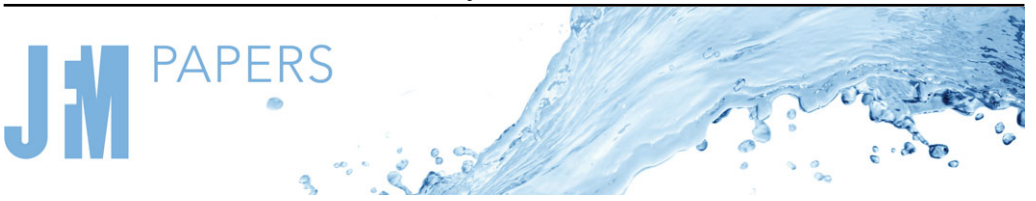
J. Fluid Mech. (2025), *vol.* 1009, A53, doi:10.1017/jfm.2025.248



Most likely noise-induced tipping of the overturning circulation in a two-dimensional Boussinesq fluid model

Jelle Soons[1], **Tobias Grafke**[2] **and Henk A. Dijkstra**[1]

[1]Institute for Marine and Atmospheric research Utrecht, Utrecht University, Princetonplein 5, Utrecht 3584 CC, The Netherlands

[2]Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

Corresponding author: Jelle Soons, j.soons@uu.nl

(Received 26 August 2024; revised 17 February 2025; accepted 19 February 2025)

There is a reasonable possibility that the present-day Atlantic Meridional Overturning Circulation is in a bistable regime, hence it is relevant to compute pathways of noise-induced transitions between the stable equilibrium states. Here, the most probable transition pathway of a noise-induced tipping of the northern overturning circulation in a spatially-continuous two-dimensional model with surface temperature and stochastic salinity forcings is computed directly using large deviation theory. This pathway reveals the fluid dynamical mechanisms of such a tipping. Paradoxically it starts off with a strengthening of the northern overturning circulation before a short but strong salinity pulse induces a second overturning cell. The increased atmospheric energy input of this two-cell configuration cannot be mixed away quickly enough, leading to the collapse of the northern overturning cell, and finally resulting in a southern overturning circulation. Additionally, the approach allows us to compare the probability of this transition under different parameters in the deterministic part of the salinity surface forcing, which quantifies the increase in transition probability as the bifurcation point of the system is approached.

Key words: ocean circulation, variational methods, buoyancy-driven instability

1. Introduction

The Atlantic Meridional Overturning Circulation (AMOC) plays a vital role in regulating Earth's climate. It transports warm upper ocean water northwards, which causes the rather mild climate in western Europe. When reaching the subpolar North Atlantic Ocean,

the relatively warm and salty water is cooled by the atmosphere and becomes the denser cold and salty North Atlantic Deep Water, which sinks and returns as a deep southward flow (Frajka-Williams *et al.* 2019). The AMOC is identified as a tipping element in the present-day climate system (Armstrong McKay *et al.* 2022), and its collapse would have severe consequences for the global climate (van Westen *et al.* 2024*b*).

Stommel (1961) already realised the AMOC's potential for tipping, as the AMOC is affected by two competing feedbacks due to the opposing effects of temperature and salinity on the density of seawater. A strengthening of the AMOC results in increased northward heat transport, causing a decrease in density of the upper subpolar North Atlantic water, which inhibits sinking. Consequently, the thermal circulation represents a negative feedback. On the other hand, AMOC strengthening also increases northward salt transport, which aids the northern deep water formation. Hence the haline circulation provides a positive feedback mechanism: the salt-advection feedback (Marotzke 2000). These feedbacks, together with the fact that ocean temperature anomalies are more strongly damped by the atmosphere than ocean salinity anomalies, result in a multiple equilibria regime for the AMOC. In the box model used by Stommel (1961), two stable equilibria exist in a range of the surface buoyancy forcing, hence transitions between these states can occur.

Since then, multiple equilibria regimes have been found in a hierarchy of AMOC models, from conceptual box models (Rooth 1982; Rahmstorf 1996; Lucarini & Stone 2005; Cimatoribus *et al.* 2014), two-dimensional (2-D) models (Quon & Ghil 1992; Thual & McWilliams 1992), three-dimensional ocean-only models (Dijkstra 2007), various Earth system models of intermediate complexity (Rahmstorf *et al.* 2005), global climate models (Hawkins *et al.* 2011), to modern Earth system models (van Westen & Dijkstra 2023). Based on observations of available stability indicators, such as the AMOC induced freshwater convergence (Dijkstra 2007), the present-day AMOC would be in such a regime (Weijer *et al.* 2019), although this claim is contested (Jackson & Wood 2018). Accordingly, there has been a growing interest in establishing the distance of the present-day AMOC to these bifurcation points, and whether it will cross a critical threshold in a bifurcation-induced tipping event. Based on early warning signals determined from reconstructed historical data, the AMOC is thought to be heading towards this critical bifurcation point (Caesar *et al.* 2018; Boers 2021), and estimates have been made for when this point will be reached (Ditlevsen & Ditlevsen 2023). However, many uncertainties, in particular in the sea surface temperature data to reconstruct the historical AMOC, remain (Ben-Yami *et al.* 2024).

In addition to a bifurcation-induced tipping, the AMOC can also undergo a noise-induced tipping, where the transition occurs due to variations in small-scale forcing processes represented as 'noise' (Dijkstra 2024). This is far more dangerous, as there are no reliable early-warning signals for a noise-induced transition as opposed to a bifurcation-induced one (Ditlevsen & Johnsen 2010), especially when considering non-Gaussian noise (Lucarini *et al.* 2022). Moreover, a bifurcation-induced tipping can occur only close to the bifurcation thresholds, whereas one induced by noise can occur as long as the system is in a multiple-equilibria regime. Several AMOC models have been developed where the ocean's surface is forced by stochastic salinity noise in order to study the statistics of these noise-induced transitions (Cessi 1994; Timmermann & Lohmann 2000). A major hurdle in this analysis is that noise-induced AMOC transitions are expected to be quite rare. There is no observational evidence for such a transition over the historical period, and computed probabilities for a transition in box models can be as low as 10^{-8} under realistic noise levels (Castellana *et al.* 2019; van Westen *et al.* 2024*a*). At these low probabilities, standard Monte Carlo techniques fail to obtain a transition path within reasonable computing

time. However, we would still like to obtain these rare transition paths as they can be qualitatively different from the more commonly studied bifurcation-induced transitions (Soons *et al.* 2024), and they may provide new early-warning signals for a noise-induced transition (Giorgini *et al.* 2020).

Here, we use a technique from the Freidlin–Wentzell theory of large deviations (Freidlin & Wentzell 1998) to directly compute the most likely transition path in the low-noise limit of a noise-induced tipping of the overturning circulation in a 2-D Boussinesq fluid model with stochastic surface salinity forcing. This model captures the salt-advection feedback mechanism well, although it does not represent the AMOC realistically. The technique entails minimising the Freidlin–Wentzell action, yielding this most likely transition path and its associated noise forcing. This minimisation is equivalent to maximising the probability of the applied stochastic forcing under the constraints that it causes a transition. This technique transforms this rare transition sampling problem into a deterministic minimisation problem. The corresponding path, also called the instanton, has been determined in other applications to study noise-induced transitions (Grafke *et al.* 2015; Grafke & Vanden-Eijnden 2019; Woillez & Bouchet 2020; Schorlepp *et al.* 2022), and we have previously computed the instantons for an AMOC collapse in a stochastic version of the Wood *et al.* (2019) box model (Soons *et al.* 2024). This work is a natural extension to that.

In the following, we introduce the 2-D Boussinesq model of thermohaline flow in § 2. Then in § 3, the methodology is described briefly, including a short introduction of the Freidlin–Wentzell theory of large deviations, which is subsequently applied to the model. The fluid dynamical mechanisms and energetics of the resulting most likely tipping of the overturning circulation are analysed in § 4. Next, the probability ratios of the transitions in the low-noise limit are presented for various surface forcings in § 5. Section 6 finally contains a summary and discussion of the results.

2. The 2-D Boussinesq model

We use a Boussinesq model of the zonally averaged thermohaline circulation in the double-hemispheric Atlantic basin with heat and salt forcing at the surface, where the latter has a stochastic component. It is considered for a rectangular basin of length L and depth H. The diffusivities of heat κ_T, salt κ_S, and momentum ν are assumed constant, and must be interpreted as eddy diffusivities. A linear equation of state holds, with thermal expansion and haline contraction coefficients α_T and α_S, respectively. The model used is a slight modification of the ones studied in Quon & Ghil (1992), Thual & McWilliams (1992) and Dijkstra & Molemaker (1997).

The governing equations are non-dimensionalised with scales H, H^2/κ_T, κ_T/H, ΔT and $\Delta S/\lambda$ for length, time, velocity, temperature and salinity, respectively. Here, ΔT and ΔS are characteristic meridional temperature and salinity differences, and λ is the buoyancy ratio ($\lambda = (\alpha_S\,\Delta S)/(\alpha_T\,\Delta T)$). The streamfunction ψ and vorticity ω are introduced, with the horizontal and vertical velocity (u, w) expressed by $u = \partial_z\psi$, $w = -\partial_x\psi$ and $\omega = \partial_x w - \partial_z u$. Hence the system is described by streamfunction $\psi(x, z, t)$, vorticity $\omega(x, z, t)$, temperature $T(x, z, t)$ and salinity $S(x, z, t)$. The non-dimensional governing equations are

$$Pr^{-1}\left(\frac{\partial\omega}{\partial t} + \frac{\partial\psi}{\partial z}\frac{\partial\omega}{\partial x} - \frac{\partial\psi}{\partial x}\frac{\partial\omega}{\partial z}\right) = \nabla^2\omega + Ra\left(\frac{\partial T}{\partial x} - \frac{\partial S}{\partial x}\right), \tag{2.1}$$

$$\omega = -\nabla^2\psi, \tag{2.2}$$

$$\frac{\partial T}{\partial t} + \frac{\partial \psi}{\partial z}\frac{\partial T}{\partial x} - \frac{\partial \psi}{\partial x}\frac{\partial T}{\partial z} = \nabla^2 T + \frac{h(z)}{\tau_T}\left(T_S(x) - T\right), \tag{2.3}$$

$$\frac{\partial S}{\partial t} + \frac{\partial \psi}{\partial z}\frac{\partial S}{\partial x} - \frac{\partial \psi}{\partial x}\frac{\partial S}{\partial z} = Le^{-1}\nabla^2 S + \frac{h(z)}{\tau_S}\left(S_S(x) + \tilde{S}_S(x,t)\right), \tag{2.4}$$

on the domain $(x, z) \in [0, A] \times [0, 1]$ for time $t \geqslant 0$. The basin's bottom is at $z = 0$, its surface is at $z = 1$, and the southern and northern ends are located at $x = 0$ and $x = A$, respectively. All the boundaries are assumed to be stress free, i.e.

$$\text{for } x = 0,\, A, \quad \psi = \omega = 0, \tag{2.5}$$

$$\text{for } z = 0,\, 1, \quad \psi = \omega = 0, \tag{2.6}$$

and additionally they are isolated and impervious to salt and temperature, so

$$\text{for } x = 0,\, A, \quad \frac{\partial T}{\partial x} = \frac{\partial S}{\partial x} = 0, \tag{2.7}$$

$$\text{for } z = 0,\, 1, \quad \frac{\partial T}{\partial z} = \frac{\partial S}{\partial z} = 0. \tag{2.8}$$

Equations (2.1) and (2.2) model the advection and diffusion of momentum, which is forced by the meridional temperature and salinity gradients. Note that using the streamfunction implies mass conservation. Equations (2.3) and (2.4) represent the advection–diffusion processes for heat and salinity, respectively. Temperature is forced by a restoring force, where it is always driven towards a fixed atmospheric temperature $T_S(x)$. Salinity, on the other hand, is forced by a flux, consisting of a deterministic and temporally constant part $S_S(x)$ and a stochastic part $\tilde{S}_S(x, t)$. All forcings have a vertical profile prescribed by

$$h(z) = \exp\left(\frac{z-1}{\delta_V}\right), \tag{2.9}$$

with $\delta_V \ll 1$ a characteristic thickness for the top boundary layer. In this way, the forcings are essentially applied only near the surface where $h(1) = 1$. Moreover, the temperature and salinity forcings have respective time scales τ_T and τ_S. Note that the stochastic forcing is acting only on the salinity component of our system, so we have degenerate noise. However, we still expect the system to be a hypo-elliptic diffusion, i.e. the noise spreads to all degrees of freedom in the system, since any perturbation to the surface salinity will propagate to the other components via the advective terms (Lucarini & Bódai 2020).

Finally, the following non-dimensional parameters are used: the Prandtl number Pr, the Lewis number Le, the thermal Rayleigh number Ra and the aspect ratio A. These are defined as

$$Pr = \frac{\nu}{\kappa_T}, \quad Le = \frac{\kappa_T}{\kappa_S}, \quad Ra = \frac{g\alpha_T\,\Delta T\,H^3}{\nu\kappa_T}, \quad A = \frac{L}{H}, \tag{2.10}$$

where g is the standard acceleration of gravity. Throughout this work, we take these to be constant: $Pr = 1$, $Le = 1$, $Ra = 4 \times 10^4$ and $A = 5$ which are based on the values found in Quon & Ghil (1992) and Dijkstra & Molemaker (1997). Moreover, we also fix $\tau_T = 0.1$, $\tau_S = 1$ and $\delta_V = 0.05$.

Note that the adopted parameter values are unrealistic for the present-day AMOC, as A and Ra should be altered by several orders of magnitude to represent the actual Atlantic basin and the observed diffusivities. However, they are chosen such that the model still has a stable diabatically driven pole-to-pole circulation despite omitting the relevant wind and tidal mechanisms that mainly drive the AMOC (Kuhlbrodt *et al.* 2007;

Johnson *et al.* 2019). The presented model, even though conceptual, is still relevant, as it captures the salt-advection feedback (Dijkstra 2024). This mechanism is responsible for the multi-stability of the AMOC and its collapse (Weijer *et al.* 2019; Vanderborght *et al.* 2024). The purpose of our model consequently is to represent the thermohaline aspect of the AMOC as opposed to including all relevant drivers (Cessi & Young 1992; Quon & Ghil 1992; Thual & McWilliams 1992; Dijkstra & Molemaker 1997). Hence our study aims to produce and analyse the most likely process in which the thermal and salt-advection feedbacks cause a tipping of this diabatic overturning circulation, which in turn is insightful for noise-induced AMOC tippings in more representable models.

2.1. *Surface forcing*

The prescribed atmospheric temperature is

$$T_S(x) = \frac{1}{2}\left(\cos\left(2\pi\left(\frac{x}{A} - \frac{1}{2}\right)\right) + 1\right), \tag{2.11}$$

which is identical to Dijkstra & Molemaker (1997) and is symmetric around the equator ($x = 1/2$), with hot tropics and cold poles. Regarding the freshwater forcing, an additional constraint is present, as we want the total salinity in the basin to be conserved. Integrating (2.4) over the complete basin yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1\int_0^A S\,\mathrm{d}x\,\mathrm{d}z = \frac{\delta}{\tau_S}\left(1 - \mathrm{e}^{-1/\delta}\right)\left(\int_0^A S_S(x)\,\mathrm{d}x + \int_0^A \tilde{S}_S(x,t)\,\mathrm{d}x\right), \tag{2.12}$$

so total salinity is conserved only if each of the integrated horizontal profiles of the deterministic and stochastic parts of the freshwater forcing are zero. For the deterministic part, we take

$$S_S(x) = 3.5\cos\left(2\pi\left(\frac{x}{A} - \frac{1}{2}\right)\right) - \beta\sin\left(\pi\left(\frac{x}{A} - \frac{1}{2}\right)\right), \tag{2.13}$$

which is asymmetric for $\beta > 0$, with the northern (sub)polar region being more freshened than its southern counterpart (and vice versa for $\beta < 0$). Note that integrated over the basin's horizontal dimension, the freshwater forcing is zero.

For the stochastic component, we assume that the noise is white in time as this simplifies the computations significantly, and it is standard for other AMOC models (Cessi 1994; Timmermann & Lohmann 2000; Castellana *et al.* 2019). Physically, this assumption implies that the time scale at which the sea surface salinities vary is much smaller than the time scale at which the large-scale circulation varies. We choose

$$\tilde{S}_S(x,t) = \sqrt{\frac{\varepsilon}{K}}\sum_{k=1}^{K}\dot{W}_k^{(1)}(t)\cos\left(\frac{2\pi}{A}kx\right) + \dot{W}_k^{(2)}(t)\sin\left(\frac{2\pi}{A}kx\right), \tag{2.14}$$

with $2K$ spatial components, variance $\varepsilon > 0$, and $2K$ independent Wiener processes $W_k^{(1)}(t)$ and $W_k^{(2)}(t)$ for $k \in \{1, \ldots, K\}$. This fulfils the salinity conservation constraint and is white in time as

$$\mathbb{E}\left(\tilde{S}_S(x,t)\,\tilde{S}_S(x,t')\right) = \varepsilon\,\delta(t - t'), \tag{2.15}$$

where $\delta(\cdot)$ is the Dirac delta distribution. The number of components K influences the noise behaviour greatly, so a considered choice for its value is required. Of course, a larger K will result in a higher resolution in the form of allowing the forcing to vary on smaller spatial scales, but it will also require additional CPU time to compute realisations and

the instanton trajectory. Following Adler *et al.* (2003, 2017) and Boot & Dijkstra (2025), the largest variation in mean annual precipitation is around the Inter tropical Convergence Zone. This band of high variance stretches for roughly 20° in latitude. Taking the Atlantic basin from 70°S to 70°N then results in a rough cut-off at $K = 7$, which we fix here in this paper. Finally, we base the size of the temporal variance ε on estimates of the annual standard deviation of the sea surface salinities in the Atlantic basin (Friedman *et al.* 2017). This yields $\sqrt{\varepsilon} \in (10^{-3}, 10^{-2})$, hence $\varepsilon \in (10^{-6}, 10^{-4})$.

2.2. *Model as a stochastic partial differential equation*

The model is rewritten as a standard stochastic partial differential equation (SPDE) in order to apply the Freidlin–Wentzell theory to it. We define the function

$$\boldsymbol{\phi} : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^3 \quad \text{with } \boldsymbol{\phi}(x, z, t) = (\omega(x, z, t), T(x, z, t), S(x, z, t))^T, \tag{2.16}$$

where $\Omega = [0, A] \times [0, 1]$ is the basin domain. Note that the streamfunction $\psi(x, z, t)$ is omitted as it has no time evolution equation, and we rewrite it as $\psi[\omega(x, z, t)]$. This is possible since the Poisson equation (2.2) with homogeneous Dirichlet boundaries has a unique well-defined $\psi(x, z, t)$ for every $\omega(x, z, t)$. So we write

$$\psi[\omega] = \mathcal{G}(\omega), \tag{2.17}$$

where $\mathcal{G}$ is the linear operator

$$\mathcal{G}(v) = -\int_\Omega G(x, z; \zeta, \xi)\, v(\zeta, \xi, t)\, \mathrm{d}\zeta\, \mathrm{d}\xi, \tag{2.18}$$

with G (Green's function) obeying

$$\nabla^2_{(x,z)} G(x, z; \zeta, \xi) = \delta(\zeta - x, \xi - z) \quad \text{for } (x, z) \in \Omega, \tag{2.19}$$

$$G(x, z; \zeta, \xi) = 0 \quad \text{for } (x, z) \in \partial\Omega. \tag{2.20}$$

Now, the deterministic part of the evolution equations is given by

$$f : \boldsymbol{\phi} \mapsto (f_1[\boldsymbol{\phi}], f_2[\boldsymbol{\phi}], f_3[\boldsymbol{\phi}])^T, \tag{2.21}$$

where

$$f_1[\boldsymbol{\phi}] = \partial_x \psi[\phi_1]\, \partial_z \phi_1 - \partial_z \psi[\phi_1]\, \partial_x \phi_1 + Pr\, \nabla^2 \phi_1 + Pr\, Ra\, (\partial_x \phi_2 - \partial_x \phi_3), \tag{2.22}$$

$$f_2[\boldsymbol{\phi}] = \partial_x \psi[\phi_1]\, \partial_z \phi_2 - \partial_z \psi[\phi_1]\, \partial_x \phi_2 + \nabla^2 \phi_2 + \frac{h(z)}{\tau_T} (T_S(x) - \phi_2), \tag{2.23}$$

$$f_3[\boldsymbol{\phi}] = \partial_x \psi[\phi_1]\, \partial_z \phi_3 - \partial_z \psi[\phi_1]\, \partial_x \phi_3 + Le^{-1}\, \nabla^2 \phi_3 + \frac{h(z)}{\tau_S} S_S(x). \tag{2.24}$$

Let $\boldsymbol{\eta}(x, z, t)$ represent the three-dimensional spatio-temporal white noise, with

$$\mathbb{E}\left(\eta_i(x, z, t)\, \eta_i(x', z', t')\right) = \delta(x - x')\, \delta(z - z')\, \delta(t - t') \quad \text{for } i \in \{1, 2, 3\}, \tag{2.25}$$

and note that

$$\left\{ \cos\left(\frac{2\pi}{A} kx\right) \cos\left(2\pi l z\right),\ \cos\left(\frac{2\pi}{A} kx\right) \sin\left(2\pi l z\right),\ \sin\left(\frac{2\pi}{A} kx\right) \cos\left(2\pi l z\right),\right.$$
$$\left. \sin\left(\frac{2\pi}{A} kx\right) \sin\left(2\pi l z\right) \right\}_{k,l \in \mathbb{N}_0} \tag{2.26}$$

is an orthonormal basis of $L^2(\Omega)$. We denote the associated basis transform operator with $\mathcal{U}$. As this is a unitary operator, we have that its adjoint is equal to its inverse ($\mathcal{U}^* = \mathcal{U}^{-1}$).

Furthermore, we define a projection operator $\mathcal{P}$ that projects a function in $L^2(\Omega)$ onto the space spanned by this basis under restrictions that $l = 0$ and $1 \leqslant k \leqslant K$. Then it holds that

$$\tilde{S}_S(x) = \sqrt{\frac{\varepsilon}{K}}\, \mathcal{U}^* \mathcal{P}\, \mathcal{U}(\boldsymbol{\eta})(x, z, t). \tag{2.27}$$

The model can then be written as the SPDE system

$$\partial_t \boldsymbol{\phi}(x, z, t) = f[\boldsymbol{\phi}(x, z, t)] + \sqrt{\varepsilon}\, \sigma(\boldsymbol{\eta}(x, z, t)), \tag{2.28}$$

where the operator σ acts on the white-in-space and white-in-time noise

$$\sigma = \begin{pmatrix} 0 \\ 0 \\ \dfrac{h(z)}{\tau_S \sqrt{K}} \mathcal{U}^* \mathcal{P} \mathcal{U} \end{pmatrix}. \tag{2.29}$$

2.3. *Deterministic equilibria*

Under the stated surface forcings, there are three distinct equilibria in the deterministic setting (i.e. $\varepsilon = 0$) that are relevant, and they are depicted in the partial bifurcation diagram in figure 1. Here, the asymmetry β of the deterministic freshwater forcing is the bifurcation parameter. There are two possible stable states, which both consist of one asymmetric overturning cell with downwelling near one pole; see figures 1(*a*,*c*). One state with downwelling near the basin's northern boundary is denoted as the ON state (as it mimics an active AMOC state), and one with downwelling in the south is denoted as the OFF state. As parameter β increases, the freshwater flux into the northern half increases as well, which inhibits downwelling there and weakens the ON state. Eventually, for $\beta > 0.11$, the freshwater flux is large enough to tip the ON state. Note that the model's dynamics is invariant under the transformations $\beta \to -\beta$ and $x \to A - x$. In other words, the ON state at β is exactly the OFF state at $-\beta$, but mirrored at the equator, $x = A/2$. Therefore, the OFF state tips for values $\beta < -0.11$. This creates a bistable regime for $|\beta| < 0.11$. Additionally, in this bistable regime, a saddle state is present, which is an unstable equilibrium with two symmetric overturning cells in each hemisphere, with downwelling near both poles; see figure 1(*b*). We restrict ourselves here to the bistable regime as this contains the relevant attractive structure for the noise-induced transition. The bifurcation structure outside of this region is in principle much more complicated, and a thorough bifurcation analysis and descriptions of the instability mechanisms can be found in Dijkstra & Molemaker (1997).

The ON and OFF states (also denoted as $\boldsymbol{\phi}_{ON}$ and $\boldsymbol{\phi}_{OFF}$) are pole-to-pole circulations that each can be qualitatively viewed as a superposition of a thermally driven symmetric circulation with deep water formation at both cold poles and upwelling near the equator, and a haline-driven circulation that sees the opposite flow with downwelling near the equator and upwelling at both fresh poles. The haline circulation is much weaker than the thermally driven one as the former is forced by a constant salinity flux while the latter is forced by a restoring boundary condition that imposes a fixed surface temperature. This causes sharper isopycnals and hence a stronger flow (Thual & McWilliams 1992). Therefore, in the pole-to-pole circulation, the thermally driven downwelling near one of the poles is stronger than the salinity-induced upwelling near the other. As volume needs to be conserved, a small vertical boundary layer for the downwelling is countered by upwelling in the rest of the basin. The horizontal length scale of this layer is approximately 0.9, which is consistent with the scaling argument presented in Quon & Ghil (1992).

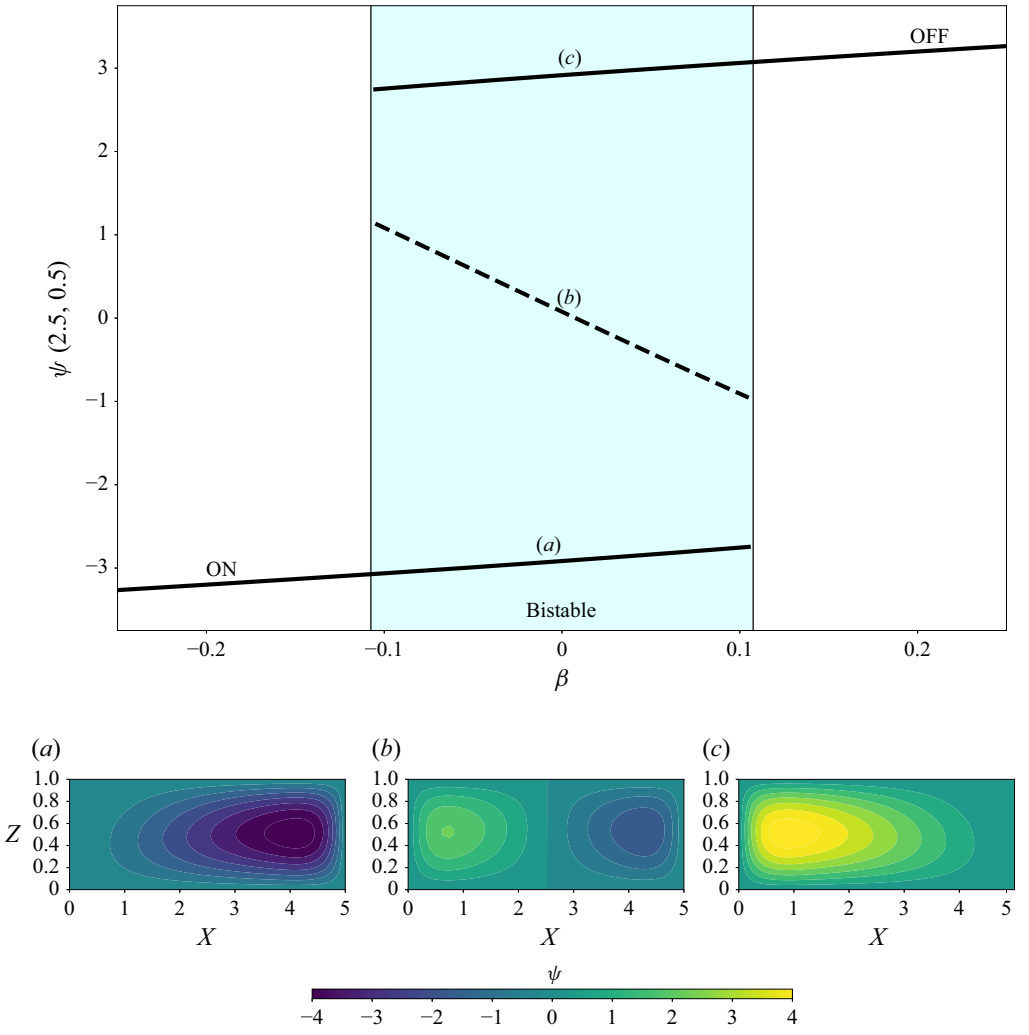


Figure 1. Partial bifurcation diagram of the 2-D Boussinesq model (top) with stable ON and OFF branches (solid) and unstable saddle states in the bistable region (dashed), with contour plots of the streamfunction ψ of examples of (*a*) the stable ON state, (*b*) the saddle, and (*c*) the stable OFF state, which are all indicated on the bifurcation diagram.

The stability of these states can be explained physically using arguments presented in Turner (1973). In the steady state, diffusion of salt and heat is balanced by buoyancy-induced mixing. If a perturbation at the surface increases the horizontal density gradient, and hence the overturning strength, then more warm and salty water will be transported to the bottom via downwelling, while more cold and fresh water wells up. This water needs to be heated and salinified by the surface forcing via diffusion. If this cannot keep up with the increased supply of deep water, then the horizontal density gradients will fall, so the overturning strength decreases again. This negative feedback mechanism stabilises the pole-to-pole circulations.

The saddle state, on the other hand, is completely thermally driven. Its instability is due to the competition between the two overturning cells: if a perturbation would cause one of the cells to be slightly larger than the other, then the larger cell would see a net input

of salt via the freshwater forcing. This would strengthen its downwelling, hence this cell would grow even more and eventually become the sole overturning cell.

3. Methodology

3.1. *Freidlin–Wentzell theory*

Consider an SPDE with additive noise

$$\partial_t \boldsymbol{U} = \mathcal{B}(\boldsymbol{U}) + \sqrt{\varepsilon}\,\mathcal{F}(\boldsymbol{\eta})(\boldsymbol{x}, t), \quad t \geqslant 0, \tag{3.1}$$

where $\boldsymbol{U} : \mathbb{R}^d \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^m$ is the m-dimensional state of the system as a function of space $\boldsymbol{x} \in \mathbb{R}^d$ and time $t \in \mathbb{R}_{\geqslant 0}$. The deterministic drift is given by the operator $\mathcal{B}(\boldsymbol{U})$, and operator $\mathcal{F}(\boldsymbol{\eta})$ acts on the Gaussian spatio-temporal white noise $\boldsymbol{\eta}$, which is scaled by smallness parameter ε. The covariance operator is denoted by $\mathcal{A} = \mathcal{F}^*\mathcal{F}$, which is the adjoint of $\mathcal{F}$ acting on itself. For simplicity, we treat only additive noise, but the theory can be extended to include multiplicative noise. It is non-trivial to show in general that this SPDE is well-posed, since the spatially rough noise necessitates carefully defining any nonlinearities present in $\mathcal{B}(\boldsymbol{U})$ (Hairer 2014). Since our choice of spatial noise covariance with a finite number of wavelengths K introduced a natural cut-off, such complications are avoided, and we can formally extend the Freidlin–Wentzell theory directly to the infinite-dimensional set-up (Grafke *et al.* 2017). Concretely, we are interested in situations where the stochastic process (3.1) realises a certain transition within a time τ starting at $\boldsymbol{U}(\boldsymbol{x}, 0) = \boldsymbol{U}_0(\boldsymbol{x})$ and ending at $\boldsymbol{U}(\boldsymbol{x}, \tau) = \boldsymbol{U}_1(\boldsymbol{x})$. This transition might be impossible in the deterministic setting ($\varepsilon = 0$), but can occur if noise is present ($\varepsilon > 0$), although it might be increasingly rare in the limit of small noise ($\varepsilon \to 0$).

Large deviation theory allows us to compute the rate at which this probability decays as the size of the noise ε approaches zero. The probability of observing a realisation close to a function $\boldsymbol{v} : \mathbb{R}^d \times [0, \tau] \to \mathbb{R}^m,\ (\boldsymbol{x}, t) \mapsto \boldsymbol{v}(\boldsymbol{x}, t)$, obeys

$$\mathbb{P}\left[\sup_{t\in[0,\tau]} \|\boldsymbol{U}(\boldsymbol{x}, t) - \boldsymbol{v}(\boldsymbol{x}, t)\|_{L^2} < \delta_R\right] \asymp \exp\left(-S_\tau[\boldsymbol{v}/\varepsilon]\right) \tag{3.2}$$

for sufficiently small $\delta_R > 0$. Here, $\asymp$ denotes log-asymptotic equivalence, and $\|\cdot\|_{L^2}$ is the L^2-norm in the spatial components. The functional $S_\tau[\boldsymbol{v}]$ is the Freidlin–Wentzell action, and is defined as

$$S_\tau[\boldsymbol{v}] = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^\tau \left\langle \partial_t \boldsymbol{v} - \mathcal{B}(\boldsymbol{v}),\ \mathcal{A}^{-1}\left(\partial_t \boldsymbol{v} - \mathcal{B}(\boldsymbol{v})\right)\right\rangle_{L^2} \mathrm{d}t & \text{if the integral converges,} \\ \infty & \text{otherwise,} \end{cases} \tag{3.3}$$

where $\langle\cdot,\cdot\rangle_{L^2}$ denotes the L^2 inner product in the spatial components, and $\mathcal{A}^{-1}$ is the inverse of the covariance operator. More technical constraints regarding the smoothness of the functions and operators at hand can be found in Freidlin & Wentzell (1998). In our case of degenerate noise, the covariance operator is not invertible, and there are several methods to circumvent the inversion (Grafke & Vanden-Eijnden 2019).

The major implication of (3.2) is that in the limit of small noise ($\varepsilon \to 0$), the trajectory $\tilde{\boldsymbol{v}}$ with the smallest action becomes the least unlikely trajectory to realise the transition. In other words,

$$\tilde{\boldsymbol{v}}(\boldsymbol{x}, t) = \underset{\boldsymbol{v}(\boldsymbol{x},0)=\boldsymbol{U}_1(\boldsymbol{x}),\ \boldsymbol{v}(\boldsymbol{x},\tau)=\boldsymbol{U}_2(\boldsymbol{x})}{\arg\min} S_\tau[\boldsymbol{v}] \tag{3.4}$$

is the maximum likelihood pathway (or instanton). Moreover, all realisations that fulfil the transition conditions will concentrate around $\tilde{\boldsymbol{v}}$ for low noise levels:

$$\lim_{\varepsilon\to 0}\mathbb{P}\left[\sup_{t\in[0,\tau]}\|\boldsymbol{U}(\boldsymbol{x},t)-\tilde{\boldsymbol{v}}(\boldsymbol{x},t)\|_{L^2}<\delta\middle|\boldsymbol{U}(\boldsymbol{x},0)=\boldsymbol{U}_1(\boldsymbol{x}),\ \boldsymbol{U}(\boldsymbol{x},\tau)=\boldsymbol{U}_2(\boldsymbol{x})\right]=1, \tag{3.5}$$

for $\delta>0$ sufficiently small. Our goal now is to compute this instanton $\tilde{\boldsymbol{v}}(\boldsymbol{x},t)$ for a tipping of the overturning circulation in our model. This will be the path whose associated forcing is the most likely to occur out of all possible stochastic forcings that induce a tipping in a time interval $t\in[0,\tau]$.

The minimisation (3.4) is a common problem in classical field theory (Altland & Simons 2010). The Lagrangian of (3.3) is

$$\mathcal{L}(\boldsymbol{v},\partial_t\boldsymbol{v})=\left\langle\partial_t\boldsymbol{v}-\mathcal{B}(\boldsymbol{v}),\ \mathcal{A}^{-1}(\partial_t\boldsymbol{v}-\mathcal{B}(\boldsymbol{v}))\right\rangle_{L^2}. \tag{3.6}$$

Then we can define the conjugate momentum to $\boldsymbol{v}$ as

$$\boldsymbol{\chi}(\boldsymbol{x},t)=\frac{\partial\mathcal{L}(\boldsymbol{v},\partial_t\boldsymbol{v})}{\partial(\partial_t\boldsymbol{v})}, \tag{3.7}$$

and so define the Hamiltonian as the Fenchel–Legendre transform of the Lagrangian:

$$\begin{aligned}\mathcal{H}(\boldsymbol{v},\boldsymbol{\chi})&=(\boldsymbol{\chi}\,\partial_t\boldsymbol{v}-\mathcal{L}(\boldsymbol{v},\partial_t\boldsymbol{v}))\Big|_{\partial_t\boldsymbol{v}=\partial_t\boldsymbol{v}(\boldsymbol{v},\boldsymbol{\chi})}\\&=\langle\mathcal{B}(\boldsymbol{v}),\boldsymbol{\chi}\rangle_{L^2}+\tfrac{1}{2}\langle\boldsymbol{\chi},\mathcal{A}(\boldsymbol{\chi})\rangle_{L^2}.\end{aligned} \tag{3.8}$$

Hence the Hamiltonian field equations, also called the instanton equations, follow directly as

$$\partial_t\boldsymbol{v}=\mathcal{B}(\boldsymbol{v})+\mathcal{A}(\boldsymbol{\chi}), \tag{3.9a}$$

$$\partial_t\boldsymbol{\chi}=-(\nabla_{\boldsymbol{v}}\mathcal{B})^*(\boldsymbol{\chi}), \tag{3.9b}$$

with boundary conditions $\boldsymbol{v}(\boldsymbol{x},0)=\boldsymbol{U}_1(\boldsymbol{x})$ and $\boldsymbol{v}(\boldsymbol{x},\tau)=\boldsymbol{U}_2(\boldsymbol{x})$. Solving these equations yields the instanton $\tilde{\boldsymbol{v}}(\boldsymbol{x},t)$. The advantage of reformulating the minimisation problem into (3.9) is that the covariance operator $\mathcal{A}$ no longer needs to be inverted. A disadvantage of the Hamiltonian framework is that we have one first-order partial differential equation with an initial value as well as a final condition, while the conjugate momentum equation has no temporal boundary conditions.

3.2. *The instanton equations for the Boussinesq flow*

We formulate the instanton equations (3.9) for the SPDE of the Boussinesq flow (2.2). Let $\boldsymbol{\phi}(x,z,t)=(\omega,T,S)^T(x,z,t)$ be the instanton path, and let $\boldsymbol{\theta}(x,z,t)=(\theta_\omega,\theta_T,\theta_S)^T(x,z,t)$ be its conjugate momentum. For the covariance operator $\mathcal{A}$, we find

$$\mathcal{A}=\sigma^*\sigma=\begin{pmatrix}0&0&0\\0&0&0\\0&0&\dfrac{(h(z))^2}{\tau_S^2K}\mathcal{U}^*\mathcal{P}\mathcal{U}\end{pmatrix}, \tag{3.10}$$

where we used that the adjoint of the basis transformation operator is its inverse ($\mathcal{U}^*=\mathcal{U}^{-1}$), and that $\mathcal{P}$ is a projection operator, so $\mathcal{P}^*=\mathcal{P}$ and $\mathcal{P}^2=\mathcal{P}$. Note that the covariance operator is clearly non-invertible, as expected: the noise in our model is degenerate.

Essentially, $\mathcal{A}(\boldsymbol{\theta})$ is a filter that yields the Fourier modes of θ_S that are constant along the vertical axis and have a horizontal wavelength A/k for $1 \leqslant k \leqslant K$. The amplitude of the noise is indicated by ε just as in (3.1). The set of partial differential equations (PDEs) describing the instanton $\boldsymbol{\phi}$ of a tipping of the overturning circulation in (2.2) is

$$\partial_t \omega + \partial_z \psi\, \partial_x \omega - \partial_x \psi\, \partial_z \omega = Pr\, \nabla^2 \omega + Pr\, Ra\, (\partial_x T - \partial_x S)\,, \tag{3.11a}$$

$$\partial_t T + \partial_z \psi\, \partial_x T - \partial_x \psi\, \partial_z T = \nabla^2 T + \frac{h(z)}{\tau_T}\,(T_S(x) - T)\,, \tag{3.11b}$$

$$\partial_t S + \partial_z \psi\, \partial_x S - \partial_x \psi\, \partial_z S = Le^{-1}\, \nabla^2 S + \frac{h(z)}{\tau_S} S_S(x) \tag{3.11c}$$

$$+ \frac{(h(z))^2}{\tau_S^2 K} \sum_{k=1}^{K} \theta_k^{(1)}(t) \cos\left(\frac{2\pi}{A} kx\right) + \theta_k^{(2)}(t) \sin\left(\frac{2\pi}{A} kx\right),$$

where

$$-\nabla^2 \psi = \omega \quad \text{with } \psi = 0 \text{ for } (x, z) \in \partial\Omega, \tag{3.12}$$

$$\theta_k^{(1)}(t) = \frac{2}{A} \int_\Omega \theta_S(x, z, t) \cos\left(\frac{2\pi}{A} kx\right) \mathrm{d}x\, \mathrm{d}z, \tag{3.13}$$

$$\theta_k^{(2)}(t) = \frac{2}{A} \int_\Omega \theta_S(x, z, t) \sin\left(\frac{2\pi}{A} kx\right) \mathrm{d}x\, \mathrm{d}z, \tag{3.14}$$

with spatial boundary conditions

$$\omega = 0 \quad \text{for } (x, z) \in \partial\Omega, \tag{3.15}$$

$$\partial_x T = \partial_x S = 0 \quad \text{for } x = 0, A, \tag{3.16}$$

$$\partial_z T = \partial_z S = 0 \quad \text{for } z = 0, 1, \tag{3.17}$$

and temporal boundary conditions

$$\boldsymbol{\phi}(x, z, 0) = \boldsymbol{\phi}_{ON}, \tag{3.18}$$

$$\boldsymbol{\phi}(x, z, \tau) = \boldsymbol{\phi}_{OFF}. \tag{3.19}$$

The set of PDEs for the conjugate momentum $\boldsymbol{\theta}$ is derived by first computing the variational derivatives of the deterministic drift f and their adjoints; see Appendix A. This results in the equations

$$\partial_t \theta_\omega + \partial_z \psi\, \partial_x \theta_\omega - \partial_x \psi\, \partial_z \theta_\omega + \nu_\omega + \nu_T + \nu_S + Pr\, \nabla^2 \theta_\omega = 0, \tag{3.20a}$$

$$\partial_t \theta_T + \partial_z \psi\ \ \partial_x \theta_T - \partial_x \psi\, \partial_z \theta_T + \nabla^2 \theta_T - \frac{h(z)}{\tau_T} \theta_T - Pr\, Ra\, \partial_x \theta_\omega = 0, \tag{3.20b}$$

$$\partial_t \theta_S + \partial_z \psi\, \partial_x \theta_S - \partial_x \psi\, \partial_z \theta_S + Le^{-1}\, \nabla^2 \theta_S + Pr\, Ra\, \partial_x \theta_\omega = 0, \tag{3.20c}$$

where the functions $\nu_\omega,\ \nu_T,\ \nu_S : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ obey the Poisson equations

$$-\nabla^2 \nu_\omega = \partial_x \omega\, \partial_z \theta_\omega - \partial_z \omega\, \partial_x \theta_\omega, \tag{3.21}$$

$$-\nabla^2 \nu_T = \partial_x T\, \partial_z \theta_T - \partial_z T\, \partial_x \theta_T, \tag{3.22}$$

$$-\nabla^2 \nu_S = \partial_x S\, \partial_z \theta_S - \partial_z S\, \partial_x \theta_S, \tag{3.23}$$

where

$$\nu_\omega = \nu_T = \nu_S = 0 \quad \text{for } (x, z) \in \partial\Omega, \tag{3.24}$$

with spatial boundary conditions

$$\theta_\omega = 0 \quad \text{for } (x, z) \in \partial\Omega, \tag{3.25}$$

$$\partial_x \theta_T = \partial_x \theta_S = 0 \quad \text{for } x = 0, A, \tag{3.26}$$

$$\partial_z \theta_T = \partial_z \theta_S = 0 \quad \text{for } z = 0, 1. \tag{3.27}$$

Altogether, PDEs (3.11) and (3.20) will yield the most likely path $\boldsymbol{\phi}$ from ON to OFF state within time $t \in [0, \tau]$ in the low-noise limit as $\varepsilon \to 0$. The associated forcing that brings this path about is then $\mathcal{A}(\boldsymbol{\theta})$.

3.3. *Numerical implementation*

As mentioned before, the problem with the previously described PDEs is that (3.11) describing $\boldsymbol{\phi}$ have two temporal boundary conditions, while (3.20) regarding the conjugate momentum $\boldsymbol{\theta}$ have none. To deal with this, we use the augmented Lagrangian method in which these equations are reformulated as a control problem (Hestenes 1969; Schorlepp *et al.* 2022); see Appendix B for more technical details. The PDEs there are solved using finite difference methods. We use a first-order semi-implicit Euler method to integrate the path and the conjugate momentum equations in time. The spatial derivatives are approximated with a second-order central difference approach (Veldman & Rinzema 1992). We use a non-equidistant grid to discretise Ω similarly as in Dijkstra *et al.* (1995). There are $M + 1$ grid points in the horizontal direction, and $N + 1$ points in the vertical. The grid points are defined as

$$x_m = \frac{m}{M} A \qquad \text{for } m \in \{0, 1, \ldots, M\}, \tag{3.28}$$

$$z_n = 0.5 + \tanh\left(q\left(\frac{n}{N} - \frac{1}{2}\right)\right) \Big/ \left(2\tanh\left(\frac{q}{2}\right)\right) \quad \text{for } q = 3 \text{ and } n \in \{0, 1, \ldots, N\}. \tag{3.29}$$

This way, we have a finer grid near the surface where the forcing is applied. For the spatial resolution, we take $(M, N) = (40, 80)$ unless otherwise stated. Here, priority is given to a high vertical resolution in order to accurately resolve the top boundary layer. Now the bifurcations of the system will shift under different spatial resolutions, but the same qualitative structure will remain (Dijkstra & Molemaker 1997). Similarly, the instanton will show only a quantitative shift under different resolutions, but the physical mechanisms will remain the same. Hence the following analyses can be carried out similarly for any resolution. For the time discretisation, we employ a step size $\Delta t = 0.01$.

In order to verify our method, we compare the computed instanton with several realised transitions at low noise levels. As it is computationally demanding to generate these transitions at the standard resolution, we compute the instanton and the realisations at the lower resolution $(M, N) = (15, 30)$. In figure 2, the resulting instanton is shown in several projected phase spaces together with multiple realised transitions. The instanton lies at the centre of the tube of transitions in salinity and streamfunction spaces, which is a good indication that it is indeed correct and representative of observed stochastic transitions. However, in the temperature variable, this agreement is less obvious; transitions projected onto the temperature space are quite erratic compared with those projected onto the salinity and streamfunction spaces. This illustrates that the underlying quasi-potential is relatively flat there, which can cause the most likely transition path for finite noise to deviate from the Freidlin–Wentzell instanton (Börner *et al.* 2024). If we compare the transitions for noise level $\varepsilon = 0.005$ to those with $\varepsilon = 0.0035$, then we see that the latter lie closer and around the instanton. This indicates that for even lower noise, the transition tube and instanton also agree in temperature space. Generating unbiased transition paths for even

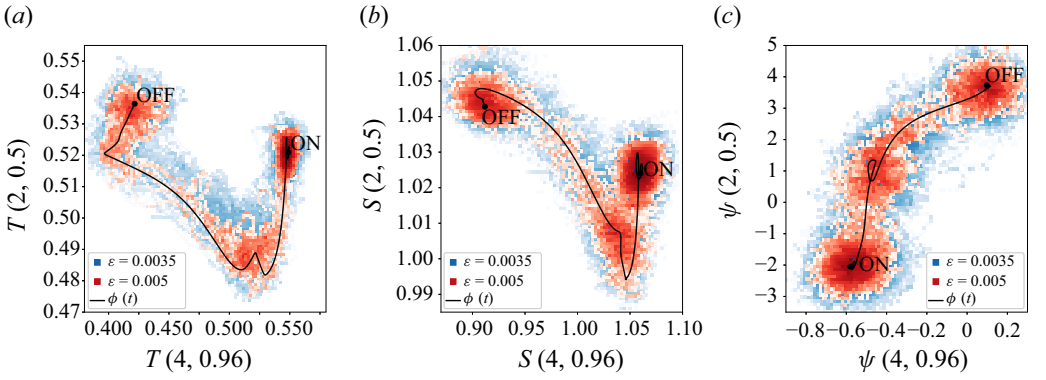


Figure 2. Instanton (black) and histograms of 55 realised transitions (blue) at noise level $\varepsilon = 0.005$, and 10 realised transitions (red) at noise level $\varepsilon = 0.0035$, both with logarithmic scaling in the various phase spaces (*a*) $T\,(4, 0.96)$, $T\,(2, 0.5)$, (*b*) $S(4, 0.96)$, $S(2, 0.5)$ and (*c*) $\psi\,(4, 0.96)$, $\psi\,(2, 0.5)$, with the ON and OFF states indicated. Model parameters are $\beta = 0.1$ and $(M, N) = (15, 30)$.

lower noise levels is, however, unfeasible as for $\varepsilon = 0.0035$, the Monte Carlo estimated probability of tipping within time $\tau = 20$ is approximately 10^{-3}. All in all, we can justify that the computed path is indeed the Freidlin–Wentzell instanton from the ON to OFF states.

In addition to the method's verification, we also need to argue its validity. The instanton will be a fair representation of a noise-induced AMOC tipping if two assumptions hold: (i) the noise is indeed white in time and spatially uniformly represented by the first K Fourier components; and (ii) the system is in the limit of low noise. Assumption (ii) has already been argued for the model used in Soons *et al.* (2024), and the noise compared to the AMOC strength should not depend too much on the model choice. Moreover, using the previously mentioned estimates for the variance ε, we find the transition probability in the low-resolution version of the model to be already less than 10^{-3}. Hence we can safely assume that the low-noise limit also holds for this model. The choice of noise structure (i) has already been argued in § 2. A thorough discussion of the noise structure of the atmospheric temperature and freshwater flux in the Atlantic Basin can be found in Boot & Dijkstra (2025).

4. The most likely tipping of the overturning circulation

In this section, we present the most likely transition path from the ON to the OFF state for $\beta = 0.1$. The value of β is chosen close to the bifurcation point $\beta = 0.11$, because a noise-induced transition is more probable close to the bifurcation point, and the resulting trajectory therefore more relevant. We allocated a time interval $\tau = 20$ for the forced part of the trajectory up to the separatrix. For larger time intervals, our procedure yields identical dynamics as a result, but with the trajectory initially spending additional time in the ON state. This is evidence that we are already in the long-time limit, and no further effects are expected by allowing a longer transition time. The resulting instanton is shown in figure 3, together with the associated optimal forcing in the Hovmöller diagram in figure 4. Here, we have restricted the instanton to the interval $t \in [0, 30]$, where end and beginning are sufficiently close to the OFF and ON states.

From these figures, we get a general idea of the tipping. The most likely freshwater forcing consists of a repeating pattern of first a freshening in the south ($x \lesssim 1.5$) followed by a salinification, while in the rest of the basin compensation occurs. The largest forcing

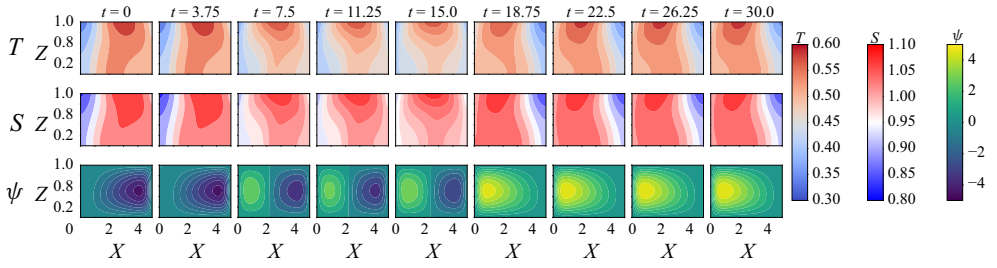


Figure 3. Instanton from $\boldsymbol{\phi}_{ON}$ to $\boldsymbol{\phi}_{OFF}$ for $\beta = 0.1$ at times $t \in \{0.0, 3.75, 7.5, 11.25, 15.0, 18.75, 22.5, 26.25, 30.0\}$ (left to right), with temperature T (top), salinity S (middle) and streamfunction ψ (bottom).

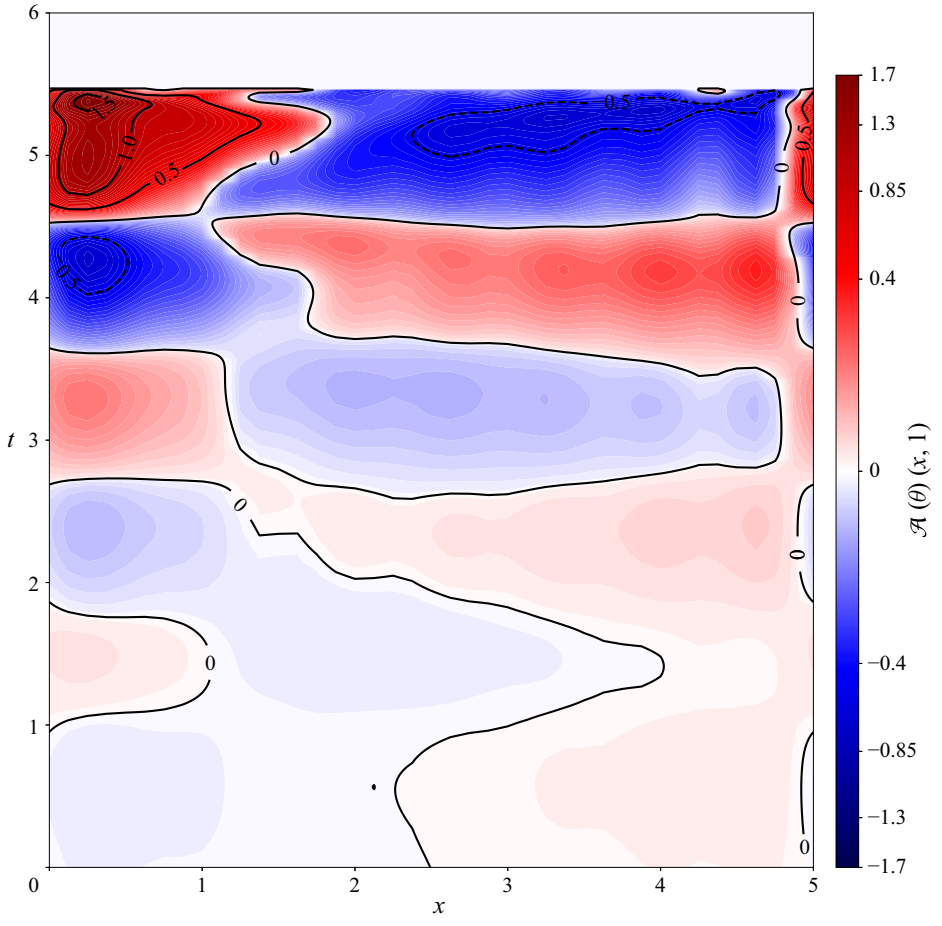


Figure 4. Hovmöller diagram of the salinity forcing at the surface ($\mathcal{A}(\boldsymbol{\theta})(x, 1)$) for $t \in [0, 6]$. Note that the colour mapping is nonlinear.

peak is reached at $t = 5.38$, and the forcing ceases completely at $t = 5.5$, at which point the separatrix is reached. At this point, the forcing has created a new additional overturning cell in the south. From there on, the system is in a symmetric two-cell overturning configuration. Note that the temperature and salinity distribution are still skewed towards the north. Only at $t \approx 15$ does the instanton approach the saddle state where both tracers

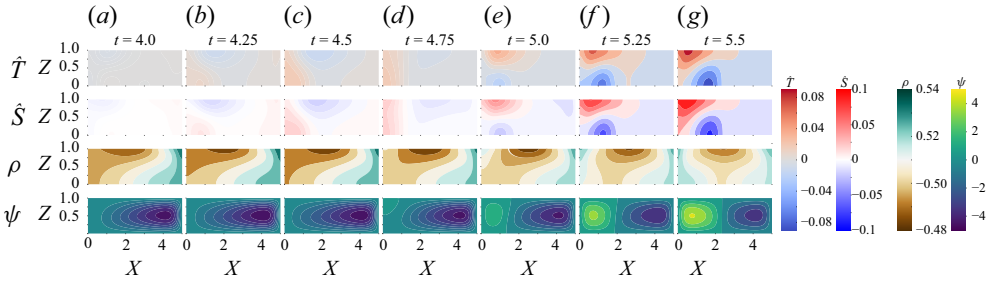


Figure 5. Instanton from $\boldsymbol{\phi}_{ON}$ to $\boldsymbol{\phi}_{OFF}$ for $\beta = 0.1$ at times $t \in \{4.0, 4.25, 4.5, 4.75, 5.0, 5.25, 5.5\}$ (left to right), with temperature anomaly $\hat{T} = T - T_{ON}$ (top), salinity anomaly $\hat{S} = S - S_{ON}$ (second row), density ρ (third row) and streamfunction ψ (bottom).

and the streamfunction have a symmetric distribution (the latter is actually antisymmetric). Subsequently, after passing the saddle, the original northern cell collapses, and the new southern cell grows. Eventually, the system converges to the stable OFF state. We will discuss this process in more detail, where we distinguish three stages: the forcing to create the new cell ($t \in [0, 5.5]$), the two-cell configuration leading to the saddle ($t \in [5.5, 15]$), and the collapse of the original cell ($t \in [15, 30]$). These will be analysed from mechanical and energetics perspectives.

4.1. *Fluid dynamical mechanism of the tipping*

4.1.1. *The forcing*

In figure 4, it is shown that the forcing consists of three repeating sequences. We will discuss only the last one ($t \in [4.0, 5.5]$), as this is the largest and similar in shape to the other two. The corresponding instanton trajectory is shown in figure 5, where salinity and temperature anomalies ($\hat{S} = S - S_{ON}$, $\hat{T} = T - T_{ON}$) and the dimensionless density ($\rho = S - T$) are depicted. Paradoxically, the most likely transition starts off with a strengthening of the pole-to-pole circulation. For $t \in [3.8, 4.6]$, a freshwater pulse is added to the southern part of the basin, while the rest of the basin is salinified. This increases the horizontal density gradients and hence the strength of the overturning cell too, which causes this freshwater anomaly to be transported from the south to north, and this salinity anomaly is transported to the bottom; see figures 5(*b*,*c*). If we take the minimum of ψ as the overturning strength of the northern cell, then the original strength is $\psi_{min,ON} \approx -4.25$, it peaks at $t = 4.3$ with $\psi_{min} \approx -4.71$, and returns to its original level at $t = 4.74$. However, this latter state is a transient state, and its anomalies with respect to the stable ON state are shown in figures 5(*c*,*d*). These show that in this state, the positive salinity and temperature anomalies have traversed the overturning circulation partially and are now near the bottom of the southern boundary, while their adverse anomalies are near the equator. Looking at the dimensionless density, we see that this transient state is less stable: the vertical density gradient in the south has risen to nearly zero.

After the initial southern freshwater pulse and strengthening, a strong salinity forcing is now applied to the southern region for $t \in [4.6, 5.5]$, while the rest of the basin is freshened as compensation. One can see in figure 5(*d*) that as the initial positive salinity anomaly upwells in the south, the strong salinity perturbation is applied there at the surface. This way, a large salinification can be achieved there using multiple smaller perturbations, as we can see from figure 4, which are evidently more likely to occur than one large perturbation. Meanwhile, due to the now weakened upwelling in the south, relatively fresh and cool

water stays behind near the bottom, causing negative anomalies there (figure 5*e*). The effects of this large salinification pulse start to appear at $t \approx 5.0$ (figure 5*e*): a positive temperature and salinity anomaly in the south on top of a negative anomaly. These dipoles induce a new second overturning cell in the south as the heavier water sinks and the lighter water rises. Note that this temperature dipole pattern has a dampening effect on the new cell, while the salinity dipole aids the new cell. At $t = 5.5$ (figure 5*g*), the new cell is strong enough to maintain its own density gradients, and the two-cell configuration is established. The forcing is no longer needed, and from here on, the system will reach the OFF state in a deterministic way.

Regarding the applied optimal forcing, we will discuss three aspects: its magnitude, its timing and its location. As the instanton is the path that is the most efficient to reach the OFF state in terms of applied forcing, we expect the created density anomalies to be just large enough to generate the second cell. We consider the volume integrated vorticity equation (2.1):

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \omega \,\mathrm{d}x\,\mathrm{d}z = Pr\left[\int_0^1 \partial_x\omega\big|_{x=A} - \partial_x\omega\big|_{x=0}\,\mathrm{d}z + \int_0^A \partial_z\omega\big|_{z=1} - \partial_z\omega\big|_{z=0}\,\mathrm{d}x\right] + Pr\,Ra\int_0^1 (S-T)\big|_{x=0} - (S-T)\big|_{x=A}\,\mathrm{d}z, \tag{4.1}$$

with the two terms on the right-hand side representing diffusion and buoyancy forcing, respectively. To reach a two-cell configuration, we need a second cell that is approximately as strong as the first one, i.e. the negative vorticity blob in the north has to be countered by a similar but positive vorticity blob in the south. Due to this symmetry, the terms representing the diffusion cancel out, and the volume-integrated vorticity balance for the two-cell configuration becomes

$$0 \approx Pr\,Ra\int_0^1 \rho\big|_{x=0} - \rho\big|_{x=A}\,\mathrm{d}z, \tag{4.2}$$

with dimensionless density $\rho = S - T$. So the required density anomalies are such that the cumulative density at the southern boundary is as large as that at the northern boundary. Intuitively, one can argue that equal densities at the meridional boundaries imply downwelling of equal strength near the poles and hence two overturning cells of equal strength. In figure 6, these depth-integrated densities near the poles are shown together with the total forcing $\langle \boldsymbol{\theta}, \mathcal{A}(\boldsymbol{\theta})\rangle_{L^2}$. It can be seen that as soon as the southern depth-integrated density is as large as the northern one (at $t = 5.35$), the forcing rapidly decreases to zero; the newly formed cell is strong enough, and the forcing is no longer needed. Note that as the forcing decreases, the southern density overshoots the northern density, and at $t = 5.5$ – as the forcing has ceased – the cumulative densities are equal again. From there on, as the instanton continues deterministically, the southern density is again lower than the northern density, but it is now significantly higher than originally in the ON state, as there is now downwelling near the southern pole. This continues until they are again equal when the saddle state is reached at $t \approx 15$, after which the northern cell collapses.

Note that beforehand we see the previous initial strengthening in action, with the cumulative density in the south reaching a local minimum at $t \approx 4.3$. For $t \in [5.5, 15]$, we have $\int_0^1 \rho_{x=0}\,\mathrm{d}z < \int_0^1 \rho_{x=A}\,\mathrm{d}z$, which seems to partially contradict the previous reasoning, but the cause is that the two-cell configuration is still slightly asymmetric in the vorticity,

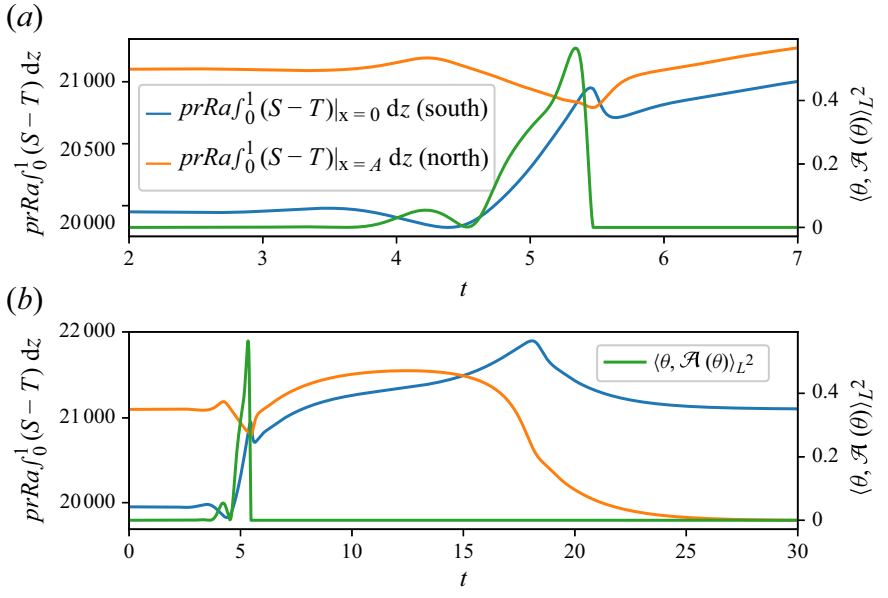


Figure 6. The cumulative density $Pr\,Ra\int_0^1 \rho\,\mathrm{d}z$ in the south ($x=0$, blue) and the north ($x=A$, orange) (left-hand axis), and the total forcing ($\langle\theta, \mathcal{A}(\theta)\rangle_{L^2}$, green) (right-hand axis) for $t\in(2,7)$ (top) and $t\in(0,30)$ (bottom).

so the assumptions do not hold completely. During this time interval, the new cell is slightly smaller than the original one, but it can still sustain itself after its establishment.

Considering the timing of the forcing, it is found that the alternating pulses each last for approximately 0.90 ± 0.05. As discussed, a positive (negative) salinity anomaly created by a pulse in the southern part of the basin weakens (strengthens) the original cell, resulting in an decreased (increased) salinity concentration at the bottom of the southern part. This upwells as a new alternate pulse is applied to the southern surface. In this way, two alternating pulses constitute the effect of one big perturbation. As seen in figures 5(*b–e*), the upwelling of the anomaly in the south in the ON state takes approximately 0.9 in time, which explains the observed frequency.

Finally, regarding the location of the forcing, we see that the forcing alternates sign around $x\approx1.5$, with the eventual goal of creating a large salinity concentration in this part of the domain (figure 5). Now, in order to create a second overturning cell in the south, we need a positive vorticity blob, which is induced by strong negative salinity gradients. For $\beta=0.11$, the deterministic salinity forcing gradient $\partial_x S_S$ has its minimum for $x\approx1.3$. The optimal stochastic salinity forcing has its largest gradient also around this point as it switches sign, exacerbating the negative salinity gradient and hence increasing the vorticity in the southern part of the basin.

4.1.2. *The two-cell configuration*

From $t=5.5$ to $t=15.5$, the instanton persists in a two-cell configuration. An overview is provided in figure 7, where ψ_{min} and ψ_{max} indicate the strengths of the northern and southern overturning cells, respectively, and x_s is the vertically averaged horizontal coordinate away from the boundaries where $\psi\approx0$, i.e. it is the approximate boundary between the two overturning cells. For the saddle state, it can be deduced that $x_s\approx0.49A$ for $\beta=0.1$ since at this value, the deterministic salinity forcings to each cell are equal.

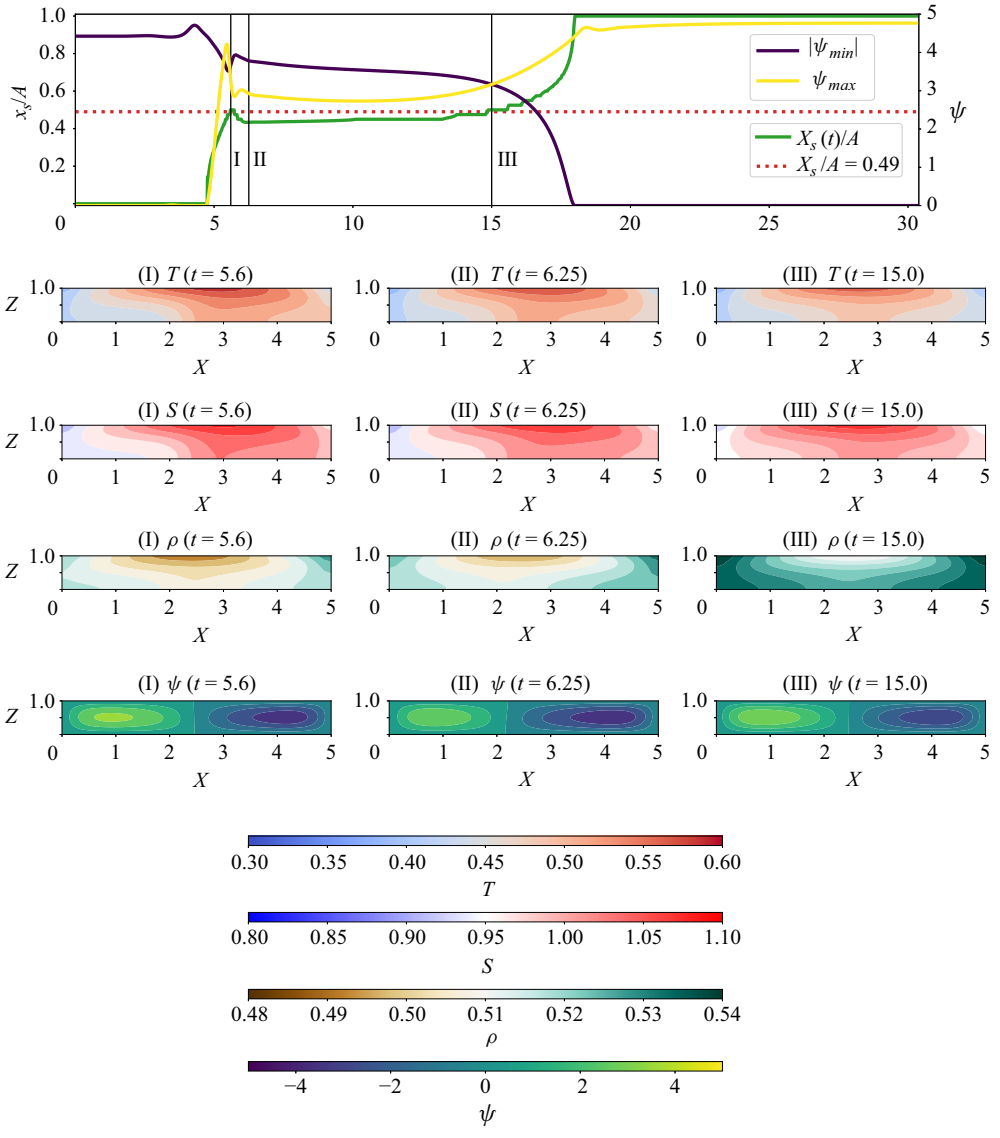


Figure 7. Top: $x_s(t)/A$ (green) along the instanton together with saddle-state values $x_s/A = 0.49$ (red, dashed) on the left-hand axis, and $\|\psi_{min}\|$ (purple) and ψ_{max} (yellow) indicating the strength of the northern and southern cells, respectively, on the right-hand axis. States I, II and III are indicated and visualised in the columns (left to right respectively), with temperature T (top row), salinity S (second row), density ρ (third row), and streamfunction ψ (bottom row).

From figure 7, it follows that initially, after the stochastic forcing has ceased, the new southern cell reaches a local maximum in size and strength before weakening; see state I at $t = 5.6$. To start the new cell, the density near the southern surface has been increased by the salinity forcing. As this sinks, its strength peaks followed by a peak in cell size. Now this surface water is not directly replaced by new equally dense water, as can be seen in the density distribution of state I, where dense water has sunk to the bottom in the south without dense water following. This causes the intermediate weakening. Once at $t \approx 6$, the densities in the southern cell have redistributed into this persistent state II.

Interestingly, at state II, the southern cell is both smaller and weaker than the northern cell, but it nevertheless grows, so that eventually the instanton reaches the symmetric saddle state III at $t = 15$. Note that in state II, the salinity and temperature distribution are still skewed towards the north, but they complement each other such that the density distribution and consequently the streamfunction are approximately symmetric around the equator. Moreover, the overall density is lower than in the saddle state III. To examine how the weaker and smaller southern cell can grow, we formulate the various transports of density into both cells. With $\Omega_S = [0, x_s] \times [0, 1]$ and $\Omega_N = \Omega - \Omega_S$, we denote the southern and northern cells, respectively, and F^i and Q^i stand for the salinity and heat transport into cell i. For the salinity transports we have

$$\text{deterministic surface salinity transport} \quad F_d^S = \frac{1}{\tau_S} \int_{\Omega_S} h(z)\, S_S(x)\, \mathrm{d}x\, \mathrm{d}z \approx \frac{\delta}{\tau_S} \int_0^{x_s} S_S(x)\, \mathrm{d}x, \tag{4.3}$$

$$\text{advective salinity transport} \quad F_a^S = \int_0^1 \partial_z \psi\, S\big|_{x=x_s}\, \mathrm{d}z, \tag{4.4}$$

$$\text{diffusive salinity transport} \quad F_f^S = \int_0^1 \partial_x S\big|_{x=x_s}\, \mathrm{d}z, \tag{4.5}$$

$$\text{stochastic salinity transport} \quad F_s^S = \int_{\Omega_S} \mathcal{A}(\theta)\, \mathrm{d}x\, \mathrm{d}z, \tag{4.6}$$

where it holds that $F_j^S = -F_j^N$ due to conservation of salinity, where $j \in \{d, a, f, s\}$ indicates the transport process. Moreover, as $\psi(x = x_s) \approx 0$, we have that the advective transport is negligible. For the heat transports Q, we have similar formulations. Note that here we can also neglect advective transport, and $Q_f^S = -Q_f^N$, but Q_d^S and Q_d^N do not necessarily sum to zero due to the nature of the restoring boundary condition. The total salinity and heat transport into Ω_N are then $F^N = -F_d^S - F_f^S - F_s^S$ and $Q^N = Q_d^N - Q_f^S$.

The results are shown in figure 8, which illustrates that the salinity dynamics is responsible for reaching the two-cell configuration of state I, while the temperature dynamics is the most significant in reaching the saddle state III. Examining figure 8(*a*) shows that during the two-cell interval ($t \in [5.5, 15]$), there is diffusive transport of salt from north to south (as most of the salt is still left in the northern part of the basin), which is almost completely compensated by the deterministic salinity flux at the surface (as $x_s < 0.49A$). As the cell grows, the deterministic salt flux at the surface and the diffusive exchange approach zero and then switch sign, where the former slightly outpaces the latter, so that eventually there is a net salt influx from the north to the south. However, the almost zero net salinity exchange ($F^S \approx 10^{-4}$) between the two cells during the bigger part of the interval cannot cause the growth of the southern overturning cell. The temperature dynamics in figure 8(*b*) are therefore more interesting, where we can see that from II to III, there is a net southward diffusive transport of heat as most of it is still in the north as a remnant of the original ON state, similar to salinity. However, the southern cell sees a cooling by the surface forcing, while the northern cell heats up, which for both cells outweighs the diffusive transport. This is a result of the northern cell's surface being exposed to the hot equator, with only the north pole cooling it, whereas the southern cell is exposed only to the cold south pole. As the latter cell grows, we see that both Q_d^S and Q_d^N approach zero and then switch signs.

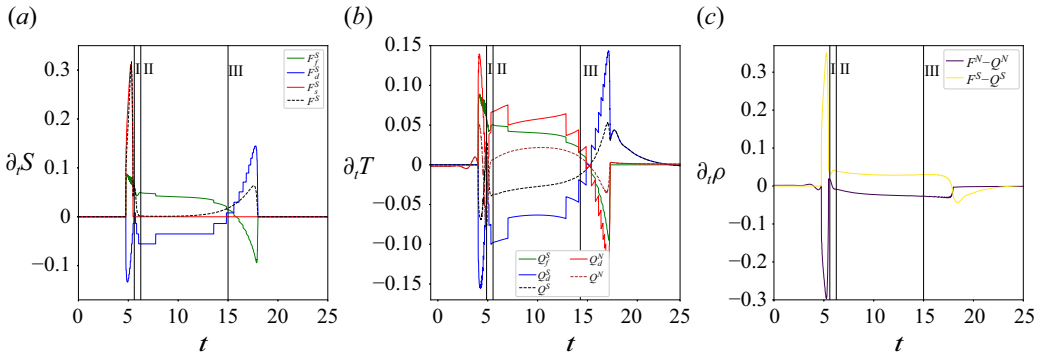


Figure 8. (*a*) The salinity transports into Ω_S with diffusive transport F_f^S (green), deterministic transport F_d^S (blue), stochastic transport F_s^S (red) and the total F^S (black dashed). (b) The heat transports into Ω_S with diffusive transport Q_f^S (green), deterministic transport Q_d^S (blue) and total Q^S (black dashed), and into Ω_N with deterministic transport Q_d^N (red) and total Q^N (brown dashed). (*c*) The total transport of density into the northern cell $F^N - Q^N$ (yellow) and into the southern cell $F^S - Q^S$ (purple). States I, II and III as in figure 7 are indicated.

The net effect of both heat and salinity transports is that already starting at state II there is a net density input into the southern cell and a slight density transport out of the northern cell. The smaller southern cell has a disadvantage with regard to the freshwater surface forcing, but this is completely compensated by the diffusive southward salinity transport. The surface heating, on the other hand, will increase the smaller cell's density, and decrease that of the larger cell. This forcing cannot be compensated by diffusive transport. The key difference is that the salinity forcing is a constant flux applied to the surface, whereas the surface heating is a restoring force. All in all, the new weaker cell sees a net mass import, hence it can grow in size and strength. Eventually, the cumulative densities at each pole grow to be equal (see figure 6) and the saddle state is reached.

Interestingly, note that at II, the most destabilising tracer is temperature. The streamfunction is close to the symmetric two-cell configuration, and the salt disparity between the two cells is held up by the constant surface forcing. Only the restoring temperature forcing nudges the state towards the saddle. Moreover, as we approach the saddle, there has been an overall mass increase. Due to the two-cell state, less flow has passed through the equator, hence the heating has been less effective, so the whole basin in total experiences a net cooling. Indeed, as seen in figure 7, the density distribution in III is still symmetric but elevated.

4.1.3. *The collapse of the northern cell*

At $t \approx 15$, the instanton passes near the saddle state. Beforehand, this cannot necessarily be expected: as discussed in Soons *et al.* (2024) the finite-time instanton might be a better representation of the transition path in the low-noise limit than the infinite-time instanton. This latter one will pass through the saddle, but – in case the transitions avoid the saddle (Börner *et al.* 2024) – is not necessarily a good exemplar. For this model, the transitions nicely follow the instanton as it passes near the saddle, which can be seen in figure 2, where the three distinct equilibria are the regions where the realisations linger. Therefore, saddle-avoidance is no issue here, and the saddle state indeed mediates the transition.

Around $t \approx 18$, the original northern cell collapses as $x_s \to A$. In figures 8(*a*,*b*), we can see that leading up to this collapse, now both heat and salt transport play a role. As the southern cell expands beyond the equator, the surface salinity forcing aids it, while the

surface heating inhibits it. Both are again countered by diffusive transports, but now the diffusive salt transport cannot keep up with the surface forcing as the salinity distribution is no longer skewed but symmetric, unlike earlier (during $t \in [5.5, 15]$). So as the southern cell grows, more salt and heat are transported southwards, and their distributions get more skewed, as is visible in figure 3. The salinity contribution is larger ($F^S > |Q^S|$), so the southern surface densifies, which aids downwelling there and hence increases the southward transport, which is a signature of the salt-advection feedback. Eventually, the southern cell has grown to fill the whole basin, and the OFF state is almost reached, with upwelling in the north where the water is cold and fresh, and downwelling in the south where the water is warm and saline.

From $t \approx 18$, the pole-to-pole circulation converges to the OFF state by a net heating ($Q_S > 0$). During the two-cell configuration, less water passed through the equator, which provided a net cooling of the whole basin. Now that there is again a cross-equatorial flow, a net heating occurs, and the overall density of the system decreases (see figure 8*c*). The overall strength of the OFF state is slightly larger than that of the ON state ($\psi_{max}\big|_{t=30} \approx 4.77$ versus $|\psi_{min}|\big|_{t=0} \approx 4.42$) as the salinity forcing is skewed towards the south.

Finally, note that after the collapse of the northern cell, there is a slight overshoot in overturning strength, with a local maximum $\psi_{max}\big|_{t=18.4} \approx 4.65$ that drops to $\psi_{max}\big|_{t=18.9} \approx 4.59$ before reaching the OFF state. This decrease is caused by the relatively flat isopycnals in the north just after the collapse, which inhibit upwelling there. Once these anomalies are mixed out, and a net heating has taken place, the OFF state is attained.

4.2. *Energetics of the tipping*

The instanton trajectory is also analysed in terms of its energetics. We use the framework devised by Winters *et al.* (1995) for density-stratified Boussinesq fluid flows. Similar approaches have been applied to the meridional overturning circulation (Hughes *et al.* 2009; Hogg *et al.* 2013). We derive the following energy balances:

$$\frac{\mathrm{d}}{\mathrm{d}t}E_k = \Phi_z - \mathcal{D}, \tag{4.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E_p = \Phi_i - \Phi_z - Q_T, \tag{4.8}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E_b = \Phi_a + \Phi_d, \tag{4.9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}E_a = \Phi_i - \Phi_z - \Phi_a - \Phi_d - Q_T, \tag{4.10}$$

where E_k, E_p, E_b and E_a are the kinetic energy, potential energy, background potential energy and available potential energy, respectively, with Φ_z, $\mathcal{D}$, Φ_i, Q_T, Φ_a and Φ_d indicating the buoyancy flux, the dissipation, the conversion of internal energy to potential energy, surface heating, rate of change of E_b due to surface forcings, and the energy flux by diapycnal mixing, respectively. Their definitions and derivations can be found in Appendix C.

An overview of these energies and fluxes along the transition is shown in figure 9. The kinetic energy is driven solely by the buoyancy forcing (Φ_z), which is followed by a slightly delayed dissipation ($\mathcal{D}$). Initially, it peaks at $t = 4.3$ as the original pole-to-pole circulation is strengthened, followed by another peak at $t = 5.5$ as the second cell has just formed. Both peaks coincide with the peaks in the forcings Φ_z and $\mathcal{A}(\boldsymbol{\theta})$ (figure 6). As these cease, E_k falls and state I is reached. A slight adjustment follows to state II, and from here on, both the buoyancy forcing and dissipation decrease slowly as the tracers become more

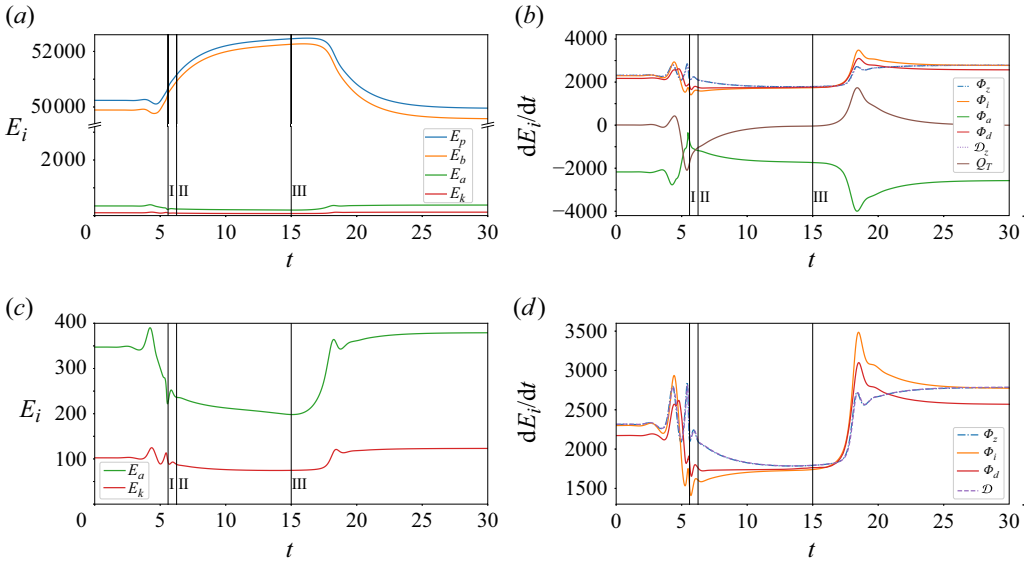


Figure 9. (*a*) The energies E_p (blue), E_b (yellow), E_a (green) and E_k (red), with (*c*) zoomed-in portion with E_a and E_k. (*b*) The fluxes Φ_z (blue dashed), Φ_i (yellow), Φ_a (green), Φ_d (red), $\mathcal{D}$ (purple dashed) and Q_T (brown) with (*d*) zoomed-in portion with Φ_z, Φ_i, Φ_d and $\mathcal{D}$. States I, II and III as in figure 7 are indicated.

evenly distributed. The kinetic energy stays level, and is lower in the two-cell configuration than in the pole-to-pole setting as flow speeds are generally reduced. The kinetic energy budget does not change as the saddle is passed. When the northern cell collapses, the buoyancy forcing rises and overshoots since now the circulation is again thermally and haline driven. With it, the dissipation and kinetic energy also follow.

The evolution of the potential energy E_p is relatively simple. At the beginning ($t = 4.7$), it attains a minimum during the transient state as the vertical density gradients in the south approach zero. From there on, it rises steadily until the collapse of the northern cell at $t \approx 18$. Looking at the fluxes, we can conclude that the initial minimum is caused by a rise in heating Q_T, while the increasing buoyancy forcing and internal mixing Φ_i approximately cancel each other out. This rise in heating is caused by the surface water being cooler than in the ON state (figure 5(*d*), at approximately $t = 4.7$). From there on – because there is no cross-equatorial flow– a net cooling occurs, so the overall basin mass and E_p increase. The slightly lower Φ_z and Φ_i play a lesser role in this rise. The internal mixing Φ_i is lowered as diffusion is not as large when the basin contains two overturning cells. At approximately $t = 18$, this build-up reservoir of potential energy is released in the collapse. Via the risen Φ_i, part of this reservoir is converted into internal energy (Batchelor 2000), while the increased Φ_z converts another part into kinetic energy, and a final part is lost to the atmosphere via surface heating $Q_T > 0$. We note that the increase in potential energy of the buoyancy-driven flow near its separatrix has also been found for more complex AMOC models (Lohmann & Lucarini 2024).

A more detailed view of the energy transfer is obtained by the split of the potential energy into available and background potential energy. The background potential energy has approximately the same evolution as the potential energy, meaning that indeed most of the E_p variation is caused by changes in the basin's mass. This is turn is caused by changes in the energy flux from the surface Φ_a, which includes the atmospheric heating. And for the diffusive flux Φ_d, we see similar behaviour as for the internal mixing Φ_i: both peak during the forcing stage and are lowered during the two-cell stage as diffusion is

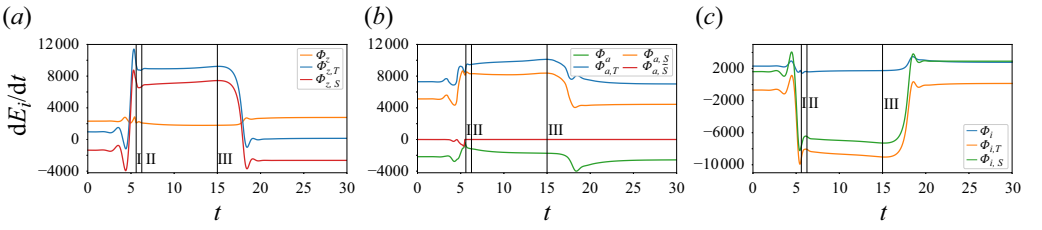


Figure 10. Fluxes (*a*) Φ_z, (*b*) Φ_a and (*c*) Φ_i split into a salinity and temperature contributions, with the former split again into stochastic and deterministic components. States I, II and III as in figure 7 are indicated.

less prevalent. The build-up of E_b is released during the collapse via the atmosphere and diffusive mixing $\Phi_a + \Phi_d$ as available potential energy. The available potential energy E_a shows a peak when E_b has a minimum at $t = 4.7$ as the vertical density gradients in the south are small, i.e. there is relatively dense water near the surface. This build-up of E_a is released when the second overturning cell is formed and state I is reached. During the two-cell stage, E_a slowly decreases as the density distribution becomes more symmetric and the isopycnals become sharper. At the collapse and the emergence of the pole-to-pole circulation, we see dense water moving across the equator, and E_a increases.

In summary, the main energetics during the equilibria are an energy input from the atmosphere as available potential energy via surface flux Φ_a and via the same flux removal of background potential energy to the atmosphere. Then via diffusive fluxes Φ_i and Φ_d, available potential energy is again removed as either internal energy or background potential energy. A last part is removed as buoyancy forcing Φ_z, which then finally sustains the actual circulation and provides kinetic energy. This last one is then again dissipated away as internal energy. Now the forcing puts this system in an imbalance as there is now additional input from the surface via Φ_a and heating Q_T that cannot be removed fast enough via either diffusive processes or buoyancy fluxes. Eventually, right before the collapse, an unstable balance is reached where the surface input has decreased to a level that can be upheld again by diffusion and buoyancy fluxes. However, a small perturbation to this two-cell configuration increases diffusion, which in turn releases the built-up potential energy, which is converted into kinetic energy and eventually causes the collapse.

The effects of both tracers on the energetics of the collapse can be determined by splitting the fluxes Φ_z, Φ_i and Φ_a into contributions due to salinity S and temperature T, where Φ_a also has deterministic and stochastic salinity components $\Phi_{a,S}$ and $\Phi_{a,\tilde{S}}$. We denote them as

$$\Phi_z = \Phi_{z,T} - \Phi_{z,S}, \tag{4.11}$$

$$\Phi_a = \Phi_{a,S} - \Phi_{a,T} + \Phi_{a,\tilde{S}}, \tag{4.12}$$

$$\Phi_i = \Phi_{i,T} - \Phi_{i,S}, \tag{4.13}$$

and their dynamics is shown in figure 10. During the pole-to-pole circulations, we have $\Phi_{z,T} > 0$ and $\Phi_{z,S} < 0$, confirming that these circulations are driven by both salinity and temperature. Salinity aids the upwelling near one pole, while temperature aids the downwelling near the other. During the two-cell configuration both cells are only thermally driven, and indeed the salinity component $\Phi_{z,S}$ has switched sign, opposing the creation of kinetic energy. For the components of Φ_i, we observe similar behaviour. The diffusion of salt (heat) aids the transfer of internal energy to potential energy when $\Phi_{i,S} > 0$ ($\Phi_{i,T} < 0$), and vice versa. So during the pole-to-pole circulations, both components help, while for two cells, only heat diffusion aids. During the two-cell stage, the surface is (spatially

averaged) much saltier than the bottom as less salt is transported downward by the cells from the equator. Therefore, vertical diffusion of salt would reduce the potential energy of the system, hence less energy would be available for the driving buoyancy flux. On the other hand, the surface is much cooler, so vertical diffusion of heat would still aid the potential energy. For the pole-to-pole case, both salt and heat diffusion help as apparently enough salt is transported downwards while the surface remains on average slightly cooler than the bottom. Now for both $\Phi_{z,T}$ and $\Phi_{i,T}$, the only time they work against the conversion to either E_k or E_p is during the initial strengthening ($t \approx 4.3$). At that moment, there is increased upwelling in the south due to the added freshwater flux, while at the same time the upwelling is inhibited by the cooling of the surface there. Regarding Φ_a, we see that during all stages, background energy is added via salinity and removed via heat. This follows from the fact that the freshwater surface forcings always add density (salinity) to the light water away from the downwelling regions, while the surface heating always adds density (cooling) to the heavy water in the downwelling region. The effect is especially pronounced for the two overturning cells. Finally, the stochastic component $\Phi_{a,\tilde{S}}$ only removes a little background energy by salinifying the southern downwelling region. It is striking that the total energy perturbed by the stochastic component $|\int \Phi_{a,\tilde{S}}\,\mathrm{d}t| \approx 575$ is only minor compared to the other energy fluxes.

5. Ratios of tipping probabilities

Our bifurcation parameter β describes the asymmetry of the freshwater forcing, hence a different instanton, i.e. most likely transition, and optimal forcing have to be found for every value of β. While we do not necessarily expect qualitative changes in physical mechanisms mediating the transition, its likelihood will be strongly affected by β. It is therefore instructive to compare the transition probabilities for different values of the bifurcation parameter.

The leading-order term of the ratios of probabilities in the small noise limit can be determined using the large deviation principle (3.2). Consider two trajectories $\boldsymbol{\phi}_A(x, z, t)$ and $\boldsymbol{\phi}_B(x, z, t)$ obeying the same model (3.1) with Freidlin–Wentzell actions S_A and S_B, respectively. Their relative likelihood is given by

$$\frac{\mathbb{P}\left(\boldsymbol{\phi}_B\right)}{\mathbb{P}\left(\boldsymbol{\phi}_A\right)} \asymp \exp\left[(S_A - S_B)/\varepsilon\right]. \tag{5.1}$$

A derivation can be found in the appendix of Soons *et al.* (2024). Now the actions as well as the ratios of transition probabilities for various values of bifurcation parameter β are shown in figure 11 for a range of noise amplitudes ε. The actions are computed as

$$S\left[\boldsymbol{\theta}(x, z, t)\right] = \int_0^\tau \left\langle \boldsymbol{\theta}(x, z, t), \mathcal{A}(\boldsymbol{\theta})(x, z, t)\right\rangle_{L^2(\omega)}\,\mathrm{d}t. \tag{5.2}$$

In figure 11, one can see a monotone decrease in the action as the deterministic salinity forcing freshens the north of the basin as β increases. This can be expected; as more freshwater is added to the southern surface, the stochastic salinity forcing has to increase in order to get the surface water sufficiently dense for downwelling. Note that the downward trend of $S[\boldsymbol{\theta}]$ is gradual; there are no qualitative changes in the instanton trajectory as β varies, and all trajectories show the same dynamics, as described in the previous section.

Regarding the probability ratios, we see that the probability of a tipping increases steeply as we approach the bifurcation point. A similar result has been found by Baars *et al.* (2021)

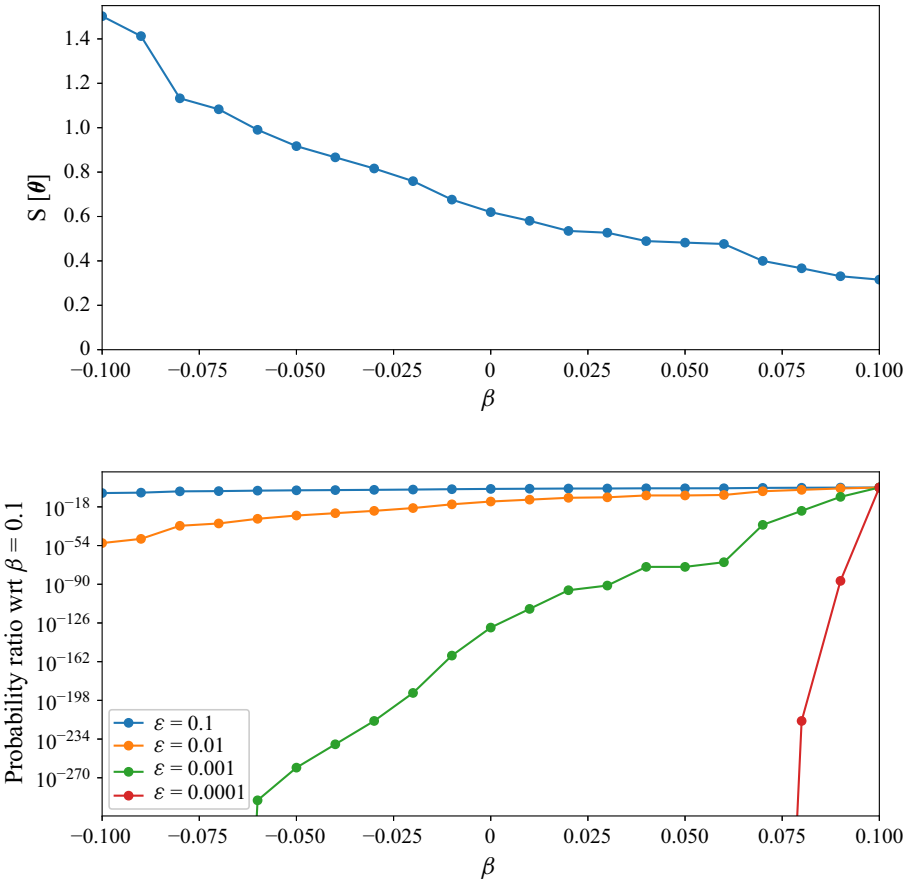


Figure 11. (*a*) The actions $S[\boldsymbol{\theta}(x, z, t)]$ of the instanton trajectory of the AMOC tipping for several $\beta \in [-0.1, 0.1]$. (*b*) The resulting probability ratios of these tippings with respect to the tipping under $\beta = 0.1$ for various noise levels ε.

in a comparable model, where the probabilities were computed numerically using a rare-event algorithm. For example, here a small rise in β from 0.09 to 0.1 increases the tipping probability with factor 1.2, 7.3, 4.9×10^8 and 7.2×10^{86} for the respective noise levels $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$, where only the last noise value can be considered realistic. So a small but permanent shift in the large-scale precipitation or evaporation patterns above the Atlantic can alter the probability of a noise-induced transition of the AMOC greatly.

6. Summary and discussion

The most likely paths of a tipping of the overturning circulation in a two-dimensional (2-D) Boussinesq fluid model for various asymmetric surface salinity forcings were determined. This model is a member of the hierarchy of models of the Atlantic Meridional Overturning Circulation (AMOC) where the zonally averaged Atlantic Ocean flow is represented with constant eddy diffusivities and small stochastic freshwater surface forcing (Quon & Ghil 1992; Thual & McWilliams 1992). Using large deviation theory, the most probable noise-induced transitions from a northern overturning circulation (mimicking an AMOC ON state) to a southern overturning circulation state in the low-noise limit were computed.

Curiously, this transition starts with a strengthening of the northern overturning cell. This phenomenon was also found in previous work (Soons *et al.* 2024) studying the instanton of an AMOC collapse in the box model by Wood *et al.* (2019). For both models, the strengthening puts the system in a less stable transient state by transporting salt quickly to the bottom of the domain. In this way, in the box model, a smaller surface salinity perturbation is needed to tip the AMOC, whereas in this spatially continuous Boussinesq model, the upwelling salinity anomaly is combined with a surface forcing to create one bigger surface perturbation. The net effect is strong enough vertical salinity gradients to induce a second overturning cell in the south. This two-cell configuration is in energy imbalance, where the elevated atmospheric energy input cannot be sufficiently diffused away. This causes a build-up of background potential energy, which is released rapidly as the original northern overturning cell collapses and puts the system into the reversed pole-to-pole circulation. Not only does this method provide these rare transitions, but it also allows us to compare probabilities in the low-noise limit between various forcing scenarios. It shows the steep increase in probability of a noise-induced overturning circulation tipping in the low-noise limit as the bifurcation point is approached.

The advantages of our method are that the problem of finding these rare stochastic events is transformed into a deterministic optimisation problem where no choices regarding surface salinity forcing protocols need to be made. The step forward compared to the instantons on box models (Soons *et al.* 2024) is that spatial patterns are now identified, and the fluid dynamical mechanisms and energetics of these transitions can be more clearly determined.

Obviously, this 2-D model is still a low member in the hierarchy of AMOC models, as the overturning here is driven solely by buoyancy forcing and interior mixing, while for the AMOC, also wind and tidal forcings are important (Kuhlbrodt *et al.* 2007). Moreover, in order to reproduce the required present-day ocean stratification, an open zonal channel in the southern hemisphere is essential (Henning & Vallis 2005; Wolfe & Cessi 2010). As a consequence of these omissions, we have had to adopt highly unrealistic parameter values in order to still have an overturning circulation. In particular, the aspect ratio A and thermal Rayleigh number Ra are off by several orders of magnitude from values that are realistic for the Atlantic Basin and seawater. As a result, our circulation is completely diabatic and non-hydrostatic, as opposed to the actual AMOC. Therefore, we need to stress that this model is only informative for the AMOC, in that it only captures the thermal and salt-advection feedback well, with the latter being the responsible mechanism for the multi-stability of the AMOC (Weijer *et al.* 2019). Consequently, our results provide insight into how these feedbacks function during a noise-induced transition. Since these feedbacks are relevant throughout the model hierarchy (Dijkstra 2024), our results can in turn be relevant for noise-induced collapses in more detailed AMOC models.

Regarding the stochastic salinity surface forcing, a number of assumptions had to be made. In particular, its spatial spectrum was taken to be uniform up to a cut-off frequency, where the cut-off is heuristically based on annual synoptic precipitation patterns. For its temporal structure we chose additive white noise, as this is conceptually the most simple and a common assumption in stochastic AMOC models (Cessi 1994; Timmermann & Lohmann 2000). However, the noise may be coloured and multiplicative, especially under climate change (Boot & Dijkstra 2025). The Freidlin–Wentzell theory can be extended to include this type of noise (Grafke & Vanden-Eijnden 2019). The computed instantons are, of course, a representable trajectory of the noise-induced tipping only if the noise is indeed small. As this is the case in other AMOC models (Castellana *et al.* 2019; Soons *et al.* 2024), and since the noise amplitude relative to the AMOC strength should not depend too much on model choice, we can justify this to be the case for the model used here.

All in all, the merits of this study are the direct computation of a rare overturning tipping in a Boussinesq fluid and the associated probability ratios between various surface forcings in the low-noise limit. As our employed model has severe limitations to represent the AMOC, these results should be viewed as only a stepping stone towards understanding a noise-induced AMOC collapse. We showed that the Freidlin–Wentzell theory can also be applied to spatially continuous models of buoyancy-driven flows where the salt-advection feedback causes a bistable bifurcation structure. A next step would be to apply it to a more representative AMOC model such as a three-dimensional ocean-only model, implemented in e.g. Dijkstra (2007). Furthermore, large deviation theory can be used to investigate the distinct ways in which a combined rate- and noise-induced tipping event can occur (Slyman & Jones 2023), and it would be interesting to examine this in AMOC models of varying complexity.

Acknowledgements. We thank P. Bernuzzi, V. Lucarini and the two anonymous reviewers for their insightful comments.

Funding. J.S. and H.A.D. are funded by the European Research Council through ERC-AdG project TAOC (project 101055096, PI: Dijkstra). T.G. acknowledges support from EPSRC projects EP/T011866/1 and EP/V013319/1.

Declaration of interests. The authors report no conflict of interest.

Data availability statement. The results can be readily reproduced using the described method and model.

Appendix A. The variational derivatives and their adjoints

A.1 Variational derivatives of the deterministic drift

Consider the deterministic drift functions

$$f_1[\boldsymbol{\phi}] = \partial_x \psi[\omega]\, \partial_z \omega - \partial_z \psi[\omega]\, \partial_x \omega + Pr\, \nabla^2 \omega + Pr\, Ra\, (\partial_x T - \partial_x S), \tag{A1}$$

$$f_2[\boldsymbol{\phi}] = \partial_x \psi[\omega]\, \partial_z T - \partial_z \psi[\omega]\, \partial_x T + \nabla^2 T + \frac{h(z)}{\tau_T}\, (T_S(x) - T), \tag{A2}$$

$$f_3[\boldsymbol{\phi}] = \partial_x \psi[\omega]\, \partial_z S - \partial_z \psi[\omega]\, \partial_x S + Le^{-1}\, \nabla^2 S + \frac{h(z)}{\tau_S} S_S(x), \tag{A3}$$

where $\boldsymbol{\phi} : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^3$ with $\boldsymbol{\phi}(x, z, t) = (\omega, T, S)^T(x, z, t)$ and spatial boundary conditions as in § 2. The variational derivatives with respect to ω acting on a function $v_1 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$D_\omega f_1[\boldsymbol{\phi}](v_1) = \partial_x \psi[\omega]\, \partial_z v_1 + \partial_x \mathcal{G}(v_1)\, \partial_z \omega - \partial_z \psi[\omega]\, \partial_x v_1 - \partial_z \mathcal{G}(v_1)\, \partial_x \omega + Pr\, \nabla^2 v_1, \tag{A4}$$

$$D_\omega f_2[\boldsymbol{\phi}](v_1) = \partial_x \mathcal{G}(v_1)\, \partial_z T - \partial_z \mathcal{G}(v_1)\, \partial_x T, \tag{A5}$$

$$D_\omega f_3[\boldsymbol{\phi}](v_1) = \partial_x \mathcal{G}(v_1)\, \partial_z S - \partial_z \mathcal{G}(v_1)\, \partial_x S, \tag{A6}$$

where we have boundary condition $v_1(x, z, t) = 0$ for $(x, z) \in \partial\Omega$. The variational derivatives with respect to T acting on a function $v_2 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$D_T f_1[\boldsymbol{\phi}](v_2) = Pr\, Ra\, \partial_x v_2, \tag{A7}$$

$$D_T f_2[\boldsymbol{\phi}](v_2) = \partial_x \psi[\omega]\, \partial_z v_2 - \partial_z \psi[\omega]\, \partial_x v_2 + \nabla^2 v_2 - \frac{h(z)}{\tau_T} v_2, \tag{A8}$$

$$D_T f_3[\boldsymbol{\phi}](v_2) = 0, \tag{A9}$$

where we have boundary conditions $\partial_x v_2(x, z, t) = 0$ for $x = 0$ or $x = A$, and $\partial_z v_2(x, z, t) = 0$ for $z = 0$ or $z = 1$. The variational derivatives with respect to S acting on a function $v_3 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$D_S f_1[\boldsymbol{\phi}](v_3) = -Pr\, Ra\, \partial_x v_3, \tag{A10}$$

$$D_S f_2[\boldsymbol{\phi}](v_3) = 0, \tag{A11}$$

$$D_S f_3[\boldsymbol{\phi}](v_3) = \partial_x \psi[\omega]\, \partial_z v_3 - \partial_z \psi[\omega]\, \partial_x v_3 + Le^{-1}\, \nabla^2 v_3, \tag{A12}$$

where we have boundary conditions $\partial_x v_3(x, z, t) = 0$ for $x = 0$ or $x = A$, and $\partial_z v_3(x, z, t) = 0$ for $z = 0$ or $z = 1$.

A.2 The adjoints of the variational derivatives

The adjoints of the derivatives of the first component acting on a function $u_1 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$(D_\omega f_1[\boldsymbol{\phi}])^* (u_1) = \partial_z \psi[\omega]\, \partial_x u_1 - \partial_x \psi[\omega] \partial_z u_1 - \mathcal{G}\, (\partial_x u_1\, \partial_z \omega - \partial_z u_1\, \partial_x \omega) + Pr\, \nabla^2 u_1, \tag{A13}$$

$$(D_T f_1[\boldsymbol{\phi}])^* (u_1) = -Pr\, Ra\, \partial_x u_1, \tag{A14}$$

$$(D_S f_1[\boldsymbol{\phi}])^* (u_1) = Pr\, Ra\, \partial_x u_1, \tag{A15}$$

where we have boundary condition $u_1(x, z, t) = 0$ for $(x, z) \in \partial\Omega$. The adjoints of the derivatives of the second component acting on a function $u_2 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$(D_\omega f_2[\boldsymbol{\phi}])^* (u_2) = \mathcal{G}\, (\partial_z u_2\, \partial_x T - \partial_x u_2\, \partial_z T)\,, \tag{A16}$$

$$(D_T f_2[\boldsymbol{\phi}])^* (u_2) = \partial_z \psi[\omega]\, \partial_x u_2 - \partial_x \psi[\omega]\, \partial_z u_2 + \nabla^2 u_2 - \frac{h(z)}{\tau_t} u_2, \tag{A17}$$

$$(D_S f_2[\boldsymbol{\phi}])^* (u_2) = 0, \tag{A18}$$

where we have boundary conditions $\partial_x u_2(x, z, t) = 0$ for $x = 0$ or $x = A$, and $\partial_z u_2(x, z, t) = 0$ for $z = 0$ or $z = 1$. The adjoints of the derivatives of the third component acting on a function $u_3 : \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ are

$$(D_\omega f_3[\boldsymbol{\phi}])^* (u_3) = \mathcal{G}\, (\partial_z u_3\, \partial_x S - \partial_x u_3\, \partial_z S)\,, \tag{A19}$$

$$(D_T f_3[\boldsymbol{\phi}])^* (u_3) = 0, \tag{A20}$$

$$(D_S f_3[\boldsymbol{\phi}])^* (u_3) = \partial_z \psi[\omega]\, \partial_x u_3 - \partial_x \psi[\omega]\, \partial_z u_3 + Le^{-1}\, \nabla^2 u_3, \tag{A21}$$

where we have boundary conditions $\partial_x u_3(x, z, t) = 0$ for $x = 0$ or $x = A$, and $\partial_z u_3(x, z, t) = 0$ for $z = 0$ or $z = 1$.

Appendix B. The augmented Lagrangian method

Using the augmented Lagrangian method (ALM), (3.11) and (3.20) are solved. We define the cost function

$$\begin{aligned} &J[\boldsymbol{\phi}(x,z,t),\theta_S(x,z,t),\boldsymbol{\mu}(x,z,t),\boldsymbol{\gamma},\lambda] \\ &= \int_0^\tau \int_\Omega \frac{(h(z))^2}{2K\tau_S^2}\theta_S \mathcal{U}^*\mathcal{P}\mathcal{U}(\theta_S)\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}t + \int_0^\tau \int_\Omega \Bigg(\mu_\omega\left(\partial_t\omega - f_1(\boldsymbol{\phi})\right) \\ &\quad + \mu_T\left(\partial_t T - f_3(\boldsymbol{\phi})\right) + \mu_S\left(\partial_t S - f_2(\boldsymbol{\phi}) - \frac{(h(z))^2}{K\tau_S^2}\mathcal{U}^*\mathcal{P}\mathcal{U}(\theta_S)\right)\Bigg)\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}t \\ &\quad + \lambda\langle\boldsymbol{\phi}(x,z,\tau),\boldsymbol{\phi}_{OFF}(x,z)\rangle_{L^2} + \langle\boldsymbol{\gamma},\boldsymbol{\phi}(x,z,\tau)-\boldsymbol{\phi}_{OFF}(x,z)\rangle_{L^2}, \end{aligned} \tag{B1}$$

where $\boldsymbol{\mu}:\Omega\times\mathbb{R}_{\geqslant 0}\to\mathbb{R}^3$ with $\boldsymbol{\mu}(x,z,t)=(\mu_\omega,\mu_T,\mu_S)^T(x,z,t)$ is our control variable, and $\lambda\in\mathbb{R}_{\geqslant 0}$ and $\boldsymbol{\gamma}:\Omega\to\mathbb{R}^3$ with $\boldsymbol{\gamma}(x,z)=(\gamma_\omega,\gamma_T,\gamma_S)^T(x,z)$ are penalty parameters to enforce the end conditions. The conjugate momenta θ_ω and θ_T are omitted as there is no direct stochastic forcing onto the vorticity or the temperature, so no cost is assigned to them. Now, minimising the cost function is equivalent to solving the instanton equations, since we have

$$\partial_{\boldsymbol{\phi}}J=0 \iff \partial_t\boldsymbol{\mu}=-\left(\nabla_{\boldsymbol{\phi}}f\right)^*\boldsymbol{\mu}, \tag{B2}$$

$$\partial_{\boldsymbol{\mu}}J=0 \iff \partial_t\boldsymbol{\phi}=f(\boldsymbol{\phi})+(h(z))^2/K\left(0,\ 0,\ \mathcal{U}^*\mathcal{P}\mathcal{U}(\theta_S)\right)^T, \tag{B3}$$

$$\partial_{\theta_S}J=0 \iff \theta_S=\mu_S, \tag{B4}$$

$$\partial_{\boldsymbol{\gamma}}J=0 \iff \boldsymbol{\phi}(x,z,\tau)=\boldsymbol{\phi}_{OFF}(x,z), \tag{B5}$$

$$\partial_{\lambda}J=0 \iff \boldsymbol{\phi}(x,z,\tau)=\boldsymbol{\phi}_{OFF}(x,z), \tag{B6}$$

together with the initial condition $\boldsymbol{\phi}(x,z,0)=\boldsymbol{\phi}_{ON}(x,z)$. The end condition for the control variable is found by

$$\partial_{\boldsymbol{\mu}(x,z,\tau)}J=0 \iff \boldsymbol{\mu}(x,z,\tau)=-\left(2\lambda\left(\boldsymbol{\phi}(x,z,\tau)-\boldsymbol{\phi}_{OFF}\right)+\boldsymbol{\gamma}\right). \tag{B7}$$

In order to minimise this cost function, and hence find the minimising arguments, i.e. the instanton $\tilde{\boldsymbol{\phi}}$ and its associated forcing $\tilde{\theta}_S$, we employ the following protocol. First, we pick the penalty parameters λ and $\boldsymbol{\gamma}$ and then reiterate the following scheme:

(i) forward integration

$$\begin{cases} \partial_t\boldsymbol{\phi}=f(\boldsymbol{\phi})+\dfrac{(h(z))^2}{\tau_S^2K}\left(0,\ 0,\ \mathcal{U}^*\mathcal{P}\mathcal{U}(\theta_S)\right)^T \\ \boldsymbol{\phi}(x,z,0)=\boldsymbol{\phi}_{ON}(x,z) \end{cases} \tag{B8}$$

(ii) backward integration

$$\begin{cases} \partial_t\boldsymbol{\mu}=-\left(\nabla_{\boldsymbol{\phi}}f\right)^*\boldsymbol{\mu} \\ \boldsymbol{\mu}(x,z,\tau)=-\left(2\lambda\left(\boldsymbol{\phi}(x,z,\tau)-\boldsymbol{\phi}_{OFF}(x,z)\right)+\boldsymbol{\gamma}(x,z)\right) \end{cases} \tag{B9}$$

(iii) gradient computation

$$\partial_{\theta_S}J=\theta_S-\mu_S \tag{B10}$$

(iv) updating the conjugate momentum

$$\theta_S^{new}=\theta_S+\alpha\,\partial_{\theta_S}J \tag{B11}$$

with step size $\alpha \in \mathbb{R}_{>0}$ such that $J[\boldsymbol{\phi}^{new}, \theta_S^{new}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \lambda] < J[\boldsymbol{\phi}, \theta_S, \boldsymbol{\mu}, \boldsymbol{\gamma}, \lambda]$, where $\boldsymbol{\phi}^{new}$ is the trajectory forced by θ_S^{new}. This is repeated until it can be concluded that J is minimised for the given λ and $\boldsymbol{\gamma}$, which can be indicated by e.g. $\|\partial_{\theta_S} J\|_{L^2}$ being sufficiently small. These penalty parameters are then updated as:

(i) $\boldsymbol{\gamma}^{new}(x, z) = -\mu(x, z, \tau)$
(ii) $\lambda^{new} = f_\lambda \lambda$ with a constant parameter $f_\lambda > 1$,

then the first scheme is reiterated again. This process is repeated until $\boldsymbol{\phi}(x, z, \tau)$ is sufficiently close to $\boldsymbol{\phi}_{OFF}$.

The instanton is now computed as follows. A trajectory is computed using the ALM, and this trajectory is relaxed as early as possible. This is done as the ALM converges slowly to the needed relaxation after the trajectory has crossed the separatrix. Then the ALM is used again, but now to a point on the relaxed trajectory after the separatrix instead of running it to $\boldsymbol{\phi}_{OFF}$. This newly computed trajectory can now continue unforced until the stable state $\boldsymbol{\phi}_{OFF}$. This yields the instanton that we will use. The advantage here is that we now only optimise the initial forced part of the trajectory using ALM. For the instantons in this work, $\tau = 10$ was sufficient for the duration of the forced part. Any higher τ results in a trajectory with the same dynamical behaviour, but just shifted in time, where the additional time is simply spent in the initial ON state.

Our termination conditions for the above-mentioned iterations are as follows. The minimisation for a given λ and $\boldsymbol{\gamma}$ is terminated when the cost function has decreases by less than 1 %, and the whole algorithm is terminated when $\boldsymbol{\phi}(x, z, \tau)$ is sufficiently close to $\boldsymbol{\phi}_{end}$. This is defined by

$$\max_{(x,z)\in\Omega} \|\phi_i(x, z, \tau) - \phi_{end,i}(x, z)\| \big/ \overline{\phi_{end,i}} < 10^{-3} \quad \text{for all } i \in \{1, 2, 3\}, \tag{B12}$$

where $\overline{\phi_i}$ indicates the spatial mean of ϕ_i.

Appendix C. Energetics of a density-stratified fluid

The non-dimensional volume-integrated kinetic and potential energy are

$$E_k = \frac{1}{2}\int_\Omega u^2 + w^2 \,\mathrm{d}x\,\mathrm{d}z = \frac{1}{2}\int_\Omega \psi\omega \,\mathrm{d}x\,\mathrm{d}z, \tag{C1}$$

$$E_p = Pr\,Ra \int_\Omega \rho z \,\mathrm{d}x\,\mathrm{d}z, \tag{C2}$$

where the kinetic energy has been rewritten to avoid derivatives of the streamfunction so numerical discretisation errors are minimised. We also use the background potential energy E_b, i.e. the minimum potential energy attainable through an adiabatic redistribution of the density. Let $z_*(x, z, t)$ indicate the vertical position in this reference state of the fluid (the background state). The available potential energy E_a is the potential energy released in an adiabatic transition to this background state. So

$$E_b = Pr\,Ra \int_\Omega \rho z_* \,\mathrm{d}x\,\mathrm{d}z, \tag{C3}$$

$$E_a = E_p - E_b = Pr\,Ra \int_\Omega \rho(z - z_*) \,\mathrm{d}x\,\mathrm{d}z. \tag{C4}$$

This reference position $z_*(x, z, t)$ – i.e. the vertical position of the infinitesimal element at (x, z, t) with density $\rho(x, z, t)$ after an adiabatic reshuffle to the background state – is

computed as

$$z_*(x,z,t)=\frac{1}{A}\int_\Omega \mathcal{H}\left(\rho(x',z',t)-\rho(x,z,t)\right)\mathrm{d}x'\,\mathrm{d}z', \tag{C5}$$

with the Heaviside step function

$$\mathcal{H}\left(\rho'-\rho\right)=\begin{cases}0, & \rho'<\rho,\\ \frac{1}{2}, & \rho'=\rho,\\ 1, & \rho'>\rho.\end{cases} \tag{C6}$$

The energy balances are then found by applying the same method as in Winters *et al.* (1995) to our model equations. The kinetic energy obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}E_k=\Phi_z-\mathcal{D}, \tag{C7}$$

$$\Phi_z=Pr\,Ra\int_\Omega \psi\;\partial_x\rho\;\mathrm{d}x\;\mathrm{d}z, \tag{C8}$$

$$\mathcal{D}=Pr\int_\Omega \omega^2\,\mathrm{d}x\;\mathrm{d}z, \tag{C9}$$

where Φ_z denotes the buoyancy flux, and $\mathcal{D}\geqslant 0$ is the dissipation. For the potential energy, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}E_p=\Phi_i-\Phi_z-Q_T, \tag{C10}$$

$$\Phi_i=-Pr\,Ra\left(\frac{1}{Le}\int_0^A\left(S(x,1)-S(x,0)\right)\mathrm{d}x-\int_0^A\left(T(x,1)-T(x,0)\right)\mathrm{d}x\right), \tag{C11}$$

$$\begin{aligned}Q_T&=\frac{1}{\tau_T}\int_\Omega z\,h(z)\,\left(T_S(x)-T\right)\;\mathrm{d}x\;\mathrm{d}z\\ &\approx\frac{\delta}{\tau_T}\int_0^A\left(T_S(x)-T(x,z)\right)\mathrm{d}x,\end{aligned} \tag{C12}$$

where Φ_i is the conversion rate of internal energy to potential energy, and Q_T is the net heating by the surface forcing. Then for the background potential energy we have

$$\frac{\mathrm{d}}{\mathrm{d}t}E_b=\Phi_a+\Phi_d, \tag{C13}$$

$$\begin{aligned}\Phi_a&=Pr\,Ra\int_\Omega z_*\left(\frac{h(z)}{\tau_S}S_S(x)+\mathcal{A}(\boldsymbol{\theta})-\frac{h(z)}{\tau_T}\left(T_S(x)-T\right)\right)\mathrm{d}x\;\mathrm{d}z\\ &\approx Pr\,Ra\left(\frac{\delta}{\tau_S}\int_0^A z_*(x,1)\;S_S(x)\,\mathrm{d}x-\frac{\delta}{\tau_T}\int_0^A z_*(x,1)\,\left(T_S(x)-T(x,1)\right)\mathrm{d}x\right.\\ &\quad\left.+\int_\Omega z_*\mathcal{A}(\boldsymbol{\theta})\,\mathrm{d}x\;\mathrm{d}z\right),\end{aligned} \tag{C14}$$

$$\Phi_d=Pr\,Ra\int_\Omega \partial_T z_*\|\nabla T\|^2-\frac{1}{Le}\,\partial_S z_*\|\nabla S\|^2\,\mathrm{d}x\;\mathrm{d}z, \tag{C15}$$

where Φ_a is the rate of change of E_b due to addition of salinity or heat near the surface, while Φ_d denotes its increase due to diapycnal mixing. Note that $\partial_T z_*>0$ and $\partial_S z_*<0$

as z_* is the stably stratified reference state, hence $\Phi_d \geqslant 0$ means that due to mixing, the background potential energy always increases. For the available potential energy budget, we simply subtract the background budget from the potential energy budget, yielding

$$\frac{\mathrm{d}}{\mathrm{d}t}E_a = \Phi_i - \Phi_z - \Phi_a - \Phi_d - Q_T. \tag{C16}$$

All fluxes and energies are computed directly apart from Φ_d, which we compute as a residual from the background potential energy budget.

References

Adler, R.F., Gu, G., Sapiano, M., Wang, J.-J. & Huffman, G.J. 2017 Global precipitation: means, variations and trends during the satellite era (1979–2014. *Surv. Geophys.* **38** (4), 679–699.

Adler, R.F., Kummerow, C., Bolvin, D., Curtis, S. & Kidd, C. 2003 Status of TRMM monthly estimates of tropical precipitation. In *Cloud Systems, Hurricanes, and the Tropical Rainfall Measuring Mission (TRMM): A Tribute to Dr Joanne Simpson* (ed. W.K. Tao & R. Adler), pp. 223–234. American Meteorological Society (AMS).

Altland, A. & Simons, B.D. 2010 *Condensed Matter Field Theory*. Cambridge University Press.

Armstrong McKay, D.I. *et al.* 2022 Exceeding 1.5 °C global warming could trigger multiple climate tipping points. *Science* **377** (6611), eabn7950.

Baars, S., Castellana, D., Wubs, F.W. & Dijkstra, H.A. 2021 Application of adaptive multilevel splitting to high-dimensional dynamical systems. *J. Comput. Phys.* **424**, 109876.

Batchelor, G.K. 2000 *An Introduction to Fluid Dynamics*. Cambridge University Press.

Ben-Yami, M., Morr, A., Bathiany, S. & Boers, N. 2024 Uncertainties too large to predict tipping times of major Earth system components from historical data. *Sci. Adv.* **10** (31), eadl4841.

Boers, N. 2021 Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation. *Nat. Clim. Change* **11** (8), 680–688.

Boot, A.A. & Dijkstra, H.A. 2025 Observation-based temperature and freshwater noise over the Atlantic Ocean. *Earth Syst. Dynam.* **16** (1), 115–150.

Börner, R., Deeley, R., Römer, R., Grafke, T., Lucarini, V. & Feudel, U. 2024 Saddle avoidance of noise-induced transitions in multiscale systems. *Phys. Rev. Res.* **6** (4), L042053.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. 2018 Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature* **556** (7700), 191–196.

Castellana, D., Baars, S., Wubs, F.W. & Dijkstra, H.A. 2019 Transition probabilities of noise-induced transitions of the Atlantic Ocean circulation. *Sci. Rep.* **9** (1), 20284.

Cessi, P. 1994 A simple box model of stochastically forced thermohaline flow. *J. Phys. Oceanogr.* **24** (9), 1911–1920.

Cessi, P. & Young, W.R. 1992 Multiple equilibria in two-dimensional thermohaline circulation. *J. Fluid Mech.* **241**, 291–309.

Cimatoribus, A.A., Drijfhout, S.S. & Dijkstra, H.A. 2014 Meridional overturning circulation: stability and ocean feedbacks in a box model. *Clim. Dyn.* **42** (1–2), 311–328.

Dijkstra, H.A. 2007 Characterization of the multiple equilibria regime in a global ocean model. *Tellus A: Dyn. Meteorol. Oceanogr.* **59** (5), 695–705.

Dijkstra, H.A. 2024 The role of conceptual models in climate research. *Physica D: Nonlinear Phenom.* **457**, 133984.

Dijkstra, H.A. & Molemaker, J.M. 1997 Symmetry breaking and overturning oscillations in thermohaline-driven flows. *J. Fluid Mech.* **331**, 169–198.

Dijkstra, H.A., Molemaker, M.J., Van Der Ploeg, A. & Botta, E.F.F. 1995 An efficient code to compute non-parallel steady flows and their linear stability. *Comput. Fluids* **24** (4), 415–434.

Ditlevsen, P. & Ditlevsen, S. 2023 Warning of a forthcoming collapse of the Atlantic Meridional Overturning Circulation. *Nat. Commun.* **14** (1), 1–12.

Ditlevsen, P.D. & Johnsen, S.J. 2010 Tipping points: early warning and wishful thinking. *Geophys. Res. Lett.* **37** (19), L19703.

Frajka-Williams, E. *et al.* 2019 Atlantic Meridional Overturning Circulation: observed transport and variability. *Frontiers Mar. Sci.* **6**, 260.

Freidlin, M.I. & Wentzell, A.D. 1998 *Random Perturbations*. Springer.

Friedman, A.R., Reverdin, G., Khodri, M. & Gastineau, G. 2017 A new record of Atlantic sea surface salinity from 1896 to 2013 reveals the signatures of climate variability and long-term trends. *Geophys. Res. Lett.* **44** (4), 1866–1876.

GIORGINI, L.T., LIM, S.H., MOON, W. & WETTLAUFER, J.S. 2020 Precursors to rare events in stochastic resonance. *Europhys. Lett.* **129** (4), 40003.

GRAFKE, T., GRAUER, R. & SCHÄFER, T. 2015 The instanton method and its numerical implementation in fluid mechanics. *J. Phys. A: Math. Theor.* **48** (33), 333001.

GRAFKE, T., SCHÄFER, T. & VANDEN-EIJNDEN, E. 2017 Long term effects of small random perturbations on dynamical systems: theoretical and computational tools. In *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science* (ed. R. Melnik, R. Makarov, & J. Belair), pp. 17–55. Springer.

GRAFKE, T. & VANDEN-EIJNDEN, E. 2019 Numerical computation of rare events via large deviation theory. *Chaos: Interdisciplinary J. Nonlinear Sci.* **29** (6), 063118.

HAIRER, M. 2014 A theory of regularity structures. *Invent. Math.* **198** (2), 269–504.

HAWKINS, E., SMITH, R.S., ALLISON, L.C., GREGORY, J.M., WOOLLINGS, T.J., POHLMANN, H. & DE CUEVAS, B. 2011 Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport. *Geophys. Res. Lett.* **38** (10), L10605.

HENNING, C.C. & VALLIS, G.K. 2005 The effects of mesoscale eddies on the stratification and transport of an ocean with a circumpolar channel. *J. Phys. Oceanogr.* **35** (5), 880–896.

HESTENES, M.R. 1969 Multiplier and gradient methods. *J. Optim. Theor. Applics.* **4** (5), 303–320.

HOGG, A.M.C., DIJKSTRA, H.A. & SAENZ, J.A. 2013 The energetics of a collapsing meridional overturning circulation. *J. Phys. Oceanogr.* **43** (7), 1512–1524.

HUGHES, G.O., HOGG, A.M.C. & GRIFFITHS, R.W. 2009 Available potential energy and irreversible mixing in the meridional overturning circulation. *J. Phys. Oceanogr.* **39** (12), 3130–3146.

JACKSON, L.C. & WOOD, R.A. 2018 Hysteresis and resilience of the AMOC in an eddy-permitting GCM. *Geophys. Res. Lett.* **45** (16), 8547–8556.

JOHNSON, H.L., CESSI, P., MARSHALL, D.P., SCHLOESSER, F. & SPALL, M.A. 2019 Recent contributions of theory to our understanding of the Atlantic Meridional Overturning Circulation. *J. Geophys. Res.: Oceans* **124** (8), 5376–5399.

KUHLBRODT, T., GRIESEL, A., MONTOYA, M., LEVERMANN, A., HOFMANN, M. & RAHMSTORF, S. 2007 On the driving processes of the Atlantic Meridional Overturning Circulation. *Rev. Geophys.* **45** (2), RG2001.

LOHMANN, J. & LUCARINI, V. 2024 Melancholia states of the Atlantic Meridional Overturning Circulation. *Phys. Rev. Fluids* **9** (12), 123801.

LUCARINI, V. & BÓDAI, T. 2020 Global stability properties of the climate: melancholia states, invariant measures, and phase transitions. *Nonlinearity* **33** (9), R59–R92.

LUCARINI, V., SERDUKOVA, L. & MARGAZOGLOU, G. 2022 Lévy noise versus Gaussian-noise-induced transitions in the Ghil–Sellers energy balance model. *Nonlinear Process. Geophys.* **29** (2), 183–205.

LUCARINI, V. & STONE, P.H. 2005 Thermohaline circulation stability: a box model study. Part I: Uncoupled model. *J. Clim.* **18** (4), 501–513.

MAROTZKE, J. 2000 Abrupt climate change and thermohaline circulation: mechanisms and predictability. *Proc. Natl Acad. Sci. USA* **97** (4), 1347–1350.

QUON, C. & GHIL, M. 1992 Multiple equilibria in thermosolutal convection due to salt-flux boundary conditions. *J. Fluid Mech.* **245**, 449–483.

RAHMSTORF, S. 1996 On the freshwater forcing and transport of the Atlantic thermohaline circulation. *Clim. Dyn.* **12** (12), 799–811.

RAHMSTORF, S. *et al.* 2005 Thermohaline circulation hysteresis: a model intercomparison. *Geophys. Res. Lett.* **32** (23), L23605.

ROOTH, C. 1982 Hydrology and ocean circulation. *Prog. Oceanogr.* **11** (2), 131–149.

SCHORLEPP, T., GRAFKE, T., MAY, S. & GRAUER, R. 2022 Spontaneous symmetry breaking for extreme vorticity and strain in the three-dimensional Navier–Stokes equations. *Phil. Trans. R. Soc. Lond. A* **380** (2226), 20210051.

SLYMAN, K. & JONES, C.K. 2023 Rate and noise-induced tipping working in concert. *Chaos: Interdisciplinary J. Nonlinear Sci.* **33** (1), 013119.

SOONS, J., GRAFKE, T. & DIJKSTRA, H.A. 2024 Optimal transition paths for AMOC collapse and recovery in a stochastic box model. *J. Phys. Oceanogr.* **54** (12), 2537–2552.

STOMMEL, H. 1961 Thermohaline convection with two stable regimes of flow. *Tellus* **13** (2), 224–230.

THUAL, O. & MCWILLIAMS, J.C. 1992 The catastrophe structure of thermohaline convection in a two-dimensional fluid model and a comparison with low-order box models. *Geophys. Astrophys. Fluid Dyn.* **64** (1–4), 67–95.

TIMMERMANN, A. & LOHMANN, G. 2000 Noise-induced transitions in a simplified model of the thermohaline circulation. *J. Phys. Oceanogr.* **30** (8), 1891–1900.

TURNER, J.S. 1973 *Buoyancy Effects in Fluids*. Cambridge University Press.
VANDERBORGHT, E., VAN, W., RENÉ, M. & DIJKSTRA, H.A. 2024 Feedback processes causing an AMOC collapse in the community Earth system model. arXiv preprint arXiv: 2410.03236.
VELDMAN, A.E.P. & RINZEMA, K. 1992 Playing with nonuniform grids. *J. Engng Math.* **26** (1), 119–130.
WEIJER, W., CHENG, W., DRIJFHOUT, S.S., FEDOROV, A.V., HU, A., JACKSON, L.C., LIU, W., MCDONAGH, E.L., MECKING, J.V. & ZHANG, J. 2019 Stability of the Atlantic Meridional Overturning Circulation: a review and synthesis. *J. Geophys. Res.: Oceans* **124** (8), 5336–5375.
VAN WESTEN, R.M. & DIJKSTRA, H.A. 2023 Asymmetry of AMOC hysteresis in a state-of-the-art global climate model. *Geophys. Res. Lett.* **50** (22), e2023GL106088.
VAN WESTEN, R.M., JACQUES-DUMAS, V., BOOT, A.A. & DIJKSTRA, H.A. 2024*a* The role of sea ice insulation effects on the probability of AMOC transitions. *J. Climate* **37** (23), 6269–6284.
VAN WESTEN, R.M., KLIPHUIS, M. & DIJKSTRA, H.A. 2024*b* Physics-based early warning signal shows that AMOC is on tipping course. *Sci. Adv.* **10** (6), eadk1189.
WINTERS, K.B., LOMBARD, P.N., RILEY, J.J. & D'ASARO, E.A. 1995 Available potential energy and mixing in density-stratified fluids. *J. Fluid Mech.* **289**, 115–128.
WOILLEZ, E. & BOUCHET, F. 2020 Instantons for the destabilization of the inner solar system. *Phys. Rev. Lett.* **125** (2), 021101.
WOLFE, C.L. & CESSI, P. 2010 What sets the strength of the middepth stratification and overturning circulation in eddying ocean models? *J. Phys. Oceanogr.* **40** (7), 1520–1538.
WOOD, R.A., RODRÍGUEZ, J.M., SMITH, R.S., JACKSON, L.C. & HAWKINS, E. 2019 Observable, low-order dynamical controls on thresholds of the Atlantic Meridional Overturning Circulation. *Clim. Dyn.* **53**, 6815–6834.