Uncertainty quantification for overshoots of tipping thresholds

Kerstin Lux-Gottschalk[1,*] and Paul D. L. Ritchie[2,3,*]

[1]Eindhoven University of Technology

[2]Department of Mathematics and Statistics, Faculty of Environment, Science and Economy, University of Exeter, North Park Road, Exeter, EX4 4QE, UK

[3]Global Systems Institute, Faculty of Environment, Science and Economy, University of Exeter, North Park Road, Exeter, EX4 4QE, UK

[*]These authors contributed equally to this work.

Correspondence: Kerstin Lux-Gottschalk (k.m.lux@tue.nl) and Paul D. L. Ritchie (Paul.Ritchie@exeter.ac.uk)

Abstract. Many subsystems of the Earth system, that are currently in a stable state, are at risk of undergoing abrupt transitions to a drastically different, and often less desired, state due to anthropogenic climate change. These so-called tipping events often present severe consequences for ecosystems and human livelihood that are difficult to reverse. Forcing a nonlinear system beyond a critical threshold that signifies the onset of self-amplifying feedbacks provides one possible mechanism for tipping. However, previous work has shown that it could be possible to briefly overshoot a critical threshold without tipping. For some cases, the peak overshoot distance and the time a system can spend beyond a threshold are governed by an inverse square law relationship (Ritchie et al., 2019). However, in the real world or complex models, critical thresholds and other system features, such as inherent timescales and the system's linear restoring force after perturbations, are highly uncertain. In this work, we look at how such uncertainties affect the probability of tipping in response to a temporary overshoot from the perspective of uncertainty quantification. We show the importance of constraining uncertainty in the location of the tipping threshold and the linear restoring force to the system's stable state to better constrain the uncertainty in the tipping behaviour for overshoot trajectories. We first prototypically analyse effects of an uncertain critical threshold location separately from effects due to an uncertain linear restoring force. We then perform an analysis of joint effects of uncertain system characteristics within a conceptual model for the Atlantic Meridional Overturning Circulation (AMOC). The simple box model for the AMOC shows that these uncertainties have the potential to reverse conclusions for mitigation pathways. A pathway believed to offer little danger of tipping, may become highly dangerous if the tipping threshold were to be closer than previously understood. In this conceptual model, we illustrate how constraining the highly uncertain diffusive timescale (representative of wind-driven fluxes) within this box model reduces the tipping uncertainty of the AMOC in response to overshoot scenarios.

1 Introduction

Recently, climate tipping points have gained increasing attention by scientists, policymakers, and the public (Lenton et al., 2023). Tipping events are abrupt transitions that may occur if some external forcing crosses a critical threshold (Scheffer et al., 2012). Systems that are currently in a stable state may find their current state ceasing to exist beyond the threshold, and therefore cause the system to transition (potentially irreversibly) to a drastically different state (Lenton et al., 2008). In the latter study, the authors have identified tipping elements of the climate system. A recent assessment can be found in Armstrong McKay et al. (2022), where the authors elaborate on the most important tipping elements and corresponding tipping points. These tipping points pose severe threats to ecosystems and human habitat (Lenton et al., 2023). The evolution of many subsystems of the Earth, such as the Atlantic Meridional Overturning Circulation (AMOC), are prone to exhibit tipping behaviour instead of only featuring small gradual changes (Armstrong McKay et al., 2022). A tipping of the AMOC would be likely to cause a significant cooling over Northern Europe, substantially change patterns in tropical rainfall, as well as trigger regional sea level rise (Jackson et al., 2015). Gaining a better understanding of the tipping behaviour of the AMOC is crucial for deriving efficient climate change mitigation strategies.

Awareness of the need for action is prevalent since the impact is likely to be far reaching if a system tips, see e.g. Ritchie et al. (2020). To increase the efficiency of measures, a better understanding of these tipping events and under which circumstances tipping can be prevented is crucial. To understand overshoots of tipping thresholds, it is important to understand which mechanisms can cause a system to tip. We distinguish between noise-induced, rate-induced, and bifurcation-induced tipping (Ashwin et al., 2012). So called *noise-induced tipping* can occur when fluctuations from fast processes become particularly pronounced and might cause the system to tip, see e.g. Ashwin et al. (2012); Ma et al. (2019). Alternatively, changing the external forcing too quickly can lead to a different tipping mechanism known as *rate-induced* tipping, see e.g. Ritchie et al. (2023) The mechanism of crossing critical thresholds that leads to tipping is referred to as *bifurcation-induced tipping*. In Kuehn (2013), the authors provide a mathematical framework for critical transitions in terms of bifurcation theory, see also Kuznetsov (2004); Wiggins (2003) for further mathematical details. More recently, the universal nature of the emergence of critical transitions in physical systems has been analysed in Kuehn and Bick (2021).

All three mechanisms can contribute to the uncertainty in mitigation windows. Since we would like to understand overshoots of tipping thresholds of global warming, here, as a natural restriction, our primary focus is on better constraining the mitigation window for bifurcation-induced tipping. For some subsystems of the Earth, critical thresholds have been suggested to be at low levels of global warming (Armstrong McKay et al., 2022), such that overshoots of the threshold are becoming increasingly likely. It is important to note that for some elements, climate model simulations have shown that tipping might still not occur, if the reversal of the forcing is sufficiently fast (Jackson and Wood, 2018a; Jackson et al., 2022). Since for some systems, we are already very close to a level of global warming that might trigger tipping (see Armstrong McKay et al. (2022) Figure 2), it is crucial to quantify which exceedance level of a possible threshold might allow us to return to the original state if forcing is reversed sufficiently quickly. This overshoot mechanism has already been subject to thorough investigations (Ritchie et al., 2021; O'Keeffe and Wieczorek, 2020; Wunderling et al., 2023; Bochow et al., 2023).

However, much less is known about the impact of uncertainties on overshooting tipping thresholds. In particular, the IPCC 2021 report (Masson-Delmotte et al., 2021) emphasises on high impact, low likelihood climate outcomes, to which some tipping events belong, as being part of climate risk assessments. To assess risks of tipping, a thorough handling of uncertainties is needed. One type of uncertainty related to a possible tipping of the AMOC is uncertainty in datasets such as those for sea-surface temperature and salinity. An uncertainty propagation procedure to quantify effects of dataset uncertainties on indicators of critical slowing down has recently been proposed in Ben-Yami et al. (2023). Here, we focus on uncertainties in a conceptual model for the AMOC regarding choices of model parameter values (Lux et al., 2022). These uncertainties might significantly affect the mitigation window, that is how far and for how long a system may overshoot tipping thresholds and still retain the system's original equilibrium state (Ritchie et al., 2019). Better constraining the mitigation window is a crucial task to gain a better understanding of overshoots in real world climate systems. For the AMOC, the uncertainty in the critical global warming threshold is particularly pronounced (Armstrong McKay et al., 2022). There is a need for further research on overshoots of tipping thresholds (of the AMOC) under uncertainty in model parameters to quantify the mitigation window more narrowly.

Therefore, in this work, we focus on the quantification of uncertainty in overshooting tipping thresholds resulting from uncertainty in system characteristics for a given forcing profile. In particular, we illustrate our methodology for AMOC overshoot scenarios with the aim of gaining a conceptual understanding of the mechanisms involved and in particular how uncertainty affects the mitigation window. This is not only important for the AMOC, but for many other systems as well (Ritchie et al., 2021; Meyer et al., 2022; Bochow et al., 2023).

The remainder of the paper is structured as follows: Section 2 details the problem setup, including introducing the mitigation window for overshoots without tipping. In Section 3, we present how uncertainty in system parameters affects the mitigation window in the prototypical fold bifurcation setting, inherent to many conceptual climate models. Thereby, we distinguish between uncertainty in (i) the location of the tipping threshold p_b and (ii) the linear restoring force proportionality constant κ. Section 4 provides results on the uncertainty in the mitigation window for a conceptual AMOC model, the Stommel-Cessi box model (Cessi, 1994), which exhibits two of these fold bifurcations. The uncertain diffusive timescale in this model entails uncertainty in both p_b and κ, thus exhibiting effects showcased in the previous Section. In Section 5, we expand on how the results on the probability of tipping for the presented scenarios might inform decisions about alternative mitigation pathways.

2 Problem Setup

Here, we consider various profiles for the change in an environmental parameter that have different impacts on the overshoot of the critical bifurcation (threshold) value. Most importantly, we analyse these overshoots in the presence of uncertainties in model parameters of a conceptual AMOC model. These uncertainties affect the time and the peak overshoot distance that would still facilitate a return to the original state. More precisely, we use the Stommel-Cessi model (Cessi, 1994) to conceptually illustrate the far reaching effect of uncertainty in wind-driven gyres and eddies, represented by the diffusive timescale parameter, on mitigation windows without tipping. This model exhibits a double-fold bifurcation and thus includes a range of forcing parameters where the AMOC exhibits multistability.

The aim is to develop a probabilistic extension of the work of Ritchie et al. (2019). Therein, the authors derived an inverse-square law between peak overshoot distance and exceedance time over the threshold to system characteristics that determine a mitigation window in the case of tipping via a fold bifurcation. Here, we focus on the one-dimensional case and refer the reader to Ritchie et al. (2019) for the multi-dimensional case. Consider an ordinary differential equation that takes the form

$$\dot{x} = f(x,p)$$

and exhibits a fold bifurcation at the point (x_b, p_b). We define

$$a_0 := \frac{\partial f}{\partial p}(x_b, p_b), \quad \text{and} \quad \kappa := \frac{1}{2a_0}\frac{\partial^2 f}{\partial x^2}(x_b, p_b),$$

where a_0 is related to the inverse of the system's timescale, and κ is proportional to the linear restoring force, which is a measure of the recovery rate back to the equilibrium after a perturbation of the system. The mitigation window can then be described via

$$(p_{peak} - p_b)t_{over}^2 < \frac{4}{a_0^2 \kappa}. \tag{1}$$

The condition provided in (1) specifies the mitigation window in terms of the exceedance time over the threshold t_{over} and the peak external forcing p_{peak} over the critical threshold p_b. Hence, to be within the mitigation window, the product of the squared time spent over the threshold and the peak overshoot distance needs to be smaller than a quantity that depends on system specific parameters, which can be highly uncertain in real-world applications.

3 Results for prototypical fold bifurcation model

We begin by considering a simple conceptual model for tipping via the prototypical fold bifurcation given by

$$\tau \dot{x} = p_b - p(rt) - \kappa(x - x_b)^2. \tag{2}$$

Utilising such a simple model allows us to isolate how uncertainty in either the location of the tipping threshold (fold bifurcation), p_b, or the linear restoring force of the starting state (proportional to κ) affects the tipping behaviour. The height of the fold bifurcation is chosen to be at $x_b = x_0 - \sqrt{p_b/\kappa}$ such that the system is initialised at the same starting position, $x_0 = 2.5$, regardless of the threshold location or linear restoring force. For simplicity, the timescale parameter, τ, is set to unity and is inversely proportional to a_0 ($\tau = -1/a_0 = 1$). The forcing profile used for all results based on the prototypical fold bifurcation is a symmetric overshoot given by

$$p(rt) = \Delta_p \operatorname{sech}^2(r(t - t_{peak})), \tag{3}$$

that starts and finishes at zero, has an amplitude Δ_p, and the rate of change is controlled by r. For more fundamental details on the theory of fold bifurcations, we refer the reader to Kuznetsov (2004); Wiggins (2003).

3.1 Results for uncertain location of tipping threshold

First, we consider uncertainty in the location of the tipping threshold. We will use the prototpyical fold model (2) and only consider different realisations of the parameter for the tipping threshold location, p_b, which we consider uncertain here. Note, that the linear restoring force (and distance to the basin boundary) will be different at the same forcing level, for two different threshold locations. Importantly however, all system characteristics remain unchanged when the system's are the same distance to their respective thresholds.

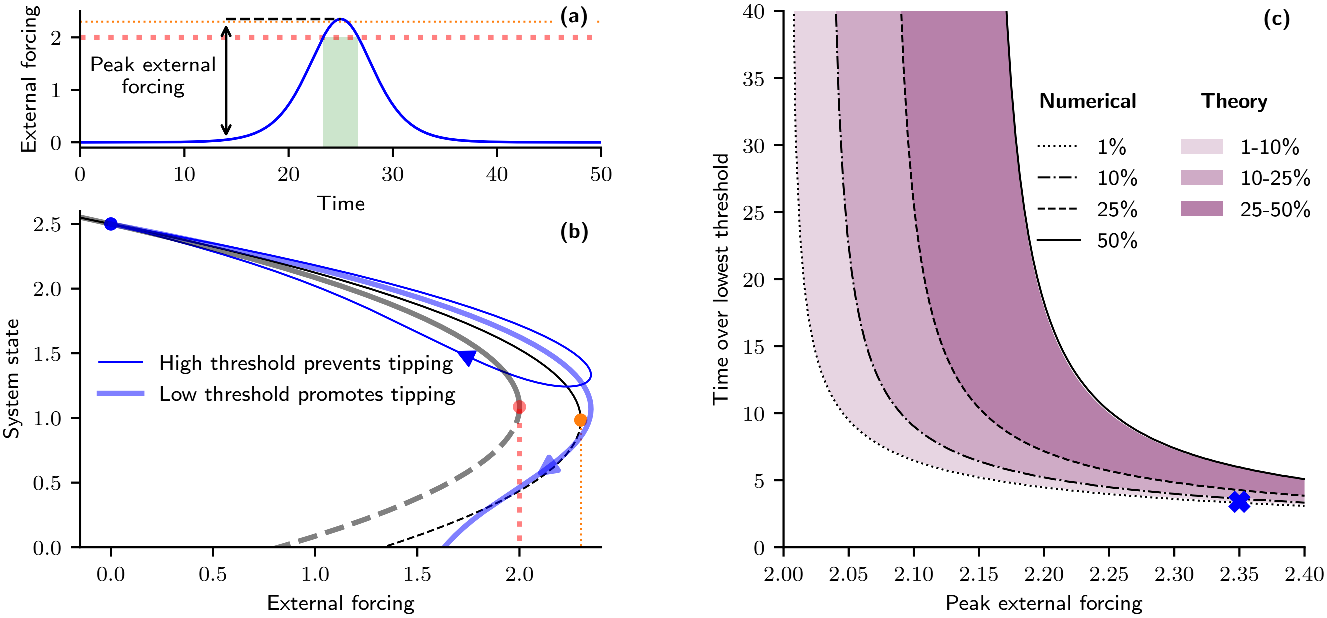

Figure 1. Probabilistic overshoots given uncertainty in location of tipping threshold. (a) Time profile of an exemplar external forcing given by equation (3) (parameters: $\Delta_p = 2.35$, $r = 0.24$, $t_{peak} = 25$). Red and orange dotted lines indicate low ($p_b = 2$) and high ($p_b = 2.3$) tipping threshold locations respectively. (b) System responses (blue) subjected to the external forcing profile given in (a) for model given by equation (2) with fixed $\kappa = 1$, but either low threshold, $p_b = 2$ (thick and translucent curves) or high threshold, $p_b = 2.3$ (thin and opaque curves). Steady states indicated by black curves, are either stable (solid) or unstable (dashed). Orange and red dots indicate threshold location (fold bifurcation) of the respective systems. (c) Tipping probability contours for overshoots characterised by the time over the lowest threshold ($p_{thr} = 2$) and peak forcing amplitude, given a uniform distribution in the threshold location, $p_b \sim \mathcal{U}[2.0, 2.3]$. Purple colour gradient shows different probability boundary levels derived from the theory, equation (4). Black curves provide the probability levels calculated numerically. Blue cross corresponds to the time profile of external forcing given in (a) with the time over the lowest threshold represented by the green shading and peak in external forcing by the black arrow and dashed line.

We now illustrate and develop this idea further in Figure 1. Let us consider a single forcing profile of the form of equation (3), which starts at some initial level of forcing, smoothly increases to a peak level before returning back to the initial level at a mirrored rate as shown in Figure 1(a). Note again that the location of the tipping threshold will determine how far (if any) and how long the system will be beyond the tipping threshold for this single forcing profile. If the threshold is low (red dotted line) then the overshoot will be large and long whereas a higher threshold (orange dotted line) means that the overshoot will be smaller and for less time.

The contrasting consequences of a system having either a low or high threshold are demonstrated in Figure 1(b). The thick translucent curves correspond to a system with a low threshold ($p_b = 2$), which causes the system to tip due to the large and long overshoot. However, for a high threshold ($p_b = 2.3$), the thin and opaque curvesshow that tipping does not occur for the same forcing profile since the overshoot is now comparatively small and for a short duration.

In Figure 1(a), (b), we have seen that for a given overshoot forcing profile, tipping can either occur or not depending on the location of the tipping threshold. If we now extend this to a continuous range between these tipping threshold locations (initially all locations are assumed to be equally likely), then a tipping probability can be determined based on this arbitrarily chosen uniform distribution, $p_b \sim \mathcal{U}[2.0, 2.3]$.

For any given overshoot profile, there will be a critical threshold location that separates tipping (lower thresholds) from not tipping (higher thresholds). Therefore, the cumulative probability density function at this critical threshold gives the probability of tipping. This probability of tipping, or more precisely the critical tipping threshold location, can either be calculated numerically (using a bisection method) or via a modification to the inverse square law. However, since we now consider a range of tipping threshold locations, we adjust the relationship to consider the time over, $t_{over,thr}$, a prescribed threshold, $p_{thr} = 2$ (here defined to be the lowest tipping threshold of the uniform distribution) resulting in

$$t_{over,thr}^2 < \frac{4(p_{peak} - p_{thr})}{a_0^2 \kappa (p_{peak} - p_b)^2}. \tag{4}$$

A full derivation from equation (1) to equation (4) can be found in Appendix A1.

The probability of tipping is plotted in Figure 1(c) for a continuum of overshoot forcing profiles characterised by the time spent over the lowest threshold ($p_{thr} = 2$) and the peak in external forcing as indicated by the green shaded region and black arrow respectively in Figure 1(a). Note however, not all of these forcing profiles will result in an overshoot of the tipping threshold, particularly if the threshold is far away.

The black curves provide contours of constant probabilities of tipping as calculated by numerical simulations. The black dotted contour corresponds to a 1% probability of tipping (*exceptionally unlikely* in IPCC terminology (Masson-Delmotte et al., 2021)); dot-dash a 10% probability of tipping (*very unlikely*); dotted 25% probability of tipping; and solid 50% probability of tipping (tipping becomes *more likely than not*). The purple shaded regions correspond to the same probability intervals but calculated by the inverse square law theory, equation (4). The numerically calculated curves display a very good agreement to the theory.

The distance (along cross sections of either the peak forcing or time over threshold) between the tipping probability contours provides an indication of the uncertainty in the tipping behaviour. A large distance between the probability contours reflects a high uncertainty in tipping behaviour. Therefore, the performance of reducing uncertainty in the tipping behaviour will be determined by the ability to reduce the distance between the probability contours upon constraining uncertainty in the system characteristics.

For small peak levels in the external forcing there is a large uncertainty in the tipping behaviour. This is due to in some cases the peak forcing not even exceeding the threshold. Therefore, if the threshold is high, regardless of the time taken for the forcing to return, tipping will not occur. Though for lower thresholds, an overshoot of the threshold will occur and therefore the reversal in the forcing needs to be sufficiently quick to ensure tipping does not happen. For larger peaks there will be an overshoot in most cases and therefore the uncertainty in tipping behaviour is reduced as there is a maximum time that the system can spend beyond the threshold before tipping would ensue.

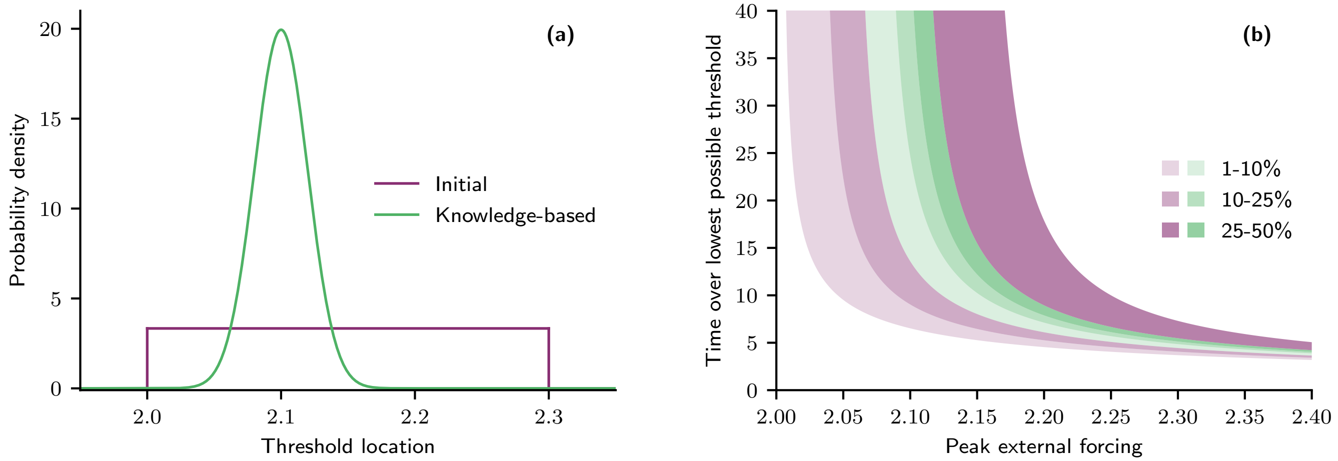

Figure 2. Constraining uncertainty in threshold location minimises uncertainty in tipping behaviour for overshoot scenarios. (a) Probability distribution functions for threshold location, p_b. A uniform distribution, $p_b \sim \mathcal{U}[2.0, 2.3]$, is used as the initial distribution (purple), whereas, the knowledge-based distribution is assumed to take the form of a normal distribution, $p_b \sim \mathcal{N}(2.1, 0.02^2)$ (green). (b) Theoretical tipping probability contours for overshoots characterised by the time over the lowest threshold ($p_{thr} = 2$) and peak external forcing, are given in colour corresponding to the distributions given in (a).

The blue cross corresponds to the overshoot profile given in panel (a), where the probability of tipping is close to the 1% boundary level for the range of tipping threshold locations considered. As previously discussed, the thick and translucent curves correspond to the lowest threshold and tipping is nearly avoided. Therefore, if the threshold is slightly higher, tipping would be prevented, as is the case for the highest threshold (thin opaque curves).

Reducing the uncertainty in the tipping threshold will constrain the uncertainty in the tipping behaviour, as shown by Figure 2. We introduce a knowledge-based distribution of the tipping threshold location, which is more constrained (for example through expert judgement or information inferred from data) than the initial, uninformed, uniform distribution. Figure 2(a) provides the comparison between the initial distribution in purple and the knowledge-based distribution, centred on a threshold location of 2.1, in green. The tipping probabilities (calculated theoretically using equation (4)) for the two different threshold distributions are given in Figure 2(b). Noticeably, the knowledge-based distribution corresponds to a much more constrained uncertainty in the tipping behaviour. For example, forcing profiles that were *very unlikely* (10% probability of tipping) for the initial distribution are now *exceptionally unlikely* ($< 1\%$) to occur given the knowledge-based distribution. Concurrently, the boundary for when tipping becomes *more likely than not* (50% level, right edge of darkest shaded region) shifts closer to the *exceptionally unlikely* (1% level, left edge of lightest shaded region) boundary and therefore signifying a reduction in the overall tipping uncertainty.

3.2 Results for uncertain linear restoring force

The previous section highlights how uncertainty in the location of the tipping threshold can affect the probability of tipping. Another important system characteristic for determining the mitigation window for overshoots is the strength of the linear restoring force – decay rate towards the stable equilibrium after making a perturbation to the system. In this section, we again use the prototypical fold model (2), but keep the tipping threshold location fixed ($p_b = 2$) and instead consider different linear restoring force proportionality factors κ.

Figure 3(a) provides an exemplary overshoot trajectory, as given by equation (3), that starts below the tipping threshold (indicated by the red dotted line), then increases such that there is a brief overshoot before reversing the forcing back to its original level. Similar to before, we can observe contrasting tipping behaviours for systems that differ only by the strength of the linear restoring force proportionality factor, see Figure 3(b). Namely, if the system has a weak linear restoring force ($\kappa = 1$) the system does not undergo tipping (thin and opaque blue curve). Whereas, if the restoring force is too strong tipping cannot be prevented. The example given by the thick and translucent curves nearly recovers but does not cross the unstable branch (representing the boundary of the basin of attraction), and so ultimately tips due to the restoring force being too strong ($\kappa = 2$). The system with the weaker linear restoring force has a weaker 'pull' towards the stable branch the system starts at, and therefore lags further behind the equilibrium of the static system than that of the system with the stronger restoring force. Consequently, when the system is forced beyond the critical threshold the system with weak restoring force will in effect take longer to realise it is over the edge and runaway to an alternative state. An additional advantageous effect of a weaker restoring force proportionality factor is that the boundary of the basin of attraction is further away (from the initial stable state). Therefore, the system with weak restoring force (thin and opaque black dashed curve) can cross the basin boundary at lower system state values compared to the system with a strong restoring force (thick and translucent black dashed curve). All these factors culminate in the system with strong restoring force tipping and the system with weak restoring force not tipping for the same forcing overshoot profile.

Following a similar approach to before, we can consider a range of restoring force proportionality parameter values that are uniformly distributed, $\kappa \sim \mathcal{U}[0.25, 03.25]$. For any given overshoot profile, according to equation (3), the critical κ can be determined and with this the probability of tipping equates to the proportion of the parameter distribution that is above this critical value. Figure 3(c) then plots the probability of tipping for a range of overshoot profiles, characterised by the time over the threshold (note unlike before this is now fixed) and the peak external forcing. As before the purple shading provides the probability intervals derived from the inverse square law theory (equation (1)) and the black curves with different linestyles are the numerically calculated boundaries (1%, 10%, 25%, 50% from dotted to solid). The numerically calculated curves again display a very good agreement to the theory especially for small and long overshoots. The blue cross corresponds to the particular overshoot profile given in panel (a), where the time over the threshold is indicated by the green shading and the peak external forcing by the black arrow and dashed lines.

Recall that for small peak external forcing levels the uncertainty in the tipping probability was large for uncertain tipping thresholds. In comparison, for uncertain restoring forces the tipping uncertainty is substantially smaller (compare Figure 1(c)

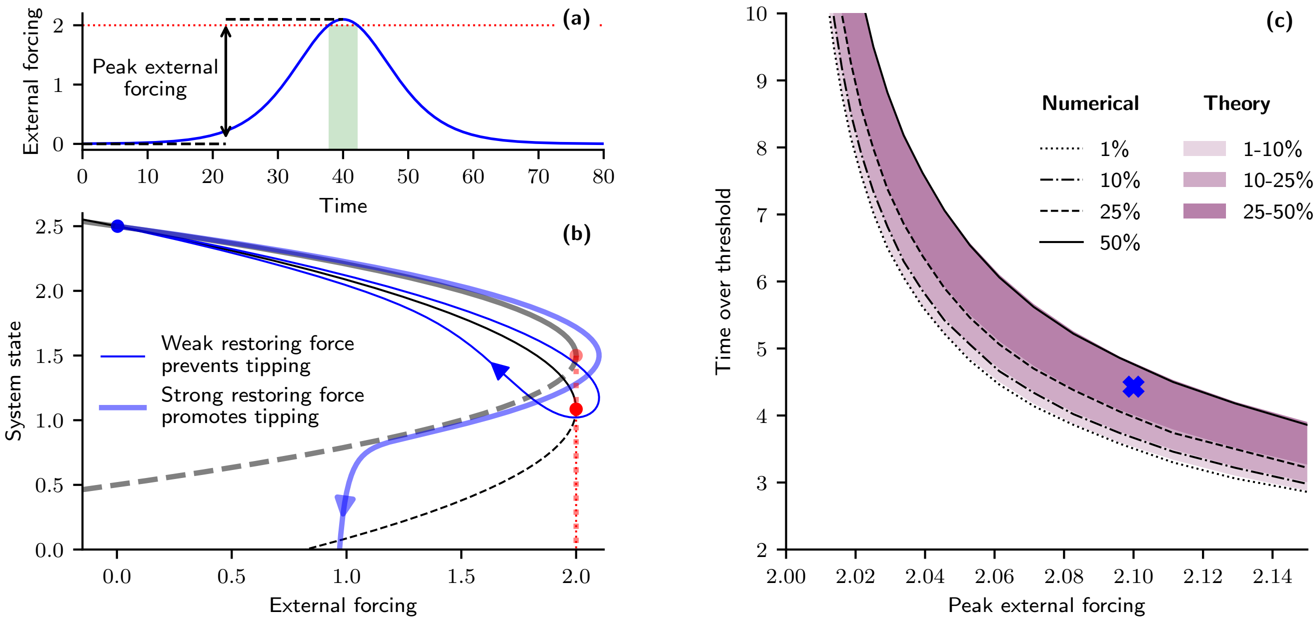

Figure 3. Probabilistic overshoots given uncertainty in strength of linear restoring force. (a) Time profile of an exemplar external forcing given by equation (3) (parameters $\Delta_p = 2.1$, $r = 0.1$, $t_{peak} = 40$). Red dotted line indicates tipping threshold location ($p_b = 2$). (b) System responses (blue) subjected to the external forcing profile given in (a) for model given by equation (2) with fixed $p_b = 2$, but either with weak restoring force, $\kappa = 1$ (thin and opaque) strong restoring force, $\kappa = 2$ (thick and translucent curves). Steady states indicated by black curves, are either stable (solid) or unstable (dashed). Red opaque and translucent dots indicate threshold location (fold bifurcation) of the respective systems. (c) Tipping probability contours for overshoots characterised by the time over the threshold ($p_b = 2$) and peak forcing amplitude, given a uniform distribution in the restoring force proportionality factor, $\kappa \sim \mathcal{U}[0.25, 3.25]$. Purple colour gradient shows different probability boundary levels derived from the theory, equation (1). Black curves provide the probability levels calculated numerically. Blue cross corresponds to the time profile of external forcing given in (a) with the time over the threshold represented by the green shading and peak in external forcing by the black arrow and dashed line.

and Figure 3(c)). For a given trajectory if the threshold is uncertain, then there may be no overshoot of the threshold, meaning the time to reverse the forcing is irrelevant given tipping is not possible (without noise). Whereas, if only the restoring force is uncertain, then it is known if the trajectory overshoots the threshold. Therefore, if it does overshoot, only a limited time can be spent over the threshold before tipping ensues.

Figure 4 shows how changing the uncertainty in the restoring force affects the tipping probability contours and therefore the uncertainty in tipping behaviour. In Figure 4(a) we again start with an initial uniform distribution for the restoring proportionality factor, i.e. $\kappa \sim \mathcal{U}[0.25, 3.25]$ (purple) and assume that the knowledge-based distribution narrows down this uncertainty. We assume a normal distribution with mean 1 and standard deviation 0.25, i.e. $\kappa \sim \mathcal{N}(1, 0.25^2)$ (green) .

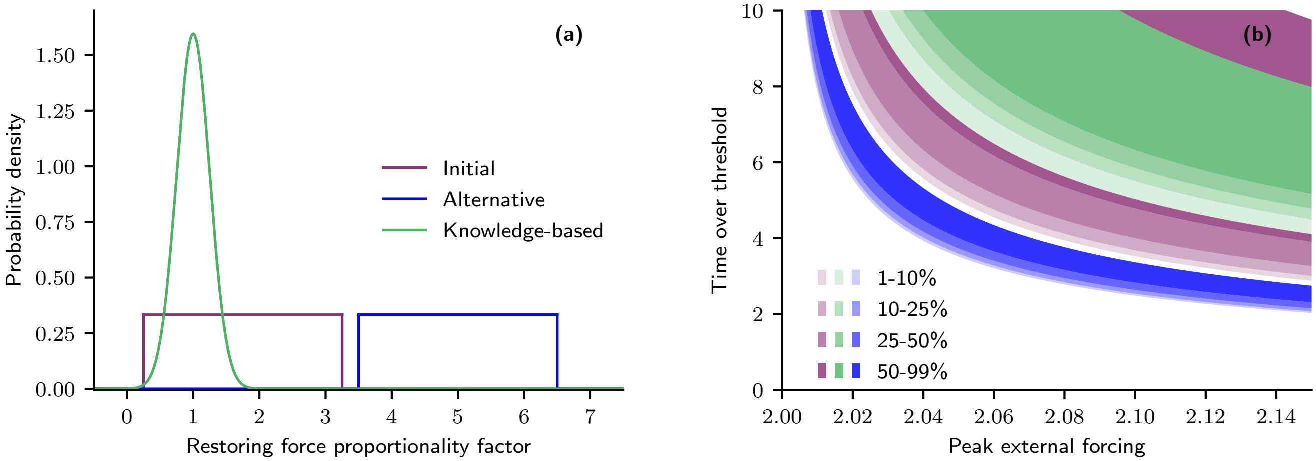

Figure 4. Constraining uncertainty in linear restoring force reduces uncertainty in tipping behaviour for overshoot scenarios. (a) Probability distribution functions for the restoring force proportionality factor, κ. A uniform distribution, $\kappa \sim \mathcal{U}[0.25, 3.25]$ is used as the initial distribution (purple), whereas, the knowledge-based distribution is assumed to take the form of a normal distribution, $\kappa \sim \mathcal{N}(1, 0.25^2)$ (green). An alternative initial uniform distribution, $\kappa \sim \mathcal{U}[3.5, 6.5]$ (blue) is also considered for same range of restoring force values but larger. (b) Theoretical tipping probability contours for overshoots characterised by the time over the threshold ($p_b = 22$) and peak external forcing, are given in colour corresponding to the distributions given in (a).

The curves for the different probability levels of tipping shift substantially, see Figure 4(b). An overshoot trajectory that sits on the 50% probability of tipping curve based on the initial distribution, would be considered to be *exceptionally unlikely* ($< 1\%$) to cause tipping given the knowledge-based distribution.

However, unlike for the threshold location, the distance (both horizontally and vertically) between the 1% and 50% tipping probability curves for the initial and knowledge-based distributions of the restoring force proportionality factor have barely changed. This seemingly counter-intuitive result can be explained by the change in the mean of the restoring force proportionality factor distribution counteracting the decrease in parameter uncertainty. To illustrate this, we consider an alternative uniform distribution, $\kappa \sim \mathcal{U}[3.5, 6.5]$ (blue), that is of the same width but has a much higher mean for the restoring force proportionality factor, see Figure 4(a). The distance between the 1% and 50% probability levels, in Figure 4(b) is much smaller for this alternative uniform distribution than the initial uniform distribution. This illustrates that an uncertainty in the restoring force for large values is less critical than at lower values. Thus, when transitioning from the initial uniform distribution to the knowledge-based distribution, the reduction in parameter uncertainty would narrow the distance between the 1% and 50% probability contours but by decreasing the mean simultaneously also widens the distance. So, for this example, little change in the distance between the 1% and 50% probability contours is observed. Importantly though, the uncertainty of the location of the critical boundary (separating tipping from not tipping) does still decrease. This can be seen by plotting the 99% probability of tipping boundary, and noticing that the separation between the 1% and 99% boundaries does indeed reduce.

So far, we have illustrated the effects of uncertainty in the tipping threshold location and the linear restoring force separately. We now come to a joint analysis of uncertainty, by considering uncertainty in a model parameter for a simple conceptual model of the Atlantic Meridional Overturning Circulation (AMOC). The uncertainty stems from the diffusive timescale parameter that jointly influences the location of the tipping threshold and the linear restoring force.

4 Results for uncertain diffusive timescale in the Stommel-Cessi model

In this section we consider how diffusive timescale uncertainties in a low-dimensional box model for the AMOC affects the uncertainty in tipping behaviour (i.e. if the AMOC collapses or not) for overshoot profiles of the freshwater flux into the North Atlantic.

The model, introduced by (Cessi, 1994), is a modification of the 2-box Stommel model (Stommel, 1961), and describes the change in the non-dimensional meridional salinity gradient, x – a proxy for the strength of the AMOC

$$\frac{\mathrm{d}x}{\mathrm{d}s} = p(s) - x(1 + \eta^2(1-x)^2), \tag{5}$$

where the parameter, $\eta^2 = t_d/t_a$, defines the ratio of the diffusive, t_d, to advective, t_a timescales and s is a time parameter. If the freshwater flux, $p(s)$, added to the North Atlantic becomes too large, it is possible to exceed a critical threshold, represented by a fold bifurcation in the model. This would cause the AMOC to tip from its current “on state” to a collapsed state, if exceeded for too long. Previously, Lux et al. (2022) showed how the location of this critical non-dimensional freshwater flux (as well as the restoring force) changes upon varying the ratio of the diffusive (mixing by wind driven gyres and eddies) to advective timescales. Specifically, the freshwater flux tipping threshold for AMOC collapse moves to lower values for smaller η^2. However, the scaling from the non-dimensional, p, to dimensional, $\tilde{F}$ (with units $[m/yr]$), freshwater flux depends on the diffusive timescale:

$$\tilde{F} = \frac{\alpha_T \theta H}{\alpha_S S_0 t_d} p = \frac{\xi}{t_d} p, \tag{6}$$

where a description of the parameters and their values can be found in the Appendix in Table A1. Therefore, depending on how the ratio, η^2, is changed (i.e. by either changing the diffusive and/or advective timescale), will affect how the tipping threshold in the dimensional freshwater flux changes. In the next subsection, we present numerical investigations for the interplay between changing t_d versus t_a, which will motivate our choice for considering uncertainties in the diffusive timescale.

4.1 Numerical results for joint influence of uncertain tipping threshold and uncertain linear restoring force

Figure 5 shows how the location of the dimensional critical freshwater fluxes (i.e. both the threshold indicating the transition to the collapsed off state and the threshold representing recovery back to the on state) and the width of the corresponding region of bistability (all denoted by colour) changes with varying the advective and diffusive timescales. The black lines provide contours of constant ratio between the diffusive and advective timescales. A sufficiently large ratio between the diffusive and advective timescales is required for the AMOC to possess a region of bistability. Hence, below this critical ratio ($\eta^2 = 3$), no tipping is possible as there exists no bistability region and hence no critical freshwater flux and so the corresponding region is coloured white.

Figure 5(a) shows the critical level of freshwater flux (denoted by colour), at which the AMOC on state terminates at a fold. Previously, the critical non-dimensional freshwater flux was shown to only depend on the ratio of timescales and that decreasing this ratio (either by decreasing the diffusive or increasing the advective timescales) moves the threshold to lower values (Lux et al., 2022). However, the scaling from dimensional to a non-dimensional freshwater flux, introduced by Cessi (1994) and given by equation (6), depends on the diffusive timescale. Hence, the tipping threshold for the dimensional freshwater flux no longer remains constant along the contours of constant ratio. Instead, the tipping threshold in the dimensional freshwater flux moves to lower values for increasing either the advective or diffusive timescale.

Similarly, the scaling dependency, equation (6), affects the location of the other fold bifurcation/threshold corresponding to the termination of the off state, see Figure 5(b). Lux et al. (2022) found that the freshwater flux threshold representing the transition from AMOC off to AMOC on was largely constant and independent of the ratio of timescales. Hence, for dimensional quantities this translates to the AMOC recovery threshold being independent of the advective timescale (i.e. for a fixed diffusive

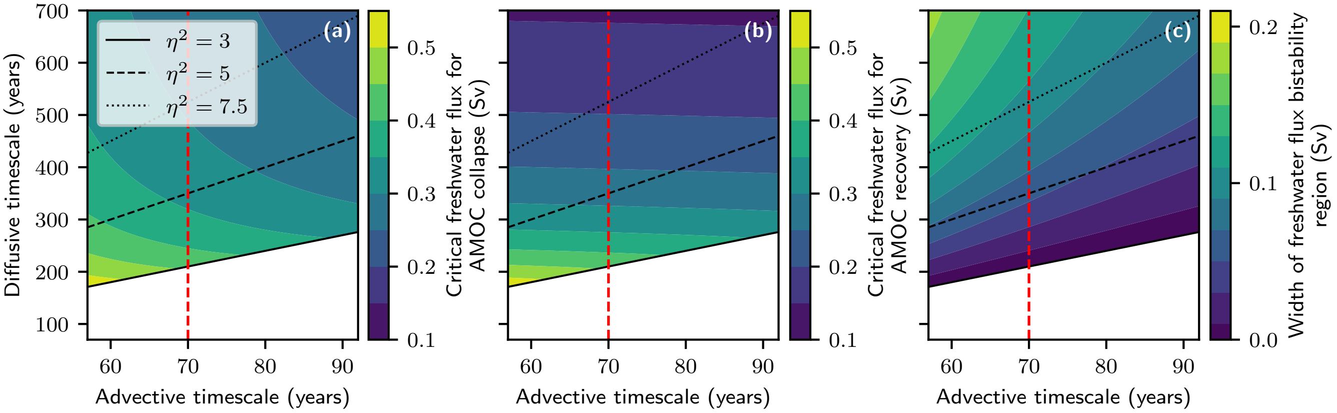

Figure 5. Critical freshwater fluxes and width of bistability region dependence on advective and diffusive timescales Colour plots for the location of the critical freshwater fluxes and the bistability region depending on the advective and diffusive timescales. (a) Location of the critical freshwater flux that triggers an AMOC collapse from the AMOC on state to the AMOC off state. (b) Location of the critical freshwater flux that triggers AMOC recovery from the off to the on state. (c) Width of bistability region, defined by the difference between the two critical freshwater fluxes. Region plotted corresponds to plausible advective and diffusive timescales as identified by (Wood et al., 2019). Black lines are contours of constant ratio between advective and diffusive timescales. Red dashed line denotes the value the advective timescale is fixed at for analysis of an uncertain diffusive timescale.

timescale, the threshold changes very little). On the other hand, the threshold decreases for increasing diffusive timescales – making it harder to restore the AMOC if it were to collapse.

The final panel, Figure 5(c), combines the first two panels by plotting in colour the difference between the two thresholds corresponding to the width of the region of bistability. Decreasing the diffusive timescale, but keeping the advective timescale fixed, decreases the width of bistability. Concurrently, both critical thresholds move to higher values, so these factors alone will make tipping less likely for any given overshoot. Additionally, the freshwater flux can be stabilised at a higher level, such that only the AMOC on state exists (since the threshold for AMOC recovery moves higher).

The chosen ranges of the advective and diffusive timescales correspond to the plausible ranges for the two different timescales, as determined by Wood et al. (2019). The advective timescale is relatively well constrained, whereas the uncertainty in the diffusive timescale is much larger. Consequently, we arbitrarily fix the advective timescale to 70 years, which is within the reasonable physical range given in Lux et al. (2022), and instead we will now focus on the uncertainty in the diffusive timescale (indicated by the red dashed line). Note that the scaling between the dimensional and non-dimensional freshwater flux in equation (6) (and the scaling of time), depends on the diffusive timescale. Therefore, the same non-dimensional freshwater flux time profile will translate into different dimensional time profiles for different diffusive timescales. This is the reason why we instead rescale time with respect to the advective timescale only, which changes equation (5) to

$$\dot{x} = p(t) - \frac{x}{t_d}(t_a + t_d(1-x)^2), \tag{7}$$

where now the scalings to the dimensional quantities, salinity difference ($\Delta S\,[psu]$), time ($t'\,[yrs]$), and freshwater flux (F now measured in Sverdrups $[Sv]$) are given by

$$\Delta S = \frac{\alpha_T \theta}{\alpha_S} x, \quad t' = t_a t, \quad F = \frac{\alpha_T \theta V_0 \gamma}{\alpha_S S_0 \beta t_a} p, \tag{8}$$

where again a description of the parameters and their values can be found in the Appendix in Table A1. The dimensional AMOC flow strength, $Q(x,t_d,V)$, with units $[S_v]$ is given by

$$Q(x,t_d,V) = \frac{\gamma V}{\beta t_a t_d}(t_a + t_d(1-x)^2), \tag{9}$$

where the ocean volume, V, is chosen such that the initial AMOC strength is equal to $Q_0 = Q(0,525,V_0)$. The reference volume V_0, given in Table A1, is an approximate value for the ocean volume based on General Circulation Models (Wood et al., 2019).

4.2 Results for uncertain diffusive timescale

We continue with analysing the AMOC tipping behaviour in the style of Figures 1 and 3. Rather than using symmetric freshwater overshoot profiles (as before), in Figure 6, we use more realistic profiles, first introduced by Huntingford et al. (2017), that take the form

$$p(t) = p_0 + \gamma t - (1 - \exp(-\mu(t)t))(\gamma t - (p_{stab} - p_0)). \tag{10}$$

Equation (10) allows the flexibility to start and finish at different levels ($p_0 = 0$, $p_{stab} = t_a/\xi$, which equates to stabilising at just below 0.25 Sv). The stabilisation level (p_{stab}) is chosen such that the freshwater always stabilises below the critical threshold for AMOC collapse, but above the threshold for AMOC recovery in most cases. The transition between the initial and stabilisation levels is determined by $\mu(t) = \mu_0 + \mu_1 t$, where μ_0 and μ_1 determine the maximum amplitude and time to converge to the stabilisation level. The parameter $\gamma = r - \mu_0(p_{stab} - p_0)$ is chosen to ensure that all profiles have the same initial rate of increase, determined by $r = 0.01$.

We now again follow a similar approach as for Figures 1 and 3 to investigate the effect of uncertainty in the diffusive timescale on the overshoots and the mitigation window. For the exemplary overshoot trajectory given in Figure 6(a), the system's response is shown for both a small ($t_d = 455$ years, thin and opaque curves) and large ($t_d = 700$ years, thick and translucent curves) diffusive timescale but fixed advective timescale ($t_a = 70$ years) in Figure 6(b). In the case of a small diffusive timescale, the tipping threshold would be far away (orange dotted lines), and so the overshoot would be small and for a short duration of time. Note also that the bistability region is small for small diffusive timescales and therefore, in this example ($t_d = 455$), the AMOC would recover (regardless of the overshoot time) if the freshwater flux is reduced back to below the lower fold at approximately 0.22 Sv.

In contrast, if the diffusive timescale is large, then the AMOC would collapse and not recover. This is by virtue of the threshold being a lot lower (red dotted lines) causing the overshoot of the tipping threshold to be much larger and for a longer period of time. These combined factors, coupled with a larger bistability region (stabilising within the bistability region) mean that the AMOC tips to its off state.

In Figure 6(c), we consider all plausible diffusive timescales (that also provide a region of bistability) with equal likelihood ($t_d \sim \mathcal{U}[210, 700]$) and plot the probability of tipping for overshoot profiles of the form given by equation 10. The overshoots are again characterised by the duration of time that the freshwater flux is above the lowest threshold (corresponding to a diffusive timescale of 700 years) and the peak freshwater flux indicated by the green shaded region and black arrow respectively in Figure 6(a). For sufficiently small diffusive timescales, recovery to the AMOC on state is guaranteed even if the AMOC temporarily collapses. This is a result of the freshwater flux stabilising below the bistability region, where the on state is the only stable equilibrium. For the uniform prior distribution, there is just under a 40% probability that the freshwater flux stabilises below the bistability region (but never above, which would guarantee tipping). Hence, we cannot rule out the possibility of AMOC recovery even for very large and long overshoots.

A good correlation is once again found between the inverse square law theory, equation (4), and the numerically calculated probability boundaries, particularly for the smaller peak freshwater overshoots. However, for larger overshoots, discrepancies arise between the numerics and theory. At the 1% level, the theory overestimates the critical boundary, caused by the asymmetry of the forcing profile. However, at the 50% level, the theory underestimates the critical boundary. The shrinking region of bistability (i.e. due to the presence of another fold bifurcation) and initialising the simulations in equilibrium provide additional sources of error. Specifically, strongly forced systems are not in equilibrium. Therefore, making the assumption that the AMOC is in equilibrium at the start of the simulation, will cause an overestimation of the numerical probability of tipping particularly for tipping thresholds that are close.

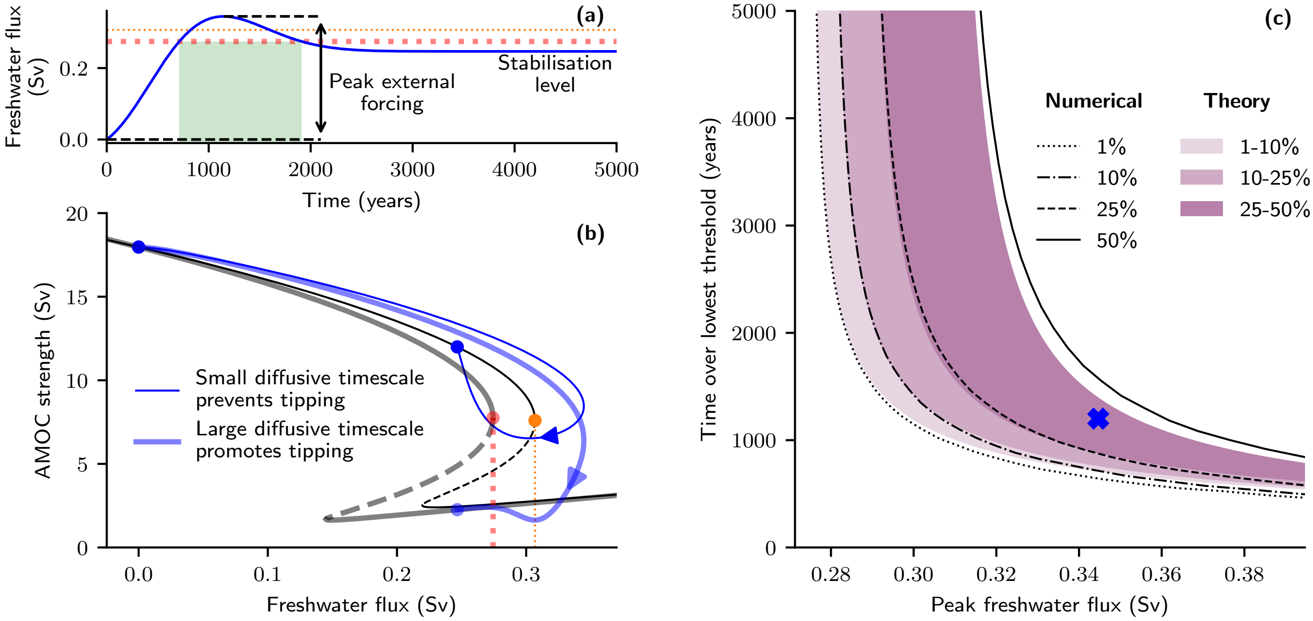

Figure 6. Probabilistic overshoots given uncertainty in diffusive timescale of AMOC Stommel-Cessi model. (a) Time profile of an exemplar freshwater flux given by equation (10) (parameters: $p_0 = 0$, $p_{\text{lim}} = 1/\xi$, $\mu_0 = -0.05$, $\mu_1 = 0.0057$, $r = 0.01$). Red and dotted lines indicate thresholds corresponding to a large ($t_d = 700$) and small ($t_d = 455$) diffusive timescale respectively. (b) System responses (blue) subjected to the freshwater flux profile given in (a) for model given by equation (7) and AMOC strength expressed by equation (9), for either a small diffusive timescale, $t_d = 455$ (thin and opaque curves) or large diffusive timescale, $t_d = 700$ (thick and translucent curves). Further system parameter values can be found in Table A1. Steady states indicated by black curves, are either stable (solid) or unstable (dashed). Orange and red dots indicate threshold location (fold bifurcation) of the respective systems. (c) Tipping probability contours for overshoots characterised by the time over the lowest threshold (corresponding to a diffusive timescale of 700 years) and peak forcing amplitude, given a uniform distribution in the diffusive timescale, $t_d \sim \mathcal{U}[210, 700]$. Purple colour gradient shows different probability mass levels derived from the theory, equation (4). Blue cross corresponds to the time profile of freshwater flux given in (a) with the time over the lowest threshold represented by the green shading and peak in external forcing by the black arrow and dashed line.

The large uncertainty in the tipping behaviour is once again caused by the large uncertainty in the system parameter, here the diffusive timescale. Therefore, it is necessary to constrain the uncertainty in the diffusive timescale to reduce the uncertainty in the tipping behaviour. The approach we use to constrain the uncertainty in the diffusive timescale is through Bayesian inference (Stuart, 2010). The procedure is analogous to the one used in Lux et al. (2022), especially in terms of the discrepancy model, the likelihood function, and the generation of the synthetic time series. Note that we use synthetically generated data due to the absence of real-world data for the AMOC to be matched to the Stommel-Cessi box model. We assume a true value for the diffusive timescale parameter $t_d = 525$ years and the underlying ODE is given by (7). We use a Markov chain Monte Carlo (MCMC) approach (Brooks et al., 2011), where the idea is to obtain the desired data-informed (posterior) distribution as the invariant distribution of the Markov chain over the prior support. We obtain the posterior distribution by running a MCMC algorithm provided in the MATLAB-based software framework UQLab (Marelli and Sudret, July 13-16, 2014), version 1.3.0. using the affine invariant ensemble sampler with 100 Markov chains with 400 steps (see the manual (Wagner et al., 2019) for a detailed documentation).

Performing Bayesian inference starting with the uniform prior distribution (purple) we are able to create a tightly constrained posterior distribution (green) centred close to the assumed diffusive timescale of 525 years, see Figure 7(a).

The tightly constrained posterior distribution results in a strong reduction in the uncertainty of the mitigation window, see comparison of purple to green in Figure 7(b). An overshoot that has a 25% probability of tipping, based on the prior distribution, would be classified as *exceptionally unlikely* with less than 1% probability of tipping given the posterior distribution. Furthermore, as can be inferred from Figure 5(b), the stabilisation level is within the bistability region (more than 99% confidence). Note that, the critical freshwater flux threshold, for AMOC recovery, is only below the stabilisation level for diffusive timescales greater than 400 years. Thus, for the posterior distribution the time taken to reverse the freshwater flux is critical to determine whether tipping occurs or not. Whereas previously, given the prior distribution knowledge, the probability for AMOC recovery would be non zero regardless of the time taken to reverse the freshwater flux.

Whereas the analysis in Figure 7(b) covers a whole spectrum of different overshoot trajectories, Figure 8 performs a more in-depth analysis of how the probability of tipping for a single overshoot trajectory changes based on the distribution of the diffusive timescale. A zoomed in view of the overshoot trajectory is given in Figure 8(a). For any overshoot there exists a critical diffusive timescale such that smaller diffusive timescales will prevent the AMOC from collapsing. The tipping threshold that corresponds to this critical diffusive timescale (for the particular overshoot given) is plotted in black. If the diffusive

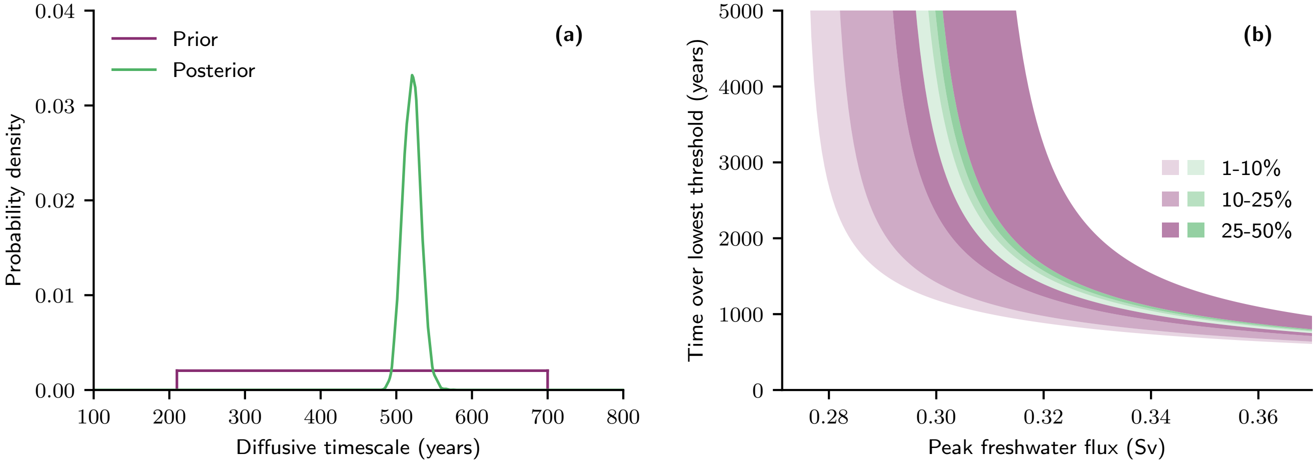

Figure 7. Constraining uncertainty in diffusive timescale minimises uncertainty in tipping behaviour for overshoot scenarios. (a) Probability distribution functions for diffusive timescale, t_d. A uniform distribution, $t_d \sim \mathcal{U}[210, 700]$, is used as a prior distribution (purple), whereas, the posterior distribution has been calculated by performing Bayesian inference on synthetic data generated for an assumed diffusive timescale of 525 years (green). (b) Theoretical tipping probability contours for overshoots characterised by the lowest possible threshold (given the prior distribution) and peak freshwater flux amplitude, are given in colour corresponding to the distributions given in (a).

timescale is smaller, the tipping threshold will be higher, meaning that the overshoot of the threshold will be smaller and for a shorter duration (compare orange line with black line). Let us now consider some uncertainty on the diffusive timescale, centred around a reference value $t_d^{\mathrm{ref}} = 450$ years that corresponds to the orange threshold $p_b^{\mathrm{ref}} = 0.308$ Sv. The orange banding represents the threshold locations p_b that arise from a nonlinear transformation of t_d within one standard deviation (125 years) of t_d^{ref}. Note that the nonlinear relation between the diffusive timescale parameter and the threshold location makes the orange band not symmetrically distributed around the orange line. Despite an assumed normal distribution on the diffusive timescale the distribution of thresholds is not normally distributed. Visibly, within one standard deviation of the mean includes both timescales above the critical level that would cause the AMOC to tip, and timescales that would avoid the system crossing the threshold altogether for the same overshoot trajectory.

In Figure 8(b) the probability of tipping is plotted based on the mean and standard deviation of the normally distributed diffusive timescale. The orange cross corresponds to the mean and standard deviation given in Figure 8(a). If the standard deviation of the distribution is zero (i.e. the diffusive timescale is known), then, without stochastic variability in the system, the probability of tipping is either zero or 100%. Increasing the standard deviation a little will create some tipping uncertainty close to the critical diffusive timescale for the specific trajectory, but still for most of the cross-section the probability of tipping will be close to zero or 100%. As the standard deviation increases further, more distributions will include the critical diffusive timescale with some non-small probability and therefore create greater tipping uncertainty. Thus, the region of tipping uncertainty spreads out for increasing standard deviation as shown numerically by the colouring.

The theoretical contours are added to Figure 8(b) as black lines of different line styles and show a good agreement to the numerically calculated probability given by the colour plot. The discrepancy arises from determining the value of the critical diffusive timescale. This is best identified on the x-axis for zero standard deviation, where the black contours converge roughly 10 years larger than the convergence of the colour.

Figure 8(b) shows that if the standard deviation is reduced, but the mean of the diffusive timescale kept fixed (i.e. follow the green arrow), then the uncertainty in the tipping behaviour reduces (probability of tipping moves further away from 50%). However, generally when reducing parameter uncertainty (standard deviation), the mean will also likely change. In some scenarios, it is possible for the uncertainty in the tipping behaviour to increase despite a reduced uncertainty in the timescale. For example, if we instead follow the red arrow then the probability of tipping changes from approximately 25% to 50%. However, here the orange line is moving down towards the black line and at the same time the banding is shrinking around the orange line. Therefore we are instead establishing that this particular overshoot is close to the critical overshoot that separates tipping from not tipping, but importantly, if all possible overshoot trajectories are considered then the overall uncertainty in the tipping behaviour will still reduce.

5 Conclusions

In this paper, we studied how uncertainty in model parameters can propagate to uncertainty in the tipping behaviour for systems subjected to overshoot trajectories. The location of the threshold and the linear restoring force are two key characteristics that were identified and individually isolated to examine their importance for the possibility of avoiding tipping. Specifically, we have found that the tipping behaviour from a single overshoot scenario can completely change based solely on either the location of the tipping threshold or the strength of the linear restoring force.

Uncertainty in the location of the tipping threshold was the characteristic found to have the most influence on the uncertainty in the tipping behaviour. For a given overshoot trajectory, the threshold location will simply determine if an overshoot of the threshold occurs. Assuming an overshoot of the threshold does occur, both the peak overshoot distance and the time spent over the threshold would be smaller for a high threshold compared to a low threshold. These properties of the overshoot feature in the left-hand side of the inverse square law, equation (1), while the right-hand side, interpreted as an upper bound for the mitigation window, remains fixed given a fixed linear restoring force. Constraining the uncertainty in the location of the

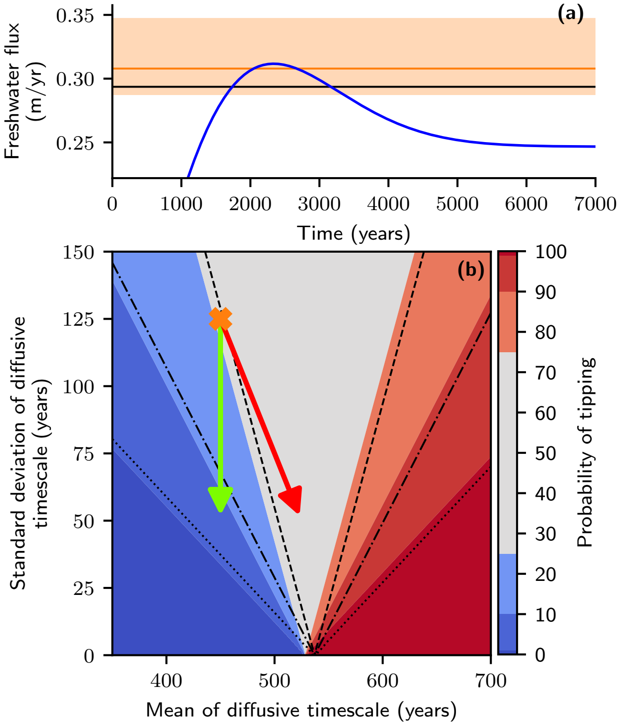

Figure 8. Probability of avoiding tipping for a single overshoot trajectory based on diffusive timescale distribution (a) Time profile of an exemplar overshoot trajectory (zoomed in) given by equation (10) (parameters: $p_0 = 0$, $p_{\lim} = 1/\xi$, $\mu_0 = 0.005$, $\mu_1 = 0.0009$, $r = 0.01$). The black horizontal line provides the location of the tipping threshold at the critical diffusive timescale separating tipping from not tipping. The orange line (mean) and banding (mean ± 1 standard deviation) show the location of the critical threshold for a diffusive timescale parameter distribution that is normally distributed with mean 450 years and standard deviation 125 years (denoted by orange cross in (b)). (b) Plot of the probability of tipping for the overshoot given in (a) depending on the characteristics of a normally distributed diffusive timescale parameter. Colour gives sampling based, numerically calculated probability of tipping and black contours the theoretical probability of tipping derived from inverse square law relationship given by equation (4).

tipping threshold will constrain the uncertainty in both the peak overshoot distance and duration of an overshoot and therefore considerably reduce the uncertainty in the tipping behaviour.

Another source of uncertainty can be in the strength of the linear restoring force to the stable equilibrium, which propagates into uncertainty in the upper bound for the mitigation window (the right-hand side of equation (1)). The linear restoring force features in the denominator of the right-hand side, and therefore a weaker restoring force will help prevent tipping by increasing the upper bound for the mitigation window. This can be intuitively understood by a weaker restoring force causing a system to further lag its stable equilibrium (of the static system) under a change in external forcing. Therefore reducing the uncertainty in the restoring force will reduce the uncertainty in the tipping delay from a system crossing its threshold and therefore how quickly the forcing needs to be reversed to prevent tipping.

We utilised a simple model for the Atlantic Meridional Overturning Circulation (AMOC) to demonstrate how uncertainty in the diffusive timescale parameter propagates simultaneously to uncertainty in the location of the threshold and the linear restoring force. Although the advective timescale is well constrained across climate models, a large uncertainty remains in the diffusive timescale. This ultimately results in uncertainty in the tipping behaviour for overshoot scenarios. For the AMOC, this translates to a large uncertainty in the probability of the AMOC collapsing.

Constraining parameter uncertainty, for instance by performing Bayesian inference on observational data, can greatly reduce the uncertainty in the tipping behaviour. For a parameter distribution with a known cumulative distribution function, the probability of tipping can be efficiently calculated in terms of computational costs by using the inverse square law relationship. This avoids a sampling based approach. Techniques presented in this manuscript might carry over to overshoots in more complex AMOC models such as the five box AMOC model from Wood et al. (2019). For this AMOC model, instead of using synthetic data for the Bayesian inference procedure, it is possible to use box-averaged time series data of general circulation model runs, which puts the parameter inference on a more realistic base. Note however, that a remaining challenge is to account for the presence of a Hopf bifurcation for some model parameter configurations, where tipping via the Hopf bifurcation (or rate-induced tipping) can occur before the system actually undergoes a fold bifurcation. An extension of the inverse square law theory would be required since it is currently designed to only consider overshoots of a fold bifurcation. Moreover, further research is required to understand how the theory can be applied to multiple uncertain parameters in more complex models.

Considering the added possibilities of tipping via other mechanisms can further change the conclusions. Note that the different mechanisms of tipping may also interact with each other: see the studies Ritchie and Sieber (2017); Slyman and Jones (2023) for the interplay between rate- and noise-induced tipping, and e.g. O'Keeffe and Wieczorek (2020); Alkhayuon et al. (2019) for the combination of rate- and bifurcation-induced tipping. For instance, if variability is considered (i.e. with the added possibility of noise-induced tipping), the longer a system spends close to, or beyond the threshold, the easier it is for a system to be triggered into tipping (Ritchie et al., 2019). This further emphasises that minimising the duration of any overshoot of a tipping threshold is paramount to preventing tipping. Furthermore, if rate-induced tipping was possible, then multiple critical rates can arise for the same peak external forcing (Ritchie et al., 2023) and therefore constraining system uncertainties becomes even more critical when considering overshoot scenarios.

Alternative profile shapes may also reduce the distinction between large but short and small but long overshoots (Enache et al., 2024). Moreover, uncertainties in the overshoot profile characteristics also need to be considered. For example, the precise peak overshoot distance and time spent over the tipping threshold is likely to be uncertain.

For a simple conceptual model of the AMOC, we find that for any sized overshoot, provided the duration is less than 800 years, tipping would very likely be avoided. Importantly, this encompasses most policy-relevant overshoots under consideration (see e.g. Kikstra et al. (2022)), and therefore this low dimensional box model would suggest AMOC tipping is very unlikely. However, it is important to note that these timescales are likely to be much longer than those observed in climate models (or the real world) (Jackson et al., 2022). Box models for the AMOC, such as that used for this study, tend to omit important advective responses that would otherwise make the response time faster (Jackson and Wood, 2018b).

Current rates of anthropogenic emissions make crossing climate tipping thresholds increasingly likely, despite not knowing their exact location. This study revealed the high influence of the tipping threshold location, and the strength of the linear restoring force, on the uncertainty in the tipping behaviour in response to possible overshoot trajectories. We have seen that constraining the uncertainty in these system characteristics enables us to better constrain the mitigation window, which is crucial if we want to avoid the tipping of elements of the climate system under overshoot scenarios.

Appendix A: Methods

A1 Overshoot theory for arbitrary threshold

The overshoot theory, as given by equation (1), details the time allowed over the tipping threshold. However, if the threshold location is uncertain, we would like to generalise this to the time over an arbitrary threshold, p_{thr} – in our case we use the lowest tipping threshold according to the initial distribution. We follow a similar approach used in the original derivation in Ritchie et al. (2019), starting with the overshoot theory given by:

$$p_{peak} - p_b < \frac{1}{a_0}\sqrt{-\frac{\ddot{p}(t_{peak})}{2\kappa}}. \tag{A1}$$

The Taylor expansion of the forcing profile, $p(t)$, about the peak level of forcing p_{peak} at time t_{peak}, is given by

$$p(t) \approx p_{peak} + \frac{1}{2}\ddot{p}(t_{peak})(t - t_{peak})^2. \tag{A2}$$

Using (A2), let us consider the time over, $t_{over,thr}$, a prescribed threshold, $p_{thr} < p_{peak}$:

$$t_{over,thr} = 2\sqrt{\frac{2(p_{peak} - p_{thr})}{-\ddot{p}(t_{peak})}}. \tag{A3}$$

Rearranging (A3) for $\ddot{p}(t_{peak})$ and substituting into (A1) gives the expression for the time allowed over an arbitrary threshold

$$t_{over,thr}^2 < \frac{4(p_{peak} - p_{thr})}{a_0^2 \kappa (p_{peak} - p_b)^2}, \tag{A4}$$

as given in the main text in equation (4). Importantly, if the threshold is chosen to be the tipping threshold ($p_{thr} = p_b$) then (A4) reduces to the original inverse square law, (1).

Author contributions. K.L-G. and P.D.L.R. contributed to the design of the study and wrote the manuscript. K.L-G. performed the Bayesian inference and P.D.L.R. carried out the numerical analysis.

Competing interests. There are no competing interests

Acknowledgements. K.L.-G. acknowledges support of the EU within the TiPES project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 820970., and funding by the Irène Curie fellowship. P.D.L.R. was supported by the European Research Council "Emergent Constraints on Climate-Land feedbacks in the Earth System (ECCLES)" project, grant agreement number 742472, and by the Optimal High Resolution Earth System Models for Exploring Future Climate Changes (OptimESM) project, grant agreement number 101081193. The authors would like to thank Richard Wood for his valuable comments that helped improving an earlier version.

Table A1. Description of parameters and their values used in the Stommel-Cessi model

Parameter	Description	Value (Units)
t_a	Adevective timescale	70 (yrs)
α_T	Thermal expansion coefficient	10^{-4} (K^{-1})
α_S	Haline contraction coefficient	7.6×10^{-4} (psu^{-1})
θ	Meridional temperature difference	25 (K)
H	Mean ocean depth	$4,500$ (m)
S_0	Reference salinity	35 (psu)
β	Seconds in a year	3.1536×10^7 $(s\ yr^{-1})$
γ	$m^3\ s^{-1}$ to Sv conversion	10^{-6} $(Sv\ s\ m^{-3})$
V_0	Reference volume	3.5×10^{16} (m^3)

References

H. Alkhayuon, P. Ashwin, L. C. Jackson, C. Quinn, and R. A. Wood. Basin bifurcations, oscillatory instability and rate-induced thresholds for atlantic meridional overturning circulation in a global oceanic box model. *Proceedings of the Royal Society A*, 475(2225):20190051, 2019.

D. I. Armstrong McKay, A. Staal, J. F. Abrams, R. Winkelmann, B. Sakschewski, S. Loriani, I. Fetzer, S. E. Cornell, J. Rockström, and T. M. Lenton. Exceeding 1.5 c global warming could trigger multiple climate tipping points. *Science*, 377(6611):eabn7950, 2022. https://doi.org/10.1126/science.abn7950. URL https://www.science.org/doi/abs/10.1126/science.abn7950.

P. Ashwin, S. Wieczorek, R. Vitolo, and P. Cox. Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1962): 1166–1184, 2012. https://doi.org/10.1098/rsta.2011.0306. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0306.

M. Ben-Yami, V. Skiba, S. Bathiany, and N. Boers. Uncertainties in critical slowing down indicators of observation-based fingerprints of the Atlantic Overturning Circulation. *Nature Communications*, 14(8344), 2023. https://doi.org/10.1038/s41467-023-44046-9.

N. Bochow, A. Poltronieri, A. Robinson, M. Rypdal, and N. Boers. Overshooting the critical threshold for the greenland ice sheet. *Nature*, 2023.

S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8. https://doi.org/10.1201/b10905.

P. Cessi. A simple box model of stochastically forced thermohaline flow. *Journal of Physical Oceanography*, 24(9):1911–1920, 1994.

E. Enache, O. Kozak, N. Wunderling, and J. Vollmer. Constraining safe and unsafe overshoots in saddle-node bifurcations. *arXiv preprint arXiv:2401.07712*, 2024.

C. Huntingford, H. Yang, A. Harper, P. M. Cox, N. Gedney, E. J. Burke, J. A. Lowe, G. Hayman, W. J. Collins, S. M. Smith, et al. Flexible parameter-sparse global temperature time profiles that stabilise at 1.5 and 2.0° c. *Earth System Dynamics*, 8(3):617–626, 2017.

L. Jackson and R. Wood. Hysteresis and resilience of the amoc in an eddy-permitting gcm. *Geophysical Research Letters*, 45(16):8547–8556, 2018a.

L. Jackson, R. Kahana, T. Graham, M. Ringer, T. Woollings, J. Mecking, and R. Wood. Global and european climate impacts of a slowdown of the amoc in a high resolution gcm. *Climate Dyn.*, 25:3299–3316, 2015. https://doi.org/10.1007/s00382-015-2540-2.

L. C. Jackson and R. A. Wood. Timescales of amoc decline in response to fresh water forcing. *Climate Dynamics*, 51(4):1333–1350, 2018b.

L. C. Jackson, E. Alastrué de Asenjo, K. Bellomo, G. Danabasoglu, H. Haak, A. Hu, J. Jungclaus, W. Lee, V. L. Meccia, O. Saenko, et al. Understanding amoc stability: the north atlantic hosing model intercomparison project. *Geoscientific Model Development Discussions*, 2022:1–32, 2022.

J. S. Kikstra, Z. R. J. Nicholls, C. J. Smith, J. Lewis, R. D. Lamboll, E. Byers, M. Sandstad, M. Meinshausen, M. J. Gidden, J. Rogelj, E. Kriegler, G. P. Peters, J. S. Fuglestvedt, R. B. Skeie, B. H. Samset, L. Wienpahl, D. P. van Vuuren, K.-I. van der Wijst, A. Al Khourdajie, P. M. Forster, A. Reisinger, R. Schaeffer, and K. Riahi. The ipcc sixth assessment report wgiii climate assessment of mitigation pathways: from emissions to global temperatures. *Geoscientific Model Development*, 15(24):9075–9109, 2022. https://doi.org/10.5194/gmd-15-9075-2022. URL https://gmd.copernicus.org/articles/15/9075/2022/.

C. Kuehn. A mathematical framework for critical transitions: normal forms, variance and applications. *J. Nonlinear Sci.*, 23(3): 457–510, 2013. ISSN 0938-8974. https://doi.org/10.1007/s00332-012-9158-x. URL https://doi-org.eaccess.ub.tum.de/10.1007/s00332-012-9158-x.

C. Kuehn and C. Bick. A universal route to explosive phenomena. *Science Advances*, 7(16), 2021. https://doi.org/10.1126/sciadv.abe3824.

Y. A. Kuznetsov. *Elements of applied bifurcation theory*, volume 112 of *Applied Mathematical Sciences*. Springer-Verlag, New York, third edition, 2004. ISBN 0-387-21906-4. https://doi.org/10.1007/978-1-4757-3978-7. URL https://doi-org.eaccess.ub.tum.de/10.1007/978-1-4757-3978-7.

T. Lenton, D. Armstrong McKay, S. Loriani, J. Abrams, S. Lade, J. F. Donges, M. Milkoreit, T. Powell, S. Smith, C. Zimm, et al. The global tipping points report 2023, 2023.

T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber. Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):1786–1793, 2008. ISSN 0027-8424. https://doi.org/10.1073/pnas.0705414105.

K. Lux, P. Ashwin, R. Wood, and C. Kuehn. Assessing the impact of parametric uncertainty on tipping points of the atlantic meridional overturning circulation. *Environmental Research Letters*, 17(7):075002, 2022.

J. Ma, Y. Xu, Y. Li, R. Tian, and J. Kurths. Predicting noise-induced critical transitions in bistable systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8):081102, 08 2019. ISSN 1054-1500. https://doi.org/10.1063/1.5115348. URL https://doi.org/10.1063/1.5115348.

S. Marelli and B. Sudret. UQLab: A Framework for Uncertainty Quantification in MATLAB. In *The 2nd International Conference on Vulnerability and Risk Analysis and Management (ICVRAM 2014)*, pages 2554–2563, University of Liverpool, United Kingdom, July 13-16, 2014. American Society of Civil Engineers. https://doi.org/10.1061/9780784413609.257.

V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2021.

A. L. Meyer, J. Bentley, R. C. Odoulami, A. L. Pigot, and C. H. Trisos. Risks to biodiversity from temperature overshoot pathways. *Philosophical Transactions of the Royal Society B*, 377(1857):20210394, 2022.

P. E. O'Keeffe and S. Wieczorek. Tipping phenomena and points of no return in ecosystems: beyond classical bifurcations. *SIAM Journal on Applied Dynamical Systems*, 19(4):2371–2402, 2020.

P. Ritchie and J. Sieber. Probability of noise- and rate-induced tipping. *Phys. Rev. E*, 95:052209, May 2017. https://doi.org/10.1103/PhysRevE.95.052209. URL https://link.aps.org/doi/10.1103/PhysRevE.95.052209.

P. Ritchie, O. Karabacak, and J. Sieber. Inverse-square law between time and amplitude for crossing tipping thresholds. *Proc. A.*, 475(2222): 20180504, 19, 2019. ISSN 1364-5021. https://doi.org/10.1098/rspa.2018.0504. URL https://doi.org/10.1098/rspa.2018.0504.

P. D. Ritchie, G. S. Smith, K. J. Davis, C. Fezzi, S. Halleck-Vega, A. B. Harper, C. A. Boulton, A. R. Binner, B. H. Day, A. V. Gallego-Sala, et al. Shifts in national land use and food production in great britain after a climate tipping point. *Nature Food*, 1(1):76–83, 2020.

P. D. Ritchie, J. J. Clarke, P. M. Cox, and C. Huntingford. Overshooting tipping point thresholds in a changing climate. *Nature*, 592(7855): 517–523, 2021.

P. D. L. Ritchie, H. Alkhayuon, P. M. Cox, and S. Wieczorek. Rate-induced tipping in natural and human systems. *Earth System Dynamics*, 14(3):669–683, 2023. https://doi.org/10.5194/esd-14-669-2023. URL https://esd.copernicus.org/articles/14/669/2023/.

M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. Van de Koppel, I. A. Van de Leemput, S. A. Levin, E. H. Van Nes, et al. Anticipating critical transitions. *science*, 338(6105):344–348, 2012.

K. Slyman and C. K. Jones. Rate and noise-induced tipping working in concert. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(1):013119, 01 2023. ISSN 1054-1500. https://doi.org/10.1063/5.0129341. URL https://doi.org/10.1063/5.0129341.

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

A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010. https://doi.org/10.1017/S0962492910000061.

P.-R. Wagner, J. Nagel, S. Marelli, and B. Sudret. UQLab user manual – Bayesian inversion for model calibration and validation. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, 2019. Report # UQLab-V1.3-113.

S. Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2003. ISBN 0-387-00177-8.

R. A. Wood, J. M. Rodríguez, R. S. Smith, L. C. Jackson, and E. Hawkins. Observable, low-order dynamical controls on thresholds of the atlantic meridional overturning circulation. *Climate Dynamics*, 53:6815–6834, 2019.

N. Wunderling, R. Winkelmann, J. Rockström, S. Loriani, D. I. Armstrong McKay, P. D. Ritchie, B. Sakschewski, and J. F. Donges. Global warming overshoots increase risks of climate tipping cascades in a network model. *Nature Climate Change*, 13(1):75–82, 2023.