream-could-collapse-as-early-as-2025-study-suggests?, visited 5 January 2025.

decades could in fact lead towards a colder and drier climate (Ritchie et al. 2020). Due to the intrinsic limits to predictability illuminated in this thesis, we should not wait for the perfect prediction. As Ditlevsen and Johnsen (2010) pointed out, early warning might be wishful thinking. Instead, the different futures present possible storylines society should prepare for. Despite irreducible uncertainties, we hoped to demonstrate with this work that the mathematics and physics of planet Earth can make an important contribution to informing risk management.

In the University of Reading Fairbrother Lecture, I compared the situation with a game of "climate roulette". Imagine two guns, each with many bullet chambers. One gun is loaded with only a few bullets, representing a low-emissions future. The other gun contains many more bullets and corresponds to a high-emissions future. In other words, the more greenhouse gases we emit, the more bullets we are loading into the gun. Given that we are inevitably playing climate roulette, which gun would you play with? In either case it is uncertain whether the round will fire, but in the high-emissions case, the probability is significantly higher. This shows that despite uncertainty, we have the agency to reduce the risk of climate transitions by cutting greenhouse gas emissions. In that regard, a deeper, nuanced mathematical understanding of climate dynamics can also help finding a balance between denial and doomism.

Appendix A

Supplemental figures

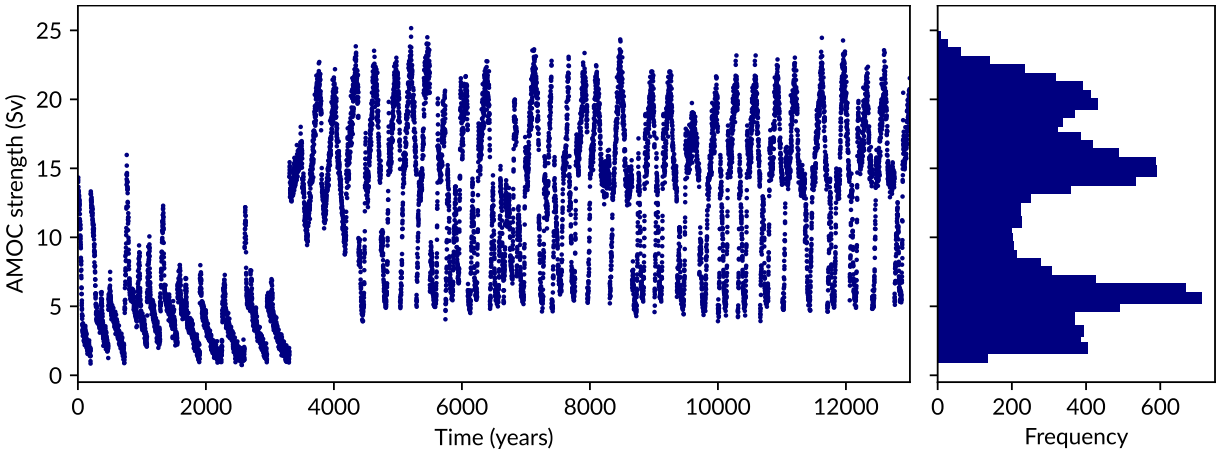

Figure A.1 | **Edge tracking simulations used for the EOF analysis** shown in Fig. 6.7. Left: Concatenated AMOC timeseries for all simulations used. Right: Histogram of AMOC strength including all data points shown on the left panel (annual means). Reproduced from Börner et al. (2025a), Supplemental Information.

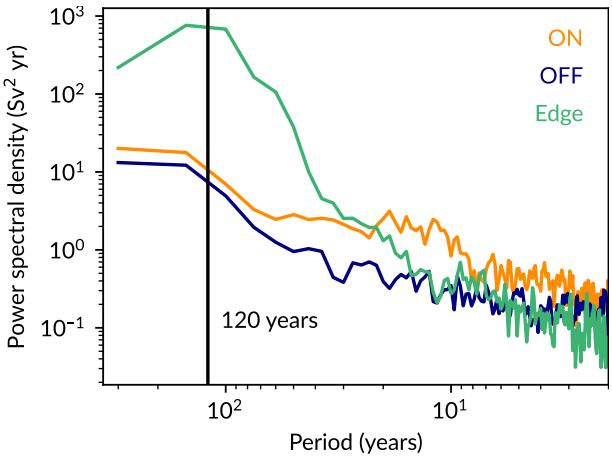

Figure A.2 | **Power spectra of the AMOC states** in PlaSim-LSG at 360 ppm CO_2. The edge state (green) has a higher centennial variability (peaking at a period of around 120 years) compared to the ON (orange) and OFF (blue) states, reflecting the large AMOC oscillations. The ON state exhibits slightly elevated decadal variability compared to the OFF and edge states.

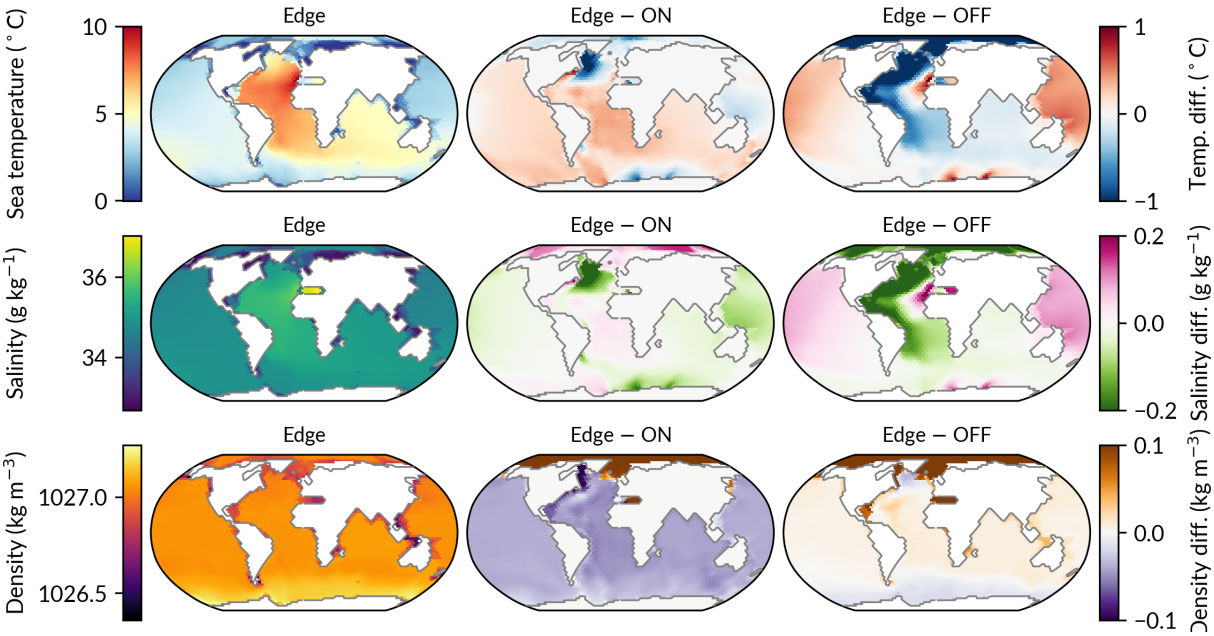

Figure A.3 | **Deep sea properties of the edge state** in PlaSim-LSG at 360 ppm CO_2. Same as the first three rows of Fig. 6.10, but for the deep ocean (averaged over all depths between 1000 and 3000 m).

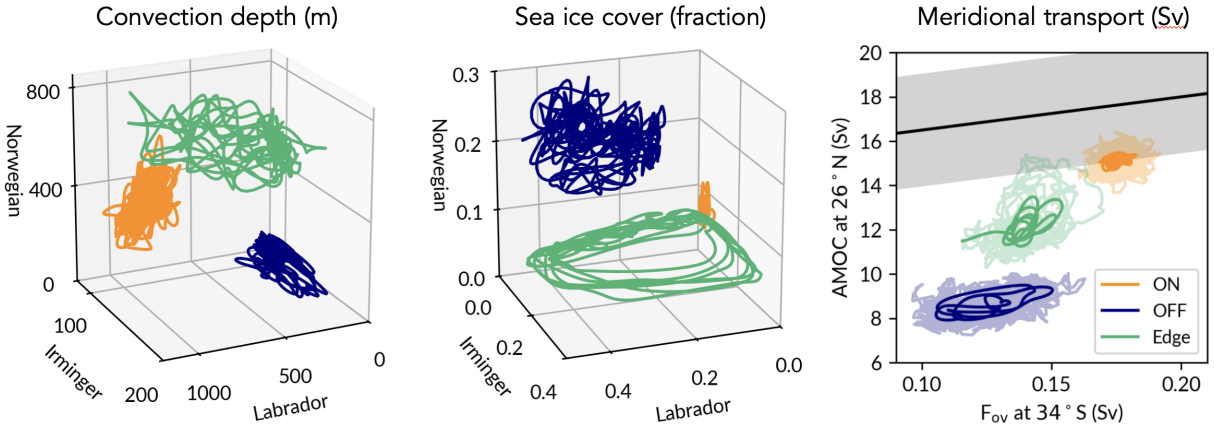

Figure A.4 | **Additional phase space projections of the AMOC states** in PlaSim-LSG. Left: Convection depth by deep water formation region (LabS - Labrador, IrmS - Irminger, and NorS - Norwegian). Center: Similar to the left panel, but for sea ice cover fraction. Right: AMOC strength at 26°N against F_{ovS}, the freshwater transport due to overturning into the Atlantic at 34°S (compare with Fig. 4D of van Westen et al. (2024)). The black line indicates the linear regression across CMIP6 models in the time range 1994-2020, with the gray shading representing one standard deviation.

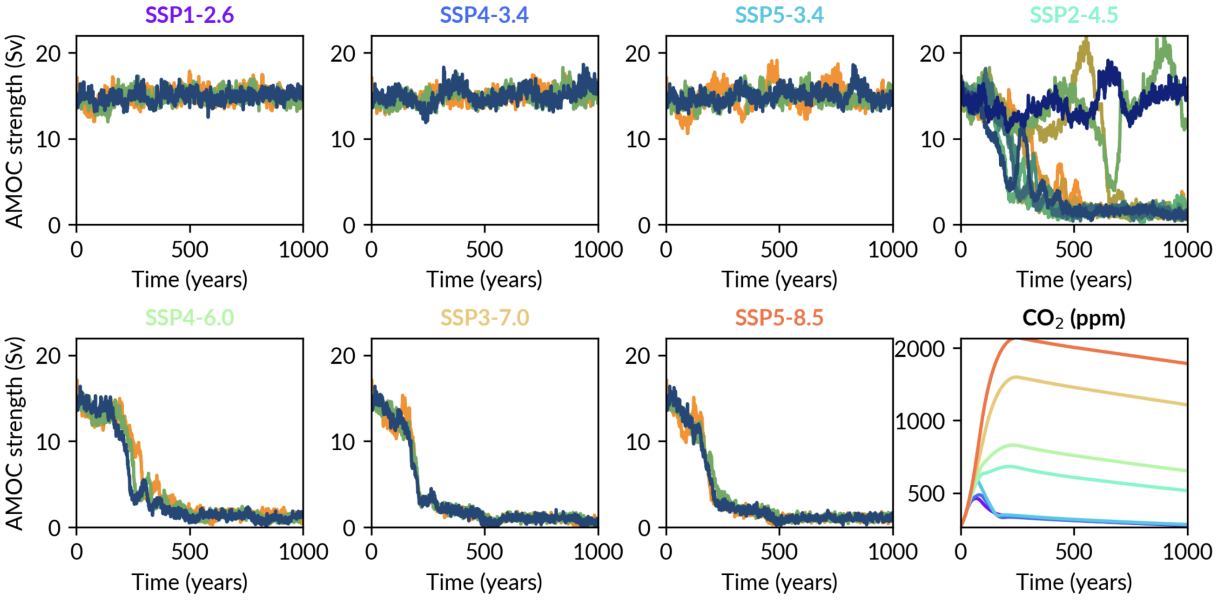

Figure A.5 | **AMOC response to extended SSP scenarios** in PlaSim-LSG. Each panel shows three ensemble members, except for SSP2-4.5 where 10 ensemble members are shown. The CO_2 concentration as a function of time is shown for each scenario on a logarithmic CO_2 scale (bottom right; coloring corresponds to subpanel title color). See also Fig. 5.13.

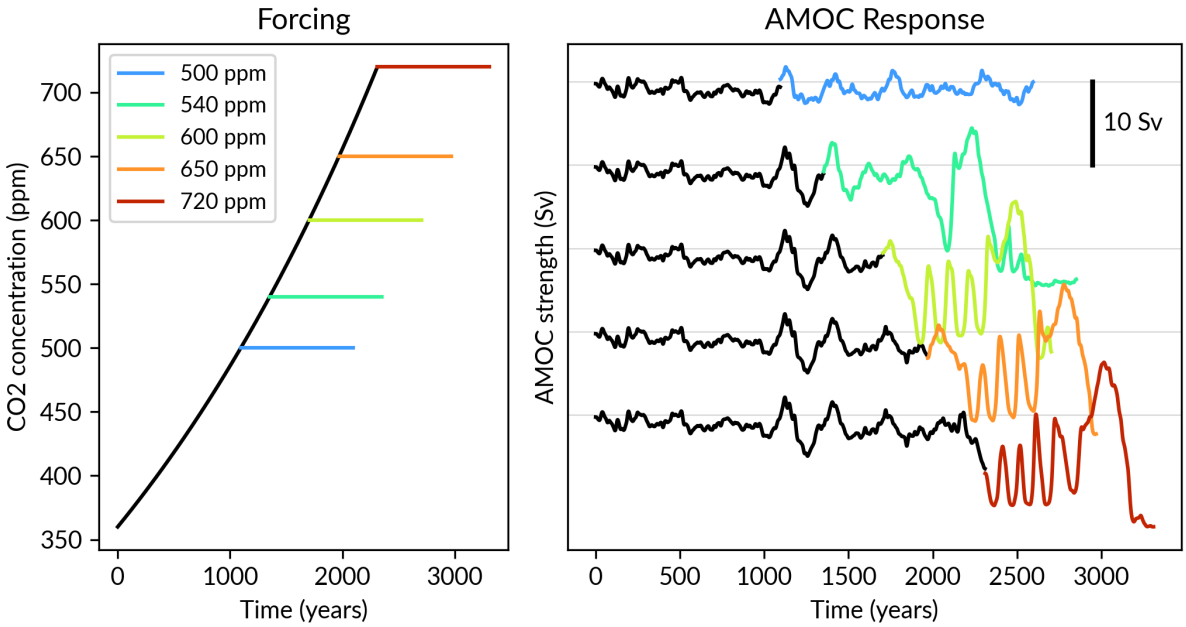

Figure A.6 | **AMOC response to CO_2 ramping and subsequent stabilization** at different CO_2 levels in PlaSim-LSG. Left: CO_2 forcing protocols with an exponential ramp-up phase (black) conducted at a rate of 0.03% increase per year and continuation under constant forcing for 1000 years (colored). Right: AMOC response under each protocol, showing the AMOC strength at 46-66N. Timeseries are vertically shifted by 10 Sv each for visibility.

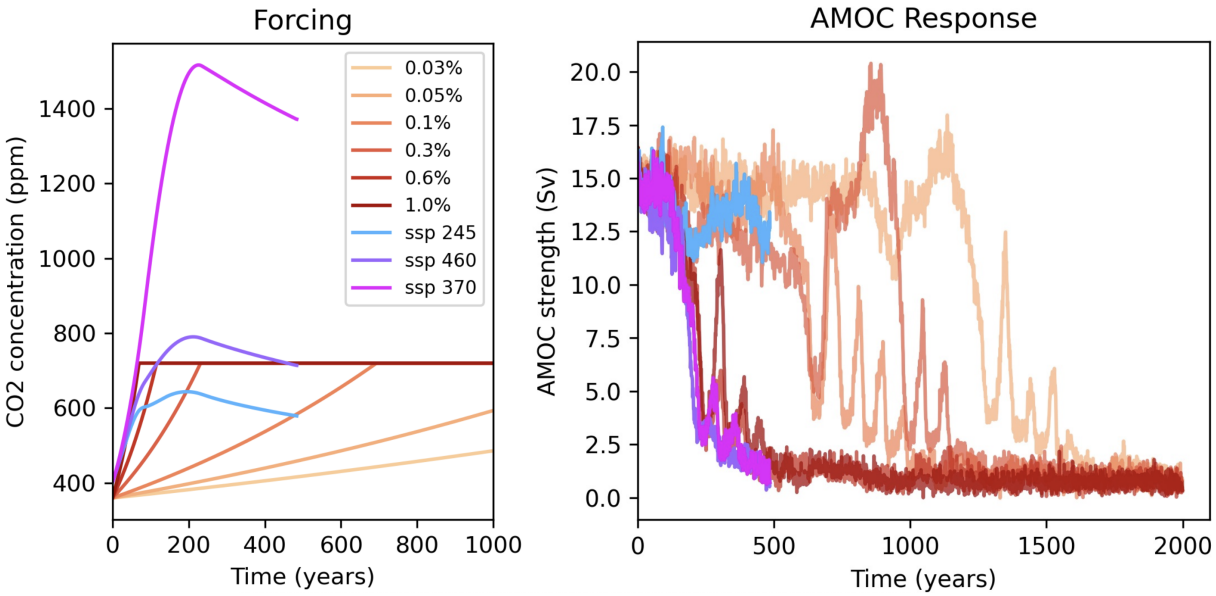

Figure A.7 | **AMOC response to CO_2 ramping at different rates** in PlaSim-LSG. Left: Zoom-in into the forcing protocols until year 1000. Idealized exponential ramping (orange/red tones) was conducted up to CO_2 doubling (from 360 to 720 ppm) for different rates; extended SSP scenarios (2-4.5, blue; 3-7.0, pink; 4-6.0, purple; see Meinshausen et al. (2020)) were run for 505 years (1995-2500 C.E.). Right: AMOC response under each protocol (as colored, 1 simulation each), showing the AMOC strength at 46-66N.

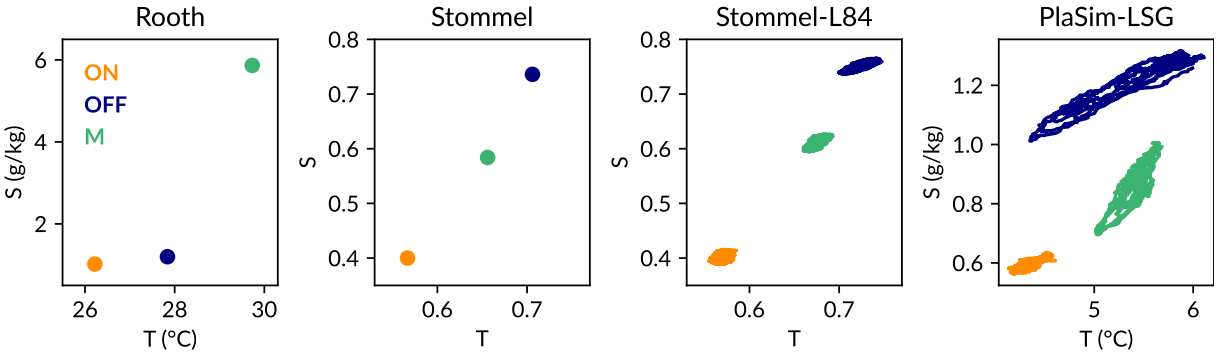

Figure A.8 | **Stommel projection** of AMOC equilibrium states in the models studied in this thesis, showing the ON (orange), OFF (blue), and Melancholia (M, green) states in the projected phase space of meridional temperature gradient T and salinity gradient S. From left to right: Rooth model with $F_1/F_3 = 1.5$ (chapter 3), uncoupled Stommel model for $\sigma_0 = 0.9$ (chapter 5), Stommel-Lorenz84 model (chapter 5, $\sigma_0 = 0.9$) and PlaSim-LSG at 360 ppm CO_2 (chapter 6).

Acknowledgements

Since the kick-off meeting in Hornbæk three years ago, being part of the CriticalEarth network has been the best PhD experience I can imagine. First of all, I would like to thank **Valerio** Lucarini for giving me the unique opportunity to work with him. Valerio has been a fantastic mentor with a lovely ability to combine scientific excellence with fun and kindness. As his student, I enjoyed extreme flexibility, academic freedom and personal support, while he was always available to discuss science or philosophize about anything from Soviet potato farming to Berlin's nightlife. I learned a lot from Valerio that has profoundly influenced my scientific approach. Valerio has entrusted me with exciting roles that often went beyond the PhD level and treated me like a peer, which I am very grateful for.

I was privileged to have additional supervisors as a result of my two secondments in Oldenburg, Germany and Turin, Italy. Since her kind dinner invitation in Edinburgh and our first meeting in a Berlin café, **Ulrike** Feudel has been extremely dedicated not only to our scientific exchange but also to including me in her academic family (the "friends of complexity"). Thanks to the great group atmosphere, my three months at the Institute for Chemistry and Biology of the Marine Environment, Carl von Ossieztky University of Oldenburg, have been a wonderful time and produced some of the key results of this thesis. A special thanks goes to **Ryan** Deeley, with whom I co-led the saddle avoidance paper through long hours of scientific exploration and a challenging review process. Sharing a house, an office, an appetite for Fulham's best double falafel wraps and a tendency for last-minute wasp-infested night shifts has led to many fun moments in our friendship.

My second exchange took me to the Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture at Politecnico di Torino, where I spend ten lovely weeks under the supervision of **Jost** von Hardenberg. The exchange provided excellent hands-on experience in climate modelling that was essential for taking my thesis to higher rungs of the model hierarchy ladder. I am very grateful to Jost for his enthusiasm and helpful input in co-supervising the project, remembering a legendary debugging session in the scoring sun that taught me how even humans can read binary files if needed. I am especially thankful to **Oliver** Mehling, who has become a great collaborator and friend. Oliver provided many helpful tips for using PlaSim-LSG, and I always enjoyed our long science talks, be it in the office, on an Alpine hike, among British sheep, or on a night train to the next conference.

Back at headquarters in Reading, it has always been a pleasure to meet with **Jeroen** Wouters, my second PhD supervisor. Our discussions, often over an all-you-can-eat lunch in the canteen, deepened my understanding of rare events. Jeroen has been especially helpful in getting me started with large deviation theory at the beginning. Unfortunately, with my jetting around, we did not get the chance to collaborate as much as I would have liked.

It has been fantastic to meet **Tobias** Grafke, who became an additional supervisor of the saddle avoidance project. Our weekly meetings, followed by a day visit at University of Warwick, contributed a strong numerical basis for the work and taught me a lot about large deviations. Besides Tobias, **Raphael** Römer joined our project, too, with dedication, important contributions and refined trumpet skills. Towards the final phase of the PhD, I have very much enjoyed the co-working sessions with Raphael in Oxford, regardless of the lunch portion sizes.

At the beginning of the PhD, I enjoyed collaborating with **Larissa** Serdukova on the Rooth model, who always shows a lovely enthusiasm and positivity towards science. **George** Datseris

lured me into JuliaDynamics and helped me learn about good coding practices and software development. I'm lucky that **Orjan** Ameye found our CriticalTransitions.jl package on GitHub, leading to a very fruitful coding collaboration and a lovely research/sailing visit in Konstanz. Founding and co-convening the session "The Climate Model Hierarchy" at the EGU General Assembly 2024 together with Oliver, Raphael, **Maya** Ben-Yami and **Richard** Wood has been a great experience. The EGU conferences have also enabled continued discussions with **Yvan** Romé that led to a joint paper on AMOC oscillation mechanisms. I'm also happy to have started interesting collaborations with **Dániel** Janosi, **Simon** Michel and **Valentin** Chanut. Lastly, I learned a lot participating in the comment article coming out of the **YESS community**'s side event on "Early-mid career perspectives on South-North inequalities" at the World Climate Research Programme Open Science Conference 2023 in Kigali, Rwanda.

I would like to sincerely thank the **European Union** for funding my PhD as part of the Horizon 2020 Marie Skłodowska-Curie Innovative Training Programme CriticalEarth, conceived and led by **Peter** Ditlevsen. This has been an exceptional program with many privileges that I feel very lucky to have enjoyed. CriticalEarth introduced me to a wonderful community, centered around a close-knit group: **Alessandro**, **Clara**, **Ignacio**, **Jade**, **Jan**, **Kolja**, Maya, **Oisín**, Oliver, **Paolo**, Raphael, Ryan, **Sacha**, and **Valérian**.

I also got much inspiration and knowledge from **Anastasia** Romanou, **Maria** Cameron, **Thomas** Stocker, **Támas** Tel, **Sebastian** Wieczorek, Peter Ditlevsen, **Johannes** Lohmann, **Peter** Ashwin, **Melinda** Gálfi, **Davide** Faranda and **Susanna** Corti. I am grateful for having met **Kalel** Rossi, **Giada** Cerato, **Matteo** Cini, **Elena** Saggioro, **Melanie** Kobras, **Chiara** Cecilia Maiocchi, **Niccolò** Zagli, **Calvin** Nesbitt, **John** Moroney, **Alexis** Gomel and **Victor** Couplet, plus many more with whom I recall great conversations. In the Reading office, it has always been fun to meet **Ghazaleh** and **Madhawi**, who was always ready to offer a cup of delicious spiced tea.

A special highlight of my PhD was the Fairbrother Lecture 2024, which would not have been possible without **Joanna** John and her team from the Doctoral and Researcher Office. A huge thank you goes to **Ruth** Harris and **Vivien** Too from the Math Department's administrative support, as well as **Jenny** Chalk and her team from the Gray Dawes Travel Group. Their kind and continuous support with the frequent travel arrangements is much appreciated. I thank the staff of the Paddington Office, a Jazz bar in Edinburgh, the carpet floor of the Minneapolis Convention Center and various rail companies for facilitating my PhD meetings with Valerio.

Finally, it often goes unnoticed how much the support from family and friends matters to keep going and stay motivated. Thank you to my family, especially my **parents**, **Kaja** and **Dorottya** for their love and for caring about what I do – next time they could probably write the thesis themselves!

List of manuscripts and outputs

Manuscripts forming parts of this thesis

- R. Börner, R. Deeley, R. Römer, T. Grafke, V. Lucarini, and U. Feudel (2024b). "Saddle avoidance of noise-induced transitions in multiscale systems". *Physical Review Research* 6.4, p. L042053. DOI: 10.1103/PhysRevResearch.6.L042053
- O. Mehling, R. Börner, and V. Lucarini (2024). "Limits to predictability of the asymptotic state of the Atlantic Meridional Overturning Circulation in a conceptual climate model". *Physica D: Nonlinear Phenomena* 459, p. 134043. DOI: 10.1016/j.physd.2023.134043
- R. Börner, O. Mehling, J. von Hardenberg, and V. Lucarini (2025a). "Boundary crisis and long transients of the Atlantic overturning circulation mediated by an edge state". *submitted to Philosophical Transactions of the Royal Society A*. DOI: 10.48550/arXiv.2504.20002

Further co-authored manuscripts related to my PhD work

- Y. M. Romé, R. F. Ivanovic, L. J. Gregoire, D. Swingedouw, S. Sherriff-Tadano, and R. Börner (2025). "Simulated millennial-scale climate variability driven by a convection–advection oscillator". *Climate Dynamics* 63.3, p. 150. DOI: 10.1007/s00382-025-07630-x
- N. Testani, L. M. Cappelletti, L. B. Díaz, C. Prudente, V. Rabanal, J. Mindlin, R. Börner, D. David T, I. Diallo, I. M. Leyba, M. Osman, and A. Tangarife-Escobar (2025). "Balancing Earth science careers in an unequal world". *Communications Earth & Environment* 6.1. DOI: 10.1038/s43247-024-01964-w

Software

- R. Börner, R. Deeley, R. Römer, and O. Ameye (2025). "CriticalTransitions.jl – A Julia package for critical transitions in dynamical systems with time-dependent forcing". URL: https://github.com/JuliaDynamics/CriticalTransitions.jl

Public science presentations

- R. Börner (2022). "Breaking the wall of climate predictability". MSCA Falling Walls Lab competition, Paris. URL: https://youtu.be/noZHYzTL4yw?t=1015
- R. Börner (2024). "Uncertain currents: Predicting tipping points in our ocean and climate". Fairbrother Lecture 2024, University of Reading. URL: https://youtu.be/lJWHSMyvEAo

Data availability

Data supporting this thesis are publicly available in Börner et al. (2024a) (for chapter 4), Mehling and Börner (2023) (for chapter 5) and Börner et al. (2025b) (for chapter 6).

References

Adib, A. B. (2008). "Stochastic Actions for Diffusive Dynamics: Reweighting, Sampling, and Minimization". *J. Phys. Chem. B* 112.19, pp. 5910–5916. DOI: 10.1021/jp0751458.

Agmon, N. and R. Kosloff (1987). "Dynamics of two-dimensional diffusional barrier crossing". *J. Phys. Chem.* 91.7, pp. 1988–1996. DOI: 10.1021/j100291a061.

Aguilar, J., J. W. Baron, T. Galla, and R. Toral (2022). "Sampling rare trajectories using stochastic bridges". *Phys. Rev. E* 105.6, p. 064138. DOI: 10.1103/PhysRevE.105.064138.

Akabane, T. K. et al. (2024). "Weaker Atlantic overturning circulation increases the vulnerability of northern Amazon forests". *Nature Geoscience* 17.12, pp. 1284–1290. DOI: 10.1038/s41561-024-01578-z.

Alkhayuon, H., P. Ashwin, L. C. Jackson, C. Quinn, and R. A. Wood (2019). "Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2225, p. 20190051. DOI: 10.1098/rspa.2019.0051.

Alley, R. B., S. Anandakrishnan, and P. Jung (2001). "Stochastic resonance in the North Atlantic". *Paleoceanography* 16.2, pp. 190–198. DOI: 10.1029/2000PA000518.

Alley, R. B. et al. (2003). "Abrupt Climate Change". *Science* 299.5615, pp. 2005–2010. DOI: 10.1126/science.1081056.

Alley, R. B. and A. M. Ágústsdóttir (2005). "The 8k event: cause and consequences of a major Holocene abrupt climate change". *Quaternary Science Reviews* 24.10, pp. 1123–1149. DOI: 10.1016/j.quascirev.2004.12.004.

Alligood, K. T., T. D. Sauer, and J. A. Yorke (1996). *Chaos: An Introduction to Dynamical Systems*. Springer.

Angeloni, M. (2022). "Climate variability in an Earth system Model of Intermediate Complexity: from interannual to centennial timescales". PhD Thesis. Università di Bologna. DOI: 10.48676/unibo/amsdottorato/10152.

Angeloni, M., E. Palazzi, and J. Von Hardenberg (2020). *Evaluation and climate sensitivity of the PlaSim v.17 Earth SystemModel coupled with ocean model components of different complexity*. Preprint. DOI: 10.5194/gmd-2020-245.

Arakawa, A. and V. R. Lamb (1977). "Computational Design of the Basic Dynamical Processes of the UCLA General Circulation Model". *Methods in Computational Physics: Advances in Research and Applications*. Vol. 17. General Circulation Models of the Atmosphere. Elsevier, pp. 173–265. DOI: 10.1016/B978-0-12-460817-7.50009-4.

Armstrong McKay, D. I. et al. (2022). "Exceeding 1.5°C global warming could trigger multiple climate tipping points". *Science* 377.6611. DOI: 10.1126/science.abn7950.

Arnold, L. (2001). "Hasselmann's program revisited: The analysis of stochasticity in deterministic climate models". *Stochastic climate models*. Springer, pp. 141–157.

Arrhenius, S. (1889). "Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren". *Zeitschrift für Physikalische Chemie* 4.1, pp. 226–248. DOI: 10.1515/zpch-1889-0416.

Ashwin, P. and J. Newman (2021). "Physical invariant measures and tipping probabilities for chaotic attractors of asymptotically autonomous systems". *Eur. Phys. J. Spec. Top.* 230.16, pp. 3235–3248. DOI: 10.1140/epjs/s11734-021-00114-z.

Ashwin, P., J. Newman, and R. Römer (2025). "Contrasting Chaotic and Stochastic Forcing: Tipping Windows and Attractor Crises". *SIAM J. Appl. Dyn. Syst.*, pp. 277–316. DOI: 10.1137/24M1661534.

Ashwin, P., S. Wieczorek, R. Vitolo, and P. Cox (2012). "Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370.1962, pp. 1166–1184. DOI: 10.1098/rsta.2011.0306.

Axelsen, A. R., C. R. Quinn, and A. P. Bassom (2024). "Finite-Time Analysis of Crises in a Chaotically Forced Ocean Model". *J Nonlinear Sci* 34.5, p. 97. DOI: 10.1007/s00332-024-10077-9.

Baker, J. A., M. J. Bell, L. C. Jackson, G. K. Vallis, A. J. Watson, and R. A. Wood (2025). "Continued Atlantic overturning circulation even under climate extremes". *Nature* 638.8052, pp. 987–994. DOI: 10.1038/s41586-024-08544-0.

Bashkirtseva, I. and L. Ryashko (2011). "Sensitivity analysis of stochastic attractors and noise-induced transitions for population model with Allee effect". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21.4, p. 047514. DOI: 10.1063/1.3647316.

Battelino, P. M., C. Grebogi, E. Ott, J. A. Yorke, and E. D. Yorke (1988). "Multiple coexisting attractors, basin boundaries and basic sets". *Physica D: Nonlinear Phenomena* 32.2, pp. 296–305. DOI: 10.1016/0167-2789(88)90057-7.

Baumgartner, A. and E. Reichel (1975). *The world water balance: mean annual global, continental and maritime precipitation, evaporation and run-off*. Elsevier Scientific Pub. Co.

Bazykin, A. D. (1998). "Dynamics of Isolated Populations". *World Scientific Series on Nonlinear Science Series A*. Vol. 11. World Scientific, pp. 7–17. DOI: 10.1142/9789812798725_0002.

Bellman, R. E. (2010). *Dynamic Programming*. Revised edition. Princeton Landmarks in Mathematics and Physics 33. Princeton University Press.

Bellomo, K. and O. Mehling (2024). "Impacts and State-Dependence of AMOC Weakening in a Warming Climate". *Geophysical Research Letters* 51.10, e2023GL107624. DOI: 10.1029/2023GL107624.

Bellomo, K. et al. (2023). "Impacts of a weakened AMOC on precipitation over the Euro-Atlantic region in the EC-Earth3 climate model". *Climate Dynamics* 61.7, pp. 3397–3416. DOI: 10.1007/s00382-023-06754-2.

Ben-Yami, M., P. Good, et al. (2024). "Impacts of AMOC Collapse on Monsoon Rainfall: A Multi-Model Comparison". *Earth's Future* 12.9, e2023EF003959. DOI: 10.1029/2023EF003959.

Ben-Yami, M., A. Morr, S. Bathiany, and N. Boers (2024). "Uncertainties too large to predict tipping times of major Earth system components from historical data". *Science Advances* 10.31. DOI: 10.1126/sciadv.adl4841.

Benettin, G., L. Galgani, A. Giorgilli, and J.-M. Strelcyn (1980). "Lyapunov Characteristic Exponents for smooth dynamical systems and for hamiltonian systems; A method for computing all of them. Part 2: Numerical application". *Meccanica* 15.1, pp. 21–30. DOI: 10.1007/BF02128237.

Berezhkovskii, A., L. Berezhkovskii, and V. Zitzerman (1989). "The rate constant in the Kramers multidimensional theory and the saddle-point avoidance". *Chemical Physics* 130.1-3, pp. 55–63. DOI: 10.1016/0301-0104(89)87036-3.

Berger, A. (1988). "Milankovitch Theory and climate". *Reviews of Geophysics* 26.4, pp. 624–657. DOI: 10.1029/RG026i004p00624.

Berglund, N. (2013). *Kramers' law: Validity, derivations and generalisations*. DOI: 10.48550/arXiv.1106.5799.

Bernuzzi, P. and C. Kuehn (2023). "Bifurcations and Early-Warning Signs for SPDEs with Spatial Heterogeneity". *Journal of Dynamics and Differential Equations* 37.1, pp. 191–235. DOI: 10.1007/s10884-023-10274-2.

Bernuzzi, P. and T. Grafke (2024). *Large Deviation Minimisers for Stochastic Partial Differential Equations with Degenerate Noise*. DOI: 10.48550/arXiv.2409.17839.

Bintanja, R. and F. M. Selten (2014). "Future increases in Arctic precipitation linked to local evaporation and sea-ice retreat". *Nature* 509.7501, pp. 479–482. DOI: 10.1038/nature13259.

Birkhoff, G. D. (1931). "Proof of the Ergodic Theorem". *Proceedings of the National Academy of Sciences* 17.12, pp. 656–660. DOI: 10.1073/pnas.17.2.656.

Bjerknes, J. (1964). "Atlantic Air-Sea Interaction". *Advances in Geophysics*. Vol. 10. Elsevier, pp. 1–82. DOI: 10.1016/S0065-2687(08)60005-9.

Bobrovsky, B. Z. and Z. Schuss (1982). "A Singular Perturbation Method for the Computation of the Mean First Passage Time in a Nonlinear Filter". *SIAM J. Appl. Math.* 42.1, pp. 174–187. DOI: 10.1137/0142014.

Bobrovsky, B.-Z. and O. Zeitouni (1992). "Some results on the problem of exit from a domain". *Stochastic Processes and their Applications* 41.2, pp. 241–256. DOI: 10.1016/0304-4149(92)90124-9.

Bódai, T. and V. Lucarini (2020). "Rough basin boundaries in high dimension: Can we classify them experimentally?" *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.10, p. 103105. DOI: 10.1063/5.0002577.

Bódai, T., V. Lucarini, F. Lunkeit, and R. Boschi (2015). "Global instability in the Ghil–Sellers model". *Climate Dynamics* 44.11, pp. 3361–3381. DOI: 10.1007/s00382-014-2206-5.

Boers, N. (2021). "Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation". *Nature Climate Change* 11.8, pp. 680–688. DOI: 10.1038/s41558-021-01097-4.

Boers, N., M. Ghil, and T. F. Stocker (2022). "Theoretical and paleoclimatic evidence for abrupt transitions in the Earth system". *Environ. Res. Lett.* 17.9, p. 093006. DOI: 10.1088/1748-9326/ac8944.

Bonan, D. B., A. F. Thompson, T. Schneider, L. Zanna, K. C. Armour, and S. Sun (2025). "Observational constraints imply limited future Atlantic meridional overturning circulation weakening". *Nature Geoscience*, pp. 1–9. DOI: 10.1038/s41561-025-01709-0.

Bond, G. et al. (1993). "Correlations between climate records from North Atlantic sediments and Greenland ice". *Nature* 365.6442, pp. 143–147. DOI: 10.1038/365143a0.

Boot, A. A., J. Steenbeek, M. Coll, A. S. Von Der Heydt, and H. A. Dijkstra (2025). "Global Marine Ecosystem Response to a Strong AMOC Weakening Under Low and High Future Emission Scenarios". *Earth's Future* 13.1, e2024EF004741. DOI: 10.1029/2024EF004741.

Boot, A. A., A. S. Von Der Heydt, and H. A. Dijkstra (2024). "Response of atmospheric pCO_2 to a strong AMOC weakening under low and high emission scenarios". *Climate Dynamics* 62.8, pp. 7559–7574. DOI: 10.1007/s00382-024-07295-y.

Börner, R., R. Deeley, R. Römer, T. Grafke, V. Lucarini, and U. Feudel (2024a). *Saddle avoidance of noise-induced transitions in multiscale systems*. Figshare (dataset). DOI: 10.6084/m9.figshare.27844776.

— (2024b). "Saddle avoidance of noise-induced transitions in multiscale systems". *Physical Review Research* 6.4, p. L042053. DOI: 10.1103/PhysRevResearch.6.L042053.

Börner, R., V. Lucarini, and L. Serdukova (2022). "Dynamical Landscape and Noise-induced Transitions in a Box Model of the Atlantic Meridional Overturning Circulation". *EGU General Assembly Conference Abstracts*, EGU22–8332.

Börner, R., O. Mehling, J. von Hardenberg, and V. Lucarini (2025a). "Boundary crisis and long transients of the Atlantic overturning circulation mediated by an edge state". *submitted to Philosophical Transactions of the Royal Society A*. DOI: 10.48550/arXiv.2504.20002.

— (2025b). *Data repository for Börner et al. "Boundary crisis and long transients of the Altantic overturning circulation mediated by an edge state"*. Zenodo. DOI: 10.5281/zenodo.17053348.

Bouchaud, J.-P. and R. Cont (1998). "A Langevin approach to stock market fluctuations and crashes". *Eur. Phys. J. B* 6.4, pp. 543–550. DOI: 10.1007/s100510050582.

Bouchet, F. and J. Reygner (2016). "Generalisation of the Eyring–Kramers Transition Rate Formula to Irreversible Diffusion Processes". *Ann. Henri Poincaré* 17.12, pp. 3499–3532. DOI: 10.1007/s00023-016-0507-4.

— (2022). "Path integral derivation and numerical computation of large deviation prefactors for non-equilibrium dynamics through matrix Riccati equations". *Journal of Statistical Physics* 189.21. DOI: 10.1007/s10955-022-02983-7.

Bouchet, F., J. Rolland, and E. Simonnet (2019). "Rare event algorithm links transitions in turbulent flows with activated nucleations". *Physical Review Letters* 122.7, p. 074502. DOI: 10.1103/PhysRevLett.122.074502.

Bouchet, F., J. Rolland, and J. Wouters (2019). "Rare Event Sampling Methods". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.8, p. 080402. DOI: 10.1063/1.5120509.

Bressloff, P. C. (2017). "Stochastic switching in biology: from genotype to phenotype". *Journal of Physics A: Mathematical and Theoretical* 50.13, p. 133001. DOI: 10.1088/1751-8121/aa5db4.

Brinkman, B. A. W. et al. (2022). "Metastable dynamics of neural circuits and networks". *Applied Physics Reviews* 9.1, p. 011313. DOI: 10.1063/5.0062603.

Bröcker, J. (2019). "What is the correct cost functional for variational data assimilation?" *Climate Dynamics* 52.1, pp. 389–399. DOI: 10.1007/s00382-018-4146-y.

Broecker, W. S., G. Bond, M. Klas, G. Bonani, and W. Wolfli (1990). "A salt oscillator in the glacial Atlantic? 1. The concept". *Paleoceanography* 5.4, pp. 469–477. DOI: 10.1029/PA005i004p00469.

Broecker, W. S., D. M. Peteet, and D. Rind (1985). "Does the ocean–atmosphere system have more than one stable mode of operation?" *Nature* 315.6014, pp. 21–26. DOI: 10.1038/315021a0.

Broer, H., C. Simó, and R. Vitolo (2002). "Bifurcations and strange attractors in the Lorenz-84 climate model with seasonal forcing". *Nonlinearity* 15.4, p. 1205. DOI: 10.1088/0951-7715/15/4/312.

Buckley, M. W. and J. Marshall (2016). "Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: A review". *Reviews of Geophysics* 54.1, pp. 5–63. DOI: 10.1002/2015RG000493.

Budyko, M. I. (1969). "The effect of solar radiation variations on the climate of the Earth". *Tellus* 21.5, pp. 611–619. DOI: 10.3402/tellusa.v21i5.10109.

Buizert, C. and A. Schmittner (2015). "Southern Ocean control of glacial AMOC stability and Dansgaard-Oeschger interstadial duration". *Paleoceanography* 30.12, pp. 1595–1612. DOI: 10.1002/2015PA002795.

Caesar, L., S. Rahmstorf, A. Robinson, G. Feulner, and V. Saba (2018). "Observed fingerprint of a weakening Atlantic Ocean overturning circulation". *Nature* 556.7700, pp. 191–196. DOI: 10.1038/s41586-018-0006-5.

Cameron, M. K. (2012). "Finding the quasipotential for nongradient SDEs". *Physica D: Nonlinear Phenomena* 241.18, pp. 1532–1550. DOI: 10.1016/j.physd.2012.06.005.

Caraballo, T. and X. Han (2016). *Applied Nonautonomous and Random Dynamical Systems*. SpringerBriefs in Mathematics. Springer. DOI: 10.1007/978-3-319-49247-6.

Carpenter, S. R. and W. A. Brock (2006). "Rising variance: a leading indicator of ecological transition". *Ecology Letters* 9.3, pp. 311–318. DOI: 10.1111/j.1461-0248.2005.00877.x.

Castellana, D., S. Baars, F. W. Wubs, and H. A. Dijkstra (2019). "Transition Probabilities of Noise-induced Transitions of the Atlantic Ocean Circulation". *Scientific Reports* 9.1, p. 20284. DOI: 10.1038/s41598-019-56435-6.

Cessi, P. (1994). "A Simple Box Model of Stochastically Forced Thermohaline Flow". *Journal of Physical Oceanography* 24.9, pp. 1911–1920. DOI: 10.1175/1520-0485(1994)024<1911:ASBMOS>2.0.CO;2.

Chambers, J. M., C. L. Mallows, and B. W. Stuck (1976). "A Method for Simulating Stable Random Variables". *Journal of the American Statistical Association* 71.354, pp. 340–344. DOI: 10.1080/01621459.1976.10480344.

Chapman, R., P. Ashwin, J. Baker, and R. Wood (2024). *Quantifying risk of a noise-induced AMOC collapse from northern and tropical Atlantic Ocean variability*. DOI: 10.1088/2515-7620/ad8f98.

Cheng, W., J. C. H. Chiang, and D. Zhang (2013). "Atlantic Meridional Overturning Circulation (AMOC) in CMIP5 Models: RCP and Historical Simulations". *Journal of Climate* 26.18, pp. 7187–7197. DOI: 10.1175/JCLI-D-12-00496.1.

Cimatoribus, A. A., S. S. Drijfhout, V. Livina, and G. van der Schrier (2013). "Dansgaard–Oeschger events: bifurcation points in the climate system". *Climate of the Past* 9.1, pp. 323–333. DOI: 10.5194/cp-9-323-2013.

Cimatoribus, A. A., S. S. Drijfhout, M. Den Toom, and H. A. Dijkstra (2012). "Sensitivity of the Atlantic meridional overturning circulation to South Atlantic freshwater anomalies". *Climate Dynamics* 39.9-10, pp. 2291–2306. DOI: 10.1007/s00382-012-1292-5.

Cimatoribus, A. A., S. S. Drijfhout, and H. A. Dijkstra (2014). "Meridional overturning circulation: stability and ocean feedbacks in a box model". *Climate Dynamics* 42.1, pp. 311–328. DOI: 10.1007/s00382-012-1576-9.

Cini, M., G. Zappa, F. Ragone, and S. Corti (2024). "Simulating AMOC tipping driven by internal climate variability with a rare event algorithm". *npj Clim Atmos Sci* 7.1, pp. 1–10. DOI: 10.1038/s41612-024-00568-7.

— (2025). "Noise-shaped hysteresis cycles of the AMOC under increasing CO2 forcing". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 35.2, p. 023167. DOI: 10.1063/5.0241503.

Clark, P. U., N. G. Pisias, T. F. Stocker, and A. J. Weaver (2002). "The role of the thermohaline circulation in abrupt climate change". *Nature* 415.6874, pp. 863–869. DOI: 10.1038/415863a.

Claussen, M. et al. (2002). "Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models". *Climate Dynamics* 18.7, pp. 579–586. DOI: 10.1007/s00382-001-0200-1.

Clement, A. C. and L. C. Peterson (2008). "Mechanisms of abrupt climate change of the last glacial period". *Reviews of Geophysics* 46.4. DOI: 10.1029/2006RG000204.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer. DOI: 10.1007/978-1-4471-3675-0.

Crawford, J. D. (1991). "Introduction to bifurcation theory". *Rev. Mod. Phys.* 63.4, pp. 991–1037. DOI: 10.1103/RevModPhys.63.991.

Cugliandolo, L. F. and V. Lecomte (2017). "Rules of calculus in the path integral representation of white noise Langevin equations: the Onsager–Machlup approach". *J. Phys. A: Math. Theor.* 50.34, p. 345001. DOI: 10.1088/1751-8121/aa7dd6.

Cvitanović, P. (1991). "Periodic orbits as the skeleton of classical and quantum chaos". *Physica D: Nonlinear Phenomena* 51.1, pp. 138–151. DOI: 10.1016/0167-2789(91)90227-Z.

D'Errico, M. et al. (2022). "Present and future synoptic circulation patterns associated with cold and snowy spells over Italy". *Earth System Dynamics* 13.2, pp. 961–992. DOI: 10.5194/esd-13-961-2022.

Dahiya, D. and M. Cameron (2018a). "An Ordered Line Integral Method for computing the quasi-potential in the case of variable anisotropic diffusion". *Physica D: Nonlinear Phenomena* 382-383, pp. 33–45. DOI: 10.1016/j.physd.2018.07.002.

— (2018b). "Ordered Line Integral Methods for Computing the Quasi-Potential". *J Sci Comput* 75.3, pp. 1351–1384. DOI: 10.1007/s10915-017-0590-9.

Dakos, V., M. Scheffer, E. H. van Nes, V. Brovkin, V. Petoukhov, and H. Held (2008). "Slowing down as an early warning signal for abrupt climate change". *Proceedings of the National Academy of Sciences* 105.38, pp. 14308–14312. DOI: 10.1073/pnas.0802430105.

Dannenberg, P. H., J. C. Neu, and S. W. Teitsworth (2014). "Steering Most Probable Escape Paths by Varying Relative Noise Intensities". *Physical Review Letters* 113.2, p. 020601. DOI: 10.1103/PhysRevLett.113.020601.

Dansgaard, W., J. W. C. White, and S. J. Johnsen (1989). "The abrupt termination of the Younger Dryas climate event". *Nature* 339.6225, pp. 532–534. DOI: 10.1038/339532a0.

Dansgaard, W. et al. (1993). "Evidence for general instability of past climate from a 250-kyr ice-core record". *Nature* 364.6434, pp. 218–220. DOI: 10.1038/364218a0.

Datseris, G. (2018). "DynamicalSystems.jl: A Julia software library for chaos and nonlinear dynamics". *JOSS* 3.23, p. 598. DOI: 10.21105/joss.00598.

Datseris, G., K. Luiz Rossi, and A. Wagemakers (2023). "Framework for global stability analysis of dynamical systems". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 33.7, p. 073151. DOI: 10.1063/5.0159675.

Datseris, G. and U. Parlitz (2022). *Nonlinear Dynamics: A Concise Introduction Interlaced with Code*. Springer Nature.

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

Day, M. V. (1994). "Cycling and skewing of exit measures for planar systems". *Stochastics and Stochastic Reports* 48.3-4, pp. 227–247. DOI: 10.1080/17442509408833907.

De Melo Viríssimo, F., D. A. Stainforth, and J. Bröcker (2024). "The evolution of a non-autonomous chaotic system under non-periodic forcing: A climate change example". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 34.1, p. 013136. DOI: 10.1063/5.0180870.

Deco, G. and V. K. Jirsa (2012). "Ongoing Cortical Activity at Rest: Criticality, Multistability, and Ghost Attractors". *J. Neurosci.* 32.10, pp. 3366–3375. DOI: 10.1523/JNEUROSCI.2523-11.2012.

Del Amo, I. and P. Ditlevsen (2025). "Escape by jumps and diffusion by α-stable noise across the barrier in a double well potential". *Physica D: Nonlinear Phenomena* 479, p. 134709. DOI: 10.1016/j.physd.2025.134709.

Dhooge, A., W. Govaerts, and Y. A. Kuznetsov (2003). "MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs". *ACM Trans. Math. Softw.* 29.2, pp. 141–164. DOI: 10.1145/779359.779362.

Dijkstra, H. A. (2013). *Nonlinear Climate Dynamics*. Cambridge University Press.

Dijkstra, H. A. and W. Weijer (2005). "Stability of the Global Ocean Circulation: Basic Bifurcation Diagrams". *Journal of Physical Oceanography* 35.6, pp. 933–948. DOI: 10.1175/JPO2726.1.

Ditlevsen, P. and S. Ditlevsen (2023). "Warning of a forthcoming collapse of the Atlantic meridional overturning circulation". *Nature Communications* 14.1, p. 4254. DOI: 10.1038/s41467-023-39810-w.

Ditlevsen, P. D. (1999). "Observation of α-stable noise induced millennial climate changes from an ice-core record". *Geophysical Research Letters* 26.10, pp. 1441–1444. DOI: 10.1029/1999GL900252.

Ditlevsen, P. D. and S. J. Johnsen (2010). "Tipping points: early warning and wishful thinking". *Geophysical Research Letters* 37.19. DOI: 10.1029/2010GL044486.

Drótos, G., T. Bódai, and T. Tél (2015). "Probabilistic Concepts in a Changing Climate: A Snapshot Attractor Picture". *Journal of Climate* 28.8, pp. 3275–3288. DOI: 10.1175/JCLI-D-14-00459.1.

Du, Q., T. Li, X. Li, and W. Ren (2021). "The graph limit of the minimizer of the Onsager-Machlup functional and its computation". *Science China Mathematics* 64.2, pp. 239–280. DOI: 10.1007/s11425-019-1650-7.

Dürr, D. and A. Bach (1978). "The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process". *Commun. Math. Phys.* 60.2, pp. 153–170. DOI: 10.1007/BF01609446.

E, W., W. Ren, and E. Vanden-Eijnden (2004). "Minimum action method for the study of rare events". *Communications on Pure and Applied Mathematics* 57.5, pp. 637–656. DOI: 10.1002/cpa.20005.

Engel, M., M. A. Gkogkas, and C. Kuehn (2021). "Homogenization of Coupled Fast-Slow Systems via Intermediate Stochastic Regularization". *J Stat Phys* 183.2, p. 25. DOI: 10.1007/s10955-021-02765-7.

Ensing, B., M. De Vivo, Z. Liu, P. Moore, and M. L. Klein (2006). "Metadynamics as a Tool for Exploring Free Energy Landscapes of Chemical Reactions". *Acc. Chem. Res.* 39.2, pp. 73–81. DOI: 10.1021/ar040198i.

Eyring, H. (1935). "The Activated Complex in Chemical Reactions". *The Journal of Chemical Physics* 3.2, pp. 107–115. DOI: 10.1063/1.1749604.

Feng, Q. Y., J. P. Viebahn, and H. A. Dijkstra (2014). "Deep ocean early warning signals of an Atlantic MOC collapse". *Geophysical Research Letters* 41.16, pp. 6009–6015. DOI: 10.1002/2014GL061019.

Ferrell, J. E. (2012). "Bistability, Bifurcations, and Waddington's Epigenetic Landscape". *Current Biology* 22.11, R458–R466. DOI: 10.1016/j.cub.2012.03.045.

Feudel, U. (2008). "Complex dynamics in multistable systems". *Int. J. Bifurcation Chaos* 18.06, pp. 1607–1626. DOI: 10.1142/S0218127408021233.

— (2023). "Rate-induced tipping in ecosystems and climate: the role of unstable states, basin boundaries and transient dynamics". *Nonlinear Processes in Geophysics* 30.4, pp. 481–502. DOI: 10.5194/npg-30-481-2023.

Feudel, U., A. N. Pisarchik, and K. Showalter (2018). "Multistability and tipping: From mathematics and physics to climate and brain—Minireview and preface to the focus issue". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.3, p. 033501. DOI: 10.1063/1.5027718.

FitzHugh, R. (1961). "Impulses and Physiological States in Theoretical Models of Nerve Membrane". *Biophysical Journal* 1.6, pp. 445–466. DOI: 10.1016/S0006-3495(61)86902-6.

Fleurantin, E., K. Slyman, B. Barker, and C. K. R. T. Jones (2023). "A dynamical systems approach for most probable escape paths over periodic boundaries". *Physica D: Nonlinear Phenomena* 454, p. 133860. DOI: 10.1016/j.physd.2023.133860.

Fraedrich, K., H. Jansen, E. Kirk, U. Luksch, and F. Lunkeit (2005). "The Planet Simulator: Towards a user friendly model". *Meteorologische Zeitschrift* 14.3, pp. 299–304. DOI: 10.1127/0941-2948/2005/0043.

Frajka-Williams, E. et al. (2019). "Atlantic meridional overturning circulation: Observed transport and variability". *Frontiers in Marine Science* 6, p. 260. DOI: 10.3389/fmars.2019.00260.

Freidlin, M. I. and A. D. Wentzell (1998). *Random perturbations of dynamical systems*. Springer.

Galfi, V. M. and V. Lucarini (2021). "Fingerprinting Heatwaves and Cold Spells and Assessing Their Response to Climate Change Using Large Deviation Theory". *Physical Review Letters* 127.5, p. 058701. DOI: 10.1103/PhysRevLett.127.058701.

Gálfi, V. M., V. Lucarini, F. Ragone, and J. Wouters (2021). "Applications of large deviation theory in geophysical fluid dynamics and climate science". *Riv. Nuovo Cim.* 44.6, pp. 291–363. DOI: 10.1007/s40766-021-00020-z.

Ganopolski, A. and S. Rahmstorf (2002). "Abrupt Glacial Climate Changes due to Stochastic Resonance". *Physical Review Letters* 88.3, p. 038501. DOI: 10.1103/PhysRevLett.88.038501.

Gao, T., J. Duan, X. Li, and R. Song (2014). "Mean Exit Time and Escape Probability for Dynamical Systems Driven by Lévy Noises". *SIAM J. Sci. Comput.* 36.3, A887–A906. DOI: 10.1137/120897262.

Gardiner, C. (2009). *Stochastic methods – A Handbook for the Natural and Social Sciences*. Vol. 4. Springer.

Geer, A. J. (2021). "Learning earth system models from observations: machine learning or data assimilation?" *Phil. Trans. R. Soc. A.* 379.2194, p. 20200089. DOI: 10.1098/rsta.2020.0089.

Gent, P. R. (2018). "A commentary on the Atlantic meridional overturning circulation stability in climate models". *Ocean Modelling* 122, pp. 57–66. DOI: 10.1016/j.ocemod.2017.12.006.

Ghil, M. and V. Lucarini (2020). "The physics of climate variability and climate change". *Rev. Mod. Phys.* 92.3, p. 035002. DOI: 10.1103/RevModPhys.92.035002.

Ghil, M. and P. Malanotte-Rizzoli (1991). "Data Assimilation in Meteorology and Oceanography". *Advances in Geophysics*. Vol. 33. Elsevier, pp. 141–266. DOI: 10.1016/S0065-2687(08)60442-2.

Ghil, M. and D. Sciamarella (2023). "Review article: Dynamical systems, algebraic topology and the climate sciences". *Nonlinear Processes in Geophysics* 30.4, pp. 399–434. DOI: 10.5194/npg-30-399-2023.

Gladrow, J., U. F. Keyser, R. Adhikari, and J. Kappler (2021). "Experimental Measurement of Relative Path Probabilities and Stochastic Actions". *Phys. Rev. X* 11.3, p. 031022. DOI: 10.1103/PhysRevX.11.031022.

Gladwell, M. (2000). *The tipping point: how little things can make a big difference*. Little Brown.

Gottwald, G. A. (2021). "A model for Dansgaard–Oeschger events and millennial-scale abrupt climate change without external forcing". *Climate Dynamics* 56.1, pp. 227–243. DOI: 10.1007/s00382-020-05476-z.

Gottwald, G. A., D. T. Crommelin, and C. L. E. Franzke (2017). "Stochastic Climate Theory". *Nonlinear and Stochastic Climate Dynamics*. Cambridge University Press, pp. 209–240. DOI: 10.1017/9781316339251.009.

Grafke, T., R. Grauer, T. Schäfer, and E. Vanden-Eijnden (2014). "Arclength Parametrized Hamilton's Equations for the Calculation of Instantons". *Multiscale Model. Simul.* 12.2, pp. 566–580. DOI: 10.1137/130939158.

Grafke, T. and A. Laio (2024). "Metadynamics for transition paths in irreversible dynamics". *Multiscale Modeling & Simulation* 22.1, pp. 125–141. DOI: 10.1137/23M1563025.

Grafke, T. and E. Vanden-Eijnden (2019). "Numerical computation of rare events via large deviation theory". *Chaos* 29.6, p. 063118. DOI: 10.1063/1.5084025.

Graham, R. and T. Tél (1984). "Existence of a Potential for Dissipative Dynamical Systems". *Physical Review Letters* 52.1, pp. 9–12. DOI: 10.1103/PhysRevLett.52.9.

— (1986). "Nonequilibrium potential for coexisting attractors". *Physical Review A* 33.2, p. 1322. DOI: 10.1103/PhysRevA.33.1322.

Grebogi, C., E. Ott, and J. A. Yorke (1983a). "Crises, sudden changes in chaotic attractors, and transient chaos". *Physica D: Nonlinear Phenomena* 7.1, pp. 181–200. DOI: 10.1016/0167-2789(83)90126-4.

— (1983b). "Fractal Basin Boundaries, Long-Lived Chaotic Transients, and Unstable-Unstable Pair Bifurcation". *Physical Review Letters* 50.13, pp. 935–938. DOI: 10.1103/PhysRevLett.50.935.

— (1986). "Critical Exponent of Chaotic Transients in Nonlinear Dynamical Systems". *Physical Review Letters* 57.11, pp. 1284–1287. DOI: 10.1103/PhysRevLett.57.1284.

— (1988). “Unstable periodic orbits and the dimensions of multifractal chaotic attractors”. *Phys. Rev. A* 37.5, pp. 1711–1724. DOI: 10.1103/PhysRevA.37.1711.

Gu, Q. et al. (2024). “Wide range of possible trajectories of North Atlantic climate in a warming world”. *Nature Communications* 15.1, p. 4221. DOI: 10.1038/s41467-024-48401-2.

Häfner, D. et al. (2018). “Veros v0.1 – a fast and versatile ocean simulator in pure Python”. *Geoscientific Model Development* 11.8, pp. 3299–3312. DOI: 10.5194/gmd-11-3299-2018.

Hairer, M., A. M. Stuart, and J. Voss (2007). “Analysis of SPDEs arising in path sampling part II: The nonlinear case”. *The Annals of Applied Probability* 17.5-6, pp. 1657–1706. DOI: 10.1214/07-AAP441.

Halekotte, L. and U. Feudel (2020). “Minimal fatal shocks in multistable complex networks”. *Scientific Reports* 10.1, p. 11783. DOI: 10.1038/s41598-020-68805-6.

Hänggi, P., P. Talkner, and M. Borkovec (1990). “Reaction-rate theory: fifty years after Kramers”. *Rev. Mod. Phys.* 62.2, pp. 251–341. DOI: 10.1103/RevModPhys.62.251.

Harland, W. B. (1964). “Critical evidence for a great infra-Cambrian glaciation”. *Geologische Rundschau* 54.1, pp. 45–61. DOI: 10.1007/BF01821169.

Harte, J. (1988). *Consider A Spherical Cow: A Course in Environmental Problem Solving*. University Science Books.

Hasselmann, K. (1976). “Stochastic climate models Part I. Theory”. *Tellus* 28.6, pp. 473–485. DOI: 10.3402/tellusa.v28i6.11316.

— (1982). “An ocean model for climate variability studies”. *Progress in Oceanography* 11.2, pp. 69–92. DOI: 10.1016/0079-6611(82)90004-0.

Hawkins, E. et al. (2011). “Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport”. *Geophysical Research Letters* 38.10. DOI: 10.1029/2011GL047208.

Held, I. M. (2005). “The Gap between Simulation and Understanding in Climate Modeling”. *Bulletin of the American Meteorological Society* 86.11, pp. 1609–1614. DOI: 10.1175/BAMS-86-11-1609.

Heymann, M. and E. Vanden-Eijnden (2008a). “Pathways of Maximum Likelihood for Rare Events in Nonequilibrium Systems: Application to Nucleation in the Presence of Shear”. *Physical Review Letters* 100.14, p. 140601. DOI: 10.1103/PhysRevLett.100.140601.

— (2008b). “The geometric minimum action method: A least action principle on the space of curves”. *Communications on Pure and Applied Mathematics* 61.8, pp. 1052–1117. DOI: 10.1002/cpa.20238.

Hoffman, P. F., A. J. Kaufman, G. P. Halverson, and D. P. Schrag (1998). “A Neoproterozoic Snowball Earth”. *Science* 281.5381, pp. 1342–1346. DOI: 10.1126/science.281.5381.1342.

Hoffman, P. F. and D. P. Schrag (2002). “The snowball Earth hypothesis: testing the limits of global change”. *Terra Nova* 14.3, pp. 129–155. DOI: 10.1046/j.1365-3121.2002.00408.x.

Holling, C. S. (1973). “Resilience and Stability of Ecological Systems”. *Annu. Rev. Ecol. Syst.* 4.1, pp. 1–23. DOI: 10.1146/annurev.es.04.110173.000245.

Horsthemke, W. and A. Bach (1975). “Onsager-Machlup Function for one dimensional nonlinear diffusion processes”. *Z Physik B* 22.2, pp. 189–192. DOI: 10.1007/BF01322364.

Horsthemke, W. and R. Lefever (1984). “Noise-Induced Transitions in Physics, Chemistry, and Biology”. *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology*. Springer Series in Synergetics. Springer, pp. 164–200. DOI: 10.1007/3-540-36852-3_7.

Hsu, G.-H., E. Ott, and C. Grebogi (1988). “Strange saddles and the dimensions of their invariant manifolds”. *Physics Letters A* 127.4, pp. 199–204. DOI: 10.1016/0375-9601(88)90102-8.

Hu, J. and J. Duan (2020). *Onsager-Machlup action functional for stochastic partial differential equations with Lévy noise*. DOI: 10.48550/arXiv.2011.09690.

Huang, Y., S. Bathiany, P. Ashwin, and N. Boers (2024). "Deep learning for predicting rate-induced tipping". *Nat Mach Intell* 6.12, pp. 1556–1565. DOI: 10.1038/s42256-024-00937-0.

Hughes, T. M. C. and A. J. Weaver (1994). "Multiple Equilibria of an Asymmetric Two-Basin Ocean Model". *Journal of Physical Oceanography* 24.3, pp. 619–637. DOI: 10.1175/1520-0485(1994)024<0619:MEOAAT>2.0.CO;2.

Hummel, C., N. Boers, and M. Rypdal (2024). "Inconclusive Early warning signals for Dansgaard-Oeschger events across Greenland ice cores". *EGUsphere*, pp. 1–44. DOI: 10.5194/egusphere-2024-3567.

Hunt, B. R., E. Ott, and J. A. Yorke (1996). "Fractal dimensions of chaotic saddles of dynamical systems". *Phys. Rev. E* 54.5, pp. 4819–4823. DOI: 10.1103/PhysRevE.54.4819.

Hurst, H. E. (1951). "Long-Term Storage Capacity of Reservoirs". *Transactions of the American Society of Civil Engineers* 116.1, pp. 770–799. DOI: 10.1061/TACEAT.0006518.

Imkeller, P. and I. Pavlyukevich (2006a). "First exit times of SDEs driven by stable Lévy processes". *Stochastic Processes and their Applications* 116.4, pp. 611–642. DOI: 10.1016/j.spa.2005.11.006.

Imkeller, P. and I. Pavlyukevich (2006b). "Lévy flights: transitions and meta-stability". *J. Phys. A: Math. Gen.* 39.15, p. L237. DOI: 10.1088/0305-4470/39/15/L01.

IPCC (2023). "Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". *IPCC*. DOI: 10.59327/IPCC/AR6-9789291691647.

Isachsen, P. E. and K. H. Christensen (2017). "Introduction to Physical Oceanography". *Lecture Notes, University of Oslo*.

Jackson, L. C. et al. (2015). "Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM". *Climate Dynamics* 45.11, pp. 3299–3316. DOI: 10.1007/s00382-015-2540-2.

Jackson, L. C. et al. (2022). "The evolution of the North Atlantic Meridional Overturning Circulation since 1980". *Nat Rev Earth Environ* 3.4, pp. 241–254. DOI: 10.1038/s43017-022-00263-2.

Jackson, L. C. et al. (2023). "Understanding AMOC stability: the North Atlantic Hosing Model Intercomparison Project". *Geosci. Model Dev.* 16.7, pp. 1975–1995. DOI: 10.5194/gmd-16-1975-2023.

Jacques-Dumas, V., H. A. Dijkstra, and C. Kuehn (2024). "Resilience of the Atlantic meridional overturning circulation". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 34.12, p. 123162. DOI: 10.1063/5.0226410.

Jayne, S. R. et al. (2017). "The Argo Program: Present and Future". *Oceanography* 30.2, pp. 18–28. DOI: 10.5670/oceanog.2017.213.

Jiang, W., G. Gastineau, and F. Codron (2021). "Multicentennial Variability Driven by Salinity Exchanges Between the Atlantic and the Arctic Ocean in a Coupled Climate Model". *J Adv Model Earth Syst* 13.3, e2020MS002366. DOI: 10.1029/2020MS002366.

Johns, W. E. et al. (2023). "Towards two decades of Atlantic Ocean mass and heat transports at 26.5°N". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381.2262, p. 20220188. DOI: 10.1098/rsta.2022.0188.

Johnsen, S. J. et al. (1992). "Irregular glacial interstadials recorded in a new Greenland ice core". *Nature* 359.6393, pp. 311–313. DOI: 10.1038/359311a0.

Kang, S. M., I. M. Held, D. M. W. Frierson, and M. Zhao (2008). "The Response of the ITCZ to Extratropical Thermal Forcing: Idealized Slab-Ocean Experiments with a GCM". *Journal of Climate* 21.14, pp. 3521–3532. DOI: 10.1175/2007JCLI2146.1.

Kantz, H. and P. Grassberger (1985). "Repellers, semi-attractors, and long-lived chaotic transients". *Physica D: Nonlinear Phenomena* 17.1, pp. 75–86. DOI: 10.1016/0167-2789(85)90135-6.

Kappler, J., M. E. Cates, and R. Adhikari (2024). *Sojourn probabilities in tubes and pathwise irreversibility for Itô processes*. DOI: 10.48550/arXiv.2009.04250.

Kaszás, B., T. Haszpra, and M. Herein (2019). "The snowball Earth transition in a climate model with drifting parameters: Splitting of the snapshot attractor". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.11, p. 113102. DOI: 10.1063/1.5108837.

Kaufman, D. et al. (2020). "Holocene global mean surface temperature, a multi-method reconstruction approach". *Sci Data* 7.1, p. 201. DOI: 10.1038/s41597-020-0530-7.

Keeling, C. D. et al. (1976). "Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii". *Tellus* 28.6, pp. 538–551. DOI: 10.3402/tellusa.v28i6.11322.

Kelly, D. and I. Melbourne (2017). "Deterministic homogenization for fast–slow systems with chaotic noise". *Journal of Functional Analysis* 272.10, pp. 4063–4102. DOI: 10.1016/j.jfa.2017.01.015.

Kendon, M. et al. (2024). "State of the UK Climate 2023". *International Journal of Climatology* 44.S1, pp. 1–117. DOI: 10.1002/joc.8553.

Kilbourne, K. H. et al. (2022). "Atlantic circulation change still uncertain". *Nature Geoscience* 15.3, pp. 165–167. DOI: 10.1038/s41561-022-00896-4.

Kindler, P., M. Guillevic, M. Baumgartner, J. Schwander, A. Landais, and M. Leuenberger (2014). "Temperature reconstruction from 10 to 120 kyr b2k from the NGRIP ice core". *Clim. Past* 10.2, pp. 887–902. DOI: 10.5194/cp-10-887-2014.

Klemm, K., V. M. Eguíluz, R. Toral, and M. S. Miguel (2003). "Global culture: A noise-induced transition in finite systems". *Phys. Rev. E* 67.4, p. 045101. DOI: 10.1103/PhysRevE.67.045101.

Klinger, B. A. and J. Marotzke (1999). "Behavior of Double-Hemisphere Thermohaline Flows in a Single Basin". *Journal of Physical Oceanography* 29.3, pp. 382–399. DOI: 10.1175/1520-0485(1999)029<0382:BODHTF>2.0.CO;2.

Kloeden, P. E. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Springer. DOI: 10.1007/978-3-662-12616-5.

Klose, A. K., J. F. Donges, U. Feudel, and R. Winkelmann (2024). "Rate-induced tipping cascades arising from interactions between the Greenland Ice Sheet and the Atlantic Meridional Overturning Circulation". *Earth System Dynamics* 15.3, pp. 635–652. DOI: 10.5194/esd-15-635-2024.

Knietzsch, M.-A., A. Schröder, V. Lucarini, and F. Lunkeit (2015). "The impact of oceanic heat transport on the atmospheric circulation". *Earth System Dynamics* 6.2, pp. 591–615. DOI: 10.5194/esd-6-591-2015.

Knutti, R. and T. F. Stocker (2002). "Limited Predictability of the Future Thermohaline Circulation Close to an Instability Threshold". *Journal of Climate* 15.2, pp. 179–186. DOI: 10.1175/1520-0442(2002)015<0179:LPOTFT>2.0.CO;2.

Koch, D., A. Nandan, G. Ramesan, I. Tyukin, A. Gorban, and A. Koseska (2024). "Ghost Channels and Ghost Cycles Guiding Long Transients in Dynamical Systems". *Physical Review Letters* 133.4, p. 047202. DOI: 10.1103/PhysRevLett.133.047202.

Krakovská, H., C. Kuehn, and I. P. Longo (2023). "Resilience of dynamical systems". *European Journal of Applied Mathematics*, pp. 1–46. DOI: 10.1017/S0956792523000141.

Kramers, H. A. (1940). "Brownian motion in a field of force and the diffusion model of chemical reactions". *Physica* 7.4, pp. 284–304. DOI: 10.1016/S0031-8914(40)90098-2.

Kraut, S. and U. Feudel (2003). "Enhancement of noise-induced escape through the existence of a chaotic saddle". *Phys. Rev. E* 67.1, p. 015204. DOI: 10.1103/PhysRevE.67.015204.

Kuehn, C. (2015). *Multiple time scale dynamics*. Vol. 191. Applied Mathematical Sciences. Springer. DOI: 10.1007/978-3-319-12316-5.

Kuehn, C. et al. (2022). "A general view on double limits in differential equations". *Physica D: Nonlinear Phenomena* 431, p. 133105. DOI: 10.1016/j.physd.2021.133105.

Kuhlbrodt, T., A. Griesel, M. Montoya, A. Levermann, M. Hofmann, and S. Rahmstorf (2007). "On the driving processes of the Atlantic meridional overturning circulation". *Reviews of Geophysics* 45.2. DOI: 10.1029/2004RG000166.

Kuhlbrodt, T., S. Titz, U. Feudel, and S. Rahmstorf (2001). "A simple model of seasonal open ocean convection". *Ocean Dynamics* 52.1, pp. 36–49. DOI: 10.1007/s10236-001-8175-3.

Kypke, K. and P. Ditlevsen (2024). "On the representation of multiplicative noise in modeling Dansgaard–Oeschger events". *Physica D: Nonlinear Phenomena* 466, p. 134215. DOI: 10.1016/j.physd.2024.134215.

Lai, Y.-C. and T. Tél (2011). *Transient Chaos: Complex Dynamics on Finite-Time Scales*. Springer. DOI: 10.1007/978-1-4419-6987-3.

Latif, M., J. Sun, M. Visbeck, and M. Hadi Bordbar (2022). "Natural variability has dominated Atlantic Meridional Overturning Circulation since 1900". *Nature Climate Change* 12.5, pp. 455–460. DOI: 10.1038/s41558-022-01342-4.

Lenton, T. M. et al. (2008). "Tipping elements in the Earth's climate system". *Proceedings of the National Academy of Sciences* 105.6, pp. 1786–1793. DOI: 10.1073/pnas.0705414105.

Lenton, T. M. et al., eds. (2023). *The Global Tipping Points Report 2023*. University of Exeter.

Lenton, T. M. et al. (2024). "Remotely sensing potential climate change tipping points across scales". *Nature Communications* 15.1, p. 343. DOI: 10.1038/s41467-023-44609-w.

Levermann, A. and A. Born (2007). "Bistability of the Atlantic subpolar gyre in a coarse-resolution climate model". *Geophysical Research Letters* 34.24. DOI: 10.1029/2007GL031732.

Li, T. and X. Li (2021). "Gamma-Limit of the Onsager–Machlup Functional on the Space of Curves". *SIAM J. Math. Anal.* 53.1, pp. 1–31. DOI: 10.1137/20M1310539.

Lin, B., Q. Li, and W. Ren (2021). "A Data Driven Method for Computing Quasipotentials". *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Vol. 145. Proceedings of Machine Learning Research, pp. 652–670.

Liu, W., A. V. Fedorov, S.-P. Xie, and S. Hu (2020). "Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate". *Science Advances* 6.26, eaaz4876. DOI: 10.1126/sciadv.aaz4876.

Liu, W., S.-P. Xie, Z. Liu, and J. Zhu (2017). "Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate". *Science Advances* 3.1, e1601666. DOI: 10.1126/sciadv.1601666.

Livina, V. N., F. Kwasniok, and T. M. Lenton (2010). "Potential analysis reveals changing number of climate states during the last 60 kyr". *Climate of the Past* 6.1, pp. 77–82. DOI: 10.5194/cp-6-77-2010.

Lohmann, J., D. Castellana, P. D. Ditlevsen, and H. A. Dijkstra (2021). "Abrupt climate change as a rate-dependent cascading tipping point". *Earth Syst. Dynam.* 12.3, pp. 819–835. DOI: 10.5194/esd-12-819-2021.

Lohmann, J., H. A. Dijkstra, M. Jochum, V. Lucarini, and P. D. Ditlevsen (2024). "Multistability and intermediate tipping of the Atlantic Ocean circulation". *Science Advances* 10.12, eadi4253. DOI: 10.1126/sciadv.adi4253.

Lohmann, J., A. B. Hansen, A. Lovo, R. Chapman, F. Bouchet, and V. Lucarini (2025). "The role of edge states for early warning of tipping points". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 481.2312. DOI: 10.1098/rspa.2024.0753.

Lohmann, J. and V. Lucarini (2024). "Melancholia states of the Atlantic Meridional Overturning Circulation". *Phys. Rev. Fluids* 9.12, p. 123801. DOI: 10.1103/PhysRevFluids.9.123801.

Lohmann, J., B. Wuyts, P. D. Ditlevsen, and P. Ashwin (2024). "On the predictability of possible storylines for forced complex systems". *J. Phys. Complex.* 5.3, p. 035015. DOI: 10.1088/2632-072X/ad7b95.

Lorenz, E. N. (1975). "Climatic predictability". *GARP Publications Series*. The Physical Basis of Climate and Climate Modelling 16, pp. 132–136.

Lorenz, E. N. (1963). "Deterministic Nonperiodic Flow". *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

— (1984). "Irregularity: a fundamental property of the atmosphere*". *Tellus A* 36A.2, pp. 98–110. DOI: 10.1111/j.1600-0870.1984.tb00230.x.

Loriani, S. et al. (2023). "Tipping points in ocean and atmosphere circulations". *EGUsphere* 2023, pp. 1–62. DOI: 10.5194/egusphere-2023-2589.

Lucarini, V. and F. Ragone (2011). "Energetics of Climate Models: Net Energy Balance and Meridional Enthalpy Transport". *Reviews of Geophysics* 49.1. DOI: 10.1029/2009RG000323.

Lucarini, V., R. Blender, C. Herbert, F. Ragone, S. Pascale, and J. Wouters (2014). "Mathematical and physical ideas for climate science". *Reviews of Geophysics* 52.4, pp. 809–859. DOI: 10.1002/2013RG000446.

Lucarini, V. and T. Bódai (2017). "Edge states in the climate system: exploring global instabilities and critical transitions". *Nonlinearity* 30.7, R32–R66. DOI: 10.1088/1361-6544/aa6b11.

— (2019). "Transitions across Melancholia States in a Climate Model: Reconciling the Deterministic and Stochastic Points of View". *Physical Review Letters* 122.15, p. 158701. DOI: 10.1103/PhysRevLett.122.158701.

— (2020). "Global stability properties of the climate: Melancholia states, invariant measures, and phase transitions". *Nonlinearity* 33.9, R59. DOI: 10.1088/1361-6544/ab86cc.

Lucarini, V. and M. D. Chekroun (2023). "Theoretical tools for understanding the climate crisis from Hasselmann's programme and beyond". *Nature Reviews Physics* 5.12, pp. 744–765. DOI: 10.1038/s42254-023-00650-8.

— (2024). "Detecting and Attributing Change in Climate and Complex Systems: Foundations, Green's Functions, and Nonlinear Fingerprints". *Physical Review Letters* 133.24, p. 244201. DOI: 10.1103/PhysRevLett.133.244201.

Lucarini, V., V. M. Galfi, J. Riboldi, and G. Messori (2023). "Typicality of the 2021 Western North America summer heatwave". *Environ. Res. Lett.* 18.1, p. 015004. DOI: 10.1088/1748-9326/acab77.

Lucarini, V., L. Serdukova, and G. Margazoglou (2022). "Lévy noise versus Gaussian-noise-induced transitions in the Ghil–Sellers energy balance model". *Nonlinear Processes in Geophysics* 29, pp. 183–205. DOI: 10.5194/npg-29-183-2022.

Lucarini, V. and P. H. Stone (2005). "Thermohaline Circulation Stability: A Box Model Study. Part II: Coupled Atmosphere–Ocean Model". *Journal of Climate* 18.4, pp. 514–529. DOI: 10.1175/JCLI-3279.1.

Luchinsky, D. G., R. S. Maier, R. Mannella, P. V. E. McClintock, and D. L. Stein (1999). "Observation of Saddle-Point Avoidance in Noise-Induced Escape". *Physical Review Letters* 82.9, pp. 1806–1809. DOI: 10.1103/PhysRevLett.82.1806.

Lynch-Stieglitz, J. (2017). "The Atlantic Meridional Overturning Circulation and Abrupt Climate Change". *Annual Review of Marine Science* 9.Volume 9, 2017, pp. 83–104. DOI: 10.1146/annurev-marine-010816-060415.

Maier, R. S. and D. L. Stein (1993). "Escape problem for irreversible systems". *Physical Review E* 48.2, p. 931. DOI: 10.1103/PhysRevE.48.931.

— (1997). "Limiting exit location distributions in the stochastic exit problem". *SIAM Journal on Applied Mathematics* 57.3, pp. 752–790. DOI: 10.1137/S0036139994271753.

Maier-Reimer, E., U. Mikolajewicz, and K. Hasselmann (1993). "Mean Circulation of the Hamburg LSG OGCM and Its Sensitivity to the Thermohaline Surface Forcing". *Journal of Physical Oceanography* 23.4, pp. 731–757. DOI: 10.1175/1520-0485(1993)023<0731:MCOTHL>2.0.CO;2.

Maiocchi, C. C., V. Lucarini, and A. Gritsun (2022). "Decomposing the dynamics of the Lorenz 1963 model using unstable periodic orbits: Averages, transitions, and quasi-invariant sets". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.3. DOI: 10.1063/5.0067673.

Maiocchi, C. C., V. Lucarini, A. Gritsun, and Y. Sato (2024). "Heterogeneity of the attractor of the Lorenz '96 model: Lyapunov analysis, unstable periodic orbits, and shadowing properties". *Physica D: Nonlinear Phenomena* 457, p. 133970. DOI: 10.1016/j.physd.2023.133970.

Malmierca-Vallet, I., L. C. Sime, P. J. Valdes, M. Klockmann, G. Vettoretti, and J. Slattery (2024). "The Impact of CO2 and Climate State on Whether Dansgaard–Oeschger Type Oscillations Occur in Climate Models". *Geophysical Research Letters* 51.13, e2024GL110068. DOI: 10.1029/2024GL110068.

Manabe, S. and R. J. Stouffer (1988). "Two Stable Equilibria of a Coupled Ocean-Atmosphere Model". *Journal of Climate* 1.9, pp. 841–866. DOI: 10.1175/1520-0442(1988)001<0841:TSEOAC>2.0.CO;2.

Manabe, S. and R. J. Stouffer (1995). "Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean". *Nature* 378.6553, pp. 165–167. DOI: 10.1038/378165a0.

— (1999). "Are two modes of thermohaline circulation stable?" *Tellus A: Dynamic Meteorology and Oceanography* 51.3, pp. 400–411. DOI: 10.3402/tellusa.v51i3.13461.

Manabe, S. and R. T. Wetherald (1967). "Thermal equilibrium of the atmosphere with a given distribution of relative humidity". *J. Atmos. Sci* 24.3, pp. 241–259.

Margazoglou, G., T. Grafke, A. Laio, and V. Lucarini (2021). "Dynamical landscape and multistability of a climate model". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477.2250, p. 20210019. DOI: 10.1098/rspa.2021.0019.

Marotzke, J. and J. Willebrand (1991). "Multiple Equilibria of the Global Thermohaline Circulation". *Journal of Physical Oceanography* 21.9, pp. 1372–1385. DOI: 10.1175/1520-0485(1991)021<1372:MEOTGT>2.0.CO;2.

Maslin, M. (2009). "Quaternary Climate Transitions and Cycles". *Encyclopedia of Paleoclimatology and Ancient Environments*. Springer, pp. 841–855. DOI: 10.1007/978-1-4020-4411-3_198.

Masoller, C. (2002). "Noise-Induced Resonance in Delayed Feedback Systems". *Physical Review Letters* 88.3, p. 034102. DOI: 10.1103/PhysRevLett.88.034102.

Masson-Delmotte, V. et al. (2012). "Greenland climate change: from the past to the future". *WIREs Climate Change* 3.5, pp. 427–449. DOI: 10.1002/wcc.186.

Masson-Delmotte, V. et al., eds. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. DOI: 10.1017/9781009157896.

Mayewski, P. A. et al. (2004). "Holocene climate variability". *Quaternary Research* 62.3, pp. 243–255. DOI: 10.1016/j.yqres.2004.07.001.

McCarthy, G. et al. (2012). "Observed interannual variability of the Atlantic meridional overturning circulation at 26.5°N". *Geophysical Research Letters* 39.19. DOI: 10.1029/2012GL052933.

McDonald, S. W., C. Grebogi, E. Ott, and J. A. Yorke (1985). "Fractal basin boundaries". *Physica D: Nonlinear Phenomena* 17.2, pp. 125–153. DOI: 10.1016/0167-2789(85)90001-6.

Mecking, J., S. Drijfhout, L. Jackson, and M. Andrews (2017). "The effect of model bias on Atlantic freshwater transport and implications for AMOC bi-stability". *Tellus A: Dynamic Meteorology and Oceanography* 69.1, p. 1299910. DOI: 10.1080/16000870.2017.1299910.

Mehling, O., K. Bellomo, M. Angeloni, C. Pasquero, and J. von Hardenberg (2022). "High-latitude precipitation as a driver of multicentennial variability of the AMOC in a climate model of intermediate complexity". *Climate Dynamics* 61, pp. 1519–1534. DOI: 10.1007/s00382-022-06640-3.

Mehling, O. and R. Börner (2023). *Code for the coupled Stommel–L84 model*. Version 1.0. DOI: 10.5281/zenodo.10370900.

Mehling, O., R. Börner, and V. Lucarini (2024). "Limits to predictability of the asymptotic state of the Atlantic Meridional Overturning Circulation in a conceptual climate model". *Physica D: Nonlinear Phenomena* 459, p. 134043. DOI: 10.1016/j.physd.2023.134043.

Meinshausen, M. et al. (2020). "The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500". *Geoscientific Model Development* 13.8, pp. 3571–3605. DOI: 10.5194/gmd-13-3571-2020.

Meng, Y., Y.-C. Lai, and C. Grebogi (2020). "Tipping point and noise-induced transients in ecological networks". *Journal of The Royal Society Interface* 17.171, p. 20200645. DOI: 10.1098/rsif.2020.0645.

Michel, S. L. L. et al. (2022). "Early warning signal for a tipping point suggested by a millennial Atlantic Multidecadal Variability reconstruction". *Nature Communications* 13.1, p. 5176. DOI: 10.1038/s41467-022-32704-3.

Milanković, M. (1941). *Kanon der Erdbestrahlung und seine Anwendung auf das Eiszeitenproblem*. Royal Serbian Academy Special Publications. Mihaila Ćurčića.

Moinat, L., J. Kasparian, and M. Brunetti (2024). "Tipping detection using climate networks". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 34.12, p. 123161. DOI: 10.1063/5.0230848.

Moresco, P. and S. Ponce Dawson (1999). "The PIM-simplex method: an extension of the PIM-triple method to saddles with an arbitrary number of expanding directions". *Physica D: Nonlinear Phenomena* 126.1, pp. 38–48. DOI: 10.1016/S0167-2789(98)00234-6.

Morozov, A. Y., D. Almutairi, S. V. Petrovskii, and Y.-C. Lai (2023). "Long transients in discontinuous time-discrete models of population dynamics". *Chaos, Solitons & Fractals* 174, p. 113707. DOI: 10.1016/j.chaos.2023.113707.

Morr, A. and N. Boers (2024). "Detection of Approaching Critical Transitions in Natural Systems Driven by Red Noise". *Phys. Rev. X* 14.2, p. 021037. DOI: 10.1103/PhysRevX.14.021037.

Mouginot, J. et al. (2019). "Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018". *Proceedings of the National Academy of Sciences* 116.19, pp. 9239–9244. DOI: 10.1073/pnas.1904242116.

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). "An Active Pulse Transmission Line Simulating Nerve Axon". *Proc. IRE* 50.10, pp. 2061–2070. DOI: 10.1109/JRPROC.1962.288235.

Navarra, A. and V. Simoncini (2010). *A Guide to Empirical Orthogonal Functions for Climate Data Analysis*. Springer.

Nian, D. et al. (2023). "A potential collapse of the Atlantic Meridional Overturning Circulation may stabilise eastern Amazonian rainforests". *Communications Earth & Environment* 4.1, p. 470. DOI: 10.1038/s43247-023-01123-7.

NOAA (2025). *How much of the ocean has been explored?* URL: https://oceanexplorer.noaa.gov/facts/explored.html (visited on 01/27/2025).

Nolting, B. C. and K. C. Abbott (2016). "Balls, cups, and quasi-potentials: quantifying stability in stochastic systems". *Ecology* 97.4, pp. 850–864. DOI: 10.1890/15-1047.1.

Nordhaus, W. (2013). "Integrated Economic and Climate Modeling". *Handbook of Computable General Equilibrium Modeling*. Vol. 1. Elsevier, pp. 1069–1131. DOI: 10.1016/B978-0-444-59568-3.00016-X.

North, G. R., R. F. Cahalan, and J. A. Coakley Jr. (1981). "Energy balance climate models". *Reviews of Geophysics* 19.1, pp. 91–121. DOI: 10.1029/RG019i001p00091.

Northrup, S. H. and J. A. McCammon (1983). "Saddle-point avoidance in diffusional reactions". *The Journal of Chemical Physics* 78.2, pp. 987–989. DOI: 10.1063/1.444804.

Nusse, H. E. and J. A. Yorke (1989). "A procedure for finding numerical trajectories on chaotic saddles". *Physica D: Nonlinear Phenomena* 36.1, pp. 137–156. DOI: 10.1016/0167-2789(89)90253-4.

Onsager, L. and S. Machlup (1953). "Fluctuations and Irreversible Processes". *Phys. Rev.* 91.6, pp. 1505–1512. DOI: 10.1103/PhysRev.91.1505.

Orbe, C. et al. (2023). "Atmospheric Response to a Collapse of the North Atlantic Circulation under a Mid-Range Future Climate Scenario: A Regime Shift in Northern Hemisphere Dynamics". *Journal of Climate* 36.19, pp. 6669–6693. DOI: 10.1175/JCLI-D-22-0841.1.

Oseledet, I. V. (1968). "A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems". *Transactions of the Moscow Mathematical Society* 19, pp. 197–231.

Ott, E. (2002). *Chaos in Dynamical Systems*. Cambridge University Press.

Panagiotopoulos, I., J. Starke, J. Sieber, and W. Just (2023). "Continuation with Noninvasive Control Schemes: Revealing Unstable States in a Pedestrian Evacuation Scenario". *SIAM J. Appl. Dyn. Syst.* 22.1, pp. 1–36. DOI: 10.1137/22M1482032.

Pavliotis, G. A. and A. Stuart (2008). *Multiscale Methods: Averaging and Homogenization*. Springer.

Pavliotis, G. A. (2014). *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Vol. 60. Texts in Applied Mathematics. New York, NY: Springer. DOI: 10.1007/978-1-4939-1323-7.

Peixóto, J. P. and A. H. Oort (1984). "Physics of climate". *Rev. Mod. Phys.* 56.3, pp. 365–429. DOI: 10.1103/RevModPhys.56.365.

Peltier, W. R. and G. Vettoretti (2014). "Dansgaard-Oeschger oscillations predicted in a comprehensive model of glacial climate: A "kicked" salt oscillator in the Atlantic". *Geophysical Research Letters* 41.20, pp. 7306–7313. DOI: 10.1002/2014GL061413.

Pikovsky, A. and A. Politi (2016). *Lyapunov Exponents: A Tool to Explore Complex Dynamics*. Cambridge University Press.

Pinski, F. J. and A. M. Stuart (2010). "Transition paths in molecules at finite temperature". *J. Chem. Phys.* 132.18, p. 184104. DOI: 10.1063/1.3391160.

Pisarchik, A. N. and U. Feudel (2014). "Control of multistability". *Physics Reports* 540.4, pp. 167–218. DOI: 10.1016/j.physrep.2014.02.007.

Povea-Pérez, Y., É. Guilyardi, A. V. Fedorov, and B. Ferster (2024). "The central role of the Atlantic meridional overturning circulation in the Bjerknes compensation". *Climate Dynamics* 62.1, pp. 575–587. DOI: 10.1007/s00382-023-06926-0.

Rackauckas, C. and Q. Nie (2017). "DifferentialEquations.jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia". *Journal of Open Research Software* 5.1, pp. 15–15. DOI: 10.5334/jors.151.

Ragone, F., J. Wouters, and F. Bouchet (2018). "Computation of extreme heat waves in climate models using a large deviation algorithm". *Proceedings of the National Academy of Sciences* 115.1, pp. 24–29. DOI: 10.1073/pnas.1712645115.

Rahmstorf, S. (1995). "Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle". *Nature* 378.6553, pp. 145–149. DOI: 10.1038/378145a0.

— (1996). "On the freshwater forcing and transport of the Atlantic thermohaline circulation". *Climate Dynamics* 12.12, pp. 799–811. DOI: 10.1007/s003820050144.

— (2002). "Ocean circulation and climate during the past 120,000 years". *Nature* 419.6903, pp. 207–214. DOI: 10.1038/nature01090.

— (2024). "Is the Atlantic Overturning Circulation Approaching a Tipping Point?" *Oceanography* 37.3, pp. 16–29. DOI: 10.5670/oceanog.2024.501.

Rahmstorf, S. et al. (2005). "Thermohaline circulation hysteresis: A model intercomparison". *Geophysical Research Letters* 32.23. DOI: 10.1029/2005GL023655.

Rahmstorf, S. et al. (2015). "Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation". *Nature Climate Change* 5.5, pp. 475–480. DOI: 10.1038/nclimate2554.

Rantanen, M. et al. (2022). "The Arctic has warmed nearly four times faster than the globe since 1979". *Commun Earth Environ* 3.1, pp. 1–10. DOI: 10.1038/s43247-022-00498-3.

Ravishankara, A. R., D. A. Randall, and J. W. Hurrell (2022). "Complex and yet predictable: The message of the 2021 Nobel Prize in Physics". *Proc. Natl. Acad. Sci. U.S.A.* 119.2, e2120669119. DOI: 10.1073/pnas.2120669119.

Riechers, K., G. Gottwald, and N. Boers (2024). "Glacial Abrupt Climate Change as a Multiscale Phenomenon Resulting from Monostable Excitable Dynamics". *Journal of Climate* 37.8, pp. 2741–2763. DOI: 10.1175/JCLI-D-23-0308.1.

Risken, H. (1996). "Fokker-Planck Equation". *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer, pp. 63–95. DOI: 10.1007/978-3-642-61544-3_4.

Ritchie, P., Ö. Karabacak, and J. Sieber (2019). "Inverse-square law between time and amplitude for crossing tipping thresholds". *Proc. R. Soc. A.* 475.2222, p. 20180504. DOI: 10.1098/rspa.2018.0504.

Ritchie, P. and J. Sieber (2016). "Early-warning indicators for rate-induced tipping". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26.9, p. 093116. DOI: 10.1063/1.4963012.

Ritchie, P. D. L. et al. (2020). "Shifts in national land use and food production in Great Britain after a climate tipping point". *Nature Food* 1.1, pp. 76–83. DOI: 10.1038/s43016-019-0011-3.

Ritchie, P. D., J. J. Clarke, P. M. Cox, and C. Huntingford (2021). "Overshooting tipping point thresholds in a changing climate". *Nature* 592.7855, pp. 517–523. DOI: 10.1038/s41586-021-03263-2.

Rockström, J. et al. (2009). "A safe operating space for humanity". *Nature* 461.7263, pp. 472–475. DOI: 10.1038/461472a.

Roebber, P. J. (1995). "Climate variability in a low-order coupled atmosphere-ocean model". *Tellus A* 47.4, pp. 473–494. DOI: 10.1034/j.1600-0870.1995.t01-3-00006.x.

Romanou, A. et al. (2023). "Stochastic Bifurcation of the North Atlantic Circulation under a Midrange Future Climate Scenario with the NASA-GISS ModelE". *Journal of Climate* 36.18, pp. 6141–6161. DOI: 10.1175/JCLI-D-22-0536.1.

Romé, Y. M., R. F. Ivanovic, L. J. Gregoire, D. Swingedouw, S. Sherriff-Tadano, and R. Börner (2025). "Simulated millennial-scale climate variability driven by a convection–advection oscillator". *Climate Dynamics* 63.3, p. 150. DOI: 10.1007/s00382-025-07630-x.

Römer, R. and P. Ashwin (2025). *The effect of timescale separation on the tipping window for chaotically forced systems*. DOI: 10.48550/arXiv.2504.12530.

Rooth, C. (1982). "Hydrology and ocean circulation". *Progress in Oceanography* 11.2, pp. 131–149. DOI: 10.1016/0079-6611(82)90006-4.

Rossi, K. L., R. C. Budzinski, E. S. Medeiros, B. R. R. Boaretto, L. Muller, and U. Feudel (2025). "Dynamical properties and mechanisms of metastability: A perspective in neuroscience". *Phys. Rev. E* 111.2, p. 021001. DOI: 10.1103/PhysRevE.111.021001.

Rousseau, D.-D., W. Bagniewski, and V. Lucarini (2023). "A punctuated equilibrium analysis of the climate evolution of cenozoic exhibits a hierarchy of abrupt transitions". *Scientific Reports* 13.1, p. 11290. DOI: 10.1038/s41598-023-38454-6.

Santos Gutiérrez, M. and V. Lucarini (2022). "On some aspects of the response to stochastic and deterministic forcings". *J. Phys. A: Math. Theor.* 55.42, p. 425002. DOI: 10.1088/1751-8121/ac90fd.

Sauer, J., J. Demaeyer, G. Zappa, F. Massonnet, and F. Ragone (2024). "Extremes of summer Arctic sea ice reduction investigated with a rare event algorithm". *Climate Dynamics* 62.6, pp. 5219–5237. DOI: 10.1007/s00382-024-07160-y.

Schäfer, B., M. Matthiae, X. Zhang, M. Rohden, M. Timme, and D. Witthaut (2017). "Escape routes, weak links, and desynchronization in fluctuation-driven networks". *Phys. Rev. E* 95.6, p. 060203. DOI: 10.1103/PhysRevE.95.060203.

Schätzing, F. (2005). *Der Schwarm*. Fischer Taschenbuch.

Scheffer, M., S. H. Hosper, M.-L. Meijer, B. Moss, and E. Jeppesen (1993). "Alternative equilibria in shallow lakes". *Trends in Ecology & Evolution* 8.8, pp. 275–279. DOI: 10.1016/0169-5347(93)90254-M.

Scheffer, M. et al. (2009). "Early-warning signals for critical transitions". *Nature* 461.7260, pp. 53–59. DOI: 10.1038/nature08227.

Schleussner, C.-F., A. Levermann, and M. Meinshausen (2014). "Probabilistic projections of the Atlantic overturning". *Climatic Change* 127.3, pp. 579–586. DOI: 10.1007/s10584-014-1265-2.

Schleussner, C.-F. et al. (2024). "Overconfidence in climate overshoot". *Nature* 634.8033, pp. 366–373. DOI: 10.1038/s41586-024-08020-9.

Schneider, T. M., J. F. Gibson, M. Lagha, F. De Lillo, and B. Eckhardt (2008). "Laminar-turbulent boundary in plane Couette flow". *Phys. Rev. E* 78.3, p. 037301. DOI: 10.1103/PhysRevE.78.037301.

Schorlepp, T., T. Grafke, S. May, and R. Grauer (2022). "Spontaneous symmetry breaking for extreme vorticity and strain in the three-dimensional Navier–Stokes equations". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2226, p. 20210051. DOI: 10.1098/rsta.2021.0051.

Schuss, Z. and A. Spivak (1998). "Where is the exit point?" *Chemical Physics* 235.1, pp. 227–242. DOI: 10.1016/S0301-0104(98)00078-0.

Scott, J. R., J. Marotzke, and P. H. Stone (1999). "Interhemispheric thermohaline circulation in a coupled box model". *Journal of Physical Oceanography* 29.3, pp. 351–365. DOI: 10.1175/1520-0485(1999)029<0351:ITCIAC>2.0.CO;2.

Sellers, W. D. (1969). "A Global Climatic Model Based on the Energy Balance of the Earth-Atmosphere System". *Journal of Applied Meteorology (1962-1982)* 8.3, pp. 392–400. DOI: 10.1175/1520-0450(1969)008<0392:AGCMBO>2.0.CO;2.

Semtner, A. J. (1976). "A Model for the Thermodynamic Growth of Sea Ice in Numerical Investigations of Climate". *Journal of Physical Oceanography* 6.3, pp. 379–389. DOI: 10.1175/1520-0485(1976)006<0379:AMFTTG>2.0.CO;2.

Sethian, J. A. and A. Vladimirsky (2003). "Ordered Upwind Methods for Static Hamilton–Jacobi Equations: Theory and Algorithms". *SIAM J. Numer. Anal.* 41.1, pp. 325–363. DOI: 10.1137/S0036142901392742.

Sgubin, G., D. Swingedouw, S. Drijfhout, Y. Mary, and A. Bennabi (2017). "Abrupt cooling over the North Atlantic in modern climate models". *Nature Communications* 8.1, p. 14375. DOI: 10.1038/ncomms14375.

Shakun, J. D. and A. E. Carlson (2010). "A global perspective on Last Glacial Maximum to Holocene climate change". *Quaternary Science Reviews* 29.15, pp. 1801–1816. DOI: 10.1016/j.quascirev.2010.03.016.

Sieber, J., O. E. Omel'chenko, and M. Wolfrum (2014). "Controlling Unstable Chaos: Stabilizing Chimera States by Feedback". *Physical Review Letters* 112.5, p. 054102. DOI: 10.1103/PhysRevLett.112.054102.

Simonnet, E. (2023). "Computing non-equilibrium trajectories by a deep learning approach". *Journal of Computational Physics* 491, p. 112349. DOI: 10.1016/j.jcp.2023.112349.

Sinet, S., A. S. Von Der Heydt, and H. A. Dijkstra (2023). "AMOC Stabilization Under the Interaction With Tipping Polar Ice Sheets". *Geophysical Research Letters* 50.2, e2022GL100305. DOI: 10.1029/2022GL100305.

Skufca, J. D., J. A. Yorke, and B. Eckhardt (2006). "Edge of Chaos in a Parallel Shear Flow". *Physical Review Letters* 96.17, p. 174101. DOI: 10.1103/PhysRevLett.96.174101.

Slater, T. et al. (2021). "Increased variability in Greenland Ice Sheet runoff from satellite observations". *Nature Communications* 12.1, p. 6069. DOI: 10.1038/s41467-021-26229-4.

Slyman, K. and C. K. Jones (2023). "Rate and noise-induced tipping working in concert". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 33.1, p. 013119. DOI: 10.1063/5.0129341.

Soons, J., T. Grafke, and H. A. Dijkstra (2024). "Optimal Transition Paths for AMOC Collapse and Recovery in a Stochastic Box Model". *Journal of Physical Oceanography* 54.12, pp. 2537–2552. DOI: 10.1175/JPO-D-23-0234.1.

Steffen, W. et al. (2018). "Trajectories of the Earth System in the Anthropocene". *Proceedings of the National Academy of Sciences* 115.33, pp. 8252–8259. DOI: 10.1073/pnas.1810141115.

Stephens, P. A., W. J. Sutherland, and R. P. Freckleton (1999). "What Is the Allee Effect?" *Oikos* 87.1, pp. 185–190. DOI: 10.2307/3547011.

Stommel, H. (1961). "Thermohaline Convection with Two Stable Regimes of Flow". *Tellus* 13.2, pp. 224–230. DOI: 10.1111/j.2153-3490.1961.tb00079.x.

Stone, P. H. (1978). "Constraints on dynamical transports of energy on a spherical planet". *Dynamics of Atmospheres and Oceans* 2.2, pp. 123–139. DOI: 10.1016/0377-0265(78)90006-4.

Stratonovich, R. L. (1971). "On the probability functional of diffusion processes". *Selected Trans. in Math. Stat. Prob* 10, pp. 273–286.

Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze (2007). "Arctic sea ice decline: Faster than forecast". *Geophysical Research Letters* 34.9. DOI: 10.1029/2007GL029703.

Stuart, A. M., J. Voss, and P. Wilberg (2004). "Conditional path sampling of SDEs and the Langevin MCMC method". *Communications in Mathematical Sciences* 2.4, pp. 685–697. DOI: 10.4310/CMS.2004.v2.n4.a7.

Sweet, D. and E. Ott (2000). "Fractal dimension of higher-dimensional chaotic repellors". *Physica D: Nonlinear Phenomena* 139.1, pp. 1–27. DOI: 10.1016/S0167-2789(99)00222-5.

Swingedouw, D., T. Fichefet, H. Goosse, and M. F. Loutre (2009). "Impact of transient freshwater releases in the Southern Ocean on the AMOC and climate". *Climate Dynamics* 33.2, pp. 365–381. DOI: 10.1007/s00382-008-0496-1.

Swingedouw, D., M.-N. Houssais, C. Herbaut, A.-C. Blaizot, M. Devilliers, and J. Deshayes (2022). "AMOC Recent and Future Trends: A Crucial Role for Oceanic Resolution and Greenland Melting?" *Front. Clim.* 4. DOI: 10.3389/fclim.2022.838310.

Szabó, K. G. and T. Tél (1994). "Transient chaos as the backbone of dynamics on strange attractors beyond crisis". *Physics Letters A* 196.1, pp. 173–180. DOI: 10.1016/0375-9601(94)91066-9.

Tantet, A. J. J. (2016). "Ergodic theory of climate: variability, stability and response". PhD Thesis. Utrecht University.

Terhaar, J., L. Vogt, and N. P. Foukal (2025). "Atlantic overturning inferred from air-sea heat fluxes indicates no decline since the 1960s". *Nature Communications* 16.1, p. 222. DOI: 10.1038/s41467-024-55297-5.

Testani, N. et al. (2025). "Balancing Earth science careers in an unequal world". *Communications Earth & Environment* 6.1. DOI: 10.1038/s43247-024-01964-w.

Thoning, K. W., P. P. Tans, and W. D. Komhyr (1989). "Atmospheric carbon dioxide at Mauna Loa Observatory: 2. Analysis of the NOAA GMCC data, 1974–1985". *Journal of Geophysical Research: Atmospheres* 94.D6, pp. 8549–8565. DOI: 10.1029/JD094iD06p08549.

Thorneywork, A. L., J. Gladrow, U. F. Keyser, M. E. Cates, R. Adhikari, and J. Kappler (2024). *Resolution dependence of most probable pathways with state-dependent diffusivity.* DOI: 10.48550/arXiv.2402.01559.

Timmermann, A., H. Gildor, M. Schulz, and E. Tziperman (2003). "Coherent Resonant Millennial-Scale Climate Oscillations Triggered by Massive Meltwater Pulses". *Journal of Climate* 16.15, pp. 2569–2585. DOI: 10.1175/1520-0442(2003)016<2569:CRMCOT>2.0.CO;2.

Titz, S., T. Kuhlbrodt, and U. Feudel (2002). "Homoclinic bifurcation in an ocean circulation box model". *International Journal of Bifurcation and Chaos* 12.04, pp. 869–875. DOI: 10.1142/S0218127402004759.

Touchette, H. (2009). "The large deviation approach to statistical mechanics". *Physics Reports* 478.1, pp. 1–69. DOI: 10.1016/j.physrep.2009.05.002.

Treguier, A. M., J. Deshayes, C. Lique, R. Dussin, and J. M. Molines (2012). "Eddy contributions to the meridional transport of salt in the North Atlantic". *Journal of Geophysical Research: Oceans* 117.C5. DOI: 10.1029/2012JC007927.

Treguier, A. M. et al. (2014). "Meridional transport of salt in the global ocean from an eddy-resolving model". *Ocean Science* 10.2, pp. 243–255. DOI: 10.5194/os-10-243-2014.

Tsitouras, C. (2011). "Runge–Kutta pairs of order 5(4) satisfying only the first column simplifying assumption". *Computers & Mathematics with Applications* 62.2, pp. 770–775. DOI: 10.1016/j.camwa.2011.06.002.

Vallis, G. K. (2017). *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press.

van Kampen, N. G. (1981). "Itô versus Stratonovich". *J Stat Phys* 24.1, pp. 175–187. DOI: 10.1007/BF01007642.

van Westen, R. M. and H. A. Dijkstra (2023). "Asymmetry of AMOC Hysteresis in a State-Of-The-Art Global Climate Model". *Geophysical Research Letters* 50.22, e2023GL106088. DOI: 10.1029/2023GL106088.

— (2024). "Persistent climate model biases in the Atlantic Ocean's freshwater transport". *Ocean Science* 20.2, pp. 549–567. DOI: 10.5194/os-20-549-2024.

van Westen, R. M., M. Kliphuis, and H. A. Dijkstra (2024). "Physics-based early warning signal shows that AMOC is on tipping course". *Science Advances* 10.6, eadk1189. DOI: 10.1126/sciadv.adk1189.

Vettoretti, G. and W. R. Peltier (2018). "Fast Physics and Slow Physics in the Nonlinear Dansgaard–Oeschger Relaxation Oscillation". *Journal of Climate* 31.9, pp. 3423–3449. DOI: 10.1175/JCLI-D-17-0559.1.

von der Heydt, A. S. et al. (2021). "Quantification and interpretation of the climate variability record". *Global and Planetary Change* 197, p. 103399. DOI: 10.1016/j.gloplacha.2020.103399.

Waddington, C. H. (1957). *The Strategy of the Genes*. Routledge. DOI: 10.4324/9781315765471.

Wales, D. J. (2015). "Perspective: Insight into reaction coordinates and dynamics from the potential energy landscape". *The Journal of Chemical Physics* 142.13, p. 130901. DOI: 10.1063/1.4916307.

Wang, H., Y. Yu, and G. Wen (2014). "Dynamical Analysis of the Lorenz-84 Atmospheric Circulation Model". *Journal of Applied Mathematics* 2014.1, p. 296279. DOI: 10.1155/2014/296279.

Weaver, A. J. et al. (2012). "Stability of the Atlantic meridional overturning circulation: A model intercomparison". *Geophysical Research Letters* 39.20. DOI: 10.1029/2012GL053763.

Weijer, W., W. Cheng, O. A. Garuba, A. Hu, and B. T. Nadiga (2020). "CMIP6 Models Predict Significant 21st Century Decline of the Atlantic Meridional Overturning Circulation". *Geophysical Research Letters* 47.12, e2019GL086075. DOI: 10.1029/2019GL086075.

Weijer, W. et al. (2019). "Stability of the Atlantic Meridional Overturning Circulation: A Review and Synthesis". *Journal of Geophysical Research: Oceans* 124.8, pp. 5336–5375. DOI: 10.1029/2019JC015083.

Weissmann, H., N. M. Shnerb, and D. A. Kessler (2018). "Simulation of spatial systems with demographic noise". *Phys. Rev. E* 98.2, p. 022131. DOI: 10.1103/PhysRevE.98.022131.

Westerhold, T. et al. (2020). "An astronomically dated record of Earth's climate and its predictability over the last 66 million years". *Science* 369.6509, pp. 1383–1387. DOI: 10.1126/science.aba6853.

Wieczorek, S., P. Ashwin, C. M. Luke, and P. M. Cox (2010). "Excitability in ramped systems: the compost-bomb instability". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 467.2129, pp. 1243–1269. DOI: 10.1098/rspa.2010.0485.

Wieczorek, S., C. Xie, and P. Ashwin (2023). "Rate-induced tipping: thresholds, edge states and connecting orbits". *Nonlinearity* 36.6, p. 3238. DOI: 10.1088/1361-6544/accb37.

Willeit, M. and A. Ganopolski (2024). "Generalized stability landscape of the Atlantic meridional overturning circulation". *Earth System Dynamics* 15.6, pp. 1417–1434. DOI: 10.5194/esd-15-1417-2024.

Wolff, E. W., J. Chappellaz, T. Blunier, S. O. Rasmussen, and A. Svensson (2010). "Millennial-scale variability during the last glacial: The ice core record". *Quaternary Science Reviews* 29.21, pp. 2828–2838. DOI: 10.1016/j.quascirev.2009.10.013.

Wood, R. A., J. M. Rodríguez, R. S. Smith, L. C. Jackson, and E. Hawkins (2019). "Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation". *Climate Dynamics* 53.11, pp. 6815–6834. DOI: 10.1007/s00382-019-04956-1.

Woollings, T., A. Hannachi, and B. Hoskins (2010). "Variability of the North Atlantic eddy-driven jet stream: Variability of the North Atlantic Jet Stream". *Q.J.R. Meteorol. Soc.* 136.649, pp. 856–868. DOI: 10.1002/qj.625.

Wouters, J. and V. Lucarini (2012). "Disentangling multi-level systems: averaging, correlations and memory". *J. Stat. Mech.* 2012.03, P03003. DOI: 10.1088/1742-5468/2012/03/P03003.

Wouters, J., R. K. H. Schiemann, and L. C. Shaffrey (2023). "Rare Event Simulation of Extreme European Winter Rainfall in an Intermediate Complexity Climate Model". *Journal of Advances in Modeling Earth Systems* 15.4, e2022MS003537. DOI: 10.1029/2022MS003537.

Wunderling, N. et al. (2024). "Climate tipping point interactions and cascades: a review". *Earth System Dynamics* 15.1, pp. 41–74. DOI: 10.5194/esd-15-41-2024.

Yang, S., S. F. Potter, and M. K. Cameron (2019). "Computing the quasipotential for nongradient SDEs in 3D". *Journal of Computational Physics* 379, pp. 325–350. DOI: 10.1016/j.jcp.2018.12.005.

Yeung, S. M. K. and S. H. Strogatz (1998). "Nonlinear dynamics of a solid-state laser with injection". *Phys. Rev. E* 58.4, pp. 4421–4435. DOI: 10.1103/PhysRevE.58.4421.

Yorke, J. A. and E. D. Yorke (1979). "Metastable chaos: The transition to sustained chaotic behavior in the Lorenz model". *J Stat Phys* 21.3, pp. 263–277. DOI: 10.1007/BF01011469.

Young, W. R. (2010). "Dynamic Enthalpy, Conservative Temperature, and the Seawater Boussinesq Approximation". *Journal of Physical Oceanography* 40.2, pp. 394–400. DOI: 10.1175/2009JPO4294.1.

Zheng, Y., F. Yang, J. Duan, X. Sun, L. Fu, and J. Kurths (2020). "The maximum likelihood climate change for global warming under the influence of greenhouse effect and Lévy noise". *Chaos* 30.1, p. 013132. DOI: 10.1063/1.5129003.

Zhou, J. X., M. D. S. Aliyu, E. Aurell, and S. Huang (2012). "Quasi-potential landscape in complex multi-stable systems". *Journal of The Royal Society Interface* 9.77, pp. 3539–3553. DOI: 10.1098/rsif.2012.0434.

Zhou, P. and T. Li (2016). "Construction of the landscape for multi-stable systems: Potential landscape, quasi-potential, A-type integral and beyond". *The Journal of Chemical Physics* 144.9, p. 094109. DOI: 10.1063/1.4943096.

Zilberman, N. V. et al. (2023). "Observing the full ocean volume using Deep Argo floats". *Front. Mar. Sci.* 10. DOI: 10.3389/fmars.2023.1287867.

Zimmerman, C. C., T. J. W. Wagner, E. A. Maroon, and D. E. McNamara (2025). "Slowed Response of Atlantic Meridional Overturning Circulation Not a Robust Signal of Collapse". *Geophysical Research Letters* 52.2, e2024GL112415. DOI: 10.1029/2024GL112415.