Precise calculations of the existence of multiple AMOC equilibria in coupled climate models Part I: equilibrium states.

Willem P. Sijp*, Matthew H. England and Jonathan M. Gregory**.

Climate Change Research Centre (CCRC), University of New South Wales, Sydney, New South Wales, Australia

April 19, 2011

Journal of Climate

*Corresponding author address: Willem P. Sijp, Climate Change Research Centre (CCRC), University of New South Wales, Sydney, NSW 2052, Australia. E-mail: w.sijp@unsw.edu.au

** NCAS-Climate, University of Reading, UK and Met Office Hadley Centre, Exeter, UK.
Abstract

This study examines criteria for the existence of two stable states of the Atlantic Meridional Overturning Circulation (AMOC) using a combination of theory and simulations from a numerical coupled atmosphere-ocean climate model. By formulating a simple collection of state parameters and their relationships, we reconstruct the NADW "off" state behaviour under a varying external salt flux forcing. This part (Part I) of the paper examines the steady state solution, which gives insight into the mechanisms that sustain the NADW "off" state in our coupled model; Part II deals with the transient behaviour predicted by the evolution equation. The non-linear behaviour of the Antarctic Intermediate Water (AAIW) reverse cell is critical to the "off" state. Higher Atlantic salinity leads both to a reduced AAIW reverse cell and to a greater vertical salinity gradient in the South Atlantic. The former tends to reduces Atlantic salt export to the Southern Ocean, while the latter tends to increases it. These competing effects produce a non-linear response of Atlantic salinity and salt export to salt forcing, and the existence of maxima in these quantities. Thus we obtain a natural and accurate analytical saddle node condition for the maximal surface salt flux for which a NADW "off" state exists. By contrast, the bistability indicator proposed by De Vries and Weber (2005) does not generally work in our model. It is applicable only when the effect of the AAIW reverse cell on the Atlantic salt budget is weak.
1. Introduction

North Atlantic Deep Water (NADW) formation has a significant role in high-latitude climate via its capacity to control poleward heat transport. In addition, the associated deep mixed layers in winter allow vertical transport of underlying warm saline water to the surface. The absence of this process would likely result in a significant cooling of the North Atlantic (e.g. Manabe and Stouffer 1988, 1999; Vellinga and Wood 2002), and a fundamental reorganization of global ocean circulation. Stommel (1961) already put forward the idea that an ocean thermohaline circulation could have a regime of two stable states. Bryan (1986) showed that this regime is consistent with the governing equations of the three-dimensional oceanic circulation applied in a numerical model employing latitudinally symmetric forcing and geometry. Manabe and Stouffer (1988) found that a coupled ocean-atmosphere model may allow two stable equilibria, where one is characterised by a permanent absence of NADW formation (the NADW “off” state). Rahmstorf (1995, 1996) analysed the existence of two steady states, also named “bistability” or “multiple equilibria”, in the context of the total Atlantic salinity budget. Dijkstra and Weijer (2005) were the first to directly compute the equilibrium states via continuation methods applied to a fully implicit global ocean model. The state of their system is characterised by large vectors containing velocity and tracer values at each grid cell. Despite method-related simplifications such as very high viscosity and an absence of convective adjustment, the great value of their approach lies in the calculation of the unstable equilibrium branch. This reveals the precise locations of the nodes bounding the bistable regime. The asymmetrical forcing and geometry of the Atlantic basin determine
the node type to be a saddle node (Dijkstra and Weijer 2003), and a symmetric situation as described by Bryan (1986) would lead to pitchfork bifurcations instead via symmetry breaking upon the development of an asymmetric circulation (Vellinga 1995; Dijkstra and Weijer 2003). However, although these important studies were able to identify the unstable equilibrium states in a model of reasonable complexity, the deeper reasons for bistability remained obscured by the high number of state parameters. Subsequent work has focused on the identification of a bistability indicator (e.g. Dijkstra 2007; Huisman et al. 2010, see below) in this context. In the present study, using characteristics of the NADW “off” state alone, the maximal Atlantic surface salt is calculated where multiple equilibria are possible in the UVic model. The simplicity of our state parameters allows us to identify the physical reasons behind bistability in our model, and identify the associated saddle node bifurcation conditions.

Manabe and Stouffer (1988) noted that the NADW “off” state is associated with a shallow Antarctic Intermediate Water (AAIW) reverse cell inside the Atlantic, where cool fresh intermediate water replenishes the departure of warm saline thermocline water from the Atlantic. Saenko et al. (2003a), Gregory et al. (2003) and Sijp and England (2006) attribute to this reverse cell an active role in the suppression of NADW formation. Here, the reverse circulation constitutes an export mechanism of salt, keeping Atlantic salinity low enough to maintain the suppression of NADW formation. Furthermore, Gregory et al. (2003) suggest that bistability occurs because both the NADW “off” and NADW “on” states tend to reinforce the different Atlantic salinity contents associated with each state, opposing the anomalous surface salinity flux. This re-inforcement of Atlantic salinity by the AAIW reverse cell could arise from
enhanced salt export when Atlantic surface salinity is increased due to an enhanced vertical salinity gradient at 30 °S. This raises the question whether, in the absence of other similarly interactive salinity transport components, the reverse AMOC circulation is a necessary factor in maintaining the NADW “off” state. Sijp and England (2006) find that the NADW “off” state is eliminated in their models when the reverse cell is reduced by low values of vertical diffusivity only inside the Atlantic (but not outside the basin), suggesting a critical role for this cell in their model.

Generally, the addition of a sufficiently large constant anomalous surface salt flux to the Atlantic basin can cause a transition from a NADW “off” state to a NADW ”on” state. This approach is taken in so-called “hysteresis” experiments (e.g. Rahmstorf 1996), where the system is gradually taken through a series of near-steady states by means of a slowly varying anomalous salt flux applied for instance to the North Atlantic. Eventually, this anomalous salt flux typically reaches a threshold value at which the NADW “off” state is eliminated, and a monostable regime is obtained with NADW “on”. The term hysteresis here refers to the behaviour where a subsequent reduction in the anomalous salt flux to values previously associated with the “off” state now leads to a NADW “on” state, a phenomenon we refer to as “bistability” here. That is, there are two steady states under identical anomalous surface salt flux forcing (although the atmosphere may cause further changes in Atlantic surface flux). Alternatively, transitions between the two stable states can also be excited by short fresh water pulses applied to the NADW and AAIW sinking regions (e.g. Saenko et al. 2003a; Sijp and England 2006).
Saenko et al. (2003a) show that the elimination of the NADW “off” state can also be achieved by adding fresh water (FW) to the AAIW formation regions, as the strength of the AAIW cell depends on the density difference between the AAIW and the NADW formation regions, which can be eliminated in this fashion, leading to the onset of NADW formation. Gregory et al. (2003) shows that a state transition from a NADW “off” state to a NADW “on” state can be triggered by applying an implied anomalous southward atmospheric FW transport contained inside the Atlantic and without a net change in total Atlantic surface flux. This reduces the density difference between the AAIW and NADW formation regions, thus diminishing the AAIW reverse cell and allowing salinity to accumulate inside the Atlantic (their Figure 2). The difference in salinity export characteristics between the AAIW reverse cell and the NADW cell plays a key role in what the authors refer to as the ”Atlantic flip flop”.

The bistability behaviour of ocean models prompts questions about whether a NADW “off” state could exist in nature (Rahmstorf 1996), and why this state occurs in certain climate models and not in others. For example Pardaens et al. (2003) notes that the Hadley Centre coupled climate model (HadCM3) exhibits a FW transport from the Northern Hemisphere (NH) to the Southern Ocean (SO) that is strong, arising from a hydrological cycle that is too vigorous, and state that this may have consequences for the stability of NADW in this model. HadCM3 probably does not have a NADW “off” state.

Rahmstorf (1996) realized the significance of net salt import into the Atlantic basin by the AMOC, and indicated a close link between the net import/ export of salt and the stability of the overturning. In a related fashion, DeVries and Weber (2005) conducted experiments in
a model of intermediate complexity where the Atlantic salt budget is changed via modifications of the longitudinal variation of surface salinity near the southern border of the basin at 33 degs S. Results suggested that the sign of the AMOC salt import at the latitudes of the Cape of Good Hope determines the existence of a monostable or bistable regime. We will refer to this bistability indicator as the de Vries Weber (dVW) metric. Dijkstra (2007) use a fully implicit global ocean model to examine bistability arising from the existence of an OFF state, and find good performance in their model for a slightly modified version of the de Vries Weber metric, which now comprises the salt flux divergence between 30 °south and 60 °north. Huisman et al. (2010) trace the cause of the good performance of the metric in these two studies to a coupling between the net anomalous FW transport into or out of the Atlantic and the interaction between the velocity perturbation (in response to a surface FW pulse) and the steady state background salinity field. Hofman and Rahmstorf (2009) find bistability also in a low-diffusion setting.

In contrast, here it is proposed that the sign of the AMOC-related salinity transport is not always a precise bistability indicator, as in our model. Instead, it is the derivative with respect to salinity of the total oceanic salt flux in the NADW “off” state that determines the existence of this state. We derive a reliable expression for the critical Atlantic surface salt flux $F_{s}^{*}$ beyond which multiple equilibria no longer occur, further indicating the significance of the AAIW reverse cell in maintaining the OFF state in our model. Both the theory and the model yield remarkably consistent results.

The rest of this paper is divided as follows. Section 3 examines the salt import properties of
the oceanic circulation in the NADW OFF and ON state and the climatic changes associated with state transitions and changes in surface salt fluxes. We show that the OFF state AMOC exports salt and is responsive to changes in surface flux. In the following sections the discussion is confined to the OFF states alone. In Section 4 it is shown how the oceanic salt flux terms vary with the average Atlantic salinity and how the AAIW reverse cell depends on the density difference between the NADW and the AAIW formation regions. We then derive a simple evolution equation for the average Atlantic salinity to capture the OFF state existence domain and its behaviour in two steps. The first step consists of ground work, described in Section 5, involving a set of basic behaviours of the OFF state in response to different values of a fixed anomalous salt flux. The formulation of these basic OFF state properties in terms of linear equations gives an evolution equation for the average Atlantic salinity in Section 6. Simplicity of this evolution equation arises from the fact that local Atlantic salinities depend linearly on the average Atlantic salinity throughout our bistable regime, NA temperature effects are minimal, the reverse cell strength depends linearly on a meridional density difference and the overall climatic effect of changes in reverse cell strength can be encapsulated in the total surface flux.

2. The Model and Experimental Design

We use the intermediate complexity coupled model described in detail in Weaver et al. (2001). This consists of an ocean general circulation model (GFDL MOM Version 2.2 Pacanowski 1995) coupled to a simplified one-layer energy-moisture balance model for the
atmosphere and a dynamic-thermodynamic sea-ice model of global domain and horizontal resolution 3.6° longitude by 1.8° latitude. Heat and moisture transport takes place via advection and Fickian diffusion. Air-sea heat and freshwater fluxes evolve freely in the model, but a non-interactive wind field is employed. The wind forcing is taken from the NCEP/NCAR reanalysis fields (Kalnay et al. 1996), averaged over the period 1958-1997 to form a seasonal cycle from the monthly fields. Vertical mixing in the control case is represented using a diffusivity that increases with depth, taking a value of 0.3 \( cm^2/s \) at the surface and increasing to 1.3 \( cm^2/s \) at the bottom. We use version 2.8 of the UVic model. The effect of sub-grid scale eddies on tracer transport is modelled by the parameterizations of Gent and McWilliams (1990). We will refer to this model as “the numerical model” to distinguish it from our conceptual model. We use virtual salt fluxes with a reference salinity of 35 \( kg/m^3 \). We use the Flux Corrected Transport (FCT) scheme (Boris and Book 1973; Zalesak 1979; Gerdes et al. 1991) for tracer advection.

The bistability in ocean models arising from the application of a slowly varying anomalous salt flux generally leads to two branches in a diagram depicting a key model diagnostic (e.g. NADW sinking, or Atlantic surface salinity) as a function of the applied salt flux. One branch is associated with the NADW “off” state and one with the NADW “on” state (e.g. see Figure 5 in Rahmstorf 1996). Here, the procedure of a slowly varying flux is not followed. Instead, the model is run to equilibrium for a number of anomalous surface flux values \( H \) applied to the NADW formation regions (see Fig. 1). This leads to an equilibrium diagram. Naturally, this flux is compensated everywhere else in the world ocean so as to render the total anomalous salt flux zero. Here, we start off with a stable NADW “off” equilibrium
and change $H$ to obtain equilibria along the NADW “off” branch, and we start with a stable NADW “on” equilibrium to obtain the “on” branch. We seek the critical surface fluxes where each branch becomes unstable by trial and error, and denote by $H^*$ the critical $H$ beyond which only an ON state exists. We loosely refer to the existence of the NADW OFF state at a certain $H$ as “bistability”. Strictly, bistability refers to the existence of both the ON and the OFF state at a certain $H$. However, the critical flux $H^*$ where the ON state becomes unstable is not examined.

For convenience, the stable equilibria where NADW is absent are referred to as the OFF states (where NADW is “off”), and to the equilibria where NADW is present as the ON states. In addition, the term AMOC will be used also for the meridional circulation in the OFF state. We decompose the Atlantic salt budget into the net surface salt flux $F_s$ across the surface interface with the atmosphere, the AMOC-related component $F_m$ as defined by Rahmstorf (1996) (see below) and the residual flux $F_r = -F_m - F_s$, assuming a flux balance $F_r + F_m + F_s = 0$. The naming of this term is intended as a reminder of how this term is computed, but can be thought of loosely as the “gyre term” as most of the incorporated transports are advective. $F_s$ includes the forcing flux $H$. Bering Strait is closed in our model, so the only oceanic exchange out of the Atlantic occurs around 30° south at the latitude of the Cape of Good Hope. The AMOC-related term represents the net oceanic salt flux arising from salinity differences in the water masses passing across 30° south at different depths. For simplicity, $\bar{S}$ will denote the average upper Atlantic salinity, where the overbar denotes spatial averaging over the entire Atlantic north of the Cape of Good Hope and above 1250m depth. See Table 1 for an explanation of the mathematical terms. Finally, salt fluxes of positive sign
are always assumed to add salinity to the Atlantic. We use units of “negative” Sverdrup (Sv) for convenience, where a negative salt flux of $1\text{Sv}$ dilutes the Atlantic salinity by the same amount as the addition of $1\text{Sv}$ of Fresh Water. Atlantic water volume remains constant under precipitation or evaporation as if a rigid lid were applied to the ocean surface.

3. Survey of the ON and OFF states and criteria for bistability

We will first examine the salt import properties of the oceanic circulation in the NADW OFF and ON state and the climatic changes associated with state transitions and changes in surface salt fluxes. The OFF state in the numerical model is characterised by two vertically stacked overturning cells originating from the Southern Ocean, as shown in Figure 2. The upper cell is the AAIW reverse cell found by Manabe and Stouffer (1988) and discussed in Saenko et al. (2003a); Sijp and England (2006). The deep cell is the inflow of Antarctic Bottom Water into the Atlantic. Sijp and England (2005) show how the Drake Passage causes a strong separation between the two Atlantic water masses in the OFF state. In the present study all Atlantic salinity changes in response to the anomalous salt flux in the OFF state occur in the upper 1200m for this reason. We obtained the shown OFF state in our model by applying a 300 year fresh water pulse to the North Atlantic and allowing in excess of 5000 years for equilibration to a stable OFF state. Also shown is the circulation of the OFF state when the anomalous salt flux $H$ applied to the North Atlantic is near the critical $H^*$ (see
Table 1 for mathematical terms). The unperturbed case exhibits a reverse cell of around 8 Sv, whereas the case close to critical perturbation exhibits a significantly reduced cell between 4 and 5 Sv.

We now turn to the model equilibria under different values of the anomalous salt flux $H$. We found increasingly large equilibration times of the order of 50,000 years were needed when approaching $H^*$, the critical flux beyond which there is only an ON state (some experiments are integrated for more than 50,000 years). We name the maximal salinity $\bar{S}$ possible for a stable OFF state $\bar{S}^*$, the associated critical total Atlantic surface flux $F_s^*$ (similar for reverse cell strength $M^*$). This is the “bifurcation point” of the equilibrium structure where the OFF state ceases to exist.

We analyse the steady state salt budget of the Atlantic, and use a balance $F_m + F_r + F_s = 0$. Here, $F_m$ is the salt flux associated with the AMOC (in the ON and OFF state) arising from the salinity difference between the inflow and outflow branch (see also Table 1), and given by (as in Rahmstorf 1996): $F_m = \int V < S >_x dz$, where $< S >_x$ denotes the zonal averaged salinity and $V$ the zonally integrated meridional velocity in the Atlantic at 30°s. The integral is taken over the entire depth range, see Rahmstorf (1996). In our rigid lid model, this means that we can compute this quantity by integrating over a fixed depth-range. We will see that for a small salinity difference between the South Atlantic and the Southern Ocean, this term behaves as a diffusive flux between the Atlantic and the Southern Ocean akin to the action of a gyre. Indeed, the definition of $F_m$ suggests that the remainder $F_r$ should incorporate all advective transports of salt by net velocities that disappear in the zonal average (unlike
the AMOC component), as well as diffusion. Finally, large changes in $H$ and a very strong AAIW reverse cell may influence the gyre circulation and the salinity distribution inside the Atlantic, and therefore the operation of $F_r$. In our domain of interest $F_m$ and $F_r$ can be meaningfully separated into independently acting mechanisms.

The Atlantic salinity $\bar{S}$ in ON and OFF and the Southern Ocean salinity in OFF as a function of the varying anomalous flux $H$ are shown in Figure 3a. Increasing $H$ here means a larger addition of salt to the North Atlantic. The SO salinity, defined here as the entire ocean south of $32^\circ$S, remains remarkably constant. In the ON state, there is a modest decrease in Atlantic salinity $\bar{S}$ under increasing $H$. This is due to a redistribution of salt from the global surface ocean to the deep ocean in ON. The most responsive to $H$ is the OFF state salinity, where increasing $H$ leads to a progressively more saline Atlantic until stable OFF states no longer exist where the curve ends. ON state Atlantic salinity $\bar{S}$ is always higher than OFF Atlantic state SO salinity, which in turn is always higher than OFF state Atlantic salinity $\bar{S}$. The ON state values end on the left for lowest $H$ when the ON state is no longer stable.

The Atlantic AMOC-related salinity flux $F_m$ in ON and OFF are shown in Figure 3b. Here, a positive value indicates an AMOC that imports salt into the Atlantic, and a negative value indicates salt export. The AMOC generally exports salt in ON and OFF, with the exception of the ON state for low values of $H$. $F_m$ always decreases for increasing $H$ in the ON state, that is, the AMOC resists Atlantic freshening everywhere in the ON state, adding to its stability. The OFF state AMOC has a strong freshening influence on the Atlantic, with negative $F_m$ values significantly below those in ON. However, the AMOC contributes to OFF
state stability only for a certain value range of $H$. Furthermore, there is some discrepancy between the zero location of the de Vries Weber metric ($\text{sign}(F_m)$ in ON) and the upper limit of the bistable regime.

By contrast, in the OFF state, increasing the forcing salt flux $H$ produces an AMOC salt export $F_m$ which increases only up to a point. $F_m$ reaches a maximum and then decreases, shortly before the OFF state breaks down. DeVries and Weber (2005) propose that the breakdown of the OFF state should occur when the sign of $F_m$ in the ON state becomes negative, ensuring that the interaction between the mean-state salinity distribution and a velocity anomaly arising from a fresh water perturbation acts to restore the system to the ON state. In our experiments, an OFF state exists even for values where the NADW cell (ON state) exports salt (Fig. 3b).

The de Vries Weber metric is derived from ideas relating to the salt feedbacks operating in the ON state, and appears to assume no fundamentally different feedbacks for the OFF state. We conjecture that the accuracy of this metric may then rely on a sufficiently weak AMOC FW import in the OFF state (see also Section 7). To test this, we have conducted two additional versions of the original sequence of experiments shown in Fig. 3b, where, in addition to the original $H$, we now apply an extra constant anomalous surface salt flux $G$ to the entire Atlantic basin of 0.2 Sv (strong OFF state average salt transport) and -0.2 Sv (weak OFF state average salt transport). Figure 4 shows $F_m$ as a function of a $H$ (moved horizontally to place the bifurcation points at $H = 0$) for the two additional sequences of experiments, along with the original sequence (Fig. 4b). The additional positive Atlantic-
wide surface salt flux leads to a stronger average (over the domain of $H$) FW import by the AAIW reverse cell (Fig. 4a) due to an enhanced vertical salinity gradient at 30 °S. As a result, the discrepancy between the zero point of the de Vries Weber metric ($\text{sign}(F_m)$ in ON) and the actual upper limit of the bistable regime is widened significantly, leading to a significant inaccuracy of this metric. In contrast, weaker average FW import by the AMOC in the OFF state leads to a high accuracy in the de Vries Weber metric (Fig. 4c). The standard experiments lie in between these two extremes (Fig. 4b), with a noticeable error in the de Vries Weber metric. The dVW metric appears to work best in the case of a relatively weak $F_m$ in OFF (Fig. 4c). We conjecture that this metric has a specific domain of validity, namely those situations where the FW export by the OFF state AMOC is sufficiently weak.

This specific example has been included to examine the validity of the dVW metric. The reasons for the relative $H$-dependent behaviour of $F_m$ in ON and OFF under changes in $G$ is not immediately clear to us. The basin-wide addition of salt via $G$ results in a larger $|F_m|$ in OFF for a given $M$ (figure not shown). This may be due to the different way in which $G$ affects the NADW formation regions compared to $H$. However, a detailed explanation is beyond the scope of this paper, and the results presented in this paragraph serve to illustrate the domain of validity of the dVW metric.

To examine the effect of atmospheric changes on the total surface salt flux, $\Delta F_s$ (with respect to $F_s$ when $H = 0$) as a function of $H$ is shown in Figure 5a. The total Atlantic net surface salt flux $F_s$ change is somewhat non-linear. The change in $F_s$ is also greater than the change in $H$, where the slope of $\Delta F_s$ as a function of $H$ is about $(1 + \beta) = 1.25$ on average. This is due to increasing temperature with increasing $H$ (Fig. 5b) leading to increased evaporation.
(Fig. 5c), as the reverse cell and its cooling effect on the Atlantic diminish with increasing $H$ (see below). Therefore, the factor $\beta$ represents the climate amplification on a fixed change in $H$. We can approximate this effect on net evaporation of the Atlantic as $\Delta F_s = (1 + \beta)H$ under a flux parameter change $H$. This significant effect illustrates the general cooling of the Atlantic by the AAIW reverse cell and the presence of atmospheric feedbacks in the OFF state.

In this section, it has been shown that the OFF state AMOC removes salt and is responsive to changes in $H$, where salt export initially increases with increasing $H$, a response that reverses beyond a certain $H$.

4. Model OFF state circulation, salinity and salt fluxes

Figure 6 shows that the AAIW reverse cell strength $M$ increases with the difference in average density over the upper 1200m between the NADW and AAIW formation regions $\delta \rho$ (see Fig. 1 for location). This is in agreement with Saenko et al. (2003a), who find that state transitions are controlled by this density difference. In our experiments, $M$ depends linearly on $\delta \rho$ to a remarkably high degree. Moreover, a vanishing density difference corresponds approximately to a vanishing $M$. We therefore assume a linear dependence of $M$ on $\delta \rho$. Indeed, Stommel (1961) already parameterized the AMOC strength as proportional to the density gradient between the northern and equatorial box, while Rooth (1982) used the high latitude southern and northern hemisphere box. Hughes and Weaver (1994) find a
proportionality of this kind for the NADW “on” state in an ocean circulation model. Here, no attempt is made at explaining this relationship in the OFF state but we will assume its validity. We also note that the relationship holds more accurately in the OFF state than in the ON state in our experiments (figure not shown).

Figure 7 shows the residual (gyre) term $F_r$ as a function of the salinity difference between the SO and the South Atlantic. $F_r$ is small when this salinity difference is small. The negative slope of $F_r$ means it acts to restore the salinity difference to a value close to 0. This diffusive action is what would be expected from the salinity transport by the gyres at 30° south, acting to reduce the salinity difference between the SO and the South Atlantic. This linear behaviour is not prescribed (as in box models), and the term actually becomes non-linear when the SO is significantly more saline than the South Atlantic in OFF.

Both the salt flux $F_r$ arising from the gyre circulation and the flux $F_m$ arising from the AMOC are sensitive to the Atlantic salinity $\bar{S}$, and it makes sense to examine these terms as functions of $\bar{S}$ for OFF. A minimal value of $F_m$ is attained at a certain $\bar{S}$, and the slope of $F_m$ changes for higher $\bar{S}$, indicating a reversal of its sensitivity to $\bar{S}$. In contrast, the slope of $F_{circ} = F_m + F_r$ is reduced with increasing $\bar{S}$, but remains negative throughout the bistability domain. In other words, small increases in $\bar{S}$ are always countered by $F_{circ}$ in the model. We will use a conceptual framework to examine whether this condition is also sufficient for the existence of the OFF state, providing us with a critical salinity $\bar{S}^*$ and $F_s^*$ beyond which no multiple equilibria occur.

Examination of the OFF states (figure not shown) suggests that the Atlantic meridional salin-
ity gradients do not depend in first order on the strength \( M \) of the AAIW reverse cell. Instead, while the cell affects the average Atlantic salinity, salinity gradients are dominated by gradients in surface fluxes and also the horizontal (gyre) circulation throughout the Atlantic in OFF. See Appendix B for a box-analysis. Therefore, a linear relationship between local salinity and global average salinity can be expected in the Atlantic when varying \( H \). Figure 8a) shows that this holds for NADW formation regions in our model. We keep in mind that this linear relationship incorporates the effect of the location where \( H \) is applied. Note that none of these linearities exist for salinity distributions at different \( H \) in the ON state.

The next two sections examine the dependence of \( F_r, F_m \) and \( M \) on \( \bar{S} \) in the OFF state more closely and explain the behaviour of \( F_m \) and \( F_{circ} \) shown in Figure 7b.

5. Basic premises to describe the OFF state

The relative simplicity of the OFF state allows us to make some basic premises, namely:

1. The salinities of the NADW formation region (see Fig. 1) and the South Atlantic (between 30 °s and the equator) remain linearly related to \( \bar{S} \) under changes in \( H \).

2. The properties of cool fresh AAIW constituting the lower branch of the reverse cell remains relatively unchanged compared to Atlantic surface salinity under changes in \( H \).

3. The strength of the reverse cell \( M \) depends linearly on the density difference \( \delta \rho_{NADW-AAIW} \)
between the NADW and the AAIW formation region.

4. The term $F_r$ acts linearly to reduce the salinity difference between the South Atlantic and the Southern Ocean to 0, and so causes $\tilde{S}$ to tend towards a certain $\bar{S}_r$.

5. North Atlantic density changes are dominated by salinity (not salinity) for steady states forced by changes in $H$.

See for instance Figure 7 in Sijp and England (2005) for premise 5. We will show that the OFF states in the bistable regime of our equilibrium diagram have this property.

The basic premises lead to the basic formulas:

$$F_m = (v_1 \tilde{S} + v_2)(M - M_0) \quad (1)$$

$$M = d_1 \tilde{S} + d_2 \quad (2)$$

$$F_r = r_1 (S - \bar{S}_r) = r_1 \tilde{S} + r_2 \quad (3)$$

Where $v_1, v_2, d_1, d_2, r_1, r_2$ are constants. In equation (1), the assumption (1) is used that the salinity of the South Atlantic is linearly related to the average Atlantic salinity under changes in $M$, and assumption (2) that the changes in Atlantic salinity are large compared to those in the AAIW formation regions (and therefore the inflow of AAIW in the reverse
cell), as confirmed in Figure 3a. We can therefore express the vertical salinity difference between the upper and lower branch of the reverse cell as \( h_1 \bar{S} + h_2 - S_{AAIW} = v_1 \bar{S} + v_2 \).

The vertical integral \( F_m = \int V \frac{< S >_x}{dz} \) at 30°s (see above) could be approximated by \( M \Delta \bar{S}_{vertical} \), where \( \Delta \bar{S}_{vertical} \) is the salinity difference between the upper South Atlantic and the salinity of the underlying AAIW. To reduce error in this coarse approximation, a small reference overturning rate \( M_0 = 1.5Sv \) is introduced to include the effect of the reduction in the vertical extent of the cell (shrinking) as \( M \) is reduced. That is, the vertical extent of the reverse cell is expected to be reduced to within the upper thermocline depth when \( M = 1.5Sv \) compared to the vertical salinity gradient so that no net salt transport can be expected in this case. Figure 8b shows that an excellent linear fit occurs in our model over the entire range of salinity values under study. We choose the value of \( M_0 \) so as to obtain a good linear fit. The approximation renders values of \( F_m \) for \( M < M_0 \) dubious, restricting the validity of approximation (1) to values \( M > M_0 \).

In Equation (2), the fact that the average salinity of the NADW formation regions depends linearly on the average Atlantic salinity under changes in the AAIW reverse cell (premise 1) is used, and that the strength of the reverse cell \( M \) is proportional to \( \rho_{NADW} - \rho_{AAIW} \) (assumption 3) and that \( \rho_{NADW} \) depends linearly on \( \bar{S} \) where the effect of changes in temperature is neglected. This gives a chain of linear relationships leading to Equation (2), where the coefficients \( d_1 \) and \( d_2 \) could be expressed in terms of the dependence of \( \rho \) on \( \bar{S} \) and the linear coefficients of the spatial relationships for salinity. Figure 8c shows that this linear relationship occurs in our model.
In Equation (3) $F_r$ attempts to restore $\bar{S}$ to a certain $\bar{S}_r = -r_2/r_1$, so $r_1 < 0$ and $r_2 > 0$. $r_1$ is related to the strength of the gyres and the diffusion in the South Atlantic, and other factors. More generally, $\bar{S}_r$ is the Atlantic salinity expected in the absence of surface fluxes and salt transport by the AMOC. Figure 8d shows that the gyre salt flux $F_r$ exhibits a good linear dependence on $\bar{S}$ near the critical $\bar{S}^*$ in our model, with significant departures for fresh values of $\bar{S}$ away from $\bar{S}^*$. Again, caution should be taken in the interpretation of $F_r$ as a gyre term with fixed gyre strength that acts linearly when considering the OFF state at very fresh Atlantic values.

The basic premises allowed us to express the strength $M$ of the reverse cell and its action on the vertical salinity gradient in the South Atlantic linearly in $\bar{S}$ in the OFF state. Also, the gyre salt flux $F_r$ is expressed linearly in $\bar{S}$, representing a restorating action of $\bar{S}$ to a certain value. This allows us to formulate the OFF state Atlantic salt budget in terms of $\bar{S}$ alone in the next section.


We will now use the simple linear equations 1-3 from the previous section to find an expression for the Atlantic salt budget in terms of $\bar{S}$. This allows us to examine why no OFF state occurs when $H$ exceeds a certain value $H^*$ and why the OFF state is unstable for certain $\bar{S}$. The total circulation-related Atlantic salt flux $F_{circ} = F_m + F_r$ depends on local Atlantic salinities, yet this circulation flux term determines the average Atlantic salinity $\bar{S}$ for each
fixed $F_s$. We examine the solutions $\bar{S}$ as a function of $F_s$. Equations 1-3 yield an evolution equation for $\bar{S}$:

$$V_{Atl} \frac{\partial \bar{S}}{\partial t} = F_m + F_r + F_s = a\bar{S}^2 + b\bar{S} + (c + F_s)$$

(4)

Here, $V_{Atl}$ is the volume of the Atlantic. The second equality arises from the fact that a quadratic expression for $F_m$ is added to a linear expression for $F_r$, obtaining $F_{circ} = a\bar{S}^2 + b\bar{S} + c$, and $F_s$ is considered fixed and grouped with the constant term. $F_m$ can be computed by substituting the expression for $M$ in Equation (2) into the expression for $F_m$ in Equation (1), yielding a quadratic function in $\bar{S}$. The coefficients $a, b$ and $c$ in Equation (4) can be constructed from the coefficients in equations 1-3. The solutions where $\bar{S}$ does not change with time occur when $F_m + F_r + F_s = 0$. That is, for each fixed $F_s$ there are two salinities $\bar{S}_1$ and $\bar{S}_2$ where all salt fluxes are balanced:

$$\bar{S}_i = \frac{-b \pm \sqrt{b^2 - 4a(c + F_s)}}{2a}$$

(5)

We can choose the naming of $\bar{S}_i$ so that $\bar{S}_1 \geq \bar{S}_2$. Expression (5) allows us to regard $\bar{S}_i$ as a function of $F_s$. We will see later that the (small solution) salinity $\bar{S}_2(F_s)$ represents a stable solution, whereas the (larger solution) salinity $\bar{S}_1(F_s)$ is unstable.

The existence of the solutions $\bar{S}_i$ depends on a zero or positive discriminant of the right side of Equation (4). This discriminant is $D(F_s) = b^2 - 4a(c + F_s) = D_{circ} - 4aF_s$, where $D_{circ} = D(F_s = 0) = b^2 - 4ac$, the discriminant of the quadratic function $F_{circ}(\bar{S})$. The
domain of real solutions $\tilde{S}_i$ ends for a certain critical $F_s = F^*_s$ that satisfies $D(F^*_s) = 0$. This is the only case where $S_1 = \tilde{S}_2 = -\frac{b^2}{2a} = \tilde{S}^*$, where $\tilde{S}^*$ denotes the salinity at $F^*_s$. We have:

$$F^*_s = \frac{D_{circ}}{4a} = \frac{b^2}{4a} - c \quad (6)$$

Note that this point is characterized by $\frac{\partial F_{circ}}{\partial S} = 0$, and that $\frac{\partial F_{circ}}{\partial S} \leq 0$ is a necessary and sufficient condition for the existence of the OFF states. It is therefore the non-linearity of the large-scale OFF state oceanic circulation that determines the critical $\tilde{S}^*$ and $F^*_s$. Using (5), the salinity $\tilde{S}$ of the stable states can be expressed as follows (using the critical salinity $\tilde{S}^* = -\frac{b^2}{2a}$, see below):

$$\tilde{S} = \tilde{S}^* - \sqrt{\frac{F^*_s - F_s}{a}} \quad (7)$$

Similarly, the unstable equilibria are given by $\tilde{S}_{unst} = \tilde{S}^* + \sqrt{\frac{F^*_s - F_s}{a}}$. The associated reverse cell strength $M$ can be found from $\tilde{S}_{unst}$ using (2).

To see how well the stable salinity $\tilde{S}$ calculated in (5) and the associated $M$ approximate the behaviour of the model in the OFF state, Figure 9 shows $M$ and $\tilde{S}$ calculated via Equation (5) and $M$ and $\tilde{S}$ produced by the numerical model as a function of the total Atlantic surface salt flux. We calculate $M$ in the numerical model as the minimal value of the overturning streamfunction around 30° south and in the upper 1200m (see Figure 2). The curves of $M$ and $\tilde{S}$ calculated via Equation (5) ends on the right where solutions no longer exist due to $F_s > F^*_s$ (see eq 6). There is excellent agreement between the calculated solutions and
the model over the entire domain of $F_s$ for $M$. Agreement for $\bar{S}$ is also excellent, with the exception of an increasing error at the left-most extreme of the domain. Figure 8 suggests this may be due to an increasing non-linearity of $F_r$ for low values of $\bar{S}$. There is a very close agreement for the critical values $M^*$ and $\bar{S}^*$ between the calculations and the numerical model. Our Equations 1-3 were not tuned to give an agreement in critical values, supporting the validity of our approach. Figure 10a shows the flux balance terms $F_m$, $F_r$ and $F_s$ (that should balance to 0) as a function of $F_s$ for the model. Comparing this model balance to the balance arising from the calculated functions $F_m$ and $F_r$ (see Eq. 3) shown in Figure 10b are in excellent agreement. This shows that our calculated terms shown in Fig 10b behave in a way that is very similar to the behaviour of the model terms (Fig. 10c) near the critical $F_s^*$. Note that the climate amplification (see Fig. 5) only affects the transient behaviour towards the steady states shown here (see below), which are shown for the unchanged $F_s$ of the final steady state.

In both the mathematical formulation and the numerical model the slope of $F_m$ with respect to $F_s$ changes sign, where further increases in salinity $\bar{S}$ are now re-inforced by changes in $F_m$. The change in slope is because of two competing effects: 1. Increased Atlantic salinity means greater salt export in the upper branch of the AAIW reverse cell (while property changes in the inflowing lower branch remain small). 2. Increased salinity means denser NADW formation regions and therefore a reduced strength $M$ of the AAIW reverse cell. "The minimal value of $F_m$ i.e. the maximum salt export is given by the maximum product of these two effects. Our mathematical formulation uses a linear approximation of these two effects, where $F_m$ is their product. This allows us to compute the minimal $F_m$ (see
below). The gyre term $F_r$ increases for increasing $F_s$, while the OFF-state AMOC term $F_m$ begins to decrease with $F_s$ near $F_s^*$. This means that close to $F_s^*$, the gyre term $F_r$ becomes responsible for matching the increases in $F_s$ as $F_s$ approaches $F_s^*$, ensuring the continued stability of the OFF state over a small range of $F_s$. We can calculate the critical salinity $\bar{S}^* = -\frac{b}{2a} = 34.63 \text{kg/m}^3$ and the critical flux $F_s^* = \frac{D_{circ}}{4a} = 0.122 \text{Sv}$, against $F_s^* = 0.12 \text{Sv}$ diagnosed from the numerical model with $10^{-3} \text{Sv}$ accuracy. Note that $\bar{S}^*$ can also be calculated from $\left|\frac{\partial F_m}{\partial \bar{S}}\right| = \left|\frac{\partial F_r}{\partial \bar{S}}\right|$, where increases in $F_m$ with $\bar{S}$ are exactly matched by decreases in $F_r$.

The good agreement between our approximation of the model behaviour of $\bar{S}$, $M$ and the salt fluxes $F_m$ and $F_r$ arising from Equation (4) and the model, including an accurate prediction of the critical flux $F_s^*$, allows us to use the quadratic expression $F_{circ}(\bar{S})$ to examine why multiple equilibria are restricted to values $F_s < F_s^*$ in the model. The salinity $\bar{S}$ at which $F_{circ}(\bar{S})$ attains a minimal value can be obtained by setting $\frac{\partial F_{circ}}{\partial \bar{S}} = 0$, giving $\bar{S} = -\frac{b}{2a}$. The critical Atlantic salt flux $F_s^*$ (see Equation 6) occurs at $-F_{circ} = F_s^* = \frac{D_{circ}}{4a}$. Therefore, $F_s$ attains the critical value $F_s^*$ when $F_{circ}$ is minimal (calculate $F_{circ}(-b/2a)$). In other words, the sign of $\frac{\partial F_{circ}}{\partial \bar{S}}$ determines the existence of multiple equilibria in our model. Intuitively, increases of $F_s$ beyond $F_s^*$ can no longer be matched via an increase in $\bar{S}$ leading to an increase in $F_{circ}$ because the sign of $\frac{\partial F_{circ}}{\partial \bar{S}}$ has changed. This leads to a transition out of this domain and into a stable NADW ON state.

Figure 11 shows the Atlantic salinity in the model in ON and OFF as a function of the anomalous North Atlantic salt flux $H$, showing the bistability behaviour in our model. The
unstable equilibria are derived from the stable equilibria by mirroring in $\bar{S}^*$ i.e. in a line (see Equation 7). States with salinities above the unstable branch and inside the bistable regime are expected to go up and are captured by the stable ON branch, whereas initial conditions below the unstable branch are expected to evolve towards the OFF branch, provided the conditions encapsulated in equations 1-3 are met. There is no stable ON-state capture of upward moving states above the unstable branch to the left of the ON state regime. Nonetheless, OFF state dynamics require a salinity $\bar{S}$ below the unstable branch to allow a collapse onto the stable OFF state branch, and the fate of higher salinity values appears more complicated. However, the unstable branch salinity values to the left of the ON state regime are high, and represent a significant excess with respect to the stable equilibrium values associated with $H$ in OFF, and likely any transient ON states. We speculate that in this case initial conditions above the unstable branch give rise to a transient salt exporting ON state, allowing sufficient salt loss from the Atlantic to yield a state below the unstable OFF state branch. The diagram is easily recognised as a phase portrait depicting a saddle node bifurcation. The capture of upward moving states by the ON branch and the absence of stable ON states to the left of the bistable regime indicate the presence of another node, now bounding the ON state domain on the left, also identified as a saddle node or limit point by Dijkstra and Weijer (2005). Here, we do not discuss this latter node, and focus on the existence conditions for the OFF state. A further discussion is contained in the Appendix A.

In summary, it has been shown that the basic premises listed in Section 5 allow us to find stable equilibria for the evolution equation of $\bar{S}$ in Section 6. The stable equilibria require $\frac{\partial F_{\text{circ}}}{\partial \bar{S}} \leq 0$. This is confirmed in Fig. 7b. The critical $H^*$ and $S^*$ are found from $\frac{\partial F_{\text{circ}}}{\partial \bar{S}} = 0$. 25
The upper limit to $H$ and $F_s$ where an OFF state exists is determined by the minimal $F_{circ}$. The dependence of $F_{circ}$ on $\bar{S}$ allows us to calculate the $\bar{S}^*$ where this occurs. This essentially amounts to an analytical expression for a limit point to the bistable domain, also called a saddle node bifurcation. This saddle node bifurcation has been calculated numerically by Dijkstra and Weijer (2005), where the unstable branch and saddle nodes are determined. They solve for a state parameter of much larger dimension, including velocity and tracer values at each grid cell. The simplicity of our state parameter allows us to identify the physical mechanism behind the saddle node bifurcation where OFF states are eliminated.

7. Summary and Conclusions

We have deduced for the first time a natural and accurate analytical saddle node condition for the maximal surface salt flux forcing $H^*$ and maximal Atlantic salinity $\bar{S}^*$ for which an AMOC OFF state exists in a numerical climate model. The OFF state is one in which there is no NADW formation, but in our model the presence of an AAIW reverse cell is critical to the OFF state. We formulate a simple set of equations to describe the OFF state. This sheds light on the mechanisms behind OFF state stability in our model, and allow us to identify the non-linear behaviour of the AAIW reverse cell as the underlying mechanism determining the existence domain of OFF states. Our approach also yields an accurate reconstruction of OFF state numerical model behaviour over a significant range of values for the salt flux forcing $H$. Because of the simplicity of its behaviour, only two OFF states in the numerical model are required to reconstruct OFF state behaviour under varying $H$ and the saddle node $H^*$. The
bistability indicator proposed by DeVries and Weber (dVW, 2005), appears to work in our experiments only when the AAIW reverse cell impact on the Atlantic salt budget is weak. In contrast, this criterion fails when the AAIW reverse cell impact becomes strong.

We have examined the response of the salt transport terms $F_m$ (AMOC) and $F_r$ (gyre) to changes in the surface boundary salt flux $F_s$, changes in which are dominated by the forcing flux $H$. The AMOC exports salt dynamically in the OFF state ($F_m < 0$), confirming the freshening role of the AAIW reverse cell suggested by Gregory et al. (2003), Saenko et al. (2003b) and Sijp and England (2006). Changes in average stable OFF state salinity $\bar{S}$ are opposed by changes in the total oceanic salt flux $F_{circ} = F_m + F_r$, ensuring the persistence of this state, in agreement with Gregory et al. (2003), who suggest that both the NADW "on" and "off" states tend to reinforce their salinity distributions in this model. (We do not examine the existence criteria for a stable ON state; this will be pursued in a future study.)

OFF states occur only as long as the total OFF state oceanic circulation is able to balance the Atlantic surface salt flux. Hence, the OFF state is stable only when $\frac{\partial F_{circ}}{\partial \bar{S}} \leq 0$ i.e. the limit point occurs at the OFF state circulation of maximum $F_{circ}$. This condition is not only necessary but also sufficient. In our model, $F_{circ}$ has a maximum because $\frac{\partial F_{circ}}{\partial \bar{S}}$ changes sign beyond a certain $\bar{S}$, a behaviour associated with the non-linearity associated with the AAIW reverse cell. Here, increasing $H$ leads to an increasing vertical salinity contrast at $30^\circ$S but decreasing $M$, yielding opposing effects on $F_m$. The derivative $\frac{\partial F_{circ}}{\partial \bar{S}}$ can be interpreted as the salt-flux responsiveness of the oceanic system to salt additions.

The failure of the dVW metric in the regime of a strong reverse cell action on the Atlantic salt
budget suggests a specific domain of validity for this criterion. Its accuracy appears to depend on the magnitude of the reverse cell impact on the Atlantic salt budget \( (F_m \text{ in OFF}) \). Indeed, the success of our analytical OFF state condition shows the importance of the AAIW reverse cell in determining the existence of the OFF state in the numerical model, as our condition is based on the non-linear dynamics of this cell. Failure of the dVW metric in our results shows it is impossible to infer the OFF state existence from ON state characteristics relating to FW export alone when the AAIW reverse cell is strong. This is presumably because the nature of the reverse cell cannot be anticipated from the ON state.

It is significant that the OFF state is essentially a weak ON state in the results of DeVries and Weber (as in 2005), where the dVW metric works well. This suggests similar dynamics for the ON and OFF state regimes in their study, leading to a reliable dVW metric. Dijkstra and Weijer (2005); Huisman et al. (2010) also find the dVW metric to be a good indicator for the existence of OFF states. The OFF state circulation in these studies is characterised by AABW upwelling and little inflow at intermediate depth. The difference with our results (characterised by an AAIW reverse cell) may arise from the significantly different viscosity and higher vertical diffusivity employed in their model. This renders a direct comparison difficult. For example, the AABW cell in their model is likely to behave differently from our AAIW reverse cell, and would also not depend on the AAIW-NADW formation region density difference. We speculate that the salt flux associated with their AABW cell could be only weakly responsive to Atlantic salinity changes, perhaps due to a weakened vertical salinity gradient associated with high vertical diffusivity. This suggests that different models may have different types of OFF states.
The counter-example presented in our study shows that the dVW metric is not universally valid and so prompts the question whether observations of the present day system which is evidently in an ON state are sufficient to determine whether the present-day climate admits a stable OFF state. It would be useful to be able to estimate characteristics of a possible OFF state from the distribution of surface fluxes in the observed ON state, allowing a prediction to be made about behaviour in the real ocean. Future work could consist of attempts to find these relationships first in models.

We use a vertical diffusivity value $K_v = 0.3cm^2/s$ in the thermocline. An increase in $K_v$ could extend the bistable regime as the AAIW reverse cell strength $M$ of the OFF state increases with increasing $K_v$ (Sijp and England 2006). However, this is not given \textit{a priori}, as surface salt fluxes also play a role. A relatively modest AAIW reverse cell (perhaps associated with low $K_v$) could still have significant consequences for bistability whenever its effect on the Atlantic salt budget is strong due to the specific distribution of surface salt fluxes. This motivates the need for further investigation into the effect of $K_v$ on bistability via its effect on the AAIW reverse cell.

Our numerical model employs a one layer atmosphere, yet despite this simplification a significant increase in Atlantic surface salt flux occurs in the ON state compared to the OFF state due to a change in evaporation. This emphasises the importance of an atmospheric response in bistability. In addition to thermal effects, atmospheric circulation changes are expected to have a significant effect on AMOC bistability, as shown by Arzel et al. (2008). Future work should focus on the role of climate feedbacks in bistability.
Our results give new theoretical insight into the possible workings of the OFF state, allows for a precise calculation of $H^*$, suggests a domain of validity for the ON-state based indicator of dVW and introduces the question of whether a precise yet universal ON state-based indicator might be fundamentally impossible. Also, our method can be practically applied to climate models with accessible OFF states.

Appendix A. Factors affecting OFF state stability

Equation (4) and its solutions allows us to examine the end result of the evolution of various initial conditions where NADW is absent. We will now examine when initial conditions are expected to lead to an OFF state, provided the basic premises of Section 5 hold. We also briefly examine the relative roles of the reverse cell and the South Atlantic gyre in maintaining OFF state stability.

To examine the stability of possible values of $\bar{S}$, Figure 13 shows a schematic representation of $F_{\text{circ}}$ as a function of $\bar{S}$ (the vertical axis is a salt flux $F$). Stable values of $\bar{S}$ occur where a horizontal line at $F = -F_s$ intersects with $F_{\text{circ}}$. The stable equilibrium $\bar{S}$ is the intersection $(\bar{S}, -F_s)$ at the lowest $\bar{S}$, and the unstable equilibrium is the intersection $(\bar{S}_{\text{unst}}, -F_s)$ at the highest $\bar{S}$. Note that no intersection occurs for values $F_s > F_s^*$. Here, there is no Atlantic salinity where $F_{\text{circ}}(\bar{S})$ can match $F_s$, and no stable value for $\bar{S}$ occurs.

As will be seen in part II of this paper, the evolution Equation (4) describes realistic time dependent behaviour under certain conditions. The evolution equation should also apply to
small perturbations $\delta \bar{S}$ around an equilibrium state. A small positive anomaly $\delta \bar{S}$ around a stable flux equilibrium $(\bar{S}, -F_s)$ represents an excess in average Atlantic salinity associated with a small excess in salt export by the oceanic circulation, leading to freshening and therefore a reduction in the initial salt excess $\delta \bar{S}$, restoring $\bar{S}$ exponentially to its original stable value (similar for negative $\delta \bar{S}$). In contrast, all $\bar{S}_{\text{unst}} > \bar{S}^*$, so for the unstable flux equilibrium $(\bar{S}_{\text{unst}}, -F_s)$, $\frac{\partial F_{\text{circ}}}{\partial \bar{S}}(\bar{S}_{\text{unst}}) > 0$ and therefore initial exponential growth of any small perturbation $\delta \bar{S}$ on $\bar{S}_{\text{unst}}$. This is because any salt perturbation is re-inforced by the oceanic circulation.

The non-linear nature of the circulation response $F_{\text{circ}}$ in the OFF state arises solely from $F_m$, which can be expressed quadratically in $\bar{S}$ (the other component $F_r$ of $F_{\text{circ}}$ is linear).

It is therefore instructive to briefly examine the relative contributions of $F_r$ and $F_m$ to the behaviour of $F_{\text{circ}}$. We can compute the discriminant $D_m$ of $F_m$ as $D_m = D_{\text{circ}} - 2r_1 b + r_1^2 + 4a r_2$. Given that $F^*_s = \frac{D_{\text{acc}}}{4a}$ (see Equation 6), where $F_{\text{circ}}$ is minimal, and that $F_m$ takes a minimal value $F_m^{\text{min}}$ at $F_s = \frac{D_m}{4a}$, so $F^*_m = -F_m^{\text{min}} - F_r(\bar{S}^*) - \frac{r_2^2}{4a}$ (subtract $F_r$ from $F_{\text{circ}}$). It can be shown from this that $F_m^* - F_m^{\text{min}} = \frac{r_2^2}{4a}$. This is the amount of (destabilizing) increase in $F_m$ (see Fig. 10) that the system can tolerate under increasing $F_s$ before OFF becomes unstable. For weak gyre responsiveness $r_1$ relative to the the non-linearity of the reverse cell $a$, $\frac{r_1^2}{4a} = 0$ ($r_1 << a$), and OFF state existence is approximately determined by $\frac{\partial F_m}{\partial \bar{S}} < 0$. 

31
Appendix B. Linear relationships between salinities

Here, we examine the reasons behind the linear relationships stated in Section 5. The linear dependence of $F_r$ on the South Atlantic salinity (while SO salinity remains constant) and of $\rho_{NADW}$ on the North Atlantic salinity are more trivially anticipated, so we will focus only on the linear dependence of local (e.g. NA) salinity on the average Atlantic salinity $\bar{S}$ in the OFF state. To this end, we imagine an idealized upper Atlantic basin of volume $V$ (to 1200m depth) subdivided into disjunct adjacent rectangular volumes $V_k$ with average salinities $S_k$ ($k = 1, 2, \ldots$) as shown in Figure 12, each volume reaching from the surface to 1200m depth. Then, $V\bar{S} = V_1S_1 + V_2S_2 + \ldots + V_nS_n$. The surface flux $F_s$ is subdivided accordingly into $F_{s,k}$, which sum to $F_s$. Salt transport from $V_{k+1}$ to $V_k$ takes place due to gyre circulation (and diffusion), and is represented by a diffusive flux $F_{r,k} = r_k(S_{k+1} - S_k)$ for coefficients $r_k$ ($k=0,1,2,\ldots$) associated with the gyre strength at each location. Then in steady state, $F_{r,k} - F_{r,k-1} + F_{s,k} = V_k\frac{\partial S_k}{\partial t} = 0$, and we take $r_0 = 0$ to represent a northern boundary.

Representing a model of the OFF state, we denote the fixed salinity of the Southern Ocean by $S_{n+1}$, and first assume that the OFF state AMOC is only active in the southern volume by introducing an extra term so that the total oceanic salt flux between $V_n$ and $V_{n+1}$ is $(\tilde{r}_n + \gamma S_1)(S_{n+1} - S_n)$ there, where the constant $\tilde{r}_n$ is a modified version of $r_n$ to incorporate the effect of the reverse cell, as is $\gamma S_1$, where $\gamma$ a negative constant. The term containing $S_1$ arises from the dependence of $M$ on $\rho_{NADW} - \rho_{AAIW}$, and we assume the reverse cell is acting vertically across the salinity contrast between $S_2$ and $S_3$.

Linear dependence of $S_k$ on $\bar{S}$ arises when $\frac{\partial S_k}{\partial S}$ is constant for a change $\delta F_{s,1}$ in $F_{s,1} =$
H, where \( H \) corresponds to the anomalous North Atlantic surface flux \( H \) discussed in the previous sections. Here we will examine this system for \( n = 2 \) (so \( S_3 \) is the fixed SO salinity). The flux balance for \( V_1 \) allows us to express \( S_1 = S_2 + \frac{F_{s,1}}{r_1} \). Then, via \( V \bar{S} = V_1 S_1 + V_2 S_2 \), we obtain \( \bar{S} = S_1 - \frac{V_2 F_{s,1}}{V} = S_2 + \frac{V_1 F_{s,1}}{V} \). So \( \delta S_2 = \delta \bar{S} - \frac{V_1 \delta F_{s,1}}{V} \). Now, in the absence of the AMOC, we could solve for \( \bar{S} \) by noting that \( F_{r,2} + F_s = 0 \) and obtain constant ratios \( S_k \bar{S} \). However, incorporating AMOC effects, we examine the Atlantic-wide balance \( F_s + (\bar{r}_2 + \gamma S_1)(S_3 - S_2) = 0 \) by introducing a perturbation \( \delta F_s \), leading to \(-\gamma \delta S_1 (S_3 - S_2) + (\bar{r}_2 + \gamma S_1) \delta S_2 = \delta F_s \). The first term is a product of salinity differences, and for convenience we will neglect it. Substituting \( \delta S_2 \) gives \( \delta \bar{S} = \frac{V_1 \delta F_{s,1}}{V} + \frac{\delta F_s}{r_2 + \gamma S_1} \).

We now express surface flux changes in terms of \( H \) and so also incorporate the climate feedback \( \beta \). We assume \( H \) is applied solely inside \( V_1 \), while \( F_s \) is subject to a homogeneous amplification \( \delta F_s = (1 + \beta)H \) and so \( \delta F_{s,1} = (1 + \frac{V_1}{V} \beta)H \). Then: \( \frac{\delta S_2}{\delta S} = 1 - \frac{V_1(\bar{r}_2 + \gamma S_1)(V + V_1 \beta)}{V_1(V + V_1 \beta)(\bar{r}_2 + \gamma S_1) + V^2 r_1 (1 + \beta)} = 1 - (1 + \frac{V^2 (1 + \beta)}{V_1(V + V_1 \beta)(\bar{r}_2 + \gamma S_1)} r_1)^{-1} \). The last term in the expression contains the sum of a quotient of volumes times the quotient of the NH gyre strength and the sum of the SH gyre strength plus the AMOC-strength. It is reasonable to take the volume quotient close to 2, assuming volumes of similar size and small \( \beta \). Also, it is reasonable to assume the gyre strengths and the AMOC-term to be of similar magnitude, leading to an average sensitivity \( \frac{\delta S_2}{\delta S} = 0.5 \). Assuming \( \bar{r}_2 = 1.5 r_2 \), \( \gamma S_1 = 0.5 r_2 \) and perturbing \( \gamma S_1 \) with values \( \delta \gamma S_1 \) between \( \pm 0.5 \gamma S_1 \) we find less than a fraction 0.05 change in \( \frac{S_2}{S} \), a low sensitivity. Finally, a similar sensitivity is found for \( S_1 \). The main point here is that although \( \frac{\delta S_2}{\delta S} \) depends on \( S_1 \), indicating a non-linear relationship between \( S_2 \) and \( \bar{S} \), it is plausible that this non-linearity is weak, as we find in our GCM. This is unlike \( \frac{\delta F_{s,1}}{\delta S} \), where sensitivity is
large. The weak non-linearity disappears altogether in the absence of an OFF state AMOC, and models with a weaker reverse circulation in the OFF state than our model are expected to have a weak non-linearity in the relationships between local and global Atlantic salinity.

The near-linearity between $S_k$ and $\bar{S}$ is derived for an OFF state AMOC. Examination of Fig. 2 shows that the OFF state AMOC (reverse cell) horizontal mass transport between adjacent volumes inside the Atlantic is weak compared to that in a typical ON state, particularly north of the equator, justifying our restriction of the AMOC term to the southern volume. A strong AMOC introduces coefficients for all $S_k$ that depend on the overturning rate $M$ and therefore $S_1$. This leads to products $S_kS_l$ in the expression of $S_2$ as functions of $S_1$ etc, leading to stronger non-linearity. Indeed, the linear relationships do not hold in the ON state (figure not shown), where the AMOC acts on salinity gradients throughout the basin. This is due to a strong AMOC-related transport between volumes $V_k$, introducing a significant non-linearity in the relationships.

Acknowledgements. We thank the University of Victoria staff for support in usage of the their coupled climate model. This research was supported by the Australian Research Council and the Australian Antarctic Science Program. Jonathan Gregory was partly supported by the NCAS-Climate Programme, and partly by the Joint DECC and Defra Integrated Climate Programme, DECC/Defra (GA01101). We thank Olivier Arzel for commenting on a previous version of this study and Marc d’Orgeville for an internal review of the present manuscript. We thank three anonymous reviewers for helpful comments that led to an improved paper.
References


List of Figures

1  Schematic depiction of ocean areas referred to in the text. The AAIW and NADW formation regions are marked by black lines. A black line also marks our working definition of the Southern Ocean. ................................. 45

2  Atlantic meridional overturning circulation of the NADW OFF state for (a) the unperturbed OFF experiment OFF (b) the critical OFF state for high anomalous salt flux $H$. No other additional surface salt fluxes are applied. Values are given in Sv ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ sec}^{-1}$). ................................. 46

3  Results of experiments where the model is run to equilibrium for a range of constant anomalous salt fluxes $H$ applied to the North Atlantic Deep Water formation region (see fig. 1). (a) Average Atlantic salinity $\bar{S}$ for ON (black) and OFF (red). Also, the OFF state salinity of the Southern Ocean defined here as the entire ocean south of $32^\circ\text{S}$ (green). (b) the salt flux $F_m$ associated with the OFF state (red) and the ON state (black). Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where $1 \text{ Sv} = 10^6 \text{ m}^3 \text{ sec}^{-1}$). A positive salt flux represents the addition of salt to the Atlantic. ................................. 47
AMOC-salt flux $F_m$ for three different versions of the ON (black) and OFF (red) states as in Fig. 3b. Two additional versions of the original experiment shown in Fig. 3b have been conducted, where, in addition to the original $H$, we now apply a constant anomalous surface salt flux to the entire Atlantic basin of (a) 0.2 Sv (strong OFF state average salt transport), (b) 0 Sv (i.e. the original experiment, standard OFF state average salt transport) and (c) -0.2 Sv (weak OFF state average salt transport). The only varying parameter within each panel is the anomalous salt flux $H$ applied to the North Atlantic. The rectangle indicates the point where the deVries Weber metric predicts the upper limit of the bistable regime to occur. The horizontal axis is translated so as to let $H = 0$ at the bifurcation point. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.
Results of experiments where the model is run to equilibrium for a range of constant anomalous salt fluxes $H$ applied to the North Atlantic Deep Water formation region (see fig. 1). (a) The total net Atlantic surface flux change $\Delta F_s$ with respect to the unperturbed OFF equilibrium ($H = 0$) as a function of the salt flux perturbation $H$ (red). $H$ is shown for comparison (black). (b) The average Atlantic temperature as a function of $H$. The AAIW reverse cell cools the Atlantic. (c) The average Atlantic evaporation as a function of $H$. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.

(a) AAIW reverse cell strength $M$ against the density difference $\Delta \rho_{NADW-AAIW}$ between the NADW and AAIW formation regions in the OFF state. A best linear fit is included (dashed).

Oceanic flux terms in the OFF state. (a) the resultant term $F_r = -F_m - F_s$ against the salinity difference between the South Atlantic (between 30 °S and the equator) and the Southern Ocean. (b) AMOC salt flux $F_m$ (black) and total salt flux $F_{circ} = F_m + F_r$ arising from circulation (red) against the average Atlantic salinity $\bar{S}$.
Values against the Atlantic salinity $\bar{S}$ of (a) mean salinity of NADW formation region, (b) $\frac{F_m}{M-M_0} = (v_1 S + v_2)$, (c) the AAIW reverse cell strength $M$ and (d) the gyre salt flux $F_r$. Best linear fits (dashed) are added for the 5 model equilibria with highest $S$ (and closest to the critical flux $F^*_s$).

Average Atlantic salinity $\bar{S}$ (a) and AAIW reverse cell strength $M$ (b) against total net Atlantic salt flux $F_s$ for the model (black) and our equations (red).

The flux balance of $F_m$ (black) $F_r$ (green) and $F_s$ (red) such that $F_m + F_r + F_s = 0$ for the numerical model (a) and our equations (b). The horizontal axis is the total net Atlantic salt flux $F_s$. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.

Average Atlantic salinity $\bar{S}$ against the salt flux $H$ added to the North Atlantic for the ON state (black), the stable branch of the OFF state (blue) and the unstable OFF state branch (red, dashed) where $F_{circ} + F_s = 0$, but small perturbations $\delta \bar{S}$ grow exponentially (see previous figure). Arrows indicate the expected evolution of the system for initial conditions inside the domain of the arrow. Note that the unstable branch is computed from the OFF state. A saddle node occurs where the stable and unstable branches meet and terminate.
Schematic decomposition of the Atlantic basin into $n$ rectangular volumes $V_1, V_2, ..., V_n$ with salinities $S_1, S_2, S_3, ...$; the gyre-related salt transports $F_{r,k}$ between them and the surface fluxes $F_{s,k}$ into each volume $V_k$ for $k = 1, 2, ...$. The AMOC is absent. The salt flux from volume $V_{k+1}$ to $V_k$ is $F_{r,k} = r_k (S_{k+1} - S_k)$ for coefficients $r_k$ associated with the gyre strength.
Table 1: Explanation of mathematical terms. Positive sign of flux terms indicate salt added to ocean.

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{S}$</td>
<td>Basin-wide vol-average upper Atlantic salinity.</td>
</tr>
<tr>
<td>$M$</td>
<td>AAIW reverse cell strength (OFF state only).</td>
</tr>
<tr>
<td>$M_0$</td>
<td>small $M$ where $F_m$ becomes zero (Eq. 1).</td>
</tr>
<tr>
<td>$H$</td>
<td>Added salt flux to NADW region.</td>
</tr>
<tr>
<td>$H^*$</td>
<td>Maximal $H$ where stable OFF state occurs. Similar for $F_s^*$.</td>
</tr>
<tr>
<td>$\bar{S}^*$</td>
<td>$\bar{S}^* = -\frac{b}{2a}$. Maximal salinity where stable OFF state occurs. 34.63 kg/m$^3$ here.</td>
</tr>
<tr>
<td>$F_s$</td>
<td>Total net Atlantic surface salt flux.</td>
</tr>
<tr>
<td>$F_m$</td>
<td>AMOC-related salt flux (ON and OFF) $\int V &lt; S &gt;_x dz$ at 30°s.</td>
</tr>
<tr>
<td>$F_r$</td>
<td>Residual term $-F_s - F_m$, composed largely of Atlantic gyre salt transport across 30°s.</td>
</tr>
<tr>
<td>$F_{circ}$</td>
<td>Total Atlantic oceanic salt exchange $F_r + F_m$. Balances as $F_{circ} + F_s = 0$.</td>
</tr>
<tr>
<td>$a, b, c$</td>
<td>Coefficients of $F_{circ} = a\bar{S}^2 + b\bar{S} + c$ in OFF (Eq. 4).</td>
</tr>
<tr>
<td>$r_1, r_2$</td>
<td>Coefficients of $F_r = r_1\bar{S} + r_2 = r_1(\bar{S} - \bar{S}_r)$ (Eq. 3). $r_1$ encapsulates gyre strength</td>
</tr>
<tr>
<td>$\bar{S}_r$</td>
<td>idealized Atlantic salinity expected in the absence of $F_s$ and $F_m$ in OFF.</td>
</tr>
<tr>
<td>$\beta$</td>
<td>climate amplification of $H$ such that $\Delta F_s = (1 + \beta)H$.</td>
</tr>
<tr>
<td>$\bar{S}_{unst}$</td>
<td>$\bar{S}_{unst} = \frac{-b + \sqrt{b^2 - 4ac + F_s}}{2a}$ Salinity at unstable flux equilibrium.</td>
</tr>
</tbody>
</table>
Figure 1: Schematic depiction of ocean areas referred to in the text. The AAIW and NADW formation regions are marked by black lines. A black line also marks our working definition of the Southern Ocean.
Figure 2: Atlantic meridional overturning circulation of the NADW OFF state for (a) the unperturbed OFF experiment OFF (b) the critical OFF state for high anomalous salt flux $H$. No other additional surface salt fluxes are applied. Values are given in Sv ($1\text{ Sv} = 10^6 \text{ m}^3 \text{ sec}^{-1}$).
Figure 3: Results of experiments where the model is run to equilibrium for a range of constant anomalous salt fluxes $H$ applied to the North Atlantic Deep Water formation region (see fig. 1). (a) Average Atlantic salinity $\bar{S}$ for ON (black) and OFF (red). Also, the OFF state salinity of the Southern Ocean defined here as the entire ocean south of 32 °S (green). (b) the salt flux $F_m$ associated with the OFF state (red) and the ON state (black). Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where $1 \text{ Sv} = 10^6 \text{ m}^3 \text{ sec}^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.
Figure 4: AMOC-salt flux $F_m$ for three different versions of the ON (black) and OFF (red) states as in Fig. 3b. Two additional versions of the original experiment shown in Fig. 3b have been conducted, where, in addition to the original $H$, we now apply a constant anomalous surface salt flux to the entire Atlantic basin of (a) 0.2 Sv (strong OFF state average salt transport), (b) 0 Sv (i.e. the original experiment, standard OFF state average salt transport) and (c) -0.2 Sv (weak OFF state average salt transport). The only varying parameter within each panel is the anomolous salt flux $H$ applied to the North Atlantic. The rectangle indicates the point where the deVries Weber metric predicts the upper limit of the bistable regime to occur. The horizontal axis is translated so as to let $H = 0$ at the bifurcation point. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.
Figure 5: Results of experiments where the model is run to equilibrium for a range of constant anomalous salt fluxes $H$ applied to the North Atlantic Deep Water formation region (see fig. 1). (a) The total net Atlantic surface flux change $\Delta F_s$ with respect to the unperturbed OFF equilibrium ($H = 0$) as a function of the salt flux perturbation $H$ (red). $H$ is shown for comparison (black). (b) The average Atlantic temperature as a function of $H$. The AAIW reverse cell cools the Atlantic. (c) The average Atlantic evaporation as a function of $H$. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.
Figure 6: (a) AAIW reverse cell strength $M$ against the density difference $\Delta \rho_{NADW-AAIW}$ between the NADW and AAIW formation regions in the OFF state. A best linear fit is included (dashed).
Figure 7: Oceanic flux terms in the OFF state. (a) the resultant term $F_r = -F_m - F_s$ against the salinity difference between the South Atlantic (between 30 $^\circ$S and the equator) and the Southern Ocean. (b) AMOC salt flux $F_m$ (black) and total salt flux $F_{circ} = F_m + F_r$ arising from circulation (red) against the average Atlantic salinity $\bar{S}$. 
Figure 8: Values against the Atlantic salinity $\bar{S}$ of (a) mean salinity of NADW formation region, (b) $\frac{E_{in}}{M-M_0} = (v_1S + v_2)$, (c) the AAIW reverse cell strength $M$ and (d) the gyre salt flux $F_r$. Best linear fits (dashed) are added for the 5 model equilibria with highest $S$ (and closest to the critical flux $F^*_s$).
Figure 9: Average Atlantic salinity $\bar{S}$ (a) and AAIW reverse cell strength $M$ (b) against total net Atlantic salt flux $F_s$ for the model (black) and our equations (red).
Figure 10: The flux balance of $F_m$ (black) $F_r$ (green) and $F_s$ (red) such that $F_m + F_r + F_s = 0$ for the numerical model (a) and our equations (b). The horizontal axis is the total net Atlantic salt flux $F_s$. Salt flux units are in equivalent Sverdrups of fresh water flux (So called “negative Sverdrups”, where 1 Sv = $10^6$ m$^3$ sec$^{-1}$). A positive salt flux represents the addition of salt to the Atlantic.
Figure 11: Average Atlantic salinity $\bar{S}$ against the salt flux $H$ added to the North Atlantic for the ON state (black), the stable branch of the OFF state (blue) and the unstable OFF state branch (red, dashed) where $F_{circ} + F_s = 0$, but small perturbations $\delta \bar{S}$ grow exponentially (see previous figure). Arrows indicate the expected evolution of the system for initial conditions inside the domain of the arrow. Note that the unstable branch is computed from the OFF state. A saddle node occurs where the stable and unstable branches meet and terminate.
Figure 12: Schematic decomposition of the Atlantic basin into \( n \) rectangular volumes \( V_1, V_2, \ldots, V_n \) with salinities \( S_1, S_2, S_3, \ldots \); the gyre-related salt transports \( F_{r,k} \) between them and the surface fluxes \( F_{s,k} \) into each volume \( V_k \) for \( k = 1, 2, \ldots \). The AMOC is absent. The salt flux from volume \( V_{k+1} \) to \( V_k \) is \( F_{r,k} = r_k(S_{k+1} - S_k) \) for coefficients \( r_k \) associated with the gyre strength.
Figure 13: $F_{\text{circ}}$ vs. $\bar{S}$. 

\[
F_{\text{circ}} = F_m + F_r
\]