# ~~Collapse~~ Characterisation ~~of the~~ Atlantic Meridional Overturning ~~described by~~ hysteresis using Langevin dynamics

Jelle van den Berk[1], Sybren Drijfhout[1,2], and Wilco Hazeleger[2]

[1]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands
[2]Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

**Correspondence:** J. van den Berk (jelle.van.den.berk@knmi.nl)

**Abstract.** ~~Using a machine learning technique,~~ Steady-state collapse trajectories of the Atlantic Meridional Overturning Circulation (AMOC) under freshwater forcing from climate models of intermediate complexity are fitted to a simple model based on the Langevin equation. A total of six parameters are sufficient to quantitatively describe the collapses seen in these simulations~~under a freshwater forcing~~. Reversing the freshwater forcing results in asymmetric behaviour that is less well captured and ~~would~~ appears to require a more complicated model. The Langevin model allows for comparison between models that display an AMOC collapse. Variation between the climate models studied here is mainly due to the strength of the stable AMOC and the strength of the response to a freshwater forcing.

## 1 Introduction

The Atlantic Meridional Overturning Circulation (AMOC) is an important circulation in the Atlantic ocean. It is also an important part of the climate system overall due to the heat it transports from the South Atlantic to the North Atlantic (Ganachaud and Wunsch, 2000; Vellinga and Wood, 2002). The AMOC therefore has a substantial influence on the (western) European climate and a weakening of the AMOC might cause changes in the European climate and ~~weather, see Weijer et al. (2019) for a review~~weather. The AMOC has also been identified as one of Earth's 'tipping elements' where a rapid change on markedly faster times scales could take place in the (near) future (Lenton et al., 2008). The ~~Atlantic Meridional Overturning Circulation (AMOC)~~ AMOC is partly buoyancy driven by the deep water formations in the North Atlantic subpolar gyre which produces the North Atlantic Deep Water (NADW) (e.g. Rahmstorf (2000)). The AMOC might be ~~bistable~~ bi-stable in nature which means it admits an 'off' state, with little or no transport from north to south, as a counterpart to its current 'on' state (Broecker et al., 1985).

Palaeoclimate records of the last glacial ~~maximum and early holocene~~ period show a rapid switching of temperature, which might be associated with the presence/absence of a vigorous AMOC as exists today (Dansgaard et al., 1993). The possibility of a bistable AMOC being the cause of these rapid changes has been noted (Broecker et al., 1990). With the current climate warming rapidly, the stability of the AMOC is of particular interest (Collins et al., 2013) and climate modelling projections indicate the AMOC strength will decrease under an increase of $CO_2$. Recent measurements show the AMOC has decreased in

strength (Smeed et al., 2018). An understanding of the possibly bistable nature of the AMOC is therefore relevant to understand the consequences of climate change. See Weijer et al. (2019) for a review on AMOC bistability.

The Langevin equation has been posited before as ~~appropriate~~ suitable to capture the essential dynamics of an AMOC collapse (Ditlevsen and Johnsen, 2010; Berglund and Gentz, 2002). It has also been used elsewhere as the basis for describing the dynamics of climate sub-systems (Livina et al., 2010) and the AMOC in particular (Kleinen et al., 2003; Held and Kleinen, 2004). A fourth order potential function is used in Ditlevsen and Johnsen (2010); Berglund and Gentz (2002) because it is the minimum required for having three distinct solutions (double wells). This potential function has two parameters which ~~are presumed to be functions~~ are presumed to be functions of the freshwater forcing. Variation in the freshwater forcing is assumed to directly drive changes in AMOC strength by changing the potential function in the Langevin equation. Although the hysteresis loops of the AMOC include both a collapse and a resurgence point, we will only attempt to model the collapse from the stable 'on' branch to the stable 'off' branch.

Though the Langevin equation has played a role in the conceptual picture of bistability and tipping points in the climate, ~~but~~ it has not been used to actually fit the parameters to a (simulated) AMOC collapse. Here, we attempt to construct a simple model based on the Langevin equation and fit its dynamics to salt-advection driven collapse trajectories of the AMOC seen in climate models (Rahmstorf et al., 2005). The result is a set of parameters that quantitatively describe the AMOC collapse process. This derived model defines a low-dimensional manifold that captures the essential AMOC collapse characteristics. To the extent that the low-dimensional model is successful in capturing the more complex model this method could also be used to predict the parameter range where in a model a collapse would occur. At present, however, it is intended to provide a characterisation of the collapse that will allow comparison between climate models.

Section 2 sketches the theoretical background of the Langevin equation and of the salt-advection mechanism. In Section 3 we fit the proposed Langevin model to the AMOC collapse trajectories seen in a set of climate models of intermediate complexity (EMICs) taken from Rahmstorf et al. (2005). We end with a discussion and conclusions in Section 4.

## 2 The Langevin model

An increase in surface air temperatures, or an ~~increase in P-E (precipitation - evaporation) of freshwater surface flux~~increased surface freshwater flux by changes in precipitation minus evaporation, will decrease the buoyancy in the shallow layer of the deep water formation regions in the North Atlantic subpolar gyre. The deep water formation is reduced, and the southward meridional flow reduced. In principle, this mechanism can reduce the AMOC to zero gradually if fully buoyancy-driven. A salt-advection feedback mechanism that leads to a bimodal AMOC was proposed by Stommel (1961). In this mechanism, ~~the deep southward flow couples to the surface return flowvia upwelling or other mechanisms in the Southern Ocean, affecting the salinity in~~ salinity anomalies in the North Atlantic are amplified by the overturning flow, which in turn controls the North Atlantic ~~subpolar gyre. A reduction in salinitydecreases buoyancy, and this positive feedback accelerates the process of AMOC weakening and a collapse results on relatively short timescales.~~ salinity. Positive anomalies are strengthened and

negative anomalies weakened; this results in a positive feedback between the salinity anomalies and the overturning. Bistability, consisting of a strong and a weak AMOC state, and possible abrupt transitions result.

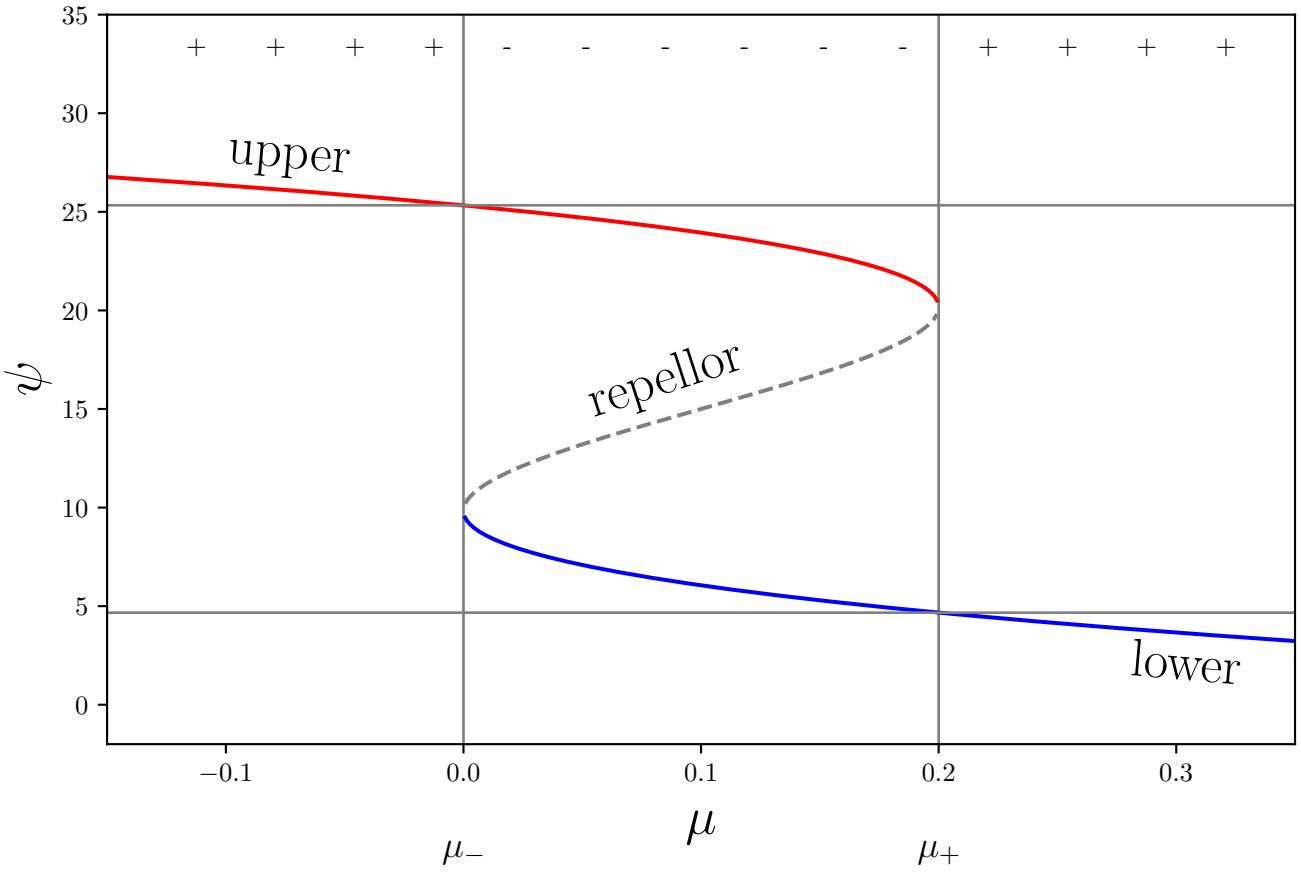


**Figure 1.** Example ~~trajectory~~ bifurcation diagram of the AMOC ($\Psi$) in response to a control variable $\mu$. The red branch is the on-state (upper), blue the off-state (lower). The ~~trajectory~~ upper branch deforms when closer to the bifurcation points which are connected though the repellor ~~that forms the trench of the distribution~~ (dashed line). The two bifurcations points are indicated as $\mu_+$ (collapse point) and $\mu_-$ (~~re-invigouration~~ resurgence point). Top $\pm$ symbols indicate unimodal ($+$) or bimodal ($-$) regime.

Fig. 1 shows a conceptual picture of the two stable AMOC (index) states. The AMOC is a ~~zero-dimensional variable arrived at~~ scalar variable obtained by integrating the overturning transport and selecting its maximum value (typically located in the subtropical North Atlantic). In red, the upper branch is drawn up to the collapse point where a bifurcation occurs. The real AMOC in the current climate moves along this branch from the left, to the right, towards its (assumed) collapse point. The branch in blue is the counterpart of the upper branch and represents the off state of the AMOC and ends in another bifurcation point to the left where the AMOC jumps back to full strength. The dashed line (repellor) separates the two basins of attraction associated with the two stable branches (attractors). At the bifurcation point one of the two basins of attraction ~~ceases to~~

exist vanishes and a qualitative change takes place in the potential function (the number of solutions for a given value of the freshwater forcing $\mu$ goes from 2 to 3 to 1).

Below we will derive a model based on the Langevin equation that captures the essential dynamics of a bimodal AMOC under a freshwater forcing $\mu$.

### 2.1 Multiple stable AMOC states

The conceptual picture of the AMOC being a zero-dimensional variable that is driven by stochastic forces trapped in a potential is similar to that of a particle's motion described by Langevin dynamics (Lemons et al., 1908). The Langevin equation (Gardiner, 2004; Ditlevsen and Johnsen, 2010),

$$\dot{x} = -\partial_x U_\mu(x) + \sigma\eta \tag{1}$$

describes the position of a noise-driven particle ($x$) trapped in a potential function $U$. The stochastic term is a white noise process ($\eta$) scaled with an intensity parameter $\sigma$. At first we will ignore the stochastic nature of the AMOC collapse process and focus on the deterministic behaviour.

The (deterministic) bistability double well potential seen in Fig. 1 has been studied mainly in a qualitative way(within catastrophe theory)extensively studied and applied, also in a quantitative way. But to our knowledge it has not been quantitatively applied to AMOC hysteresis using the Langevin equation in complex numerical climate models before.

studied mainly qualitatively in connection with the Langevin equation. AMOC bistability has, however, been studied quantitatively in e.g. Boulton et al. (2014) using transient runs. In Poston and Stewart (1978) an extensive treatment is given why, in addition to a scaling and shifting, only two parameters are sufficient to describe the bistability. These two parameters From a fourth order polynomial for $U$, the third and fourth order coefficients can be eliminated. The two remaining coefficients in the polynomial describe the critical behaviour, not just locally near the critical points, but the entire trajectory under a suitable transformation. (The behaviour at small scale is fundamentally tied to the global behaviour.) A direct consequence is that only partial information, in the form of a piece of the trajectory, should suffice to describe the entire trajectory (the full hysteresis loop).

The potential function takes the form (Gardiner, 2004; Ditlevsen and Johnsen, 2010)

$$-U(x) = -\frac{1}{4}x^4 + \frac{\beta}{2}x^2 + \alpha x. \tag{2}$$

The two parameter parameters $\alpha, \beta$ are functions of the freshwater forcing $\mu$. The AMOC state variable $\Psi$ requires an affine transformation (Cobb, 1980),

$$\alpha = \alpha(\mu)$$
$$\beta = \beta(\mu)$$
$$x = (\Psi - \lambda)/\nu.$$

To fit the model trajectories we need to find expressions for $\alpha$ and $\beta$, and suitable values for the transformation parameters $\lambda$ and $\nu$. The In the literature $\alpha$ is referred to as the normal factor, and $\beta$ the splitting factor (Poston and Stewart, 1978). In the

bifurcation diagram the value of $\nu$ is ~~roughly~~ approximately the distance in $\Psi$ between the bifurcation point on the top branch to the bifurcation point on the lower branch. ~~The~~ Similarly, the value of $\lambda$ is ~~roughly~~ approximately the $\Psi$ value between the bifurcation points at $\mu_\pm$. The transformation uses $\lambda$ to shift the trajectory and $\nu$ to scale it. Below we describe the potential visually and state additional constraints that follow from the demand that the freshwater forcing is the only variable that determines the dynamical behaviour.

### 2.2 Potential description

In Fig. 2 an overview of the qualitatively different forms of potential are shown ($-U(x)$, right panels) together with their derivative functions ($-\partial_x U$, left panels). Dots indicate the location of critical points and are related to the number of wells in the potential. The top panels show the typical bimodal form ($I$) with two stable states and one unstable one in the middle. Below these are the three possible unimodal states ($E$). These occur for forcing values to the left of $\mu_-$ and to the right of $\mu_+$. The panels $B_1$ and $B_2$ are the submanifolds that separates the unimodal regime from the bimodal regime. These two meet in the cusp point $P$, as shown in the bottom panels. See Poston and Stewart (1978) for further details.

In Fig. 3 the stability ~~landscape~~ diagram is shown where the areas indicated are those with qualitatively different behaviour seen in Fig. 2. See also Poston and Stewart (1978) for similar ~~diagrammes~~diagrams. The cusp point $P$ is the singular point where no proper solution can exist because only the trivial solution (all parameters are valued 0) is allowed here (both ~~collapse~~ bifurcation points $\mu_\pm$ and AMOC strength are at zero). The two parameters are $\alpha$ and $\beta$ and are the two coefficients in the potential function. Their values change because of their dependency on the forcing value ($\mu$).

Our aim is to arrive at a description that matches a ~~track~~ series of $\mu$ values across the stability ~~landscape~~diagram. The two parameters $\alpha, \beta$ are independent but can be ~~parametrised~~ parameterised by other variables that map them to observations. ~~In the literature $\alpha$ is referred to as the normal factor, and $\beta$ the splitting factor (Poston and Stewart, 1978). If parametrised~~ If parameterised by a single variable~~the track~~, the values of ($\alpha$, $\beta$) across the stability surface ~~is~~ are a one-dimensional subset, as suggested by the AMOC index. On one side of the cusp point, along the splitting axis ($\beta$), only a unimodal regime exists, while on the other side two regimes exist with the modes at relative distances apart.

### 2.3 Constraints

With a varying $\alpha$ there exist an interval between two critical points ($\alpha_\pm$) in between which the distribution is bimodal and unimodal outside that interval. Because the AMOC trajectory is 1-dimensional and $\mu$ is also 1-dimensional, there must be a relation between $\alpha$ and $\beta$ that reduces dimensionality from two to one dimensions. When passing through the critical point $\alpha_+$, the number of potential wells goes from two to one. Similarly, moving through $\alpha_-$ changes the number of wells from one to two (for given $\mu_\pm$). The two critical points of $\partial_x U$, $\mu_\pm$, can be found analytically for $\mu_\pm$ real and being degenerate solutions. It can be shown (Birkhoff and Mac Lane, 1970, p. 106) that the discriminant $D = 27\alpha^2 - 4\beta^3 = 0$ (i.e. real solutions) needs to be solved for $\alpha$ to obtain the two critical solutions that relate $\alpha$ and $\beta$. It is at these solutions that the number of critical points changes at forcing values $\mu_\pm$. (When $D < 0$ there are three distinct real solutions which corresponds to the bimodal regime, when $D > 0$ there is only one distinct real solution, which corresponds to the unimodal regime.) When any two of the roots are

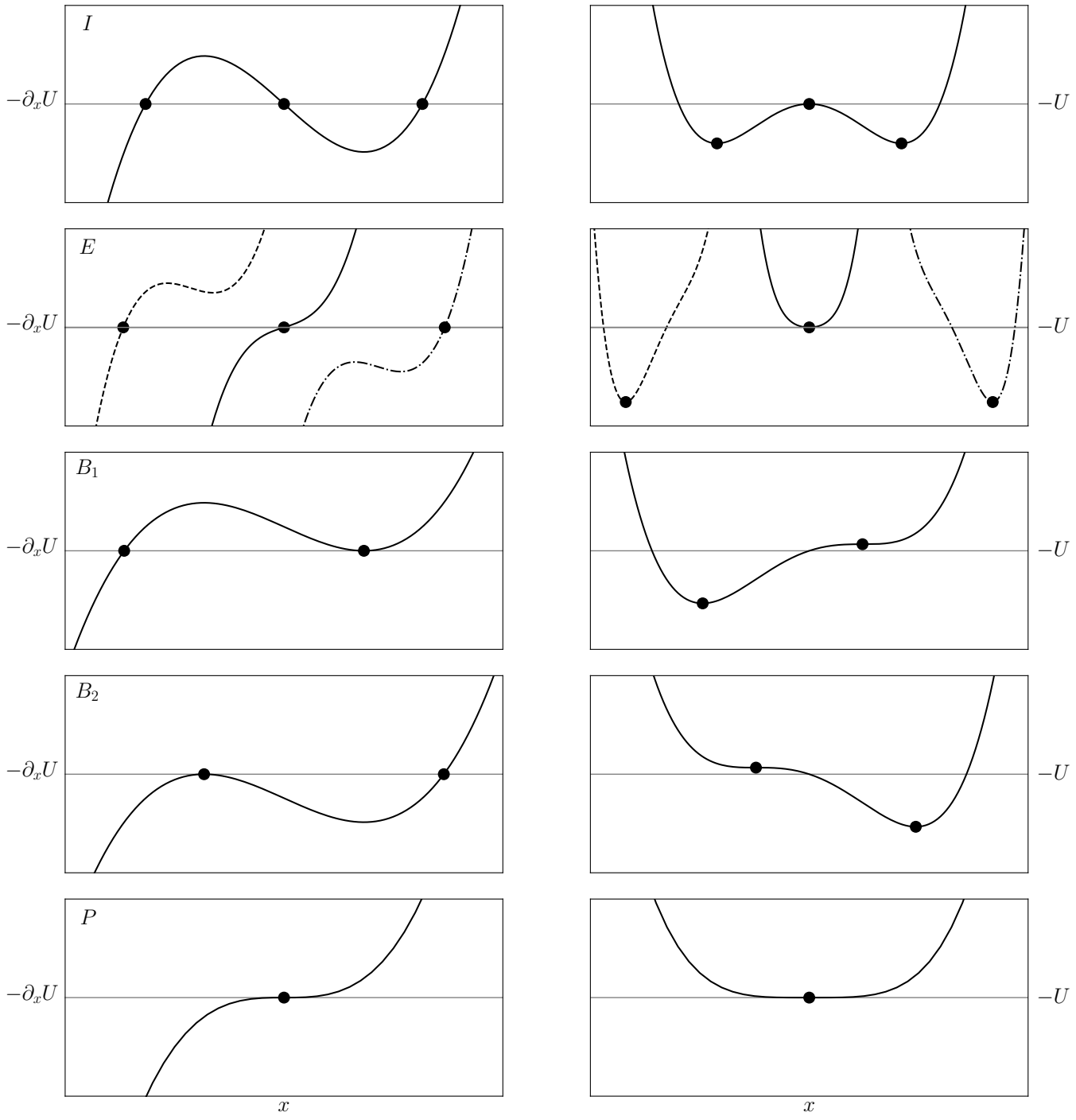


**Figure 2.** Sample potentials (right) and their derivatives (left) for (top to bottom) the three possible varieties of bimodal state ($I$), three types of unimodal state ($E$), the two pathological cases where $D = 0$ ($B_1$ and $B_2$), and the cusp catastrophe point ($P$). Dots indicate the critical points. (Scaling is not uniform between panels. Note the choice of negative sign of the potential $U$.)

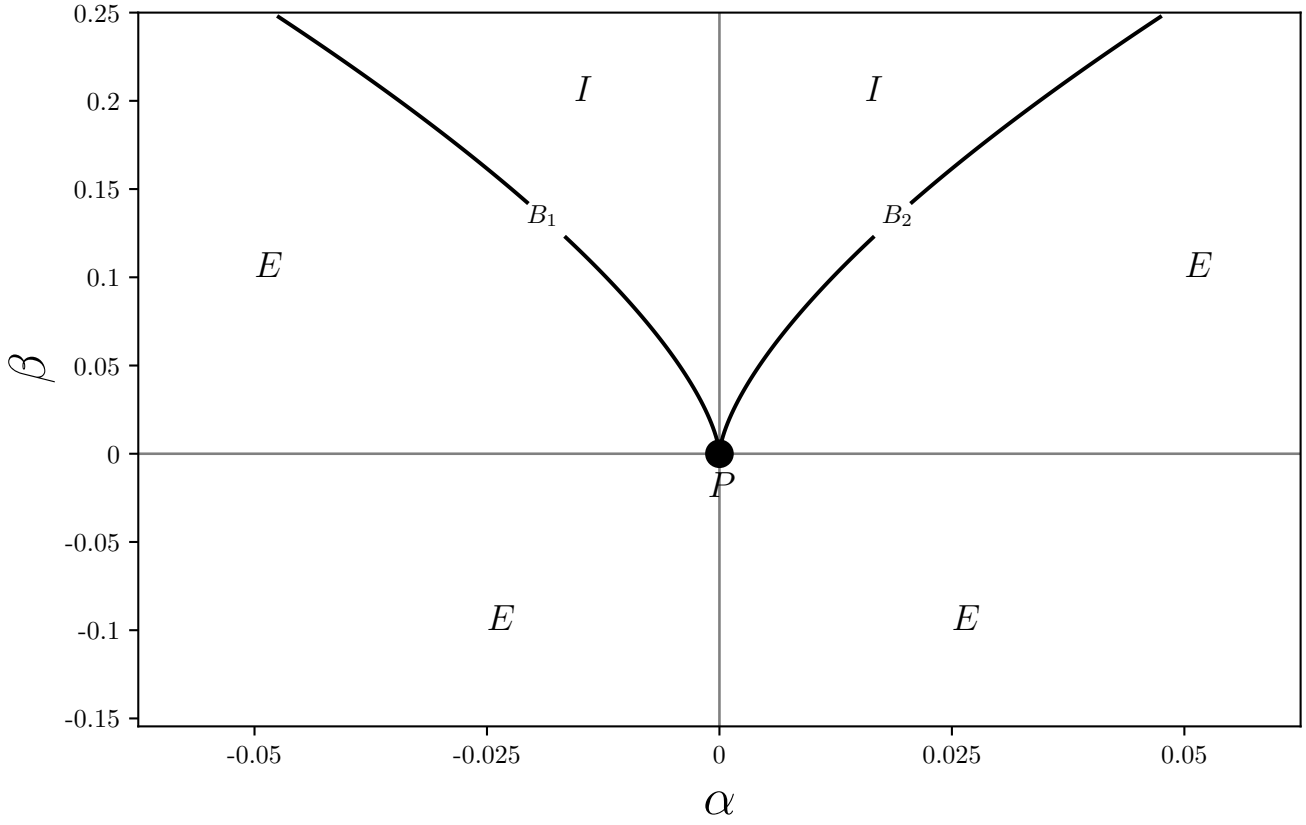


**Figure 3.** Discriminant determining the stability and number of critical points. The splitting factor $\beta$ and normal factor $\alpha$ describe the stability ~~landscape~~diagram. The bimodal regime ($I$) is separated from the unimodal regime ($E$) by two lines ($B_{1,2}$) which meet in the point $P$.

the same, the number of extrema goes from 3 to 2 (or 1 if all are the same) and the solutions become degenerate (this occurs at $B_{1,2}$ in Fig. 3).

Solving for $\alpha$ gives two solutions that are the critical values as functions of $\beta$,

$$\alpha_\pm = \pm \frac{2\sqrt{3}}{9} (\beta)^{3/2} \quad \text{or} \quad \alpha_\pm = \mp \frac{2\sqrt{3}}{9} (\beta)^{3/2},$$

with $\beta \geq 0$ for real solutions. The points $\alpha_\pm$ correspond to where the lines $B_{1,2}$ in Fig. 3 are passed when moving across the stability surface.

For $\alpha_+ < 0$ $-U(1) < 0$. This corresponds with the AMOC undergoing a collapse at $\mu_+$ from an on state to an off state, and the correct choice of sign is

$$\alpha_\pm = \mp \frac{2\sqrt{3}}{9} (\beta_\pm)^{3/2}, \tag{3}$$

with $\alpha_\pm$ and $\beta_\pm$ the values corresponding to $\mu_\pm$. ~~The track across the stability landscape is~~ Changing $\mu$ in the bifurcation diagram corresponds to moving from curve $B_2$ to curve $B_1$ and Eq. 3 relates the two stability parameters $\alpha$ and $\beta$ at the two critical forcing values $\mu_\pm$.

### 2.3.1 Linear functions $\alpha, \beta$

The value of $\beta$ does not need to be fixed (to ~~$\alpha_\pm$~~$\alpha_\pm$ and in general there is a corresponding $\beta_\pm$ at the respective critical points~~) and a varying $\beta$ corresponds to a slanted track across the stability landscape in Fig. 3.~~.We assume linear functions for $\alpha$ and $\beta$,

$$\alpha(\mu) = \alpha_0 + \mu\,\delta\alpha$$
$$\beta(\mu) = \beta_0 + \mu\,\delta\beta,$$

reducing the dependency to these four parameters. Linear functions are the ~~most simple~~ simplest non-trivial dependencies, while adding non-linear parameters introduces further unknowns, making this the most parsimonious parametrisation that captures the first order behaviour. Also, intuitively we can understand the pair (~~$\alpha$, $\beta$~~$\delta\alpha$, $\delta\beta$) as the angle under which the system moves to the bifurcation point ($B_{1,2}$) in Fig. 3), which locally only requires the values of $\alpha$ and $\beta$ up to first order. ~~Poston and Stewart (1978, p. 59) also remark that the system's local behaviour is essentially the same between critical points, which means a linear expansion should suffice for fitting the upper branch.~~ From this parametrisation we can determine the offset $\alpha_0$ and rate $\delta\alpha$ in terms of $\beta_0$ and $\delta\beta$,

$$\alpha_+ = \alpha_0 + \mu_+\delta\alpha = -\frac{2\sqrt{3}}{9}\left(\beta_+\right)^{3/2} \quad \text{and}$$
$$\alpha_- = \alpha_0 + \mu_-\delta\alpha = +\frac{2\sqrt{3}}{9}\left(\beta_-\right)^{3/2}$$

gives

$$\delta\alpha = -\frac{2\sqrt{3}}{9}\,\frac{\left(\beta_+\right)^{3/2} + \left(\beta_-\right)^{3/2}}{\mu_+ - \mu_-} \tag{4}$$
$$\alpha_0 = \alpha(\mu = 0) = \frac{\sqrt{3}}{9}\left[-\left(\beta_+\right)^{3/2} + \left(\beta_-\right)^{3/2}\right] - \frac{1}{2}\delta\alpha\,(\mu_+ + \mu_-)\,. \tag{5}$$

This constrains the values of $\alpha$, leaving only $\beta$ as a free variable, which is then ~~parametrised~~ parameterised by $\beta_0$ and $\delta\beta$. Note that only solutions with $\beta_\pm > 0$ are valid. Also, values for $\beta_0$ and $\delta\beta_0$ that result in ~~a track that crosses~~ crossing $B_2$ in another point besides $\beta_-$ are unsuitable. (The curves $B_{1,2}$ are each intersected by a straight line in at most two points, and we require intersection at a single point only.)

## 2.4 Stochastic interpretation

With the deterministic framework in place, the stochastic nature can be reintroduced.~~We obtain a distribution needed to fit the parameters of the potential function .~~ The potential function can be replaced by a distribution which is the stationary distribution in the asymptotic limit (i.e. the long term behaviour of repeated sampling of the hysteresis loop). ~~As shown by Cobb (1978), this distribution belongs to the exponential family.~~

~~For a polynomial function as the potential, the distribution obtained is from the exponential family, which is a generalisation of the exponential distribution (Balakrishnan and Nevzorov, 2004) where any function can determine the exponent value. The potential, we had already supposed to be described by~~ The potential (a fourth-order polynomial~~in the previous section,~~) gives the probability distribution (Cobb, 1978)

$$P(x,\alpha,\beta) = Ce^{-2/\sigma^2\, U(x)} = Ce^{2/\sigma^2(-1/4x^4+\beta/2x^2+\alpha x)}. \tag{6}$$

~~Note that~~ The factor $C = C(\alpha,\beta)$ ~~and~~ does not have a (known) analytical expression for the general case, but can be computed numerically (and therefore used as a likelihood function in the next section). This can be done accurately with an adaptive quadrature method (Piessens et al., 2012), though it suffers from numerical limitations. The value of $\sigma$ is a measure of intrinsic variation in the AMOC. Note that $\sigma$ is a measure of additive noise (because we assume that $\sigma$ is not dependent on $\mu$) and other choices, such as multiplicative noise, can be made (Das and Kantz, 2020). See Gardiner (2004) for a derivation of this distribution using the Fokker-Planck equation, from which also the Langevin equation can be derived. ~~Note~~ Also, note that $\sigma \to \sigma/\nu$ because ~~to~~ of the scaling with $\nu$ we introduced in Section 2.1.

~~A sample collapse trajectory~~ An example bifurcation diagram with corresponding distribution is shown in Fig. 4. The ~~grey~~ purple lines indicate the (fixed) positions of the bifurcation points. The dashed grey line marks the positions of the unstable solution (repellor) in between the two attractor branches which separates the two basins of attraction. Note that the bifurcation points are extremal in the sense that no bimodality can exists beyond them. With the trajectories being noisy and driven along the attractor, there is (always) some probability of a 'noise-induced' transition. The state shifts from one basin of attraction to the other, crossing the repellor, and the AMOC rapidly moves from one attractor to the other. For this reason, the ~~sampled~~ bimodality region might be larger than is apparent from a particular sample AMOC trajectory. A larger noise level (as seen in AMOC observations Smeed et al. (2018)) would increase the likelihood of a collapse before the AMOC reaches the bifurcation point.

The distributions in Figs 5 show that qualitatively distinct behaviour occurs when $\alpha$ or $\beta$ are varied. For both parameters, a change from a unimodal to a bimodal distribution can be seen. ~~These changes correspond directly to~~ Each distinct shape of the distribution can be identified with one of the potential functions in Fig. 2. In principle, a change in ~~on~~ only one of the two structural parameters ($\alpha$ and $\beta$) can move the distribution between unimodal and bimodal forms.

~~The normalisation of the family of distributions depends on the values of the parameters. Therefore, we are required to calculate the normalisations factors for each parameter set. This cannot be done analytically, but can be done accurately with an adaptive quadrature method (Piessens et al., 2012), though it suffers from numerical limitations.~~

We are now in a position to apply the above to collapse trajectories from climate models.

## 3 ~~Simulated~~ AMOC collapse parameter estimation

We describe how to find an optimal solution under the framework ~~arrived at~~ described in the previous section. Using a Bayesian optimisation procedure, estimated values of $\beta_0$ and $\delta\beta$ can be found, together with the scaling parameters $\nu$ and $\lambda$. We will also estimate the values for $\mu_\pm$, resulting in a six parameter list that describes (the upper branch) of an AMOC collapse.

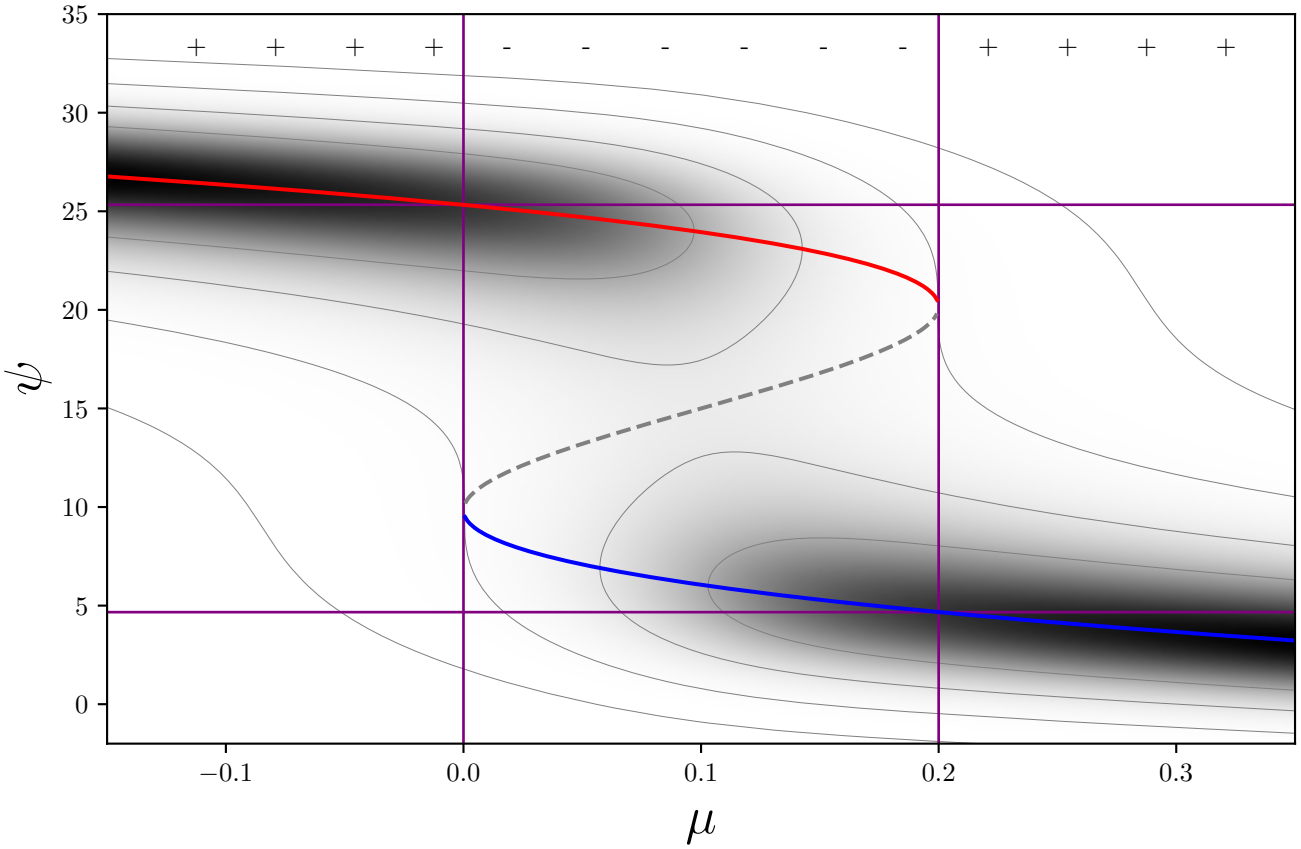


**Figure 4.** Example trajectory with corresponding distribution. ~~Parametrised~~ Parameterised by $\lambda = 15$, $\nu = 20$, $\sigma = 0.12\nu$, $\mu_+ = 0.2$, $\mu_- = 0$, $\beta_0 = 0.2$, $\delta\beta = 0$; $\alpha$ under constraints in Eqs 4 and 5. The distribution of one of the attractor branches (red: on state, blue: off state) deforms when closer to the bifurcation points which are connected though the repellor that forms the trench of the distribution (dashed line). Top $\pm$ symbols indicate unimodal $(+)$ or bimodal $(-)$ regime based on the discriminant value $(D)$.The value of $\sigma$ is relatively large and is chosen for clarity. The purple lines indicate the (fixed) positions of the bifurcation points.

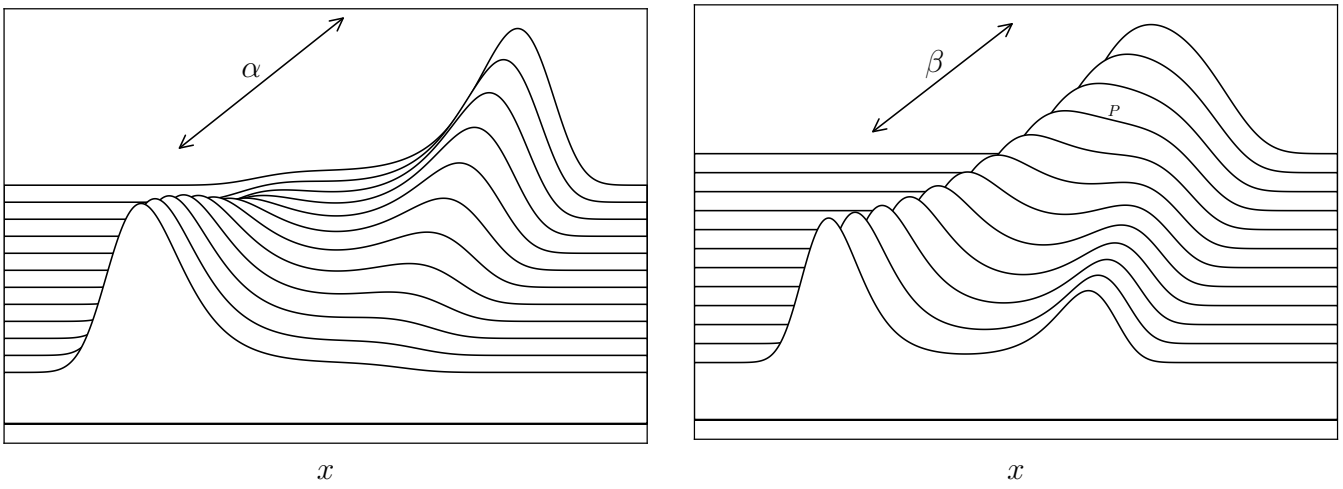


**Figure 5.** Left: Distributions from the exponential family (Eq. 6) where the parameter $\beta$ is kept at a fixed value and $\alpha$ is varied. The distribution transforms from unimodal (back), to bimodal (middle), to a different unimodal distribution (front). The ~~unimodal states have distinct singular maxima. The~~ bimodal states have a ~~dominant~~ larger and a ~~weak~~ smaller mode, depending on the position within the bimodal regime~~; in the middle and inversion from weak to dominant takes place~~. The relative strength between ~~dominant and weak~~ modes depends on $\sigma$. Right: Distributions from the exponential family (Eq. 6) where the parameter $\alpha$ is kept at a fixed value and $\beta$ is varied. A broad unimodal state (at the back) splits into distinct bimodal states (to the front). In the middle a critical point exists, called the cusp (point $P$ in Fig. 3) where the split occurs.

The parameters $\beta_0$ and $\delta\beta$ are independent ~~to~~ of each other, but need to cross the curves $B_{1,2}$ in Fig. 3) to match the corresponding values for $\mu_\pm$. This constraint is satisfied by the resulting values for $\alpha_0$ and $\delta\alpha$. (This can still lead to solution candidates that are not suitable for the collapse trajectories and are eliminated in the sampling process below.) The scaling parameters are not fully independent because $\lambda < \nu$ (the offset cannot exceed the scaling) and knowing where the upper and lower branches are located already gives a rough estimate.

### 3.1 Parameter estimation

Cobb (1978) was able to fit the distribution in Eq. 6 using optimisation techniques (which were numerically unstable and not very flexible). Though the estimates for the scaling parameters $\lambda$ and $\nu$ can be quite good with this approach, estimating the trajectory parameters $\beta_0$ and $\delta\beta$ requires a more flexible method. Knowing which distribution to use, we can ~~fit its parameters under some measure of goodness-of-fit by machine learning. Specifically, we can use Bayes' rule to maximise the likelihood of a parametrised Langevin model $L_\sigma(\nu, \lambda, \beta_0, \delta\beta, \mu_\pm)$ given a trajectory~~ estimate the posterior probability distribution of the parameters given the data $\Psi(\mu)$~~(Bolstad, 2010)~~,

$$P(\nu, \lambda, \beta_0, \delta\beta, \mu_\pm \mid \Psi),.$$

~~to arrive at the (linearised) posterior distributions of $\nu$, $\lambda$, $\beta_0$, $\delta\beta$, $\mu_\pm$ under the observed $\Psi(\mu)$.~~ Bayes' rule tells us the probability of a given observation $\Psi$ given the probability of the parameters (marginal on the left, or posterior) is proportional to the probability given the parameters (marginal on the right, or prior) and the full distribution (likelihood),

$$P(\nu, \lambda, \beta_0, \delta\beta, \mu_\pm \mid \Psi) \propto P(\Psi \mid \nu, \lambda, \beta_0, \delta\beta, \mu_\pm) \cdot P(\nu, \lambda, \beta_0, \delta\beta, \mu_\pm).$$

~~(The right-hand side of Bayes's rule is called the Bayes factor and can be normalised by the probability of the observed trajectory $P(\Psi)$ (called the evidence) to obtain an equality.)~~

Sampling different values from the parameters' prior distributions will give corresponding values for the posterior distributions. ~~These resultant posterior distributions can, in turn, be used as prior distributions, yielding a chain of sampled parameter vectors.~~ A Bayesian sampler chooses successive values that tend towards greater likelihood of the model, given the observed trajectory, and will converge towards an optimal fit. ~~This is roughly~~ Conceptually, this what an MCMC (Markov chain Monte-Carlo) optimiser does (Bolstad, 2010). A widely used sampling algorithm is the Metropolis algorithm ~~(Hastings, 1970)~~ (Hastings, 1970; Bernardo and Smith, 2009), which we also use here.[1]

The ~~models can be fit with uninformative priors, but the~~ sampling process is time consuming because the evaluation of the potential (to calculate $P(\Psi \mid \nu, \lambda, \beta_0, \delta\beta, \mu_\pm)$) requires numerical integration (using a quadrature method), which is costly to evaluate (the exponential family of distributions cannot, in general, be evaluated analytically).

### 3.1.1 Prior distributions

The prior distribution of a parameter represents all the information known about that parameter before confrontation with the observed values (Bolstad, 2010). With $\nu$ and $\lambda$ ~~introduced earlier, the state variable $x$ undergoes an affine transformation and normalises the polynomial. These~~ transform the AMOC state variable ($\Psi$) with a shift ($\lambda$) and a scaling ($\nu$). The shift $\lambda$ cannot exceed the normalisation $\nu$, giving an upper bound on $\lambda$. Also, we note the lower limit of the lower branch, meaning $\lambda$ must be larger than this minimum value. Similarly, the scaling $\nu$ cannot be larger than the maximum value of the AMOC on the upper branch. We expect the linear parametrisation of $\alpha$ and $\beta$ introduced in the previous section to be $\mathcal{O}(1)$.

We are nonetheless still faced with infinite support on the coefficients of the expansion of the parameters ($\beta_0$, $\delta\beta$). We therefore transform $\beta_0$ and $\delta\beta$, with support $(-\infty, \infty)$, using the arctan function to map to $(-\pi/2, \pi/2)$. After such a transformation, we can sample from the flat prior distribution on that interval with most of the probability mass on 'reasonable' values (i.e. $\mathcal{O}(1)$). ~~For $\beta_0$ and $\delta\beta$ this transformation will be usedand $\alpha$ will follow from the constraints in Eqs 5 and 4. An overview~~

[1] This algorithm has been implemented in many software packages.

~~of priors is~~The following prior distributions are used:

$$\nu = U(\min(\text{AMOC}), \max(\text{AMOC}))$$
$$\lambda = U(\min(\text{AMOC}), \nu)$$
$$\mu_+ = U(\mu_{\text{S+}}, \mu_{\text{UP}})$$
$$\mu_- = U(\mu_{\text{DN}}, \mu_{\text{S-}})$$
$$\tan(\beta_0) = U(-\pi/2, \pi/2)$$
$$\tan(\delta\beta) = U(-\pi/2, \pi/2),$$

with $\min(\text{AMOC})$ and $\max(\text{AMOC})$ is the minimum/maximum values in an observed collapse trajectory. $U$ is the uniform distribution on indicated intervals. The intervals values of the collapse points $\mu_\pm$ we stipulate as being bounded by where the trajectories merge ($\mu_{\text{UP}}$ and $\mu_{\text{DN}}$) and the inner values ($\mu_{\text{S-}}$ and $\mu_{\text{S+}}$) observed in the trajectories (within which bimodality is demanded, see Fig. 6). [2]

### 3.2 Fitting EMIC collapse trajectories

An AMOC collapse was induced in ~~six~~ models of intermediate complexity in Rahmstorf et al. (2005) by applying a freshwater forcing to the North-Atlantic subtropical gyre region that reduced the salinity in the subpolar gyre to its north. ~~In~~ Six of these models have a 3-D ocean components; in Fig. 6 the trajectories of those collapses are reproduced (right column, the freshwater flux has been labelled $\mu$ here) together with their numerical derivatives (left ~~column).~~ columns in the panels). In Tab. 1 the models are listed. The forcing values of $\mu$ are known and the same for each climate model. Each model was run to equilibrium for each forcing value; there is therefore no explicit time dependence in the hysteresis loops shown. Both the AMOC strength and the forcing value have units Sv (=$10^6\,\text{m}\,\text{s}^{-1}$). Note that the bifurcation points ($\mu_\pm$) must lie within the range where the trajectories appear bimodal.

The trajectories are taken from the numerical Earth System Models (EMICs) Rahmstorf et al. (2005, Fig. 2, bottom panel) by extracting the data points directly from the graphic in the electronic publication[3]. The numerical derivatives show where the AMOC changes quickest as a response to the change in freshwater forcing. Each model has two peaks where the changes are largest, one for each change between stable branches. These peaks are located at the repellor in between the two attractors (the stable branches). At the repellor only unstable solutions exist and the AMOC is driven to a ~~more~~ stable solution, away from these states.

[2] To ~~ensure the signs~~ exclude parameter values that lead to intersections of ~~the discriminants are the same~~ $B_{1,2}$ more than once, we ~~added an additional constraint as a sharp-peaked likelihood on the discriminant that follows from the proposed solutions and that follows from~~ artificially decrease the ~~observed trajectory~~likelihood of these values. ~~This helps~~ The discriminant of the ~~fitting process by explicitly excluding invalid solutions with incorrect modality~~polynomial at each forcing value indicates when this is needed.

[3] The figure we used is a vector graphic and the data set can be retrieved from it by inverting the plot matrices used to map the original data to the values in the graph. We can replicate the data in this manner.

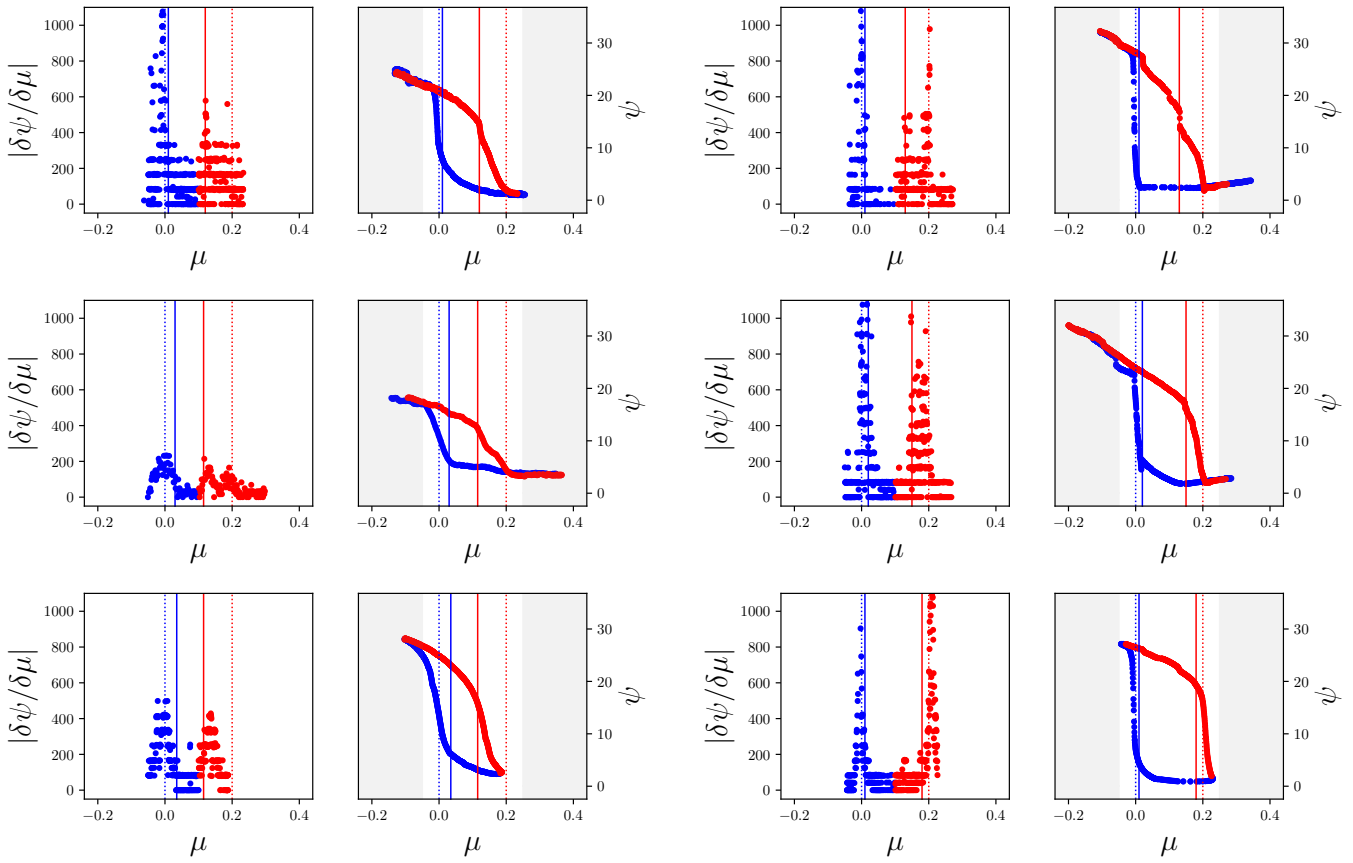


**Figure 6.** Absolute values of numerical derivatives (left) from the trajectories of AMOC strength as function of freshwater forcing to the right (taken from Rahmstorf et al. (2005, Fig. 2, bottom panel), reproduced with permission from the publisher: American Geophysical Union). In red the upper branch, blue the lower branch. ~~Top to bottom~~Left column: ~~UVic~~Bremen, ~~MOM iso~~ECBilt-CLIO, C-GOLDSTEIN; right column: MOM hor, ~~C-GOLDSTEIN, Bremen~~MOM iso, ~~ECBilt-CLIO~~UVic.Vertical solid lines mark $\mu = 0$ (blue) and $\mu = 0.2$ (red); vertical dashed lines mark the chosen boundary values for $\mu_{\pm}$. All values ~~in~~ have units ~~of~~ Sv.

| model | #data points | ocean component | atmosphere component | reference |
|---|---|---|---|---|
| Bremen | 2461 | large-scale geostrophic | energy balance | Prange et al. (2003) |
| ECBilt-CLIO | 243 | 3D primitive equations | quasi-geostrophic | Goosse et al. (2001) |
| C-GOLDSTEIN | 849 | 3D simplified | energy-moisture balance | Edwards and Marsh (2005) |
| MOM hor | 1233 | 3D primitive equations (MOM) | simple energy balance | Rahmstorf and Willebrand (1995) |
| MOM iso | 1442 | as above, with isopycnal mixing | simple energy balance | |
| UVic | 464 | 3D primitive equations (MOM) | energy-moisture balance | Weaver et al. (2001) |

**Table 1.** Overview of models used. Each data point is independent from the others because each is the result of a quasi steady state run. The number of data points used is given. The summary of the type of model component and references are taken from Rahmstorf et al. (2005).

| model | $\sigma$ | $\mu_-$ | $\mu_+$ | present da |
|---|---|---|---|---|
| ~~UVic~~ Bremen | ~~0.260~~0.181 | [~~-0.020~~-0.018, 0.010] | [~~0.188, 0.225~~0.120, 0.220] | ( ~~0.080, 25.00~~0.07( |
| ~~MOM iso~~ ECBilt-CLIO | ~~0.216~~0.176 | [~~-0.010, 0.020~~-0.044, 0.030] | [~~0.150~~0.115, 0.210] | (~~0.050, 22.8)-MOM hor 0.526-0.010, 0.0100.1~~ |
| C-GOLDSTEIN | 0.122 | [-0.100, 0.035] | [0.115, 0.190] | (-0.100, 29.( |
| ~~Bremen~~ MOM hor | ~~0.181~~0.526 | [~~-0.018,~~ -0.010, 0.010] | [~~0.120, 0.220~~0.130, 0.200] | ( ~~0.070, 18.8)~~0.11 |
| ~~ECBilt-CLIO~~ MOM iso | ~~0.176~~0.216 | [~~-0.044, 0.030~~-0.010, 0.020] | [~~0.115,~~ 0.150, 0.210] | ( ~~-0.110, 18.2)+~~0.0 |
| UVic | 0.260 | [-0.020, 0.010] | [0.188, 0.225] | ( 0.080, 25.( |

**Table 2.** Overview of models, the estimated standard deviation with the upper branch fitted to a linear function (note that the original trajectories had already been smoothed), the ranges of $\mu_\pm$, the location of present day in the models, and whether the present day value is in the unimodal regime (+) or not (-). All values ~~in~~ have units ~~of~~ Sv.

If no other mechanisms apart from the salt advection are important we expect the bifurcation points to lie beyond the observed transition points because a noise-induced transition pushes the AMOC into the off-state sooner. (Note that although the collapse points are expected to lie before these peaks, low levels of noise will obscure this effect.) The dashed lines indicate the regions where we will search for the optimum values of $\mu_\pm$. These differ from the fixed 0 and 0.2 values chosen by (Rahmstorf et al., 2005), who also shifted the trajectories to align on these values. ~~The dragged-out descent to the lower branch (e.g. the model Bremen) indicates that the salt-advection mechanism does not necessarily result in an abrupt collapse in the trajectory.~~

Before fitting, the upper and lower branches were extended to the left and right to fill the space of $-0.2 < \mu < 0.4$. A linear fit was use to produce additional values of the corresponding branches (at the same density of those points already present). All models then occupy the same freshwater forcing space. This is desirable because not all models have a lower branch that is fully sampled (specifically, UVic). The lower branch was extended with a negative rate of increase if the lower branch was moving upwards with increasing $\mu$ (MOM hor and MOM iso).

~~Because we want~~ Our main goal is to model the ~~collapse of the AMOC, only the behaviour of the upper branch and~~ transition from on-branch to off-branch, that is, the upper right half of the hysteresis curve, and not so much the dynamics that govern

the lower branch. Also, because we assume that other dynamics govern the lower branch and our simple model has to be extended to account for those dynamics. We ignore the data on the lower branch ~~beyond $\mu_+$ is of interest. We expect these~~ before the collapse point so the fits would not be influenced by these points. We expect the remaining points of the trajectory ~~do~~ to be dominated by the salt-advection mechanism. ~~The hysteresis loop is exemplified in the trajectory of C-GOLDSTEIN. This model shows a smooth path with a relatively rapid transition region–but not a collapse as such by appearance.~~

We start by identifying some characteristic points in the trajectories in Tab. 2. The $\sigma$ (variance of the process) of the models is not given in Rahmstorf et al. (2005) or elsewhere in the literature, but was estimated as the deviation with a fitted function to ~~(part of)~~ the left most the top branch~~(note~~. (Note that smoothing was already applied in Rahmstorf et al. (2005), lowering the variance of the trajectories~~)~~. Because we want to fit the collapse trajectory as given, we use the variance as evident from the data.) In principle, $\sigma$ could also be estimated as a parameter in the Bayesian optimisation, but that would unnecessarily enlarge the search space. Note that the 'off-state' of the AMOC in these models is not 0, but $\sim$ 2Sv of AMOC strength. If the salt-advection mechanism were the only operative effect, we expect this value to be $\leq 0$. If a reverse advection cell emerges as the lower hysteresis branch, this value is negative.

~~The model trajectories are apparently driven by the forcing value, which means we should be able to explain the behaviour of the hysteresis using this variable as the only driver applied to the Langevin model.~~

In Fig. 7 fitted distributions are shown ~~for linearly parametrised sample paths through the stability space~~ (also tabulated in Tab. ~~7). The $\beta$ parameter changes linearly and $\alpha$ follows from the constraints in Eqs. 4 and 5. Blue and red lines indicate the prior bounds for $\mu_-$ and $\mu_+$, respectively.~~ 3). The parameter values of these distributions are the means of the posterior distributions. The dashed grey line marks the positions of the unstable solution (repellor) in between the two attractor branches which separates the two basins of attraction.

| ~~model~~ | ~~$\nu$~~Bremen | ~~$\lambda$~~ECBilt-CLIO | ~~$\beta_0$~~C-GOLDSTEIN | ~~$\delta\beta$~~MOM hor | ~~$\mu_-$~~M |
|---|---|---|---|---|---|
| ~~UVic~~ $\nu$ | ~~23.32~~ $20.8 \pm 1.8 \cdot 10^{-1}$ | ~~10.39~~ $14.1 \pm 2.5 \cdot 10^{-1}$ | ~~0.3523~~ $24.0 \pm 2.4 \cdot 10^{-1}$ | ~~-1.051~~ $27.9 \pm 2.5 \cdot 10^{-1}$ | ~~-0.004~~ 25. |
| ~~MOM iso~~ $\lambda$ | ~~25.38~~ $8.18 \pm 6.5 \cdot 10^{-3}$ | ~~9.395~~ $8.37 \pm 2.0 \cdot 10^{-2}$ | ~~0.293~~ $10.1 \pm 1.5 \cdot 10^{-2}$ | ~~-1.259~~ $11.4 \pm 4.5 \cdot 10^{-2}$ | ~~0.020~~ 9.4 |
| ~~MOM hor~~ $\beta_0$ | ~~27.93~~ $0.278 \pm 5.2 \cdot 10^{-3}$ | ~~11.36~~ $0.250 \pm 9.8 \cdot 10^{-3}$ | ~~0.2719~~ $0.272 \pm 5.5 \cdot 10^{-3}$ | ~~-1.339~~ $0.272 \pm 4.6 \cdot 10^{-3}$ | ~~0.010~~ 0.29 |
| ~~C-GOLDSTEIN~~ $\delta\beta$ | ~~23.99~~ $-1.34 \pm 2.5 \cdot 10^{-2}$ | ~~10.12~~ $-1.30 \pm 5.0 \cdot 10^{-2}$ | ~~0.2715~~ $-1.46 \pm 3.1 \cdot 10^{-2}$ | ~~-1.458~~ $-1.34 \pm 3.2 \cdot 10^{-2}$ | ~~0.035~~ -1.2 |
| ~~Bremen 20.75~~ $\mu_-$ | ~~8.184~~ $0.010 \pm 9.4 \cdot 10^{-5}$ | ~~0.2778~~ $0.029 \pm 8.9 \cdot 10^{-4}$ | ~~-1.339~~ $0.035 \pm 5.4 \cdot 10^{-5}$ | $0.010 \pm 1.7 \cdot 10^{-4}$ | ~~0.139~~ 0.02 |
| ~~ECBilt-CLIO~~ $\mu_+$ | ~~14.1~~ $0.14 \pm 5.8 \cdot 10^{-4}$ | ~~8.369~~ $0.13 \pm 1.1 \cdot 10^{-3}$ | ~~0.2497~~ $0.13 \pm 5.8 \cdot 10^{-4}$ | ~~-1.303~~ $0.14 \pm 2.3 \cdot 10^{-3}$ | ~~0.029~~ 0.1 |

**Table 3.** ~~Parameters~~ Mean values and standard deviations of parameters corresponding to the fitted functions in Fig 7.

The fits with a linear ~~track~~ series through the $(\alpha, \beta)$ parameter space result in a mismatch between the behaviour seen on lower branches and that on the upper branches. ~~In particular, the upper branch shows a non-linear degradation, where the fitted distributions do not.~~ This is less obvious for UVic and ECBilt-CLIO, but especially apparent for the two MOM models.

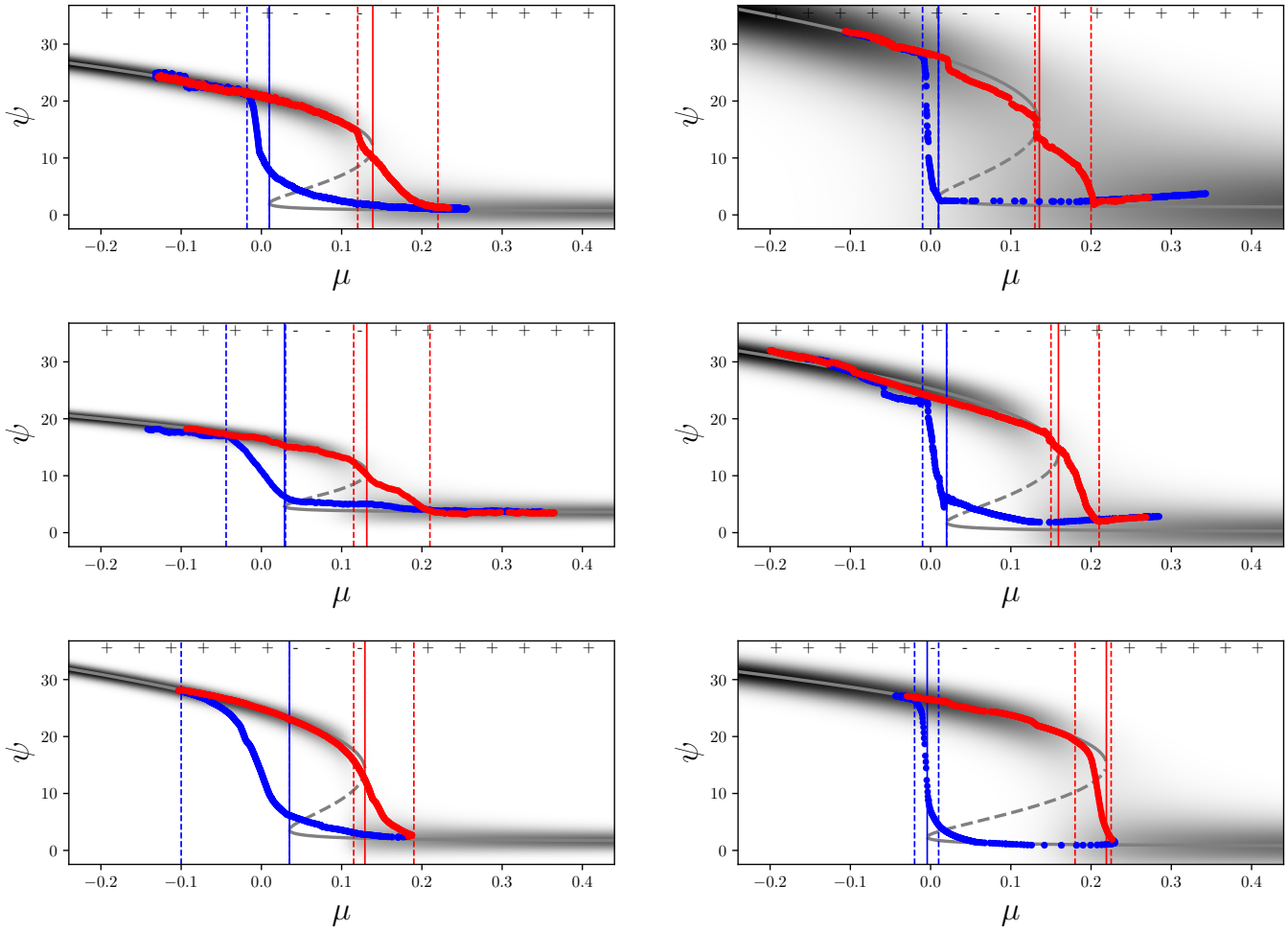


**Figure 7.** Estimated distributions under changing $\mu$. ~~Top-left to bottom-right~~Left column: ~~UVic~~Bremen, ~~MOM iso~~ECBilt-CLIO, C-GOLDSTEIN; right column: MOM hor, ~~C-GOLDSTEIN, Bremen~~MOM iso, ~~ECBilt-CLIO~~UVic. Vertical dashed lines mark the chosen boundary values for $\mu_{\pm}$, with solid lines the fit values. Grey dashed line indicates the local minimum in the distribution (trench). Top $\pm$ symbols indicate the sign of the discriminant $D$ for the fitted distribution ($+$ for unimodal, $-$ for bimodal). Distribution spreads have been inflated with a factor $\nu/2$ to make them visible. All values ~~in~~ have units ~~of~~ Sv.

## 4 Discussion and conclusion

We derived a simple model of AMOC collapse based on Langevin dynamics (Eq. 1) with a changing freshwater forcing ($\mu$) and applied this to EMIC simulated collapse trajectories taken from Rahmstorf et al. (2005). The collapse occurs at a bifurcation point $\mu_+$ which appears smaller than given in (Rahmstorf et al., 2005). A corresponding bifurcation point $\mu_-$ relates an abrupt transition back to the on-state. ~~Additionally, a linear parameterisation through state space couples to the freshwater applied to the North Atlantic subtropical gyre region.~~ The AMOC also requires an offset and scaling parameter to be fitted ($\lambda$ and $\nu$). These six parameters are sufficient to describe the abrupt collapse ~~/re-invigouration~~ of the AMOC that leads to a hysteresis loop under varying freshwater forcing. The resurgence of the AMOC is not the same as the collapse process and we did not attempt to obtain an accurate fit of that part of the hysteresis loop.

Any process which allows two stable states with rapid transitions between them and an asymmetric response to the forcing could in principle be described by our method. Other such geophysical processes might be ice sheet mass loss (e.g. Robinson et al. (2012) ), forest dieback (e.g. Staal et al. (2016)), and lake turbidity (Scheffer and van Nes, 2007).

The AMOC collapse and ~~re-invigouration~~ resurgence seen in these models cannot be completely fitted with Langevin dynamics due to the asymmetry in the lower vs the upper branch. It is, however, possible to fit the change in the upper branch of the AMOC–the 'on-state'–as it moves towards a critical point and the dominant salt-advection feedback mechanism breaks down.

We note that Rahmstorf et al. (2005) determine the AMOC strength as the maximum of the meridional volume transport in the North Atlantic and might explain the asymmetry between the two branches. If for a reverse overturning cell the wrong metric has been used then the lower branch location is not correct. It is conceivable that the Langevin model results in better fits if Rahmstorf et al. (2005) had sampled ~~$|\max(\Psi)|$~~ $\max(|\Psi|)$ instead of $\max(\Psi)$, which would have resulted in a better metric of the lower branch. With the metric used it is not apparent whether a reversed overturning cell was present or not because it was not sampled if the AMOC had taken on a negative value. Therefore, there is no obvious way to model the asymmetry between the two branches, and obtain a full description. The two branches could be separated by associating each with a different overturning cell. The upper branch is identified with the NADW-driven cell, while a reverse cell is responsible for the lower branch. If indeed a reverse overturning cell (as described in e.g. Yin and Stouffer (2007)) dominates the lower AMOC branch, two separate overturning cells are responsible for the observed trajectories, and the two branches then cannot be expected to fit with the same parameter set.

However, another possible explanation is that (two) separate mechanisms are responsible for the upper and lower branch dependency on $\mu$. Possible mechanisms include possible mechanisms include the influence of wind-stress, ~~SPG~~ North Atlantic subpolar gyre convective instability (Hofmann and Rahmstorf, 2009), or other pathways of deep water formation (Heuzé, 2017). Also, changes in the ITCZ (inter-tropical convergence zone) due to ocean-atmosphere feedbacks are possible (Green et al., 2019); these can, in turn, can affect the salinity of the North Atlantic subtropical gyre region. However, Mecking et al. (2017) showed that for a high-resolution model the salt-advection feedback was nevertheless stronger than the ITCZ effects. ~~Also, Gent (2018) notes that the EMICs in Rahmstorf et al. (2005) have reduced air-sea interaction feedbacks compared to more~~

~~modern and more complicated fully coupled ocean-atmosphere models; stronger feedbacks lead to greater AMOC stability.~~ Other wind coupling can occur further south through a coupling with the ACC (Antarctic Circumpolar Current) which is based on the thermal wind relation (Marshall and Johnson, 2017).

A third explanation is that deep water formation is a local process, and as a result an asymmetry is to be expected between the two branches. Local convection can, however, be subject to global controls and be associated with a sinking branch which occurs in conjunction with deep convection, but is not directly driven by it, see Spall and Pickart (2001) for a detailed discussion. The AMOC could develop a reverse cell where the overturning is driven by Antarctic Intermediate Water (AAIW), which is not part of the conceptual picture presented here (Yin and Stouffer, 2007; Jackson et al., 2017). The reverse cell introduces an asymmetry in the collapse trajectories because the driver of deep water formation is not in the North Atlantic, and might break our assumption that both the on and off branches are controlled by the same process. It is therefore difficult to estimate the return path of the AMOC if the lower branch has additional drivers from the dominant salt-advection mechanism of the upper branch. Forcing values appropriate for the lower branch might be different than those found for the upper branch.

Furthermore, the methodology used in this paper comes with difficulties in the numerical implementation. The fit procedure requires the normalisation of each distribution in the $\mu$ timeseries. Because no analytic solution exist a numerical approach is needed. The numerical integration adds to the computational costs of the fits. The Markov chain method is also prone to find local optima. Also, the cost of numerical integration necessitates stopping the fits at shorter chains than (perhaps) are needed, an analytic formulation of the integrand would alleviate this but none exists to our knowledge. Modern sampling algorithms allow for gradient information to be used, which is effective when sampling a higher dimensional parameter space (the Metropolis algorithm used in this paper has greater difficulty as the dimensionality of the parameter space increases). ~~It is possible for non-admissible solutions to be generated; in which case, the sampler is effectively reduced to a random searcher, till its finds a solution subspace to optimise under. (~~Tighter constraints on the prior distributions could be beneficial here. ~~)~~

As stated in Rahmstorf et al. (2005), the EMIC trajectories had already been smoothed, resulting in a smaller variance; a smaller variance leads to distributions that are more sharply peaked. This increases the computational cost of integrating the distributions numerically. Smoothing can also add to the inertia seen in the collapses, but might be due to other reasons such as stopping the EMIC simulations before equilibration of the AMOC collapse, leaving the AMOC in a winding-down state. Also, it is not clear how long the models in Rahmstorf et al. (2005) were integrated per freshwater forcing value. If the integrations were done for an insufficient amount of time, the AMOC collapse is incomplete, leaving the measured value out of equilibrium. The intermediate points in the collapse trajectories beyond the bifurcation points indicate that either the sample points are inaccurate or other processes are involved in the AMOC.

Finally, the fitted collapse trajectories were done on an ensemble of EMICs, which arguably are not sufficiently representative of the real climate. As noted by Gent (2018), the hysteresis behaviour has not been investigated fully in models of greater complexity than EMICs; the computational cost being the prohibitive. The question arises to what extent the procedure outlined in this paper can be applied to more complicated models such as those in the CMIP archives (Taylor et al., 2012). These models do not show a full collapse trajectory like those in Rahmstorf et al. (2005), which means no sample points of the lower branch are available. ~~If it is indeed possible to use direct numerical stochastic integration of the Langevin equation, no lower branch~~

needs to be sampled. Sampling from the prior distributions and optimising under the observed upper branch should also lead to robust estimates. Another advantage is that the probability distribution can be assumed Gaussian when on the upper branch. Also, CMIP provides times series of forced runs. To validate our method, a transient run requires known equilibrium bifurcation points, under a slowly changing $\mu$, and include an AMOC collapse. Using a simple box model, transition probabilities for an AMOC collapse have been determined by Castellana et al. (2019). From the CMIP ensemble a similar estimate might be obtainedusing a direct numerical stochastic integration approach, or at least the collapse characteristics of various models can be compared. Provided the CMIP models accurately capture the behaviour of the real AMOC and the freshwater forcing counterpart (our $\mu$) can be identified, an estimate can be made of the distance of the current climate state to the collapse point. Freshwater quantities such as $M_{ov}$ have been posited (e.g. Drijfhout et al. (2011)) as being suitable indicators of AMOC stability. It is possible that $M_{ov}$ relates to $\mu$ and can be used to extend our method to transient runs, but at present in is unknown whether this can be done. It is therefore still an open question how probable an AMOC collapse is in more realistic models, and reality, but with the method outlined in this paper a first step could be made in answering those questionsthis question.

*Author contributions.* S.D conceived the original idea. J.B. developed the theoretical formalism, performed the calculations, and prepared the figures. J.B., S.D., and W.H. contributed to the final version of the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This work was partially funded by the European Commission's 7th Framework Programme, under Grant Agreement number 282672, EMBRACE project. The authors thank the two anonymous referees for their valuable comments and suggestions that have improved the manuscript greatly.

## References


Balakrishnan, N. and Nevzorov, V. B.: A primer on statistical distributions, John Wiley & Sons, 2004.

Berglund, N. and Gentz, B.: Metastability in simple climate models: pathwise analysis of slowly driven Langevin equations, Stochastics and Dynamics, 2, 327–356, 2002.

Bernardo, J. and Smith, A.: Bayesian Theory, Wiley Series in Probability and Statistics, Wiley, 2009.

Birkhoff, G. and Mac Lane, S.: A survey of modern algebra, Macmillan New York, 1970.

Bolstad, W. M.: Understanding computational Bayesian statistics, vol. 644, John Wiley & Sons, 2010.

Boulton, C. A., Allison, L. C., and Lenton, T. M.: Early warning signals of Atlantic Meridional Overturning Circulation collapse in a fully coupled climate model, Nature communications, 5, 1–9, 2014.

Broecker, W. S., Peteet, D. M., and Rind, D.: Does the ocean-atmosphere system have more than one stable mode of operation?, Nature, 315, 21–26, 1985.

Broecker, W. S., Bond, G., Klas, M., Bonani, G., and Wolfli, W.: A salt oscillator in the glacial Atlantic? 1. The concept, Paleoceanography, 5, 469–477, 1990.

Castellana, D., Baars, S., Wubs, F. W., and Dijkstra, H. A.: transition probabilities of noise-induced transitions of the Atlantic ocean circulation, Scientific Reports, 9, 1–7, 2019.

Cobb, L.: Stochastic catastrophe models and multimodal distributions, Systems Research and Behavioral Science, 23, 360–374, 1978.

Cobb, L.: Estimation theory for the cusp catastrophe model, 1980.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., and Wehner, M.: Long-term climate change: Projections, commitments and irreversibility, pp. 1029–1136, Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9781107415324.024, 2013.

Dansgaard, W., Johnsen, S., Clausen, H., Dahl-Jensen, D., Gundestrup, N., Hammer, C., Hvidberg, C., Steffensen, J., Sveinbjörnsdottir, A., Jouzel, J., et al.: Evidence for general instability of past climate from a 250-kyr ice-core record, Nature, 364, 218, 1993.

Das, M. and Kantz, H.: Stochastic resonance and hysteresis in climate with state-dependent fluctuations, Phys. Rev. E, 101, 062 145, https://doi.org/10.1103/PhysRevE.101.062145, 2020.

Ditlevsen, P. D. and Johnsen, S. J.: Tipping points: Early warning and wishful thinking, Geophysical Research Letters, 37, https://doi.org/10.1029/2010GL044486, l19703, 2010.

Drijfhout, S. S., Weber, S. L., and van der Swaluw, E.: The stability of the MOC as diagnosed from model projections for pre-industrial, present and future climates, Climate Dynamics, 37, 1575–1586, 2011.

Edwards, N. R. and Marsh, R.: Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model, Climate dynamics, 24, 415–433, 2005.

Ganachaud, A. and Wunsch, C.: Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data, Nature, 408, 453, 2000.

Gardiner, C. W.: Handbook of stochastic methods for physics, chemistry and the natural sciences, vol. 13 of *Springer Series in Synergetics*, Springer-Verlag, third edn., 2004.

Gent, P. R.: A commentary on the Atlantic meridional overturning circulation stability in climate models, Ocean Modelling, 122, 57–66, 2018.

Goosse, H., Selten, F., Haarsma, R., and Opsteegh, J.: Decadal variability in high northern latitudes as simulated by an intermediate-complexity climate model, Annals of Glaciology, 33, 525–532, 2001.

Green, B., Marshall, J., and Campin, J.-M.: The 'sticky' ITCZ: ocean-moderated ITCZ shifts, Climate dynamics, 53, 1–19, 2019.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57, 97–109, https://doi.org/10.1093/biomet/57.1.97, 1970.

Held, H. and Kleinen, T.: Detection of climate system bifurcations by degenerate fingerprinting, Geophysical Research Letters, 31, 2004.

Heuzé, C.: North Atlantic deep water formation and AMOC in CMIP5 models, Ocean Science Discussions, 2017, 1–22, https://doi.org/10.5194/os-2017-2, 2017.

Hofmann, M. and Rahmstorf, S.: On the stability of the Atlantic meridional overturning circulation, Proceedings of the National Academy of Sciences, 106, 20 584–20 589, 2009.

Jackson, L., Smith, R. S., and Wood, R.: Ocean and atmosphere feedbacks affecting AMOC hysteresis in a GCM, Climate Dynamics, 49, 173–191, 2017.

Kleinen, T., Held, H., and Petschel-Held, G.: The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation, Ocean Dynamics, 53, 53–63, 2003.

Lemons, D., Gythiel, A., and Langevin's, P.: paper "Sur la théorie du mouvement brownien [On the theory of Brownian motion]", CR Acad. Sci.(Paris), 146, 530–533, 1908.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, Proceedings of the national Academy of Sciences, 105, 1786–1793, 2008.

Livina, V. N., Kwasniok, F., and Lenton, T. M.: Potential analysis reveals changing number of climate states during the last 60 kyr., Climate of the Past, 6, 2010.

Marshall, D. P. and Johnson, H. L.: Relative strength of the Antarctic Circumpolar Current and Atlantic Meridional Overturning Circulation, Tellus A: Dynamic Meteorology and Oceanography, 69, 1338 884, https://doi.org/10.1080/16000870.2017.1338884, 2017.

Mecking, J., Drijfhout, S., Jackson, L., and Andrews, M.: The effect of model bias on Atlantic freshwater transport and implications for AMOC bi-stability, Tellus A: Dynamic Meteorology and Oceanography, 69, 1299 910, 2017.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C., and Kahaner, D.: Quadpack: A Subroutine Package for Automatic Integration, Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2012.

Poston, T. and Stewart, I.: Catastrophe theory and its applications, Surveys and reference works in mathematics, Pitman, 1978.

Prange, M., Lohmann, G., and Paul, A.: Influence of vertical mixing on the thermohaline hysteresis: Analyses of an OGCM, Journal of Physical Oceanography, 33, 1707–1721, 2003.

Rahmstorf, S.: The thermohaline ocean circulation: A system with dangerous thresholds?, Climatic Change, 46, 247–256, 2000.

Rahmstorf, S. and Willebrand, J.: The role of temperature feedback in stabilizing the thermohaline circulation, Journal of Physical Oceanography, 25, 787–805, 1995.

Rahmstorf, S., Crucifix, M., Ganopolski, A., Goosse, H., Kamenkovich, I., Knutti, R., Lohmann, G., Marsh, R., Mysak, L. A., Wang, Z., et al.: Thermohaline circulation hysteresis: A model intercomparison, Geophysical Research Letters, 32, 2005.

Robinson, A., Calov, R., and Ganopolski, A.: Multistability and critical thresholds of the Greenland ice sheet, Nature Climate Change, 2, 429–432, 2012.

Scheffer, M. and van Nes, E. H.: Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size, in: Shallow lakes in a changing world, pp. 455–466, Springer, 2007.

Smeed, D., Josey, S., Beaulieu, C., Johns, W. E., Moat, B., Frajka-Williams, E., Rayner, D., Meinen, C., Baringer, M., Bryden, H., et al.: The North Atlantic Ocean is in a state of reduced overturning, Geophysical Research Letters, 45, 1527–1533, 2018.

Spall, M. A. and Pickart, R. S.: Where does dense water sink? A subpolar gyre example, Journal of Physical Oceanography, 31, 810–826, 2001.

Staal, A., Dekker, S. C., Xu, C., and van Nes, E. H.: Bistability, spatial interaction, and the distribution of tropical forests and savannas, Ecosystems, 19, 1080–1091, 2016.

Stommel, H.: Thermohaline convection with two stable regimes of flow, Tellus, 13, 224, 1961.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Vellinga, M. and Wood, R. A.: Global climatic impacts of a collapse of the Atlantic thermohaline circulation, Climatic change, 54, 251–267, 2002.

Weaver, A. J., Eby, M., Wiebe, E. C., Bitz, C. M., Duffy, P. B., Ewen, T. L., Fanning, A. F., Holland, M. M., MacFadyen, A., Matthews, H. D., et al.: The UVic Earth System Climate Model: Model description, climatology, and applications to past, present and future climates, Atmosphere-Ocean, 39, 361–428, 2001.

Weijer, W., Cheng, W., Drijfhout, S. S., Fedorov, A. V., Hu, A., Jackson, L. C., Liu, W., McDonagh, E. L., Mecking, J. V., and Zhang, J.: Stability of the Atlantic Meridional Overturning Circulation: A review and synthesis, Journal of Geophysical Research: Oceans, 124, 5336–5375, 2019.

Yin, J. and Stouffer, R. J.: Comparison of the Stability of the Atlantic Thermohaline Circulation in Two Coupled Atmosphere–Ocean General Circulation Models, Journal of Climate, 20, 4293–4315, https://doi.org/10.1175/JCLI4256.1, 2007.

The paper by van den Berk et al "Collapse of the Atlantic Meridional Overturning described by Langevin dynamics" is an interesting application of the classic analytical approach of Poston and Stewart with introduced stochasticity for modelling AMOC trajectories of the EMICs published in [Rahmstorf et al 2005]. I think the paper should be published after a minor revision.

R: We thank the reviewer for helpful comments and suggestions. Below you can find our responses.

The title should be corrected: "Modelling collapse of the Atlantic Meridional Overturning using the Langevin dynamics".

[x]R: Our suggestion would be "Characterisation of Atlantic Meridional Overturning hysteresis using Langevin dynamics" to emphasise the purpose of the paper better, that is, using a reduced set of numbers to quantitatively describe the AMOC collapse under a freshwater forcing.

As the authors admit themselves, EMICs are not sufficiently, representative of the real climate. Also, given the number of parameters the authors use to fit their model (six) and their geometrical origin (see description of ν and λ), I understand why the authors claim that only the freshwater forcing is the variable that determines the dynamical b

[x]R: Unfortunately, here seems to be a typesetting problem at ESD that renders some of the comments to be unreadable. As we understand it, the question is about using only freshwater as forcing for studying AMOC stability. The other possible forcing effect is thermal, and in principle a sufficiently large warming could also halt deep water formation and induce a collapse of the AMOC. However, in this paper we intend to explain the hysteresis behaviour shown in Rahmstorf et al. (2005), which is obtained by changing the freshwater forcing. As a result, we use this forcing as the dynamical variable that controls the stability regime of the AMOC. This point is discussed in the text now.

It would be interesting to see how the model can be used for forecast of bifurcations. The authors perform derivation of the model parameters using Bayesian framework, but once the model has been fully formed and the parameters are obtained for several EMICs, can the authors attempt forecast or hindcast of the bifurcating time series?

[x]R: We thank the reviewer for this interesting comment which could be explored in further research. A forecast from a partial AMOC weakening series would require an estimate of future freshwater forcing, and maybe making use of EMIC (or GCM) derived values as estimates. We added a paragraph in the Discussion section where we consider options along these lines for future research.

[Rahmstorf et al 2005] paper used 11 models and only hysteresis loops were presented (not actual AMOC trajectories)
https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1029/2005GL023655?download=true

[x]R: Our calculations are very time consuming. For this reason we decided to focus on the 6 models with most complete representations of the physics amongst the numerical models and disregard the 5 models without a 3-D ocean component. We consider that their characterisation is too far removed from the real world or CMIP class numerical models. This is now explained in the text (L227).

Can a figure be added with plotted time series that could be derived from the obtained model? For example, for the set of parameters averaged over a set of the selected EMICs? I wonder how realistic could be the time series and at what time scale it could forecast an AMOC bifurcation?
[x]R: This is a good suggestion, but unfortunately no timeseries were given in the published data. To derive those new runs would have to be made. The hysteresis loops are obtained by changing the forcing with small steps and then obtaining a new (quasi) equilibrium state for the changed forcing.

I understand that the framework is quite heavy computationally. Can the authors add discussion on how applicable can be this approach in other areas of geosciences where similar potential models may be used?

[x]:R: In principle, any hysteresis curve that is produced under a forcing where the lambda and nu transformations suffice to normalise the curve could be used. The calculation is indeed quite heavy in computational terms, but not more time-consuming for a hysteresis curve obtained in a full Earth System Model than for a hystresis curve from a much simpler EMIC. Other geophysical processes might be icesheet mass loss (e.g. Robinson, Alexander & Calov, Reinhard & Ganopolski, Andrey. (2012). Multistability and critical thresholds of the Greenland Ice Sheet. Nature Climate Change. 2.), forest dieback (e.g. Staal, A., Dekker, S.C., Xu, C. *et al.* Bistability, Spatial Interaction, and the Distribution of Tropical Forests and Savannas. *Ecosystems* **19,** 1080–1091 (2016)), or lake turbidity (Scheffer, M., van Nes, E.H. Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size. *Hydrobiologia* **584,** 455–466 (2007). Any process which allows two stable states with rapid transitions between them and an asymmetric response to the forcing could be described by our method.
Paragraph added (2$^{nd}$) in discussion.

The authors derived datasets from the published figures - is it allowed practice? Shouldn't they be obtained from the authors as datasets? Can the authors add information about the derived datasets in the table (number of points, etc)? Also, can more recent EMICs be used?

[x]R: We strongly support the development of open science and making data findable and accessible. Unfortunately, we have not been able to obtain the original datasets from the authors (we received no reply to our requests). The (individual points of the) measured values were retrieved from the plots by inverting the transformation matrices. This can be done for certain plots that are converted to pdf from plotting software such as Matlab. The dataset is then numerically the same as the set used to produce the plots in Rahmstorf et al (2005). The publisher allows for the use of individual graphics from their publications: we will note this in the text (L233).

[x]We have included a table with additional information about the dataset.

[x]In principle any hysteresis curve can be used, but we have not expanded the data set beyond the Rahmstorf et al set.

Further comments

The abstract should be modified to say that model is fitted to the trajectories.

[x] R: Corrected.

In the first paragraph, AMOC acronym is introduced twice.

[x] R: 2nd mention removed.

Instead of "invigoration" it is better to say "re-activation".

[x] R: Corrected with “resurgence” as suggested by our other reviewer.

Line 90 – "diagrams"

[x] R: Corrected.

Figure 2 - labels in all panels should be of the same font size

[x]R: Replotted with the same label size.

Line 124 - "the simplest"

[x] R: Corrected.

Line 152 - grey lines are mentioned in Figure 4, not clear which, maybe make them dashed? Similarly, dashed lines in Figs. 6,7 are impossible to see - enlarge these figures and all labels.

[x] R: Replotted with a colour different from grey and enlarge the labels.

Table 1 should be expanded to include more information on the selected models - countries, resolution, etc.

[x] R: Table included.

This is an interesting paper, which aims to describe the collapse and hysteresis of the AMOC observed in intermediate complexity climate models subject to freshwater forcing by low-dimensional Langevin dynamics as a stochastic bifurcation of a double well potential. Substantial revisions are necessary to improve the clarity of the manuscript and to support the conclusions.

R: We thank the reviewer for the detailed comments and valuable suggestions. Below are our revised comments and performed alternations.

General comments
1.) It is not clear what the purpose of the paper is. The authors do not state what their model is able to explain or predict.
Is the purpose to predict the exact parameter value of a collapse? Or at least to develop a method to do this?

[x] R: Our aim is to investigate whether it is possible to model the outcome of complex numerical models with a low dimensional model and thereby enhance understanding of the physics of an AMOC collapse and its hysteresis behaviour. The Langevin model defines a low-dimensional manifold that captures the essential collapse characteristics. To the extent that the low-dimensional model is successful in capturing the more complex model this investigation can indeed be seen as the development of a method to predict the parameter range where in a model a collapse would occur. The method is thus partly geared toward providing a means of prediction, but (at present) mainly to provide some characterisation of the collapse that will allow comparison between climate models. If a good fit can be found, then we can further explain the non-dynamical nature of the AMOC variations. As we show, this is partially the case. We have stated in the introduction where the purpose of the paper is defined.

Are there prospects to apply the method to observational data?

[x] R: Since the forcing is a freshwater anomaly in the North Atlantic, we would need to estimate the counterpart in the real world. Moreover, the forcing values at the bifurcations have to be known. At present, this is not the case for the real world, and in models it is model-dependent. An attempt could be made to relate the forcing to an indicator that can be linked to the bifurcation points, for instance $M_{ov}$, the freshwater transport by the overturning. This should first be tested in more comprehensive numerical models before applying it to observations. If this can be done, then from the transient change in mu and with knowledge of sigma from observed AMOC variations, a predictive model for the likelihood of an AMOC collapse could be developed. We have added a paragraph in the Discussion section with an outlook of future research.

Or is an aim to understand dynamically what is happening in realistic climate models?
This should be stated in the introduction

[x] R: We would like to see this paper as a first step to say more about the behaviour of the AMOC in complex numerical climate models such as used in the CMIP model intercomparisons. In particular, we propose a simplified low-dimensional model that is able to explain (and predict) bifurcation (tipping) points and abrupt change in the AMOC. Ultimately, this could be used to investigate abrupt behaviour in CMIP models, and how likely abrupt changes would be, even if not simulated by the model. Without a-priori knowledge of the freshwater-forcing values associated with the model's bifurcation points, first must be investigated whether mu can be linked to a general indicator like Mov (see the comment above). We have stated the purpose of the paper more clearly in the introduction and add the outlook above to the discussion section.

(P2L29ff). It is also unclear whether they want to only/mostly model the AMOC collapse (as stated at some points in the paper) or also the resurgence.

[x] R: We are mainly interested in modelling the collapse, not the resurgence. We note this in the manuscript as “Although the hysteresis loops of the AMOC include both a collapse and a resurgence, we will only attempt to model the collapse from the stable on-branch to the stable off-branch.”

2.) Regarding the conclusions, how can the authors say that the model successfully captures the dynamics?
They don’t compare with other models of higher or lower complexity, nor do they have any metric that shows goodness of fit or anything similar. This would be necessary to make such a conclusion.

[x] R: We thank the reviewer for this comment. We will add the posterior spread to show the goodness of fit. To explore and compare with other low-dimensional models than the Langevin model is beyond the scope of this paper (the six included already form a multi-model ensemble). A possible next step could be to apply the Langevin model to a transient run where mu depends on time. The AMOC should eventually collapse when increasing the freshwater forcing and we can have greater confidence in the applicability of our approach after testing it to such a transient simulation.

3.) The manuscript is not very well written and hard to follow. The terminology is often unclear. (E.g. what is a “track”, and how does the use of “stability landscape” apply here? See specific comments.) Some corrections are given under “technical corrections”, but the language and terminology has to be generally improved throughout the manuscript. Furthermore, I believe the manuscript can be shortened severely. What the authors want to get across can be said more efficiently. Many things are mentioned twice or more (see specific and technical comments for suggestions). Finally, the labels in multiple figures are unreadable.

[x] R: We critically reviewed the text. We comment on identified issues with the specific comments below. In short, the terminology has been clarified, and parts of the text that were confusing have been removed. The figures have been improved with larger and consistent fonts for the labels.

4.) The data acquisition seems problematic. I am not sure whether it is viable for this journal to present a data analysis based on visually extracted data from a figure of another publication.
Accordingly, the quality of the data is a major drawback of the study (e.g. arbitrary smoothing and AMOC metric).
Their main problem in fitting the data might be due to the specific metric that is shown in the Rahmstorf et al. (2005) figures, so it is a shame that the authors are not able to resolve that.

[x] R: Indeed, the data has been obtained from the figures of the paper. However, it is not visually extracted as the reviewer suggests. The figure we used is a vector graphic and the dataset can be retrieved from it by inverting the plot matrices used to map the original data to the values in the graph. We can exactly replicate the data used for this figure in this manner. In order to validate this method, we asked for the original data from Rahmstorf et al. (2005) but have not received a reply at present.
Less smoothing, and presumably larger noise levels, would likely show a stochastic collapse more easily.
We fully agree with this remark but want to emphasize that the main goal here is to develop and test a method to capture complex model behaviour with a simple low-order model. The methodology

described here is not affected by smoothing, only the assessment how well the method works is somewhat hindered by this.

5.) The description of their method contains many errors, and is incomplete. An explicit expression for the likelihood, as well as details of the Metropolis-Hastings implementation are missing.

[x] R: We agree that more detail could be given, but we believe a description of the Metropolis-Hastings algorithm is too much detail for this paper as it is well established and already described in many textbooks. We added a reference to the textbook of Bernardo & Smith which describes this algorithm.

In the discussion, the authors name difficulties in the numerical implementation as a possible reason for the failure of their fit to describe the lower AMOC branch, but it is for the reader not possible to assess whether this is relevant, since no details or robustness tests are given.
Furthermore, it is not stated how many data points the respective data sets contain, and it is not mentioned that the authors assume successive data points to be independent.

[x]R: Indeed, it is important to assess the robustness of our implementation. To further detail the validity methodology and outcome, we added a table with model characteristics and state that each point is independent.
Bremen: 2461
EC-Bilt-CLIO: 243
C-GOLDSTEIN: 849
MOM hor: 1233
MOM iso: 1442
UVIC: 464
In addition, we want to emphasize that our main goal is to model the transition from on-branch to off-branch, that is, the upper right half of the hysteresis curve, and not so much the dynamics that govern the lower branch, also because we assume that other dynamics govern the lower branch and our simple model has to be extended to account for those dynamics.

It is also not mentioned how the maximum of the posterior parameter distributions is picked.

[x]R: The most likely value is the mean of the posterior distribution. This does assume that the posterior distributions are unimodal. We will discuss this in the text.

6.) Finally, several questions regarding the methodology.
a) Why do the authors not try to estimate sigma with their Bayesian method?

[x]R: In principle this can be done. The variation in the hysteresis loops appears constant and can therefore be estimated more easily by other means. This does add to the computational costs and expands the search space, however, making it more difficult to find solutions. Therefore, we did not follow this method. We will mention this in the text (L261).
Why not include observational noise?

[x]R: Observational noise of the real AMOC would have a larger spread than the sigma we obtained. Synthetic series on the basis of the found parameters could indeed be generated with such a noise level. However, we intend to fit the intermediate complexity model outcome, using the data of that particular model. AMOC collapse and its likelihood at a given point in parameter space is model-dependent. One of the essential parameters in this dependency, is the model-dependent

sigma. Therefore, we prefer to use the sigma that is characteristic of each particular model. This point is now discussed in the paper (L257).

This could handle the fact that the data is filtered arbitrarily. It could also completely change the locations of the inferred bifurcation points.

[x] R: The bifurcation points are determined by the limit (non-stochastic) solutions. A noise driven transition could occur, however, and push the points that bound the hysteresis curve further inwards, towards each other. We will mention this limitation more clearly in the paper (L241).

b) To make the paper more understandable it would be good to note explicitly early in the manuscript that the movement of mu is actually known.

[x]R: We agree with the reviewer. We added "The forcing values of mu are known and the same for each climate model."
However, the values of mu+ and mu- are not, and model-dependent.

c) Why not try multiplicative noise? (see also e.g. Das/Kantz Phys. Rev. E 101, 062145, 2020) This should relatively easily give a model that describes the asymmetric behavior.

[x] R: Multiplicative noise is state dependent, while we have made the assumption that the noise level is constant; therefore, we used additive noise. We will discuss this point in the text (L160).

d) It should be noted explicitly that there is no time dependency of the data. I wonder why they choose not to fit to time series instead? This would allow to treat the non-equilibrium nature of the data. Also, it would be much more applicable to observational data and to make predictions.

[x]R: Indeed, we will add a remark that each data point corresponds to a fixed value of mu and is not time dependent. No timeseries are available, however. Each mu value represents a separate climate model run which has run (more or less) to equilibrium. We need to be able to estimate mu+ and mu- from the models, and this can only be done from a hysteresis-curve which indeed contains equilibrium solutions and no time-dependence. We agree that a logical next step is to apply the method on time-dependent runs, but to validate the model it is needed that we know the equilibrium bifurcations points as well for that model, and that the time-dependent runs are based on a slowly-changing mu and include an AMOC collapse. There are not many models available that answer all these requirements. We added a paragraph in the Discussion section where we mention this point.

e) Why not only move along alpha at a certain fixed beta? Is moving both parameters supported by the data significantly better?

[x]R: Early attempts with a fixed beta resulted in worse fits; therefore, we opted not to restrict beta. Removed (see comment P14L255)

Specific comments
Abstract: "Machine learning": To my knowledge MCMC is not considered a machine learning technique. The abstract needs to be expanded to better reflect the motivation of the study, what their method enables them to do, and their conclusions.

[x] R: We will remove mentioning machine learning. We will also add to the abstract: "The Langevin model allows for comparison between models that display an AMOC collapse. Variation between climate models studied here is mainly in the strength of the stable AMOC and the strength of the response to a freshwater forcing."

P2L42-45: This is a not a very clear explanation of the salt advection feedback. The main point is that North Atlantic salinity anomalies (positive/negative) are amplified by their effect on the overturning flow (strengthening/weakening), the strength of which controls the North Atlantic salinity.
This is thus a positive feedback and leads to bi-stability with the associated possibility of abrupt transitions.

[x]R: Rewritten as suggested.

P2L53: ". . .number of solutions for a given value of the freshwater forcing goes from 2 to 3. . . ". Should say "goes from 3 to 1" as the bifurcation point is crossed. (There are 2 solutions precisely at the bifurcation point, but I think this saddle-node fixed point is not relevant here.)

[x] R: Corrected this as suggested.

P3 Caption Fig.1: The terminology of this figure is not appropriate and furthermore not understandable at this point within the manuscript. No trajectory is shown, but a bifurcation diagram.

[x] R: Correct, renamed to "bifurcation diagram".

They have to be more specific with what they mean by a deformation of the "trajectory".
Also, at this position within the manuscript, it is completely unclear what they mean with "trench of the distribution".
Either leave out or explain in the main text.

[x] R: Use of "trench" removed from the text.

Furthermore, I suggest to use the term "resurgence point" for mu-, and use that terminology throughout the paper.

[x] R: Suggestion in agreement with our other reviewer, replaced with "resurgence point".

Note that e.g. in P5L91, mu+/- are being referred to as "collapse points".

[x] R: Replaced "collapse points" with "bifurcation points".

P4L64: Can the authors elaborate why they think a double well potential has mainly been studied qualitatively? I would argue that this simple and general mathematical model has been studied quantitatively to an exceptional degree.

[x] R: The reviewer is correct that this model has been extensively studied and applied. But to our knowledge it has not been quantitatively applied to AMOC collapse in complex numerical climate models or observational data in the way as we present here. We will replace "studied mainly in a qualitative way (within catastrophe theory)" with "the double well potential has been extensively studied and applied, also in a quantitative way. But to our knowledge it has not been quantitatively applied to AMOC hysteresis using the Langevin equation in complex numerical climate models before.
Bolton et al (Boulton, C., Allison, L. & Lenton, T. Early warning signals of Atlantic Meridional Overturning Circulation collapse in a fully coupled climate model. *Nat Commun* **5,** 5752 (2014).)

do study an AMOC hysteresis loop qualitatively, but do not mention the Langevin equation. We added Bolton et al to our references and discuss its relevance as a quantitative study of AMOC bistability, specifically, the transient run studied in that paper and how it could relate to the Langevin model.

P4L65ff: It is a bit confusing when the authors first say that 2 parameters are enough to describe bi-stability, but then use another 2 parameters to scale and shift to the AMOC variable.
Maybe it would be better to first explicitly say that by a shift and scale of the variable x, one can eliminate the third order term as well as the fourth order coefficient.

[x] R: The reviewer is correct, indeed that is their purpose: to reduce the polynomial to a smaller set such that only the minimal number of parameters remain. Added clarification as suggested (L77).

Both of these transformation do not influence the global bifurcation behavior. Then, they can state that a shift and scaling is considered when fitting to the climate model data.

[x] R: The reviewer is right, because the topology is not affected. We added a remark added as suggested (L92).

P4L78-81: Can the authors elaborate why they obtain these rough estimates for the parameters, and how they are insensitive to other parameter values?

[x] R: These are not estimates but interpretations that can be linked to the bifurcation diagram. We will replace “The value of nu is roughly … “ with “In the bifurcation diagram the value of nu is roughly … .” And likewise for lambda.

P5L90-91: When speaking about “solution” what exactly do the authors mean?

[x] R: In this case we mean that only the trivial solution exists: only 0 as the value for all variables. This is further explained in the text (L104).

P5L92-97: This section is a bit unclear. Can the authors define a “track”, and what does it mean to be one-dimensional?

[x] R: We removed the notion of a “track”. It being 1-dimensional means that because the hysteresis loop is 1-dimensional the values (alpha, beta) as a function of mu are as well.

The fact that alpha and beta are called normal and splitting factor is better mentioned earlier.

[x] R: Sentence moved up.

A more clear distinction of “parameter” and “variable” would be appropriate.

[x] R: We have clarified the text such that the Rahmstorf set has data mu and psi; alpha and beta are the stability parameters (which in turn are expressed as a rate and offset); nu and lambda are the scaling parameters, and mu+/- the bifurcation points.

P5L101: This argument is unclear to me. The fact that the AMOC is scalar variable should not constrain the path through the stability landscape in any way.

Do the authors rather want to say that in the climate model experiments there is only a single control parameter mu, and that by assuming a linear dependency of both alpha and beta on mu, they can express some parameters by the extremal values of mu?

[x] R: Indeed, mu is the only control parameter. By assuming linearity a reduced set of equations can be determined later on. This is now explained in the text (L139).

P8L127-129: Maybe the authors can elaborate more specifically on why these arguments are relevant in order to neglect a non-linear change of either mu or alpha/beta?

[x] R: We will remove these lines, they are not needed for the argument.

P8L141-146: Improve this explanation. When introducing stochasticity, the asymptotic dynamics for each parameter value give rise to a stationary density.

[x] R: The reviewer is correct, we added “The potential function can be replaced by a distribution which is the stationary distribution in the asymptotic limit (i.e. the long term behaviour of repeated sampling of the hysteresis loop).”.

In the case of the scalar potential, this distribution can be given analytically up to a normalization factor. Thus, the distribution can be used as a likelihood function (if I understand correctly) for parameter inference with MCMC.

[x] R: The reviewer is correct, we added. “ but can be computed numerically (and therefore used as a likelihood function in the next section)”

P9L157: What is the “sampled” bi-modality region?

[x] R: Sampled as in where the dataset has values. We dropped “sampled”.

P9L160: “These changes correspond directly to the potential functions in Fig. 2.” What is meant by this?

[x] R: We mean that the distribution functions can be linked to the different characteristic shapes of the polynomial catalogued in fig 2. We will rewrite as “Each distinct shape of the distribution can be linked to one of the potential functions in Fig 2.”.

P10L174: Can the authors explain why lambda < nu in general?

[x] R: The offset (lambda) cannot exceed the scaling factor (nu) because the offset needs to be roughly in the middle of the two stable branches.

Caption Figure 5: The terminology is unclear. What is meant by a “singular” maximum?

[x] R: Removed (superfluous)

What is meant by the dominant and the weak mode, and what is the inversion?

[x] R: For a bimodal distribution there is a mode with more probability mass than the other which we call the dominant mode and the model with less mass the weak mode. And inversion is where these modes switch in strength: the dominant turn weak and the weak turns dominant. Replaced with “small” and “large”.

There are also grammatical errors ("...in the middle and inversion from weak. . .").

[x] R: "in the middle and inversion from weak to dominant takes place" removed from caption.

P11L180: I think a more precise statement would be that they estimate the posterior probability distribution of the parameters, given the data psi(mu).
Furthermore, the following equation does not define the likelihood but the posterior distribution.

[x]R: Yes, the reviewer is correct; rewritten as suggested: the equations states that the posterior distribution is proportional to the likelihood multiplies with the prior distribution.

P11L183: Why linearised?

[x]R: Removed, this related to linearisation of beta0, delta beta, etc

P11L184-188: This statement of Bayes' rule is not correct, please revise. The right hand side is not called Bayes' factor (which arises in model comparison).

[x]R: We removed L187-188.

P12L209-219: The constants muS+ etc. are not properly introduced and should be shown in one of the figures.

[x] R: These are given in the caption of fig 6. Reference to the figure given in the text.

The footnote 2 needs to be explained better.

[x] R: This is a technicality in how we defined the optimiser. Rewritten to explain that this is useful to avoid solutions that intersect the B1 or B2 twice (see also P17L308-310 below).

Caption Table 1: Why is a linear function used and not a higher order polynomial?
This does not seem to be very suitable to the data.

[x] R: We only fitted the beginning (left part) of the upper branch and assumed a constant sigma. clarification to the text added (L255).

P13L228-230: This is not very precise wording. What do the authors mean with "un-stable" and "more stable" solutions?

[x] R: Removed "more". Stability relates solely to the attractors and repellor.

P14L242-245: What do the authors want to say here? It comes as a surprise to me that suddenly only the data for mu > mu+ should be relevant?

[x] R: We ignore the data on the lower branch before the collapse point because we did not want it to influence the fits, especially because we are only interested in the collapse from the upper branch. This is now mentioned in the text (L32,L249).

And why do they now claim that the model C-GOLDSTEIN does not "appear" to show a collapse?

[x] R: Removed: our intent was to comment of the smoothness of the trajectory, but it is unnecessary,

P14L252-253: Unclear what the authors are trying to say.

[x] R: We removed these lines, they are redundant.

P14L259: What is meant by “non-linear degradation”?

[x] R: We mean the part of the hysteresis loop after the collapse point: before that point the change was fairly linear, but after it is strongly non-linear. Removed

P15L267: In what way is the model sufficient to describe the data? Certainly the “re-invigouration” is not well captured.

[x] R: The aim was to model the collapse; the resurgence appears more difficult. Added a clarification to emphasise this point and estimate a goodness of fit.

P17L308-310: Unclear what is meant here. What are “non-admissible” solutions?

[x] R: An inadmissible solution is one where the curve through (alpha, beta) space intersects one of the subspace B1,2 twice. Because B1,2 are concave this is a possibility. Removed from text

P17L313-315: Unclear. Smoothing might be due to other reasons?

[x] R: Perhaps, but is not mentioned in Rahmstorf (2005).

P17L323: What is meant by “direct numerical stochastic integration”?

[x] R: This remark was originally intended to point out another way to perform the calculations: by solving the SDE directly. We now believe this remark to be redundant and removed it.

P17L330: How exactly does this paper present a step forward to assess the likelihood of a future collapse of the AMOC?
The method presented here relies on previously modeled collapses of the AMOC with realistic climate models.
How does the method generate additional information?

[x] R: If the characterisation has predictive values, more complex climate models can be used to derive a collapse point if the freshwater forcing at the bifurcation points can be estimated. It is also a way to compare the collapse characteristics of various models. If the freshwater forcing at the bifurcation points can be generalised and linked to a robust indicator (such as, perhaps, Mov), the method can be applied to the real world as well. We agree, there are quit some steps in between the method outlined here and its extension to the modelled and observed timeseries. We will expand the discussion on this point.

Technical corrections

[x] R: We are in agreement with all corrections below

P1L15: last glacial maximum and early holocene -> last glacial period

[x] R: Replaced as suggested.

P2L26: . . . which are presumed to be functions . . .

[x] R: Corrected.

C6P2L30: ...and tipping points in the climate, it has not been . . .

[x] R: Corrected.

P2L38: . . . or an increased surface freshwater flux by changes in precipitation minus evaporation. . .

[x] R: Corrected.

P2L46: scalar variable obtained by integrating …
[x] Corrected.

P2L52: . . . one of the two basins of attraction vanishes . . .

[x] R: Corrected.

P7L126: rather “(delta alpha, delta beta)”?

[x] R: Correct: changes in alpha, beta: corrected.
P8L142: Remove: As shown by Cobb (1978), this distribution belongs to the exponential family.

[x] R: We removed this.

P8L144: The polynomial potential introduced in the previous section, we had already. . ., gives the probability . . .

[x]R: Corrected.

P8L148: Note that C = C(alpha,beta), which does not have…

[x]R: Corrected.

P8L150: . . . because of the scaling …

[x] R: corrected.

P8L152 and Fig. 4 caption: In what way is this a sample collapse trajectory, or an example trajectory?

[x] R: This should indeed be example, not sample.

P9L161: . . . a change in only one . . .

[x] R: Corrected

P9L163-165: Why not say this at P8L148? It is a bit redundant otherwise.

[x] R: We moved these lines to P8L148.

P10L168: arrived at -> described

[x] R: Corrected

P10L171: independent of each other

[x] R: Corrected.

P11L180: The method used is not considered machine learning.

[x] R: Removed.

P11L190: Does not seem to be relevant, as it is not done here.

[x] R: This sentence relates to "These resultant posterior distributions can, in turn, be used as prior distributions, yielding a chain of sampled parameter vectors." It is roughly how the sampler works, but we removed this because it is redundant.

P11L194-196: This is partly redundant, and it is not clear why the authors mean that the model can be fit with uninformative priors.

[x] R: We removed the part about uninformative priors: it is unnecessary.

P11L199-200: Redundant.

[x] R: We replaced "With ν and λ introduced earlier, the state variable x undergoes an affine transformation and normalises the polynomial. These ..." with "The parameters ν and λ … "

P11L207: Redundant.

[x] R: Removed.

P11L207-208: An overview of priors is: The following prior distributions are used:

[x]R: Corrected.

P14L243: do -> to

[x]R: Corrected.

P14L254: Why "sample" paths? The authors are showing distributions, which is exactly contrary to showing sample paths.

[x]R: We removed " … for linearly parametrised sample paths through the stability space" … "

P14L255: Redundant.

[x]R: We removed “The β parameter changes linearly and α follows from the constraints in Eqs. 4 and 5. Blue and red lines indicate the prior bounds for μ − and μ + , respectively.”

P15L265: couples -> models. Why “additionally”?

[x]R: This is not needed:  we removed “Additionally, a linear parameterisation through state space couples to the freshwater applied to the North Atlantic subtropical gyre region.”

P16L275: Do they mean arg(max(|psi|)) ?

[x] R: We mean max(|psi|),  corrected.

P17L309 till its -> until it

[x] R: Sentence removed