Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation

Article

Accepted Version

Wood, R. A., Rodríguez, J. M., Smith, R. S. ORCID: https://orcid.org/0000-0001-7479-7778, Jackson, L. C. and Hawkins, E. ORCID: https://orcid.org/0000-0001-9477-3677 (2019) Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation. Climate Dynamics, 53 (11). pp. 6815-6834. ISSN 0930-7575 doi: https://doi.org/10.1007/s00382-019-04956-1 Available at https://centaur.reading.ac.uk/87278/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1007/s00382-019-04956-1

Publisher: Springer

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

University of Reading

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Observable, low-order dynamical controls on thresholds of the Atlantic Meridional Overturning Circulation

Richard A. Wood[1], José M. Rodríguez[1], Robin S. Smith[2], Laura C. Jackson[1] and Ed Hawkins[2]

[1] Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK
[2] National Centre for Atmospheric Science, University of Reading, Whiteknights, Reading RG6 7BE, UK

Acknowledgements:

RAW, JMR and LCJ were supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101).

Submitted to *Climate Dynamics*
October 2018
Revised July 2019

Corresponding author: Richard Wood
Email: richard.wood@metoffice.gov.uk
Telephone: +44(0)1392 886641
ORCID: 0000-0002-3960-9513

Abstract

We examine the dynamics of thresholds of the Atlantic Meridional Overturning Circulation (AMOC) in an Atmosphere-Ocean General Circulation Model (AOGCM) and a simple box model. We show that AMOC thresholds in the AOGCM are controlled by low-order dynamics encapsulated in the box model. In both models, AMOC collapse is primarily initiated by the development of a strong salinity advection feedback in the North Atlantic.

The box model parameters are potentially observable properties of the unperturbed (present day) ocean state, and when calibrated to a range of AOGCM states predict (within some error bars) the critical rate of fresh water input (H_{crit}) needed to turn off the AMOC in the AOGCM. In contrast, the meridional fresh water transport by the MOC (M_{OV}, a widely-used diagnostic of AMOC bi-stability) on its own is a poor predictor of H_{crit}.

When the AOGCM is run with increased atmospheric carbon dioxide, H_{crit} increases. We use the dynamical understanding from the box model to show that this increase is due partly to intensification of the global hydrological cycle and heat penetration into the near-surface ocean, both robust features of climate change projections. However changes in the gyre fresh water transport efficiency (a less robustly modelled process) are also important.

Key Words

Atlantic Meridional Overturning Circulation

Thresholds

Climate Change

Dynamics

Fresh water

1. Introduction

The Atlantic Meridional Overturning Circulation (AMOC) plays an important role in the climate of the Northern hemisphere through its transport of heat into the North Atlantic (Bryden and Imawaki 2001, Vellinga and Wood 2002, Jackson et al. 2015). Stommel (1961) identified the AMOC's potential to have multiple stable states, due to a simple salinity advection feedback mechanism. Beyond a certain threshold in the freshwater forcing of the North Atlantic, the AMOC becomes unsustainable and collapses. If freshwater forcing then returns to below the threshold value, the AMOC does not restart. If the AMOC were close to such a threshold, a small additional freshwater input to the Atlantic (e.g. from accelerated melting of the Greenland ice sheet) could trigger AMOC collapse (Fichefet et al. 2003).

Such theoretical AMOC behaviour has been demonstrated in a range of models, including more complex box models (e.g. Rahmstorf 1996, Lucarini and Stone 2005), intermediate complexity climate models (e.g. Rahmstorf et al. 2005, Lenton et al. 2007) and ocean general circulation models (GCMs) (Rahmstorf 1996, Dijkstra 2007, Hofmann and Rahmstorf 2009). It has also been proposed to be relevant to a number of transitions seen in the palaeoclimatic record (e.g. Alley 2003). Evidence of similar behaviour has been seen in some coupled atmosphere-ocean GCMs (AOGCMs) (Manabe and Stouffer 1988, Mikolajewicz et al. 2007), but due to computational constraints a full AMOC hysteresis curve has to date only been calculated for one, low resolution AOGCM (FAMOUS) for conditions of pre-industrial atmospheric carbon dioxide (CO_2) (Hawkins et al. 2011, hereafter H11). In H11 and many previous studies using simpler models, the thresholds are explored through a 'hosing' experiment in which a standard model equilibrium state is perturbed by adding an extra source of fresh water, H, to the North Atlantic. The strength of the hosing H is increased very slowly, with the aim of allowing the model to adjust towards its equilibrium state for each value of H. Hence a model run of several thousand years is required, and even then as shown in H11 a full equilibrium is not reached. Typically in such experiments, once H passes a critical value H_{crit} the AMOC collapses. H is then slowly reduced again, but in general the AMOC does not recover when H crosses back below H_{crit}. Instead AMOC recovery occurs at a lower (or even negative) value of H, giving a hysteresis in the AMOC strength and a range of values of H for which the AMOC is bistable (both strong and weak/reversed AMOC states are possible). Recently Jackson et al. (2017, hereafter J17) have analysed the detailed dynamics of the AMOC thresholds seen in the H11 study, showing that the salinity budget of the North Atlantic can be used to understand the dynamics of the thresholds.

The region of *H* values for which two stable states exist is bounded by bifurcation points beyond which either only the strong AMOC (small or negative *H*), or only the weak AMOC state (large *H*) is sustainable. Many studies have pointed to the importance of the fresh water budget of the Atlantic basin (north of 34°S) in determining the bistable region, and in particular the importance of the fresh water transport across 34°S due to the AMOC itself (denoted here by M_{OV}, *deVries and Weber 2005*, Drijfhout et al 2011). If $M_{OV}<0$ there is a positive salinity advection feedback in which negative anomalies in the AMOC induce a freshening of the Atlantic basin and hence further AMOC weakening. It has been suggested that current AOGCMs are biased towards an over-stable AMOC, due to a common positive bias in M_{OV} (e.g. Weber et al. 2007, Valdes 2010, Mecking et al. 2017). However Sijp (2012) pointed out that other feedbacks, specifically anomalous fresh water transports due to advection of salinity anomalies by the mean AMOC (<q>S') and the gyre/eddy components, are always stabilising, so $M_{OV}<0$ is not a sufficient condition for instability. It is therefore likely that the location of AMOC thresholds or bifurcation points is not simply determined by M_{OV}, but by a more complex set of feedbacks involving the fresh water budget of the Atlantic or North Atlantic basins. Recently Cheng et al (2018) have shown that in two AOGCM control runs the salinity advection feedback is *not* the dominant factor in variability of the North Atlantic AMOC, again emphasising the more complex nature of the processes controlling AMOC dynamics.

To quantify how far the AMOC is from a threshold, based on AOGCM hosing results, would require a wider range of AOGCM runs than is currently possible, although advances in computational power are beginning to enable a more thorough investigation of thresholds in current generation climate models including eddy-permitting ocean components (Jackson and Wood 2018). Dijkstra et al (2004) propose an alternative approach involving energetic analysis of the discrete GCM equations; however this involves a very large matrix inversion problem which is also likely to present computational challenges as model resolution and complexity increase. In this study we explore a new approach to quantifying AMOC thresholds: we hypothesise that AMOC thresholds are controlled by low-order dynamical processes which are quantitatively captured by a simple but physically-based box model. The box model structure is motivated by well-established understanding of the leading order water mass structure of the current AMOC. The crucial novelties of this model, compared to previous AMOC box models, are that the model is designed to represent a physically closed global circulation/water mass system, and that the model's control parameters can be simply determined from observable, large-scale properties of the present day ($H=0$) ocean state. Hence the box model cannot be 'tuned' to have a particular threshold – rather it is calibrated to the $H=0$ ocean state and *predicts* where the threshold H_{crit} will lie. To test the chosen dynamics of the box model we calibrate it to the unperturbed ocean state simulated using the FAMOUS AOGCM of H11 and J17. We demonstrate that the box model captures the leading mechanisms in the threshold dynamics of FAMOUS, as analysed by J17, particularly well for the first ('ramp-up') threshold in the hosing experiment described above. The box model dynamics are in this sense traceable to those of the AOGCM. Our calibration method implies that the present day ocean state contains sufficient information to determine the threshold hosing H_{crit} (to within errors which we quantify). We test this claim by repeating the H11 hosing experiment using a modified version of the AOGCM and various atmospheric CO_2 concentrations, yielding various values of H_{crit}. We calibrate the box model to the various baseline ($H=0$) AOGCM states and test its ability to predict the different values of H_{crit}.

The box model also provides a simple diagnostic framework that allows us to identify the key processes and ocean properties that determine the position of the AMOC threshold over a range of modelled states, and so acts as an 'emergent constraint' (e.g. Hall and Qu 2006, Cox et al. 2018), allowing the threshold position to be estimated by calibrating the box model to present day observations. Here (Section 6) we calibrate the box model to a data-assimilating ocean reanalysis to provide a preliminary estimate of H_{crit} for the present day ocean. However a more in-depth analysis would be needed to generate a robust estimate including error bars.

The question of whether increasing greenhouse gases will bring the AMOC closer to a threshold has not to date been directly addressed using AOGCMs. Schneider et al. (2007) concluded from a variety of studies (including expert elicitations) that increasing greenhouse gases will increase the likelihood of substantial AMOC responses. Drijfhout et al. (2011) studied the response of M_{OV} to increasing greenhouse gases, finding a complex response with M_{OV} generally decreasing and the strongest change at medium levels of greenhouse gas increase; however it is not clear whether M_{OV} has a close relationship to the threshold position, and they did not calculate the changes in AMOC thresholds explicitly. Here we directly calculate the AMOC hysteresis curve in FAMOUS, for a climate state with increased atmospheric CO_2. We find that for this AOGCM the amount of freshwater H_{crit} needed to provoke AMOC collapse is greater with elevated CO_2. This change is reproduced by the box model when we calibrate it to the higher CO_2 AOGCM state. We then use the dynamical understanding provided by the box model to assess whether this change is likely to be robust or merely an artefact of the particular AOGCM used.

Section 2 provides a brief description of the FAMOUS AOGCM, introduces the box model, and explains how the box model parameters are calibrated to the AOGCM state. Section 3 explores the processes behind AMOC thresholds in the AOGCM and box model, showing that the box model captures the essential dynamics of the AOGCM thresholds to within quantifiable errors. Section 4 explores the sensitivity of the AMOC collapse threshold to box model parameters, pointing to key features of the ocean state that determine the threshold position, and uses this insight to understand why H_{crit} increases under increased CO_2 in FAMOUS. Section 5 discusses limitations of the traceability between the box model and AOGCM. Section 6 draws together the results and discusses their implications for monitoring and early warning of AMOC thresholds, and the likely implications of climate change for future AMOC stability.

2. Model descriptions

2.1 The AOGCM

FAMOUS (Smith et al. 2008, Smith 2012) is a coarse resolution AOGCM based on the widely used HadCM3 model (Gordon et al. 2000). The atmospheric component has a horizontal resolution of $5^\circ \times 7.5^\circ$ with 11 vertical levels, while the ocean has a horizontal resolution of $2.5^\circ \times 3.75^\circ$ with 20 vertical levels. The model provides a three-dimensional simulation of atmosphere and ocean, with physically detailed representations of processes such as clouds, precipitation and atmosphere-ocean feedbacks. FAMOUS does not employ artificial flux adjustments, which are known to distort the AMOC hysteresis behaviour (Marotzke and Stone 1995, Dijkstra and Neelin 1999). We use two versions here: the first ['XDBUA', Smith et al. 2008, hereafter $FAMOUS_A$] is the version used by H11, while the second is an updated version including a range of minor changes [version 'XFXWB', Smith 2012, hereafter $FAMOUS_B$]. These model changes result in a change in the position of the AMOC threshold, and will provide an additional test of our model hierarchy.

2.2 The box model

Our box model is represented in Figure 1a. Its five boxes represent large contiguous regions of the global ocean, corresponding to large scale water mass structures (Talley et al 2011) (Figure 1b): the 'T' box represents the Atlantic thermocline; the 'N' box the North Atlantic Deep Water (NADW) formation region and Arctic; the 'B' box the southward propagating NADW and its upwelling in the Southern Ocean as Circumpolar Deep Water; the 'S' box fresh Southern Ocean near-surface waters and their return into the Atlantic as Antarctic Intermediate Water; and the 'IP' box the Indo-Pacific thermocline. The boxes are connected by pipes of negligible volume that carry the flow. The flow is separated into a 'cold water path’ (CWP), representing AMOC return flow via the South Pacific and Drake Passage, and a ‘warm water path’ (WWP), representing AMOC return via the Indo-Pacific thermocline and Agulhas leakage.

The box model physics is governed by salt conservation in each box, and a linear dependence of the overturning circulation on the density difference of the North Atlantic and Southern Ocean boxes:

$$q = \lambda\, [\alpha(T_S - T_N) + \beta(S_N - S_S)] \qquad (1)$$

where q is the AMOC flow and λ is a constant. A linear equation of state is used, with thermal and haline coefficients α=0.12 $kgm^{-3}K^{-1}$ and β=0.79 $kgm^{-3}(psu)^{-1}$. T and S denote mean temperature and salinity over the boxes. Such a relationship has previously been demonstrated in a range of models (e.g. Hughes and Weaver 1994, Rahmstorf 1996, Thorpe et al 2001, Sijp 2012), and we find it holds in our FAMOUS runs over the entire hysteresis loop described below (Figure 2a), justifying its use in our box model *a posteriori*.

The salinities of the five boxes are governed by salt conservation:

$q \geq 0$:

$$V_N \frac{\mathrm{d}\, S_N}{\mathrm{d}\, t} = q(S_T - S_N) + K_N(S_T - S_N) - F_N S_0 \qquad (2)$$

$$V_T \frac{\mathrm{d}\, S_T}{\mathrm{d}\, t} = q[\gamma S_S + (1-\gamma) S_{IP} - S_T] + K_S(S_S - S_T) + K_N(S_N - S_T) - F_T S_0 \qquad (3)$$

$$V_S \frac{\mathrm{d}\, S_S}{\mathrm{d}\, t} = \gamma q(S_B - S_S) + K_{IP}(S_{IP} - S_S) + K_S(S_T - S_S) + \eta(S_B - S_S) - F_S S_0 \qquad (4)$$

$$V_{IP} \frac{\mathrm{d}\, S_{IP}}{\mathrm{d}\, t} = (1-\gamma) q(S_B - S_{IP}) + K_{IP}(S_S - S_{IP}) - F_{IP} S_0 \qquad (5)$$

$$V_B \frac{\mathrm{d}\, S_B}{\mathrm{d}\, t} = q(S_N - S_B) + \eta(S_S - S_B) \qquad (6)$$

$q < 0$:

$$V_N \frac{\mathrm{d}\, S_N}{\mathrm{d}\, t} = |q|(S_B - S_N) + K_N(S_T - S_N) - F_N S_0 \tag{7}$$

$$V_T \frac{\mathrm{d}\, S_T}{\mathrm{d}\, t} = |q|(S_N - S_T) + K_S(S_S - S_T) + K_N(S_N - S_T) - F_T S_0 \tag{8}$$

$$V_S \frac{\mathrm{d}\, S_S}{\mathrm{d}\, t} = \gamma|q|(S_T - S_S) + K_{IP}(S_{IP} - S_S) + K_S(S_T - S_S) + \eta(S_B - S_S) - F_S S_0 \tag{9}$$

$$V_{IP} \frac{\mathrm{d}\, S_{IP}}{\mathrm{d}\, t} = (1-\gamma)|q|(S_T - S_{IP}) + K_{IP}(S_S - S_{IP}) - F_{IP} S_0 \tag{10}$$

$$V_B \frac{\mathrm{d}\, S_B}{\mathrm{d}\, t} = \gamma|q|S_S + (1-\gamma)|q|S_{IP} - |q|S_B + \eta(S_S - S_B) \tag{11}$$

where V_i is the volume of box i, γ denotes the proportion of the cold water path, and η is a S-B box mixing parameter, representing mixing of NADW with fresher waters as it passes around the global circulation. Oceanographically η represents the mixing of Circumpolar Deep Water with fresher surface water masses in the Southern Ocean (Talley et al. 2011). Wind driven salinity transports between boxes are represented by a diffusive flux with coefficients K_N , K_S, K_{IP} associated with the gyre strengths.

The box volumes V_i, gyre coefficients K_i, surface freshwater fluxes F_i, along with λ, η and γ are specified, time-invariant parameters. S_0 is a reference salinity set to 0.035. We assume that the mean temperature T_N of the North Atlantic box increases linearly with AMOC strength, reflecting the role of the AMOC in transporting heat into the North Atlantic:

$$T_N = \mu q + T_0 \tag{12}$$

The other box temperatures are fixed. While not as tight as the *q vs.* density relationship (1) over the whole hysteresis loop, there is nonetheless a close linear relationship between q and T_N, over the portion of the curve between the un-hosed state and the first threshold crossing, which is the part of the experiment which we will focus on in our analysis below (Figure 2b). We found empirically that allowing for this variation in T_N slightly increases the sharpness of the transition to the off state near the threshold, but temperature variations only play a minor role in density variations in these experiments (Figure 4a) and there is little sensitivity of H_{crit} to the value of μ (see discussion in Section 4.1). A more sophisticated treatment of temperature effects would be needed for thermally driven scenarios such as the response of the AMOC to transient global warming.

Our model adopts a similar broad approach to the box model of Rahmstorf (1996), but with several important additions:

i. Our model is designed to achieve a degree of quantitative, as well as qualitative agreement with corresponding AOGCM experiments. For this reason our boxes represent contiguous regions that span the majority of the global ocean, and are assigned different volumes that are identified with the largest scale water masses;
ii. The choice of separate N and B boxes was partly driven by the desire for quantitative comparison with the AOGCM: in an earlier prototype of the model where the N and B boxes were merged, the relationship between the density difference and MOC strength (Fig. 2a) was less tight, leading to large quantitative errors in the hysteresis loop. In the Rahmstorf model the B box (Rahmstorf's Box 4) is essentially passive and isolated (S_4=S_2 at equilibrium), whereas here we allow for mixing between the B box and the surface ocean (S box);
iii. Our model explicitly represents a closed global circulation and its associated fresh water transports, including the different roles of the cold and warm water paths. In contrast, in the Rahmstorf 1996 model the closure of the MOC outside the Atlantic basin (Rahmstorf's Box 1), and the role of gyre transports, must be specified through the concept of a fixed 'active fresh water flux' which is hard to associate with a specific observable quantity and does not respond to the evolving salinity fields. The additional physics in our model allows it to generate self-consistent solutions that can be identified with physical variables.

Our representation of the WWP/CWP has limitations: due to the large extent of the IP box the water coming back into the Atlantic basin through the WWP is not as saline as the real Agulhas return flow. Therefore our model may underestimate the importance of the WWP/CWP parameter γ. We note that for the parameter values studied here, variations in S_S and S_B are small compared to the other boxes. This means that a 3-box reduction of the model (with S_S and S_B fixed) is possible that contains the essential dynamical behaviour of the 5-box model in the most relevant parameter ranges, at the cost of some quantitative fidelity. Even the 3-box reduction has one extra degree of freedom compared with the Stommel 1961 and Rahmstorf 1996 models, allowing a much richer dynamical structure including homoclinic and Hopf bifurcations in addition to the saddle-node bifurcations that are seen in the simpler models (Alkhayuon et al. 2019).

Our model has several similarities to the model of Johnson et al (2007), which showed how more recent theories of the AMOC which emphasise closure of the potential energy budget through Southern Ocean winds and interior diapycnal mixing (e.g. Gnanadesikan 1999) can be reconciled with salinity-budget considerations and bistability as emphasised by the Stommel (1961) model. However our model differs from that of Johnson et al. 2007 in that we do not attempt to parametrise the processes that determine the transformation of NADW to cold, fresh Antarctic Intermediate Water or warm, salty thermocline water, and then solve for the pycnocline structure and AMOC. Instead in our model these transformations, and the basic geometry of the water masses are to some extent prescribed through the model parameters and the specified box boundaries. Our emphasis is on describing the dynamical mechanisms that occur when the AMOC passes from a strong ('on') state to a weak or reversed state (i.e. when the current strong AMOC state becomes unsustainable), on demonstrating that the box model dynamics accurately describe the dynamics of this transition in the AOGCM, and on identifying observable properties of the ocean circulation that determine where the transition lies.

2.3 Calibration of the box model to the AOGCM

To calibrate the box model to a GCM such as FAMOUS we use decadal mean variables diagnosed purely from large scale properties of the GCM's unperturbed equilibrium state (red dot in Figure 3c), without knowledge of the GCM's response to hosing. First, box boundaries are chosen to reflect approximate water mass boundaries in the GCM salinity field (Figure 1b). Once the box volumes are fixed, all but one of the control parameters of the box model can be diagnosed from emergent properties of FAMOUS (box average temperature and salinity, surface fluxes and section freshwater transports), and so could also in principle be diagnosed from observations. Box mean salinities, temperature and surface fresh water fluxes are obtained directly from the GCM. K_N, K_S and K_{IP} are determined by diagnosing the gyre salt transport M in the GCM across the corresponding box boundaries:

$$K_{ij} = (M \times 1000) / \rho_0 (S_i - S_j) \tag{13}$$

where ρ_0 is the mean seawater density. The K_{ij} above are in units of $m^3 s^{-1}$, M in $kg\ s^{-1}$ and the salinities in psu.

The flow constant λ is calculated from (1), after diagnosing q from the GCM as the maximum of the Atlantic overturning streamfunction at 30ºS.

The parameters μ and T_0 are calibrated by comparison with the North Pacific, a basin without a strong overturning circulation: we diagnose T_0 as the mean oceanic temperature of a full-depth box covering the North Pacific and choose μ to balance (12) using the diagnosed values of T_N and q. Finally γ, the proportion of the return AMOC flow carried by the cold water path, is chosen in the range $0 \leq \gamma \leq 1$ to optimise the model fit to the box average salinities in the GCM control state. We find γ in the range 0.39 to 0.85 in the cases considered here, somewhat larger than the values diagnosed directly from ocean GCMs by Döös (1995) and Speich et al. (2001). The sensitivity of the AMOC threshold to γ is discussed in Section 4. In this paper we calibrate the box model to a number of AOGCM states, discussed below. The resulting parameter values are shown in Table 1.

3. AMOC thresholds in the GCM and box model

3.1 Dynamics of the hysteresis

The AMOC hysteresis structure and thresholds were assessed in FAMOUS_A in a series of 'hosing' experiments by [*H11*]. A freshwater flux *H* was artificially applied to the North Atlantic surface between 20° N − 50° N. The same flux was removed uniformly from the rest of the ocean surface to conserve global salinity. The AMOC response is sensitive to the region to which *H* is applied (Smith and Gregory 2009), and other regions may be more appropriate if the goal were to simulate, say, additional fresh water discharge from the Greenland Ice Sheet (Swingedouw et al. 2015, Bakker et al. 2016). However our focus here is on elucidating the dynamics of the AMOC thresholds so we stick to a single region of application for consistency with the existing AOGCM experiment.

H was gradually increased at a rate of $5\text{x}10^{-4}$ Sv/year (1 Sv = 10^6 m^3s^{-1}), allowing the AMOC to adjust towards equilibrium with the hosing at any time. When *H* reached 1 Sv (after 2000 years), it was gradually reduced until it reached −0.4 Sv. In the period of increasing hosing, the AMOC collapsed when *H* reached about 0.55 Sv (Figure 3c, dotted curve). When *H* was reduced, the AMOC stayed collapsed, only recovering once *H* became less than about −0.1 Sv.

Even though H is increased and decreased slowly, the experiments do not capture fully equilibrated AMOC solutions. This was shown in H11, which demonstrated that the region of bistable equilibrium solutions in $FAMOUS_A$ is narrower than the hysteresis region that appears in response to the slow increase then decrease of H. However in what follows we adopt a pragmatic definition of the 'AMOC threshold' as the value H_{crit} of the additional freshwater flux H when the AMOC strength first reaches zero in the 'ramp-up' phase of the experiment (see dashed lines in Figure 3c). Further discussion of the response of the box model to time-varying H, including rate-dependent tipping responses, can be found in Alkhayuon et al. (2019).

The dynamics driving the AMOC thresholds in $FAMOUS_A$ are captured by the simple physics of the box model. When the same hosing experiment is performed with the box model calibrated to $FAMOUS_A$, box-average salinities in the regions represented by the box model evolve similarly in $FAMOUS_A$ and the box model (Figure 3a,b). The box model's AMOC shows hysteresis similar to that in $FAMOUS_A$ (Figure 3c), collapsing at a similar hosing value (0.48 Sv). Together the salinities and AMOC in the box model represent its full state vector. This strongly suggests that the dynamics of AMOC hysteresis in the AOGCM are described to leading order by the dynamics of the box model. This will be confirmed below by a comparison of the box model dynamics with the detailed analysis of the $FAMOUS_A$ run by J17.

We note that our measure of the AMOC in AOGCMs is the maximum (negative value) of the overturning streamfunction at 30ºS, which has been proposed as the key latitude at which the salinity advection feedback operates (e.g. Rahmstorf 1996, Drijfhout et al. 2011), rather than taking the maximum over the whole Atlantic, or around 30°N, as used by many previous studies. This explains why the $FAMOUS_A$ AMOC is negative in the collapsed state in Figure 3, rather than close to zero as shown in H11 and J17 (whose Figure 5a shows the maximum streamfunction at 26°N). The collapsed state in $FAMOUS_A$ has a reverse overturning cell that is largely confined to the South Atlantic and so not seen in the streamfuction at 26°N (see J17 Figure 3c or H11 Figure1). The use of 30°S gives a tighter and more linear relationship between the density difference and the AMOC (compare Figure 2a with Figure 5a of J17, which defines the AMOC at 26°N), and the relationship passes through the origin, whereas if 26°N were used an offset would need to be added to Equation (1) to obtain a good fit (J17), and it would be hard to calibrate the offset from the un-hosed state alone. The threshold values of H diagnosed for the AOGCM do not differ much whether either latitude is used (compare Figure 3c with Figure 2a of J17).

The agreement between box model and AOGCM is particularly good in the initial 'ramp-up' part of the hosing experiment, up to the point where the right-hand threshold is crossed (after about 1100 years, Figure 3), although the decline of the AMOC as *H* is increased is more gradual in the box model. We show in Section 5.3 below that the more gradual AMOC decline in the box model is a consequence of the limited vertical resolution of the box model, with surface fluxes being distributed over the full depth of the boxes. Once the collapsed AMOC state is established, changes in AOGCM water mass structure (*see* J17) result in larger quantitative differences between the box model and AOGCM solutions. We discuss these differences briefly in Section 5.2, but our focus in this paper is primarily on the 'ramp-up' stage and the right-hand threshold, as this is the most relevant for assessing the resilience of the current AMOC.

3.2 Detailed dynamics of the 'ramp-up' threshold

The AMOC threshold behaviour in the $FAMOUS_A$ experiment has been analysed in detail by J17, in terms of the salinity budget of the North Atlantic/Arctic from 40º - 90ºN, the same region as the N box in our box model calibration. AMOC changes in $FAMOUS_A$ are driven primarily by changes in the salinity component of density in this region. We therefore compare here the salinity budget of the N box (equations 2 and 7) with the corresponding budget in $FAMOUS_A$ from J17, as the right-hand threshold is crossed, to obtain a more detailed understanding of how well the box model captures the threshold dynamics of the AOGCM[1]. Having demonstrated very similar dynamics in the box model and AOGCM we exploit the simplicity of the box model to gain further insight into the threshold dynamics.

Figure 4a shows terms in the N box salinity budget for $FAMOUS_A$, during the 'ramp up' part of the experiment, adapted from J17. During most of the ramp-up phase the North Atlantic freshens slowly in response to the increasing hosing (red). However the freshening is partly offset by increasing salinification due to advection by the gyre component of the flow, which transports the fresh anomalies out across 40ºN (blue). Advection by the overturning component of the flow (green) is remarkably constant for most of the ramp-up phase. However as the threshold is approached (from about 800 years into the run) two factors act to accelerate the freshening. First, atmospheric feedbacks act to increase the surface fresh water flux into the North Atlantic (seen as a slight increase in the slope of the red line in Figure 4a from about t=800 years), attributed by J17 to a spinup of the Pacific MOC and

[1] The main $FAMOUS_A$ experiment, discussed here and in H11, is denoted SCOMP in J17. We briefly discuss a second $FAMOUS_A$ experiment, denoted VCOMP in J17, in Section 5.3 below.

consequent increase in inter-basin atmospheric water transport. Secondly a strong salinity advection feedback begins to operate, leading to a rapid decrease in the salinity advection by the overturning component of the flow (green line). These two processes lead to rapid freshening of the North Atlantic and collapse of the AMOC. The box model does not include the atmospheric feedback on fresh water fluxes since its surface fresh water flux is fixed. So the question arises whether this atmospheric feedback plays a critical qualitative or quantitative role in the AMOC threshold. Figure 4a suggests that the atmospheric feedback (which can be seen more clearly in Figure 6e of J17) is relatively small.

Figure 4b shows the corresponding salinity budget terms for the box model. We see quantitatively similar behaviour to $FAMOUS_A$ for all the budget terms, in the first 800 years. The salinity advection by the overturning is again roughly constant. From year 800, the box model surface fluxes do not include the atmospheric feedback described for $FAMOUS_A$ above. However the salinity advection by the MOC does decrease from this point in the box model just as in $FAMOUS_A$, leading to AMOC collapse. Hence the atmospheric feedback identified by J17 does not appear to be an essential element in the AMOC collapse, which instead is primarily due to the sudden collapse of the salinity advection by the MOC. However the atmospheric feedback may be expected to hasten the AMOC collapse, as suggested by J17. To confirm this we have rerun the box model with time-varying F_N diagnosed from the $FAMOUS_A$ run; the value of H_{crit} diagnosed with time-varying F_N is 0.40 Sv, compared with 0.48 Sv for the constant F_N case. The total fresh water input (hosing plus increase in F_N) at collapse is approximately the same in both cases, suggesting that the additional water input from the atmospheric feedback behaves simply as an additional hosing.

To elucidate the sudden reduction in the salinity advection by the MOC, we rewrite the salinity advection term in (2) by substituting for q from (1) and reformulating in terms of (S_T-S_N):

$$q(S_T-S_N) = \lambda[\alpha\,(T_S-T_N) + \beta(S_T-S_S)]\,(S_T - S_N) - \lambda\beta(S_T-S_N)^2 \quad (14)$$

Noting that over the first 800 years, salinity changes are dominated by changes in S_N (Figure 3b), we can approximate S_T and S_S as constant over this period. As S_T-S_N increases due to freshening of S_N, the $-\lambda\,\beta(S_T-S_N)^2$ term eventually dominates, resulting in the eventual rapid collapse of $q(S_T-S_N)$.

Note that $-q(S_T-S_N)$, the fresh water transport by the AMOC across 40ºN by the MOC, is the equivalent at 40ºN of the diagnostic commonly associated with AMOC stability through a linear salinity advection feedback argument (often referred to as M_{OV} or F_{OV}, e.g. Rahmstorf 1996, Mecking et al. 2017). We will use the notation $^{L}M_{OV}$ to denote M_{OV} at latitude L, where necessary for clarity. The linear feedback argument requires $^{L}M_{OV}$ to be negative at latitude L for the salinity advection feedback to become positive/destabilising at that latitude. However, as pointed out by Sijp (2012), what is important for stability is not M_{OV} but $\partial M_{OV}/\partial q$; positive $\partial M_{OV}/\partial q$ implies a negative (stabilising) feedback. In the initial phase (years 0-800), decreases in q are offset by increases in (S_T-S_N) as the hosing freshens the North Atlantic (Figure 4c). So although $^{40N}M_{OV}$ is negative in the initial state, the net salinity advection feedback $\partial^{40N}M_{OV}/\partial q$ is approximately zero until the $(S_T-S_N)^2$ term begins to dominate around year 800.

3.3 The 'ramp up' threshold in other AOGCM states

To test the ability of the box model to provide quantitative insight into the position of the right-hand threshold, we have performed two new hosing experiments with FAMOUS. For these we use the more recent model version $FAMOUS_B$. The baseline state for the first new experiment is the basic $FAMOUS_B$ model spun up from rest with pre-industrial CO_2 (Smith 2012), while for the second experiment CO_2 is doubled from pre-industrial values and the model is spun up for 920 years to adjust to the higher CO_2 forcing. We then repeat the hosing experiments, starting from these two new baseline states. The first of these experiments is identical to the experiment of H11, except for the use of $FAMOUS_B$ rather than $FAMOUS_A$, while the second experiment, also using $FAMOUS_B$, starts from a different climate state representing a climate with increased greenhouse gas concentrations.

First we repeat the 'ramp up' part of the hosing experiment using $FAMOUS_B$, with preindustrial CO_2. The model change from $FAMOUS_A$ to $FAMOUS_B$ results in a reduction of H_{crit} by about 0.1 Sv (Figure 5a). This change is captured by the box model when calibrated to the different climate states of the two FAMOUS versions (Figure 5b), providing further confidence in the box model. The different box model parameters for the $FAMOUS_A$ and $FAMOUS_B$ states are shown in Table 1.

As a further test of the ability of the box model to estimate H_{crit} for different ocean states, we have rerun the $FAMOUS_B$ hosing experiment, but now starting from a state reached after 920 years of integration at twice preindustrial CO_2. We find that around 0.35 Sv more freshwater input is needed to shut down the AMOC in the 2×CO_2 state, compared with the pre-industrial state (Figure 5a). The same simulation is done with the box model, recalibrated to the un-hosed 2×CO_2 state of $FAMOUS_B$. The box model response to increased CO_2 is qualitatively similar to that of $FAMOUS_B$, with 0.23 Sv more hosing required than in the preindustrial state (Figure 5b).

Overall the box model, when calibrated to different AOGCM states, appears to provide quantitative information on the value of H_{crit}. This implies that large scale, emergent properties of the unperturbed ocean state contain enough information to constrain H_{crit}. The simplicity of the box model allows us to understand the key factors and processes that determine H_{crit}., and we pursue this in Section 4 through a set of parameter sensitivity studies.

4. Parameter sensitivity of the box model

In this section we examine the sensitivity of the 'ramp-up' threshold H_{crit} to changes in individual box model parameters, and provide a physical interpretation of those sensitivities. We then discuss whether the fresh water transport by the AMOC in the baseline state (M_{OV}) is a good predictor of the value of H_{crit}, and assess the impact of the parameter changes seen at increased CO_2.

4.1 Parameter sensitivity of the threshold

Figure 6a shows the value of hosing H_{crit} at which q crosses zero in the ramp-up phase, as a function of the various box model parameters. Each parameter is varied individually with other parameters held fixed at their baseline values for the $FAMOUS_A$ experiment. Most parameters have been set to zero, one half and two times their baseline values, except where this did not make physical sense. We also varied the strength of the global atmospheric water cycle by simultaneously scaling all the surface fresh water fluxes F_i by 0.5 and 1.5 (thus mantaining zero global mean flux in each case).

The physical mechanisms of the different parameter sensitivities during the ramp-up phase can be understood in terms of the analysis of the fresh water budget of the North Atlantic (N box) in Section 3 above. Rewriting equation (1) as

$$q = \lambda\, [\alpha(T_S-T_0) + \beta(S_N-S_S)] / (1 + \lambda\alpha\mu) \qquad (15)$$

we see that the temperature driving of the flow is constant in time (and positive, Table 1). Figure 3a shows that the salinity driving is also initially positive (S_N>S_S), and that the freshening of S_N is much greater than variations in S_S during the ramp-up phase. As the hosing increases, S_N eventually becomes less than S_S (Figure 3a) and the salinity driving becomes sufficiently negative to counteract the temperature driving, giving $q=0$. We use this framework to interpret the parameter sensitivities in the following.

K_N: Higher values of K_N result in a larger H_{crit}. As K_N increases there is an increasingly strong negative feedback through salting of the N box by the gyre term as S_N freshens, counteracting and delaying the positive salinity advection feedback due to advection by the MOC ($\lambda\beta(S_T\text{-}S_N)^2$ in (14)). This can be seen by comparing the N box salinity budget in the case where $K_N=0$ (Figure 7a) with the corresponding figure in the baseline case (Figure 4b). Without the negative feedback from K_N the salinity advection feedback is much sharper (green line), leading to an earlier and more abrupt collapse of the AMOC. A similar sensitivity has recently been reported in simulations of the Last Glacial Maximum using the UVic intermediate complexity climate model (Muglia et al. 2018): applying the stronger North Atlantic wind stress typical of the LGM (equivalent to increasing the gyre strength and hence K_N) results in a stronger fresh water perturbation being required to shut down the AMOC.

K_S: Larger values of K_S result in a smaller H_{crit}. Increasing K_S increases S_S , and so reduces $(S_N – S_S)$ in the un-hosed state. Hence less freshening of S_N is needed to bring q to zero. This can be seen in Figure 7b, which shows the case with doubled K_S. The cases of doubled K_S and zero K_N (Figure 7a) therefore result in similar values of H_{crit} but for different physical reasons.

K_{IP}: Larger values of K_{IP} result in a smaller H_{crit}. This sensitivity is the only one where we find significant nonlinearity: it is particularly strong at low values of K_{IP} because as K_{IP} becomes small the only mechanism available to balance the net evaporation from the Indo-Pacific in (5) is the advective flux convergence $(1\text{-}\gamma)q(S_B\text{-}S_{IP})$. So as q decreases S_{IP} must increase rapidly to maintain the same advective flux convergence. This can be seen in the different evolution of S_{IP} in runs with low and high K_{IP} (Figure 8). For low K_{IP}, the rapid increase of S_{IP} results in a *negative* feedback on q: weakening q results in saltier Indo-Pacific water, which then enters the Atlantic via the warm water path. This negative feedback from the warm water path swamps the more commonly emphasised positive salinity advection feedback (e.g. Rahmstorf 1996); the positive feedback results from advection of the mean salinity by the anomalous flow (q'<S>), whereas the negative feedback that we identify here results from advection of anomalous salinity by the mean flow (<q>S', Sijp 2012). Advection of anomalous salinity was also found to make a significant contribution to the natural internal variability of M_{OV} and the AMOC in two modern AOGCMs by Cheng et al (2018). In the low K_{IP} situation it is likely that the consequent large increase in S_{IP} (Figure 8a) would result in

changes to the Indo-Pacific circulation (e.g. the Pacific MOC, see J17), with possible oceanic or atmospheric feedbacks that are not included in the box model. So the strong sensitivity to K_{IP} seen here may to some extent be an artefact of the limited Pacific Ocean and atmospheric processes in the box model.

T_S-T_0: Larger values imply stronger temperature driving of the flow. Hence greater freshening of S_N (stronger hosing) is needed to before the salinity gradient is strong enough to counteract the temperature gradient in (15).

μ: In this case as μ was varied, T_S-T_0 was adjusted to keep the same value of q in the baseline state. Larger values of μ imply larger values of T_S-T_0 ,and hence the same sign of sensitivity as was seen to T_S-T_0.. If μ is instead changed without adjusting T_S-T_0, there is virtually no sensitivity of H_{crit} to μ, since the amount of North Atlantic freshening (hosing) required to bring the density gradient to zero in (15) is not directly changed. Thus the apparent sensitivity to μ is mostly due to sensitivity to the invariant part of the temperature gradient T_S-T_0.

λ: The sensitivity is weak because a change in λ does not directly change the North Atlantic freshening (hosing) needed to bring the N-S density difference to zero in (15). Although increased λ produces a stronger baseline flow, there is a balancing change in the amount that q changes for a given density change.

η: Sensitivity to η is weak. η effectively relaxes S_S toward the salinity of the large deep water reservoir S_B, resulting the small variation in S_S seen in the baseline experiment (Figure 3a). For small η, S_S is free to vary more in response to advection by the changing q, but these salinity variations are simply advected around the CWP and cause corresponding changes in S_T and S_N. So the overall variations in $(S_N$-$S_S)$ in (15) are not much different from the baseline case.

γ: Larger values of γ have smaller values of H_{crit}. Large values of γ imply a dominant CWP. In this case the Atlantic is fresher and the Southern Ocean saltier than in the low γ (WWP) case. In terms of (15), $(S_N$-$S_S)$ begins at a lower value and so less freshening is required to reverse the density gradient.

F_i: Here all the surface fresh water fluxes are scaled by a factor of 0.5 or 1.5, maintaining zero global mean flux in each case. A stronger mean hydrological cycle results in a larger initial salinity difference $(S_N\text{-}S_S)$ in (15). Hence more hosing is needed to reverse the density gradient, and larger fresh water fluxes result in a larger H_{crit}.

Overall, we see that H_{crit} is sensitive to many of the box model parameters, including those involving the thermohaline forcing ($T_S\text{-}T_0$, F_i, μ), and those involving wind-driven gyre exchange (K_i). It is perhaps surprising (but explained by the analysis above) that the sensitivity to parameters involving internal dynamics of the AMOC (λ, γ, η) is relatively weak. The parameter sensitivity is generally linear in the range considered, except for K_{IP}, where the strong nonlinearity at low values may be a consequence of the simplicity of the box model dynamics.

4.2 Role of the AMOC fresh water transport M_{OV}

The fresh water transport into the Atlantic basin across the southern boundary of the basin (around 34°S) by the AMOC itself (often denoted M_{OV} or F_{OV}) has been proposed as an important diagnostic of AMOC bi-stability at equilibrium, with negative M_{OV} implying that the AMOC is in a bi-stable regime, and positive M_{OV} implying a mono-stable AMOC (Rahmstorf 1996; deVries and Weber 2005; Mecking et al. 2017). M_{OV} also plays a role in the transient response of the AMOC to hosing: modifying M_{OV} by applying flux adjustments at the Southern boundary or throughout the Atlantic can change the response of the AMOC in AOGCM hosing experiments (Cimatoribus et al. 2012, Jackson 2013, Liu et al. 2017). The sign of M_{OV} has been associated with the sign of the salinity advection feedback, with positive M_{OV} implying a negative (stabilising) feedback and negative M_{OV} implying a positive (destabilising) feedback on AMOC changes (Stommel 1961, Rahmstorf 1996). However the relationship between the role of M_{OV} in AMOC bistability (a property of the equilibrium state) and the salinity advection feedback (a transient process) is unclear.

The role of M_{OV} in AMOC feedbacks and stability was shown by Sijp (2012) to be more complicated than the above advection feedback argument. In the standard argument a negative M_{OV} at a given latitude implies that the AMOC is removing fresh water from the Atlantic basin north of that latitude. A weakening of the AMOC leads to less fresh water removal and hence a fresher Atlantic basin and further AMOC weakening. This feedback focuses on fresh water transport anomalies arising from advection of the mean salinity field by the anomalous flow ($q'\langle S\rangle$); however as noted by Sijp (2012), advection of salinity anomalies by the mean flow ($\langle q\rangle S'$) can also be an important term, is stabilising whatever the sign of M_{OV} in the un-hosed state, and can be larger than the first term. A compensation between these two terms can be seen (for M_{OV} at 40ºN) in Figure 4c. Further, the gyre/eddy components of fresh water transport are always down-gradient and are expected to be stabilising. Hence there are both stabilising and destabilising feedbacks, and a stable AMOC is possible even when $M_{OV} < 0$, as is believed to be the case in the real present-day ocean.

Given the theoretical importance of and interest in M_{OV} as a diagnostic of AMOC bi-stability, we ask whether M_{OV} in the un-hosed state contains any information about the distance of the AMOC from the right hand stability threshold, H_{crit}. This distance does not *a priori* depend on whether the unperturbed AMOC is in a mono- or bi-stable régime. Our box model does not contain a physical boundary at 34°S, so we examine three alternative definitions of the fresh water transport by the AMOC into the Atlantic basin:

$$N_{OV} = -q\ (S_T\text{-}S_N)\ /\ S_0 \qquad (16)$$

is the transport into the N box (equivalent to the value of M_{OV} at around 40°N in FAMOUS, and close to the North Atlantic region used for analysis of the $\mathrm{FAMOUS_A}$ run in J17);

$$T_{OV} = -q\ [(\gamma(S_S + (1-\gamma)S_{IP}\ - S_N]\ /\ S_0 \qquad (17)$$

is the transport into the combined T and N boxes (North Atlantic above the NADW layer); and

$$B_{OV} = -q\ [(\gamma(S_S - S_B) + (1-\gamma)(S_{IP} - S_B)]\ /\ S_0 \qquad (18)$$

is the transport into the combined T, N and B boxes (whole Atlantic plus the global NADW/CDW water mass). B_{OV} is the closest box model equivalent to the conventional ${}^{34S}\mathrm{M_{OV}}$, if we assume that the southward transport across 34°S is qS_B. The first term on the right hand side is positive, representing northward fresh water transport by the CWP, and the second term is negative, representing southward transport by the WWP.

The dependence of H_{crit} on the un-hosed value of N_{OV}, T_{OV} and B_{OV}, for the box model parameter sensitivity experiments described above, is shown in Figure 6b. We see that none of these diagnostics has a clear relationship with H_{crit} overall. This is unsurprising given the variety of mechanisms by which parameter changes result in changes in H_{crit}, as discussed in Section 4.1. For example, the sensitivity of H_{crit} to K_N is a consequence of changes in N_{OV} (see discussion in Section 4.1 and Figure 7a), and the 'expected' relationship between H_{crit} and N_{OV} (i.e. larger H_{crit} as N_{OV} increases) is seen in Figure 6b. On the other hand, the sensitivity of H_{crit} to K_{IP} is primarily due to changes in the salinity of the Indo-Pacific water (Section 4.1), and we see large changes in H_{crit} in response to changes in K_{IP}, despite only small changes in the un-hosed value of any of N_{OV}, T_{OV} and B_{OV} (Figure 6b).

Overall we conclude that while the advection of fresh water by the AMOC (quantified by M_{OV}) plays an important role in the stability of the AMOC, the distance of the unperturbed AMOC from the threshold (H_{crit}) is sensitive to a number of processes, so that the unperturbed value of M_{OV} does not in itself provide a reliable indicator of H_{crit}.

4.3 Parameter changes at increased CO_2 concentration

Comparing the two $FAMOUS_B$ experiments with pre-industrial and doubled CO_2, we see that increased CO_2 results in an increase in H_{crit} by several tenths of a Sverdrup. The different box model parameters for the two states are given in Table 1, and we have performed further box model parameter sensitivity studies changing each of these parameters individually from its $1{\times}CO_2$ to its $2{\times}CO_2$ value, to determine the main causes of the threshold shift under increased CO_2. From these sensitivity studies we find that the dominant factors contributing to the increase in H_{crit} are:

a) An increase in the average temperature difference between the North Pacific and the S box, T_S-T_0. Causes increase in H_{crit} of 0.16 *Sv*.
b) an increase in the overall strength of the global water cycle, particularly an increase in net Atlantic evaporation $-(F_N + F_T)$. Causes increase in H_{crit} of 0.12 *Sv*.
c) changes in the efficiency of the 'gyre' freshwater transports in the Atlantic (K_S, K_N). These roughly cancel, leaving an overall increase in H_{crit} of 0.02 *Sv*.

The enhanced atmospheric water cycle at increased CO_2 (b) is a robust feature of climate model simulations (Collins et al 2013). The increase in $T_S - T_0$ (a) is also likely to be a robust result: most of the ocean warming occurs in the upper layers (*cf.* Gregory 2000, Landerer et al. 2007), so for the same change in heat content the box-mean temperature T_S (covering only the top 1000m or so of the ocean) changes more than T_0 (for which a full-depth North Pacific box is used). Changes in gyre transports (c) are less well understood.

To explore whether the increase in H_{crit} with increasing CO_2 is likely to be robust, we have calibrated the box model to the more recent (CMIP5-generation) AOGCM HadGEM2-AO (Martin et al. 2011), in quasi-equilibrium states with 1×, 2×, and 4× pre-industrial CO_2, and performed hosing experiments to determine H_{crit}. Parameter values for these three calibrations are given in Table 1. For HadGEM2-AO we find that H_{crit} increases by 0.27 Sv and 0.43 Sv at 2×, and $4{\times}CO_2$ respectively, compared to the $1{\times}CO_2$ state (Fig. 5c). As was seen for $FAMOUS_B$, a strengthened fresh water cycle (b) and increased temperature driving (a) both contribute to the increase in H_{crit} ; however for the HadGEM2-AO calibrations, increases in K_N dominate the changes in the ‘gyre’ components (c), and make a large contribution to the increase in H_{crit}. Changes to gyre exchange are less well understood than the other factors above so more uncertainty remains about this contribution. We also see a flattening of the response curve, with a less sharp threshold at higher CO_2 in HadGEM2 but not in $FAMOUS_B$. Through single-parameter perturbation experiments (not shown), we find that the flattening is due to the increase of K_N at higher CO_2, in HadGEM2.

5. Limits of traceability

An advantage of our box modelling approach is that since all the box model state variables and control parameters can be diagnosed directly from GCM solutions (and in principle from observations), the box model provides a low order dynamical framework to analyse the GCM; we can examine discrepancies between the box model and GCM solutions directly, and so understand where the box model breaks down. Indeed we used this process in the development of the box model. For example an earlier, four-box version of the model treated the N and B boxes as a single box. While this provided solutions that were qualitatively similar to the GCM, quite large quantitative discrepancies arose, and diagnosis of the discrepancies pointed to the relationship between density and circulation strength (1), which was not as tight as in Figure 2a when the density of the merged N and B boxes was used rather than the N box alone. In this section we examine aspects of the solution where quantitative agreement between box model and GCM solutions remains less good, and diagnose the reasons behind these discrepancies.

5.1 Atmospheric fresh water feedbacks

As discussed in Section 3 above and in J17, the climate variations associated with AMOC changes through the $FAMOUS_A$ hosing experiment result in a slight increase in the surface fresh water flux into the North Atlantic, which accelerates the AMOC weakening. This atmospheric feedback is not included in our box model but by re-running the box model using the time-dependent surface fluxes diagnosed from the $FAMOUS_A$ run we assessed that the atmospheric feedback reduces the value of H_{crit} by about 0.08 *Sv* in $FAMOUS_A$. In principle the atmospheric feedback could be parametrised in the box model. However, when we assessed the impact of the feedback in the same way for the $FAMOUS_B$ 2xCO_2 run we found that in this case it resulted in an *increase* in H_{crit} (again by around 0.08 *Sv*). This suggests that the atmospheric feedback on fresh water flux may be noisy and/or difficult to parametrise, so we do not attempt this here but rather consider it an error term in the box model leading to an uncertainty of ±0.08 *Sv* in H_{crit} as estimated by the box model.

5.2 Left hand threshold

We note that in Figure 3 the left hand ('ramp down') threshold appears to be less accurately captured than the right hand ('ramp up') threshold. This can be understood as an inherent limitation of the box model, based on the analysis of $FAMOUS_A$ by J17. J17 interpreted the AMOC recovery in the ramp-down phase in terms of the North Atlantic salinity budget, as for the ramp up phase. The AMOC-off state and ramp down phase are characterised by a weak reverse overturning circulation (-4 Sv at 26°N), and the recovery is driven by advection of salinity anomalies by this circulation. However in the *South* Atlantic the reverse overturning circulation in the off state is much stronger (-8 Sv, see Figure 3 and J17 Figure 3c). The box model does not differentiate between the AMOC in the North and South Atlantic, and its 'off' state has a strong reverse circulation (-14 Sv) which extends into the North Atlantic boxes, introducing quantitative errors in the salinity advection feedbacks there (note the stronger salinity advection term in the box model than in $FAMOUS_A$ during the ramp-down phase, green lines in Figure 9 a,b). We conclude that the box model is more quantitatively accurate for the 'ramp up' threshold (which is the threshold of most direct interest for future changes), and that the quantitative errors in the 'ramp down' threshold are structural errors that could only be reduced by the addition of extra complexity in the box model (providing meridional structure in the reversed MOC cell).

5.3 Sensitivity to the method of applying fresh water perturbations

In our baseline $FAMOUS_A$ hosing hysteresis experiment, as analysed by H11 and J17, the hosing is compensated by an opposite surface fresh water extraction over the rest of the ocean surface, to maintain zero global mean fresh water flux (this experiment is called 'SCOMP' in J17). J17 also analyse an alternative $FAMOUS_A$ experiment in which the hosing is compensated by fresh water extraction distributed over the entire ocean *volume* (designated 'VCOMP'). The VCOMP experiment behaves somewhat differently to SCOMP, showing:

a) a more gradual weakening of the AMOC in VCOMP during the ramp-up phase, although the value of H_{crit} is similar to SCOMP. J17 attribute this difference to increased near-surface salinities in the subtropical Atlantic in SCOMP (due to the surface hosing compensation) being advected northwards by the MOC ($\langle q \rangle S'$, where

‹ › denotes the unhosed state and a prime denotes departures from it) and so counteracting the freshening effect of the Stommel advection feedback ($q'\langle S\rangle$). In VCOMP the near-surface freshening is not present, as the compensation is distributed through the water column, so the $\langle q\rangle S'$ term is smaller and the AMOC weakens more gradually as H increases (compare the total fresh water advection by the MOC in $\mathrm{FAMOUS_A}$, green curves in Figures 4a (SCOMP) and 10a (VCOMP)).

b) The left hand (ramp-down) threshold occurs at a much higher value of H in VCOMP, resulting in a very narrow hysteresis region in the ramp-up/ramp-down experiment, and possibly an almost completely monostable AMOC when more equilibrated solutions are considered (J17 Fig. 2b). This is attributed by J17 to the different South Atlantic reverse cells in the 'off' state in SCOMP and VCOMP.

We have emulated the VCOMP experiment in the box model by distributing the hosing compensation over the whole box model volume. We find only small differences from the box model SCOMP solution in the hysteresis loop and in the detail of the salinity budgets (Figure 10, compare with Figures 3c and 4b). We attribute the lack of impact on the sharpness of the threshold ((a) above) to the limited vertical resolution of the box model: a change in surface flux into the T box in the box model is necessarily spread over a depth of around 1000m, limiting the surface-intensified $\langle q\rangle S'$ feedback which delays AMOC weakening in the FAMOUS. In fact this difference explains why the standard SCOMP box model solution has a more gradual AMOC reduction than seen in FAMOUS (Fig. 3c); in this respect the box model SCOMP solution is intermediate between the FAMOUS SCOMP and VCOMP solutions. This limited vertical resolution is a fundamental structural bias in the box model, when used to emulate SCOMP-type hosing experiments. Turning to the differences (b) between the left-hand thresholds in VCOMP and SCOMP, we have already noted in Section 5.2 that the 'off' state involves changes in the inter-hemispheric structure of the MOC that are not represented by the box model, so it is not surprising that these differences found in $\mathrm{FAMOUS_A}$ by J17 are not present in the box model ramp-down phase.

5.4 Discussion of differences between box model and FAMOUS solutions

Overall we conclude that the box model tends to under-estimate the FAMOUS H_{crit} by around 0.1 - 0.2 Sv. Some of this bias is attributable to the lack of feedbacks through atmospheric fresh water fluxes (Section 5.1), and some to the limited vertical resolution of the box model, which reduces a stabilising advection feedback in the SCOMP experiment (Section 5.3). However the box model does include the primary driver of the rapid MOC decline near the ramp-up threshold, namely the quadratic dependence of the salinity advection by the MOC, on the North Atlantic salinity itself. This means that the box model is able to pick up the qualitative (and to some extent quantitative) differences in H_{crit} between different ocean states, and provide a simple framework to understand the main factors determining H_{crit}.

The box model also produces a more gradual AMOC decline in the ramp-up phase than is seen in the surface-compensated FAMOUS hosing experiments (SCOMP). This reflects the limited vertical resolution of the box model (Section 5.3).

By calibrating the box model to different decades in FAMOUS (not shown) and in an ocean reanalysis (Figure 5d), we estimate an additional uncertainty in the right-hand threshold position of at least ±0.04 Sv due to decadal ocean variability in the calibration variables.

The quantitative biases are greater for the left hand (ramp-down) threshold, due to water mass reorganisations in the FAMOUS off state that are not captured by the limited vertical and hemispheric resolution of the box model. However the qualitative similarity between Figures 9 a,b suggests that the box model may still provide useful qualitative insights into the dynamics of the left-hand threshold.

6. Discussion and conclusions

Our results show that the AMOC threshold and hysteresis behaviour in the FAMOUS AOGCM is controlled by low order dynamics, as represented by a 5-box dynamical model. The agreement between the box model and FAMOUS is particularly good for the 'ramp-up' threshold, which is the most relevant for future climate change. The box model parameters are determined by calibration to the baseline (un-hosed) ocean state, implying that the current ocean state contains sufficient information to estimate how far it is from threshold behaviour (e.g. in response to future fresh water input from the Greenland ice sheet).

The simplicity of the box model allows us to identify the factors in the ocean state that determine the position of the threshold H_{crit}. Because the overturning is strongly correlated with the North Atlantic density, we focus here on the salinity budget of the North Atlantic rather than the whole Atlantic basin, following Jackson et al. 2017. As in many previous studies the approach to the threshold is dependent on the 'salinity advection feedback', which involves a quadratic dependence of the AMOC on the North Atlantic salinity (eqn 14). However the exact value of H_{crit} depends on a balance between the salinity advection feedback and other processes. The un-hosed ('present day') value of M_{OV} at either the southern boundary of the Atlantic or in the northern subtropical Atlantic is not in itself a good predictor of H_{crit}. Other factors often play more important roles in determining H_{crit}, including the overall strength of the surface fresh water fluxes (hydrological cycle), the strength of the temperature driving of the flow, and the strength of the 'gyre' (i.e. non-AMOC) exchanges between the different water masses.

In our FAMOUS run with increased CO_2 concentrations, H_{crit} increases by several tenths of a Sverdrup compared to the state with pre-industrial CO_2. To the best of our knowledge this is the first time that the AMOC threshold has been evaluated explicitly with increased greenhouse gases. Analysis of the box model calibrated to the FAMOUS runs identifies three main factors driving the increase in H_{crit}, of which two (surface-intensified ocean warming and a strengthening global water cycle) are likely to be robust features of climate change. The intensified global water cycle means that even though more fresh water is delivered to the deep water formation region, the Atlantic basin as a whole becomes more evaporative (F_N + F_T becomes more negative, Table 1), leading to the increase in H_{crit}. The same warming and water cycle sensitivities are also seen when the box model is calibrated to a more advanced AOGCM, HadGEM2-AO, with various CO_2 concentrations. However, changes in the gyre mixing efficiencies also influence the value of H_{crit} at increased CO_2, and these changes appear less robust between models, perhaps because they result from changes in the wind field that are model-dependent. Analysis of more AOGCMs would be needed to understand how robust is the increase in H_{crit} with increased CO_2.

The box model can be calibrated to any AOGCM solution, and therefore opens up the possibility of obtaining a dynamical understanding of the different responses to hosing seen across different AOGCMs (e.g. Rahmstorf et al. 2005, Stouffer et al. 2006, Kageyama et al. 2013). Hysteresis experiments with other AOGCMs will also provide an important test of our model hierarchy, testing the robustness of our conclusions about the dominant AMOC stability mechanisms and allowing the importance of other modelling factors such as Bering Straits throughflow (Hu et al. 2012) or higher resolution (Jungclaus et al. 2013, den Toom et al 2014, Cheng et al. 2018) to be considered. Hysteresis experiments with eddy-resolving coupled models are computationally prohibitive at present but potentially feasible in future; a partial exploration of the hysteresis structure in a current generation (prototype-CMIP6) AOGCM, including an eddy-permitting ocean, has recently been carried out by Jackson and Wood (2018) and will be the subject of future study.

We stress that our study focuses on the response of the AMOC to slowly-varying fresh water forcing. Other processes, beyond those currently included in the box model, may come into play when considering the transient AMOC response to more rapidly varying forcing.

such as transient greenhouse gas increase (e.g. Stocker and Schmittner 1997; Thorpe et al. 2001; Gregory et al. 2005; Lucarini and Stone 2005). Such scenarios will be considered in a future study. We note that even the present box model exhibits a range of rate-dependent and duration-dependent responses to rapid changes in fresh water forcing (Alkhayuon et al. 2019).

While uncertainty remains over the quantitative modelling of changes in the AMOC threshold under increased greenhouse gases, our model hierarchy approach has identified some simple, low order dynamical controls on the threshold that can in principle be determined from observations (directly or through data-assimilating reanalyses). These observations provide a dynamically-based 'emergent constraint' (Hall and Qu 2006; Cox et al. 2018) on the position of the threshold. Hence it may be possible to monitor whether the threshold is becoming closer or further away, using large-scale oceanographic observations, to provide early warning of any approaching regime shift. This is particularly important because, as with many AOGCMs, FAMOUS and HadGEM2-AO overestimate the northward freshwater flux M_{OV} carried across 34ºS by the AMOC (Huisman et al. 2010; H11; Rodríguez et al. 2011; Mecking et al. 2017). While we showed in Section 4.3 that M_{OV} is not a direct indicator of H_{crit}, this bias suggests that the salinity advection feedback may excessively stabilise the AMOC in our AOGCMs (*Drijfhout et al. 2011; Cimatoribus et al. 2012; Jackson 2013*). So, even if it were possible to perform hosing runs with all current AOGCMs, relying on the current ensemble of AOGCMs to estimate H_{crit} may give a biased result. To obtain a preliminary estimate of H_{crit}, based on observations we have calibrated the box model to ocean states derived from an ocean reanalysis (*Smith et al. 2007*), which has M_{OV} around -0.2 *Sv*, close to observational estimates (*H11*) (Figure 5d). This yields an AMOC threshold at about 0.35 *Sv*, suggesting that the GCMs studied here (FAMOUS_A, FAMOUS_B and HadGEM2-AO) may all be slightly further from an AMOC threshold than the real ocean. Calibration of the box model to a wider range of both AOGCMs and ocean analyses, and a thorough uncertainty analysis of the observational constraints, are needed to provide a robust result; this will be the subject of a future study.

References:

Alkhayuon, H., P. Ashwin, L.C. Jackson, C. Quinn and R.A. Wood, 2019: Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model. *Proc. R. Soc. A,* **475**: 20190051, http:dx.doi.org/10.1098/rspa.2019.0051

Alley, R.B., 2003: Palaeoclimatic insights into future climate challenges. *Phil. Trans. Roy Soc. A,* **361**, 1831-1848.

Bakker, P., A. Schmittner, J.T.M. Lenaerts, A. Abe-Ouchi, D. Bi, M.R.van den Broeke, W.L. Chan, A. Hu, R.L. Beadling, S.J. Marsland, S.H. Mernild, O.A. Saenko, D. Swingedouw, A. Sullivan and J. Yin, 2016: Fate of Atlantic Meridional Overturning Circulation: Strong decline under continued warming and Greenland melting. *Geophys. Res. Lett.*, **43**, 12252-12260, doi:10.1002/2016GL070457.

Bryden, H. L. & S. Imawaki 2001 Ocean heat transport, in *Ocean Circulation and Climate*, edited by G. Siedler, J. Church & J. Gould, Academic Press, pp 455-474.

Cheng, w., W. Weijer, W.M. Kim, G. Danabasoglu, S.G. Yeager, P.R. Gent, D. Zhang, J.C.H. Chang and J. Zhang, 2018: Cn the salt advection feedback be detected in internal variability of the Atlantic Meridional Overturning Circulation? J. Climate, 31, 6649-6667, doi: 10.1175/JCLI-D-17-0825.1

Cimatoribus, A.A., S.S. Drijfhout, M. den Toom and H.A. Dijkstra, 2012: Sensitivity of the Atlantic meridional overturning circulation to South Atlantic freshwater anomalies. *Climate Dyn.*, **39**, 2291-2306, doi: 10.1007/s00382-012-1292-5.

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver and M. Wehner, 2013: Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Cox, P.M., C. Huntingford and M.S. Williamson, 2018: Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, **553**, 319-322.

Den Toom, M., H.A. Dijkstra, W.Weijer, M.W. Hecht, M.E. Maltrud and E. Van Sebile, 2014: Response of a Strongly Eddying Global Ocean to North Atlantic Freshwater Perturbations. *J. Phys. Ocenaogr.*,**44**, 464-481, DOI:10.1175/JPO-D-12-0155.1

deVries, P. and S.L. Weber, 2005: The Atlantic fresh water budget as a diagnostic for the existence of a stable shut down of the meridional overturning circulation. *Geophys. Res. Lett.,* 32, doi:10.1029/2004GL021450.

Dijkstra, H. A., 2007, Characterization of the multiple equilibria regime in a global ocean model. *Tellus A,* **59,** 695-705, DOI: 10.1111/j.1600-0870.2007.00267.x

Dijkstra, H.A. and Neelin, J.D., 1999: Imperfections of the thermohaline circulation: multiple equilibria and flux correction. *J. Climate,* **12,** 1382-1392.

Dijkstra, H.A., L. Te Raa and W. Weijer (2004): A systematic approach to determine thresholds of the ocean's thermohaline circulation. *Tellus A*, **56**, 362-370, doi:10.1111/j.1600-0870.00058.x

Döös, K., 1995: Interocean exchange of water masses. *J. Geophys. Res.*, **100**, 13499-13514.

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

Fichefet, T, C. Poncin, H. Goosse, P. Huybrechts, I. Janssens and H. Le Treut, 2003: Implications of changes in freshwater flux from the Greenland ice sheet for the climate of the 21st century. *Geophys. Res. Lett.,* **30,** doi:10.1029/2003GL017826.

Gnanadesikan, A., 1999: A simple predictive model for the structure of the oceanic pycnocline. *Science*, **283**, 2077-2079.

Gordon,C., C. Cooper, C.A. Senior, H.T. Banks, J.M. Gregory, T.C. Johns, J.F.B. Mitchell and R.A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.,* **16**, 147-168.

Gregory, J.M., 2000: Vertical heat transports in the ocean and their effect on time-dependent climate change. *Climate Dyn.*, **16**, 501-515.

Gregory, J.M., O.A. Saenko and A.J. Weaver, 2003: The role of the Atlantic freshwater balance in the hysteresis of the meridional overturning circulation. *Climate Dyn.,* **21**, 707-717

Gregory, J.M. et al. 2005: A model intercomparison of changes in the thermohaline circulation in response to increasing atmospheric CO_2 concentration. *Geophys. Res. Lett*., **32,** doi:10.1029/2005GL023209.

Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.* **33,** L03502

Hawkins, E., R. S. Smith, L. C. Allison, J. M. Gregory, T. J. Woollings, H. Pohlmann, and B. de Cuevas, 2011: Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport, *Geophys. Res. Lett.,* **38,** L10605, doi:10.1029/2011GL047208.

Hofmann, M. and S. Rahmstorf, 2009: On the stability of the Atlantic meridional overturning circulation. *Proc. Natl. Acad. Sci*., doi: 10.1073/pnas.0909146106

Hu, A. G.A. Meehl, W. Han, A. Abe-Ouchi, C. Morrill, Y. Ozaki and M.O. Chikamoto, 2012: The Pacific-Atlantic seesaw and the Bering Strait. *Geophys. Res. Lett*. **39**, L03702, doi: 10.1029/2011GL050567.

Hughes, T.M.C and A.J. Weaver, 1994: Multiple equilibria of an asymmetric 2-basin ocean model. *J. Phys. Oceanogr.*, **24**, 619-637.

Huisman, S. E., M. Den Toom, H. A. Dijkstra and S. Drijfhout, 2010: An indicator of the multiple equilibria regime of the Atlantic meridional overturning circulation. *J. Phys`. Oceanogr.,* **40,** 551–567. doi: 10.1175/2009JPO4215.1

Jackson, L.C., 2013: Shutdown and recovery of the AMOC in a coupled global climate model: The role of the advective feedback. *Geophys. Res. Lett*., **40**, 1182-1188, doi: 10.1002/grl.50289

Jackson, L., R.Kahana, T. Graham, M.A. Ringer, T. Woolings, J.V. Mecking and R.A. Wood, 2015: Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM. *Clim. Dyn.*, **45**, 3299-3316, doi: 10.1007/s00382-015-2540-2.

Jackson, L.C., R.S. Smith and R.A. Wood, 2017: Ocean and atmosphere feedbacks affecting AMOC hysteresis in a GCM. *Climate Dyn.*, doi: 10.1007/s00382-016-3336-9

Jackson, L.C. and R.A. Wood, 2018: Hysteresis and resilience of the AMOC in an eddy-permitting GCM. Geophys. Res. Lett., doi: 10.1029/2018GL078104.

Johnson, H.L., D.P. Marshall and D.A.J. Sproson, 2007: Reconciling theories of a mechanically driven meridional overturning circulation with thermohaline forcing and multiple equilibria. *Climate Dyn.*, **29**, 821-836, doi: 10.1007/s00382-007-026249.

Jungclaus, J.H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, J. S. von Storch, 2013: Characteristics of the ocean simulations in the Max Planck Institute Ocean

Model (MPIOM) the ocean component of the MPI-Earth system model. *J. Adv. in Modelling Earth Systems*, **5**, 422-446

Kageyama , M., U. Merkel, B. Otto-Bliesner, M. Prange, A. Abe-Ouchi, G. Lohmann, R. Ohgaito, D. M. Roche, J. Singarayer, D. Swingedouw, and X Zhang, 2013: Climatic impacts of fresh water hosing under Last Glacial Maximum conditions: a multi-model study. *Clim. Past*, **9**, 935–953, doi:10.5194/cp-9-935-2013

Landerer, F.W., J.H. Jungclaus and J. Marotzke, 2007: Regional dynamic and steric sea level change in response to the IPCC-A1B scenario. *J. Phys. Oceanogr.*, **37**, 296-312.

Lenton, T.M. et al., 2007: Effects of atmospheric dynamics and ocean resolution on bi-stability of the thermohaline circulation examined using the Grid ENabled Integrated Earth system modelling (GENIE) framework. *Climate Dyn.,* **29,** 591-613.

Liu, W., S. Xie, Z. Liu and J. Zhu, 2017: Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate. *Sci. Adv.*, **3**, e1601666,

Lucarini, V. and P.H.Stone, 2005: Thermohaline circulation stability: a box model study. Part I: uncoupled model. *J. Phys. Oceanogr.*, **18**, 501-513.

Manabe, S. and Stouffer, R.J., 1988: Two stable equilibria of a coupled ocean-atmosphere model. *J. Climate,* **1,** 841-863.

Marotzke, J. and Stone, P.H., 1995: Atmospheric transports, the thermohaline circulation, and flux adjustments in a simple coupled model. *J. Phys. Oceanogr.,* **25,** 1350-1364.

Martin, G.M. et al., 2011: The HadGEM2 family of Met Office Unified Model climate configurations. *Geosci. Model Dev.,* **4,** 723-757.

Mecking, J.V., S.S.Drijfhout, L.C. Jackson and M.B.Andrews, 2017: Theeffect of model bias on Atlantic freshwater transport and implications for AMOC bi-stability. Tellus A, 69:1, doi:10.1080/16000870.2017.1299910.

Mikolajewicz, U. et al., 2007: Long-term effects of anthropogenic CO2 emissions simulated with a complex earth system model. *Climate Dyn.,* **6,** 599-631.

Muglia, J., L.C. Skinner and A. Schmittner,2018: Weak overturning circulation and high Southern Ocean nutrient utilization maximised glacial ocean carbon. *Earth plan. Sci. Lett.,* **496**, 47-56, doi:10.1016/j.epsl.2018.05.038

Pardaens, A.K., Banks, H.T., Gregory, J.M. and Rowntree, P.R., 2003: Freshwater transports in HadCM3. *Clim. Dyn.,* **21,** 177-195.

Pfeffer, W.T., Harper, J.T. and O'Neel, S., 2008: Kinematic constraints on glacier contributions to 21st-cnetury sea-level rise. *Science,* **321,** 1340-1343.

Rahmstorf, S., 1996: On the Freshwater Forcing and Transport of the Atlantic Thermohaline Circulation, *Climate Dyn.,* **12,** 799–811, DOI: 10.1007/s003820050144

Rahmstorf, S. et al., 2005: Thermohaline circulation hysteresis: A model intercomparison. Geophys. Res. Lett., 32, L23605, doi:10.1029/2005GL023655

Rodríguez, J.A., T.C. Johns, R.B. Thorpe and A. Wiltshire, 2011: Using moisture conservation to evaluate oceanic surface freshwater fluxes in climate models. *Climate Dyn.*, **37**, 205-219.

Schneider, S.H., S. Semenov, A. Patwardhan, I. Burton, C.H.D. Magadza, M. Oppenheimer, A.BV. Pittock, A. Rahman, J.B. Smith, A. Suarez and F. Yamin, 2007: Assessing key vulnerabilities and the risk from climate change. In Climate Change 2007: Impacts, adaptation and vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate

Change, M.L. Parry, O.F. Canziani, J.P. palutikof, P.J. van der Linden and C.E. Hansen, eds., Cambridge University Press, Cambridge, UK, 779-810.

Sijp, W.P., 2012: Characterising meridional overturning bistability using a minimal set of state variables. *Climate Dyn.,* **39**, 2127-2142.

Smith, D. M. et al.,2007: Improved surface temperature prediction for the coming decade from a global climate model, *Science*, **317**, 796–799.

Smith, R.S., 2012: The FAMOUS climate model (versions XFXWB and XFHCC): description and update to version XDBUA. *Geosci. Model Dev.,* **5,** 269-276.

Smith, R.S. and Gregory, J.M., 2009: A study of the sensitivity of ocean overturning circulation and climate to freshwater input in different regions of the North Atlantic. *Geophys. Res. Lett.*, **36**, doi:10.1029/2009GL038607.

Smith, R.S., J.M. Gregory and A. Osprey, 2008:A description of the FAMOUS (version XDBUA) climate model and control run. *Geosci. Model Dev.*, **1**, 53-68.

Speich, S., B. Blanke and G. Madec, 2001:Warm and cold water routes of an O.G.C.M. thermohaline conveyor belt. *Geophys. Res. Lett.*, **28**,311-314.

Stocker, T.F. and A. Schmittner, 1997: Influence of CO_2 emission rates on the stability of the thermohaline circulation. *Nature*, **388**, 862-864.

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

Stouffer, R.J., K. W. Dixon, M. J. Spelman, W. Hurlin, J. Yin, J. M. Gregory, A. J. Weaver, M. Eby, G. M. Flato, D. Y. Robitaille, H. Hasumi, A. Oka, A. Hu, J. H. Jungclaus, I. V. Kamenkovich, A. Levermann, S. Nawrath, M. Montoya, S. Murakami, W. R. Peltier, G. Vettoretti, A. Sokolov, and S. L. Weber, 2006: Investigating the causes of the response of the thermohaline circulation to past and future climate changes. *Journal of Climate*, **19**(8):1365–1387.

Swingedouw, D., C.B. Rodehacke, S.M. Olsen, M. Menary, Y. Gao, U. Mikolajewicz and J. Mignot, 2015: On the reduced sensitivity of the Atlantic overturning to Greenland ice sheet melting in projections: a multi-model assessment. *Clim. Dyn.*, **44**, 3261-3279, doi: 10.1007/s00382-014-2270-x.

Talley, L.D., G.L. Pickard, W.J. Emery and J.H. Swift, 2011: *Descriptive Physical Oceanography: An Introduction.* Sixth Edition. Academic Press, Oxford, UK, 555 pp.

Thorpe, R., J.M. Gregory, T.C. Johns, R.A. Wood and J.F.B. Mitchell, 2001: Mechanisms determining the Atlantic thermohaline circulation response to greenhouse gas forcing in a non-flux-adjusted coupled climate model. *J. Climate,* **14,** 3102-3116.

Valdes, P., 2011: Built for stability? Nature Geosci., 4, 414-416.

Vellinga, M. & Wood, R.A., 2002: Global climate impacts of a collapse of the Atlantic thermohaline circulation. *Climatic Change*, **54,** 251-267.

Vellinga, M., R.A. Wood & J.M. Gregory, 2002: Coupled ocean-atmosphere feedbacks governing the recovery of a perturbed thermohaline circulation. *J. Climate*, **15**, 764-780.

Weber, S.L., S.S. Drijfhout, A. Abe-Ouchi, M. Crucifix, M. Eby, A. Ganopolski, S. Murakami, B. Otto-Bliesner and W.R. Peltier, 2007: The modern and glacial overturning circulation in the Atlantic Ocean in PMIP coupled model simulations. Clim. Past, **3**, 51-64.

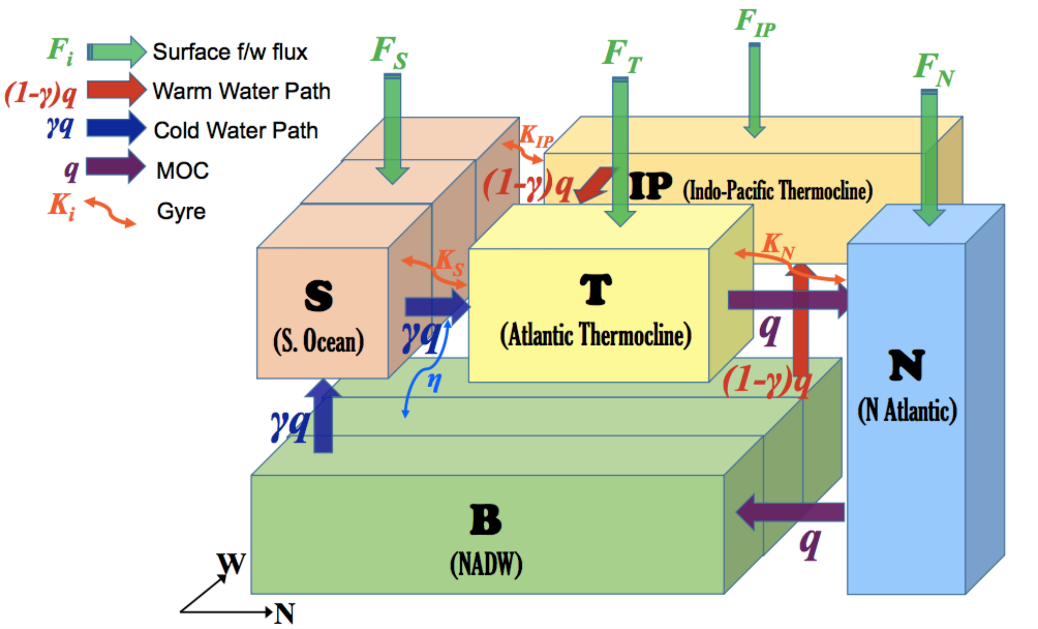
Parameter	$FAMOUS_A$ 1× CO_2	$FAMOUS_B$ 1×CO_2	$FAMOUS_B$ 2×CO_2	HadGEM2-AO 1×CO_2	HadGEM2-AO 2×CO_2	HadGEM2-AO 4×CO_2	DePreSys 1999-2008
V_N (m^3 x10^{16})	3.683	3.261	3.683	3.557	5.259	5.257	4.854
V_T (m^3 x10^{16})	5.151	7.777	5.418	8.908	7.400	7.454	7.583
V_S (m^3 x10^{16})	10.28	8.897	6.097	10.330	9.336	9.462	17.247
V_{IP} (m^3 x10^{16})	21.29	22.02	14.86	19.219	19.220	19.155	38.856
V_B (m^3 x10^{16})	88.12	86.490	99.25	90.23	89.90	90.78	73.55
A_N	0.194	0.070	0.131	0.117	0.285	0.197	0.194
A_T	0.597	0.752	0.696	0.703	0.522	0.620	0.608
A_S	-0.226	-0.257	-0.263	-0.303	-0.299	-0.326	-0.282
A_{IP}	-0.565	-0.565	-0.564	-0.517	-0.508	-0.491	-0.519
F_N (Sv)	0.375	0.384	0.486	0.453	0.496	0.577	0.531
F_S (Sv)	1.014	1.078	1.265	0.901	1.021	1.114	0.849
F_T (Sv)	-0.723	-0.723	-0.997	-0.798	-0.921	-1.099	-0.743
F_{IP} (Sv)	-0.666	-0.739	-0.754	-0.556	-0.596	-0.592	-0.637
T_S (°C)	5.571	4.773	7.919	6.456	7.424	8.710	4.385
T_0 (°C)	3.26	2.65	3.87	2.71	3.29	3.70	2.12
μ (°Cm^{-3}s x10^{-8})	7.0	5.5	22.0	1.4	16.0	28.0	2.7
λ ($m^6kg^{-1}s^{-1}$ x10^7)	2.66	2.79	1.62	2.17	1.66	1.28	3.53
K_N (Sv)	5.439	5.456	1.762	5.601	15.890	20.954	17.07
K_S (Sv)	1.880	5.447	1.872	7.169	6.828	8.384	3.546
K_{IP} (Sv)	89.778	96.817	99.977	459.095	1029.641	477.332	192.649
η(Sv)	66.061	74.492	33.264	3.758	9.871	6.773	19.689
γ	0.58	0.39	0.36	0.85	0.73	0.39	0.33

Table 1

Box model parameter values for all calibrations used in this paper. The parameters A_N, A_T, A_S and A_{IP} are multiplicative factors for the hosing for their respective boxes and depend on the latitudes of the box boundaries. In the AOGCM the hosing is added to the region 20-50°N of the Atlantic, with a compensating fresh water removal from the rest of the global ocean surface. Typically the AOGCM hosing region spans some of the N box and some of the T box. The A's are chosen to give the same total fresh water flux $H.A_i$ into each box as in the corresponding AOGCM run ($A_N + A_T + A_S + A_{IP} = 0$).

FIGURES:

a.



b.

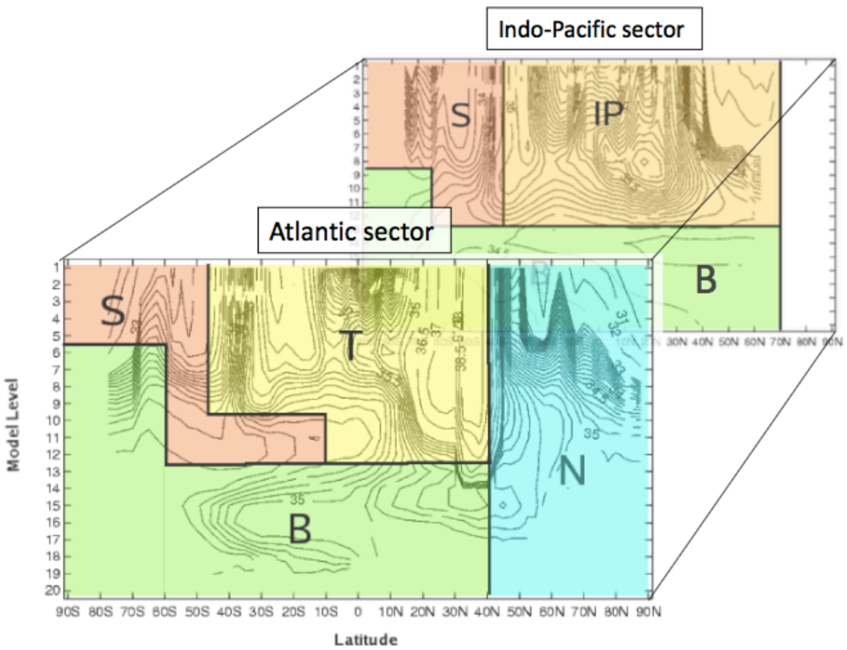
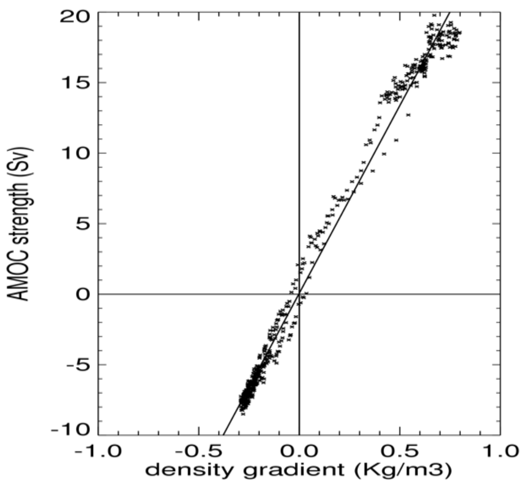


Fig. 1 Box model definition

(a) Schematic representation of the box model. The control parameters of the model are the temperature difference between N and S boxes, the pipe constant (λ), the surface freshwater fluxes (F_i), the wind-driven transport constants (K_i), the S-B box mixing parameter (η) and the proportion of the cold water path (γ). All parameters except γ can be diagnosed from any GCM state, or in principle from observations. (b): Boundaries of model boxes used in the calibration of the box model to the $FAMOUS_A$ pre-industrial ($1xCO_2$) run, superimposed on the zonal average of the $FAMOUS_A$ salinity distribution across the Atlantic and Indo-Pacific Oceans

a.



b.

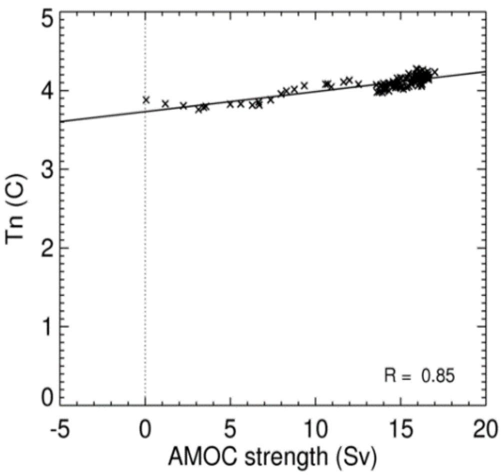
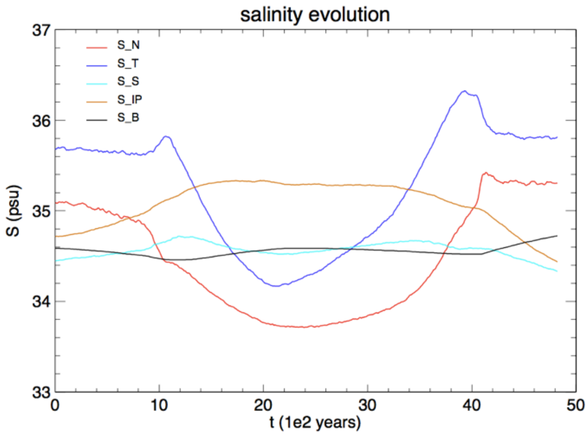


Fig. 2

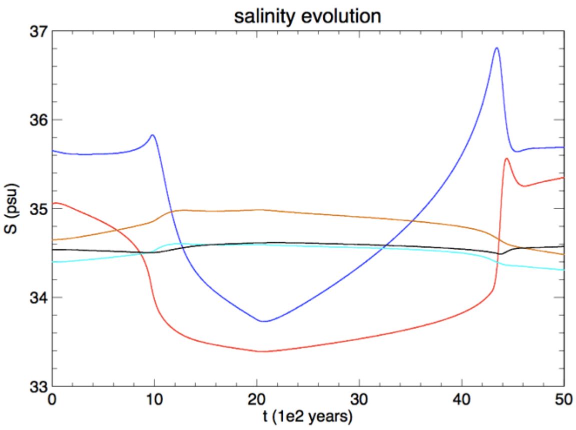
(a) AMOC strength as function of N-S density difference. Scatter plot of FAMOUS_A AMOC strength vs. density difference between the two portions of the ocean that define the N and S boxes in the box model. The points shown cover the entire hysteresis run with preindustrial CO_2.

(b) Temperature of N box as a function of AMOC strength**.** Scatter plot of FAMOUS_A box-mean temperature T_N vs. AMOC strength q. The points shown cover the part of hysteresis between the unhosed state and the first threshold crossing, for the run with preindustrial CO_2.

a.



b.



c.

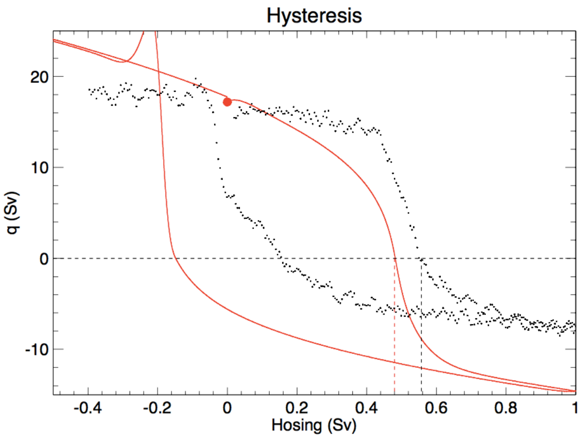
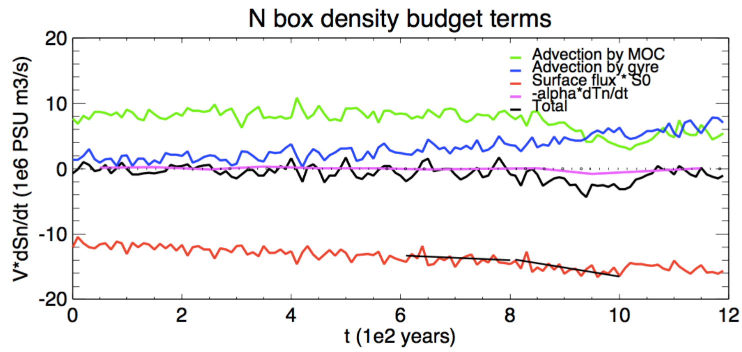


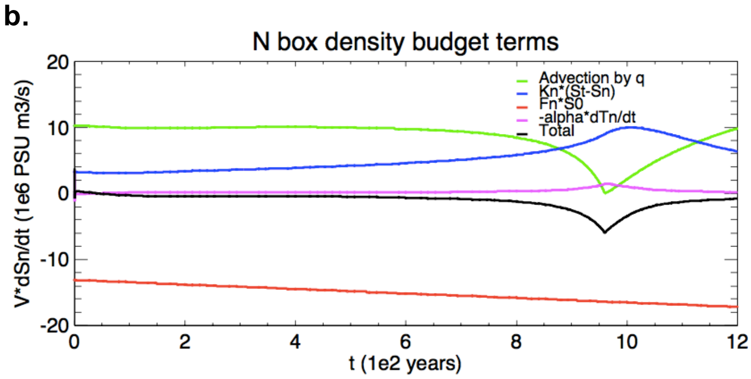
Fig. 3: Comparison between FAMOUS_A and box model simulations

(a) Salinity evolution in the five model boxes through the 5000 years of the FAMOUS_A hosing experiment [*H11*] (b) As (a) but for the corresponding box model experiment. The same rate of increase of hosing is used for both experiments. (c) AMOC strength as function of hosing applied. Dots: FAMOUS_A (decadal means). Red line: box model. The box model has been calibrated solely to the unperturbed initial state of $FAMOUS_A$ (shown by the red dot). The dashed lines show the critical hosing value H_{crit}.

a.



b.



c.

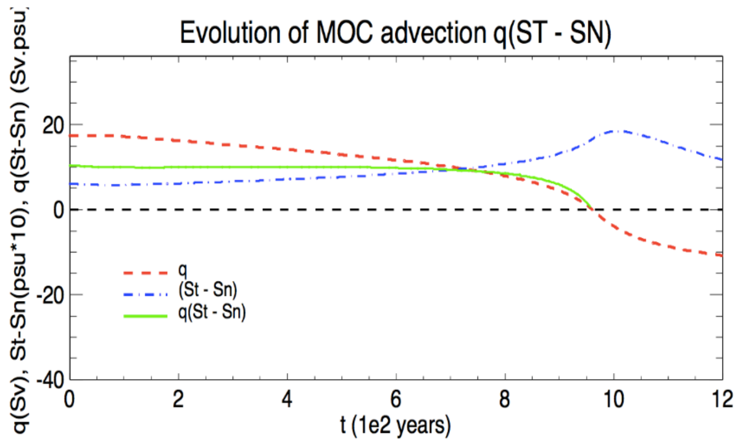


Fig. 4

Salinity budget terms for the North Atlantic box in years 0-1200, for (a) $FAMOUS_A$ (adapted from J17), (b) box model. Black: dS_N/dt; red: surface flux (including hosing); green; advection by MOC; blue: advection by gyre(FAMOUS)/diffusion by K_N (box model). Also shown is the density change due to temperature response to the AMOC, converted into an equivalent salinity change (pink). Average slope lines for years 601-800 and 801-1000 are shown for the surface flux term in (a) to illustrate the atmospheric water flux feedback. The individual components of the fresh water transport by the MOC, $-q(S_T\text{-}S_N)$, are shown for the box model in (c) [Red: q (Sv); blue: $(S_T\text{-}S_N)$ (psu * 10); Green: $-q(S_T\text{-}S_N)$ (Sv.psu)].

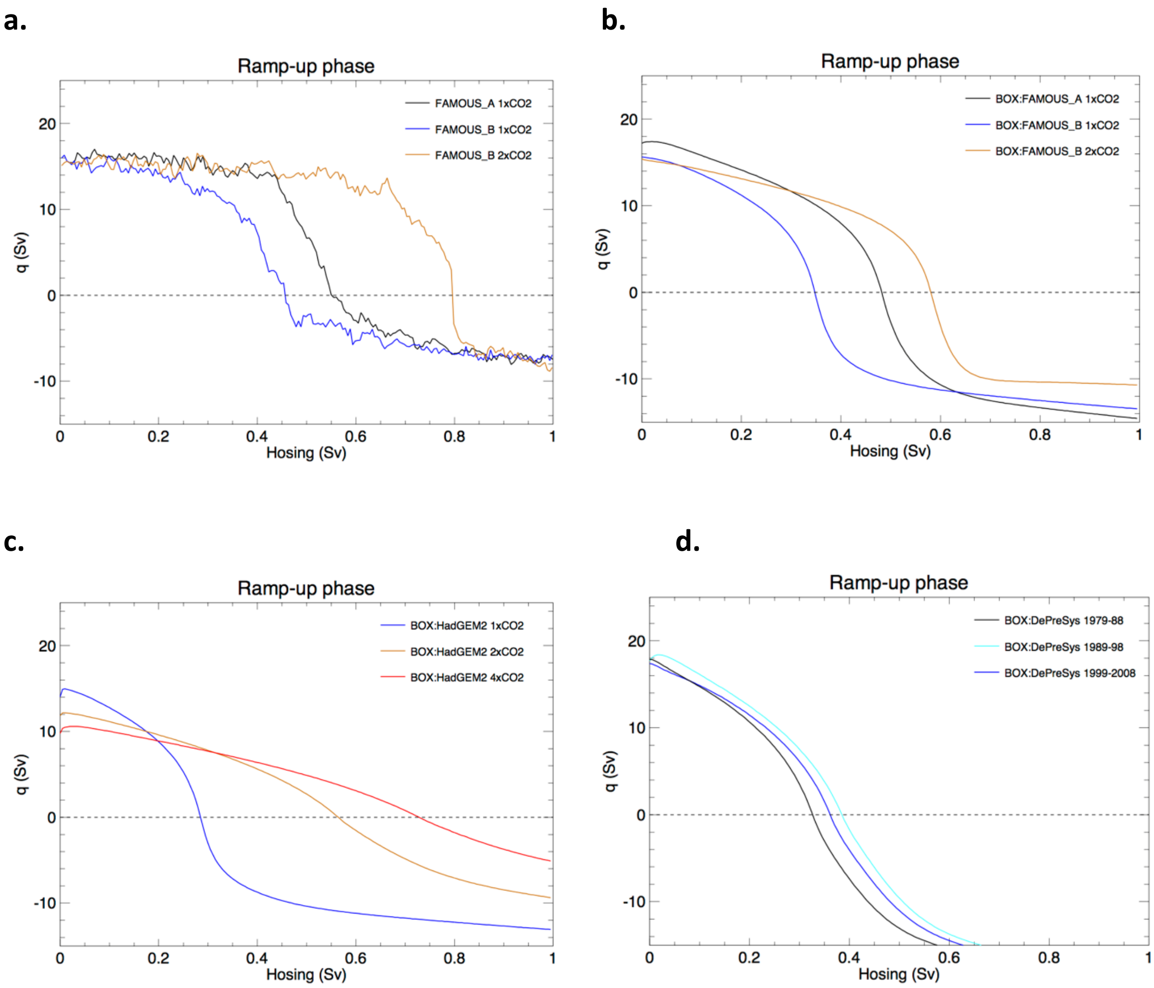
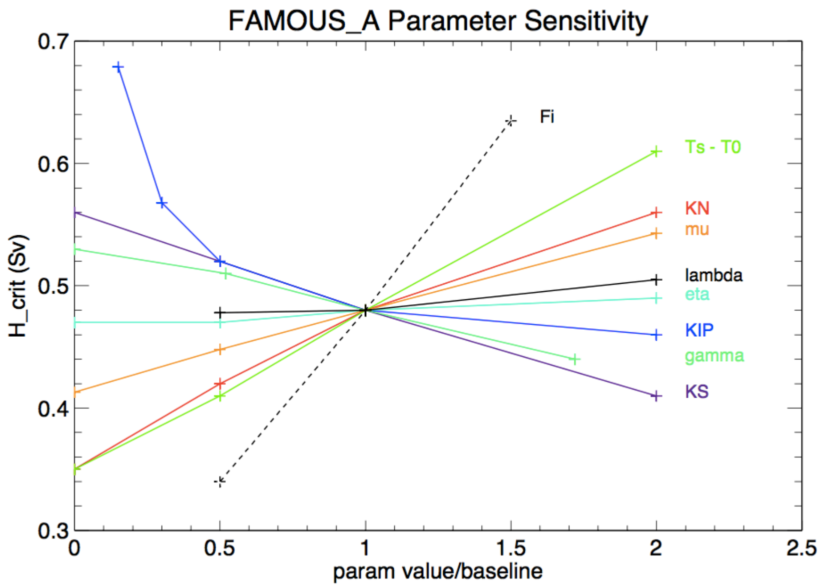


Fig. 5 AMOC thresholds in preindustrial and increased CO_2 simulations
AMOC strength as function of hosing applied in transient experiments from various near-equilibrated CO_2 states. Only the 'ramp-up' part of the experiment (hosing increasing up to 1.0 Sv) is shown. (a) $FAMOUS_A$ at pre-industrial CO_2 (black), $FAMOUS_B$ at pre-industrial (blue) and $2\times CO_2$ (brown); (b) box model calibrated to the three FAMOUS runs shown in (a); (c) box model calibrated to HadGEM2-AO at preindustrial (blue), $2\times CO_2$ (brown) and $4\times CO_2$ (red); (d) box model calibrated to *Smith et al.* [2007] ocean reanalyses for the decades 1979-89 (black), 1989-99 (cyan), 2000-2009 (blue).

a.



b.

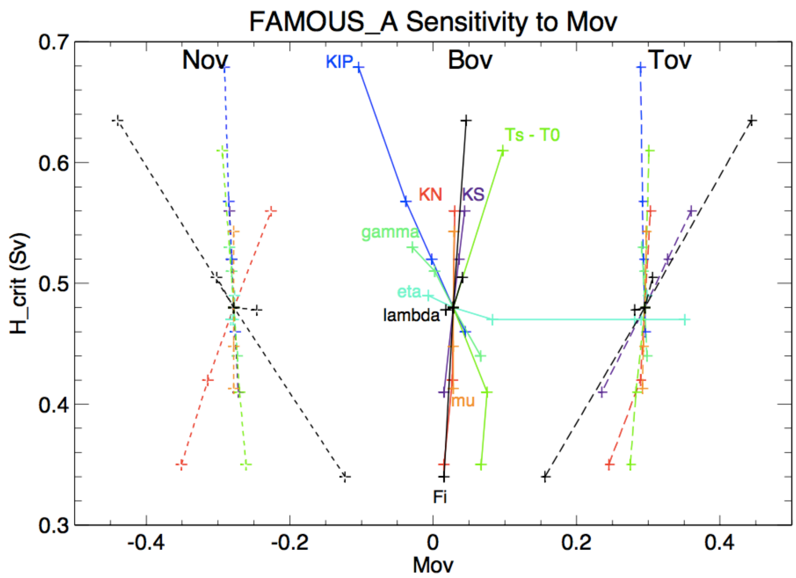
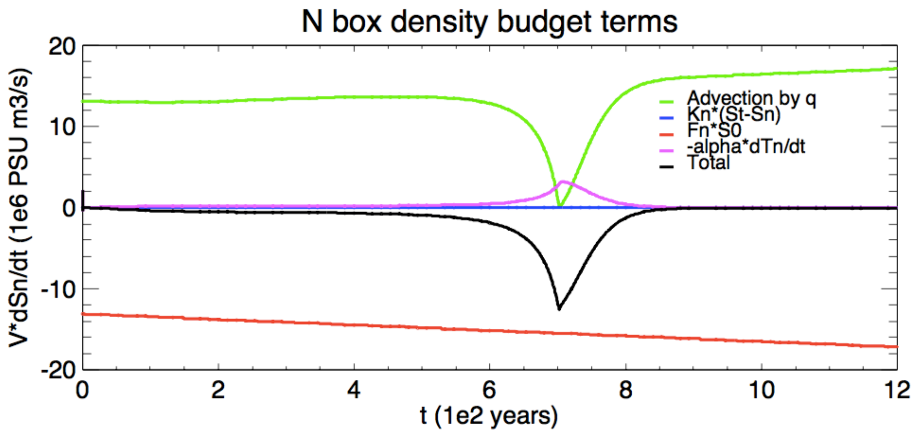


Fig. 6. Sensitivity of H_{crit} to box model parameters

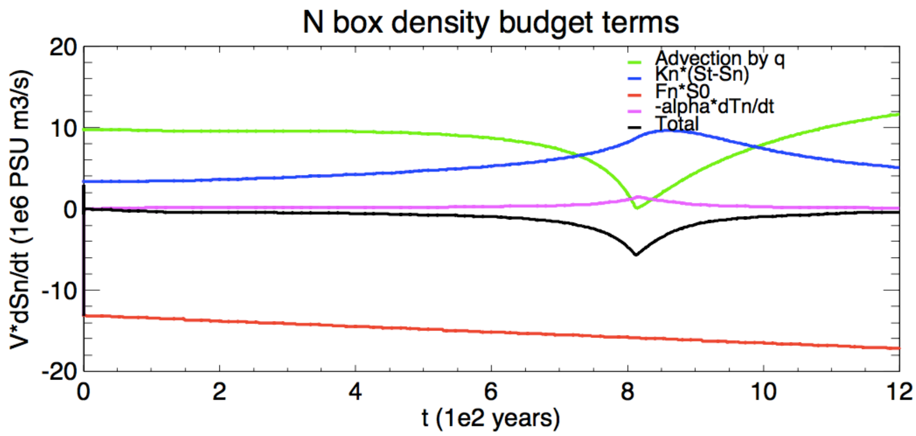
(a) Sensitivity of H_{crit} to changes in the values of a single box model parameter, relative to a baseline state calibrated to the $FAMOUS_A$ AOGCM experiment. The baseline parameter values are given in Table 1, and the parameter changes are shown along the horizontal axis as a proportion of the baseline value.

(b) For same box model parameter sensitivity experiments as in (a), sensitivity of H_{crit} to the value of the fresh water transport by the AMOC (Sv) in the un-hosed state, for the three diagnostics N_{OV} (short dashed, left), T_{OV} (long dashed, right) and B_{OV} (solid, centre) – units: Sv.

a.



b.



c.

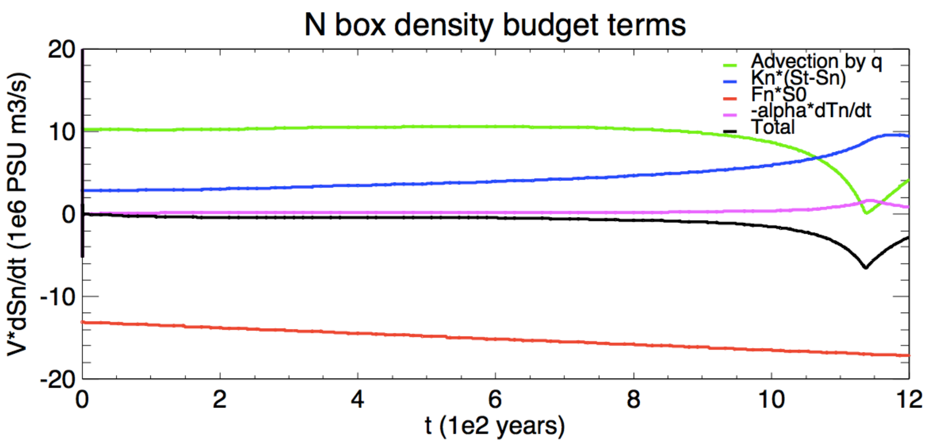
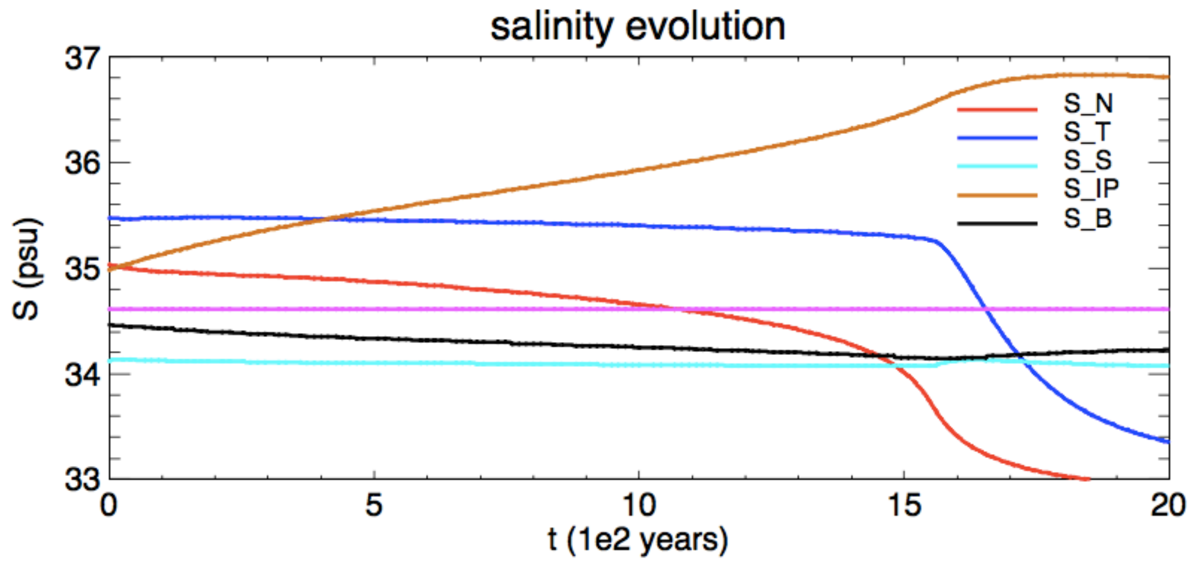


Fig. 7

N box salinity budget for selected box model parameter sensitivity tests relative to the baseline $FAMOUS_A$ calibration: (a) $K_N=0$, (b) K_S = 2 × baseline value , (c) K_{IP} = 0.3 × baseline value. Legend as for Fig. 4b.

a.



b.

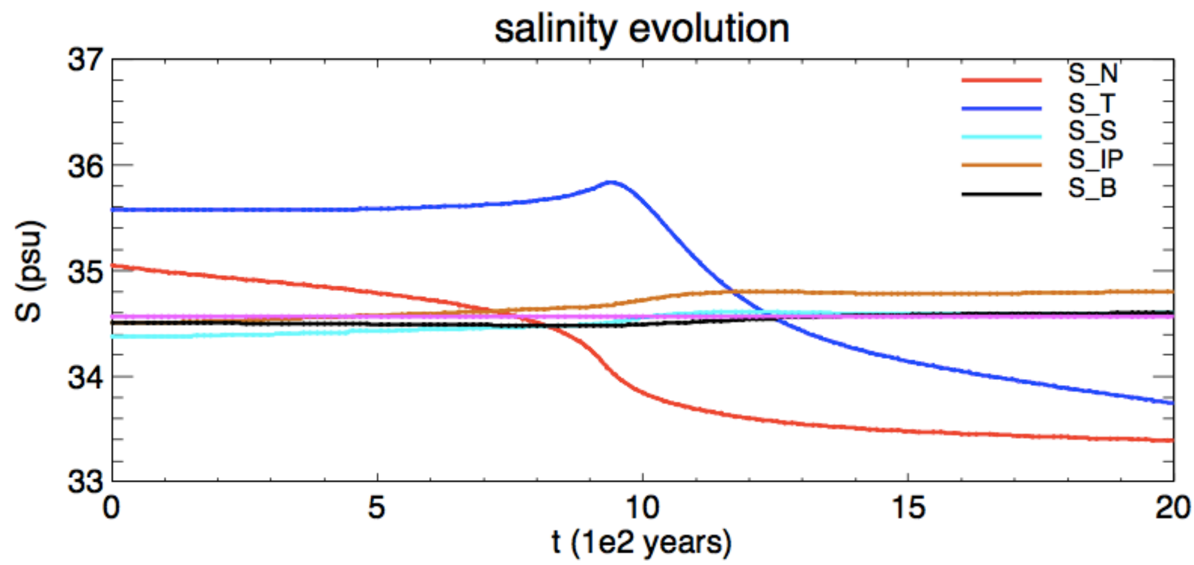
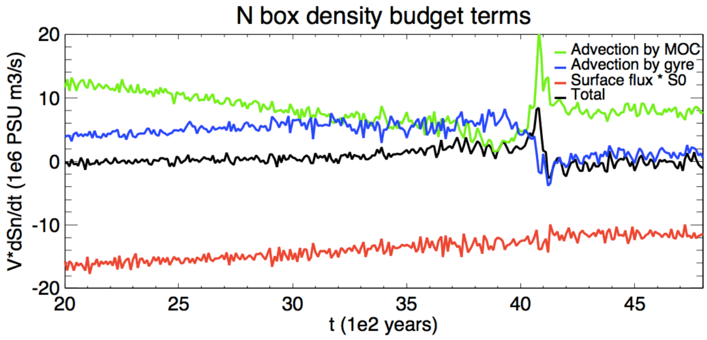


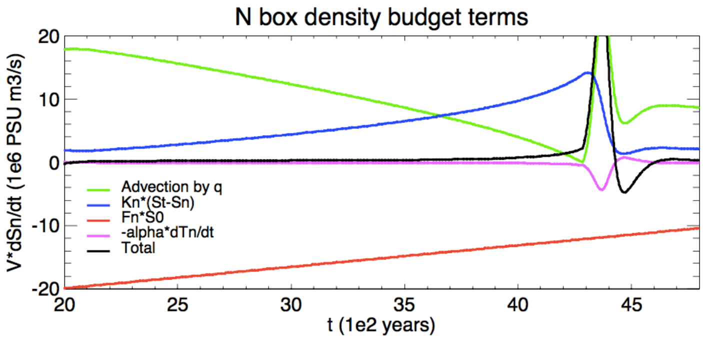
Fig. 8

Box model salinity evolution over the ramp-up stage in the parameter sensitivity studies for (a) K_{IP}=8.9778 Sv (0.1 x baseline value) and (b) K_{IP}=179.556 Sv (2 x baseline value).

a.



b.



c.

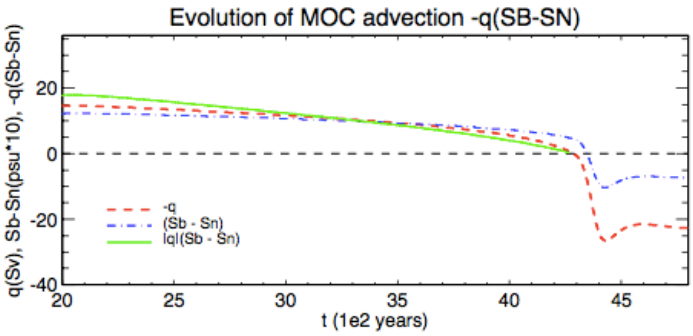
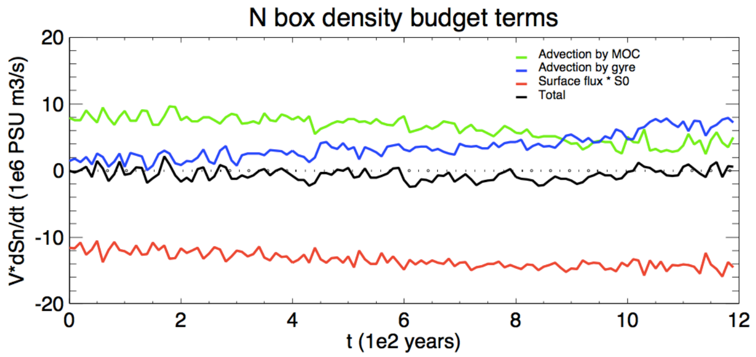


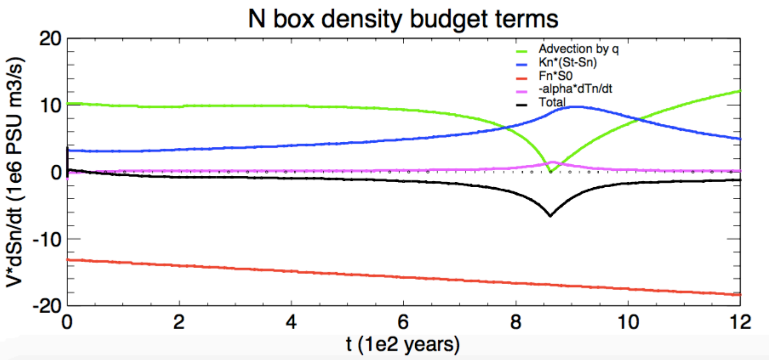
Fig. 9

As Fig. 4, but for the ramp-down phase from year 2000 (H = 1.0 *Sv*) to year 4800 (H = -0.4 *Sv*).

a.



b.



c

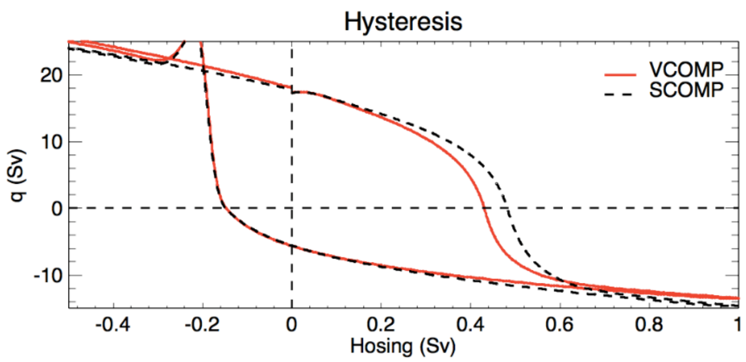


Fig. 10

AMOC hysteresis in the VCOMP version of $FAMOUS_A$ and the corresponding box model. Shown in (a) and (b) are the $FAMOUS_A$ and box model salinity budgets for the N box in the ramp-up phase (cf. Fig. 4 a,b for SCOMP), while (c) shows the whole hysteresis loop (red), with the corresponding loop from the SCOMP run in black dashed (reproduced from Fig. 3c)