Stochastic stability of open-ocean deep convection

Till Kuhlbrodt

kuhlbrodt@pik-potsdam.de

Institute of Physics, University of Potsdam

Am Neuen Palais 10, 14469 Potsdam, Germany

and

Potsdam Institute for Climate Impact Research

P.O.Box 60 12 03, 14412 Potsdam, Germany

Adam Hugh Monahan

monahana@uvic.ca

School of Earth and Ocean Sciences, University of Victoria

P.O.Box 3055, STN CSC, Victoria BC, V8P 5C2, Canada

and

Canadian Institute for Advanced Research, Earth System Evolution Program

In press, Journal of Physical Oceanography

May 30, 2003

Abstract

Open-ocean deep convection is a highly variable and strongly nonlinear process that plays an essential role in the global ocean circulation. A new view of its stability is presented here, in which variability, as parameterised by stochastic forcing, is central. The use of an idealised deep convection box model allows analytical solutions and straightforward conceptual understanding, while retaining the main features of deep convection dynamics. In contrast to the generally abrupt stability changes in deterministic systems, measures of stochastic stability change smoothly in response to varying forcing parameters. These stochastic stability measures depend chiefly on the residence times of the system in different regions of phase space, which need not contain a stable steady state in the deterministic sense. Deep convection can occur frequently even for parameter ranges in which it is deterministically unstable; this effect is denoted wandering unimodality. The stochastic stability concepts are readily applied to other components of the climate system. The results highlight the need to take climate variability into account when analysing the stability of a climate state.

1 Introduction

1.1 Climate and stability

In our changing climate, the sensitivity of oceanic and atmospheric circulations to perturbations is of vital interest. Under the assumption that a circulation pattern is a steady state, classical stability theory is applicable and gives information about its sensitivity. Climate, however, is variable on a broad range of time scales. In most cases, a climate "state" can only be defined through an average in space and time over the observed circulation patterns. For instance, the deep water formation in the North Atlantic, which is often characterised as a stable state of the present climate (Manabe and Stouffer, 1999; Rahmstorf,

2000), arises as an average over many different processes, including intermittent deep convection events. Moreover, in a nonlinear system several stable states may coexist, such that variability leads to transitions between them. Climate change, then, can appear as a shift in the preference of different states, or climate regimes (Houghton et al., 2001; Palmer, 1999). A change of this kind can only be understood if we consider not only the mean state, but also the variability around it.

Our aim in this paper is to study how including environmental variability, parameterised as a stochastic process, changes the stability of deep convection in a conceptual box model. This leads to a discussion of general measures of stochastic stability. While the model is developed to characterise deep convection in the North Atlantic, its simplicity allows a wider application of the results.

1.2 Deep convection in the North Atlantic

As a consequence of various forcing processes (e.g. freshwater fluxes or cyclonic wind forcing) the vertical stratification of the North Atlantic is particularly weak in two small regions, one in the Labrador Sea and the other in the Greenland Sea. In these regions, strong surface cooling in winter, along with wind forcing, can lead to a vanishing vertical density gradient. This starts a vigorous vertical mixing process. Such a deep convection event occurs in patches of about 100 km diameter, extends to depths of 2 to 3 km, and lasts for a couple of days (see Marshall and Schott (1999) for a comprehensive review, and the 2002 JPO issue 32 (2) dedicated to deep convection in the Labrador Sea).

The cold and dense water masses formed by deep convection eventually sink to depth and flow southwards, feeding the deep branch of the thermohaline circulation (THC). In return, warmer waters flow northwards at the surface. This relative heat transport achieved by the THC is an important contribution to global-scale meridional transports of heat (Ganachaud and Wunsch, 2000), and in particular to Europe's mild climate (Manabe and Stouffer, 1988).

Observational data show considerable variability in the occurrence and depth of deep convection events. Especially in the Labrador Sea, the so-called Great Salinity Anomalies (GSA) are well-documented (Dickson et al., 1988; Belkin et al., 1998). During these events, the upper layer becomes anomalously fresh, thus enhancing the vertical density gradient and suppressing deep convection. Lazier's (1980) data, spanning the years 1964 – 1974, explicitly show the absence of deep convection during a GSA in 1968 – 1972.

Obviously deep convection is sensitive to the forcing of the upper layer of the ocean (Dickson et al., 1996; Lilly et al., 1999). Recent studies suggest that the ocean's upper layer, through a positive salinity feedback, responds actively to anomalies in the forcing (Houghton and Visbeck, 2002). Simulations from coupled climate models indicate that a long-term shutdown of deep convection may lead to a reduced heat transport by the THC, inducing climatic changes in the North Atlantic region (Wood et al., 1999; Hall and Stouffer, 2001; Schaeffer et al., 2002). It is our aim to contribute to a deeper understanding of this sensitivity in order to assess possible future circulation changes.

1.3 Box models of deep convection

Welander (1982) introduced a simple box model to study how slowly changing horizontal fluxes caused by eddy diffusion and advection interact with short and vigorous convective mixing events. Since then, many studies have applied and extended his model (e.g. Lenderink and Haarsma (1994, 1996); Cessi (1996); Hirschi et al. (1999)). Recently, Rahmstorf (2001) and Kuhlbrodt et al. (2001) (hereafter cited as K01) extended Welander's model to a system of four variables, considering the coupled salinity and temperature dynamics of two interacting boxes, one representing a well-mixed surface layer and the other the deep waters below. The model, referred to as the 2TS model, consists of four equations:

$$\frac{dT_1}{dt} = \frac{1}{h^* \tau_c(\Delta\rho)}(T_2 - T_1) + \frac{1}{\tau_{1T}}\left(T_1^* - A_T \cos(2\pi t) - T_1\right) \qquad (1)$$

$$\frac{dS_1}{dt} = \frac{1}{h^*\tau_c(\Delta\rho)}(S_2 - S_1) + \frac{1}{\tau_{1S}}\left(S_1^* + A_S\cos(2\pi t + \psi) - S_1\right) \quad (2)$$

$$\frac{dT_2}{dt} = \frac{1}{\tau_c(\Delta\rho)}(T_1 - T_2) + \frac{1}{\tau_2}(T_2^* - T_2) \quad (3)$$

$$\frac{dS_2}{dt} = \frac{1}{\tau_c(\Delta\rho)}(S_1 - S_2) + \frac{1}{\tau_2}(S_2^* - S_2)\,. \quad (4)$$

The box depths are assumed to have a constant ratio h^*. The variables T_1 and T_2 represent respectively the temperatures of the upper and lower boxes; similarly, S_1 and S_2 are the salinities of these boxes. These variables are relaxed towards prescribed relaxation temperatures and salinities T_1^*, S_1^*, T_2^*, S_2^*, parameterising various processes for the individual variables. The relaxation processes in the upper box represent lateral heat and salt exchanges with surrounding waters, mostly by eddy mixing. As well, the upper box temperature T_1 is coupled strongly to the atmosphere through surface heat fluxes, while the upper box salinity S_1 evolves without a feedback to the atmosphere. This fundamental difference is accounted for by using two different relaxation time scales: τ_{1T} and τ_{1S}. The deep box temperature T_2 and salinity S_2 are assumed to be determined by eddy transfer fluxes at depth, motivating a common restoring time scale τ_2. Observations show strong seasonal cycles in the mixed layer variables; these cycles are captured in the model through annually-varying forcings with amplitudes A_T and A_S and a phase shift ψ. Note that the time variable has been scaled in equations (1)-(4) so that one time unit corresponds to one year. The vertical exchange time scale τ_c is a function of the vertical density difference

$$\Delta\rho = \rho_1 - \rho_2 = -\alpha(T_1 - T_2) + \beta(S_1 - S_2)\,, \quad (5)$$

where α and β are the thermal and haline expansion coefficients of the linearised equation of state for seawater. The vertical exchange time scale is very large for $\Delta\rho \leq 0$, but for $\Delta\rho > 0$ convective mixing starts, and τ_c is of the order of a few days.

It was demonstrated in K01 that if the model parameters are chosen to represent the conditions in the Labrador Sea, the pronounced nonlinearity of

the vertical mixing leads to a *bistability* of deep convection. The model then has two stable steady states, one without any deep convection ("off") and one with deep convection occurring regularly ("on").

1.4 Fluctuating forcing

What impact do forcing anomalies have on the bistability of deep convection? Single short-lived anomalies can induce transitions between the "on" and the "off" states. In particular, due to the positive salinity feedback, the vertical density gradient is strengthened during a nonconvecting phase, making it increasingly harder to interrupt such a phase the longer it lasts (K01, see also Houghton and Visbeck (2002)).

The ubiquitous presence of variability in the atmospheric forcing can be parameterised by stochastic processes (Hasselmann (1976); see Imkeller and Monahan (2002) for a recent review). In the 2TS model, stochastic forcing, parameterising heat flux variability, leads to frequent jumps between the neighbourhoods of the two model states (K01). The jump frequency, conveniently measured by the mean residence time in a neighbourhood, is a smooth function of the model parameters. In contrast, the deterministic stability of a state disappears abruptly if a model parameter is moved beyond a bifurcation point. The need for stochastic stability concepts to address these facts was suggested in K01.

The theory of stochastic stability, although well developed (Freidlin and Wentzell, 1998), has not often been used in ocean dynamics. Cessi (1994) computed the mean residence times to characterise the variability in a simple box model of the THC. Timmermann and Lohmann (2000) suggested that multiplicative noise might excite additional stochastically stable model states in Cessi's (1994) model, but this suggestion was erroneous (Monahan et al., 2002). Very recently, Monahan (2002a,b) worked out a number of differences between deterministic and stochastic stability using a bistable THC box model similar to Cessi's (1994). Shifting the focus to deep convection, we establish in

this paper a concept of stochastic stability that is suitable for broad application in climate dynamics. It is built on the temporal character of jumps between different model regimes.

In the following we move progressively from the example of deep convection to the general concept of stochastic stability. In the next section a highly simplified box model of deep convection is derived (the "1S" model) which is both analytically tractable and open to straightforward conceptual understanding. Despite its simplicity, it reproduces the main dynamical features of the 2TS model. Section 3 deals with the mean residence times in the model regimes, both analytically and numerically; these are crucial for defining stochastic stability. For the analysis of the 1S model output a coarse-grained statistics is developed in section 4. In section 5 a general concept of stochastic stability is established and is applied to the 1S model. In comparing deterministic and stochastic stability, the concepts of effective and wandering unimodality are introduced. Finally, the concluding section 6 features the wider applicability of the stochastic stability concepts.

2 The 1S model

2.1 Model Reduction

As a first step we develop a minimal conceptual model of deep convection that is sufficiently simple to be open to analytical understanding, yet retains the essential features of more complex models of the system. Starting from the 2TS box model with four variables (eqs. (1) to (4)), we will end up with a 1-box model of the mixed layer with salinity as the only variable; this is dubbed the 1S model. To begin the simplification, we note the fact that the variations in the deep ocean are about one order of magnitude smaller than those in the upper layer. This motivates setting the deep box temperature and salinity to constant values T_2^* and S_2^*. Next, the seasonal cycle is not considered; K01

showed that its presence does not change the basic stability properties. Finally, noting that the temperature relaxation time scale of the upper box is much shorter than that of salinity (by a factor of about 20 for the parameter values from observations taken at Ocean Weather Ship Bravo [see K01]), we assume a constant upper box temperature $T_1 = T_1^*$. Together, these assumptions leave us with one single equation for the upper box salinity:

$$\frac{dS_1}{dt} = \frac{1}{\tau_c(\Delta\rho)}(S_2^* - S_1) + \frac{1}{\tau_{1S}}(S_1^* - S_1)\,, \tag{6}$$

with the same function for τ_c as above. The vertical density gradient now depends on S_1 only:

$$\Delta\rho = \rho_1 - \rho_2 = -\alpha(T_1^* - T_2^*) + \beta(S_1 - S_2^*)\,. \tag{7}$$

We can rewrite the two equations (6) and (7) after a transformation of the variables, and switching from time scales to exchange coefficients:

$$\begin{aligned} \frac{dy}{dt} &= -k\,y + k_S(y^* - y)\,, \text{where} \qquad (8)\\ & \quad k = 0 \quad \text{for} \quad y \le y_0 \\ & \quad k = k_c \quad \text{for} \quad y > y_0, \end{aligned}$$

and

$$\begin{aligned} y &= S_1 - S_2^* && (9)\\ y^* &= S_1^* - S_2^* && (10)\\ y_0 &= \frac{\alpha}{\beta}(T_1^* - T_2^*) && (11)\\ k_S &= 1/\tau_{1S} && (12)\\ k &= 1/\tau_c(\Delta\rho)\,, && (13) \end{aligned}$$

with $1/k_c$ of the order of a few days. For later use we define

$$K = k_c/k_S + 1\,. \tag{14}$$

The model is sketched in Figure 1.

The only variable of the 1S model, the vertical salinity gradient y, is restored to a reference value y^*, parameterising primarily the effect of eddy mixing (other processes acting on the upper layer are less important; Houghton and Visbeck (2002)). If the upper layer salinity becomes sufficiently high, then the vertical salinity gradient y overcomes the fixed vertical temperature gradient y_0 ($y - y_0$ being the vertical density gradient), and convective mixing starts. Hence, y_0 plays an important role as a threshold that separates the two *regimes* of the model. In the convecting regime ($y > y_0$), the upper box is coupled to the deep box very strongly; in the nonconvecting regime ($y \leq y_0$) the two boxes are independent. The function $k = 1/\tau_c(\Delta\rho)$ has thus been specified to be a step function with either $k = k_c$ for convection or $k = 0$ in the absence of convection (corresponding to $\tau_c(\Delta\rho) = \infty$).

The two stable steady states $y_{st}^{(n,c)}$ follow immediately from the model equation (8): a nonconvecting, or "off" state at $y_{st}^{(n)} = y^*$, existing if $y^* \leq y_0$, and a convecting, or "on" state at $y_{st}^{(c)} = y^*/K$, existing if $y^* > Ky_0$. For $Ky_0 < y^* \leq y_0$ both states exist; the model is bistable in this parameter range. The stability diagram (Fig. 2) shows that the stability properties of the 1S model are very similar to those of the 2TS model (Fig. 5 in K01). In a manner analogous to the 2TS model, we will study stability changes as a function of two model parameters: the fixed vertical temperature gradient y_0 and the reference vertical salinity gradient y^*.

There are clearly defined borders in parameter space for the existence of the two deterministically stable *states* (Fig. 2), but it is important to note that the two *regimes* always exist. Even if one of the regimes does not contain a stable state, it may be accessed temporarily by the model trajectory, due to perturbations to the system.

Because of its simplicity, equation (8) can be expressed in terms of the potential:

$$U(y) \;=\; \frac{k_S}{2}\left((y-y^*)^2 - (y_0-y^*)^2\right) \quad \text{if } y \leq y_0\,, \tag{15}$$

$$U(y) \;=\; \frac{k_S\,K}{2}\left(\left(y-\frac{y^*}{K}\right)^2-\left(y_0-\frac{y^*}{K}\right)^2\right) \quad \text{if } y>y_0, \tag{16}$$

as

$$\frac{dy}{dt}=-\frac{dU}{dy} \tag{17}$$

where we have set $U(y_0)=0$. This potential (see Fig. 3) is different from a classical double-well potential because the convection threshold in (8) results in a continuous, but non-differentiable, point at y_0. For parameter values in the bistable domain, there is no unstable steady state at this "kink".

The vertical salinity gradient y can be pictured as an overdamped ball moving in the potential landscape U. When it exists, the nonconvecting state is associated with a broad potential well; in contrast, the convecting state appears as a narrower well. The two wells are connected at the convection threshold, where the vertical density gradient is zero. A crossing of the threshold is associated with a transition from one model regime to the other. For instance, if the ball is initially in the nonconvecting well, a strong salinity anomaly may reduce the density gradient until convection starts. Such a perturbation pushes the ball away from the broad well over the threshold into the narrow well of the convecting state. If the model parameters are such that a stable convecting state does not exist ("off" panel of Fig. 3), then the respective potential well is replaced by a mere upward sloping potential curve; in such a case, the ball cannot stay for long on this slope, and will roll back into the nonconvecting state. It is a crucial feature of the model that perturbations can drive the system across the threshold to convect temporarily.

2.2 Stochastic forcing

We now proceed to study the impact on the upper layer of anomalies in the forcing, represented by a stochastic term in the model equation. In the stochastic 2TS model such a term represented the synoptic heat flux variability in the surface fluxes. In the present model, the focus is on the freshwater flux variability in the lateral eddy mixing; such variability is an important contribution

to the long-term evolution of the background stratification of the water column (Houghton and Visbeck, 2002). In the simple picture of the ball in the potential landscape, the stochastic forcing continuously pushes the ball around, both within one well and over the convection threshold between the regimes.

The 1S model equation (8) is thus extended to include a red noise process ξ with decorrelation time τ_ξ and variance $\sigma^2/2\tau_\xi$:

$$\frac{dy}{dt} = -k\,y + k_S(y^* - y) + \xi \tag{18}$$

$$\frac{d\xi}{dt} = -\frac{1}{\tau_\xi}\xi + \frac{\sigma}{\tau_\xi}\zeta_t \tag{19}$$

$$k = 0 \quad \text{for} \quad y \le y_0$$

$$k = k_c \quad \text{for} \quad y > y_0$$

where ζ_t is a Gaussian white noise process. In the limit $\tau_\xi \to 0$ the noise process $\xi(t)$ becomes white noise as well. In what follows, we will consider the dynamics of the system in the $\tau_\xi \to 0$ limit, unless explicitly stated otherwise.

The model's parameters are estimated from the 2TS parameter values (given in Table 2 of K01) and directly from observational data, yielding an "estimated" parameter set. In addition, a "tutorial" parameter set is defined that helps to clarify basic model properties. The estimation is carried out in detail in the Appendix, and the parameter sets are given in Table 1. For the sake of brevity, the units of these parameters are suppressed from this point on.

A typical trajectory of the model (Fig. 4) illustrates how the model fluctuates around the two states for parameter values in the deterministically bistable range, crossing the separating threshold from time to time. The relative frequencies of occurrence of different ranges in y are characterised by the stationary probability density p_s. With the potential (15) and (16) and white noise forcing it is straightforward to use the Fokker-Planck equation (see, for instance, Gardiner (2002)) to determine p_s:

$$p_s(y) = N\,b_n \exp\left[-\frac{k_S}{\sigma^2}\,(y - y^*)^2\right] \quad \text{if } y \le y_0 \tag{20}$$

$$p_s(y) = N\,b_c \exp\left[-\frac{k_S K}{\sigma^2}\left(y - \frac{y^*}{K}\right)^2\right] \quad \text{if } y > y_0\,, \tag{21}$$

where

$$b_n = \exp\left[\frac{k_S}{\sigma^2}\left(y_0 - y^*\right)^2\right], \tag{22}$$

$$b_c = \exp\left[\frac{k_S K}{\sigma^2}\left(y_0 - \frac{y^*}{K}\right)^2\right], \tag{23}$$

and the constant N is determined by the normalisation condition

$$\int_{-\infty}^{\infty} p_s(y)dy = 1\,. \tag{24}$$

Probability density functions (pdfs) are shown at three representative sets of parameter values in Fig. 3. From the exponential functions in (20) and (21) it is obvious that the pdf peak is sharper for a deeper potential well (large k_S values) and for weaker noise (small σ values). It is useful to define the probabilities for the model to be in the convecting regime or in the nonconvecting regime. These are respectively:

$$P_c = \int_{y_0}^{\infty} p_s(y)dy \tag{25}$$

and

$$P_n = \int_{-\infty}^{y_0} p_s(y)dy\,. \tag{26}$$

There is a full correspondence between potential wells and pdf peaks only in the limit of white noise forcing. In a system driven by red noise (with $\tau_\xi > 0$) additional pdf peaks may appear that do not correspond to a potential well, as was shown by Monahan et al. (2002) and Monahan (2002a) in a simple box model of the thermohaline circulation. Fig. 5 reveals that such additional pdf peaks can also appear in the 1S model. In the ball-and-well picture this effect is easily understood. The red noise process will contain phases in which all its values are positive, pushing the ball constantly in the positive direction. For parameter values in the nonconvecting monostable domain, the potential rises steeply in the convecting regime (see Fig. 3); the ball cannot climb up this slope far beyond the threshold of the convecting regime y_0. Consequently, a pdf peak accumulates just at the threshold. The larger the value of τ_ξ is (holding the variance of ξ fixed), the longer are positive-only phases in the red

noise process, yielding a larger pdf peak at the threshold. If σ is held fixed as τ_ξ is increased (as in Figure 5), the peak around y_0 eventually disappears again as the standard deviation of the process ξ (equal to $\sigma/\sqrt{2\tau_\xi}$) becomes so small that the system rarely makes excursions to the threshold of convection.

3 Residence times

The residence time t_r is defined as the time that a trajectory spends uninterruptedly in one regime, regardless of whether a stable steady state exists or not. The *mean* residence time $\langle t_r \rangle$ is an important dynamical feature of a stochastic dynamical system, providing information on how often jumps between the regimes occur. In contrast, the probabilities P_c and P_n only indicate how much time in total the system has spent in either regime. Inspection of Fig. 4 shows that the definition of $\langle t_r \rangle$ includes long episodes of several years in which the trajectory does not approach the convection threshold—but also short episodes of a few time steps' length where the trajectory stays close to the threshold y_0.

Another approach is needed for an approximate analytical calculation of the residence times. Here an expression for the mean *escape* time can be obtained. Provided that a deterministically stable state exists, the escape time t_e is defined as the time the trajectory spends in the corresponding pdf peak before it hits a given threshold for the first time. For a system moving in a potential U, it is an exact result (Gardiner, 2002) that the mean escape time $\langle t_e \rangle$ from a potential minimum (here at y^*) to a threshold (here at y_0, with $y_0 > y^*$) is

$$\langle t_e \rangle = \frac{2}{\sigma^2} \int_{y^*}^{y_0} \exp\left[\frac{2}{\sigma^2} U(y)\right] \left(\int_{-\infty}^{y} \exp\left[-\frac{2}{\sigma^2} U(y')\right] dy' \right) dy \,. \qquad (27)$$

This equation can be used to obtain an approximate analytic form for the mean escape time $\langle t_{e,n} \rangle$ from the nonconvecting state. If we assume small noise, then the first exponential is large only close to y_0, while the second has significant magnitude only near y^*. Hence the contribution from the second integral is relevant for y close to y_0 only, and will not vary strongly for these y

values. Therefore we can set y to y_0 in the upper bound of the second integral and approximate the two integrals as independent. The first integral can be evaluated by linearising $U(y)$ around y_0. With these assumptions, equation (27) becomes

$$\langle t_{e,n} \rangle \approx \frac{\sqrt{\pi}}{k_S} \left(\frac{k_S}{\sigma^2} (y_0 - y^*)^2 \right)^{-\frac{1}{2}} \exp \left[\frac{k_S}{\sigma^2} (y_0 - y^*)^2 \right] . \tag{28}$$

This analytical expression for the mean escape time is a generalisation of a result already achieved by Kramers (1940) in that it applies to a bistable potential without an intermediate unstable fixed point. It is a good approximation to the mean residence time $\langle t_{r,n} \rangle$ if there is a sharp pdf peak around the "off" state (cf. eq. 20). This is achieved with either a strong restoring coefficient k_S, a small noise σ or a large positive difference $(y_0 - y^*)$, as can be seen in comparing Fig. 6a with Fig. 6c. For a small difference $(y_0 - y^*)$, the term $(k_S (y_0 - y^*)^2 \sigma^{-2})^{-1/2}$ in (28) diverges and the approximation breaks down (right-hand parts of Fig. 6a and 6c).

Having considered the mean residence time in a regime, we now turn to the distribution of residence times and to the probability for the residence time to exceed a particular threshold. For a random process with vanishing autocorrelation it is known (Leadbetter et al., 1983; von Storch and Zwiers, 1999) that the residence time t_r in a given regime is exponentially distributed:

$$p_s(t_r) = \frac{1}{\langle t_r \rangle} \exp \left[-\frac{t_r}{\langle t_r \rangle} \right] , \tag{29}$$

with $\langle t_r \rangle$ the mean residence time. The cumulative probability distribution for t_r is then:

$$P(t_r < t_x) = \int_0^{t_x} p_s(t_r) dt_r = 1 - \exp \left[-\frac{t_x}{\langle t_r \rangle} \right] \tag{30}$$

It is easy now to compute, for instance, the probability that the residence time $t_{r,n}$ exceeds one year:

$$P(t_{r,n} > 1) = 1 - P(t_{r,n} < 1) = \exp \left[-\frac{1}{\langle t_{r,n} \rangle} \right] \tag{31}$$

In fact, the process y will have a nonzero autocorrelation, due on the one hand to the relaxation parameters k_S and k_c and on the other hand, if $\tau_\xi > 0$, to

autocorrelation in the forcing. If the resulting autocorrelation time for y is still clearly smaller than the mean residence time $\langle t_{r,n} \rangle$, then (31) is a reasonable approximation (Khatiwala et al., 2001), although (29) does not hold generically (e.g. K01).

4 Coarse-grained statistics

For the sake of simplicity we have avoided introducing an explicit seasonal cycle in the 1S model. Deep convection events thus may occur at any time in a given year, not only around the seasonal cycle's extremum in winter. Of course, in the real ocean the physically most relevant feature of a deep convection event is the mixing of the deep waters with the cold surface layer waters. Yet, in the absence of a seasonal cycle in the model, it does not really matter when exactly in a given year deep convection occurs, since the diffusive and advective time scales of the deep ocean are larger than one year, and since there is virtually no seasonal cycle in the deep waters. Under statistically stationary boundary conditions, the relevant physical question is whether convection did occur in a given year at all. This motivates the introduction of a coarse-grained measure of convection frequency. A given year of the model output is called a "convecting year" if there was at least one convection event during this year. In the ball-and-well picture, this means that the ball has left the nonconvecting state and has been pushed across the convection threshold on at least one occasion. Otherwise, a year is said to be "nonconvecting". As an illustrative example, consider the time series displayed in Fig. 4. The model years 24, 28, 29, 30, 31, 36, 37, and 38 are "convecting years" in the above sense.

The coarse-grained measure of convection we will consider is n_c, the probability for convection to occur in a given year, which was introduced in K01. This measure is coarse-grained in the sense that it neglects any short back-and-forth changes between the convecting and the nonconvecting regime. An approximate analytical expression for n_c is readily derived. Suppose that the

nonconvecting stable state exists, and that (28) is valid. Obviously,

$$n_c = 1 - P(y < y_0 \text{ during one year}) \,. \tag{32}$$

The probability $P(y < y_0$ during one year) for convection not to occur in a year is the probability to be in the nonconvecting regime in the beginning of this year times the probability for the residence time in the nonconvecting regime to exceed one year:

$$P(y < y_0 \text{ during one year}) \tag{33}$$

$$= P(y(t_0) < y_0) \cdot P(y(t) < y_0 \text{ for all } t_0 < t \leq t_0 + 1) \tag{34}$$

$$= \int_{-\infty}^{y_0} p_s(y)dy \cdot \exp\left[-\frac{1}{\langle t_{r,n}\rangle}\right] . \tag{35}$$

Finally,

$$n_c = 1 - P_n \cdot \exp\left[-\frac{1}{\langle t_{r,n}\rangle}\right] . \tag{36}$$

Fig. 7a shows a comparison of n_c estimated from long numerical integrations of (18) to the approximate analytic expression (36). The analytical approximation obviously follows the numerical values quite closely over a large range of y^* values, for the (representative) value of y_0 considered. Numerical estimates of n_c over broad ranges of y^* and y_0 are illustrated in Figure 7b.

It is instructive to compare n_c, the probability for a year to be convecting, with P_c, the probability to be in the convecting regime. Fig. 7a already reveals that n_c can exceed P_c significantly. Figure 8 contours both P_c and n_c over broad ranges of y_0 and y^*. Consider first the bistable domain in the lower left-hand section of the panels. Here, the values of P_c and n_c are very similar, as both potential wells are rather deep and the mean residence times are large. Consequently, the exponential in (36) is close to one, so $n_c \approx 1 - P_n \cdot 1 \approx P_c$. Now consider the upper right-hand section of the panels, in the "off" domain. Here the differences of P_c and n_c are considerable. The P_c values are small, as the stable convecting state does not exist, while the n_c values are large (even exceeding 0.5 for some parameter values). The explanation lies in the "off" potential well being shallow here: the small residence time

implies that there are frequent excursions from the neighbourhood of the non-convecting steady state, beyond the convection threshold. The pdf peak is broad, as can be seen in the "off" panel of Fig. 3. Since $P_n \approx 1$, we have $n_c \approx 1 - 1 \cdot \exp[-1/\langle t_{r,n} \rangle]$. In the "off" domain, n_c depends on the mean residence time only. This feature comes out clearly in Fig. 7b: for large y^*, the isolines of n_c are parallel to the "off" domain borders. These parallels are the lines of equal $(y_0 - y^*)$, and therefore isopotential (and iso-residence time) lines in the nonconvecting regime. Note that the analytical expression (36) gives a good qualitative understanding although the broad pdf peak (cf. Fig. 3) indicates a relatively strong noise intensity.

The qualitative differences between P_c and n_c are the same when using the more realistic "estimated" parameters, rather than the "tutorial" parameters. Fig. 9 shows that with the estimated parameters, the dependence of P_c and n_c in the bistable domain on y^* is weaker than in Fig. 8, but the dependence on y_0 is stronger. This leads to kinks in the contours along the line where the stable deterministic convecting state vanishes. The reason lies in the larger ratio K of the two potential wells in either parameter set (see (15) and (16)). Nevertheless, in the "off" domain the values of n_c still exceed the P_c values.

The difference in physical interpretation between P_c and n_c is noteworthy. The frequent occurrence of convecting years need not depend on the existence of the convecting state, because short excursions over the convection threshold into the convecting regime are sufficient to achieve convection events. The curves of n_c run smoothly (or at least continuously) across the point where the convecting stable state ceases to exist.

5 Stochastic stability of deep convection

5.1 The concept of stochastic stability

This study has direct bearing on the concept of stability of a climate state. In particular, it highlights a pronounced difference between deterministic stability and stochastic stability. Note that the dynamic stability we deal with here is not to be confused with the static stability of a stratified fluid.

In a deterministic dynamical system, where no random perturbations are present, we call a state *stable* under two conditions: (i) it is a steady state, meaning that it does not change in time (except possibly due to external modulations like the seasonal cycle), and (ii), if small perturbations are applied, these asymptotically decay and the steady state is reached again (although transient growth is possible).

In a stochastic dynamical system driven by Gaussian noise, the random perturbations will sooner or later carry the system away from any neighbourhood of a steady state. If there are two or more coexisting stable states, the system will jump between their neighbourhoods, or *regimes* in our terminology. These regimes are then described as being metastable. The deterministic stability definition, focusing on one stable state and its neighbourhood, is thus no longer applicable. In such a bi- or multistable system, a definition of stochastic stability has to take these transitions between metastable regimes into account. If the lifetimes of the metastable regimes are not longer than the longest physically-relevant time scales in the phenomenon under consideration, then it is the relative stability of these regimes that is of primary interest (Monahan, 2002b). There are several natural measures of the relative stochastic stability of metastable regimes (Freidlin and Wentzell, 1998):

1. In a system with a potential, work must be done moving from one state to the other. Intuitively, less work is needed to escape the shallower well than the deeper, so the relative well depth $\Delta U = U^{(1)} - U^{(2)}$ (where $U^{(k)}$ is the potential at the bottom of the k-th well) is a measure of the relative

stability of the metastable regimes containing these states.

2. The ratio of mean escape times, $\langle t_e^{(1)} \rangle / \langle t_e^{(2)} \rangle$ is a second measure. Intuitively, a metastable regime with a longer mean escape time should be more stable than one with a shorter mean escape time.

3. A third natural measure of relative stability is the stationary probability of being in each of the metastable regimes, obtained from integrations of the stationary probability density function (see eq. (25) and (26)). This is a direct measure of the relative fractions of time spent by the system in each regime; a regime that is occupied more often is intuitively more stable.

4. Finally, as we have seen above, the specific problem may motivate consideration of the relative frequency of occupation of regimes coarse-grained by a basic time unit (e.g. one day or one year). This defines the fourth of our measures of stability.

The first three of these measures are naturally interrelated. This follows immediately from the classical result by Kramers (1940) that the mean escape time from a metastable regime depends exponentially on the potential difference that has to be overcome to leave it:

$$\langle t_e^{(k)} \rangle \sim e^{(U^{(0)} - U^{(k)})/\sigma^2} \,, \tag{37}$$

where $U^{(0)}$ is the potential at the point separating the two wells, and σ the noise intensity. Clearly, then, the potential difference ΔU will be directly proportional to the logarithm of the ratio of the mean escape times:

$$\sigma^2 \ln \left(\frac{\langle t_e^{(1)} \rangle}{\langle t_e^{(2)} \rangle} \right) \sim -\Delta U \,, \tag{38}$$

or similarly of the mean residence times. Furthermore, as the stationary distribution is given by

$$p_s(y) \sim e^{-U(y)/\sigma^2}, \tag{39}$$

the ratio of the heights of the two peaks of the distribution is logarithmically related to ΔU:

$$\sigma^2 \ln\left(\frac{p_s(y^{(1)})}{p_s(y^{(2)})}\right) \sim -\Delta U. \tag{40}$$

If the peaks corresponding to the two metastable regimes are well-separated, the ratio of their heights should be proportional to that of their areas:

$$\frac{p_s(y^{(1)})}{p_s(y^{(2)})} \sim \frac{P^{(1)}}{P^{(2)}} = \frac{\int_{-\infty}^{y0} p_s(y)\, dy}{\int_{y0}^{\infty} p_s(y)\, dy}, \tag{41}$$

which relates the third of the stability measures to the first two. In contrast, the coarse-grained fourth measure of stability combines probability and residence time information, and as we have seen above can give quite a different picture of stochastic stability than the first three measures.

The discussion of relative stability has so far assumed the existence of a deterministically stable steady state within each metastable regime. In fact, the second, third, and fourth of the above measures can also be used to compare the relative stability of two regions of state space if they do not contain a deterministic stable state. Partitioning the state space in this way is natural when the system contains thresholds separating regimes with different physical behaviour, such as the convection threshold in the 1S model.

5.2 Stochastic stability in the 1S model and in the 2TS model

The measures of relative stability for 1S model simulations are shown in Fig. 10 a, b, c, and e. As expected, the first three of these provide similar results. For a symmetric potential, the lines of $\Delta U = 0$, $\log(\langle t_{r,n}\rangle/\langle t_{r,c}\rangle) = 0$, and $P_c = 0.5$ would coincide. The asymmetry of the 1S model explains the deviations. The contours of P_c, $\log(\langle t_{r,n}\rangle/\langle t_{r,c}\rangle)$, and n_c run smoothly across the deterministic stability boundaries, reflecting that those measures are not confined to deterministic stability domains.

Note that there is only a narrow band in the (y_0, y^*) parameter plane in which the probabilities to be in either regime do not differ by at least

one order of magnitude. This is consistent with the narrow band in which $|\log(\langle t_{r,n}\rangle/\langle t_{r,c}\rangle)| < 1$. In the largest part of the bistable domain, to either side of this band, the probability to be in one regime is very close to one, and the other regime is rarely visited (E_1 and E_2 in Fig. 10a and b). Although the system has two potential wells, one well is so large that the pdf peak corresponding to the other well becomes very small. The deterministic bistability is turned into *effective unimodality* (Fig. 10d). Effective unimodality is a known feature of stochastic bistable systems (Gardiner, 2002), and has recently been studied in a box model of the thermohaline circulation (Monahan (2002a,b)).

If one considers the coarse-grained probability n_c, the probability for a convecting year, then a new stochastic stability effect comes up. It has been shown that in parts of the "off" domain (near point W in Fig. 10b), the probability for a convecting year is still high, although the steady convecting state does not exist. The high values of n_c are explained by short excursions of the model trajectory into the convecting regime. Seen in the ball-and-well picture, the ball now spends most of the time in the broad potential well of the nonconvecting state. The short mean residence time in this broad well however (Fig. 6b) makes the ball wandering across the convection threshold y_0 and into the convecting regime rather often (see the pdf shape in Fig. 10f). We call this *wandering unimodality:* while the pdf has only one peak that is in the nonconvecting regime, the coarse-grained probability n_c for a convecting year is significantly larger than P_c (and may even exceed 0.5).

The n_c stability diagram from the 1S model compares well with the n_c stability diagram from the 2TS model (Fig. 11; see K01 for details on n_c in the 2TS model). Both stability diagrams show the areas of effective unimodality in the flanks of the bistable domain, and both stability diagrams show wandering unimodality in parts of the "off" domain.

Wandering unimodality is an important effect because the coarse-grained probability n_c is well motivated by the physics of deep convection. In the real ocean like in the two box models, the convection events are quite short, but

still achieve the vertical mixing of the water column. The forcing of the upper layer must bring the vertical density gradient just beyond the point of neutral stratification to start a convection event. In the deterministic 2TS model this is achieved by the seasonal cycle; in the stochastic 2TS model convection events may be additionally triggered (or suppressed) by the stochastic forcing; and in the 1S model the stochastic forcing alone causes the convection events.

The presence of noise has two opposite effects: deterministic bistability is turned into effective unimodality, and due to the coarse-grained probability n_c areas of deterministic monostability show wandering unimodality, with both regimes being visited. The effect of noise is crucial, too, for the interpretation of the position of the optimal parameter set that the represents the conditions in the Labrador Sea (the asterisk in Fig. 11a). In the deterministic setting, a shift of the parameters to lower S_1^* or larger T_1^* looks quite dramatic due to the associated loss of stability for the "on" state. In the stochastic model, however, the same parameter changes lead to quantitative, but not qualitative, changes in the frequency of deep convection. In other words, the presence of variability in the climate system has a moderating influence here, since the *stochastic* model can repeatedly make excursions to a regime where no *deterministically* stable state exists.

6 Conclusions

Open-ocean deep convection is a highly variable process. Moreover, it is an essential part of the present day circulation in the North Atlantic, making questions about its sensitivity to changes in climate (be they of natural or of human origin) of global significance. Furthermore, its manifest nonlinearity and the well-recognised significance of environmental variability to its dynamics make it a prototypical system for considering the stochastic stability of a system with multiple regimes of behaviour.

In this study, we have introduced the idea of stochastic stability in terms of

the relative frequency of occupation of different climate regimes. In contrast to the classical deterministic concept of stability, the variability of the system is central to the idea of stochastic stability. The relative frequency of occupation of a regime can be measured by its mean residence time, by the total time spent there, or by the depth of the associated potential well (should one exist). Changing model parameters smoothly changes these frequencies of occupation and thus the stochastic stabilities; deterministic stability, by contrast, generally changes abruptly in a bifurcation. This smooth dependence of stochastic stability on system parameters reflects the fact that stability can be defined for any region in phase space, and not only for neighbourhoods of deterministic steady states.

The use of a highly simplified box model—the 1S model—for this study carries several advantages. First of all, many of the results are obtained analytically, which fosters a conceptual understanding of the model behaviour. Still, since the 1S model was derived from a more comprehensive box model of deep convection (K01), it retains the most relevant dynamical features: the bistability of deep convection in a certain domain in parameter space, the positive salinity feedback, and also the stochastic stability properties. Of course, the 1S model is a highly simplified representation of the phenomenon of deep convection, and cannot be expected to be quantitatively accurate in its fine details. A natural extension of the present study would be the investigation of the response to environmental fluctuations of models with a higher degree of fidelity to the natural world. The value of conceptual models—such as the 1S model—lies in that they provide a useful conceptual vocabulary with which the dynamics of more complex models, and the natural system, can be understood. It is worth noting that the dynamics of the 1S model are in fact not much simpler than those used to parameterise convection in many complex Ocean General Circulation Models.

The presence of the seasonal cycle motivates a coarse-grained analysis of the 1S time series. This is done by analysing whether the convection threshold

is hit in any one year. The probability n_c for such a convecting year to occur can be large even if the probability for convection itself is low. This effect is dubbed wandering unimodality: there is only one pdf peak in the nonconvecting regime, but the trajectory of the stochastically forced model frequently wanders to the physical threshold where convection starts. In this way the deep water formation process continues, although unstable from a deterministic perspective. The analytical expression for n_c explains wandering unimodality through relatively short residence times in the nonconvecting regime.

Another phenomenon of stochastic stability is called effective unimodality (see also Monahan (2002a,b)). Here the stochastic forcing has the effect that one of two existing potential wells is almost never visited. Both effects, effective unimodality and wandering unimodality, highlight the need to take climate variability into account when analysing the stability of climate states. The deterministic picture of one or more distinct stable climate states is replaced by the stochastic picture of transitions between different regimes. It is this wandering which has to be taken as the overall climate state, and not only the averages of the observed quantities. Changes in the statistical properties of this wandering may then reflect a changed regime preference in the sense of Palmer (1999) and Khatiwala et al. (2001).

In the 1S model the mean length of nonconvecting phases increases exponentially under a surface freshening or warming. However, it can easily be shown (see Kuhlbrodt (2002) for details) that the probability of these phases to be longer than a given length shows an exponential growth only initially, when this probability is small, but then shifts to a linear increase. Since the simplicity of the 1S model allows a generalisation of these results, they are an extension to the study of Khatiwala et al. (2001). At least in the context of reduced climate models, the probability of persistent climate events (e.g. the climate state being locked in one regime for an extended period) is not necessarily an exponential function of the forcing parameters.

We expect that our results are still valid if feedbacks with the surroundings

are included. Preliminary calculations indicate that coupling the noise strength to the model variable leads to an enlarged domain of bimodality. As well, abrupt shutdowns of deep convection, persisting for decadal or centennial time scales and due to environmental fluctuations, have been observed in coupled General Circulation Models (Hall and Stouffer, 2001; Goosse et al., 2002). Also, the results can easily be carried over to other physical systems with a relevant threshold. For instance, simple stochastic models have been successfully applied to convection in the tropical atmosphere (Lin and Neelin, 2000; Yano et al., 2001; Palmer, 2001). There is evidence that large-scale atmospheric variability may be seen as a wandering between different regimes (e.g. Corti et al. (1999), Monahan et al. (2001), although see Hsu and Zwiers (2001); see Sura (2002) for a recent model study). Because of the many time and space scales involved in these phenomena, concepts of stochastic stability provide a natural framework for advancing their understanding.

Appendix

The parameter estimation for the 1S model is presented here in detail. Starting from the parameter values from the 2TS model (see Table 2 in K01), the 1S parameters are obtained according to eqs. (10) to (12). This yields $y^* \approx -1.5$ psu and $y_0 = 0.04\,\mathrm{psu}$. With the time unit being one year, we have $k_S = 0.125\,\mathrm{yr}^{-1}$ from $\tau_{1S} = 8$ yr. The convective mixing rate is taken to be $k_c = 50\,\mathrm{yr}^{-1}$, corresponding to a time scale of one week.

An independent check of these parameters is possible with data from Houghton and Visbeck (2002), hereafter cited as HV02 (see also Khatiwala et al. (2002)). They specify a value of 0.2 Sv total mean freshwater flux into the Labrador Sea. Most of this freshwater is transported by boundary currents, such that only 20%, or 0.04 Sv, reaches the interior Labrador Sea upper layer through lateral eddy mixing. HV02 further assume a volume of $V = (0.6 \cdot 10^6\,\mathrm{km}^2) \cdot 300\,\mathrm{m}$ for the interior of the Labrador Sea, and a reference

salinity of $S_0 = 35$ psu. One obtains a mean freshwater flux of

$$\Phi_{FW} = \frac{S_0}{V} \cdot 0.04\,\mathrm{Sv} = 0.25\,\mathrm{psu\,yr^{-1}}\,. \tag{42}$$

According to HV02 the error of this estimated freshwater flux is 50%. Thus, the value of Φ_{FW} is consistent with the observation that, if convection is absent, the upper layer salinity decreases at about half of this rate (Fig. 5 of HV02). In the 1S model, the initial salinity decrease immediately after the end of convection ($y(t_0) = y_0$) is

$$\frac{dy}{dt} = k_S(y^* - y(t_0)) = 0.125 \cdot (-1.5 - 0.04)\,\mathrm{psu\,yr^{-1}} \approx -0.2\,\mathrm{psu\,yr^{-1}}\,, \tag{43}$$

which is consistent with the above value of Φ_{FW}.

We now estimate the decorrelation time τ_ξ and the standard deviation std(ξ) of the stochastic term ξ in (18). The decorrelation time scale of the freshwater flux is difficult to estimate from observations because long time series are sparse, and the freshwater flux has many sources (such as continental runoff, sea ice advection and melt, precipitation). Proxy time series from models suggest a decorrelation time of half a year to three years. Such time series include the sea ice export through Fram strait (H. Haak, pers. comm.) and the sea ice volume in Baffin Bay (M. Maqueda, pers. comm.). We use here $\tau_\xi = 2$ years. In this way, the stochastic freshwater forcing includes interannual anomalies.

According to HV02, the anomalous freshwater flux associated with the Great Salinity Anomaly is 20% of the mean freshwater flux. Assuming Great Salinity Anomalies to be typical for the interannual salinity fluctuations, this yields a standard deviation std(ξ) = $0.05\,\mathrm{psu\,yr^{-1}}$. The anomalous freshwater flux amounts to 20% of the processes that drive the seasonal cycle (HV02). To ensure a realistic amount of variability in the model, the variance of the seasonal cycle is subsumed in that of the noise; a sensible choice is std(ξ) = $0.25\,\mathrm{psu\,yr^{-1}}$. Since we will also use white noise forcing, it is useful to determine the noise intensity σ in (19) from $\sigma = \sqrt{2\tau_\xi}\,\mathrm{std}(\xi)$ (Gardiner, 2002). With the above values this yields $\sigma = 0.5\,\mathrm{psu\,yr^{-1/2}}$.

Acknowledgements

Vladimir Petoukhov's comments were very helpful to clarify the concepts presented in this paper. We are grateful to Stefan Rahmstorf for supporting our collaboration. Suggestions by two anonymous reviewers led to some important clarifications. TK's work was funded by the Deutsche Forschungsgemeinschaft (Sfb 555). AM is supported by the Natural Sciences and Engineering Research Council of Canada, by the Canadian Foundation for Climate and Atmospheric Sciences, and by the Canadian Institute for Advanced Research.

References

Belkin, I. M., Levitus, S., Antonov, J., and Malmberg, S.-A. (1998). Great Salinity Anomalies in the North Atlantic. *Prog. Oceanog.*, 41:1–68.

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

Cessi, P. (1996). Convective adjustment and thermohaline excitability. *J. Phys. Oceanogr.*, 26:481–491.

Corti, S., Molteni, F., and Palmer, T. N. (1999). Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, 398:799–802.

Dickson, R. R., Lazier, J., Meincke, J., Rhines, P., and Swift, J. (1996). Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanog.*, 38:241–295.

Dickson, R. R., Meincke, J., Malmberg, S.-A., and Lee, A. J. (1988). The "Great Salinity Anomaly" in the Northern North Atlantic 1968–1972. *Prog. Oceanog.*, 20:103–151.

Freidlin, M. I. and Wentzell, A. D. (1998). *Random Perturbations of Dynamical Systems*. Springer, New York, 2nd edition.

Ganachaud, A. and Wunsch, C. (2000). Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data. *Nature*, 408:453–456.

Gardiner, C. W. (2002). *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer series in Synergetics*. Springer, Berlin, 2nd edition.

Goosse, H., Renssen, H., Selten, F. M., Haarsma, R. J., and Opsteegh, J. D. (2002). Potential causes of abrupt climate events: a numerical study with a three-dimensional climate model. *Geophys. Res. Lett.*, 29:10.1029/2002GL014993.

Hall, A. and Stouffer, R. J. (2001). An abrupt climate event in a coupled ocean-atmosphere simulation without external forcing. *Nature*, 409:171–174.

Hasselmann, K. (1976). Stochastic climate models, Part I: Theory. *Tellus*, 28:473–485.

Hirschi, J., Sander, J., and Stocker, T. F. (1999). Intermittent convection, mixed boundary conditions and the stability of the thermohaline circulation. *Clim. Dyn.*, 15:277–291.

Houghton, J. T., Ding, Y., Griggs, D., Noguer, M., van der Linden, P. J., and Xiaosu, D., editors (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge University Press, Cambridge.

Houghton, R. W. and Visbeck, M. H. (2002). Quasi-decadal salinity fluctuations in the Labrador Sea. *J. Phys. Oceanogr.*, 32(2):687–701.

Hsu, C. and Zwiers, F. (2001). Climate change in recurrent regimes and modes of atmospheric variability. *J. Geophys. Res.*, 106:21045–20159.

Imkeller, P. and Monahan, A. H. (2002). Conceptual stochastic climate models. *Stochastics and Dynamics*, 2:311–326.

Khatiwala, S., Schlosser, P., and Visbeck, M. (2002). Rates and mechanisms of water mass transformations in the Labrador Sea as inferred from tracer observations. *J. Phys. Oceanogr.*, 32(2):666–686.

Khatiwala, S., Shaw, B. E., and Cane, M. A. (2001). Enhanced sensitivity of persistent events to weak forcing in dynamical and stochastic systems: Implications for climate change. *Geophys. Res. Lett.*, 28:2633–2636.

Kramers, H. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304.

Kuhlbrodt, T. (2002). *Stability and variability of open-ocean deep convection in deterministic and stochastic simple models*. PhD thesis, University of Potsdam, Germany.

Kuhlbrodt, T., Titz, S., Feudel, U., and Rahmstorf, S. (2001). A simple model of seasonal open ocean convection. Part II: Labrador Sea stability and stochastic forcing. *Ocean Dynamics*, 52(1):36–49.

Lazier, J. R. N. (1980). Oceanographic conditions at Ocean Weather Ship Bravo, 1964–1974. *Atmosphere-Ocean*, 18:227–238.

Leadbetter, M., Lindgren, G., and Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York, Heidelberg, Berlin.

Lenderink, G. and Haarsma, R. J. (1994). Variability and multiple equilibria of the thermohaline circulation associated with deep-water formation. *J. Phys. Oceanogr.*, 24:1480–1493.

Lenderink, G. and Haarsma, R. J. (1996). Modeling convective transitions in the presence of sea ice. *J. Phys. Oceanogr.*, 36(8):1448–1467.

Lilly, J. M., Rhines, P. B., Visbeck, M., Davis, R., Lazier, J. R. N., Schott, F., and Farmer, D. (1999). Observing deep convection in the Labrador Sea during winter 1994–1995. *J. Phys. Oceanogr.*, 29:2065–2098.

Lin, J. W.-B. and Neelin, J. D. (2000). Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.*, 27(22):3691–3694.

Manabe, S. and Stouffer, R. J. (1988). Two stable equilibria of a coupled ocean-atmosphere model. *J. Clim.*, 1:841–866.

Manabe, S. and Stouffer, R. J. (1999). Are two modes of thermohaline circulation stable? *Tellus*, 51A(3):400–411.

Marshall, J. and Schott, F. (1999). Open-ocean convection: Observations, theory, and models. *Rev. Geophys.*, 37:1–64.

Monahan, A. H. (2002a). Correlation effects in a simple model of the thermohaline circulation. *Stochastics and Dynamics*, 2:437–462.

Monahan, A. H. (2002b). Stabilisation by noise of climate regimes in a simple model: implications for stability of the thermohaline circulation. *J. Phys. Oceanogr.*, 32:2072–2085.

Monahan, A. H., Pandolfo, L., and Fyfe, J. C. (2001). The preferred structure of variability of the northern hemisphere atmospheric circulation. *Geophys. Res. Lett.*, 28:1019–1022.

Monahan, A. H., Timmermann, A., and Lohmann, G. (2002). Comments on "Noise-induced transitions in a simplified model of the thermohaline circulation". *J. Phys. Oceanogr.*, 32(3):1112–1116.

Palmer, T. N. (1999). A nonlinear dynamical perspective on climate prediction. *J. Clim.*, 12:575–591.

Palmer, T. N. (2001). A nonlinear dynamical perspective on model error: a proposal for nonlocal stochastic-dynamic parameterisation in weather and climate prediction models. *Quart. J. Roy. Meteorol. Soc.*, 127(572):279–304.

Rahmstorf, S. (2000). The thermohaline ocean circulation: a system with dangerous thresholds? *Climatic Change*, 46:247–256.

Rahmstorf, S. (2001). A simple model of seasonal open ocean convection. Part I: Theory. *Ocean Dynamics*, 52(1):26–35.

Schaeffer, M., Selten, F. M., Opsteegh, J. D., and Goosse, H. (2002). Intrinsic limits to predictability of abrupt regional climate change in IPCC SRES scenarios. *Geophys. Res. Lett.*, 29(16):10.1029/2002GL015254.

Sura, P. (2002). Noise-induced transitions in a barotropic beta-plane model. *J. Atmos. Sci.*, 59:97–110.

Timmermann, A. and Lohmann, G. (2000). Noise-induced transitions in a simplified model of the thermohaline circulation. *J. Phys. Oceanogr.*, 30:1891–1900.

von Storch, H. and Zwiers, F. (1999). *Statistical analysis in climate research*. Cambridge University Press, Cambridge.

Welander, P. (1982). A simple heat-salt oscillator. *Dyn. Atmos. Oceans*, 6:233–242.

Wood, R. A., Keen, A. B., Mitchell, J. F. B., and Gregory, J. M. (1999). Changing spatial structure of the thermohaline circulation in response to atmospheric CO_2 forcing in a climate model. *Nature*, 399:572–575.

Yano, J. I., Fraedrich, K., and Blender, R. (2001). Tropical convective variability as $1/f$ noise. *J. Clim.*, 14(17):3608–3616.

Parameter set	k_S (yr^{-1})	k_c (yr^{-1})	σ ($\mathrm{psu\,yr}^{-1/2}$)	τ_ξ (yr)	y^* (psu)	y_0 (psu)
"estimated"	0.125	50	0.5	2.0	-1.5	0.04
"tutorial"	1.0	10	0.8	0	–	–

Table 1: *Parameter sets for 1S model simulations. For the tutorial parameter set no particular values for y^* and y_0 are specified.*

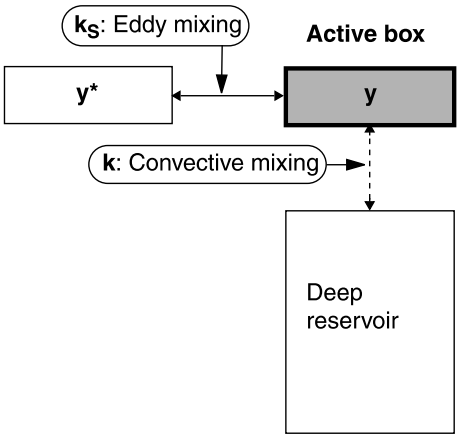

Figure 1: *Sketch of the 1S model with one single active box. The only variable is the vertical salinity gradient y. Restoring to the reference value y^* represents mixing with surrounding waters. The deep box is considered to be an infinite reservoir of water with constant salinity.*

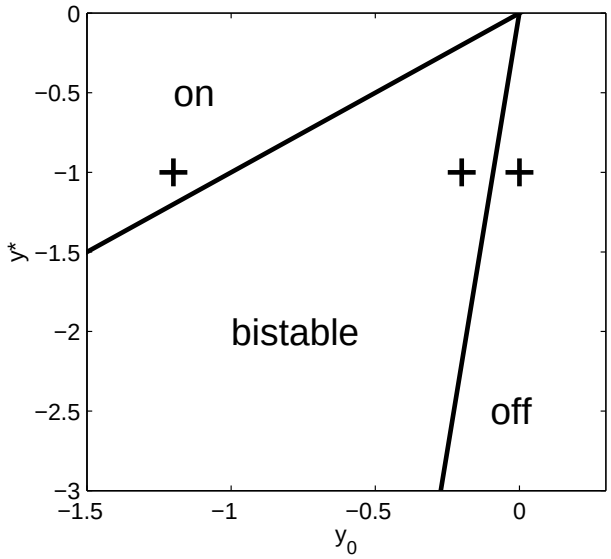

Figure 2: *Stability diagram of the 1S model for the parameters $k_S = 1$ and $k_c = 10$. Depending on the parameters y^* and y_0 there exist one or both stable states, with convection being "on" or "off". The lines define the borders of the respective domains. The crosses show the parameter sets used for the panels in Fig. 3.*

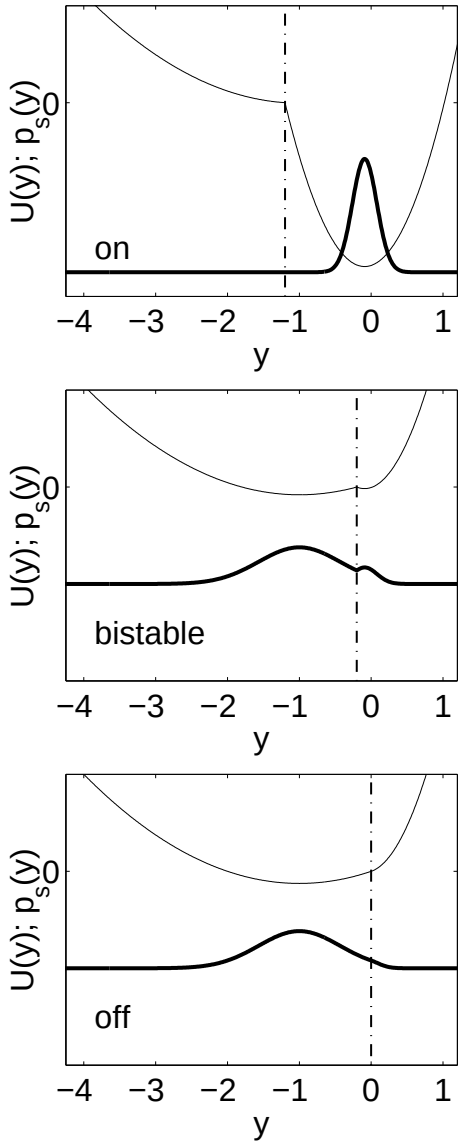

Figure 3: *Thin lines: potential U (in arbitrary units) as a function of y as given by eqs. (15) and (16). The potential is shown for three cases: convecting monostable ("on"), bistable, and nonconvecting monostable ("off"), corresponding to the crosses in Fig. 2. Tutorial parameters with $y^* = -1$ and $y_0 = -1.2; -0.2; 0.0$, respectively. The convection threshold y_0 (dash-dotted) separates the convecting and the nonconvecting regime. Thick lines: probability density function p_s (in arbitrary units) corresponding to the potential curves. The lower panel shows that there is a non-zero probability for the "on" regime to be occupied even if the stable "on" state does not exist.*

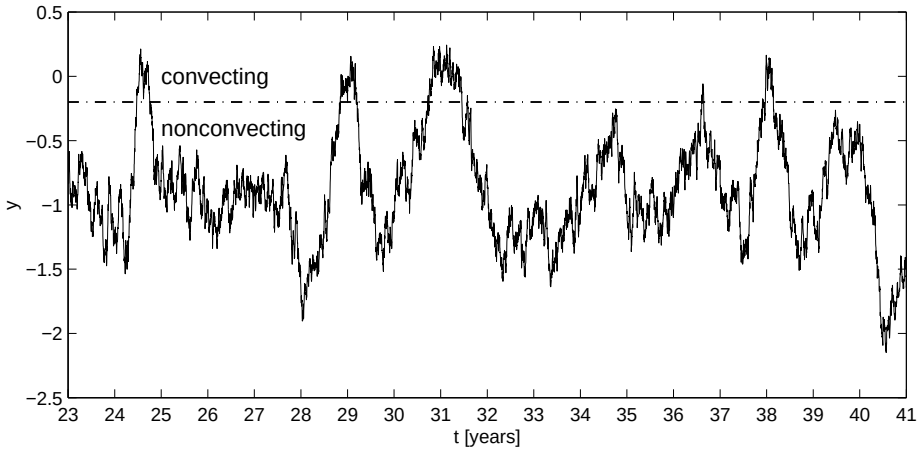

Figure 4: *Time series from the 1S model using the tutorial parameter set and* $y^* = -1$, $y_0 = -0.2$. *The parameters correspond to the middle cross in Fig. 2. The model is in the bistable domain. The dash-dotted line denotes the threshold* y_0 *that separates the convecting regime* $(y > y_0)$ *from the nonconvecting regime* $(y \leq y_0)$. *Convection occurs in the model years 24, 28, 29, 30, 31, 36, 37, and 38.*

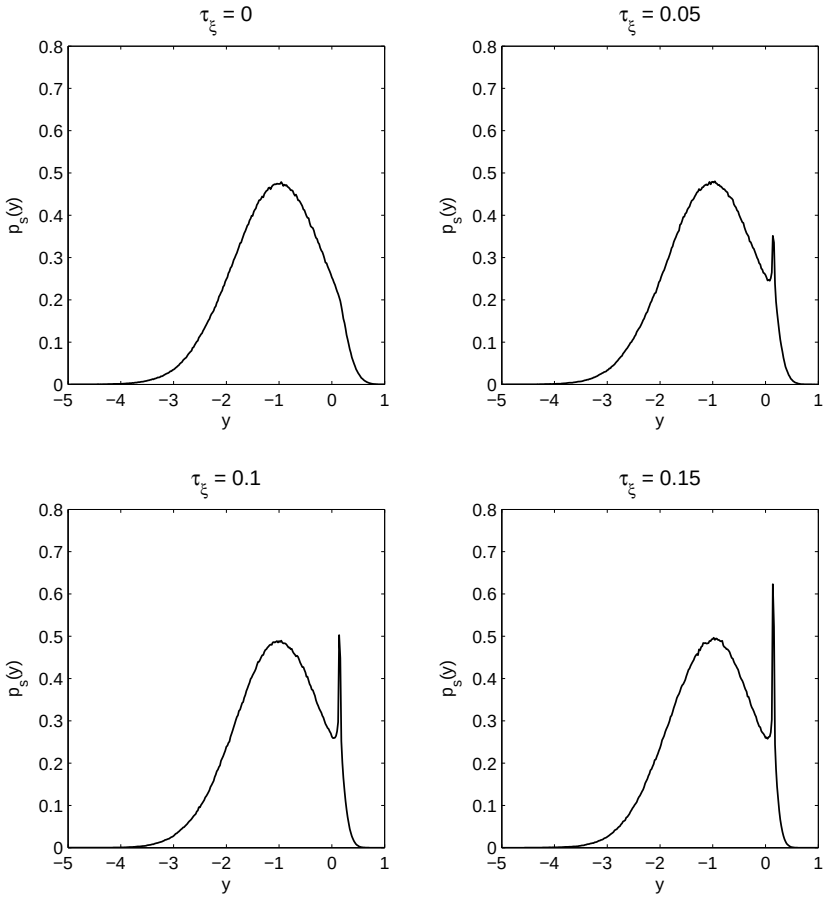

Figure 5: *Dependence of the pdf shape on the noise decorrelation time τ_ξ for tutorial parameters with $\sigma = 1.25$, $y^* = -1.0$, and $y_0 = 0.15$. The model is in the nonconvecting monostable domain. The upper left hand panel shows the white noise limit case with one pdf peak. With increasing τ_ξ (such that p_s retains significant probability around $y = y_0$) a second pdf peak emerges, although there is no deterministic bistability. As τ_ξ is further increased, the probability around $y = y_0$ shrinks and the second peak disappears.*

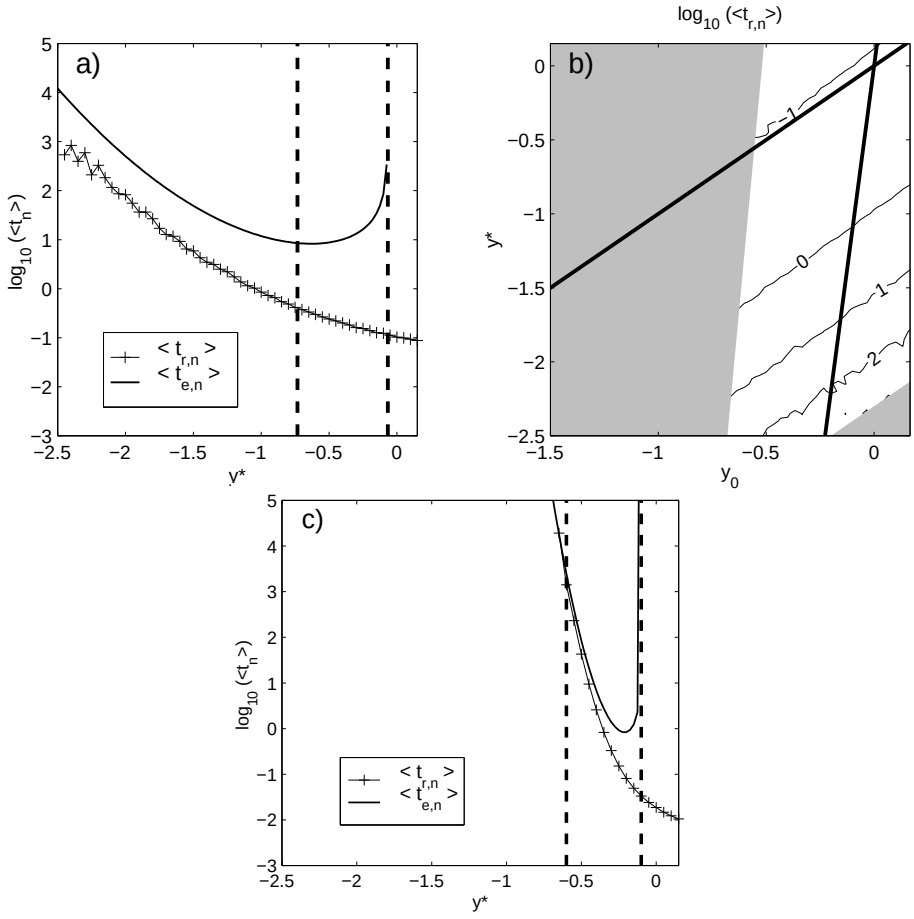

Figure 6: *(a) Comparison of the mean residence time in the nonconvecting regime* $\langle t_{r,n} \rangle$ *(logarithmically scaled, time unit is one year) from simulations (crosses) with the analytically computed mean escape time* $\langle t_{e,n} \rangle$ *(solid) as a function of* y^* *for fixed* $y_0 = -0.067$*. The dashed lines enclose the bistable domain. (b) Contour plot of the mean residence time (logarithmically scaled) as a function of the two model parameters* (y_0, y^*)*. Due to the finite length of the model simulations the contours are not perfectly smooth and could not be computed for the whole parameter plane. Shading indicates where they miss. The thick lines are the deterministic stability borders as in Fig. 2. For panels (a) and (b) the tutorial parameters were used. (c) shows the same comparison like in (a), but with other parameters:* $k_S = 10, k_c = 50, \sigma = 0.5$*, and* $y_0 = -0.1$*.*

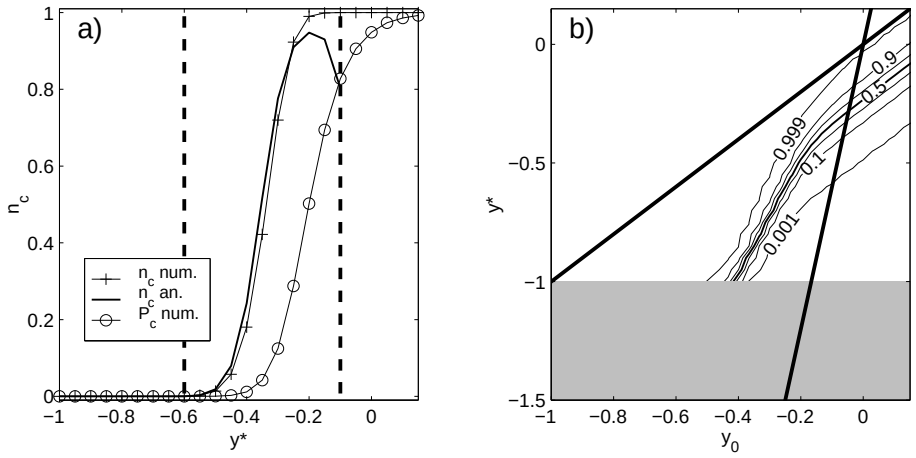

Figure 7: *(a) Comparison of numerical (crosses) and analytical (thick line) computation of n_c with P_c (circles). The parameter values are $k_S = 10, k_c = 50, \sigma = 0.5$, and $y_0 = -0.1$. The vertical dashed lines denote the boundaries of deterministic bistability. The analytical approximation of n_c breaks down where the analytical $\langle t_{e,n} \rangle$ has its minimum (cf. Fig. 6c). (b) Contours of numerically simulated n_c for the same parameter values, with varying y_0. For small y^* values the contours follow the $\langle t_{r,n} \rangle$ contours (cf. Fig. 6b), or the isolines of the potential function of the "off" regime.*

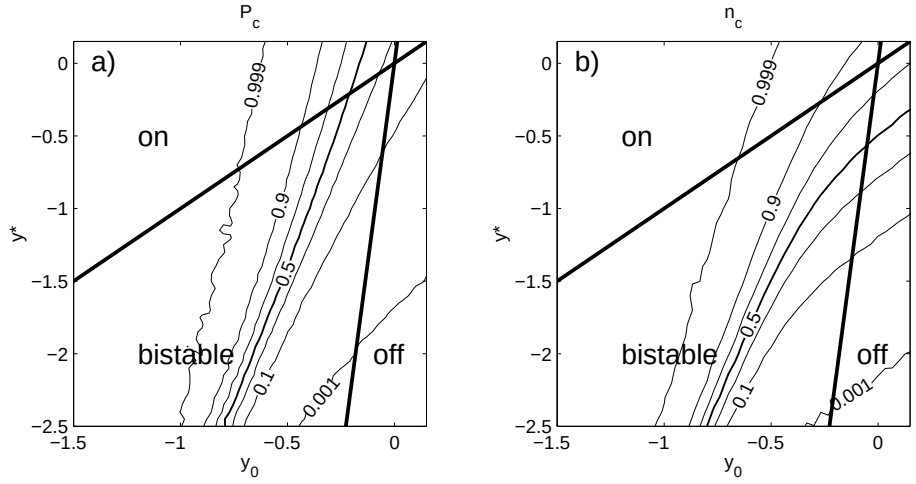

Figure 8: *Estimates of (a) the probability P_c to be in the convecting regime and (b) the probability n_c for a convecting year from numerical simulations of the 1S model. Contours show P_c and n_c as a function of the forcing parameters y^* and y_0, using the tutorial parameters. Thick lines denote the deterministic stability domain borders as in previous figures. Note that the contours run smoothly through these deterministic stability borders.*

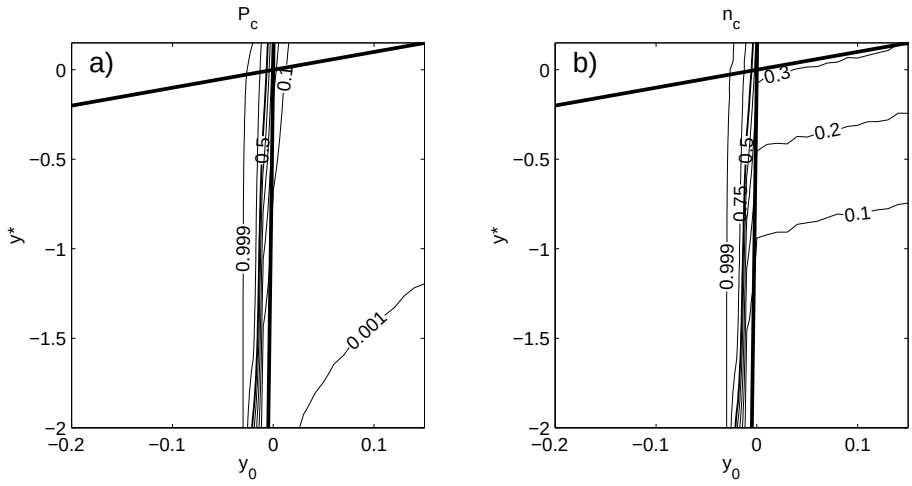

Figure 9: *Contours of (a) P_c and (b) n_c for the estimated parameter set. The difference between P_c and n_c is still significant, particularly in the "off" domain ($y_0 > 0$).*

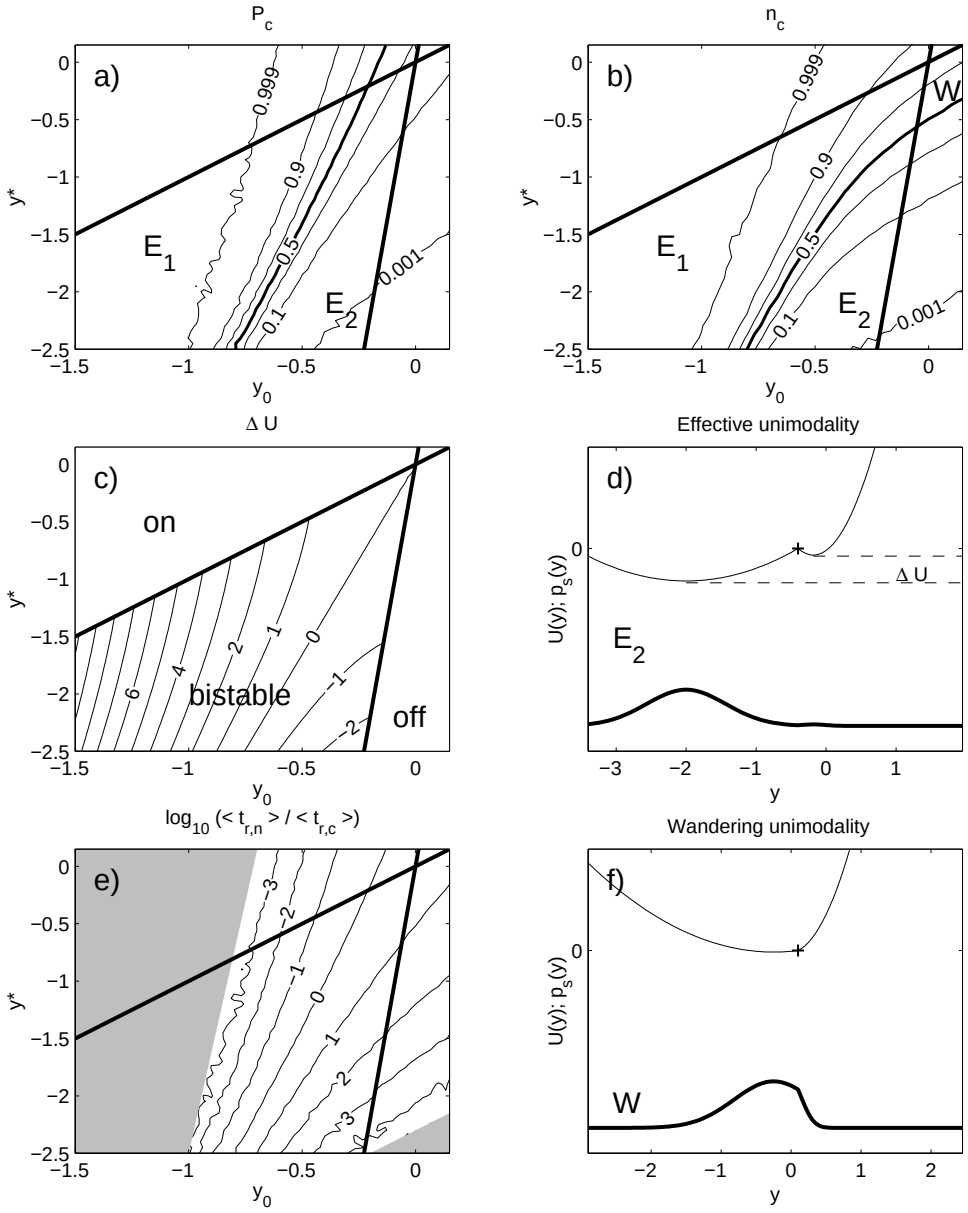

Figure 10: *Stochastic stability measures, effective unimodality, and wandering unimodality. Panel (a) and (b) show P_c and n_c like in Fig. 8. They are compared with two other stability measures: the potential well depth difference (c) and the logarithm of the ratio of the mean residence times (e). The tutorial parameters are used for all panels. Shading in (e) is as in previous figures. Panels (d) and (f) show the potential (thin) and the pdf (thick) as a function of y, with y^* and y_0 corresponding to the position of the letters E_2 and W in (b). Two pdf peaks with strongly different size lead to effective unimodality (d), whereas a single pdf peak that leaks into the other regime is associated with wandering unimodality (f).*

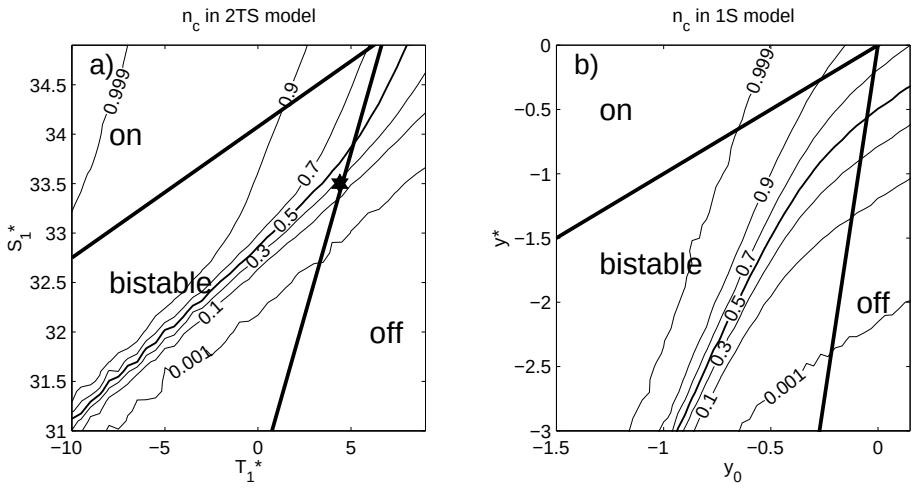

Figure 11: *Comparison of the n_c values for (a) the 2TS model and (b) the 1S model. Parameter values are for 1S the tutorial ones, and for 2TS the optimal parameter set as determined in K01, where T_1^* and S_1^* vary. The asterisk in (a) denotes the position of the optimal parameters.*