Clim Dyn
DOI 10.1007/s00382-007-0290-5

The dynamics of learning about a climate threshold

Klaus Keller · David McInerney

Received: 27 July 2006 / Accepted: 30 May 2007

Abstract Anthropogenic greenhouse gas emissions may trigger threshold responses of the climate system. One relevant example of such a potential threshold response is a shutdown of the North Atlantic meridional overturning circulation (MOC). Numerous studies have analyzed the problem of early MOC change detection (i.e., detection before the forcing has committed the system to a threshold response). Here we analyze the early MOC prediction problem. To this end, we virtually deploy an MOC observation system into a simple model that mimics potential future MOC responses and analyze the timing of confident detection and prediction. Our analysis suggests that a confident prediction of a potential threshold response can require century time scales, considerably longer that the time required for confident detection. The signal enabling early prediction of an approaching MOC threshold in our model study is associated with the rate at which the MOC intensity decreases for a given forcing. A faster MOC weakening implies a higher MOC sensitivity to forcing. An MOC sensitivity exceeding a critical level results in a threshold response. Determining whether an observed MOC trend in our model differs in a statistically significant way from an unforced scenario (the detection problem) imposes lower requirements on an observation system than the determination whether the MOC will shut down in the future (the prediction problem). As a result, the virtual observation systems designed in our model for early detection of MOC changes might well fail at the task of early and confident prediction. Transferring this conclusion to the real world requires a considerably refined MOC model, as well as a more complete consideration of relevant observational constraints.

K. Keller (✉) · D. McInerney
Department of Geosciences, Penn State University,
208 Deike Building, University Park, PA 16802-2714, USA
e-mail: kkeller@geosc.psu.edu

1 Introduction

The geological record shows that the North Atlantic Meridional Overturning Circulation (MOC) can respond abruptly to climate forcing (Alley et al. 2003; McManus et al. 2004). Model simulations suggest that anthropogenic greenhouse gas emissions might trigger such a threshold response in the future (Fichefet et al. 2003; Marsh et al. 2004; Stocker 1999). The projected economic and ecological impacts of an MOC threshold response are uncertain, but potentially nontrivial (Keller et al. 2000; Link and Tol 2006; Nordhaus 1994). An early prediction of a potential threshold response with undesirable consequences has the potential to improve the design of risk management strategies and can have considerable economic value (Keller et al. 2004; Peterson et al. 2003).

Current predictions about the long-term fate of the MOC are deeply uncertain (Cubasch et al. 2001; Keller et al. 2007b; Lempert 2002; Wunsch 2006; Zickfeld et al. 2007). Some models predict a considerable anthropogenic weakening, while other models predict a virtually insensitive MOC (Gregory et al. 2005; Latif et al. 2000). This divergence in model predictions is due to differences in representations of positive and negative feedbacks (Stouffer et al. 2006). Expert elicitations yield diverging estimates of the MOC sensitivity to anthropogenic forcing and suggest a sizeable probability that an MOC shutdown could be triggered within this century (Zickfeld et al. 2007). Reducing this predictive uncertainty requires improving MOC

models, gathering new observations, and assimilating additional observational constraints. Here we use a simple MOC model to analyze whether and how new MOC observations might enable the confident and early prediction of a potential MOC threshold response.

Previous studies analyzing the tasks of MOC change detection and prediction have broken important new ground, but are silent on a range of questions. Most assessments of potential MOC observation systems have focused on the detection problem (Baehr et al. 2007a; Hu et al. 2004; Keller et al. 2007a; Santer et al. 1995; Vellinga and Wood 2004). These studies characterize potential observation systems by virtually deploying them into models and then analyze the simulated signals. These studies suggest that high frequency (approximately annual) observations that would result in MOC reconstructions with relatively high precision (errors ≤ 2 Sv; 1 Sv = 10^6 m^{-3} s^{-1}) would enable a statistically significant ($p < 0.05$) detection of anthropogenic MOC changes within a few decades. One possible example of such a system is the MOC observation array recently implemented at 26°N (Marotzke et al. 2002; Baehr et al. 2007a). Detection asks whether the observed trends in the past are outside the range of natural variability. Detection of past changes is not equivalent to a confident prediction of a future MOC threshold crossing.

Climate change prediction can be grouped into two tasks (Lorenz 1975): (1) predicting the trajectory of the climate system (predictions of the first kind) and (2) predicting the statistical properties of an average climate state (predictions of the second kind). Predictions of the first kind face the challenge that the climate system is chaotic (Lorenz 1963, 1975). In a chaotic system, small uncertainties about the initial conditions can result in large variations in the long-term state. MOC predictions of the first kind lose skill on a decadal time scale (Collins and Sinha 2003; Griffies and Bryan 1997; Knight et al. 2005; Lohmann and Schneider 1999). MOC predictions of the second kind may be skillful over a longer time scale, as they are concerned only with an average climate state (Cubasch et al. 2001; Lorenz 1975). We are interested in MOC predictions of the second kind. Specifically, we analyze how the average state of the MOC changes in the long term as a function of external forcing. In particular, we are interested in the likelihood that a given external forcing will trigger a future MOC shutdown. The few studies that have derived MOC predictions of the second kind (e.g., Challenor et al. 2006; Knutti et al. 2003) do not analyze the timing of confident and early prediction of a potential MOC threshold response.

We use a simple MOC model to construct a temporal pattern (specifically, a timeseries) of the expected MOC strength as a function of external temperature forcing. We simulate different hypothetical MOC observation systems, accounting in a simplified manner for unresolved internal variability and observation errors. We then assess the ability of these virtual observations to inform the tasks of detecting anthropogenic MOC changes and predicting a potential anthropogenic MOC threshold response. The detection method follows previous work pioneered by Santer et al. (1995) and refined by Baehr et al. (2007a). The prediction method uses a Bayesian data assimilation technique (Hastings 1970; Metropolis et al. 1953). Following Santer et al. (1995), we define detection time as the first time an anthropogenic change has been detected with statistical significance ($p < 0.05$). Prediction time is defined as the first time when 95% of the predictions about a potential MOC shutdown are correct. We explore the implications of choosing a different quality criterion (e.g., 50 or 80%) and analyze the effects of different prediction skills in a model of decision-making under uncertainty.

Our theoretical study uses a simple model to analyze the question how a hypothetical observation system may inform detection and prediction. Using this model, we show that early and confident prediction of a MOC threshold response can impose considerably higher requirements on an observation system compared to the task of confident and early detection. This is because the early warning signs of an approaching threshold response can be subtle compared to the signs of anthropogenic MOC changes. Enhancing the hypothetical observation system by adding observations that would provide an overall improved signal-to-noise ratio (cf. Hasselmann 1993; Baehr et al. 2007b; Brennan et al. 2007) would enable an earlier prediction in our simple model. Our analysis is silent on the, so far, open question of which observation and prediction system would likely provide a reliable and early prediction of a potential MOC shutdown not just in a simple model but in the real world (Keller et al. 2007a, b).

2 Model

We use a simple box-model to mimic the response of the MOC to climate change (Zickfeld et al. 2004). This model is an extension of Stommel's (1961) pioneering study, and has three important advantages. First, the model is among the simplest possible choices for representing a potential anthropogenic MOC threshold response (Marotzke 2000). Second, the model parameters have been tuned to approximate important properties (the unperturbed MOC intensity and the location of the MOC bifurcation point with respect to freshwater forcing) of a more complex model (Ganopolski and Rahmstorf 2001). Third, the model is computationally efficient, thus allowing the use of a data assimilation technique that does not require potentially

biasing assumptions about the underlying probability density functions or the convexity of the likelihood functions (Kim et al. 2003; Warnes 2001). The model is described in Zickfeld et al. (2004). Here we give a brief conceptual overview.

The model consists of four well-mixed and interconnected boxes representing the northern, deep, southern, and tropical Atlantic oceans. These boxes are interconnected by advective flows, allowing for an MOC pattern. Atmospheric boundary conditions drive surface fluxes of heat and freshwater following a restoring condition. The MOC intensity is a function of the density difference between the northern and southern Atlantic. Water densities are a function of temperature and salinity, which are affected by the intensity of the MOC and the surface fluxes of heat and freshwater. The model is forced by a linear increase in global mean surface temperatures of 4.5°C over a period of 150 years with a subsequent stabilization. The global mean temperature changes are translated into the boundary conditions for the different boxes using a linear pattern scaling approach (Zickfeld et al. 2004). Atmospheric warming affects the MOC by increasing heat and freshwater fluxes into the surface boxes. One key parameter that influences the sensitivity of the MOC to climate change is the North Atlantic hydrological sensitivity (h, in units of Sv $°C^{-1}$). The hydrological sensitivity is the rate at which freshwater fluxes into the North Atlantic increase as the average surface temperature in the Northern Hemisphere rises.

We use this box-model to approximate a potential response pattern of the MOC intensity to anthropogenic forcing. However, the signal one might expect from an MOC observation system has two additional properties affecting the detection and prediction task. First, the box-model does not reproduce the interannual variability observed in more complex atmosphere–ocean general circulation models (AOGCM) (e.g., Delworth and Mann 2000). The MOC intensity in AOGCMs can show rather complex variability (e.g., Zhu et al. 2006; Delworth and Mann 2000; Dai et al. 2005; Holland et al. 2000). The power spectrum on seasonal to annual timescales is typically constant, then increases with increasing time-scales (e.g., Holland et al. 2000, Zhu et al. 2006). In addition, the MOC in several AOGCMs shows a distinct multidecadal variability (e.g., Dai et al. 2005; Delworth and Greatbatch 2000). Furthermore, the MOC variability in model simulations can be affected by forcing (e.g., Holland et al. 2000; Kleinen et al. 2003; Held and Kleinen 2004). Resolving this complex behaviour of the MOC variability would require an Earth System Model. A statistically sound assessment of potential MOC observation system using an Earth System Model and resolving the effects of uncertain initial conditions is beyond the scope of this conceptual study. We hence approximate the MOC variability by a simple time series model.

Second, the MOC box model does not resolve the fact that MOC observations are subject to considerable observation errors (Ganachaud 2003). We approximate the effects of unresolved internal variability and observations errors by superimposing autocorrelated noise onto the box-model predictions. Specifically, we use a time series model to approximate potential observations of the MOC strength at year t as:

$$y_t = M_t(\theta, F) + u_t, \tag{1}$$

where $M_t(\theta, F)$ is the smooth MOC strength computed from the boxmodel as a function of the model parameters (θ) and forcing (F). The error term (u_t) is approximated by:

$$u_t = \rho u_{t-1} + \varepsilon_t. \tag{2}$$

This error term is autocorrelated noise following a first-order autoregressive model, where ρ is the autocorrelation coefficient and ε_t is a random variable independently drawn from a normal distribution with zero mean and a standard deviation of σ. We choose these parameters to (admittedly crudely) approximate the potential properties of future MOC observations. The autoregressive properties of the MOC intensity in two atmosphere–ocean general circulation models (AOGCM) may provide some useful guidance in this respect.

Fitting a first-order autoregressive time series model to the annual values for the MOC intensity in the control runs described in Manabe and Stouffer (1994) and Roeckner et al. (2003) using the method described by Schneider and Neumaier (2001) results in ρ values of 0.8 and 0.3, and σ values of 0.5 and 1 Sv, respectively. The simple autoregressive time series model approximates the MOC variability in the two much more complex general circulation models reasonably well (Fig. 1). Of course, the simple autoregressive model does not capture potentially important oscillations. Detecting potential multi-decadal to century scale oscillations in the observational record poses nontrivial challenges.

To approximate the additional effect of observation errors (which result in a decreased ρ and an increased σ) we choose $\rho = 0.5$ and $\sigma = 1.5$ Sv as a base case and perform an extensive sensitivity study with respect to these parameters. It is important to stress that the resulting time series is a very simplified approximation to the likely much more complex signal provided by potential real MOC observation systems. The objective is simply to analyze the detection and prediction problem in a simple model.

Examples of the simulated MOC trajectories for different temperature forcing scenarios and hydrological sensitivities are shown in Fig. 2a. Without temperature forcing, the model MOC fluctuates about its initial strength

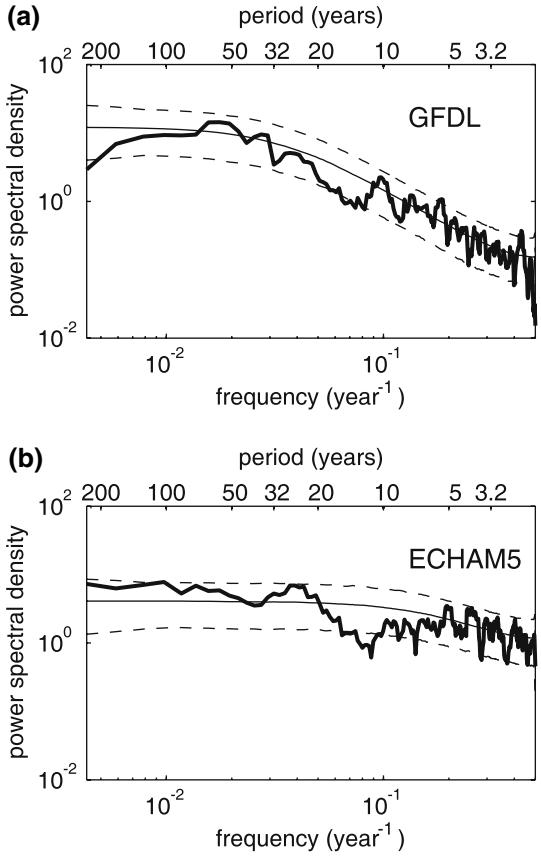

Fig. 1 Power spectral densities of the MOC strength time series from 467-years of the control runs of the **a** GFDL (Manabe and Stouffer 1994) and **b** ECHAM5 (Roeckner et al. 2003) models (*thick lines*). The spectra were computed using the Thomson multitaper method. Also depicted in each plot are the mean spectrum (*thin line*) and 95% confidence interval (*dashed lines*) for 104 realizations of an AR(1) model fit to the two control runs ($\rho = 0.8$, $\sigma = 0.5$ Sv for GFDL; $\rho = 0.3$, $\sigma = 1$ Sv for ECHAM5)

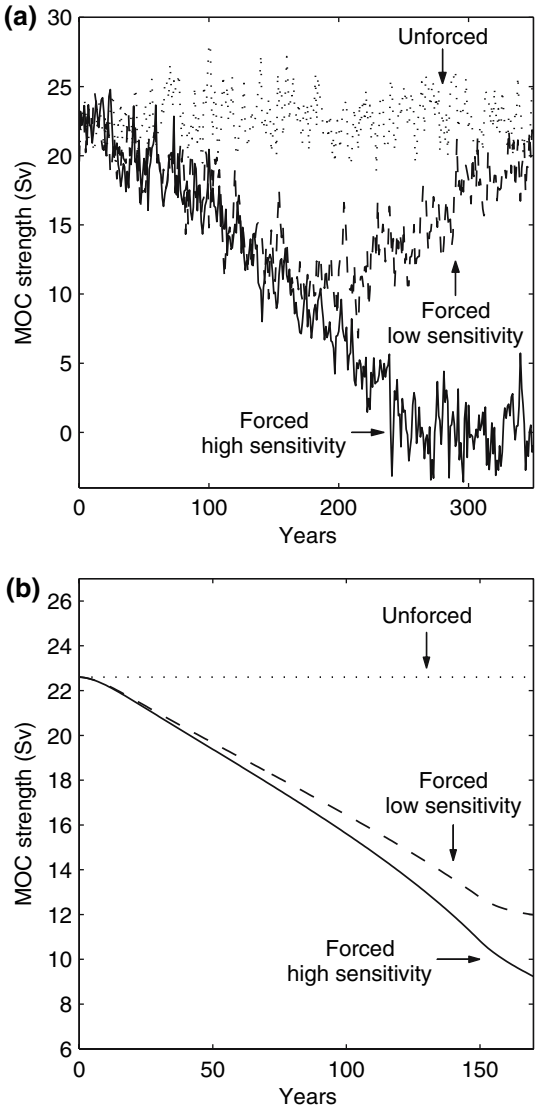

Fig. 2 Examples of unforced and forced MOC model responses. "Low" and "high" sensitivity refer to North Atlantic hydrological sensitivities of 0.04 and 0.05 Sv °C^{-1}, respectively. The *upper panel* (**a**) shows a long-term scenario for the expected observation signal with the superimposed errors due to observation uncertainties and unresolved internal variability (Eq. 2). The *lower panel* (**b**) shows the same scenarios for a shorter duration and without the superimposed errors

of 23 Sv. With forcing, the MOC weakens. The long-term MOC state depends on the extent of the forcing and on the value of the hydrological sensitivity in this analysis. For a hydrological sensitivity of $h = 0.04$ Sv °C^{-1} the MOC weakens but eventually returns to its initial state. For a hydrological sensitivity of $h = 0.05$ Sv °C^{-1} the MOC shuts down and does not recover. Note that the critical value beyond which the MOC shuts down depends on the temperature forcing. Our analysis neglects this effect, as we consider only a single temperature-forcing scenario.

In the model, virtual observations taken before the forcing threshold can be used to improve the predictions of the MOC threshold response. This is because observing the rate of MOC decrease in the model provides information about the hydrological sensitivity, which in turn contains information about the long-term fate of the MOC. A higher observed rate of MOC weakening implies a higher hydrological sensitivity (Fig. 2b). A hydrological sensitivity exceeding a critical value results in an MOC shutdown in the model. The relationship between the rate of MOC weakening and the likelihood of an MOC shutdown seen in this simple model is also seen in more complex models (Rahmstorf and Ganopolski 1999; Webster et al. 2006). It should be noted, however, that our conceptual study neglects several key sources of uncertainties (e.g., about the parameter uncertainties besides the hydrological sensitivity or the appropriate the model structure) that can play an important role in detection and prediction (Hargreaves and

Annan 2006; Marsh et al. 2004; Stouffer et al. 2006; Keller et al. 2007a).

3 Methods

We choose detection and prediction methods to mimic the currently used methods in the literature (Baehr et al. 2007a; Hargreaves and Annan 2006; Keller et al. 2007a; Santer et al. 1995).

3.1 Detection

We adopt the detection method developed by Baehr et al. (2007a). A control run (a simulation without anthropogenic forcing, see Fig. 2a) is used to approximate the uncertainty about the MOC trends given the available virtual observations. We use a bootstrap technique to estimate the uncertainties in the linear least-squares MOC trends. The probability density function (PDF) of this bootstrap sample defines the bounds of the 95% confidence region for the slopes derived from the virtual observations. Trends in the forced signal are then estimated for virtual observation periods beginning at the start of the forcing and continuing for a specific virtual observation period. The null hypothesis is that the MOC slope is zero, as it is the case for the unforced MOC (Fig. 2). The bootstrap analysis tests this null hypothesis given the virtual observations.

3.2 Prediction

We use a Bayesian data assimilation method to evaluate the ability of hypothetical observations to inform MOC predictions. A Bayesian approach combines prior information about model parameters with the information contained in observations to derive posterior parameter estimates and the associated probabilistic predictions.

We analyze learning about a single (arguably key) parameter, the hydrological sensitivity. Considering a larger number of uncertain parameters would likely result in more uncertain predictions, later prediction times, and hence likely strengthen our forthcoming conclusions. We will return to the effects of this and other methodological choices in the section "caveats", below. Uncertainties about the hydrological sensitivity are key drivers of the uncertainties in MOC predictions (Dixon et al. 1999; Ganopolski and Rahmstorf 2001; Zickfeld et al. 2004, Gregory et al. 2005). Estimates for this parameter differ greatly. Ganopolski and Rahmstorf (2001) use a value of $h = 0.013$ Sv °C^{-1}, while the model of Manabe and Stouffer (1994) has a stronger hydrological response with $h = 0.053$ Sv °C^{-1}. Zickfeld et al. (2004) examine a range of the hydrological sensitivity between 0 and 0.06 Sv °C^{-1}. We adopt a uniform prior bounded by this range. The resulting likelihood of an MOC shutdown in the model is then approximately 25%. This likelihood estimate is higher than the assessment derived from AOGCM runs, but within the range spanned by expert elicitations and results from Earth System Models of Intermediate Complexity (Challenor et al. 2006; Rahmstorf and Zickfeld 2005; Zickfeld et al. 2007).

We implement the Bayesian assimilation using a modified form of the Metropolis–Hastings algorithm (Hastings 1970; Metropolis et al. 1953). This algorithm is slightly more complex than alternative assimilation methods such as Bayes Monte Carlo (Sohn et al. 2000), but also numerically more efficient (Qian et al. 2003). The algorithm constructs a Markov chain, which converges to the posterior PDF. The Markov chain begins with an initial h value chosen from the prior range. The Markov chain is then constructed using the Metropolis–Hastings algorithm. We select a candidate value h^* by perturbing the current h value. Specifically, h^* is sampled from the uniformly distributed generating function

$$U[\max(h_L, h - \Delta h), \min(h + \Delta h, h_U)], \tag{3}$$

where h_L and h_U are the lower and upper bounds of the prior distribution ($h_L = -\infty$ and $h_U = \infty$ for unbounded priors, e.g., normally distributed priors), and Δh is the maximum step-size. This candidate is accepted as the next element in the chain with a probability given by:

$$\alpha = \min\left(\frac{p(h^*|y)\ q(h^*, h)}{p(h|y)\ q(h, h^*)}, 1\right), \tag{4}$$

where $q(h, h^*)$ is the probability of jumping from h to h^*, based on the generating distribution (3), and $p(h|y)$ is the probability density of hypothesis h given observations y (explained below). The ratio of the transition probabilities between the candidate and the initial value is

$$\frac{q(h^*, h)}{q(h, h^*)} = \frac{\min(h + \Delta h, h_U) - \max(h_L, h - \Delta h)}{\min(h^* + \Delta h, h_U) - \max(h_L, h^* - \Delta h)}. \tag{5}$$

For an unbounded prior, this ratio implies that $q(h^*, h) = q(h, h^*)$, and the original Metropolis algorithm (Hastings 1970; Metropolis et al. 1953) is recovered. Note that random-walk Metropolis algorithms that are not designed to restrict the candidate value to within the bounds of the prior distribution (e.g., Harmon and Challenor 1997) can result in posterior PDFs that contain biases near the edges of the posterior distribution. Our approach considerably improves the assimilation results for simple test cases with uniform priors (results not shown).

The probability densities $p(h|y)$ in Eq. (4) are determined using Bayes' rule:

$$p(h|y) = \frac{p(y|h)p(h)}{\int_{h_l}^{h_u} p(y|x)p(x)\mathrm{d}x}. \quad (6)$$

The probability density $p(y|h)$ can be considered to be a function of h, rather than y and is referred to as the likelihood function of h given y, $L(h|y)$ (Box and Tiao 2002). We derive the likelihood function from Bence (1995) for known σ and ρ (*cf.* Eq. 2) as:

$$L(h|y) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[(1-\rho^2)(y_1 - M_1)^2 + \sum_{t=2}^{N}[y_t - M_t - \rho(y_{t-1} - M_{t-1})]^2\right]\right\}, \quad (7)$$

where M_t is the MOC strength computed using the box-model. The ratio of probability densities in Eq. (4) is then given by:

$$\frac{p(h^*|y)}{p(h|y)} = \frac{L(h^*|y)}{L(h|y)}, \quad (8)$$

for a uniformly distributed prior.

The process for selecting and accepting elements in the Markov chain is repeated until the chain has converged to a steady state. We choose a step-size that results in an acceptance rate of approximately 25% (Harmon and Challenor 1997). We follow Raftery and Lewis (1995) in determining the number of iterations required for convergence and the number of initial "burn-in" iterations to be discarded from the chain to remove the effect of initial conditions. The assimilation method succeeded in recovering a known, analytical solution from a simple test problem. Previous tests of the algorithm (Keller et al. 2007c) show that repeated assimilation experiments with independent random initial conditions converge to virtually indistinguishable posterior estimates. The elements from the Markov chain are then used to derive the probabilistic MOC predictions as a function of the available information.

4 Results and discussion

We first illustrate the relative timing of detection and prediction for a single realization of the MOC variability. We begin by constructing a time series of MOC observations from the "high sensitivity" simulation in Fig. 2a. In this case, the North Atlantic hydrological sensitivity is assumed to be 0.05 Sv °C^{-1}, and the MOC shuts down. Prior to any observations (i.e., just using the prior information) the estimate of the hydrological sensitivity is given by the prior PDF and follows a uniform distribution bounded by 0 and 0.06 Sv °C^{-1} (Fig. 3a).

The predictions about an MOC shutdown associated with this prior are correct in roughly 25% of the cases. The remaining cases (roughly 75%) make the wrong prediction (i.e., they predict no shutdown) (Fig. 3 b). The prediction skill based on the rather diffusive prior is rather poor. Assimilating 100 years of observations into the model shifts the probability mass of h toward the correct (assumed) value, and the predicted probability of an MOC shutdown rises to 80%. After observing the MOC for 150 years, the PDF for h has tightened considerably around the true value of 0.05 Sv °C^{-1}, and the predicted probability of an MOC shutdown is 95%. A reliable prediction of the MOC shutdown (i.e., when 95% of the MOC shutdown predictions are correct) is achieved for the first time in the previous model year. In contrast, a statistically significant change in the MOC is detected after approximately 30 years; more than 100 years prior to the prediction time.

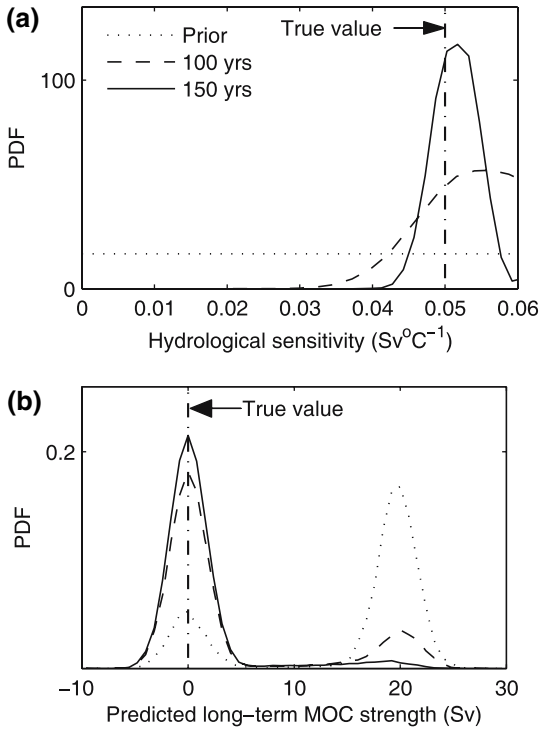

Fig. 3 Probability density functions (*PDFs*) for **a** the North Atlantic hydrological sensitivity, and **b** the predicted long-term MOC strength (after 350 years). Shown are estimates for the North Atlantic hydrological sensitivity prior to observing and after 100 and 150 years of observations. The true value in the analyzed case is 0.05 Sv °C^{-1}. The slight discrepancy between the "true" value (*vertical line*, panel **a**) and the modes of the parameter estimates stem from the superimposed random errors

Neglecting the effects of autocorrelation in the model residuals can lead to overconfident parameter estimates and model predictions (Zellner and Tian 1964). This effect can be sizeable, as demonstrated by the comparison of the posterior estimates of the hydrological sensitivity after 125 years of observations with and without a correction for the autocorrelation effects (Fig. 4). Conclusions of previous studies neglecting these autocorrelation effects (e.g., Hargreaves and Annan 2002) may hence need to be reevaluated.

We now ask whether this observation system would enable a confident and early prediction. To answer this question, we first determine the time when the temperature-forcing threshold is crossed (this depends on the rate at which the temperature forcing is changing). We determine this time for our model using a sensitivity study with respect to the duration of the forced temperature increase. If the temperature forcing continues past the forcing threshold, the MOC shuts down in this simple model (Fig. 5). For the high sensitivity simulation in Fig. 2 the forcing threshold occurs at 143 years; therefore MOC changes are detected, but not predicted, before the forcing threshold has been crossed.

The detection and prediction times are random variables as they are affected by random variability (Eqs. 1 and 2). We analyze the effects of this variability by generating 1,000 MOC realizations with the same parameter value for the hydrological sensitivity (h = 0.05 Sv °C^{-1}), but with different random realizations of the noise (Eq. 2). Prediction times range from roughly 75 to 220 years, with a median around 160 years (Fig. 6). Detection times, in comparison, are much shorter, with a median around 30 years. For each realization in the considered ensemble detection occurs before prediction. For each realization a change in the MOC is detected prior to crossing the forcing threshold, while an early and confident MOC prediction occurs in less than 40% of the cases.

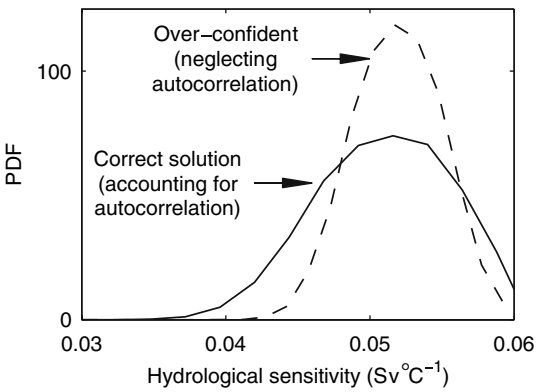

Fig. 4 Illustration of the overconfidence in parameter estimates introduced by neglecting the effects of autocorrelated model residuals. Shown are estimates for the hydrological sensitivity after 125 years with (*solid line*) and without (*dashed line*) accounting for the autocorrelation effects in the likelihood function

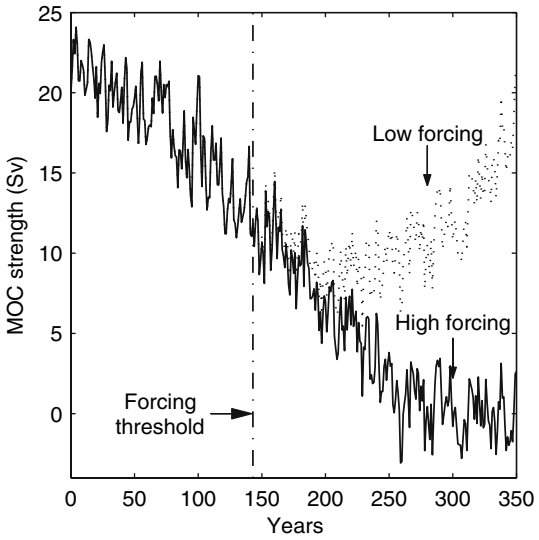

Fig. 5 Determination of the threshold crossing time. A scenario of a linear increase in global mean temperature over 140 years results in a weakening MOC that eventually recovers. Extending this forcing period to 150 years results in an MOC shutdown. The threshold crossing time for this specific example is 143 years

The characterization of a "sufficient" prediction skill depends on the use of the prediction information (Adams et al. 1984; Simon 1959). So far, we have characterized the skill of the prediction by the time at which 95% of the predictions about an MOC shutdown are correct. This cut-off for a critical prediction skill is motivated by the objective of minimizing the false alarm rate, i.e., predicting

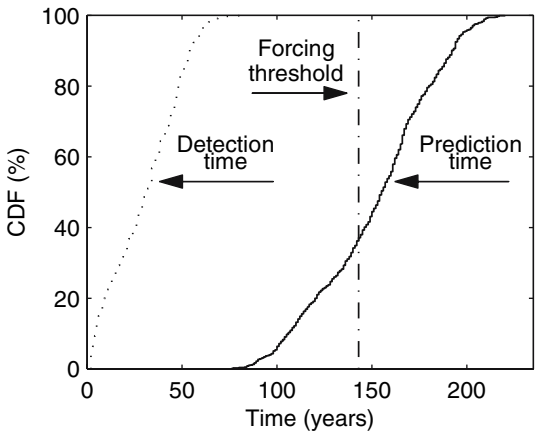

Fig. 6 Cumulative density functions (*CDFs*) for detection and prediction times using a North Atlantic hydrological sensitivity of 0.05 Sv °C^{-1}. The forcing threshold is the time when global temperature must level off to a constant value to avoid an MOC shutdown (cf. Fig. 3). Simulations assume variability with a lag–1 autocorrelation coefficient of 0.5 and independent errors of 1.5 Sv, as shown in Fig. 2a

an MOC shutdown when it is not happening. The estimated prediction time is sensitive to the choice of this arbitrary cut-off (Fig. 7). Accepting a higher rate of false alarms results in earlier prediction times.

There is a trade-off between the false alarm rate and the rate at which the prediction misses the event (Swets 1973). Both errors incur risks depending on the associated (and not necessarily equal) impacts. A decision-maker might want to balance these risks. One typical approach to analyzing this issue is to consider an extremely simplified decision problem where a binary choice is either to take or not to take action to avoid an undesirable realization of a binary outcome (Wilks 2001). The protective action in this case could be interpreted as reducing anthropogenic climate forcing (at cost C) and the undesirable outcome would be an MOC shutdown (at loss L). A strategy to minimize the expected costs would take the protective action if the likelihood of the undesirable outcome exceeds a critical value, the cost–loss ratio, defined as C/L. For a hypothetical cost–loss ratio of 0.1, the predicted likelihood of an MOC shutdown based on the prior information alone (equal to 0.25 and not requiring any new observation) would suffice to choose the protective action. For cost–loss ratios exceeding 0.25, the dynamics of learning about the potentially approaching climate threshold (Fig. 8) become important in this simple example. For a cost–loss ratio of 0.5 and observations derived from the situation of a high hydrological sensitivity (Fig. 8a), the stylized decision-maker would choose the protective action on average after more than six decades of observations. For reference, the cost–loss ratio based on a simple economic model and a subset of estimated economic damages due to an MOC shutdown is approximately 0.9 (Keller et al. 2007a). It is important to stress that current estimates of this cost–loss ratio are deeply uncertain and hinge additionally on value judgements about intergenerational welfare distribution (Keller et al. 2007a, b).

The time-scale at which these predictions cross the halfway point between the prior based prediction and the true state is on the order of a century. Note that some predictions move early on in the wrong direction. This is due to the stochastic nature of the unresolved variability, which can introduce signals in the wrong direction. In the long run, when the posteriors have sharpened considerably, this effect weakens and the long-term predictions are tightly constrained.

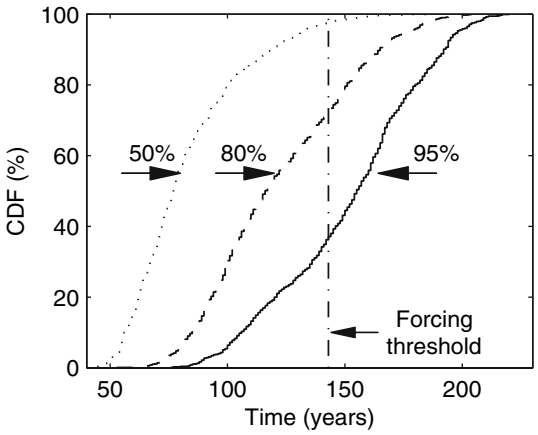

Fig. 7 Sensitivity of the prediction time with respect to the confidence requirement (the probability of the MOC shutdown prediction being correct at the prediction time). Shown are results for 95, 80 and 50% confidence in the predictions. Simulations assume a North Atlantic hydrological sensitivity of 0.05 Sv $°C^{-1}$, and variability with a lag-1 autocorrelation coefficient of 0.5 and independent errors of 1.5 Sv, as shown in Fig. 2a

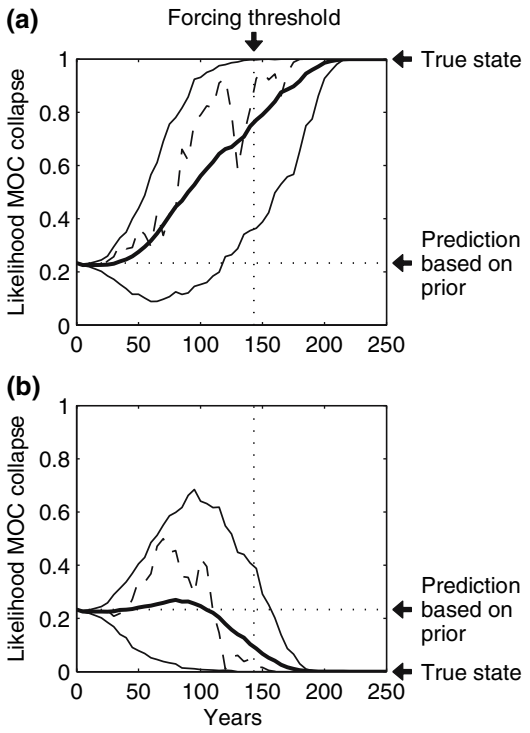

Fig. 8 Dynamics of learning about the likelihood of an MOC threshold response for **a** high ($h = 0.05$ Sv $°C^{-1}$) and (b) low ($h = 0.04$ Sv $°C^{-1}$) hydrological sensitivity. The horizontal line is the prediction based on the prior uncertainty about the hydrological sensitivity. Simulations are for a variability with a lag-1 autocorrelation coefficient of 0.5 and independent errors of 1.5 Sv (as shown in Fig. 2a) and 500 Monte Carlo realizations from the unresolved noise (Eq. 2). The expected values for the predictions are shown by the *thick solid lines*. The 5th and 95th percentiles of the predictions are shown by the *thin solid lines*. The *dashed lines* are the predictions associated with the realizations shown in Fig. 1a. Prediction at 95% confidence occurs when 95% of the predicted probability mass of the potential MOC shutdown are correct (i.e., MOC shutdown or not)

5 Caveats

Our forthcoming conclusions are subject to several caveats. First, the specific detection and prediction times depend on factors such as the adopted forcing, model structure, Bayesian priors on parameter values, the characteristics of the unresolved variability (Eq. 2), or potential biases in MOC reconstructions (Baehr et al. 2007b). Choosing different Bayesian priors, for example to represent the divergent expert assessments of the MOC sensitivity to temperature forcing (Zickfeld et al. 2007), would change the results. Furthermore, our analysis considers, so far, just a single combination of the standard deviation of the uncorrelated noise (σ) and the lag–1 autocorrelation coefficient (ρ). The combination of these parameters that would most accurately approximate a real MOC observation is uncertain. A sensitivity study with respect to these two parameter shows however, that the prediction times in the considered model are consistently above the detection times (Fig. 9). The detection and prediction times generally increase with increasing autocorrelation coefficients and standard deviation of the unresolved variability (Fig. 9). The ratio of prediction to detection time ranges approximately between 3 and 5.

Second, the adopted model and boundary conditions are very crude approximations of the true MOC response. The model lacks, for example, interactions between forcing, internal MOC variability and potential MOC threshold responses (e.g., Timmermann and Lohmann 2000; Held and Kleinen 2004). This effect could be represented by adding stochastic variability to the MOC model. Resolving the interactions between forcing, internal MOC variability, and a potential MOC shutdown would change the prediction times (Held and Kleinen 2004; Monahan 2002).

Third, our approximation of the prediction problem ignores likely important sources of uncertainties. For example, the adopted perfect model scenario neglects structural uncertainties (Draper 1995) and many sources of stochastic variability (e.g., Sato et al. 1993). A case in point is the neglected structural uncertainty introduced by potential long-term MOC oscillations. Representing these potential oscillations would complicate the detection and prediction tasks. We hypothesize that considering these uncertainties would likely result in delayed prediction times. A quantification of the effects of potential multi-decadal to century-scale oscillations on detection and prediction time is beyond the scope of this analysis.

Last, but not least, the adopted model has a very coarse spatial resolution and lacks hydrographic tracers (besides temperature and salinity) that may provide useful constraints. Incorporating additional observations of tracer trends that are mechanistically linked to MOC changes may provide important observational constraints to improve

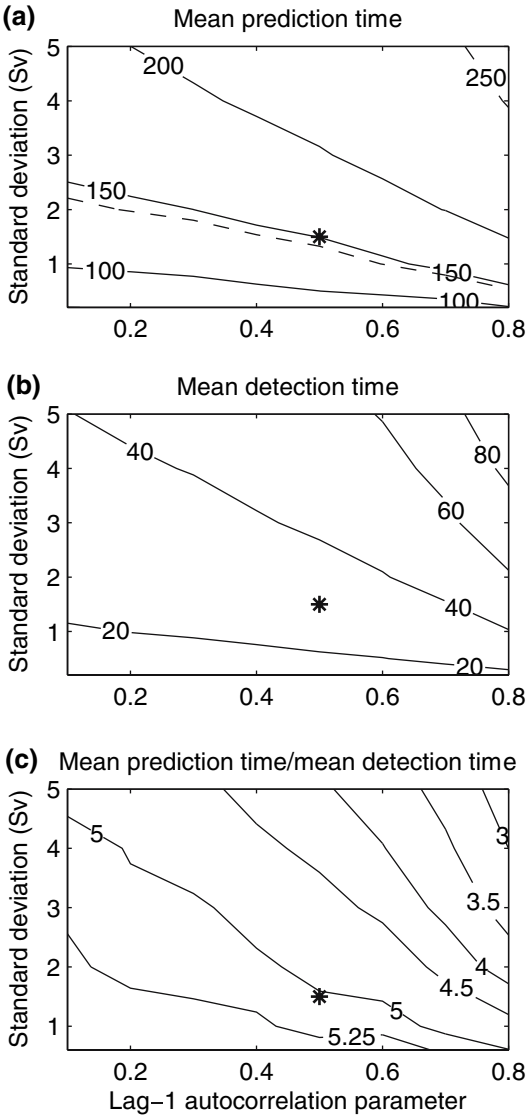

Fig. 9 Sensitivity of **a** the expected prediction time (at 95% confidence), **b** the expected detection time, and **c** the ratio of prediction time to detection time, with respect to the standard deviation (σ) and the autocorrelation coefficient (ρ) of the unresolved internal variability. Simulations assume a North Atlantic hydrological sensitivity of 0.05 Sv $°C^{-1}$. The *dashed contour* in panel **a** represents the forcing threshold

MOC predictions. Potential examples of such tracers include apparent oxygen utilization (Joos et al. 2003; Keller et al. 2002) or chlorofluorocarbons (Smethie and Fine 2001). Considering this information would require the use of more refined models. The associated increase in computational demands would render the applied inversion method practically infeasible. This problem has been approached by approximating the model and/or the underlying likelihood function (Goldstein and Rougier 2006; Hankin 2005; Hargreaves and Annan 2006; Knutti et al. 2003; Lea et al. 2000). Note that the consideration of a bifurcation results in

a multimodal PDF of the response (Fig. 3). In addition, the use of a uniform prior (to represent the uncertainty range given by previous studies) results in posterior PDFs that are poorly approximated by normal distributions. Predictions based on normal (Hargreaves et al. 2004) or linear tangent approximations (Lea et al. 2000) may be considerably biased and overconfident in such situations (Evensen 1997). The MOC predictions derived in this study provide a test case to evaluate these biases.

6 Conclusion

It is important to stress that our analysis is based on a simple model and a subset of the available information about the MOC. Subject to these and the aforementioned caveats, our analysis suggests two main conclusions. First, early prediction of an anthropogenic MOC shutdown (i.e., before it is triggered) imposes considerably higher requirements on the observation system than detecting anthropogenic MOC changes. As a result, observation systems that would detect anthropogenic MOC changes may well fail at the task of early prediction of an MOC shutdown. Second, an MOC observation system that crudely mimics annual estimates of the MOC intensity (a subset of the currently available information) fails in the task of early and confident prediction of an MOC shutdown in an extremely simplified model. Whether the currently implemented MOC observation system would succeed in delivering an actionable early warning sign in the form of a confident early prediction is, at this time, an open question.

Acknowledgments We thank J. Annan, B. Haupt, J. Baehr, D. Ludwig, N. Urban, M. Oppenheimer, and D. Ricciuto for helpful discussions. Nathan Urban calculated the spectra shown in Fig. 1. Any potential remaining errors and omissions are, of course, ours. K. Zickfeld kindly provided the code for the adopted MOC model. Careful reviews by R. Stouffer and H. Held considerably improved the presentation of the manuscript. We gratefully acknowledge support from the National Science Foundation (SES #0345925). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding entity.

References

Adams RM, Crocker TD, Katz RW (1984) Assessing the adequacy of natural science information—a Bayesian approach. Rev Econ Stat 66:568–575

Alley RB, Marotzke J, Nordhaus WD, Overpeck JT, Peteet DM, Pielke RA, Pierrehumbert RT, Rhines PB, Stocker TF, Talley LD, Wallace JM (2003) Abrupt climate change. Science 299:2005–2010

Baehr J, Keller K, Marotzke J (2007a) Detecting potential changes in the meridional overturning circulation at 26 °N in the Atlantic. Clim Change, http://www.dx.doi.org/10.1007/s10584-006-9153-z

Baehr J, McInerney D, Keller K, Marotzke J (2007) Global optimization of an observing system design for the North Atlantic meridional overturning circulation, J Atmos Oceanic Technol (in review). Available at: http://www.geosc.psu.edu/~kkeller/publications.html

Brennan C, Matear R, Keller K (2007) Observing oxygen improves the detection capabilities of an ocean observation array. J Geophys Res Oceans (in review). Available at: http://www.geosc.psu.edu/~kkeller/publications.html

Bence JR (1995) Analysis of short-time series—correcting for autocorrelation. Ecology 76:628–639

Box GEP, Tiao GC (2002) Bayesian inference in statistical analysis. Wiley, New York

Cubasch U et al (2001) Projections of future climate change. Climate change 2001—the scientific basis. Contribution of working group I of the third assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 526–582

Challenor PG, Hankin RKS, March R (2006) Towards the probability of rapid climate change. In: Schellnhuber HJ (ed) Avoiding dangerous climate change. Cambridge University Press, Cambridge

Collins M, Sinha B (2003) Predictability of decadal variations in the thermohaline circulation and climate. Geophys Res Lett 30:1306

Dai AG, Hu A, Meehl GA, Washington WM, Strand WG (2005) Atlantic thermohaline circulation in a coupled general circulation model: unforced variations versus forced changes. J Clim 18(16):3270–3293

Delworth TL, Mann ME (2000) Observed and simulated multidecadal variability in the Northern Hemisphere. Clim Dyn 16:661–676

Delworth TL, Greatbatch RJ (2000) Multidecadal thermohaline circulation variability driven by atmospheric surface flux forcing. J Clim 13(9):1481–1495

Dixon KW, Delworth TL, Spelman MJ, Stouffer RJ (1999) The influence of transient surface fluxes on North Atlantic overturning in a coupled GCM climate change experiment. Geophys Res Lett 26:2749–2752

Draper D (1995) Assessment and propagation of model uncertainty. J Roy Stat Soc Ser B Methodol 57:45–97

Evensen G (1997) Advanced data assimilation for strongly nonlinear dynamics. Monthly Weather Rev 125:1342–1354

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

Ganachaud A (2003) Error budget of inverse box models: the North Atlantic. J Atmos Oceanic Technol 20:1641–1655

Ganopolski A, Rahmstorf S (2001) Rapid changes of glacial climate simulated in a coupled climate model. Nature 409:153–158

Goldstein M, Rougier JC (2006) Bayes linear calibrated prediction for complex systems. J Am Stat Assoc 101:1132–1143

Gregory JM, Dixon KW, Stouffer RJ, Weaver AJ, Driesschaert E, Eby M, Fichefet T, Hasumi H, Hu A, Jungclaus JH, Kamenkovich IV, Levermann A, Montoya M, Murakami S, Nawrath S, Oka A, Sokolov AP, Thorpe RB (2005) A model intercomparison of changes in the Atlantic thermohaline circulation in response to increasing atmospheric CO_2 concentration. Geophys Res Lett 32:L12703

Griffies SM, Bryan K (1997) Predictability of North Atlantic multidecadal climate variability. Science 275:181–184

Hankin RKS (2005) Introducing BACCO, an R bundle for Bayesian analysis of computer code output. J Stat Softw 14(16):21

Hargreaves JC, Annan JD (2002) Assimilation of paleo-data in a simple Earth system model. Clim Dyn 19:371–381

Hargreaves JC, Annan JD (2006) Using ensemble prediction methods to examine regional climate variation under global warming scenarios. Ocean Model 11:174–192

Hargreaves JC, Annan JD, Edwards NR, Marsh R (2004) An efficient climate forecasting method using an intermediate complexity Earth System Model and the ensemble Kalman filter. Clim Dyn 23:745–760

Harmon R, Challenor P (1997) A Markov chain Monte Carlo method for estimation and assimilation into models. Ecol Model 101:41–59

Hastings WK (1970) Monte-Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Hasselmann K (1993) Optimal fingerprints for the detection of time-dependent climate-change. J Clim 6(10):1957–1971

Held H, Kleinen T (2004) Detection of climate system bifurcations by degenerate fingerprinting. Geophys Res Lett 31(23):L23207

Holland MM, Brasket AJ, Weaver AJ (2000) The impact of rising atmospheric CO_2 on simulated sea ice induced thermohaline circulation variability. Geophys Res Lett 27(10):1519–1522

Hu AX, Meehl GA, Han WQ (2004) Detecting thermohaline circulation changes from ocean properties in a coupled model. Geophys Res Lett 31:L13204

Joos F, Plattner G-K, Stocker TF, Körtzinger A, Wallace DWR (2003) Trends in marine dissolved oxygen: implications for ocean circulation changes and the carbon budget. EOS 84:197–204

Keller K, Bolker BM, Bradford DF (2004) Uncertain climate thresholds and optimal economic growth. J Environ Econ Manage 48:723–741

Keller K, Deutsch C, Hall MG, Bradford DF (2007a) Early detection of changes in the North Atlantic meridional overturning circulation: implications for the design of ocean observation systems. J Clim 20:145–157

Keller K, Schlesinger M, Yohe G (2007b) Managing the risks of climate thresholds: Uncertainties and information needs. Clim Change. http://www.dx.doi.org/10.1007/s10584-006-9114-6

Keller K, Miltich LI, Robinson A, Tol RSJ (2007c) How overconfident are current projections of carbon dioxide emissions? Working Paper Series, Research Unit Sustainability and Global Change, Hamburg University. FNU-124: http://www.ideas.repec.org/s/sgc/wpaper.html

Keller K, Slater R, Bender M, Key RM (2002) Possible biological or physical explanations for decadal scale trends in North Pacific nutrient concentrations and oxygen utilization. Deep-Sea-Res II 49:345–362

Keller K, Tan K, Morel FMM, Bradford DF (2000) Preserving the ocean circulation: implications for climate policy. Clim Change 47:17–43

Kim S, Eyink GL, Restrepo JM, Alexander FJ, Johnson G (2003) Ensemble filtering for nonlinear dynamics. Monthly Weather Rev 131:2586–2594

Kleinen T, Held H, Petschel-Held G (2003) The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. Ocean Dyn 53:53–63

Knight JR, Allan RJ, Folland CK, Vellinga M, Mann ME (2005) A signature of persistent natural thermohaline circulation cycles in observed climate. Geophys Res Lett 32:L20708

Knutti R, Stocker TF, Joos F, Plattner GK (2003) Probabilistic climate change projections using neural networks. Clim Dyn 21:257–272

Latif M, Roeckner E, Mikolajewski U, Voss R (2000) Tropical stabilization of the thermohaline circulation in a greenhouse warming simulation. J Clim 13:1809–1813

Lea DJ, Allen MR, Haine TWN (2000) Sensitivity analysis of the climate of a chaotic system. Tellus Series A-Dyn Meteorol Oceanogr 52:523–532

Lempert RJ (2002) A new decision sciences for complex systems. Proc Natl Acad Sci USA 99:7309–7313

Link PM, Tol RSJ (2006) The economic impact of a slowdown of the thermohaline circulation: an application of fund. Working paper, Hamburg University, Department of Economics, Research Unit Sustainability and Global Change, Center for Marine and Atmospheric Sciences (ZMAW), pp 14. Available at http://www.uni-hamburg.de/Wiss/FB/15/Sustainability/Working_Papers.htm

Lohmann G, Schneider J (1999) Dynamics and predictability of Stommel's box model. A phase-space perspective with implications for decadal climate variability. Tellus Series A-Dyn Meteorol Oceanogr 51:326–336

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141

Lorenz EN (1975) Climate Predictability. In: The physical basis of climate modeling, World Meteorological Organization, Washington

Manabe S, Stouffer RJ (1994) Multiple-century response of a coupled ocean–atmosphere model to an increase of atmospheric carbon dioxide. J Clim 7:5–23

Marotzke J (2000) Abrupt climate change and thermohaline circulation: mechanisms and predictability. Proc Natl Acad Sci USA 97:1347–1350

Marotzke J, Cunningham SA, Bryden HL (2002) Monitoring the Atlantic meridional overturning circulation at 26.5 N., Proposal accepted by the Natural Environment Research Council (UK). Available at: http://www.noc.soton.ac.uk/rapidmoc/

Marsh R, Yool A, Lenton TM, Gulamati MY, Edwards NR, Shepherd JG, Krznaric M, Newhouse S, Cox SJ (2004) Bistability of the thermohaline circulation identified through comprehensive 2-parameter sweeps of an efficient climate model. Clim Dyn 23:761–777

McManus JF, Francois R, Gherardi JM, Keigwin LD, Brown-Leger S (2004) Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. Nature 428:834–837

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Monahan AH (2002) Stabilization of climate regimes by noise in a simple model of the thermohaline circulation. J Phys Oceanogr 32:2072–2085

Nordhaus WD (1994) Managing the global commons: the economics of climate change. The MIT press, Cambridge

Qian SS, Stow CA, Borsuk ME (2003) On Monte Carlo methods for Bayesian inference. Ecol Model 159(2–3):269–277

Peterson GD, Carpenter SR, Brock WA (2003) Uncertainty and the management of multistate ecosystems: an apparently rational route to collapse. Ecology 84:1403–1411

Raftery AE, Lewis SM (1995) The number of iterations, convergence diagnostics, and generic Metropolis algorithms. In: Gilks WR, Spiegelhalter DJ, Richardson S (eds) Markov Chain Monte Carlo in Praxis. Chapman and Hall, London, pp 115–130

Rahmstorf S, Ganopolski A (1999) Long-term global warming scenarios computed with an efficient coupled climate model. Clim Change 43:353–367

Rahmstorf S, Zickfeld K (2005) Thermohaline circulation changes: a question of risk assessment—an editorial review essay. Clim Change 68:241–247

Roeckner E, Baeuml G, Bonaventura L, Brokopf R, Esch M, Giorgetta M, Hagemann S, Kirchner I, Kornblueh L, Manzinie E, Rhodin A, Schlese U, Schulzweida U, Tompkins A (2003) The atmospheric general circulation model ECHAM5, part I: Model description, Technical Report 349, Max Planck Institute for Meteorology

Santer BD, Mikolajewicz U, Bruggemann W, Cubasch U, Hasselmann K, Hock H, Maierreimer E, Wigley TML (1995) Ocean variability and its influence on the detectability of greenhouse warming signals. J Geophys Res-Oceans 100:10693–10725

Sato M, Hansen JE, McCormick MP, Pollack JB (1993) Stratospheric aerosol optical depths, 1850–1990. J Geophys Res-Atmos 98(D12):22987–22994

Simon HA (1959) Theories of decision-making in economics and behavioral-science. Am Econ Rev 49:253–283

Schneider T, Neumaier A (2001) Algorithm 808: Arfit—a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans Math Softw 27:58–65

Smethie WM, Fine RA (2001) Rates of North Atlantic Deep water formation calculated from chlorofluorocarbon inventories. Deep Sea Res 48:189–215

Sohn MD, Small MJ, Pantazidou M (2000) Reducing uncertainty in site characterization using Bayes Monte Carlo methods. J Environ Eng ASCE 126(10):893–902

Stocker TF (1999) Abrupt climate changes: from the past to the future—a review. Int J Earth Sci 88:365–374

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

Stouffer RJ, Yin J, Gregory JM, Dixon KW, Spelman MJ, Hurlin W, Weaver AJ, Ebyd M, Flato GM, Hasumi H, Hu A, Jungclaus JH, Kamenkovich IV, Levermann A, Montoya M, Murakami S, Nawrath S, Oka A, Peltier WR, Robitaille DY, Sokolov A, Vettoretti G, Weber SL (2006) Investigating the causes of the response of the thermohaline circulation to past and future climate changes. J Clim 19:1365–1387

Swets JA (1973) Relative operating characteristic in psychology. Science 182:990–1000

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

Vellinga M, Wood RA (2004) Timely detection of anthropogenic change in the Atlantic meridional overturning circulation. Geophys Res Lett 31:L14203

Warnes GR (2001) The Normal Kernel Coupler: an adaptive Markov Chain Monte Carlo method for efficient sampling from multi-modal distributions, in University of Washington, Department of Statistics Working Papers, Report No. 39

Webster M, Scott J, Sokolov A, Stone P (2006) Estimating the probability distributions from complex models with bifurcations: The case of ocean circulation collapse, in Joint Program on the Science and Policy of Global Change, Report Number 133. MIT, Boston

Wilks DS (2001) A skill score based on economic value for probability forecasts. Meteorol Appl 8:209–219

Wunsch C (2006) Abrupt climate change: an alternative view. Quaternary Res 65:191–203

Zellner A, Tian GC (1964) Bayesian analysis of the regression model with autocorrelated errors. J Am Stat Assoc 59:763–778

Zhu XH, Fraedrich K, Blender R (2006) Variability regimes of simulated Atlantic MOC. Geophys Res Lett 33(21, L21603 Artn l21603)

Zickfeld K, Levermann A, Morgan MG, Kuhlbrodt T, Rahmstorf S, Keith DW (2007) Present state and future fate of the Atlantic meridional overturning circulation as viewed by experts. Clim Change 82:235–265

Zickfeld K, Slawig T, Rahmstorf S (2004) A low-order model for the response of the Atlantic thermohaline circulation to climate change. Ocean Dyn 54:8–26