DYNAMICS OF THE THERMOHALINE CIRCULATION - cko Lectures

Jochem Marotzke

September 15, 2007
## Contents

1 Thermohaline Circulation and Salinity Effects 5  
1.1 Preliminaries .................................................. 5  
1.2 Deep western boundary currents .......................... 6  
1.3 Salinity effects ............................................... 15  

2 Box Models of the THC 19  
2.1 Multiple Equilibria of the THC ......................... 19  
  2.1.1 Introduction - 2-box model .......................... 19  
  2.1.2 Equilibrium solutions ................................. 23  
  2.1.3 Stability .................................................. 28  
      2.1.3.1 Lyapunov potential .............................. 31  
      2.1.3.2 Feedbacks .......................................... 32  
  2.1.4 Time-dependent solution ............................. 36  
2.2 Box Models: Interhemispheric Flow ..................... 40  
  2.2.1 Preliminaries .......................................... 40  
  2.2.2 Rooth’s model: Formulation ......................... 42  
  2.2.3 Rooth’s model: Instability of the symmetric state .. 42  
  2.2.4 Rooth’s model: Steady states and their stability .... 43  

3 THC Theory 49  
3.1 Preliminaries ............................................... 49  
3.2 Early approaches: Scaling ................................... 50  
3.3 Two-dimensional Models .................................... 52  
3.4 Boundary-layer Approaches ............................... 53  

4 Climate Variability: Stochastic Climate Models 61  
4.1 Preliminaries ............................................... 61
4.2 Pedagogical example: The Game of Peter and Paul (after Wunsch (1992))

Bibliography
Chapter 1

Thermohaline Circulation and Salinity Effects

1.1 Preliminaries

The thermohaline circulation (THC) is driven by surface fluxes of heat (changing the temperature - “thermo”) and freshwater (changing the salinity - “haline”), which combined change water density and hence pressure. The terms “thermohaline circulation” and “buoyancy-driven flow” (buoyancy = gravitational acceleration times density anomaly) are, strictly speaking, synonymous. But there has been a tendency to use the term THC in a more restricted sense, for that part of the ocean circulation associated with convection and sinking, upwelling from depth, and the horizontal flows feeding these vertical motions. There is no reason to prefer one expression to the other, except perhaps when one wishes to emphasise that temperature and salinity interact very differently with the atmosphere (see Chapter 6). In that case, referring explicitly to temperature and salinity through the term “THC” is in part a reminder of these different interactions. A more important issue of notation arises because ocean dynamics are nonlinear, making it impossible rigorously to separate wind-driven from buoyancy-driven circulations. For simplicity, we will nevertheless often pretend that they are separable.
1.2 Deep western boundary currents

What are the dynamics controlling the global THC, as depicted, for example, in Gordon’s (1986) cartoon? Even non-oceanographers know that the surface flows are concentrated in western boundary currents, which are poleward in the subtropics and middle latitudes. In the North Atlantic, much of the net northward near-surface mass transport of the THC (e.g., Fig. 11, Macdonald, Prog. Oceanogr., 41, 281-382, 1998) occurs in the Gulf Stream, which is thus a prime example of a current that is both wind- and buoyancy-driven.

It is less well known that the deep flow is likewise concentrated near the western boundary. After it was discovered in the early 19th century that deep water was cold even at low latitudes and hence had to originate from high latitudes, the deep equatorward flow was thought to spread over the entire basin. But Stommel and Arons (1960; see Warren, 1981, for a beautiful discussion of the history) predicted that strong boundary currents had to exist in the deep ocean as well. Hence, the flow of North Atlantic Deep Water occurs mainly in such a deep western boundary current (DWBC).

We now go through a very simple rendition of the Stommel-Arons theory...
(Sketch). Assume that the ocean is flat-bottomed and consider the vertically integrated flow in a bottom layer. Assume horizontally uniform upwelling from this layer, compensating strong, localised downwelling at the convection sites. (N.B.: All these assumptions have since been severely discredited. Ocean bottom topography is important, upwelling - probably associated with regions of vigorous mixing - might be highly localised rather than widespread, and downwelling is unlikely to occur at the convection sites. Still, the prediction of DWBC’s by the Stommel-Arons theory is striking, and the vast majority of oceanographers consider it the standard theory of deep circulation).

Starting from geostrophy (we use Cartesian coordinates for simplicity),

\[-fv = -\frac{1}{\rho_0} \partial_x p, \]  
\[fu = -\frac{1}{\rho_0} \partial_y p, \]  
\[\text{(1.1)}\]

\[\text{(1.2)}\]

cross differentiation and subtraction of the equations yields

\[-f(\partial_x u + \partial_y v) - v \frac{df}{dy} = 0. \]  
\[\text{(1.3)}\]

With the continuity equation

\[\partial_x u + \partial_y v + \partial_z w = 0 \]  
\[\text{(1.4)}\]

and the definition \(\beta \equiv \frac{df}{dy}\), this gives

\[\beta v = f \partial_z w, \]  
\[\text{(1.5)}\]

which is called the planetary (or linear) vorticity equation. Integrating (1.5) over the bottom layer gives (the flat bottom implies that \(w = 0\) there)

\[\beta V = fw, \]  
\[\text{(1.6)}\]

which says that upwelling out of the deep layer implies poleward horizontal flow (toward the source!), in contrast to the 19th-century expectation. Making the standard “\(\beta\)-plane” approximation in (1.6), \(f = \beta y\), \(\beta\) constant, leads to

\[V = yw \]  
\[\text{(1.7)}\]
and thus

$$\partial_y V = w$$  \hspace{1cm} (1.8)

since $w$ is assumed constant. Not only is deep flow poleward, according to (1.6), but this poleward flow even increases with latitude, as (1.8) shows. If planetary vorticity conservation holds at all longitudes, (1.8) implies for the zonal integral across the entire basin that

$$\int_0^L \partial_y V \, dx = \int_0^L w \, dx.$$  \hspace{1cm} (1.9)

But the vertically integrated continuity equation leads to a different conclusion. From (1.4),

$$\partial_x U + \partial_y V + w = 0.$$  \hspace{1cm} (1.10)

Integration across the basin yields, noting that the zonal flow vanishes at the eastern and western boundaries if these follow longitude lines,

$$\int_0^L \partial_y V \, dx = - \int_0^L w \, dx,$$  \hspace{1cm} (1.11)

the exact opposite of (1.9), which is derived from vorticity conservation! Both equations cannot be right, and since mass conservation is correct to high accuracy (here with the very sensible interpretation that the upwelling is fed from the decreasing northward transport), we conclude that (1.9) cannot be true everywhere. A boundary current must exist with a dynamical balance different from the linear vorticity relation. This boundary current must supply both the interior increase in $V$ with latitude and the upwelling out of the deep, which are equal if the width of the boundary current is much less than $L$. That the boundary current must be on the western side follows from arguments similar to those used for the westward intensification of surface currents, and we will not go into any detail here.

If one assumes a mass source in the northwestern corner of a basin, a southward boundary current emanates, which weakens as it progresses southward, due to “leakage” to the east. Depending on the strength of the source, the DWBC loses all its mass at some latitude or reaches the Antarctic Circumpolar Current (ACC). Half of this mass lost upwells over the basin, while
Figure 1.2: (Warren, Fig. 1.1): Deep circulation in a schematic world ocean driven by uniform upwelling with sources at the counterparts of the North Pole and Weddell Sea. See text. Transports measured in units of about $6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$. (Kuo and Veronis, 1973.)
the other half reaches the northern boundary, from where it recirculates (according to unspecified dynamics) to the northwestern corner.

The details of the ensuing picture (Fig. 1.2) depend on the assumed distribution of mass sources, but two conclusions are robust:

i) Flow away from the source occurs in the DWBC only; all interior flow is poleward.

ii) Cross-equatorial flow occurs only in the DWBC.

How well does the Stommel-Arons theory describe the real ocean? Generally, one should be very critical, but the existence of DWBCs is a robust phenomenon: We see them in the real ocean (Fig. 1.3); Dickson et al. (1990) made direct current measurements off the East Greenland coast. The DWBC shows up as a very fast (speed > 30cm/s) and narrow (width ca.40 km) flow of very high-density water “hugging” the western topography. This water is
Figure 1.4: (Smethie and Swift, 1989, J. Geophys. Res., 94, 8265-8275, Fig. 2) Observed potential temperature (in °C) from the TTO section just south of the Denmark Strait (from 65°N, 35°W to 63°N, 28°W) [from Smethie and Swift (1989)].
also very cold (Fig. 1.4, from Smethie and Swift 1989). Equatorward in the Atlantic, the DWBC is most easily identified through the Chlorofluorocarbons (CFCs), which are purely anthropogenic and show waters that recently were in contact with the atmosphere (see Fig. 8 in Smethie, 1993, Prog. Oceanogr., 31, 51-99). In the South Atlantic, there is a very clear signature of the DWBC in all properties: High temperature, high salinity, high oxygen, low silica (Fig. 1.10 from Warren, 1981).

Curiously at first, the DWBC now is a buoyant, rather than dense, anomaly when compared with the ambient water at the same latitude. This has to be so from thermal-wind considerations, meaning that the deep zonal pressure and hence density gradients must change sign at the equator as the DWBC crosses the equator. But even a rudimentary dynamical explanation is very complicated (Marotzke and Klinger, J. Phys. Oceanogr., 30, 2000, 955-970).

The high salinity and low silica characteristic of North Atlantic Deep water can be found in the South Indian and South Pacific Oceans to about the equator; then the trace is lost.
Figure 1.6: (Warren, Fig. 1.10B)

Figure 1.7: (Warren, Fig. 1.10C)
Figure 1.8: (Warren, 1.10D): Sections of (A) potential temperature ($^\circ$C), (B) salinity (0/00), and the concentrations of (C) dissolved oxygen (ml l$^{-1}$) and (D) silica ($\mu$ M l$^{-1}$) along roughly lat. 30$^\circ$S from South America (left) to the Mid-Atlantic Ridge, illustrating the two deep western boundary currents of the South Atlantic, namely, the northward-flowing Antarctic Bottom Water and the southward-flowing North Atlantic Deep Water above.
1.3 Salinity effects

Seawater consists of freshwater and salt. The dissolved salts have a nearly universal mixing ratio; for physical purposes, it is sufficient to know the total salt mass. Temperature and salinity have influences on density that are of the same order of magnitude, in contrast to the other familiar two-component fluid, the atmosphere, where water vapour has a noticeable influence on density only in the tropics. The thermal expansion coefficient, $\alpha$, is defined as

$$\alpha \equiv \frac{-1}{\rho} \frac{\partial \rho}{\partial T}, \quad (1.12)$$

and is a function of temperature, salinity, and pressure (see Gill, 1982, Table A3.1). Unlike freshwater, seawater has a density maximum below the freezing point (Fig. 1.9), but this unattainable density maximum can be felt at low temperatures where $\alpha$ is very small (Fig. 1.10. At the surface, $\alpha$ varies by one order of magnitude, from $0.25 \times 10^{-4} K^{-1}$ at $-2^\circ C$ to $3.4 \times 10^{-4} K^{-1}$ at $31^\circ C$. When we assume an intermediate value of $\alpha(13^\circ C, 35 \text{ psu}) = 2 \times 10^{-4} K^{-1}$ as representative, the observed temperature range of $25^\circ C$ gives a thermally-induced density range of $\Delta \rho_T = 5 \text{ kg m}^{-3}$.

The haline expansion coefficient, $\beta$, is defined as

$$\beta \equiv \frac{1}{\rho} \frac{\partial \rho}{\partial S}, \quad (1.13)$$

and varies by only about 10 %, so a value of $\beta = 8 \times 10^{-4}$ is appropriate for seawater. With a salinity range of perhaps 2.5, this leads to a salinity-induced density range of $\Delta \rho_S = 2 \text{ kg m}^{-3}$ — less than, but of the same order of magnitude as, $\Delta \rho_T$.

At high latitudes, however, where $\alpha$ is very small, salinity fluctuations can dominate density fluctuations. In particular, surface salinity decides whether in a given winter deep convection occurs (see Fig. 1.11): Typically, cool and fresh water overlays warm and saline water. The surface salinity then decides about whether winter cooling makes water dense enough to overturn convectively, or whether cooling even to the freezing point leaves surface water too buoyant. In the latter case, see ice forms, which insulated the ocean very effectively against further heat loss, and convection is suppressed. (N.B.: If ice gets exported, for example by wind drift, the continuously forming new sea ice might eventually leave enough salt behind in the surface
layer that convection occurs later in the winter). As a rule, high surface salinity is needed for convection to occur regularly; the North Pacific is so low in salinity that it is never convects to great depths (see, e.g., the Levitus, 1982, atlas). The contrast in sea surface salinity determines the different roles of Atlantic and Pacific in the global THC (Warren, 1983).

But why are the two oceans so different in surface salinity? To understand this, we must ask “what if” questions, for example, “What would it take to make the Atlantic look like the Pacific, and vice versa?” This requires the use of models, because “The exchange of letters in Nature shows how futile simple verbal arguments can be in discussing such issues. The reader with a morbid interest in fallacious verbal theories may find it entertaining to look over the work of the English eccentric William Leighton Jordan” (From footnote 3 in Robinson and Stommel, Tellus, 11, 295-308, 1959). The next lecture deals with how to force models of the THC. After we understand that, we will be ready to simplify the models to the extent that a conceptual model can tell us why oceans can behave as differently as North Atlantic and North Pacific.
Figure 1.9: Dietrich et al., Allgemeine Meereskunde, 3rd ed., Borntraeger, Berlin, Fig. 2.07: Density $\sigma_T$ of seawater, as a function of temperature $T$ and salinity $S$. 
Figure 1.10: (Dietrich et al., Allgemeine Meereskunde, 3rd ed., Borntraeger, Berlin, Fig. 2.08): Thermal expansion coefficient $\alpha \times 10^{-4}$ of seawater at atmospheric pressure, as a function of temperature $T$ and salinity $S$. $\delta$: Freezing point of seawater.

Figure 1.11: Fig. 6 from Dickson et al., Prog. Oceanogr., 20, 103-151, 1988: Schematic diagram illustrating the suppression of convection north of Iceland as upper-ocean salinities decrease below 34.7.
Chapter 2

Box Models of the THC

2.1 Multiple Equilibria of the THC

2.1.1 Introduction - 2-box model

We have, in the previous lectures, laid the groundwork for this one, which is arguably the central lecture concerning a conceptual understanding of the role of ocean circulation in climate dynamics. We introduce a box model, which represents the North Atlantic thermohaline circulation (THC) in its simplest possible form: The entirety of the low latitudes is represented by a single, well mixed box, as are the entire high latitudes. The model was introduced by Stommel over forty years ago (Stommel (1961)); we use here the simplification of Marotzke (1990). Despite its simplicity, the model displays an astonishing range of phenomena, many of which are central to a general theoretical understanding of dynamical systems\(^1\). All aspects of this model can be calculated analytically, and exactly, with the exception of the explicit time-dependent behaviour under time-varying forcing.

Heuristically, we assume that the atmosphere controls the ocean temper-

\(^1\)I think that this model plays - or should play - a role in understanding the THC and some classes of complex systems that is comparable to the role of the linear harmonic oscillator in basic physics. To compare it to another important paradigm: When I was a beginning graduate student, a very wise lecturer, Ulf Larsen, told us over and over again how important it was to illustrate the principles of quantum statistical mechanics with the simplest example, an isolated spin-1/2 particle, having just two quantum states (spin up and spin down). Being young and foolish, we used to chuckle, feeling that we wanted hard problems, not simple ones. I have since come to recognise our folly for what it was.
Figure 2.1: Stommel’s box model
ature and the surface fresh-water loss or gain, $E$ (in m/s). In the preceding lecture, we saw that this approximation is equivalent to assuming a Haney restoring law for heat flux with infinitely strong coupling; or, we use the extreme case of mixed thermohaline boundary conditions. We will see in the lectures on the coupled box model, later in this course, how to view this approximation as the limiting case of the coupled system. For now, let us proceed with the assumption that $T_1$, $T_2$, and $E$ are prescribed as external parameters\(^2\).

Again, as in the preceding lecture, we will use a virtual surface salinity flux, $H_S$:

$$H_S = S_0 E/D$$

where $D$ is depth and $S_0$ a reference salinity.

The boxes are connected by pipes near the surface and the bottom; the pipes are assumed to have vanishing volume but are conduits for the flow. The thermohaline circulation strength is denoted by $q$ (strictly speaking, $q$ represents THC/Volume; $q$ has units of $s^{-1}$). We use the sign convention that $q > 0$ denotes poleward surface flow, implying equatorward bottom flow and, conceptually, sinking at high latitudes. This is the picture that we are used to when thinking about the North Atlantic THC. Conversely, $q < 0$ means equatorward surface flow and poleward bottom flow. We assume a very simple flow law for $q$, namely, that $q$ depends linearly on the density difference between high and low latitudes:

$$q = \frac{k}{\rho_0} [\rho_1 - \rho_2]$$

where $\rho_0$ is a reference density and $k$ is a hydraulic constant, which contains all dynamics, that is, the connection between density and the flow field. The equation of state is

$$\rho_i = \rho_0 (1 - \alpha T_i + \beta S_i); i = 1, 2,$$

where $\alpha$ and $\beta$ are, respectively, the thermal and haline expansion coefficients,

\(^2\)This is where we depart from Stommel (1961) and instead follow Marotzke (1990). Stommel (1961) used Haney-type conditions for both temperature and salinity, but with a longer restoring timescale for salinity. As a consequence, the original Stommel box model cannot readily be solved analytically.
\[ \alpha \equiv -(1/\rho_0)\partial_T \rho; \quad \beta \equiv (1/\rho_0)\partial_S \rho. \]  

(2.4)

For simplicity, we employ a linear equation of state; that is, both \( \alpha \) and \( \beta \) are assumed constant. The flow law, 2.2, thus becomes, using 2.3,

\[ q = k[\alpha(T_2 - T_1) - \beta(S_2 - S_1)] \]  

(2.5)

As we assume that the temperatures are fixed by the atmosphere and enter the problem as external parameters, we need not formulate a heat conservation equation. The salt conservation equations for the Stommel model are

\[
\begin{align*}
\dot{S}_1 &= -H_S + |q|(S_2 - S_1) \\
\dot{S}_2 &= H_S - |q|(S_2 - S_1)
\end{align*}
\]  

(2.6)  

(2.7)

which may require a little explanation. We postulate that flow into a box carries with it the properties, in particular the salinity, of the originating box. (We note in passing that this is equivalent to “upstream differencing”). So, if \( q > 0 \), the upper pipe brings water with salinity \( S_2 \) into Box 1, while the lower pipe takes water with \( S_1 \) out of Box 1. If \( q < 0 \), it is the lower pipe that imports \( S_2 \) into Box 1, while the upper pipe exports \( S_1 \) out of Box 1. In either case, \( S_2 \) is imported into Box 1, while \( S_1 \) is exported out of Box 1, both at a rate given by the modulus of \( q \). This is what (2.6) expresses. Mutatis mutandis, the same holds for Box 2 and (2.7).

We introduce the following abbreviations for meridional differences of temperature, salinity, and density:

\[ T \equiv T_2 - T_1; \quad S \equiv S_2 - S_1; \quad \rho \equiv \rho_1 - \rho_2 \]  

(2.8)

which implies that

\[ q = \frac{k}{\rho_0} \rho = k[\alpha T - \beta S] \]  

(2.9)

Under normal conditions, net evaporation occurs at the warmer low latitudes and net precipitation at the colder high latitudes; in other words, temperature and salinity are both expected to be high at low latitudes and low at high latitudes. Concerning which of temperature and salinity influences the THC most strongly, two cases can be distinguished. When the temperature
difference dominates the salinity difference in their influence on density, high-latitude density is greater than the low-latitude density. Therefore, \( q > 0 \), and the surface flow is poleward. One can say that the temperature difference, \( T \), drives the THC and the salinity difference, \( S \), brakes the THC, as seen from

\[
q > 0 : |q| = q = k[\alpha T - \beta S] \quad (2.10)
\]

Conversely, when the salinity difference dominates the temperature difference, high-latitude density is lower than the low-latitude density, \( q < 0 \), the surface flow is equatorward. Now, \( S \) drives the THC, and \( T \) brakes it:

\[
q < 0 : |q| = -q = k[\beta S - \alpha T] \quad (2.11)
\]

The sum of the salt conservation equations 2.6 and 2.7 gives

\[
\dot{S}_1 + \dot{S}_2 = 0 \quad (2.12)
\]

reflecting that total salt mass is conserved. (One consequence of this simplification is that we cannot determine the mean salinity from the set of equations we use here. Processes other than evaporation, precipitation, and oceanic transport of salinity must be invoked for the determination of the total oceanic salt content.) Because of the constancy of total salt mass, (2.12), equivalent to the constancy of global mean salinity, we need only consider the difference, \( S \), between \( S_2 \) and \( S_1 \). The difference of the salt conservation equations (2.6) and (2.7) gives an equation for \( S \):

\[
\dot{S}_2 - \dot{S}_1 = \dot{S} = 2H_S - 2|q|S \quad (2.13)
\]

or, using the flow law (2.9),

\[
\dot{S} = 2H_S - 2k|\alpha T - \beta S|S \quad (2.14)
\]

which completes the formulation of the model; its behaviour is completely characterised by (2.14).

### 2.1.2 Equilibrium solutions

As the first step in our analysis of (2.14), governing the evolution of the salinity difference between the low and high latitude boxes, we look for steady-state or equilibrium solutions, defined by a vanishing of the time derivative:
\[ H_S - k |\alpha T - \beta \bar{S}| \bar{S} = 0 \]  

where the overbar marks a steady-state quantity. We must consider separately the cases where the argument of the modulus is positive or negative.

**Case I:**

\[ \bar{q} > 0, \ \alpha T > \beta \bar{S} \]  

We can simply replace the modulus signs by brackets, giving

\[ H_S - k(\alpha T - \beta \bar{S})\bar{S} = 0, \]  

or

\[ (\beta \bar{S})^2 - (\beta \bar{S})(\alpha T) + \beta H_S/k = 0, \]  

which has the roots

\[ (\beta \bar{S})_{1/2} = (\alpha T) \left\{ \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{\beta H_S}{k(\alpha T)^2}} \right\} \]  

For a positive radicand, defined by

\[ \frac{\beta H_S}{k(\alpha T)^2} < \frac{1}{4} \]  

the model has two equilibrium solutions for poleward near-surface flow. These solutions can also be characterised as thermally dominated or, in the language of atmospheric science, “thermally direct” (meaning that rising motion occurs at the location of heating, and subsidence at the location of cooling). If the freshwater flux forcing exceeds the threshold defined by (2.20), no thermally-driven equilibrium exists.

**Case II:**

\[ \bar{q} < 0, \ \alpha T < \beta \bar{S} \]  

Now, we must insert a minus sign when replacing the modulus signs by brackets,
\[ H_s + k(\alpha T - \beta \bar{S})\bar{S} = 0 \] (2.22)

which gives

\[(\beta \bar{S})^2 - (\beta \bar{S})(\alpha T) - \beta H_s/k = 0, \] (2.23)

and the single root

\[(\beta \bar{S})_3 = (\alpha T) \left\{ \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\beta H_s}{k(\alpha T)^2}} \right\} \] (2.24)

Notice that we must discard the negative root; the radicand is greater than 1/4, so that the negative root would imply \( \bar{S} < 0 \), in contradiction to the condition (2.21). The solution (2.24) has equatorward near-surface flow and can be characterised as salinity dominated or “thermally indirect”. It exists for all (positive) values of the freshwater flux forcing.

Fig. 2.2 shows the equilibrium solutions as a function of the freshwater flux forcing. In summary, we find the remarkable result that this simplest non-trivial model of the THC, represented in steady state by the pair of quadratic equations, (2.15) and (2.22), has three steady state solutions, provided that the freshwater flux forcing is not too strong [cf.,(2.20)]. Two equilibria have \( \bar{q} > 0 \) (poleward surface flow); they are characterised by either a small salinity contrast and strong flow \((\beta \bar{S} < (1/2)\alpha T, \bar{q} > (1/2)k\alpha T)\), or by a large salinity contrast and weak flow \((\beta \bar{S} > (1/2)\alpha T, \bar{q} < (1/2)k\alpha T)\). These steady states exist only if \( \frac{\beta H_s}{k(\alpha T)^2} < 1/4 \). The model has one steady-state solution with \( \bar{q} < 0 \) (equatorward surface flow), characterised by a very large salinity contrast \((\beta \bar{S} > \alpha T, \bar{q} < 0)\). This solution always exists, and is the only one if \( \frac{\beta H_s}{k(\alpha T)^2} > 1/4 \).

What is the physical reason behind the vanishing of the thermally direct solution if \( \frac{\beta H_s}{k(\alpha T)^2} > 1/4 \)? Stronger surface salinity flux must by balanced by stronger salinity advection, \( qS \). This can be accomplished either by increasing the salinity difference, \( S \), between low and high latitudes, or by increasing the flow strength, \( q \). But increasing \( S \) has the dynamical consequence of weakening the flow — (2.9) expresses that \( q \) decreases linearly with \( S \). Obviously, the product, \( qS \), is zero for either \( S = 0 \) or \( q = 0 \) (the
Figure 2.2: Solution portrait of the box model in phase space. Dimensionless salinity difference is denoted $\delta \equiv \beta S/\alpha T$; dimensionless surface salinity flux is $E \equiv \beta H S / k (\alpha T)^2$. The curves mark the equilibrium solutions, $\delta(E)$, while the arrows show the tendencies in phase space. Notice the existence of three steady states for $E < 1/4$. 
latter implying $\beta \bar{S} = \alpha T$; $q S$ is positive for intermediate values and attains a maximum at $\beta \bar{S} = (1/2)\alpha T$ (see phase space diagram, Fig. 2.2). At this point, $q \bar{S} = 1/4 \beta^2 (\alpha T)^2$, which marks the critical freshwater flux forcing, that is, the strongest forcing that can be balanced by salinity advection through thermally direct flow. For even greater $H_S$, balance is impossible.

An even deeper question than the one starting the preceding paragraph is, what makes the multiple equilibria possible in the first place? Two crucial ingredients are required. First is the advective nonlinearity: The flow advecting salinity is itself influenced by salinity gradients, through density. Without this nonlinearity the model would have a unique solution (or none at all). But there is a second requirement, that of different coupling of temperature and salinity to the atmosphere. We assume that the atmosphere controls temperature but the salinity flux. Imagine, instead, two extreme cases of equal coupling:

i Temperature \textit{and} salinity prescribed: Then, density is prescribed as well, meaning that the flow prescribed. Trivially, no multiple equilibria are possible.

ii Heat \textit{and} freshwater flux prescribed:

Then, the surface density (or surface buoyancy) flux is prescribed and, hence, the steady-state horizontal density transport, $\frac{k}{\bar{\rho}} |\bar{\rho}| \bar{\rho}$. As $k$ and $|\bar{\rho}|$ are positive, the sign of $|\bar{\rho}|$ is uniquely determined by the sign of the surface buoyancy flux: If the low latitude box receives buoyancy from the atmosphere, it is less dense than the high latitude box, and $\bar{\rho}$ and $q$ are both positive (thermally direct circulation). The converse is true for prescribed buoyancy loss at low latitudes. Hence, the steady-state circulation is uniquely determined.

Exercise

1. Loss of multiple steady states: What steady-state solutions are possible in the 2-box model if the flow field is given as an external parameter (that is, depends neither on temperature nor on salinity)? Hint: Plot the salinity difference as a function of freshwater forcing, with $q$ given and constant.
2. Loss of multiple steady states: Prove the sequences i. and ii. outlined just above, by using the appropriate modifications of the equations for the Stommel box model, 2.2-2.7.

3. Loss of multiple steady states: Suppose that the surface heat and salt fluxes are formulated as restoring laws, as originally done by Stommel, i.e., the equations are

\[
\begin{align*}
\dot{T}_1 &= \lambda_T(T_1^* - T_1) + |q|(T_2 - T_1) \quad (2.25) \\
\dot{T}_2 &= \lambda_T(T_2^* - T_2) - |q|(T_2 - T_1) \quad (2.26) \\
\dot{S}_1 &= \lambda_S(S_1^* - S_1) + |q|(S_2 - S_1) \quad (2.27) \\
\dot{S}_2 &= \lambda_S(S_2^* - S_2) - |q|(S_2 - S_1) \quad (2.28)
\end{align*}
\]

where the starred quantities are the target values. Assume that \(\lambda_T = \lambda_S\) and construct a single ordinary differential equation for \(q\). What are the physically meaningful steady-state solutions now? What would change if \(\lambda_T \neq \lambda_S\)? N.B.: Do not solve the entire problem for \(\lambda_T \neq \lambda_S\).

2.1.3 Stability

We have identified three equilibria of the 2-box model of the THC in a certain parameter range. Now, we concern ourselves with the stability of the equilibria more precisely, with the “linear” stability. This means that we want to understand what happens if the equilibrium is perturbed by a tiny amount, either in the forcing, \(H_S\), or in the solution, \(S\). We will use a variety of techniques, each of which is important generally in the analysis of dynamical systems, and each of which illuminates one or several characteristics.

We start by investigating in more detail the equilibrium curves in phase space, Fig. 2.2. From the steady-state conditions, as expressed in eqs. 2.18 and 2.23, we obtain through a slight modification,

\[
\begin{align*}
\bar{q} > 0, \frac{\beta S}{\alpha T} &< 1 : \quad \frac{\beta H_S}{k(\alpha T)^2} = -\left(\frac{\beta S}{\alpha T}\right)^2 + \left(\frac{\beta S}{\alpha T}\right), \quad (2.29) \\
\bar{q} < 0, \frac{\beta S}{\alpha T} &> 1 : \quad \frac{\beta H_S}{k(\alpha T)^2} = +\left(\frac{\beta S}{\alpha T}\right)^2 - \left(\frac{\beta S}{\alpha T}\right), \quad (2.30)
\end{align*}
\]
which expresses the dimensionless salinity gradient, \( \delta \equiv \beta S/\alpha T \), as a function of the dimensionless surface salinity flux, \( E \equiv \beta H_S/k(\alpha T)^2 \). Thus, we can write 2.29 and 2.30 in dimensionless form as

\[
\delta \leq 1 : E = -\delta^2 + \delta = \delta(1 - \delta) \tag{2.31}
\]

\[
\delta \geq 1 : E = \delta^2 - \delta = \delta(\delta - 1) \tag{2.32}
\]

This pair of equations represents two sideways parabolas, with opposite orientation, intersecting at \( \delta \equiv \beta S/\alpha T = 0 \) (no salinity difference) and \( \delta = 1(\alpha T = \beta S; \text{no flow}) \). In either case, the forcing must vanish \( (E = 0) \). The curves depicted in Fig. 2.2 are the zeros of the salinity conservation equation 2.14, rewritten in dimensionless form as

\[
\frac{1}{2k\alpha T} \frac{d}{dt} \left( \frac{\beta S}{\alpha T} \right) = \frac{\beta H_S}{k(\alpha T)^2} - \left| 1 - \left( \frac{\beta S}{\alpha T} \right) \right| \left( \frac{\beta S}{\alpha T} \right). \tag{2.33}
\]

Notice that 2.33 implies an advective timescale, suitable for nondimensionalisation, of \((2k\alpha T)^{-1}\), and a nondimensional overturning strength of \( \tilde{q} = 1 - \delta \). We can thus rewrite 2.33 as

\[
\dot{\delta} = E - |1 - \delta|\delta \tag{2.34}
\]

**Exercise**

4. Prove the statement in the sentence following (2.33). Hint: Write \( t = \tilde{t}\tilde{\tilde{t}}, q = \tilde{q}\tilde{\tilde{q}} \) etc., where the caret denotes the scale and the tilde the non-dimensional quantity.

5. Find the steady-state solutions of 2.34, that is, perform the procedure leading to (2.19) and (2.24), but using non-dimensional quantities from the outset.

From either (2.33) or (2.34), we can read off the following. On the equilibrium curve, the tendency (time rate of change) of the salinity difference between high and low latitudes vanishes. But to the left of the curve, \( E \) or \( H_S \) is smaller than required by the equilibrium condition. Hence, \( \dot{S} < 0 \), and \( S \) decreases, as indicated by the downward pointing arrows in Fig. 2.2. In fact, the arrows were calculated from the right-hand sides of (2.34). To the
right of the curve, $E$ or $H_S$ is greater than required for equilibrium, hence $\dot{S} > 0$, and $S$ increases. Notice that for every given $\delta$ in Fig. 2.2, there belongs a unique $E$, so “left” and “right” of the equilibrium curve are defined unambiguously.

By visual inspection of Fig. 2.2, we can now read off the stability properties of the solutions. If, by any initial perturbation or change in forcing, we find ourselves to the left of the equilibrium curve, the evolution depends critically on which solution branch we started from. On the top ($\delta > 1$) and bottom ($\delta < 1/2$) branches in Fig. 2.2 (salinity dominated and thermally dominated-strong flow, respectively), the system moves downward, back towards the equilibrium curve. But if one starts from the middle branch ($1/2 < \delta < 1$), which runs from top-left to bottom-right in Fig. 2.2, the system does not return, but instead undergoes a transition towards the lower, thermally dominated branch. If the initial perturbation or change in forcing leaves the system to the right of the equilibrium curve, the system moves upward, again back towards the equilibrium curve, if it started from the top or the bottom branch. But if it started from the middle branch, it would make a transition toward the salinity-dominated equilibrium. Hence we conclude that the salinity-dominated steady state is always stable, the strong-flow thermally dominated steady state is stable (if it exists), while the weak-flow thermally dominated steady state is unstable to insimal perturbations. There exists a tell-tale sign allowing one to infer this instability even without investigating the full time-dependent equation. As one follows the unstable branch in Fig. 2.2 ($1/2 < \delta < 1$), from left to right, say, an increase in $E$ implies a decrease in $\delta$. Thus, an increase in forcing leads to a decrease in the steady-state response, which is, to my knowledge, an unfailing indication of instability.

Two points deserve special mention, since they are semistable, meaning that the system approaches them if it is on one side in phase space, but moves away from them if it is on the other side. These points are ($E = 0, \delta = 1$), where the two parabolas meet, and ($E = 1/4, \delta = 1/2$), the point beyond which no thermally direct steady state is possible. (In the language of dynamical systems, this is called a saddle node bifurcation.) Both these points show interesting mathematical behaviour, but they are not of great physical interest because this behaviour is not robust to small perturbations, such as a small amount of random noise.
2.1.3.1 Lyapunov potential

A powerful illustration of the stability properties discussed in the preceding paragraphs comes from a mathematical construct called the "Lyapunov potential". In loose analogy to, say, the relationship between gravitational force and gravitational potential, the time rate of change of dimensionless salinity, $\dot{\delta}$, (cf., 2.34), is written as the negative gradient of the Lyapunov potential, $L$, such that

$$-\frac{\partial L}{\partial \delta} = \dot{\delta} = E - |1 - \delta|\delta$$

(2.35)

By construction, the steady states of the system coincide with the extrema (maximum or minimum) of the Lyapunov potential. But we can say more: Plotting $L(\delta)$ immediately indicates the stability properties of the equilibria; indeed one can interpret the stability as if a bead was sliding on a wire under the influence of gravity: A minimum in $L$ is a stable equilibrium, while a maximum is an unstable equilibrium. We first illustrate this graphically, before showing it mathematically.

It is readily shown that

$$L = -E\delta - \frac{1}{3}\delta^3 + \frac{1}{2}\delta^2; \quad \delta \leq 1$$

(2.36a)

$$L = -E\delta + \frac{1}{3}\delta^3 - \frac{1}{2}\delta^2 + \frac{1}{3}; \quad \delta \geq 1$$

(2.36b)

fulfils 2.35, including the (arbitrary) condition of $L(0) = 0$ and the (non-arbitrary) condition of continuity at $\delta = 1$. Fig. 2.3 shows the Lyapunov potential, as a function of $\delta$, for a variety of choices for $E$. The case, $E = 0$, has one minimum at $\delta = 0$ and a double extremum (level turning point) at $\delta = 1$. The former is stable, according to Fig. 2.2, while the latter is semistable (approached from the right, moved away from on the left). Thus, we can visualise the evolution of the system as the inertia-less sliding of a bead on the “wire” $L(\delta)$. As $E$ is nonzero but less than $1/4$, the minimum at the left moves from zero to higher values, while another minimum appears for $\delta > 1$ and growing. Since $L(\delta)$ is continuous, the two minima must be separated by a maximum. In other words, two stable equilibria must have an unstable equilibrium between them.

As $E$ approaches $1/4$, the minimum at $\delta > 1$ becomes deeper than the one at $\delta < 1/2$, until, at $E = 1/4$, the two equilibria with $\delta < 1$ merge to
form a level turning point. This is the second semistable point discussed in Fig. 2.2. For even greater $E$, the thermally dominated ($\delta < 1$) equilibrium vanishes altogether, although its vicinity can still be felt through the very small time rates of change nearby.

After gaining an intuitive understanding of how to interpret $L$, we can now derive mathematically how its shape reflects stability properties. At any point, if $L$ increases with $\delta$, the left-hand side of (2.35) is negative, $\dot{\delta} < 0$, and $\delta$ decreases. In the $L(\delta)$ phase plot, Fig. 2.3, one slides toward the left. The converse is true if $L$ decreases with $\delta$. In the vicinity of a minimum, hence, any deviation to the right ($L$ increasing with $\delta$) is followed by motion to the left, back toward the minimum. Likewise, any deviation to the left will be followed by motion back to the minimum. Near a maximum, instead, a deviation to the right, say, means that $L$ decreases with $\delta$, the left-hand side of (2.35) is positive, $\dot{\delta} > 0$, and $\delta$ increases further, that is, the system moves further to the right, away from the equilibrium. For deviations to the left of a maximum, $\delta < 0$, and $\delta$ decreases further, again moving away from the equilibrium. Hence, if we can construct a Lyapunov potential as in (2.36), we can immediately read off the plot the stable and unstable steady states, in a completely intuitive manner.

Notice that in a case such as depicted in Fig. 2.3, sometimes the nomenclature is adopted to call the stable equilibrium with the shallower potential well “metastable”, reserving the term “stable” only for the steady state with the globally lowest potential. Here, we will largely only concern ourselves with distinguishing between stability and instability to infinitesimal perturbations.

**Exercise**

6. Prove that (2.36) is the correct Lyapunov potential for the system described by (2.34).

7. At what value of $\delta$ are the two minima in Fig. 2.3 equally deep [\(L(\delta)\) equal values]?

**2.1.3.2 Feedbacks**

We have found, from either the phase space plot, 10.2, or the Lyapunov potential, 10.3, how to characterise the multiple equilibria of the 2-box model as either stable or unstable. But what are the processes that lead to stability
Figure 2.3: Lyapunov potential as defined by 2.36, for a variety of choices for $E$. 
or instability? To this end, we now analyse the model equations in the vicinity of the steady states, employing a powerful technique applicable near any equilibrium state. The trick is to approximate the full, nonlinear equation through a linear one, such that the approximation ("linearisation") is good in the vicinity of the steady state (notice that one must linearise separately about every distinct equilibrium).

For this exercise, we return to the original, dimensional equation for the salinity difference between low and high latitudes, (2.13), and the flow law (2.9). We write all quantities as the sum of the steady-state value, marked again by an overbar, and a deviation thereof, marked by a prime, such that

\[ S = \bar{S} + S', \quad q = \bar{q} + q' \quad (2.37) \]

This separation is interesting in our case (or complicated, depending on taste), owing to the appearance of the modulus of \( q \) in the salinity advection. Care must be taken, and we again must distinguish between positive and negative \( q \):

\[
|q| = |\bar{q} + q'| = |k\alpha T - k\beta (\bar{S} + S')| \\
= \left[ k\alpha T - k\beta (\bar{S} + S') \right] = |\bar{q}| - k\beta S'; \quad \bar{q} > 0, \quad (2.38) \\
= \left[ k\beta (\bar{S} + S') - k\alpha T \right] = |\bar{q}| + k\beta S'; \quad \bar{q} < 0,
\]

where it has been used that

\[ q' = -k\beta S' \quad (2.39) \]

because \( T \) is an external parameter. The salinity conservation equation, (2.13), is now written, using the expansion (2.37),

\[
\dot{S} = \left( \dot{\bar{S}} + \dot{S}' \right) = \dot{S}' = 2H_S - 2|q| S = 2H_S - 2 (|\bar{q}| \mp k\beta S') (\bar{S} + S') \quad (2.40) \\
- : \bar{q} > 0; \quad + : \bar{q} < 0
\]

where we have used that the steady-state value does not change with time. We can subtract from (2.40) the steady-state condition, (2.15), leaving

\[
\dot{S}' = -2|\bar{q}| S' \pm 2k\beta S' (\bar{S} + S'); \quad + : \bar{q} > 0; \quad - : \bar{q} < 0. \quad (2.41)
\]
Notice that so far, we have not introduced any approximation yet, but merely rewritten the original equation in an inflated form. Now, however, we introduce the assumption that

$$|S'| \ll \bar{S}; \quad |q'| \ll \|\bar{q}\|,$$  \hspace{1cm} (2.42)

that is, the deviations from the equilibrium values are small compared to the equilibrium values themselves. In other words, we remain close to the steady-state. In this case, we can neglect the term containing the product of two perturbations quantities, leaving behind only terms that are linear in primed quantities (hence the term linearisation),

$$\dot{S}' = -2 |\bar{q}| S' \mp 2q' \bar{S}; \quad +: \bar{q} > 0; -: \bar{q} < 0,$$  \hspace{1cm} (2.43a)

or

$$\dot{S}' = -2 |\bar{q}| S' \pm 2k\beta S' \bar{S}; \quad +: \bar{q} > 0; -: \bar{q} < 0.$$  \hspace{1cm} (2.43b)

On this approximated equation, or any other obtained through this approach, we can launch the full power of systematic solutions of linear differential equations. We know that if the coefficient multiplying \( S' \) on the right-hand side is negative, the perturbation \( S' \) is exponentially damped toward zero; the system returns to the steady state, which hence is stable. In contrast, if the coefficient multiplying \( S' \) on the right-hand side is positive, the perturbation \( S' \) grows exponentially, the system does not return to its equilibrium, which hence is unstable. Notice that \( S' \) does not go to infinity; instead, as it grows too large, the assumption, (2.42), behind the linearisation breaks down, and one has to resort to the full nonlinear analysis.

Which processes determine whether the steady state is stable or unstable? We must analyse (2.43) to determine the contributors to the coefficient of \( S' \). Each of the terms represents a feedback, meaning a contribution to a tendency in \( S' \) that is caused by \( S' \) itself. The first term represents the advection of an anomaly in salinity difference by the time-mean flow, and can hence be called the mean flow feedback: Assume that, from whatever cause, \( S' > 0 \). The first term on the right-hand side of (2.43) contributes negatively, so that \( \dot{S}' < 0 \), so \( S' \) is reduced by this term. In other words, the mean flow feedback works against the original anomaly, hence stabilises the equilibrium - which is the definition of a negative feedback. It is readily shown that negative anomalies (original \( S' < 0 \)) are damped as well. Notice that the mean flow feedback works identically in the thermally dominated and haline dominated equilibria, though with different strengths.
The second term on the right-hand side of (2.43) represents the advection of mean salinity gradient by the perturbation flow, and can hence be called the salinity transport feedback. (Notice that advection of perturbation salinity gradient by perturbation flow is neglected in this linear approximation). The sign of the salinity transport feedback depends on the steady-state flow direction. If $\bar{q} > 0$, and $S' > 0$ (say), then $\dot{S}' > 0$, and the initial perturbation is further increased. Again, it is readily shown that this amplification is independent of the sign of the initial anomaly. If $\bar{q} > 0$, hence, the salinity transport feedback is a positive feedback, destabilising the equilibrium.

The situation is different for the haline dominated equilibrium, $\bar{q} < 0$. If $S' > 0$, then $\dot{S} < 0$, from the contribution by the second term on the right-hand side of (2.43), and the initial perturbation is reduced. The salinity transport feedback is a negative, stabilising feedback. In summary, we identify two negative feedbacks for the thermally indirect or haline dominated circulation, $\bar{q} < 0$. As all feedbacks are negative, this equilibrium is always stable to infinitesimal perturbations.

In contrast, the thermally direct circulation, $\bar{q} > 0$, has one positive feedback and one negative feedback. To determine the stability of the equilibrium, the relative strengths of the competing feedbacks must be evaluated. Using the dynamic flow law, (2.9), for $\bar{q}$ in the salinity perturbation equation (2.41), gives

$$\dot{S}' \equiv -2\bar{q}S' + 2k\beta S = -2k (\alpha T - 2\beta S) S'. \quad (2.44)$$

Hence, if $\beta \bar{S} < (1/2)\alpha T$, the coefficient multiplying $S'$ is negative, and the equilibrium is stable. In contrast, if $1/2\alpha T < \beta \bar{S} < \alpha T$, the coefficient multiplying $S'$ is positive, and the equilibrium is unstable. In the former case, the stabilising mean flow feedback dominates, whereas in the latter, the destabilising salinity transport feedback dominates.

**Exercise**

8. Complete the discussion of feedback loops for all cases and show that the sign of the feedback is independent of the sign of the initial anomaly.

### 2.1.4 Time-dependent solution

At the beginning of this lecture, I made a rather oblique remark concerning the exceptions to the statement that we can completely calculate the
solution to the simplified Stommel model. Of course, one can always invent forcing histories, such as $E(t)$ in the dimensionless salt conservation equation 2.34, that an analytical solution can only be given symbolically. But 2.34 permits the exact, and relatively simple, analytical solution of its full time-dependence. As of first writing these notes (January 2002), I am unaware of any published account of this solution. And since the solution provides a perspective that cannot be obtained from the previous approaches, it is given here.

With some help from Matlab’s Symbolic Math toolbox, one readily finds as the solution to 2.34:

$$\delta(t) = \frac{1}{2} - \sqrt{\frac{1}{4} - E} \tanh \left\{ t \sqrt{\frac{1}{4} - E} + \tanh \frac{\frac{1}{2} - \delta(0)}{\sqrt{\frac{1}{4} - E}} \right\}; \quad \delta \leq 1 \quad (2.45)$$

$$\delta(t) = \frac{1}{2} + \sqrt{\frac{1}{4} + E} \tanh \left\{ t \sqrt{\frac{1}{4} + E} + \tanh \frac{-\frac{1}{2} + \delta(0)}{\sqrt{\frac{1}{4} + E}} \right\}; \quad \delta \geq 1 \quad (2.46)$$

where atanh is the inverse of the hyperbolic tangent, tanh, and $\delta(0)$ is the initial condition. Using $\frac{d}{dx} \tanh x = 1 - \tanh^2 x$ and noticing that the derivative of the argument of the tanh gives an additional factor of $\sqrt{\frac{1}{4} - E}$, we obtain from 2.45 that

$$\dot{\delta}(t) = -\left( \frac{1}{4} - E \right) \left( 1 - \tanh^2 \{ \ldots \} \right) ; \quad \delta \leq 1 \quad (2.47)$$

The validity of (2.34) is then readily shown by substitution of $\delta(t)$ and $\delta^2(t)$. That (2.45) is valid for $t = 0$ is almost trivial.

**Exercise**

9. Prove that 2.45 and 2.46 are the correct solutions of 2.34.

In addition to showing mathematical validity, (2.45) and (2.46) offer other interesting aspects. The long-term behaviour is very simple; for large $t$, the first term dominates the argument of the tanh (the initial condition is forgotten), and since tanh approximates 1 for large argument, we recover the two stable equilibria,

$$t \to \infty : \quad \delta(t) \to \frac{1}{2} - \sqrt{\frac{1}{4} - E}; \quad E < \frac{1}{4} \quad (2.48)$$

$$t \to \infty : \quad \delta(t) \to \frac{1}{2} + \sqrt{\frac{1}{4} + E}; \quad E > \frac{1}{4} \quad (2.49)$$
Notice that there is no trace of the unstable equilibrium left in the time-dependent solution, reflecting the fact that time evolution is always away from the unstable steady state. [Writing $\delta(t) = \frac{1}{2} + \sqrt{\frac{1}{4} - E \tanh \{ \ldots \}}$ etc. in (2.45) would not fulfil 2.34 - try it!]. Notice, further, that (2.45) is perfectly valid even for $E > \frac{1}{4}$; indeed, using that $\tanh ix = i \tan x$ etc., indicates that if $E > \frac{1}{4}$, $\delta(t)$ grows until it becomes greater than one, and 2.46 must be used.

Fig. 2.4 shows evaluations of the full time-dependent solutions to the 2-box model, (2.45) and (2.46), as functions of initial conditions and time. Notice that, if the solutions crosses the $\delta(t) = 1$ threshold from below, at time $t_c$, use of (2.45) must be discontinued and (2.46) must be used instead, with initial condition $\delta(t_c) = 1$. The first row shows the solutions for $E = 0.2$. Three types of behaviour are discernible in Fig. 2.4a. Low and high initial conditions lead to rapid convergence to the stable thermally and haline dominated equilibria, respectively. Intermediate-size initial conditions mean that the solutions hover near the unstable equilibrium for a while, before departing from it and approaching one of the stable steady states. Fig. 2.4b illustrates this behaviour in a contour plot. Moving horizontally to the right indicates the solution changing in time as one crosses colour separations. For long times, the two stable equilibria fill out the entire phase space, as witnessed by the ever expanding areas of orange and blue. The transition between the two values becomes sharper as time progresses and indicates the ever shrinking region in phase space from where the system has not yet exited to one of the stable equilibria ("attractors"). The case, $E = 0.24$, close to the bifurcation point, shows this general behaviour in more pronounced form. (It is readily shown that the equilibria are $\delta = 0.4$, 0.6, and 1.2, which means that they fall on the boundaries between colours in the intervals chosen). Finally, if $E = 0.26$, and there is no thermally dominated equilibrium any more, some of the trajectories approach the (now unique) equilibrium quickly, while those starting from a small initial value hover near the (now vanished) steady state, its influence still there. But one by one, the trajectories undergo a rapid transition (Fig. (2.4)e). The transition region between red and blue colours is not horizontal any more, as it was for $E < 0.25$, indicating that sooner or later, all initial conditions lead to the haline dominated equilibrium Fig. 2.4f.

References
Figure 2.4: Solutions of 2-box model, as a function of dimensionless time and initial conditions. Left column: Time series of solutions. Right column: Contour plot of solutions.
2.2 Box Models: Interhemispheric Flow

2.2.1 Preliminaries

Arguably, the most glaring deficiency in Stommel’s box model of the THC is its confinement to a single hemisphere of a single ocean basin, ignoring that all oceans are connected. Conceptually, the simplest way of including a greater portion of the World Ocean is perhaps to use “back-to-back” Stommel models (Fig. 2.5), an approach first pursued by Welander (1986, Willebrand & Anderson, Eds., NATO ASI Series, C190, Kluwer, 163-200). In steady state, for every pair of boxes connected by pipes, the same considerations apply as to the original Stommel model, and there are two stable steady states for the flow between every pair of adjacent boxes. Hence, if the Atlantic is viewed as two back-to-back Stommel models, one finds 4 stable equilibria, one of which corresponds to the observed “Northern Sinking” solution with a pole-to-pole circulation. In principle, this idea can be further extended to include the Pacific Ocean as well, yielding $2^4 = 16$ stable equilibria for the 4 box pairs (Marotzke, 1990).

There is, however, a huge problem in interpreting the single Atlantic THC cell as two back-to-back Stommel models. The box model would require that the (surface) density in the Atlantic be a monotonic function of latitude – greatest in the north, intermediate at low latitudes, smallest in the south. This is in blatant contradiction to observations: Surface densities at

Figure 2.5: Geometry of Welander’s generalisation of Stommel’s model. Shown is the “northern sinking” pole-to-pole equilibrium solution.
both northern and southern high latitudes are much greater than around the equator, as indicated by the observed SST (Levitus SST). It appears worth its while to construct a box model that produces interhemispheric flow while maintaining the observed density minimum in the tropics. In discussing a prototype model of this kind, we encounter, along the way, an interesting example of how tortuous the path of progress in science can be.

Stommel’s model was first formulated in 1961 but the paper, celebrated as it may be today, went virtually unnoticed for 25 years (see Marotzke, 1994, for a historical account of how Stommel’s paper was received – or ignored! – by the community). Meanwhile, in 1982, another box model was independently proposed (Rooth, 1982; see Fig. 2.6), which explained how a 2-hemispheric THC symmetric about the equator might go unstable. This result inspired what is arguably the most influential study of the THC (Bryan, 1986), but faded out of the public eye owing to the “rediscovery” of Stommel (1961). Curiously, the profound difference between the dynamics in Stommel’s and Rooth’s models did not attract attention for a long time. As one consequence, it took more than 10 years before the model in Fig. 2.6 was extensively applied to the steady-state pole-to-pole circulation (Rahmstorf, 1996, Clim. Dyn. 12, 799-811; Scott et al., 1999) – which took up only one half-sentence in Rooth (1982). We will now consider the two applications of Rooth’s model – first to the instability of the equatorially symmetric state, then to the pole-to-pole steady state.
2.2.2 Rooth’s model: Formulation

The model is in its physical laws equivalent to Stommel’s, with the crucial difference that the flow is driven by the *pole-to-pole density difference*. In general, the equivalent surface salinity fluxes in the two hemispheres differ ($H_N$ and $H_S$ in northern and southern hemispheres, respectively). For simplicity, we assume that temperature is fixed and symmetric about the equator. Assuming flow directions as in Fig. 2.6, the equations are

\[
\begin{align*}
\dot{S}_1 &= -H_S + q(S_3 - S_1) \\
\dot{S}_2 &= H_S + H_N - q(S_2 - S_1) \\
\dot{S}_3 &= -H_N + q(S_2 - S_3) \\
q &= k'(\rho_3 - \rho_1) = k'\beta(S_3 - S_1)
\end{align*}
\]  

where $k'$ is a hydraulic constant that is different from the one used in Stommel’s model. Typical density differences between high latitudes are much smaller than the pole-equator density contrast, so $k'$ must be correspondingly larger to obtain the same flow strength.

2.2.3 Rooth’s model: Instability of the symmetric state

A symmetric state requires symmetric forcing, so we set $H_N = H_S = \phi$. Inspection of (2.50) through (2.53) shows that, starting from an isohaline state at $S_0$, the equations are solved by

\[
\begin{align*}
\bar{S}_1 &= \bar{S}_3 = -\phi t; \quad \bar{S}_2 = 2\phi t; \quad \bar{q} = 0.
\end{align*}
\]  

Here, we have arbitrarily set that the initial salinities are uniformly zero. Equation (2.54) describes a rather weird reference state, obviously not an equilibrium, but with vanishing flow at all times. We can now look at the linear perturbation expansion about this reference state, by writing

\[
\begin{align*}
q &= \bar{q} + q' = k'\beta(S'_3 - S'_1), \\
S_{1/3} &= \bar{S}_{1/3} + S'_{1/3} = -\phi t + S'_{1/3}, \\
S_2 &= \bar{S}_2 + S'_2 = 2\phi t + S'_2
\end{align*}
\]  

We take the time derivative of the dynamical equation (2.53),
\[ \dot{q} = k' \beta (\dot{S}_3 - \dot{S}_1) = k' \beta [-H_N + H_S + q(S_2 - S_3) - q(S_3 - S_1)], \quad (2.58) \]

use that the forcing is symmetric, and insert (2.54) through (2.57), to obtain to first order in primed quantities,

\[ \dot{q}' = 3k' \beta \phi t q'. \quad (2.59) \]

This perturbation expansion requires that the no-flow solution existed long enough so that a considerable salinity difference can build up between low and high latitudes. Any small perturbation away from the reference state grows according to

\[ q'(t) = q'(0) \exp \left\{ \frac{3}{2} k' \beta \phi t^2 \right\}, \quad (2.60) \]

so the no-flow solution is unconditionally unstable. (Thanks to Jeff Blundell for pointing out the solution.) The physical interpretation is as follows. A small salinity excess in the northern box leads to weak flow as indicated in the figure, which further increases salinity in the northern box – advecting high salinity, \( S_2 \), in and low salinity, \( S_3 \), out. In contrast, the salinity in the southern box does not change – waters advected in and out have the same salinity. As a result, the small initial excess in northern salinity over southern salinity is amplified – a positive feedback is at work. An asymmetric state develops despite the symmetric forcing – symmetry breaking.

2.2.4 Rooth’s model: Steady states and their stability

Rooth (1982) mentioned in passing that his model also had a steady state, with flow strength proportional to \( \sqrt{\phi} \). The steady-state aspects of Rooth’s model were, however, not considered until Rahmstorf (1996) noticed that the steady state is as readily found in the more general case of asymmetric forcing. Insertion of (2.53) into (2.50) shows that the steady state flow strength is

\[ \bar{q} = \sqrt{k' \beta H_S}. \quad (2.61) \]

This result has several remarkable properties. First, the steady-state THC increases with increased freshwater flux forcing, in stark contrast with
the single-hemispheric box model but consistent with at least some steady-state 3-dimensional model results, as long as one compares (2.61) against the Atlantic component of a global model (Wang et al., 1999, J. Climate 12, 71-82).

Second, and most remarkably, the Atlantic THC only depends on Southern Hemisphere atmospheric moisture flux (Rahmstorf, 1996). The other elements of the solution are

\[
\bar{S}_3 - \bar{S}_1 = \sqrt{H_S / k \beta}. \tag{2.62}
\]

Insertion of the solution (2.61) for the flow into the steady-state version of (2.52), the salinity conservation equation for box 3, gives

\[
\sqrt{k \beta H_S (\bar{S}_2 - \bar{S}_3)} = H_N \tag{2.63}
\]

or

\[
\bar{S}_2 - \bar{S}_3 = \frac{H_N}{\sqrt{k \beta H_S}}. \tag{2.64}
\]

In particular, we find that

\[
\frac{\bar{S}_2 - \bar{S}_3}{\bar{S}_3 - \bar{S}_1} = \frac{H_N}{H_S} \equiv \Gamma. \tag{2.65}
\]

The salinity difference between equator and the southern box 1 can be inferred from (2.62) and (2.64) as

\[
\bar{S}_2 - \bar{S}_1 = \bar{S}_2 - \bar{S}_3 + \bar{S}_3 - \bar{S}_1 \tag{2.66}
\]

\[
= \frac{H_N}{\sqrt{k \beta H_S}} + \sqrt{H_S / k \beta}
\]

\[
= \frac{(H_S + H_N)}{\sqrt{k \beta H_S}}.
\]

Under symmetric conditions, \(\bar{S}_2 - \bar{S}_1 = 2(\bar{S}_2 - \bar{S}_3) = 2(\bar{S}_3 - \bar{S}_1)\), the salinity drop is equal between, in turn, equator, northern box, and southern box.

That the northern sinking THC strength only depends on southern hemisphere salinity forcing can be understood by the following argument. Rooth’s model is essentially unidirectional pipe flow. At any given point (box), the difference between incoming and outgoing salinity is given by the ratio of salt
forcing over flow strength, as follows from the kinematic steady-state condition. Box 1 is then singled out because flow strength is proportional to the difference between box 1’s incoming and outgoing salinity. Both kinematic and dynamic equations for box 1 hence only include local properties (surface salt forcing, entering salinity $S_3$, outgoing salinity $S_1$).

Everything appears quite simple now: There is a single northern-sinking state; by symmetry, there is also a single southern-sinking state, the strength of which is determined solely by northern hemisphere moisture flux. But there is more to it than meets the eye, as discovered by a graduate student at MIT, Jeff Scott (Scott et al., 1999). In numerical solutions of a more complicated version of the Rooth box model, he did not always find the putative northern sinking solution. A linear perturbation expansion about the steady state (2.61) reveals why. Notice, first, that

$$S_2' = -S_1' - S_3'$$}

(2.67)

because global-mean salinity is a constant. Linear expansion of the perturbation equations for $S_1$ and $S_3$ gives, using (2.61), (2.62), (2.64), and (2.67),

$$\dot{S}_1' = q(S'_2 - S'_1) + q'(S_3 - \bar{S}_1)$$

(2.68)

$$= 2\sqrt{k'\beta H_S}(S'_3 - S'_1)$$

$$\dot{S}_3' = q'(S_2 - S_3) + \bar{q}(S'_2 - S'_3)$$

(2.69)

$$= k'\beta(S'_3 - S'_1)H_N/\sqrt{k'\beta H_S} - \sqrt{k'\beta H_S}(S'_1 + 2S'_3)$$

$$= \sqrt{k'\beta H_S} \left[ H_N/H_S (S'_3 - S'_1) - (S'_1 + 2S'_3) \right]$$

Equations (2.68) and (2.69) can be rewritten in matrix form as

$$\begin{pmatrix} \dot{S}_1' \\ \dot{S}_3' \end{pmatrix} = A \begin{pmatrix} S'_1 \\ S'_3 \end{pmatrix}$$

(2.70)

with

$$A \equiv \sqrt{k'\beta H_S} \begin{pmatrix} -2 & 2 \\ -\Gamma - 1 & \Gamma - 2 \end{pmatrix},$$

(2.71)
where $\Gamma$ is the ratio of northern to southern hemisphere salinity forcing, defined in (2.65). The stability of the steady state is determined by the sign of the real part of the eigenvalues of $A$, which are obtained from

\[
(\lambda + 2)(\lambda + 2 - \psi) + 2(\psi + 1) = \lambda^2 - (\psi - 4)\lambda + 6 = 0 \tag{2.72}
\]

\[
\lambda_{1/2} = \frac{1}{2}(\psi - 4) \pm \sqrt{\frac{1}{4}(\psi - 4)^2 - 6}. \tag{2.73}
\]

For $\psi$ less than about 9, the eigenvalues are complex, and their real part is negative if $\psi < 4$. For a $\Gamma$ of more than 4, the northern sinking solution is unstable to infinitesimal perturbations, and the only stable equilibrium is the southern sinking one. If relatively too much freshwater is dumped into the North Atlantic, the ”northern sinking” THC cannot be sustained. The sketch shows a phase diagram with the equilibrium solutions as a function of $H_N$, for given $H_S$.

The feedbacks present are most readily identified by using (2.68) and (2.69) to write

\[
\frac{q'}{k'\beta} = S'_3 - S'_1 = q'(\bar{S}_2 - \bar{S}_3) + \bar{q}(S'_2 - S'_3) - \bar{q}(S'_3 - S'_1) - q(\bar{S}_3 - \bar{S}_1). \tag{2.74}
\]

With (2.65) and (2.67), this gives

\[
\frac{q'}{k'\beta} = -3S'_3\bar{q} + q'(\bar{S}_3 - \bar{S}_1)(\Gamma - 1). \tag{2.75}
\]

The first term represents the ubiquitous ”mean flow feedback” (mean flow eliminates anomalies), the second term the feedback associated with anomalous flow. If $\Gamma > 1$, the coefficient multiplying $q'$ is positive, and the term contributes to exponential growth, so the salinity transport feedback is positive. The opposite applies if $\Gamma < 1$.

A positive flow perturbation will cause salinity perturbations proportional to $\bar{S}_2 - \bar{S}_3$ and $\bar{S}_3 - \bar{S}_1$ in boxes 3 and 1, respectively; both perturbations are positive. If the $S_3$ perturbation is smaller than the $S_1$ perturbation (resulting from $\Gamma < 1$), the flow is weakened, meaning that the salinity advection feedback is negative (stabilising). Both feedbacks are negative and the equilibrium is stable. Conversely, if $\Gamma > 1$, then $\bar{S}_2 - \bar{S}_3 > \bar{S}_3 - \bar{S}_1$, the salinity
Figure 2.7: Analytical flow solutions to the Rooth model as a function of $H_N, H_S = 0.9 \times 10^{-10} \text{psus}^{-1}$ (fixed). Temperature is symmetric about the equator. Black, red, blue, and green curves mark increasing horizontal diffusion (mimicking horizontal gyre transport) in the model, with black having no diffusion at all. Solid lines are stable solutions, dotted lines unstable; stability is determined analytically or by numerical integration. From Longworth, H., J. Marotzke, and T. F. Stocker, 2005: Ocean gyres and abrupt change in the thermohaline circulation: A conceptual analysis. Journal of Climate, 18, 2403-2416.

Advection feedback is destabilising; if $H_N > 4H_S$, the salinity advection feedback overcomes the negative mean flow feedback, and the steady state is unstable.

The strengths of Rooth’s box model are probably not recognisable except for experts. It looks rather pathetic, owing mainly to its exclusion of equatorial upwelling (in addition to being a box model); consequently, it has not been used much over the years except lending the initial inspiration to Frank Bryan’s experiments (making Rooth’s paper the truly seminal one on the multiple equilibria of the thermohaline circulation). Recently, however, a number of GCM phenomena have been explained with the model; in particular, the connection between interhemispheric flow and pole-to-pole
density differences has been firmly established. We will return to this point in Chapter 8 (THC Theory).
Chapter 3

THC Theory

3.1 Preliminaries

In many previous lectures, we have merrily used variants of Stommel’s box model of the THC, but we have persistently ducked the issue of why we could simply relate the THC strength to the meridional density difference. In other words, we have never explained what went into the definition of the "hydraulic" parameter, $k$. This avoidance has not been gratuitous, however. Rather, it reflects the difficulty of the problem, as was eloquently expressed by Alain Colin de Verdière (Colin de Verdière 1998) in his review of Joseph Pedlosky’s book "Ocean Circulation Theory": "The thermohaline circulation problem, on the other hand, requires the parallel computation of both density and velocity fields and is only briefly touched upon [in Pedlosky’s book]. Most recent advances on the latter topic motivated by the explosive interest in climate have come from numerical simulations and there are still many steps to be ascended on the stairway linking these numerical results and first principles."

One can add that thermocline theories, trying to explain the vertical structure of both density and velocity in the top kilometre of the ocean, share the "nonlinearity problem" with theories of the THC. What makes THC theory especially hard is that we are considering the superposition of two flow regimes, the poleward western boundary current and the equatorward interior gyre flow. The near-surface THC is the potentially small residual of large, compensating transports in these regimes. This lecture will give an introduction into THC theory, which is an active research area. Hence, we
will only cover a few of the important aspects, concentrating on the dynamics
(force balance) - in simplistic terms, on "what sets $k$?"

### 3.2 Early approaches: Scaling

Frank Bryan was the first to find multiple equilibria of the THC in a GCM and, in a separate paper (Bryan 1987), the first to confirm that the THC in a GCM was sensitive to the assumed degree of vertical mixing. He also presented a scaling argument for the THC strength, essentially applying to the meridional velocity an earlier derivation by Welander (1971) for the zonal flow. If the surface density increase from equator to pole, $\delta \rho$, is assumed given, one obtains from geostrophy and thermal wind,

$$fu = -\frac{1}{\rho_0} \partial_y p,$$

(3.1)

$$f \partial_z u = g/\rho_0 \partial_y \rho,$$

(3.2)

the scaling

$$f \frac{U}{D} = \frac{g}{\rho_0} \frac{\Delta \rho}{L},$$

(3.3)

where $U$, $D$, and $L$ are typical scales for zonal flow, thermocline depth, and meridional extent, respectively. Furthermore, we assume that vertical advective-diffusive balance determines thermocline depth (a nontrivial statement but perhaps defensible in the absence of wind forcing),

$$w \partial_z \rho = k_v \partial_{zz} \rho,$$

(3.4)

where $k_v$ is vertical diffusivity. We obtain for a scaling

$$\frac{W \Delta \rho}{D} = \frac{k_v \Delta \rho}{D^2},$$

(3.5)

or, for thermocline depth, $D$,

$$D = \frac{k_v}{W}.$$  

(3.6)

The last equation to be used is mass conservation in the form
\[ \partial_x u + \partial_y v + \partial_z w = 0, \tag{3.7} \]

which poses the greatest conceptual difficulties in scaling. We are interested in the zonally averaged flow, that is, a scaling for

\[ \partial_y \bar{v} + \partial_z \bar{w} = 0, \tag{3.8} \]

where the overbar marks zonal average. But (3.8) does not contain the zonal flow any more, for which we have an expression, (3.3), based on thermal wind. A simple relationship is obtained only if one assumes horizontal isotropy, that is, the scales of zonal and meridional flows are the same. This, however, seems a poor assumption, a priori, given the aforementioned compensation of meridional flows at the same depth. Frank Bryan clearly acknowledged that he assumed, without justification, that the zonally averaged meridional flow scaled as the zonal flow, which allowed him to scale (3.8) as

\[ \frac{U}{L} = \frac{W}{D}. \tag{3.9} \]

Equation (3.9) also assumes that zonal and meridional extents are comparable, and that \( D \) is the appropriate vertical scale for variations in flow as well as in stratification [cf., eq. (3.6)]. This procedure gives three scaling equations, (3.3), (3.6), and (3.9) for the three unknowns \( U, W, \) and \( D \), which can be solved (first insert \( D \) from (3.6)) to give

\[ W = \left( \frac{g \Delta \rho k_v^2}{f \rho_0 L^2} \right)^{1/3}, \tag{3.10} \]

\[ D = \left( \frac{f \rho_0 L^2 k_v}{g \Delta \rho} \right)^{1/3}, \tag{3.11} \]

\[ U = \left( \frac{g^2 \Delta \rho^2 k_v}{f^2 \rho_0^2 L} \right)^{1/3}, \tag{3.12} \]

which combined give the overturning scaling

\[ \Psi = UDL = WL^2 = \left( \frac{g \Delta \rho L^4 k_v^2}{f \rho_0} \right)^{1/3}. \tag{3.13} \]

In particular, the THC depends on the 2/3 power of vertical diffusivity, the 1/3 power of meridional density contrast, and the 4/3 power of linear
basin size (because the area over which mixing can act increases). With sensible numbers \((\Delta \rho/\rho_0 = 4 \times 10^{-3} \text{ and } k_v = 10^{-4} \text{m}^2\text{s}^{-1})\), this gives 13 Sv for \(\Psi - a\) remarkably good result, which might have been responsible for people looking no further. Later researchers displayed less candour than Frank Bryan, and used (3.3) directly for \(\bar{v}\), without even mentioning that meridional flow is not in thermal wind balance with the meridional density gradient. Sometimes (perhaps when forced by a reviewer?) a remark was inserted that the scaling (3.13) is not particularly well founded, but it nevertheless enjoys widespread popularity.

### 3.3 Two-dimensional Models

For a long time, the only community that grappled - albeit indirectly - with the theoretical issue of what provides the force balance of the THC was a fairly specialist group of modellers who wanted to construct two-dimensional (latitude-depth) models of the THC, mostly for computational efficiency. The first (and simplest) of these was the one by Marotzke et al. (1988) who argued as follows. If geostrophy plus some vertical friction is assumed, the zonally averaged momentum equations are

\[
-f \bar{v} = -\frac{P_E - P_W}{\rho_0 L} + A \partial_{zz} \bar{u}, 
\]

\[
f \bar{u} = -\frac{\partial_y \tilde{p}}{\rho_0} + A \partial_{zz} \bar{v},
\]

where \(A\) is a vertical viscosity and \(p_E\) and \(p_W\) are pressure at the eastern and western boundaries, respectively. The appearance of the boundary pressure terms presents a fundamental problem since the goal is to express everything in zonally averaged form.

The large-scale flow is nearly geostrophic, so the first two terms in each equation dominate. However, Marotzke et al. (1988) blatantly asserted that neglecting the Coriolis terms (formally assuming a nonrotating system) gave sensible results, albeit for reasons not understood. They used eq. (3.15) with \(f = 0\) to calculate the flow from the density and pressure distribution, assuming a very large \(A\) to give reasonable values. This procedure implies a simple, local and linear relationship between zonal and meridional pressure gradients, which is possible but by no means guaranteed.
Wright and Stocker (1991) constructed what is probably the best known, and arguably the most sophisticated, of all two-dimensional models. Subsequently, they devised elaborate procedures for the zonal closure (Wright et al. 1995; Wright et al. 1998). They used the results from 3-dimensional models to relate, with quite some success, zonal and meridional density differences (Fig. 3.3), and to determine the necessary coefficients. However, all attempts to justify, theoretically, the closures suffered from the difficulty that, just as Marotzke et al. (1988) had done, they never used the equation that actually determines $\bar{v}$ under three-dimensional, rotating dynamics, namely (3.14). Instead, they relied on ever more complicated versions of (3.15).

The same applies to the approach of Sakai and Peltier (1995); nevertheless, a somewhat acrimonious debate ensued (Wright et al. 1998). Warren (1994) presented a closure that started from Stommel-Arons theory, but he, too, asserted that the THC had to be associated with non-geostrophic (frictional) effects.

### 3.4 Boundary-layer Approaches

While it is not generally accepted in the community, I am convinced that the traditional two-dimensional closures represent a conceptual dead end, because they do not deal with the thermal-wind balance of the meridional flow. (To what extent they reproduce the parameter sensitivity of 3-D models still remains to be seen, on a number of important points). Instead, one should explicitly investigate the pressure and density distributions along the eastern and western boundaries, and calculate $\bar{v}$ from those.

Qualitatively, the argument goes as follows (Zhang et al. 1992; Colin de Verdière 1993). Surface waters have high density at high latitudes and low density at low latitudes. Consequently, sea level is low at high latitudes and high at low latitudes (Fig. 3.4). The resulting surface circulation is eastward. This causes a pile-up of water at the eastern boundary (a secondary high, marked $H'$ in Fig. 3.4), and moreover downwelling. At the western boundary, sea level would be low (marked $L'$), and upwelling prevails. Between $H'$ and $L'$, northward geostrophic flow ensues. A zonal section of the conceptual set-up is shown in Fig. 3.4, taken from Colin de Verdière (1993). Superimposed on the zonal overturning implied by Fig. 3.4 are the circula-
Figure 3.1: Zonally integrated flow, in $m^2 s^{-1}$, as implied by the 2-D closure (abscissa), against the flow in a GCM (ordinate). From Wright et al. (1995).
Figure 3.2: Conceptual picture of the set-up of zonally integrated, buoyancy-driven flow.

Five fundamental assumptions are made, in addition to the standard approximations (hydrostatic and geostrophic balance):

i) Surface density is given and is a function of latitude only; the abyss uniformly has the properties of the densest surface water.

ii) The western boundary water is assumed to be stably stratified, following an exponential with scale height $D$ (to be determined as part of the solution).

iii) Density in the grid cells at the lateral boundary is governed by vertical advective-diffusive balance, 3.4, except where convection is present, which then also enters the balance.

iv) Since there is no wind stress in this model, no zonal pressure gradient can be supported at the equator; in other words, isopycnals are level along
v) Along the eastern boundary, convection occurs down to a depth $z$ (to be determined as part of the solution), which is a function of latitude. In other words, the isopycnal = const. is vertical at its outcrop latitude. Equatorward, it is assumed level; likewise, it is assumed that Rossby wave activity has eliminated all zonal isopycnal slopes except in the western boundary current.

Assumption (i) was used before by Welander (1971) and Bryan (1987); (ii) and (iii) are standard assumptions, which underlie the Bryan (1987) scaling and indirectly the Stommel-Arons picture. Assumption (iv) is a corollary of the force balance between wind stress and thermocline slope, traditionally assumed in equatorial oceanography. Assumption (v) is probably the most unorthodox; it is based on the physical picture that warm water generally moves to the northeast; subsurface advection of a certain density can only occur until the outcrop latitude is reached. That the isopycnals should be level equatorward of the outcrop latitude could be caused by Kelvin waves, but this is neither strictly required nor indeed fully confirmed by numerical experiments. Nevertheless, this ”stacked boxes” concept (Jeff Scott; see Fig. 3.4) allows one to calculate the meridional overturning circulation, once the basic stratification parameter - thermocline depth, $D$ - is known.
Figure 3.4: "Stacked boxes" underlying the THC theory. From Marotzke (1997).
The logic is as follows. Assume that $D$ is known. This implies that the density is known all along the western boundary. In particular, the depth of any isopycnal at the equator is known. As there is zero isopycnal slope along the equator, this depth, $z_\rho$, is known for all longitudes. It is also known along the eastern boundary, all the way northward to the outcrop latitude. Hence, density along the eastern wall is known, and one can calculate the east-west density difference. This allows us to calculate the overturning, assuming a sensible reference level (not a trivial assumption). Overall, this logical sequence allows us to calculate the flow, including $w$, given $D$. But there is a second relationship between $w$ and $D$, based on the advective-diffusive balance, and both can be determined. It turns out that in this way, we can determine not only scales for $w$ and $D$, but the complete dependence of overturning strength on latitude (Fig. 3.4). Indeed, the theory matches the numerical solution reasonably well.
Figure 3.5: Theory vs. numerical experiment. From Marotzke (1997)
Chapter 4

Climate Variability: Stochastic Climate Models

4.1 Preliminaries

Up to now, we have considered either steady-state models or the response to a steady (or slow) perturbation. Now we look at climate variability; true to our philosophy, we identify the, arguably, simplest theoretical model of climate variability. Hasselmann (1976) introduced the concept of stochastic climate models. He postulated that one could distinguish between two timescales in the climate system, one defining weather (here, the atmosphere; rapid fluctuations), the other defining climate (here, the ocean; slow response). Instead of treating the actual nonlinear interactions between frequency regimes explicitly, we consider “weather” as a \textit{stochastic forcing of climate}.

If, symbolically, $x$ denotes weather and $y$ denotes climate, we can write

$$
\frac{d}{dt} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} N_x(x, y) \\ N_y(x, y) \end{pmatrix}.
$$

(4.1)

Because of nonlinearities, $N_y \neq N_y(y)$, that is, climate evolution is not solely governed by the climate state but depends on weather. Consequently, climate is not a dynamical system, as stressed by Ed Lorenz (pers. comm.). But here we are interested only in climate. Instead of trying a closure (parameterisation of weather effects on climate purely in terms of climate variables), we write
Figure 4.1: Numerical realization of the “Game of Peter and Paul” (Coin tossing). Left: 50,000 tosses, right: 500,000 tosses, with the first 50,000 tosses the same as on the left.

\[
\frac{dy}{dt} = -\alpha y + f, \tag{4.2}
\]

where \(f\) is white noise and where we have assumed the simplest possible “climate dynamics” (linear damping). Concerning weather-climate interactions, we have replaced the nonlinear, chaotic description with a linear, stochastic one. This replacement is also known as a Langevin model or a fluctuation-dissipation approach.

### 4.2 Pedagogical example: The Game of Peter and Paul (after Wunsch (1992)).

A true coin is tossed; if the coin shows heads, Peter pays Paul 1\$; if it shows tails, Paul pays Peter. Figure 4.1 shows Paul’s net earnings as a function of time, first for 50,000, then for 500,000 tosses (executed on the computer, of course).

Counter to expectation, which would have the net earnings hover around zero for most of the time, there are large fluctuations away from the break-even point, and one player or the other is far ahead for most of the time. This result may well have a bearing on the real world; the observations by Schott et al. (1988) of mass transport through the Florida Strait (Fig. 4.2) show some similarity to the coin tossing time series.
Figure 4.2: Time series of Florida Strait transport, after Schott et al. 1988.

What is going on? Why does our intuition fail us so profoundly in predicting the outcome of “Peter and Paul”? First, the game shown in Fig. 4.1 is indeed a fair one; Paul is ahead for much of the time not because of a flaw in the execution of the game but because of an inevitable (as we will show) random succession of tosses in his favour. To gain insight, let us first consider a slightly more general case. Write

\[ y(t) = \sum_{t' = 1}^{t'} \xi(t'), \tag{4.3} \]

where \( t' \) takes integer values and \( \xi \) is a random process with vanishing ensemble mean, \( \langle \xi(t) \rangle = 0 \). Notice that the ensemble mean is taken over all possible outcomes, at any given time. Moreover, \( \xi \) is “white” (uncorrelated in time) and has variance \( \sigma^2 \), that is, \( \langle \xi(t)\xi(t') \rangle = \sigma^2 \delta_{tt} \). Here, \( \delta_{tt} \) is the Kronecker delta, with the value of zero if \( t \neq t' \) and the value of one if \( t = t' \).

Equation (4.3) reflects the summation of a sequence of random events; in the special case of \( \xi = 1 \) or \( \xi = -1 \), we recover “Peter and Paul”, with total number of coin tosses of \( t \). Writing (4.3) once again but for \( t - 1 \) tosses
shows that $y(t)$ is the solution of the very simple difference equation,

$$y(t) = y(t-1) + \xi(t), \quad (4.4)$$

the special case ($a = 1$) of a “first-order autoregressive process”, abbreviated AR(1):

$$y(t) = ay(t-1) + \xi(t); \quad 0 \leq a \leq 1. \quad (4.5)$$

For simplicity (and to be fair to both Peter and Paul), we assume that $y(0) = 0$. We see that, if $a > 0$ (that is, the state at a previous timestep matters) $y(t)$ has a “memory”.

Since

$$y(t - 1) = ay(t - 2) + \xi(t - 1), \quad (4.6)$$

we can guess that

$$y(t) = a^2 y(t - 2) + a\xi(t - 1) + \xi(t), \quad (4.7)$$

we can guess that

$$y(t) = \sum_{t'=1}^{t} a^{t-t'}\xi(t') = \sum_{t''=0}^{t-1} a^{t''} \xi(t - t''). \quad (4.8)$$

This shows that “old” $\xi(t)$ are eventually “forgotten”, if $a < 1$.

[Proof of (4.8): Insert into difference equations

$$y(t) - ay(t - 1) = \sum_{t'=1}^{t} a^{t-t'}\xi(t') - a\sum_{t'=1}^{t-1} a^{t-1-t'}\xi(t')$$

$$= \sum_{t'=1}^{t} a^{t-t'}\xi(t') - \sum_{t'=1}^{t-1} a^{t-t'}\xi(t') = \xi(t). \quad (4.9)$$

q.e.d.]

We now consider some important statistical properties of $y(t)$. First, the mean (expectation value) of any player’s (here: Paul’s) earnings is given as

$$\langle y(t) \rangle = \sum_{t'=1}^{t=t} a^{t-t'} \langle \xi(t') \rangle = 0, \quad \text{vanishing mean.} \quad (4.10)$$
Figure 4.3: Ensemble of 20 realisations of “Peter and Paul”, over 2000 tosses each.

This shows that indeed no one player has any advantage in the game, although in the particular outcome shown in Fig. 4.1 Paul seemed so privileged. Figure 4.3 confirms this point empirically. Shown are 20 realisations over 2000 tosses each. Sometimes Paul comes out ahead in the end, sometimes Peter does; over many games, no player wins. Figure 4.4 shows the average over the empirical ensemble of Fig. 4.3; over 20 realisations, no player is ahead by more than about $3, a trivial amount. However, if we are asking by how far one player or the other is ahead, the situation is different. Figure 4.4 shows the variance of Paul’s earnings; the variance counts positive and negative deviations from the expected value equally. It appears as if the square of Paul’s deviation from the break-even point grows by an amount roughly proportional to the number of tosses, consistent with the ever wider spread of the curves in Fig. 4.3.
Figure 4.4: Ensemble mean over the 20 realisations of “Peter and Paul” of Fig. 4.3.

Figure 4.5: Ensemble variance over the 20 realisations of “Peter and Paul” of Fig. 4.3.
Theoretically, the variance comes out as

\[
\langle y(t)y(t) \rangle = \left\langle \sum_{t'=t}^{t''=t} a^{t-t'} \xi(t') \sum_{t'''=1}^{t''=t} a^{t-t'''} \xi(t''') \right\rangle
\]

\[
= \sum_{t'=1}^{t''=t} \sum_{t'''=1} \left( a^{t-t'} \sigma^2 \delta_{t'''} \right) \xi(t') \xi(t''')
\]

\[
= \sum_{t'=1}^{t''=t} \sum_{t'''=1} \left( a^{t-t'} \left( \frac{\sigma^2}{\delta_{t'''}} \right) \right) \xi(t') \xi(t''')
\]

\[
= \sum_{t'=1} \left( a^{t-t} \sigma^2 \right) = \sum_{\tau=0}^{\tau=t-1} a^{2\tau}
\]

\[
= \left\{ \begin{array}{ll}
\frac{\sigma^2}{1-a^2} & \text{as } a < 1 \\
\frac{\sigma^2}{1-a} & \text{as } a = 1
\end{array} \right.
\]

For perfect memory, \( a = 1 \), as in the Game of Peter and Paul (no one loses money except to the other), the variance grows out of bounds, as \( t \) goes to infinity. Since \( y(t) \) represents Paul’s net earnings, this means that the amount by which one player leads, is expected to grow as the square root of time. As they both play with finite resources, sooner or later one must go bankrupt. This result is mathematically equivalent to the analysis of random walk first performed by G.I. Taylor in 1921.

If \( a < 1 \), meaning there is less than perfect memory, the variance of \( y(t) \) about its mean, zero, goes toward a constant, which is larger as \( a \) is closer to unity. The damping checks the growth of the variance. Still, the variance of \( y \) is greater than that of \( \xi \), owing to the finite memory. It is thus not true that the presence of a “slow” component reduces variability, on the contrary, it increases variability.

As an aside, it does not help to average over \( y(t) \) to get rid of the unbounded growth of variance. Define the sample mean over \( N \) timesteps,

\[
\bar{y}_N \equiv \frac{1}{N} \sum_{t=0}^{N} y(t),
\]

the variance of which about the mean can be shown to grow linearly with \( N \):
Exercise

1. Show that for the AR(1) process defined by (4.5), nonvanishing memory \((a > 0)\) means that the solution \(y(t)\) is correlated in time.

2. Prove (4.13). N.B.: Wunsch (1992) has the factor \(N/2\), which is not borne out by checks by countless students.

Now, we turn to the low-frequency variability of \(y(t)\), clearly shown by the power spectrum, which is “red” (more power at low frequencies, Fig. 4.6), typical of geophysical spectra. The power spectral density is proportional to \(\omega^{-2}\), as can be shown by first rewriting the difference equation (4.5) as

\[
y(t) - y(t - 1) = -(1 - a)y(t - 1) + \xi(t)
\]  

\[
\langle y_N^2 \rangle \xrightarrow{N \to \infty} \sigma^2 \frac{N}{3}.
\]  

Figure 4.6: Power spectral density of the 500,000-tosses realization of “Peter and Paul” of Fig. 4.1.
which plausibly approximates the differential equation

\[ \dot{y}(t) = - (1 - a) y(t) + \xi(t) \equiv - \kappa y(t) + \xi(t), \quad (4.15) \]

provided that the damping parameter, \( \kappa \), is much less than unity (damping small over one timestep). Notice that in a continuum formulation, noise can never be truly white, but we will ignore this complication. We will take the Fourier transform, defined by

\[ y(t) = \int_{-\infty}^{\infty} \hat{y}(\omega)e^{-i\omega t}d\omega. \quad (4.16) \]

The Fourier transform of white noise (the Dirac delta-function) is a constant, \( c \). The Fourier transform of (4.15) is

\[ -i\omega \hat{y}(\omega) = -\kappa \hat{y}(\omega) + c, \quad (4.17) \]

yielding

\[ \hat{y}(\omega) = \frac{c}{\kappa - i\omega}. \quad (4.18) \]

We are interested in the real part of the solution; most conveniently we calculate the square of the modulus,

\[ |\hat{y}(\omega)|^2 = \frac{c^2}{\kappa^2 + \omega^2} \]

\[ \begin{cases} \kappa \to 0 & c^2 \over \omega^2 \\ \omega \to 0 & c^2 \over \kappa^2 \\ \omega^2 \gg \kappa^2 & c^2 \over \omega^2 \end{cases} \quad (4.19) \]

The completely undamped system has a power spectrum proportional to \( \omega^{-2} \), again confirmed by the numerical example from the coin tossing (Fig. 4.6). With finite damping, the high-frequency part of the spectrum falls off as \( \omega^{-2} \), while the low-frequency part tends toward a constant, which is dependent on the damping.

We will not go into further detail, but notice that the appropriate “null-hypothesis” for explaining low-frequency variability in an oceanographic record (e.g., Rockall Channel time series, Fig. 4.7, which is Fig. 12 from Dickson et al. (1988) is forcing by quasi-random atmospheric perturbations. It might be vain to search for a deterministic cause. This lesson is often
Figure 4.7: Time series, 1972 - 1982, of near-surface salinity in the Rockall Channel, after Dickson et al. 1988

hard to swallow, because it is sometimes perceived that identification of a specific cause represents a higher level of understanding. But Figs. 4.8 and 4.9, which are Figs. 5 and 6 of Mikolajewicz and Maier-Reimer (1990), clearly show that nontrivial large-scale behaviour can be caused by random surface forcing. Figure 4.9 also shows examples of the theoretical spectrum, (4.19), for various choices of damping. The stochastic theory of climate of Hasselmann (1976) has thus proven to be a very important paradigm for explaining climate variability.
Figure 4.8: From Mikolajewicz and Maier-Reimer (1990)

Annual mean time series of net freshwater flux in the Southern Ocean (south of 30°S), the mass transport through Drake Passage, the atmosphere-ocean heat exchange (heat loss of the ocean gives negative values), in the Southern Ocean and in the North Atlantic (Atlantic north of 30° N and Arctic) and the outflow of NADW at 30° S from the Atlantic to the Southern Ocean, negative values indicating flow to the south. The vertical lines represent the reference times of the defined events used for the composites.
Figure 4.9: From Mikolajewicz and Maier-Reimer (1990)

Variance spectrum of the mass transport through Drake Passage and the net freshwater flux of the Southern Ocean (dashed line). Data are scaled by their standard deviation. The thin lines represent the results to be expected from simple linear stochastic climate models with time constants of 50 years, 500 years and infinite. For periods less than 30 years the shape of the spectrum can be approximated by a stochastic climate model, whereas for longer periods the spectral shape is significantly different due to the internal dynamics of the ocean. The bar indicates the 95% confidence interval.
Bibliography


