Tellus (2005), 57A, 676–690
Printed in Singapore. All rights reserved

TELLUS

Can limited ocean mixing buffer rapid climate change?

By KEVIN I. C. OLIVER[1]*, ANDREW J. WATSON[1] and DAVID P. STEVENS[2], [1]*School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom;* [2]*School of Mathematics, University of East Anglia, Norwich, United Kingdom*

(Manuscript received 7 May 2004; in final form 18 November 2004)

ABSTRACT

It has been argued that diapycnal mixing has a strongly stabilizing role in the global thermohaline circulation (THC). Negative feedback between THC transport and low-latitude buoyancy distribution is present in theory based on thermocline scaling, but is absent from Stommel's classical model. Here, it is demonstrated that these two models can be viewed as opposite limits of a single theory. Stommel's model represents unlimited diapycnal mixing, whereas the thermocline scaling represents weak mixing. The latter limit is more applicable to the modern ocean, and previous studies suggest that it is associated with a more stable THC. A new box model, which can operate near either limit, is developed to enable explicit analysis of the transient behaviour. The model is perturbed from equilibrium with an increase in surface freshwater forcing, and initially behaves as if the only feedbacks are those present in Stommel's model. The response is buffered by any upper ocean horizontal mixing, then by propagation of salinity anomalies, each of which are stabilizing mechanisms. However, negative feedback associated with limited diapycnal mixing only prevents thermohaline catastrophe in a modest parameter domain. This is because the time-scale associated with vertical advective-diffusive balance is much longer than the time required for the THC to change mode. The model is then tuned to allow equilibrium THC transport to be independent of the rate of mixing. The equilibrium surface salinity difference controls the classical THC-transport/salinity positive feedback, whereas the equilibrium interior density difference controls the mean-flow negative feedback. When mixing is strong, unrealistic vertical homogenization occurs, causing a convergence in surface and interior meridional gradients. This reduces positive feedback, and increases stability, in the tuned model. Therefore, Stommel's model appears to overestimate, rather than underestimate, THC stability to high-frequency changes in forcing.

1. Introduction

Stommel (1961) used a simple box model (hereafter Stommel's model) to argue that a bifurcation structure is intrinsic to thermohaline circulation (THC) in a body of water if buoyancy forcing due to surface heat and freshwater exchange are of the opposite sign and have different time-scales. The significance of this was appreciated much later, when ice- and marine-core records (e.g. Dansgaard et al., 1993) and GCMs (e.g. Bryan, 1986) provided evidence for abrupt and dramatic transitions in the meridional overturning circulation (MOC), which comprises the global THC as well as overturning that is directly driven by wind. These transitions are typically associated with an increase in freshwater flux to high latitudes, either from ice-melt or from atmospheric forcing. The decrease in salinity at high latitudes causes a decrease in meridional pressure gradient and therefore THC transport, thus reducing the transport of saline water from low to high latitudes. For a large change in forcing, this positive feedback mechanism, first identified in Stommel's model, may yield a non-linear response resulting in 'thermohaline catastrophe', with a reversed haline-driven overturning. This reversed circulation can remain stable under the initial conditions because of the different ways in which oceanic surface temperature and salinity are forced. The heat flux is closely related to the temperature difference at the air–sea interface, whereas the freshwater flux does not directly depend on ocean salinity. Therefore, the dependence of the equilibrium meridional salinity difference on the rate of overturning is greater than that of the equilibrium meridional temperature difference, and weak overturning favours a haline-driven circulation.

Despite the insights it provides, the simplifications in Stommel's model remove processes that are fundamental to the THC. One of the most important limitations is the absence of an appropriate parametrization for the effect of diapycnal mixing. Theory and GCM simulations (e.g. Bryan, 1987; Gnandesikan, 1999; Nilsson et al., 2003) suggest a weak non-linear equilibrium dependence of THC transport on meridional density gradient rather than the linear relationship in Stommel's model. The effect of such a parametrization at equilibrium has been

*Corresponding author.
e-mail: K.Oliver@uea.ac.uk

investigated with conceptual models (Park, 1999, hereafter P99; Nilsson and Walin, 2001, hereafter NW01) and stability of the thermally driven THC is increased in these studies compared with Stommel's model. This may suggest that the probability of thermohaline catastrophe is less than previously believed. However, there has been an absence of studies into the transient state with models incorporating diapycnal mixing. This is an important omission, considering that the primary motivation for studying the THC is anticipating its response to anthropogenic climate change. Furthermore, existing models make assumptions that are unlikely to be valid in the transient state, such as vertical advective-diffusive balance (P99) or that the high-latitude ocean and the deep low-latitude ocean can together be considered a single reservoir (NW01).

Here, a three-box single-hemisphere model is developed with the goal of starting to fill this gap in understanding. There is strong evidence that circulation driven by interhemispheric gradients is important (e.g. Rooth, 1982; Rahmstorf, 1996; Marotzke and Klinger, 2000), and that symmetric cells in each hemisphere are unstable to asymmetric perturbations (Bryan, 1986; Vellinga, 1996; Weijer and Dijkstra, 2001; Nilsson et al., 2004). Conceptual models that have helped to explain this exhibit dynamics that are explicitly excluded when considering one hemisphere in isolation. Nevertheless, simple one-hemisphere models have proven valuable in understanding aspects of the THC, and we proceed with such a model for the sake of simplicity. Further details of the choice of model are explained in Section 2. Thermally driven equilibria are analysed in Section 3, and the model's response to changes in freshwater forcing is examined in Section 4. In Section 5, the results are placed in the context of previous work, including GCM studies, and their significance is discussed.

2. Choosing a model

2.1. *Geostrophy, thermohaline circulation transport and diapycnal mixing*

Theoretical and GCM-based evidence (e.g. Wright et al., 1995; Marotzke, 1997; Park and Bryan, 2000) suggests a linear equilibrium relationship between meridional and zonal density differences. Combined with geostrophy, this provides a linear relationship between twice-depth-integrated meridional density difference and THC transport. There is further evidence from GCM studies (Hughes and Weaver, 1994; Rahmstorf, 1996; Thorpe et al., 2001) linearly relating the rate of overturning to meridional density differences between the high-latitude North Atlantic and the subtropical South Atlantic, rather than the equator (indicating support for an interhemispheric flow). Of these studies, only Hughes and Weaver (1994) used the twice-depth-integrated density difference; Rahmstorf (1996) used the deep density difference, and Thorpe et al. (2001) used the once-depth-integrated density gradient. How quickly the ocean adjusts to an equilibrium relationship between meridional and zonal gradients, in response to perturbations, depends on the time required to propagate anomalies. Various studies suggest as little as several months within a hemisphere (e.g. Kawase, 1987; Johnson and Marshall, 2002) or as much as several decades (McDermott, 1996; Marotzke and Klinger, 2000). However, the relationship between meridional gradients and overturning transport appears to remain valid in the transient state in at least one GCM (Thorpe et al., 2001).

A linear relationship between zonal and meridional density differences would imply that the maximum in the thermohaline overturning streamfunction, ψ, in a rectilinear basin, can be written

$$\psi = L \int_{z_0}^{0} V \, \mathrm{d}z, \tag{1}$$

where

$$V(z) = \frac{c_\rho g}{\rho_0 f L} \left(\int_{-h}^{z} \rho_{\mathrm{mer}} \, \mathrm{d}Z - \frac{1}{h} \int_{-h}^{0} \int_{-h}^{0} \rho_{\mathrm{mer}} \, \mathrm{d}Z \, \mathrm{d}Z \right).$$

Here, z_0 is the level of no motion [$V(z_0) = 0$], g is acceleration due to gravity, ρ_0 and f are representative values for density and the Coriolis parameter, respectively, h is the depth of the ocean, c_ρ is a dimensionless coefficient relating zonal and meridional density difference and incorporating basin geometry, L is the horizontal scale of the basin, and ρ_{mer} is the meridional density difference as a function of depth. The second term arises because it is necessary to obtain zero net top-to-bottom volume transport in order to conserve volume in an enclosed basin, if it is assumed that basin-integrated wind-driven meridional flow is zero. If $\rho_{\mathrm{mer}} \geq 0$ in the domain $-h \leq z \leq 0$, and $\rho_{\mathrm{mer}} > 0$ in at least part of this domain, exactly one level of no motion is obtained.

There are several approximations available to simplify this to a form suitable for box models. Consider a two-layer ocean with $\rho_{\mathrm{mer}} = \Delta\rho$ above the pycnocline depth, H, and $\rho_{\mathrm{mer}} = 0$ below this level. Equation (1) becomes

$$\psi = C_H \Delta\rho H^2, \qquad C_H = \frac{c_\rho g}{\rho_0 f} \left(1 - \frac{H}{2h}\right)^2. \tag{2}$$

Studies in which Stommel's model is applied to the global ocean usually extend the surface density difference throughout the water column. This can be represented here by stating $H = h$ and $\Delta\rho = \rho_{\mathrm{s}} =$ the surface density difference. However, the THC coefficient must then be scaled to yield reasonable ψ for reasonable ρ_{s}

$$\psi = C_h \rho_{\mathrm{s}}, \qquad C_h = \frac{H^*}{h} \frac{c_\rho g h^2}{4\rho_0 f} = \text{constant}, \tag{3}$$

where H^*/h is an estimate of the proportion of the water column over which the meridional density difference exists in the ocean. Because of the lack of vertical partitioning in Stommel's model, the low-latitude ocean is kept vertically homogeneous despite the addition of buoyancy from the above and density at depth. For this to be applicable in the ocean, an energy source would

be required to flux buoyancy to the deep low-latitude ocean to balance the upward buoyancy flux associated with overturning. A process by which this can occur is diapycnal mixing. [Wind-driven upwelling of dense water in the Southern Ocean (e.g. Toggweiler and Samuels, 1998) and geothermal heating of the deep ocean (Adcroft et al., 2001) are also likely to contribute, but these are neglected here.] It is implicit in Stommel's model that this energy source is never less than the energy required to keep the low-latitude ocean homogeneous, which is not the case in the modern ocean.

In the classical 'thermocline scaling' (Bryan and Cox, 1967; Welander, 1971; Bryan, 1987), the more reasonable approximation of negligible velocity below the thermocline is made. This is equivalent to stating that mixing is sufficiently weak that there is a shallow pycnocline (i.e. $H \ll h$). Therefore, C_H is insensitive to H:

$$\psi = C_H \Delta\rho H^2, \qquad C_H = \frac{c_\rho g}{\rho_0 f} = \text{constant}. \tag{4}$$

The vertical advective-diffusive balance, assuming spatially uniform diffusivity (Munk, 1966), is then

$$\bar{\psi}/A = \bar{w} = \bar{\kappa}/\bar{H}, \tag{5}$$

where A is the basin's surface area, w is upwelling velocity, κ is diapycnal diffusivity, and overbars indicate equilibrium values for time-dependent parameters (κ is a time-dependent parameter if it depends on density structure, which is not the case in this study). Simultaneous solution of eq. (4) with eq. (5) yields

$$\bar{\psi} = C_H^{1/3} A^{2/3} \bar{\kappa}^{2/3} \overline{\Delta\rho}^{1/3}. \tag{6}$$

P99 replaced eq. (3) in Stommel's model with eq. (6) and applied the latter equation when considering small perturbations from equilibrium, using constant κ. Stable thermally driven solutions existed within a greater parameter domain, leading to the argument that greater salinity forcing is required to initiate thermohaline catastrophe than in Stommel's model. NW01 used a two-layer model with isopycnal coordinates (dense low-latitude water in the same box as high-latitude water) and introduced a fixed energy source ($\kappa \sim \Delta\rho^{-1}$), as well as alternative parametrizations, to the advective-diffusive balance. If energy consumed in mixing was kept constant, rather than the rate of mixing itself, the THC was strengthened by a weaker surface meridional density gradient in their model as well as in a GCM (Nilsson et al., 2003). This led to stable thermally driven equilibria for any parameter values.

The P99 and NW01 models provide insights into the equilibrium THC, but are not well suited to investigation of the transient state without modification. The P99 model assumes vertical advective-diffusive balance; because the time-scale of this process in the low-latitude abyssal ocean is of the order of a millennium (Munk, 1966), such an assumption is only valid at equilibrium. The structure of the NW01 model means that any changes in high-latitude density forcing are diluted through the deep low-latitude ocean (i.e. most of the global ocean), which is again valid only at equilibrium. One way of resolving these problems would be to subdivide the isopycnal dense layer in the NW01 model into a high-latitude region and low-latitude deep region. This requires the use of isopycnal coordinates, with the lowest possible resolution, where individual layer densities are varying. A parametrization is then required to conserve mass, and it has been shown that the stability of the system depends to first order on the choice of parametrization (Oliver, 2003). For this reason, a depth-coordinate model, containing an alternative simplification of eq. (1), is developed here.

2.2. A three-box model

Figure 1 is a schematic diagram of the model used in this study. Three boxes represent (1) the low-latitude mixed layer, (2) the high-latitude ocean and (3) low-latitude pycnocline water and the deep ocean. Due to thermohaline overturning, ψ, water departs the low-latitude ocean from box 1, passes through box 2, returns in box 3, and upwells to complete the loop. The structure is thus far similar to one-half of the two-hemisphere Joyce (1991) model, or the oceanic part of the Nakamura et al. (1994) and Rivin and Tziperman (1997) models. However, we wish to introduce diapycnal mixing. This necessitates the exclusion of pycnocline water from box 1, so that mixing controls the flux of buoyancy from the mixed layer into the pycnocline, which is the most dynamically active region in a one-hemisphere THC. Therefore, even before diapycnal mixing is introduced, our model is not directly comparable with those employed by Joyce (1991), Nakamura et al. (1994) and Rivin and Tziperman (1997). Diapycnal mixing is represented by a vertical exchange flux, $K = \kappa A_1/h$, where A_1 is the surface area of the low-latitude boxes.

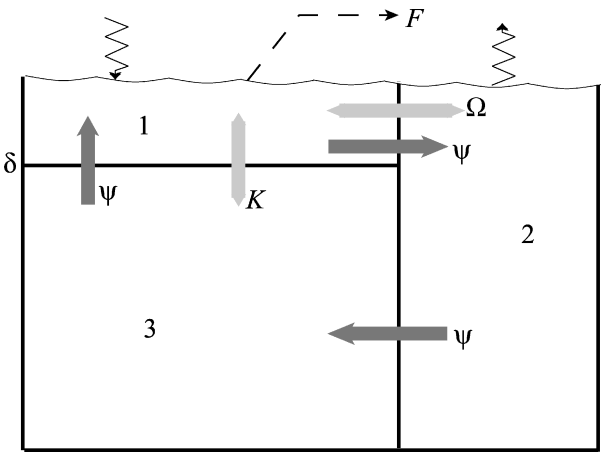

Fig 1. Schematic diagram of the three-box model. THC transport is denoted by ψ, vertical and horizontal exchange by K and Ω, respectively, atmospheric freshwater transport by F, and zigzag lines indicate air–sea heat exchange.

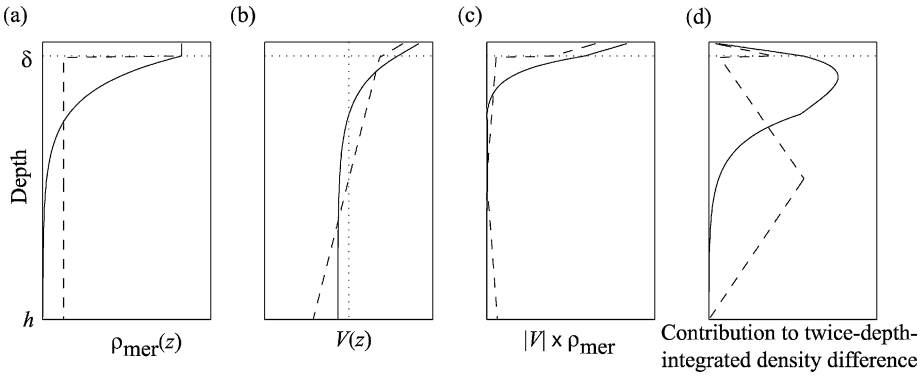

Fig 2. Depth profiles for an idealized ocean (solid) and its representation in the model (dashed) of (a) meridional density difference, ρ_{mer}, (b) meridional velocity, V, (c) a measure of meridional density transport, $|V| \times \rho_{\mathrm{mer}}$, and (d) contribution of vertically local density difference to ψ (i.e. the difference between ψ obtained from the density profile in a and the value of ψ which would be obtained if vertically local ρ_{mer} were zero). The quantities in (b), (c) and (d) are derived from eq. (1).

The ocean depth h is chosen as the scale depth because the model implicitly assumes that water crossing the interface is instantaneously well mixed throughout the ocean. (It is assumed that $\delta \ll h$, where δ is the depth of the surface layer.) A similar approach to diapycnal mixing was employed by Gargett and Ferron (1996), although they also considered double-diffusion, and by Shaffer and Olsen (2001). Additionally, there is a prescribed surface horizontal volume exchange, Ω, representing wind-driven circulation. There is evidence from another box model study (Shaffer and Olsen, 2001) that such horizontal mixing stabilizes the thermally driven THC. Note that the structure of the model presumes a thermally driven THC. Therefore, we investigate whether thermohaline catastrophe occurs, but not the evolution of the resulting haline-driven circulation or the existence and stability of haline-driven solutions.

It has been noted before that any surface density gradient must be communicated to greater depths if a substantial overturning is to be maintained (Munk and Wunsch, 1998). Equation (1) shows that ψ is insensitive to ρ_{mer} near the surface. In the thermocline scaling also, the quadratic dependence of ψ on H means that deeper pycnocline water is more dynamically active. Therefore, a simplifying approximation can be applied that ψ depends only on the density difference between boxes 2 and 3, ρ_{d}:

$$\psi = C_h \rho_{\mathrm{d}}, \qquad C_h = \frac{c_\rho g h^2}{4\rho_0 f} = \text{constant}. \tag{7}$$

[Note that eq. (7) differs from eq. (3) only in that the mean subsurface density difference is applied, and therefore the factor of H^*/h is not required.] Consistent with the thermocline scaling, the dynamically active low-latitude region contains the pycnocline.

Employing a dynamically inactive surface box may appear inconsistent with the assumption that the source of the poleward limb of the overturning cell is box 1 rather than box 3. However, the depth range over which the majority of poleward volume flux occurs is different from the depth range of the meridional density gradient that drives the THC. Figure 2 shows that, on applying eq. (1) exactly to both an idealized meridional density difference profile and to the representation of that profile in the model, much of the poleward volume and buoyancy transport occurs near the surface, even though the THC primarily depends on density gradients in the ocean interior. Nevertheless, zero thermohaline volume transport from box 3 to box 2 is not entirely justified, but it is assumed because it facilitates analytical solution and interpretation of the model.

The terms T_n, S_n and ρ_n indicate temperature, salinity and density, respectively, in box n. Subscripts are introduced such that, for example, $T_{\mathrm{s}} = T_1 - T_2$, $T_{\mathrm{d}} = T_3 - T_2$ and $T_{\mathrm{v}} = T_1 - T_3$. (For ρ_{s}, ρ_{d} and ρ_{v}, the sign is reversed, so that each of these terms would be positive for the modern ocean.) Instantaneous temperature restoring is assumed for the boxes in contact with the atmosphere, so that these boxes have fixed temperature: $\dot{T}_1 = \dot{T}_2 = 0$. The virtual salt flux approximation is applied for the fixed freshwater flux, F; there is an equatorward atmospheric salt transport rather than a poleward atmospheric freshwater transport. The equations for the evolution of temperature and salinity are

$$V_3 \dot{T}_3 = K T_{\mathrm{v}} - \psi T_{\mathrm{d}}, \tag{8}$$

$$V_1 \dot{S}_1 = S_0 F - (\psi + K) S_{\mathrm{v}} - \Omega S_{\mathrm{s}}, \tag{9}$$

$$V_2 \dot{S}_2 = (\psi + \Omega) S_{\mathrm{s}} - S_0 F, \tag{10}$$

$$V_3 \dot{S}_3 = K S_{\mathrm{v}} - \psi S_{\mathrm{d}}, \tag{11}$$

where, for example, $\dot{T}_n$ denotes the time derivative of T_n, and S_0 is a representative ocean salinity. A linear equation of state is used to determine density

$$\rho - \rho_0 = -a(T - T_0) + b(S - S_0), \tag{12}$$

where a is the thermal expansion coefficient and b is the haline contraction coefficient.

Table 1. Fixed parameters in the model

Parameter	Value
A_1	2×10^{14} m^2
A_2	1×10^{13} m^2
a	0.15 kg m^{-3} °C^{-1}
b	0.8 kg m^{-3}
C_h	30×10^6 m^6 kg^{-1} s^{-1}
δ	200 m
h	4000 m
S_0	35
T_1	20°C
T_2	0°C

Fixed parameters, used later for numerical solution of the model, are introduced in Table 1. A_2 and T_2 are chosen to represent the regions of downwelling in the northern North Atlantic. Choosing the area of the North Atlantic basin for A_1 would be inappropriate because it is unlikely that the greater proportion of upwelling and diapycnal mixing occurs in this basin. Instead, A_1 represents the majority of the global ocean, which is the area over which the canonical diffusivity of 1 cm^2 s^{-1} (Munk, 1966) would need to act to balance upward transport. This can be rationalized in terms of residence times. The implied mean upwelling velocity in the model is ψ/A_1; if A_1 is too small, then the upwelling rate is too large and the residence time of water in box 3 is too small. Therefore, unreasonably high diapycnal mixing would be needed to give a reasonable exchange flux between boxes 1 and 3 and therefore a reasonable equilibrium solution. The depths δ and h are 200 and 4000 m, respectively, so $V_1/V_2 = 1$. High-latitude downwelling regions (represented by box 2) contribute a small proportion of the global ocean area, whereas the low-latitude ocean in rapid communication with the atmosphere (box 1) occupies a narrow layer. Equal volumes provide analytical simplicity, in the absence of evidence that a different ratio of volumes would be preferable. An equal volume for box 3 would not be reasonable; this is prescribed to be 19 times greater than that of each of the other boxes. The value of the THC coefficient applicable in this model, C_h, is chosen to yield a THC of ~16 Sv when the deep temperature difference is 4°C and the deep salinity difference is 0.1. Although it is described as a fixed parameter, it is modified in Section 4.3.

3. Equilibrium solution

3.1. Thermohaline circulation dependence on surface meridional density difference

Because many studies determine THC transport as a function of surface meridional density difference, it is useful to begin with this derivation. First, from the definitions of ρ_s, ρ_v and ρ_d

$$\rho_s = \rho_v + \rho_d. \tag{13}$$

At equilibrium, the buoyancy that box 3 gains through diapycnal mixing is balanced by buoyancy lost through thermohaline overturning:

$$K\bar{\rho}_v = \bar{\psi}\bar{\rho}_d \tag{14}$$

(overbars indicate equilibrium values for time-dependent variables). Eliminating ρ_v and ρ_d from eqs. (7), (13) and (14), the quadratic equation

$$\bar{\psi}^2 + K\bar{\psi} - KC_h\bar{\rho}_s = 0, \tag{15}$$

with the positive root

$$\bar{\psi} = \frac{K}{2}\left[-1 + \left(1 + \frac{4C_h\bar{\rho}_s}{K}\right)^{1/2}\right], \tag{16}$$

is obtained.

Using eqs. (7), (13) and (14) again, the result is obtained that

$$\frac{C_h\bar{\rho}_s}{K} = \frac{\bar{\psi}}{K}\left(\frac{\bar{\psi}}{K} + 1\right) = \frac{\bar{\rho}_v}{\bar{\rho}_d}\left(\frac{\bar{\rho}_v}{\bar{\rho}_d} + 1\right). \tag{17}$$

In the modern ocean, the mean density of subsurface water is closer to that of high-latitude water than low-latitude surface water. If $2\bar{\rho}_v/\bar{\rho}_d \gg 1$, an approximation (similar to $H \ll h$ in the thermocline scaling) can be made:

$$\bar{\psi} \approx (KC_h\bar{\rho}_s)^{1/2}. \tag{18}$$

The relationship between $\bar{\rho}_s$ and $\bar{\psi}$ is less than linear because the residence time of water in the deep box is dependent on $\bar{\psi}$. More buoyancy can be stored in the deep ocean, due to diapycnal mixing, when $\bar{\psi}$ is smaller. This acts to increase the deep density gradient and therefore $\bar{\psi}$, a negative feedback mechanism. A similar redistribution of low-latitude buoyancy also provides the negative feedback in the classical thermocline scaling, although in isopycnal coordinates this is expressed as an increase in pycnocline depth rather than a decrease in subsurface density.

The power laws are 1/2 for surface density difference, and 1/2 for diapycnal mixing also, rather than 1/3 and 2/3, respectively, in the thermocline scaling eq. (6). (Note, however, that $\bar{\rho}_s$ is not entirely equivalent to $\Delta\rho$ in the thermocline scaling.) The power laws from our model would be obtained in the thermocline scaling if ψ in eq. (4) were proportional to H rather than H^2. This is equivalent to using the once-depth-integrated meridional density difference rather than the twice-depth-integrated difference. There is GCM evidence in support for such an approach (Thorpe et al., 2001), which has been employed previously in box models (Joyce, 1991; Lyle, 1997). Nevertheless, we attribute the discrepancy between our model and the classical scaling to a weakness in our model. In two depth-coordinate boxes, the depth of penetration of buoyancy is not resolved. Therefore, the dependence of ψ on the location, as opposed to the quantity, of subsurface low-latitude buoyancy is not diagnosed. As a result, the equilibrium THC transport in the model presented here might be considered slightly too dependent on $\bar{\rho}_s$ and not dependent enough on

K. However, this is only a quantitative discrepancy, and the nonlinear dependence on $\bar{\rho}_s$ results from a similar negative feedback mechanism that is present in the thermocline scaling.

In the unlikely limit of strong mixing, $4\bar{\rho}_v/\bar{\rho}_d \ll 1$, the approximation $(1+x)^n \approx 1 + nx$ can be applied and the solution is obtained:

$$\bar{\psi} \approx C_h \bar{\rho}_s. \tag{19}$$

This simply states that if low-latitude diapycnal mixing is sufficiently intense, it ceases to be rate-limiting in the THC, the low-latitude ocean is homogenized, and Stommel's model applies.

3.2. *Equilibria and stability dependent on surface buoyancy forcing*

The model is now solved with the prescribed (time-independent) input parameters. At equilibrium, we have from eq. (10) that $(\bar{\psi}+\Omega)\bar{S}_s = S_0 F$. Therefore, the poleward buoyancy transport, $(\bar{\psi}+\Omega)\bar{\rho}_s$, is given by

$$(\bar{\psi}+\Omega)\bar{\rho}_s = a(\bar{\psi}+\Omega)T_s - bS_0F. \tag{20}$$

Substituting for $\bar{\rho}_s$ in eq. (15) yields a cubic equation in $\bar{\psi}$:

$$\begin{aligned}\bar{\psi}^3 &+ (K+\Omega)\bar{\psi}^2 - K(C_h a T_s - \Omega)\bar{\psi} \\ &+ KC_h(bS_0F - aT_s\Omega) = 0.\end{aligned} \tag{21}$$

Note that T_s is a fixed parameter. With fixed parameters as defined in Table 1, $C_h a T_s = 90$ Sv; therefore, it is highly unlikely that the linear coefficient is non-negative. Because the cubic and linear coefficients have opposite signs, there is an extreme in the function given by the left-hand side of eq. (21) either side of $\bar{\psi} = 0$. There are therefore up to two real positive roots (any negative roots are meaningless).

Figure 3 is a plot of these roots in $\bar{\psi}$–F space for $\Omega = 0$ and for large Ω (5 Sv), for two values of K: 2 and 5 Sv. The stability of the thermally driven equilibria is not indicated because, unlike in Stommel's model, the greater positive equilibrium is not always stable (no cases of a stable smaller equilibrium have been found, however). Approximations for the limit of stable solutions, for large and small K, respectively, are derived in Appendix A. The maximum freshwater forcing sustainable in a stable thermally driven equilibrium THC, F_{max}, dependent on K and Ω, is plotted in Fig. 4. The existence of stable solutions is favoured both by large K and large Ω. The primary effect of increasing K is to increase $\bar{\psi}$. This increases negative feedback due to removal of anomalies by the mean flow. Increasing Ω directly increases this feedback, independent of the effect it has on $\bar{\psi}$. The stabilization by stronger diapycnal and horizontal mixing is consistent with that obtained by Shaffer and Olsen (2001). Like Shaffer and Olsen, we find that horizontal mixing alone can sustain a thermally driven overturning with vanishing diapycnal mixing (eq. A10). However, the resulting overturning transport is very small.

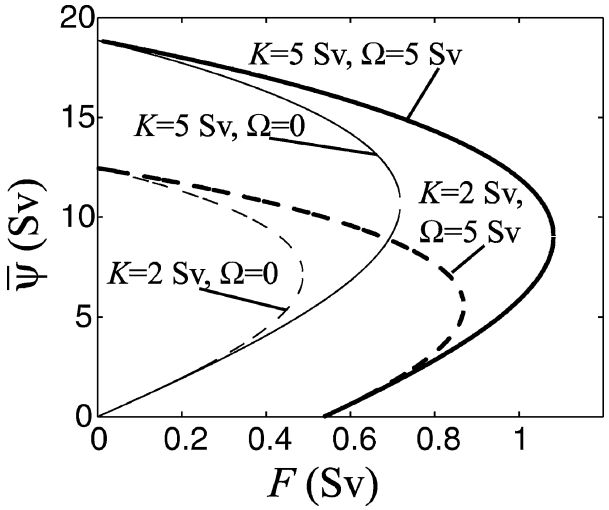

Fig 3. Dependence of $\bar{\psi}$ on F for $\Omega = 0$ (thin lines) and $\Omega = 5$ Sv (thick lines), and for $K = 5$ Sv (solid lines) and $K = 2$ Sv (dashed lines) in the model. Thermally driven equilibria are plotted, regardless of stability.

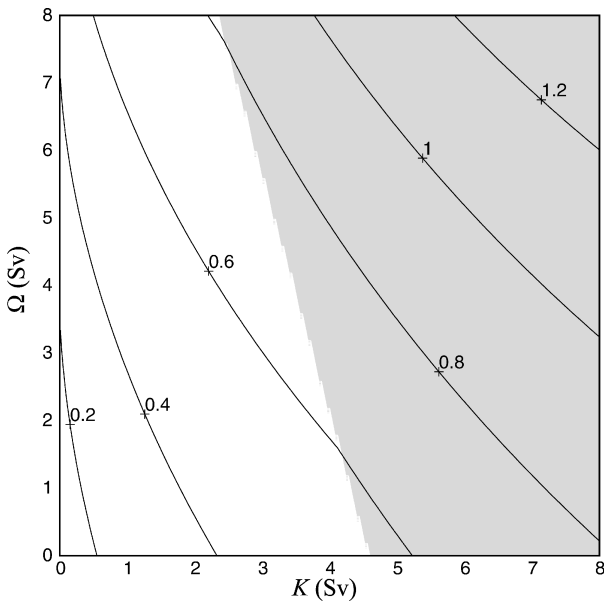

Fig 4. Contour plot of maximum freshwater forcing sustainable in a stable thermally driven THC, F_{max} (Sv), dependent on K and Ω, in the model. In the shaded region, stability is limited by eq. (A4): $\Lambda_1\Lambda_2 \geq 0$. In the unshaded region, stability is limited by eq. (A3): $\Lambda_1 + \Lambda_2 \leq 0$.

In Appendix A it is shown that, provided $V_3 \gg V_2$, neither instability nor oscillatory behaviour can be introduced by buoyancy storage in the deep ocean. However, in the absence of horizontal mixing, any stable positive equilibrium is a spiral point, indicating that internal oscillations are possible. Such oscillations do not exist in Stommel's model (Cessi, 1994; Ruddick and Zhang, 1996), due in part to its symmetry, which prohibits a phase lag in the response. By introducing a vertical dimension, this

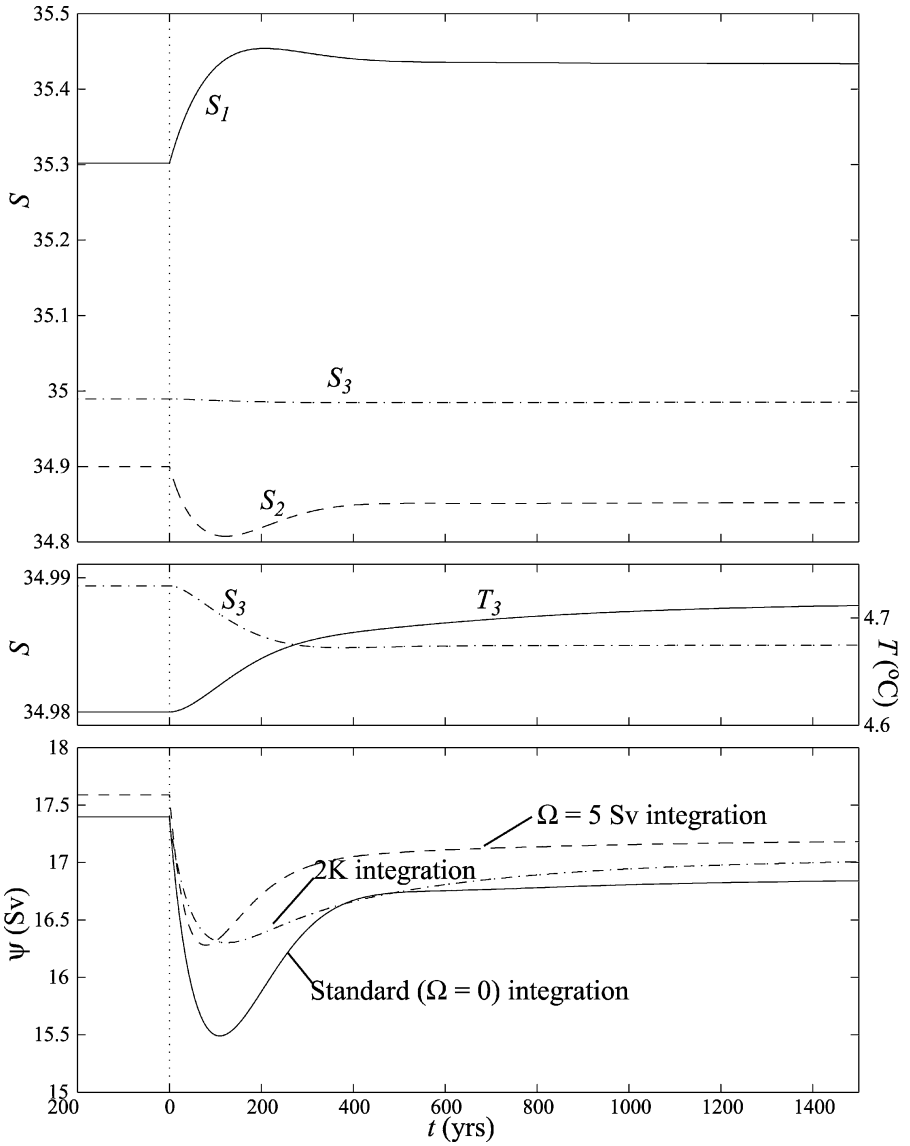

Fig 5. Evolution of the model with $K = 5$ Sv, $\Omega = 0$. F is increased from 0.2 to 0.28 Sv at $t = 0$. S_1 (solid), S_2 (dashed) and S_3 (dot-dashed) are plotted in (a). A different scale is used to plot S_3 (dot-dashed) in (b); T_3 (solid) is also plotted. ψ (solid) is plotted in (c). The dashed line in (c) shows the evolution of ψ in an identical run except that $\Omega = 5$ Sv. The dot-dashed line in (c) shows the evolution of ψ for the '2K' experiment described in Section 4.3 [$K = 10$ Sv; C_h ($\approx$18 m^6 kg^{-1} s^{-1}) tuned to yield unchanged $\bar{\psi}$; $\Omega = 0$; other terms as described above].

symmetry is removed here; further consequences of this are explored in Section 4.2. Horizontal mixing, but not diapycnal mixing, tends to suppress oscillations, whereas an increased meridional salinity difference tends to enhance them.

4. Transient behaviour

4.1. Response to a change in freshwater forcing

We wish to understand how the inclusion of limited diapycnal mixing affects the modelled ocean's response to changes in atmospheric forcing, such as those that may be occurring due to anthropogenic climate change. As a simple starting point, we present an example model run where the model is perturbed from equilibrium by an increase in freshwater forcing. In addition to the fixed parameters in Table 1, we use $K = 5$ Sv, $\Omega = 0$. Initial F is 0.2 Sv, but this is increased by $\Delta F = 0.08$ Sv to 0.28 Sv at $t = 0$ yr.

The evolution of the model parameters is plotted in Fig. 5. The immediate response is for the salinity of the two boxes in contact with the surface to rapidly diverge, resulting in a decrease in ψ, because the properties of the deep box hardly change on

this short time-scale. Positive feedback between high-latitude salinity and THC transport (similar to that in Stommel's model) enhances this change. At the same time, the properties of the deep ocean are slowly evolving. The inflow of cold water from box 2 becomes weaker, but the input of warm water from box 1 has not changed, resulting in heat storage in box 3. The amount of salt storage in the deep ocean is also changing. The salinity of the inflow to box 3 from box 2 is decreasing. Although box 3 also receives water from box 1, which is increasing in salinity, $\psi > K$, so the effect of the box 2 source is more important. The deep ocean is a large reservoir, so the effect on S_3 is small. However, because freshwater (more accurately virtual salt) is conserved, a decrease in salinity of the deep ocean is associated with an increase in salinity of the surface ocean. Perturbations to box 1 are communicated to box 2 relatively rapidly; therefore, the effect of this is to increase ρ_d, and therefore ψ. Buoyancy storage, in the form of heat, in the deep ocean also tends to increase ψ. Thus, a 'recovery phase' commences in which the THC tends towards a new equilibrium. The recovery phase is accelerated by the positive feedback between S_2 and ψ, similarly to the initial phase.

This behaviour is representative of runs in which thermohaline catastrophe does not occur (i.e. runs in which ρ_d does not change sign). Figure 6a shows the maximum amplitude response of the THC for a range of values of initial F and ΔF. Where changes to the THC remain subcritical, the maximum amplitude response is greater than the equilibrium response by a factor of 3–4 typically for positive ΔF (slightly less for negative ΔF). It is not surprising, therefore, that thermohaline catastrophe can occur even when the final freshwater forcing is less than F_{max}. Thermohaline catastrophe is favoured by large ΔF, so that the critical final freshwater forcing is much weaker if the initial forcing is also weak. If such a result is transferable to the ocean, it suggests that thermohaline catastrophe could occur whether or not the present state is near the stability limit, even without considering the high-frequency variability in forcing that is absent from this study.

The experiment in Fig. 5 was repeated but with non-zero horizontal mixing ($\Omega = 5$ Sv). The behaviour is similar, but the response of ψ, plotted as a dashed line in Fig. 5, is dampened. This is because the salinity anomalies in boxes 1 and 2, of opposite signs, are communicated to one another by horizontal mixing (not shown), acting as a negative feedback. Figure 6b shows that non-zero Ω also diminishes the possibility of thermohaline catastrophe.

4.2. *What initiates the recovery phase?*

The THC in the model decreases rapidly in response to an incremental increase in freshwater forcing, even for a subcritical change. It has been argued (Johnson and Marshall, 2002) that the effect of buoyancy transport to the deep ocean by diapycnal mixing can be ignored on this time-scale, because the density of the deep ocean changes slowly. It is also worth noting that, because $(\bar{\psi} + K)\bar{T}_v = \bar{\psi}\bar{T}_s$ and $(\bar{\psi} + K)\bar{S}_v = \bar{\psi}\bar{S}_s$, the initial perturbations in oceanic heat and salt fluxes into boxes 1 and 2 are nearly symmetric (see Fig. 5). The only asymmetry in response to a small perturbation is caused by a perturbation in ψ while K remains constant. We therefore consider the consequences of introducing a simplification by which the properties of box 3 do not change and the evolution of S_1 is equal and opposite to the evolution of S_2. This simplified model is not intended to be physically plausible, but to provide a basis from which the behaviour of the unsimplified model can be understood.

In the simplified model we have $\dot{S}_s = -2\dot{S}_2$:

$$V_2\dot{S}_s = 2S_0F - 2(\psi + \Omega)S_s, \tag{22}$$

If the freshwater forcing is suddenly increased by ΔF, perturbing the model from equilibrium, then eq. (22) can be rewritten in terms of then mean, $\bar{S}_s$, and perturbation, S'_s in S_s

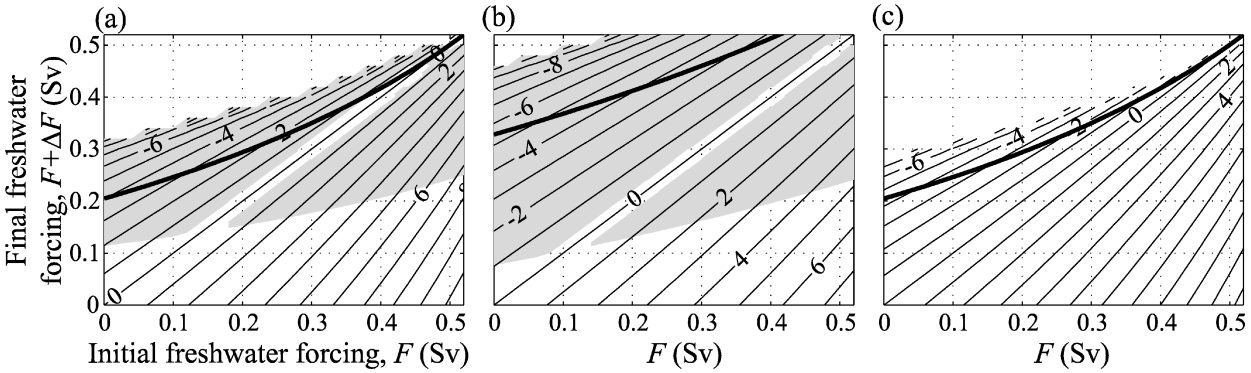

Fig 6. Maximum amplitude of change in THC transport (Sv) in response to a change in freshwater forcing from F to $F + \Delta F$, for: (a) $K = 5$ Sv, $\Omega = 0$; (b) $K = 5$ Sv, $\Omega = 5$ Sv; (c) same as (a) but with a very large deep box. Shading indicates that the maximum amplitude response is more than a factor of 3 greater than the equilibrium response; this is absent from (c) because a true equilibrium is not reached. The uncontoured region indicates that thermohaline catastrophe occurs. The thick lines indicate the threshold of ΔF in eq. (25), above which thermohaline catastrophe would be predicted by the simplified model.

$$V_2\dot{S}_s = 2S_0\Delta F - 2(\bar{\psi}+\Omega)S'_s - 2\bar{S}_s\psi' - 2\psi' S'_s$$
$$= 2S_0\Delta F - [2(\bar{\psi}+\Omega) - C_h b\bar{S}_s]S'_s + C_h b(S'_s)^2, \quad (23)$$

where $\psi' = -C_h S'_d = -(1/2)C_h S'_s$ has been used because $\dot{S}_3$ and $\dot{T}_3$ are zero in the simplified model. The non-linear term is retained because finite amplitude changes are under consideration. The linear terms contain negative feedback due to the mean flow, and the familiar meridional-salinity-difference/THC-transport positive feedback. Equation (23) is similar to Stommel's model at the limit of instantaneous temperature restoring and with a fixed freshwater flux (extended from Marotzke, 1990):

$$V_2\dot{S}_s = 2S_0\Delta F - 2\left(\bar{\psi} - C_h b\bar{S}_s\right)S'_s + 2C_h b(S'_s)^2, \quad (24)$$

where eqs. (3) and (7) indicate a greater value of C_h in our model than in Stommel's model. With $\Omega = 0$, eqs. (23) and (24) differ only in that the coefficients in the positive feedback and non-linear terms take different values. Therefore, the feedbacks that are present in the simplified model are those present in Stommel's model.

The simplified model, eq. (23), is numerically integrated (with $K = 5$ Sv, $\Omega = 0$, $F = 0.2$ Sv, $\Delta F = 0.08$ Sv) and compared with the unsimplified model in Fig. 7. As would be expected, the approximation exhibits no recovery phase and fails in long integrations. However, it performs well on decadal time-scales, and therefore offers insights into the short-term response in the unsimplified model containing diapycnal mixing.

The second-order term always acts to increase S_s and decrease ψ, causing increased damping if ΔF is negative (Fig. 6). It also provides the mechanism for thermohaline catastrophe with positive ΔF; once the second-order term becomes important, high-latitude density will decrease increasingly rapidly. This will not occur if equilibrium (in the simplified model) is reached while the second-order term is still small. Because, with positive ΔF, $\dot{S}_s$ is positive for both zero and infinite $\Delta S'_s$, such an equilibrium can only be reached if eq. (23) has two real negative roots when the left-hand side is zero. This is not the case, and the simplified model predicts thermohaline catastrophe, if

$$8C_h b S_0 \Delta F > \left[2(\bar{\psi}+\Omega) - C_h b\bar{S}_s\right]^2. \quad (25)$$

This predicted threshold is plotted in Fig. 6a. The minimum value of ΔF required for thermohaline catastrophe is significantly greater than that predicted by eq. (25). A reasonable hypothesis is that the model is stabilized by buoyancy storage in the deep ocean, due to diapycnal mixing. The hypothesis can be tested by removing this mechanism. This is achieved by repeating the standard integration with a very large deep box (V_3 increased by a factor of 10^9), so that heat and salt may still be stored in the deep ocean, but without changing the properties of box 3. Figure 7 shows that $\dot{S}_s$ changes sign in such a run at a very similar point as it does when box 3 has its standard volume. Figure 6c shows a repeat of the suite plotted in Fig. 6a, but with a very large deep box. The parameter domain in which thermohaline catastrophe occurs is slightly increased, but the simplified model still underestimates critical ΔF.

Therefore, much of the stabilization is provided by an alternative mechanism, associated with the asymmetric responses of boxes 1 and 2 (a symmetric response and no change to the properties of box 3 were the two simplifications made in eq. 23). In response to an increase in freshwater forcing, the boxes initially

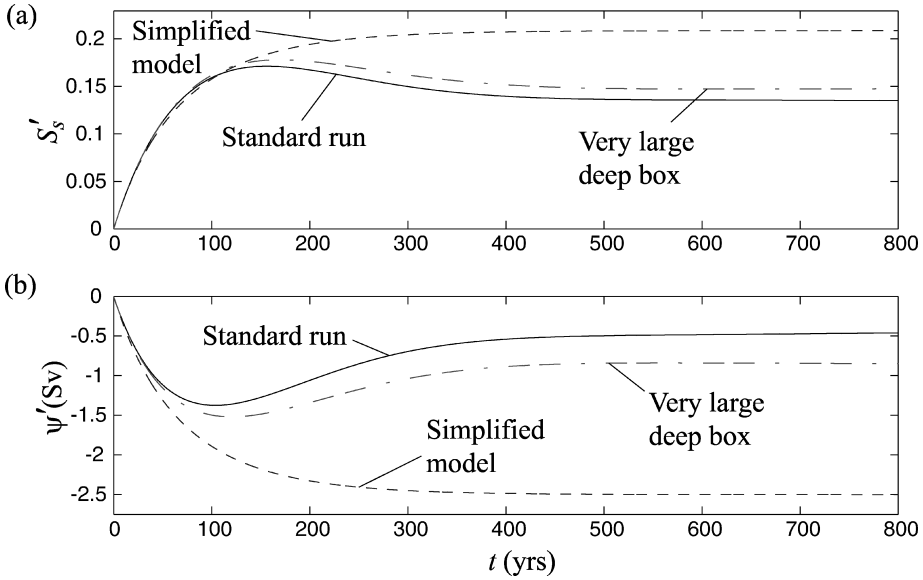

Fig 7. Evolution of (a) S'_s and (b) ψ' in simplifications of the model. Repeated for the standard run (solid), the simplified model eq. (23) (dashed) and the standard run with a very large deep box (dot-dashed).

respond nearly symmetrically. Because ψ decreases, but K remains constant, the fresh inflow becomes greater than the saline inflow to box 2, causing S_s to become slightly greater than that predicted by the simplified model after several decades (Fig. 7). This is a small effect. Of greater importance is that the source of water to box 2 (namely box 1) is increasing in salinity, whereas the salinity of the source of water to box 1 (box 3 only, because $\Omega = 0$ in this example) is barely changing. Therefore, when $\dot{S}_2$ changes sign, $\dot{S}_1$ initially remains positive. As a result, the mean salinity of the boxes in contact with the atmosphere increases in response to an increase in F. This response is physically reasonable because additional freshwater in the high-latitude ocean is communicated to the deep ocean more rapidly than the additional salt in the low-latitude ocean. The asymmetric response of boxes 1 and 2 tends to increase the density of box 2 relative to box 3 (for positive ΔF), and therefore tends to increase ψ. It is therefore a stabilizing process. This mechanism has similarities with the non-linear oscillatory mechanism found by Griffies and Tziperman (1995) and Rivin and Tziperman (1997) in box models with vertical partitioning, although in their studies the surface density gradient is the primary dynamic control and the time-scale of the oscillation is not influenced by advective-diffusive balance.

The time-scale over which the above process acts is governed by the time-scale over which the salinities of boxes 1 and 2 diverge. Integration of eq. (23) without the non-linear term, and for $F = 0$, so $\bar{S}_s = 0$, gives exponential decay with an e-folding time-scale of $V_2/2(\bar{\psi} + \Omega) \approx 40$ yr. A comparable time-scale for the evolution of deep ocean density is estimated at ~300 yr (see Fig. 5). This separation of time-scales explains the limited role of diapycnal mixing. For larger values of initial F, so that the negative and positive linear feedbacks nearly cancel, the e-folding time-scale associated with eq. (23) is significantly lengthened. This explains the greater importance of diapycnal mixing, and diminished role of the propagation of salinity anomalies, in this part of the parameter domain.

4.3. Dependence of feedback on diapycnal mixing

Quantification of the diapycnal mixing that ultimately maintains meridional overturning is typically attempted on the basis of an assumed rate of overturning (e.g. Munk and Wunsch, 1998). It follows that the quantity represented here by $\bar{\psi}$ (overturning streamfunction) is better known for the modern ocean than the quantity represented by K (exchange flux due to diapycnal mixing). If the model is tuned to obtain the same value of $\bar{\psi}$ with a different K under the same forcing, any modification to the behaviour is therefore of interest. This is only possible by compensating for changes to K by tuning C_h. Equation (21) can thus be rearranged to yield C_h for known K, $\bar{\psi}$, Ω, and surface buoyancy forcing:

$$C_h = \frac{\bar{\psi} + K}{K} \left[\frac{\bar{\psi}(\bar{\psi} + \Omega)}{aT_s(\bar{\psi} + \Omega) - bS_0F} \right]. \tag{26}$$

We begin by considering the effect of this tuning on the simplified model (23), which approximates the initial response of the model well near the weak mixing limit. The linear stability criterion of this model is that

$$2(\bar{\psi} + \Omega) \geq C_h b \bar{S}_s, \tag{27}$$

where the left-hand side is the negative feedback and the right-hand side is the positive feedback. If $\Omega = 0$, then eqs. (7) and (27) yield

$$2(a\bar{T}_d - b\bar{S}_d) \geq b\bar{S}_s. \tag{28}$$

The deep meridional density difference is compared with the haline contribution to the surface meridional density difference to establish stability. Because surface gradients are typically stronger than deep gradients, this is a stringent stability test. Interestingly, eq. (28) is equivalent to the criterion obtained by Nilsson et al. (2004) at the limit of weak mixing (their eq. 29), in an interhemispheric extension of NW01. They found that the 'thermocline-depth adjustment' stabilizing feedback was weak in response to asymmetric interhemispheric perturbations in their model; our results suggest that this is true also of intrahemispheric perturbations. Additionally, from eqs. (13) and (14)

$$\frac{T_s}{\bar{T}_d} = \frac{\bar{S}_s}{\bar{S}_d} = \frac{\bar{\psi}}{K} + 1. \tag{29}$$

Increasing diapycnal mixing (reducing $\bar{\psi}/K$) in the tuned model decreases the contrast between surface and deep gradients at equilibrium, by acting to vertically homogenize the ocean, and thus increases stability.

The behaviour can also be rationalized in terms of the positive feedback term. If $\bar{\psi}$ is fixed, the left-hand side (negative feedback term) of eq. (27) is also fixed; changes in $\bar{T}_d$ and $\bar{S}_d$ are compensated by changes in C_h. $\bar{S}_s$ (obtained by setting the left-hand side of eq. 10 equal to zero) is also independent of K. However, eq. (26) yields $C_h \sim 1/K$ if $K \ll \bar{\psi}$. If mixing is increased, the decrease in C_h decreases the right-hand side of eq. (27), stabilizing the model. An example of this is plotted in Fig. 5 for the experiment '2K': $K = 10$ Sv, C_h ($\approx 18 \times 10^6$ m^6 kg^{-1} s^{-1}) as derived from eq. (26) without changing $\bar{\psi}$, and all other parameters identical to the experiment described in Section 4.1 ($\Omega = 0$, $F = 0.2$ Sv, $\Delta F = 0.08$ Sv).

The above analysis neglects inaccuracies in the assumptions made in eq. (23). Any decrease in the time-scale associated with buoyancy storage in the deep ocean, caused by enhanced diapycnal mixing, will increase the stabilization at large K. The effect of more rapid removal of salinity anomalies from box 1 to box 3, resulting from an increase in K, has the opposite sign. However, the predicted qualitative effect of modifying K (and tuning C_h so $\bar{\psi}$ does not vary) on model stability is borne out in the weak mixing domain by the ensemble of model runs in Fig. 8a. The integration described above was repeated for a range of values of K between 0.1 and 10 Sv, and for a range of $F + \Delta F$ between 0 and 0.6 Sv (initial F is 0.2 Sv in each case). Because tuning

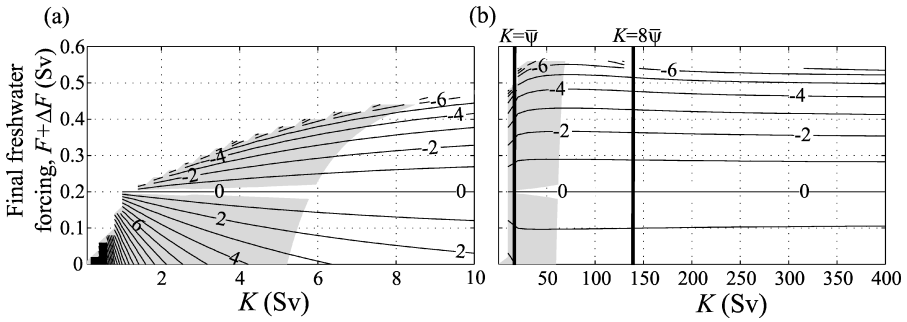

Fig 8. Maximum amplitude of change in THC transport (Sv) in response to a change in freshwater forcing from 0.2 Sv to $F + \Delta F$, in the ranges (a) $0.1 \leq K \leq 10$ Sv and (b) $0.1 \leq K \leq 400$ Sv. $\Omega = 0$, and C_h is tuned to yield $\bar{\psi}$ (for initial F) that is independent of K. The uncontoured region indicates that thermohaline catastrophe occurs. Shading indicates that the maximum amplitude response is more than a factor of 3 greater than the equilibrium response in (a), and more than a factor of 1.1 greater than the equilibrium response in (b). In (b), thick vertical lines distinguish between weak mixing ($K < 17.5$ Sv), intermediate mixing (17.5 Sv $\leq K \leq 140$ Sv) and strong mixing ($K > 140$ Sv).

the model provides a stronger positive feedback when mixing is weaker, the THC is less stable as a result. For very weak mixing, the initial equilibrium is unstable, so a significant decrease in freshwater forcing is required to prevent thermohaline catastrophe.

We now consider the limit at which Stommel's model implicitly operates: $K \gg 4\bar{\psi}$. The simplification $\dot{S}_\mathrm{s} = -2\dot{S}_2$, made in Section 4.2, is not valid at this limit, because mixing is rapid and the low-latitude ocean is homogenized. Instead, if $V_1 + V_3 \gg V_2$, $\dot{S}_S = \dot{S}_\mathrm{d} \approx -\dot{S}_2$. This is because symmetric changes in the freshwater content of the high- and low-latitude oceans result in a much greater change in salinity in the high-latitude ocean, if the low-latitude reservoir is much larger than the high-latitude reservoir. Therefore, in this limit, eq. (23) is replaced by

$$V_2\dot{S}_\mathrm{s} = S_0\Delta F - [(\bar{\psi} + \Omega) - C_h b\bar{S}_\mathrm{s}]S'_\mathrm{s} + C_h b(S'_\mathrm{s})^2. \qquad (30)$$

Although we use different volumes for high and low latitudes, unlike most studies based on Stommel's model, the only difference between eqs. (24) and (30), with $\Omega = 0$, is a factor of 2 throughout the right-hand side. Therefore, the stability criteria and critical ΔF in our model at the strong mixing limit are identical to those in Stommel's model.

The linear stability criterion is $\psi + \Omega \geq C_h b\bar{S}_\mathrm{s}$, so the left-hand sides of eqs. (27) and (28) are reduced by a factor of 2. This reduction in negative feedback, caused by the instantaneous dilution of box 1 salinity anomalies by diapycnal mixing, might suggest an even more stringent stability condition. However, because $S_\mathrm{d} = S_\mathrm{s}$ and $T_\mathrm{d} = T_\mathrm{s}$ we can rewrite the criterion as

$$a\bar{T}_\mathrm{s} \geq 2b\bar{S}_\mathrm{s}. \qquad (31)$$

Vertical homogenization means that, at this limit, the ratio of surface to interior meridional gradients, proportional to the ratio of positive to negative feedback terms, is decreased to 1. This can also be interpreted in terms of C_h. Equation (26) shows that the value of C_h (required to yield the same value of $\bar{\psi}$) is insensitive to K at this limit. However, C_h is much lower than that required when $K \ll \bar{\psi}$, so the linear positive feedback term is greatly reduced in the strong mixing limit.

Figure 8b is a repeat of Fig. 8a, but for a range of K between 0 and 400 Sv, to show the transition in stability characteristics between the extreme states of weak mixing and strong mixing. With reference to eqs. (16) and (17), we consider that the model starts to approach the weak mixing limit when $\bar{\psi}/K > 1$ and starts to approach the strong mixing limit when $8\bar{\psi}/K < 1$. In Fig. 8b, this yields the approximate domains $K < 17.5$ Sv for weak mixing, $17.5 \leq K \leq 140$ Sv for intermediate mixing, and $K > 140$ Sv for strong mixing (i.e. Stommel's model). Near the limit provided by Stommel's model, there is no significant recovery phase. This is to be expected, because in Stommel's model itself, the maximum response is identical to the equilibrium response. The THC is slightly more stable in the intermediate mixing domain than it is in the strong mixing domain, because of a decrease in negative feedback as mixing becomes stronger. In both the strong and intermediate domains, stability is highly insensitive to K. However, the THC is considerably less stable in the weak mixing domain. This is because the destabilizing increase in the difference between surface and interior meridional gradients, when mixing is weak, overwhelms the stabilizing processes discussed in Section 4.2.

The condition $K \geq \bar{\psi}$ (i.e. $\bar{\rho}_\mathrm{d} \geq \bar{\rho}_\mathrm{v}$) is not a plausible representation of the modern ocean. Therefore, the ratio of positive feedback to negative feedback is underestimated in models (such as Stommel's) that assume unlimited mixing but provide a reasonable rate of overturning. As a result, the stability of the THC to rapid changes in forcing appears to be overestimated in Stommel's model, and not underestimated as previous studies have suggested.

5. Discussion

Stommel's model and the thermocline scaling law (e.g. Bryan, 1987) represent the strong and weak mixing limits, respectively,

of a single one-hemisphere theory for thermohaline flow, the central assumption of which is that zonal and meridional velocities are linearly related. The simple model presented here can operate anywhere between the limits, although at the weak mixing limit the predicted non-linear power laws differ slightly from the thermocline scaling. Other studies (P99; NW01) have provided arguments to suggest that the THC is much more stable at the weak mixing limit than is at the strong mixing limit. Mixing is rate limiting at equilibrium when it is weak, but not when it is sufficient to keep the low-latitude ocean homogeneous. Therefore, the equilibrium dependence of overturning on buoyancy forcing is weaker when mixing is weak. However, diapycnal mixing can only affect the THC by modifying pressure gradients in the ocean interior. If diapycnal mixing is to provide a stabilizing role in the THC, then it must provide a negative feedback to changes in buoyancy forcing within the time-scale that these changes take effect in the ocean. Observations of changes over several decades in dense water in the Nordic Seas and the North Atlantic (Dickson et al., 2002) show that a time-scale for such externally forced changes of ∼40 yr, as provided by the model presented here, is reasonable.

Limited diapycnal mixing does contribute to a negative feedback in the model presented here, as it does in the thermocline scaling. Buoyancy accumulates in the low-latitude subsurface ocean in response to a weakening of the THC because the dense inflow can no longer balance the buoyancy input due to mixing. However, the process is one of vertical advective-diffusive balance and the time-scale is of several centuries, comparable to the residence time of water in the deep ocean. This explains why the evolution of the model on a decadal time-scale can be approximated solely in terms of the feedback mechanisms present in Stommel's model. Diapycnal mixing prevents thermohaline catastrophe within a small parameter domain only.

Nevertheless, the behaviour of the model diverges from that of Stommel's model before diapycnal mixing contributes significantly, and the thermally driven THC is stabilized as a result. The mechanism is the removal of any additional freshwater input at high latitudes to the deep ocean. The deep ocean is a large reservoir, so the effect on salinity there is small. However, the associated increase in salinity of the rapidly communicating high-latitude and surface low-latitude ocean becomes significant within 100 yr. Because the dynamical role of the high-latitude density in the THC is much greater than that of the low-latitude surface density (which plays no direct dynamical role in the model), the build up of salinity in these boxes initiates a 'recovery phase' in the THC, and in doing so can prevent thermohaline catastrophe. Positive feedback between high-latitude salinity and THC transport, similar to Stommel's classical mechanism, contributes to the recovery as well as the initial phase. Over a longer time-scale, the magnitude of the recovery phase, as well as the net change in THC transport, is controlled by the advective-diffusive balance between diapycnal mixing and thermohaline overturning.

As with any box model, we must consider the consequences of several simplifying assumptions. We have found that the volume of the deep reservoir must be much larger than that of the other boxes to obtain the correct time-scale for diapycnal mixing, but we have made an arbitrary assumption of $V_1 = V_2$. The ratio V_1/V_2 affects the separation between the two time-scales for salinity signals and diapycnal mixing. The surface low-latitude box, from which the high salinity signal is transmitted to high latitudes, is very narrow (200 m), which is necessary to allow density in that box to be dynamically inactive. As a result, the positive salinity anomaly is concentrated, and therefore requires less time to become significant. If the volume of this box is significantly increased, without changing the volume of the other boxes, the onset of the recovery phase occurs later (not shown), potentially providing more opportunity for thermohaline catastrophe to occur. This is not associated with a significant change in the diapycnal mixing time-scale. A further simplification is the use of instantaneous temperature restoring, and no coupling with the atmosphere/cryosphere. The case with Newtonian temperature restoring was considered in Oliver (2003). The domain of stable solutions is reduced in this case, but slow restoring tends to stabilize the model's response to changes in forcing through negative feedback between high-latitude temperature and overturning strength. However, these are small effects if reasonable restoring time-scales are used.

Potentially, a more important assumption is the nature of the mechanism by which buoyancy is fluxed to the deep ocean. The roles of wind driven upwelling of dense water in the Southern Ocean, and of geothermal heating, cannot be incorporated into this model without removing the analytical simplicity. The use of depth coordinates slightly reduces the equilibrium dependence of THC transport on mixing and increases the dependence on density difference, relative to the thermocline scaling. Diffusivity that is independent of stratification was also assumed. NW01 applied a more justifiable assumption of constant energy available for mixing, and found that the equilibrium dependence on density difference changes sign. This cannot be applied here because of the diabatic flow from box 2 to box 3. However, the primary purpose of this study is to improve understanding of the transient response, which is likely to be of much larger magnitude than the equilibrium response. The separation of time-scales necessary for the two-phase response, not present in the P99 or NW01 models, is associated with the relative size of the deep ocean reservoir and is unlikely to be removed by using a different parametrization for mixing.

There is equivocal support for the robustness of the time-scale separation from GCMs of differing complexity, forced by increasing high-latitude freshwater input that is imposed either directly or as a result of increased CO_2 levels. In a GCM integration with fixed mixing, forced by a doubling CO_2 over 70 yr, Manabe and Stouffer (1999) found a subcritical decline of about 50% in the MOC ∼150 yr after the CO_2 increase begins, followed by a recovery to near initial levels after 500 yr. (A direct

freshwater discharge experiment led to no recovery within 500 yr; this does not conflict with our results because the response was supercritical in that integration.) Imposing an incremental threefold increase in high-latitude freshwater input to a GCM with $\kappa \sim N^{-1}$ (buoyancy frequency, N, is proportional to the square root of vertical density gradient), Otterå et al. (2003) and Otterå et al. (2004) obtained a 30% decline in overturning over ~50 yr. The MOC then recovered to near initial values after 150 yr (the end of the integration), although an upward trend in MOC transport in a control simulation suggests the magnitude of this recovery is exaggerated. Differences between the two studies could have many causes, but in both cases the two-phase response in the rate of overturning, proposed here, is present. Determining the cause of the recovery phase in a GCM is less straightforward than in a box model. Manabe and Stouffer (1999) attributed the slow recovery to warming of the deep low-latitude ocean, amplified by a non-linear equation of state (a linear equation of state was applied here), caused by a reduced flow of cold water. The box model predicts this, but also predicts that the early stages of the recovery would be associated with a build up of salt near the surface at low latitudes, which either did not exist or was not noted in the Manabe and Stouffer study. In a more rapid recovery, Otterå et al. (2004) cited the role of diapycnal mixing, northward transport of saline waters (there was a low-latitude positive salinity anomaly extending to ~600 m near the end of their integration), and the maintenance of a near-constant wind-driven poleward flow of Atlantic water between the Faroe Islands and Scotland. All of these mechanisms are consistent with our results, but we would not predict a large contribution from diapycnal mixing. It could be argued that a large contribution would be expected because the GCM study employed stratification-dependent mixing, whereas mixing is constant here. Otterå et al. (2004) estimated that the rate of diapycnal upwelling increased by a maximum of 1 Sv in their experiment (the total recovery was ~6 Sv, including any underlying drift). The NW01 model suggests that, for mixing to cause an increase in the rate overturning (by the process of increasing subsurface buoyancy storage), the rate of upwelling across isopycnals must exceed the rate of overturning. Therefore, 1 Sv is an upper limit on the possible contribution of increased diapycnal mixing to the recovery phase in their study.

The model presented here is more stable when mixing is strong than when mixing is weak, even if the model is tuned to yield the same THC transport in each case. The reason is that strong mixing favours a strong THC, which can only be compensated by reducing the coefficient relating meridional density difference to THC transport, and therefore positive feedback. This suggests that Stommel's model overestimates stability, contrary to the predictions of the P99 and NW01 models. A stabilizing adjustment that can made in the NW01 model, but not here, is that of a constant energy source for diapycnal mixing. However, both P99 and NW01 obtained strong stabilization without this process. The negative feedback that they propose can only be strongly stabilizing in response to changes in forcing that occur on a similar, or longer, time-scale than of vertical advective-diffusive balance. More importantly, employing limited diapycnal mixing increases the sensitivity of the THC to changes on a shorter time-scale, increasing the probability of thermohaline catastrophe.

6. Acknowledgements

A RAPID programme grant (NER/T/S/2002/00446), funded by the Natural Environment Research Council (NERC), supported the writing of this paper. The authors thank Karen Heywood and two anonymous reviewers for valuable feedback.

Appendix A: Linear stability

The linear stability of the system is derived here. The combination of eqs. (9) and (10) can be rewritten

$$V_2\dot{S}_\mathrm{s} = 2\bar{S}F - (2\psi + 2\Omega + K)S_\mathrm{s} + (\psi + K)S_\mathrm{d}, \tag{A1}$$

if $V_1 = V_2$ and using $S_\mathrm{v} = S_\mathrm{s} - S_\mathrm{d}$. If, for example, S_d is broken down into $\bar{S}_\mathrm{d}$ and S'_d, which is a small perturbation from $\bar{S}_\mathrm{d}$ so that terms containing ${S'_\mathrm{d}}^2$ can be ignored, eq. (A1) can be written

$$\begin{aligned} V_2\dot{S}_\mathrm{s} &= -(2\bar{\psi} + 2\Omega + K)S'_\mathrm{s} - (2\bar{S}_\mathrm{s} - \bar{S}_\mathrm{d})\psi' + (\bar{\psi} + K)S'_\mathrm{d} \\ &= -(2\bar{\psi} + 2\Omega + K)S'_\mathrm{s} - \frac{C_h\bar{S}_\mathrm{s}(2\bar{\psi} + K)}{\bar{\psi} + K}aT'_\mathrm{d} \\ &\quad + \left[\bar{\psi} + K + \frac{C_h b\bar{S}_\mathrm{s}(2\bar{\psi} + K)}{\bar{\psi} + K}\right]S'_\mathrm{d}. \end{aligned}$$

Here, the results that the right-hand side of eq. (A1) is zero at equilibrium, $\psi' = C_h(aT'_\mathrm{d} - bS'_\mathrm{d})$ and $\bar{S}_\mathrm{d}/\bar{S}_\mathrm{s} = K/(\bar{\psi} + K)$, have been used.

If a similar process is used to derive $\dot{S}_\mathrm{d}$ and $\dot{T}_\mathrm{d}$, the first-order response of the model to perturbations is fully described by

$$V_2\begin{pmatrix}\dot{S}_\mathrm{s}\\ \dot{S}_\mathrm{d}\\ \dot{T}_\mathrm{d}\end{pmatrix} = \begin{pmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{pmatrix}\begin{pmatrix}S'_\mathrm{s}\\ S'_\mathrm{d}\\ T'_\mathrm{d}\end{pmatrix}, \tag{A2}$$

with

$$\begin{aligned} x_{11} &\equiv -(2\bar{\psi} + 2\Omega + K), \\ x_{12} &\equiv \bar{\psi} + K + \frac{C_h b\bar{S}_\mathrm{s}(2\bar{\psi} + K)}{\bar{\psi} + K}, \\ x_{13} &\equiv -\frac{C_h a\bar{S}_\mathrm{s}(2\bar{\psi} + K)}{\bar{\psi} + K}, \\ x_{21} &\equiv -(\bar{\psi} + \Omega) + \frac{V_2}{V_3}(K), \\ x_{22} &\equiv C_h b\bar{S}_\mathrm{s} - \frac{V_2}{V_3}\left(\bar{\psi} + K - \frac{C_h b\bar{S}_\mathrm{s}K}{\bar{\psi} + K}\right), \\ x_{23} &\equiv -C_h a\bar{S}_\mathrm{s} - \frac{V_2}{V_3}\left(\frac{C_h a\bar{S}_\mathrm{s}K}{\bar{\psi} + K}\right), \end{aligned}$$

$$x_{31} \equiv 0,$$
$$x_{32} \equiv \frac{V_2}{V_3}\left(\frac{C_h b T_s K}{\bar{\psi}+K}\right),$$
$$x_{33} \equiv -\frac{V_2}{V_3}\left(\bar{\psi}+K+\frac{C_h a T_s K}{\bar{\psi}+K}\right).$$

The Lyapunov test for stability is that there must be no eigenvalues with a positive real part, which would represent a growth term. The equations for the eigenvalues (Λ_1, Λ_2 and Λ_3) are too long to be transcribed here. However, in the limit $V_2/V_3 \rightarrow 0$, two eigenvalues tend towards the eigenvalues in a 2×2 matrix with the last row and column removed (i.e. assuming $\dot{T}_d = 0; T'_d = 0$). The model operates near this limit because $V_2/V_3 = 1/19$. We thus first consider the stability of the simplified system. In a 2×2 matrix, both eigenvalues have negative real parts if the sum of the leading diagonal (i.e. $\Lambda_1 + \Lambda_2$) is negative and the determinant ($\Lambda_1\Lambda_2$) is positive. These yield, respectively,

$$2\bar{\psi} + K + 2\Omega \geq C_h b \bar{S}_s, \tag{A3}$$

and

$$\frac{(\bar{\psi}+\Omega)(\bar{\psi}+K)^2}{K(2\bar{\psi}+K+\Omega)} \geq C_h b \bar{S}_s. \tag{A4}$$

Neither of these conditions are trivially satisfied. The former condition limits stability if

$$\bar{\psi}^3 - 2K\bar{\psi}^2 - 3K^2\bar{\psi} - K^3 + \Omega[\bar{\psi}^2 - 2K(\bar{\psi}+K+\Omega)] > 0. \tag{A5}$$

When $\Omega = 0$, this leads to $\bar{\psi}/K >\sim 3$. It is expected that K is of the order of 5 Sv and $\bar{\psi}$ is of the order of 16 Sv, so neither condition can be ignored; the former condition limits stability at small K and the latter condition limits stability at large K.

Stability limits can be obtained by substituting eqs. (A3) or (A4) into eq. (15) to yield $\psi_{\min}$, the minimum thermally driven THC transport that is stable, and solving simultaneously with eq. (21) to obtain $F_{\max}$, the freshwater forcing limit. Using eq. (A3) for small K, we have

$$\bar{\psi}^2_{\min} + 3K\bar{\psi}_{\min} - K(C_h a T_s - K - 2\Omega) = 0, \tag{A6}$$

and

$$F_{\max} = \frac{1}{C_h b S_0}\left[\left(2-\frac{\Omega}{K}\right)\bar{\psi}^2_{\min} + (K+\Omega)\bar{\psi}_{\min} + \Omega C_h a T_s\right]. \tag{A7}$$

Using eq. (A4) for large K we have

$$3\bar{\psi}^3_{\min} + (5K+2\Omega)\bar{\psi}^2_{\min} - K(2C_h a T_s - 2K - 3\Omega)\bar{\psi}_{\min} - K[(K+\Omega)C_h a T_s - \Omega K] = 0, \tag{A8}$$

and

$$F_{\max} = \frac{1}{3C_h b S_0}\left[\left(2-\frac{\Omega}{K}\right)\bar{\psi}^2_{\min} + (C_h a T_s + 2K)\bar{\psi}_{\min} + \Omega K + (2\Omega - K)C_h a T_s\right]. \tag{A9}$$

Figure 4 shows $F_{\max}$, dependent on K and Ω. Both diapycnal and horizontal mixing tend to increase $F_{\max}$ throughout the domain. Equations (A7) and (A9) apply in different parts of the domain. With vanishing K in the domain considered, eq. (A7) applies. Assuming $\Omega \gg K$ and $\Omega \gg \bar{\psi}$, simultaneous solution with eq. (10) at steady state yields

$$F_{\max} \approx \frac{2\Omega^2}{C_h b S_o}. \tag{A10}$$

Horizontal mixing is powerfully stabilizing at this limit. However, the resulting overturning, given by eq. (18), would be very weak because both K and $\bar{\rho}_s$ would be small.

The equilibrium is a spiral point, and oscillations can occur, if the eigenvalues have an imaginary part. This is the case if $(\Lambda_1 + \Lambda_2)^2 - 4\Lambda_1\Lambda_2 < 0$, which leads to

$$\bar{\psi}\Omega - C_h b \bar{S}_s\left(\bar{\psi}+K+\frac{K^2}{\bar{\psi}+\Omega}\right) < 0. \tag{A11}$$

When $\Omega = 0$ and $F > 0$, the equilibrium is always a spiral point. Horizontal mixing tends to suppress oscillations, whereas greater freshwater forcing tends to increase the probability of their existence.

Because cubic equations have either one or three real roots and Λ_1 and Λ_2 are either both real or both complex when V_2/V_3 is small, Λ_3 is real. This indicates that oscillations are unlikely to be introduced to the system by slow changes to the properties of the deep ocean. The determinant of the 3×3 matrix ($\Lambda_1\Lambda_2\Lambda_3$) must not be positive in a stable system. Using $\Lambda_1\Lambda_2$ as the determinant of the 2×2 matrix, it can be deduced that

$$\Lambda_1\Lambda_2\Lambda_3 = -\frac{V_2}{V_3}\left[\left(\bar{\psi}+K+\frac{C_h a T_s K}{\bar{\psi}+K}\right)\Lambda_1\Lambda_2 + \frac{C_h^2 a T_s b \bar{S}_s K(K+\Omega)}{\bar{\psi}+K}\right]. \tag{A12}$$

It is readily apparent that $\Lambda_1\Lambda_2 > 0 \Rightarrow \Lambda_3 < 0$ (assuming that $F \geq 0$). Therefore, when V_2/V_3 is small, the limit of system stability is not affected by changes in box 3. With $V_2/V_3 = 1/19$, instability associated with changes in the deep ocean has not been observed in long integrations.

References

Adcroft, A., Scott, J. R. and Marotzke, J. 2001. Impact of geothermal heating on the global ocean circulation. *Geophys. Res. Lett.* **28**, 1735–1738.

Bryan, F. 1986. High-latitude salinity effects and interhemispheric thermohaline circulations. *Nature* **323**, 301–323.

Bryan, F. 1987. Parameter sensitivity of primitive equation ocean general circulation models. *J. Phys. Oceanogr.* **17**, 970–985.

Bryan, K. and Cox, M. D. 1967. A numerical investigation of the oceanic general circulation. *Tellus* **19**, 54–80.

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

Dansgaard, W., Johnsen, S. J., Clausen, H. B., Dahl-Jensen, D., Gundestrup, N. S. and co-authors. 1993. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–220.

Dickson, B., Yashayaev, I., Meincke, J., Turrell, B., Dye, S. and Holfort, J. 2002. Rapid freshening of the deep North Atlantic Ocean over the past four decades. *Nature* **416**, 832–837.

Gargett, A. E. and Ferron, B. 1996. The effects of differential vertical diffusion of T and S in a box model of thermohaline circulation. *J. Marine Res.* **54**, 827–866.

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

Griffies, S. M. and Tziperman, E. 1995. A linear thermohaline oscillator driven by stochastic atmospheric forcing. *J. Climate* **8**, 2440–2453.

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

Johnson, H. L. and Marshall, D. P. 2002. A theory for the surface Atlantic response to thermohaline variability. *J. Phys. Oceanogr.* **283**, 1121–1132.

Joyce, T. M. 1991. Thermohaline catastrophe in a simple 4-box model of the ocean climate. *J. Geophys. Res.* **96**(C11), 20 393–20 402.

Kawase, M. 1987. Establishment of deep ocean circulation driven by deep-water production. *J. Phys. Oceanogr.* **17**, 2294–2317.

Lyle, M. 1997. Could early Cenozoic thermohaline circulation have warmed the poles? *Paleoceanography* **12**, 161–167.

McDermott, D. A. 1996. The regulation of northern overturning by southern hemisphere winds. *J. Phys. Oceanogr.* **26**, 1234–1255.

Manabe, S. and Stouffer, R. J. 1999. The role of thermohaline circulation in climate. *Tellus* **51A**, 91–109.

Marotzke, J. 1990. *Instability and multiple equilibria of the thermohaline circulation*. PhD thesis, Berichte aus dem Institut für Meereskunde, 126 pp.

Marotzke, J. 1997. Boundary mixing and the dynamics of three-dimensional thermohaline circulation. *J. Phys. Oceanogr.* **27**, 1713–1728.

Marotzke, J. and Klinger, B. A. 2000. The dynamics of equatorially asymmetric thermohaline circulations. *J. Phys. Oceanogr.* **30**, 955–970.

Munk, W. H. 1966. Abyssal recipes. *Deep-Sea Res.* **13**, 707–730.

Munk, W. H. and Wunsch, C. 1998. Abyssal recipes II: energetics of tidal and wind mixing. *Deep-Sea Res. I* **45**, 1977–2010.

Nakamura, M., Stone, P. and Marotzke, J. 1994. Destabilization of the thermohaline circulation by atmospheric eddy transports. *J. Climate* **7**, 1870–1882.

Nilsson, J. and Walin, G. 2001. Freshwater forcing as a booster of thermohaline circulation. *Tellus* **53A**, 629–641, (NW01).

Nilsson, J., Broström, G. and Walin, G. 2003. The thermohaline circulation and vertical mixing: does weaker density stratification give stronger overturning? *J. Phys. Oceanogr.* **33**, 2781–2795.

Nilsson, J., Broström, G. and Walin, G. 2004. On the spontaneous transition to asymmetric thermohaline circulation. *Tellus* **56A**, 68–78.

Oliver, K. I. C. 2003. *Elements of the thermohaline circulation: high-latitude buoyancy forcing and low-latitude mixing*. PhD thesis, School of Environmental Sciences, University of East Anglia, Norwich, UK, 202 pp.

Otterå, O. H., Drange, H., Bentsen, M., Kvamstø, N. G. and Jiang, D. 2003. The sensitivity of the present-day Atlantic meridional overturning circulation to freshwater forcing. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017578.

Otterå, O. H., Drange, H., Bentsen, M., Kvamstø, N. G. and Jiang, D. 2004. Transient response of the Atlantic Meridional Overturning Circulation to enhanced freshwater input to the Nordic Seas–Arctic Ocean in the Bergen Climate Model. *Tellus*, **56A** 342–361.

Park, Y.-G. 1999. The stability of thermohaline circulation in a two-box model. *J. Phys. Oceanogr.* **29**, 3101–3110, (P99).

Park, Y.-G. and Bryan, K. 2000. Comparison of thermally driven circulations from a depth coordinate model and an isopycnal layer model. Part I: a scaling law – sensitivity to vertical diffusivity. *J. Phys. Oceanogr.* **30**, 590–605.

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

Rivin, I. and Tziperman, E. 1997. Linear versus self-sustained interdecadal thermohaline variability in a coupled box model. *J. Phys. Oceanogr.* **27**, 1216–1232.

Rooth, C. 1982. Hydrology and ocean circulation. *Prog. Oceanogr.* **11**, 131–149.

Ruddick, B. and Zhang, L. Q. 1996. Qualitative behavior and non-oscillation of Stommel's thermohaline box model. *J. Climate* **9**, 2768–2777.

Shaffer, G. and Olsen, S. M. 2001. Sensitivity of the thermohaline circulation and climate to ocean exchanges in a simple coupled model. *Clim. Dyn.* **17**, 433–444.

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

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

Toggweiler, J. R. and Samuels, B. 1998. On the ocean's large-scale circulation near the limit of no vertical mixing. *J. Phys. Oceanogr.* **28**, 1832–1852.

Vellinga, M. 1996. Instability of two-dimensional thermohaline circulation. *J. Phys. Oceanogr.* **26**, 305–319.

Welander, P. 1971. The thermocline problem. *Phil. Trans. R. Soc. London* **21**, 415–421.

Weijer, W. and Dijkstra, H. A. 2001. A bifurcation study of the three-dimensional thermohaline circulation: the double hemispheric case. *J. Marine Res.* **59**, 599–631.

Wright, D. G., Vreugdenhil, V. G. and Hughes, T. M. C. 1995. Vorticity dynamics and zonally averaged ocean circulation models. *J. Phys. Oceanogr.* **25**, 2142–2154.