Warning of a forthcoming collapse of the Atlantic meridional overturning circulation

Peter Ditlevsen (pditlev@nbi.ku.dk)
University of Copenhagen https://orcid.org/0000-0003-2120-7732
Susanne Ditlevsen
University of Copenhagen https://orcid.org/0000-0002-1998-2783

Physical Sciences - Article

Keywords:

Posted Date: September 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2034845/v1

Warning of a forthcoming collapse of the Atlantic meridional overturning circulation

P. Ditlevsen[1] and S. Ditlevsen[2]
1. Niels Bohr Institute, University of Copenhagen
2. Institute of Mathematical Sciences, University of Copenhagen

Abstract

Tipping to an undesired state in the climate, viewed as a complex system, when a control parameter slowly approaches a critical value ($\lambda(t) \to \lambda_c$) is a growing concern with increasing greenhouse gas concentrations. Predictions can rely on detecting early warning signals (EWS) in observations of the system. The primary EWS are increase in variance, due to loss of resilience, and increased autocorrelation or critical slow down. These measures are statistical in nature, which implies that the reliability and statistical significance of the detection depends on the sample size in observations and the magnitude of the change away from the base value prior to the approach to the tipping point. Thus, the possibility of providing useful early warning depends on the relative magnitude of several interdependent time scales in the system. These are (a) the time before the critical value λ_c is reached, (b) the (inverse) rate of approach to the bifurcation point, (c) the size of the time window required to detect a significant change in the EWS and finally, (d) the escape time for noise-induced transition (prior to the bifurcation). Conditions for early warning of tipping of the AMOC are marginally fulfilled for the existing past ∼150 years of proxy observations where indicators of tipping have recently been reported. Here we provide statistical significance and estimate a collapse of the AMOC to occur around year 2050.

1 Main

A forthcoming collapse of the Atlantic meridional overturning circulation (AMOC) is a major concern as it is one of the most important tipping elements in Earth's climate system [1, 2, 3]. In recent years, model studies and paleoclimatic reconstructions indicate that the strongest abrupt climate fluctuations, the Dansgaard-Oeschger events [4], are connected to the bimodal nature of the AMOC [5, 6]. Numerous climate model studies show a hysteresis behaviour, where changing a control parameter, typically the freshwater input into the Northern Atlantic, makes the AMOC bifurcate through a set of co-dimension one saddle-node bifurcation [7, 8]. State-of-the-art Earth-system models can reproduce such a scenario, but inter-model spread is large and the critical threshold is poorly constrained [9, 10]. When complex systems undergo critical transitions by changing a control parameter λ through a critical value λ_c, a structural change in the dynamics happens. The previously statistically stable state ceases to exist and the system moves to a different statistically stable state. The system undergoes a bifurcation, which for λ sufficiently close to λ_c can happen in a limited number of ways rather independent from the details in the governing dynamics [11]. Beside a decline of the AMOC before the critical transition, there are statistical quantities, the so-called Early Warning Signals (EWS), which also change before the tipping happens. These are critical slow down (increased autocorrelation) and, from the Fluctuation-Dissipation Theorem, increased variance in the signal [12, 13, 14]. The latter is also termed "loss of resilience", especially in the context of ecological collapse [15]. The two EWS are statistical equilibrium concepts, thus using them as actual predictors of a forthcoming transition, rely on the assumption of quasi-stationary dynamics. The AMOC has only been monitored continuously since 2004, through combined measurements from moored instruments, induced electrical currents in submarine cables and satellite surface measurements [16]. Over the period 2004-2012 a decline in the AMOC has been observed, but longer records are necessary in order to assess the significance. For that, careful finger printing techniques have been applied to longer records of sea surface temperature (SST), which, backed by survey of a large ensemble of climate model simulations, have found the SST in the Subpolar gyre (SG) region of

the North Atlantic (Area marked with a black contour in Fig. 1**a**) to contain an optimal fingerprint of the strength of the AMOC [17, 18, 19]. In order to obtain the AMOC fingerprint, two steps are required: The seasonal cycle in the SST is governed by the surface radiation independent from the circulation and thus removed by considering the monthly anomalies, where the mean over the period of recording of the month is removed. Secondly, there is an ongoing positive linear trend in the SST related to global warming, which is also not related to the circulation. This is compensated for by subtracting two times the global mean (GM) SST anomaly (small seasonal cycle removed). The factor two is a conservative estimate for the polar amplification [20] of global warming in the SG region. Fig. 1**b** shows the SG - and the GM SST obtained from the Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) [21]. Fig. 1**c** shows the SG anomaly and Fig.1**d** shows the GM anomaly with a clear global warming trend in the last half of the record. The AMOC fingerprint obtained from the HadISST for the period 1870-2020 is shown in Fig.1**e**. This is the basis for the analysis. It has been reported [22, 23] that this, and similar AMOC indices show significant trends in the variance and autocorrelation, indicating early-warning of a shutdown of the AMOC. However, a trend in the EWSs within a limited period of observation could be a random fluctuation within a steady state statistics. Thus, for a robust assessment of the tipping point, it is necessary to establish a statistical confidence level for the change above natural fluctuations. This is not easily done given only one, the observed, realization of the transition. Here we establish such a measure of the confidence and find that the transition is most likely to occur in 2051 with the 95% confidence interval 2026-2071. The strategy is to infer the evolution of the AMOC solely on observed changes in mean and variance. The autocorrelation can be included, but that is not necessary, as we demonstrate that variance is the more reliable EWS of the two. The typical choice of control parameter is the flux of freshwater into the North Atlantic. River runoff, Greenland ice melt and export from the Arctic ocean are not well constrained [24], thus we do not assume the control parameter known. Boers [23] assumes the global mean temperature T to represent the control parameter. T increases roughly linear with time since $\sim$ 1920 (Fig.1**d**). All we assume here is that the AMOC is in an equilibrium state prior to a change towards the transition. The simplest, uninformed, assumption is that the change is sufficiently slow and that the control parameter approach the (unknown) critical value linearly with time.

2 Modeling the critical transition

For ease of notation, we denote the observed AMOC variable as $x(t)$ (Fig.1**e**). We assume X_t to be a representative observable of a system, which, depending on a control parameter $\lambda < 0$, is in risk of undergoing a critical transition through a saddle-node bifurcation for $\lambda = \lambda_c = 0$. The system is initially in a statistically stable state, i.e., it follows some stationary distribution. We are uninformed about the dynamics governing the evolution of X_t, but can assume an effective dynamics, which, with λ sufficiently close to the critical value $\lambda_c = 0$, can be described by the stochastic differential equation (SDE):

$$dX_t = -((X_t - m)^2 + \lambda)dt + \sigma dB_t, \tag{1}$$

where $m + \sqrt{|\lambda|}$ represents the mean level of the observed record. Disregarding the noise, this is the normal form of the co-dimension one saddle-node bifurcation [11]. The detection of a forthcoming transition using statistical measures involves several time scales. The primary internal time scale is the autocorrelation time τ of $x(t)$ in the steady state. The period over which the control parameter changes from the steady state value to the critical value sets an external time scale, τ_r ('r' for 'ramping'). The mean and variance are calculated from the observations as the control parameter $\lambda(t)$ is changing. These quantities are inherently equilibrium concepts and statistical, thus a time-window, T_w, of a certain size is required for a reliable estimate. The further away from the steady state value (baseline), the shorter is the window required for detecting a change. On the other hand, the closer to transition, critical slow down decreases the number of independent points within a window, thus calls for a larger window for a reliable detection. Within a short enough window, $[t - T_w/2, t + T_w/2]$, we may assume $\lambda(t)$ to be constant and the noise small enough, that the process is well approximated by the linear process $dX_t = -\alpha(X_t - \mu)dt + \sigma dB_t$, where μ is the mean and α is the inverse correlation time. The EWSs are obtained from the observed time series by maximum likelihood estimation (MLE), which for mean μ, one-lag autocorrelation $\rho = \exp(-\alpha\Delta t)$ and variance $\gamma^2 = \sigma^2/2\alpha$ are given in closed forms (see Methods) [25].

3 Uncertainty on Early Warning Signals

The uncertainty is expressed through the variances of the estimates $\hat{\gamma}^2$ and $\hat{\rho}$ obtained from the observations within a time window T_w. These are $\mathrm{Var}(\hat{\gamma}^2) \approx \sigma^4/(2\alpha^3 T_w)$ and $\mathrm{Var}(\hat{\rho}) \approx 2\alpha\Delta t^2/T_w$, where $T_w = n\Delta t$ is the observation window, n is the number of observations within the window, Δt is the time step between observations, and α is a function of λ: $\alpha(\lambda) = 2\sqrt{|\lambda|}$ (see Methods). The uncertainties can be made arbitrarily small by observing over a long time window T_w. We may therefore assume that $\rho_0 = \exp(-\alpha_0\Delta t)$ and $\gamma_0^2 = \sigma^2/(2\alpha_0)$ are known, where $\alpha_0 = 2\sqrt{|\lambda_0|}$ and λ_0 is the baseline value before t_0. At t_0, $\lambda(t)$ begins to change linearly towards $\lambda_c = \lambda(t_c) = 0$: $\lambda(t) = \lambda_0(1 - \Theta[t - t_0](t - t_0)/\tau_r)$, where $\Theta[t]$ is the Heaviside function and $\tau_r = t_c - t_0 > 0$ is the ramping time up to time t_c, whereafter the transition eventually will occur. Knowing $\alpha(t)$, and deciding the q-percentile (such as 95% or 99% confidence level) the required time window $T_w(q, \alpha)$, to detect a change from baseline in EWSs at the given confidence level is given in closed form in Methods (7) for variance and in (8) for autocorrelation. As the transition is approached, the risk of a noise-induced tipping (n-tipping) prior to t_c is increasing and at some point making the EWS irrelevant for predicting the tipping. The probability for n-tipping can in the small noise limit be calculated in closed form, $P(t, \lambda) = 1 - \exp(-t/\tau_n(\lambda))$, with mean waiting time $\tau_n(\lambda) = (\pi/\sqrt{|\lambda|})\exp(8|\lambda|^{\frac{3}{2}}/3\sigma^2)$ (see Methods). To summarize the involved time scales, in Fig. 2**a** the required window size T_w at the 95% confidence level is plotted as a function of $\lambda(t)$ for the EWS variance (red curve) and autocorrelation (yellow curve). These are plotted together with the mean waiting time for n-tipping (blue curve). With a chosen data window size of 50yrs, increased variance can only be detected after the time when $\lambda(t) \approx -1.1$ (crossing of red and red-dashed curves). At that time a window of approximately 75yrs is required to detect an increase in autocorrelation, making variance the better EWS of the two. When $\lambda \approx -0.4$ the mean waiting time for n-tipping is smaller than the data window size. Thus, the increased variance can be used as a reliable EWS in the range $-1.1 <\approx \lambda(t) <\approx -0.4$ indicated by the green band. How timely an early warning this is depends on the speed at which $\lambda(t)$ is changing from λ_0 to λ_c, i.e., the ramping time τ_r. A set of 1000 realizations have been simulated with $\lambda_0 = -2.56$ and $\tau_r = 140$yrs, indicated by the time labels on top of Fig. 2**a**. Ten of these realizations are shown in Fig. 2**b** on top of the stable and unstable branches of fixed points of the model (1) (this is the bifurcation diagram of the model Eq. (1)). Fig. 2**c** (**d**) shows the variance (autocorrelation) calculated from the realizations within a running 50yrs window (shown in Fig. 2**c**). The solid black line is the baseline value, while the solid blue line is the increasing value. The calculated 95% confidence level for the measurement of the EWS within the running window is shown by the dashed black and blue lines, respectively. The corresponding light blue curves are obtained numerically from the 1000 realizations. The green band in Fig. 2**c** corresponds to the green band in Fig. 2**a** and shows where early warning is possible in this case.

4 Predicting a forthcoming collapse of the AMOC

The AMOC fingerprint shown in Fig. 1**e** (replotted in Fig. 3**a**) shows an increased variance and autocorrelation, plotted in Fig. 3**b** and **c** as functions of the mid-point of a 50yrs running window, i.e., the EWS obtained in 2020 is assigned to year 1995. With the results above, we can now estimate the significance of the increase: Firstly, within the running window, we can obtain the parameters of the linearized dynamics, α (Fig. 3**d**) and σ^2 (Fig. 3**e**). These are consistent with a linear decrease of $\lambda(t)$ beginning from a constant level $\lambda_0 = -2.56$ at 1920, ramping linearly with $\tau_r = 130$yrs, and a constant noise level $\sigma^2 = 0.28$. This is shown by the red lines in Fig. 3**a**, **d** and **e**. With these parameter values the model is completely determined and the confidence levels can be calculated, thus the two-sigma levels around the baseline values of the EWS are shown by purple-dashed lines in Fig. 3**b** and **c**. Thus, both EWS show increases beyond the two-sigma level. Once a change in λ has been established, the ramping time can be estimated by MLE from data after time t_0 in the approximate model (see Methods). This leads to an estimate of 131yrs using $t_0 = 1920$. This is how the tipping time in year 2051 is estimated, shown in Fig. 3**f**. The estimate of t_c is robust with respect to the estimate of t_0: Assuming t_0 in 1900-1950 provides estimates between 2047 and 2056 for t_c. Observing the estimated λ_0 (Fig 4**c**) show the linear decrease beginning in 1920, suggesting this to be the best estimate for t_0. In order to obtain an estimate of the uncertainty in the derived critical time of collapse t_c, a set of 1000 realizations have been simulated, using the parameters estimated from the AMOC data starting in $t_0 = 1920$. From each realization the parameters λ_0, σ^2, m and t_c are calculated by MLE (see Methods).

From this the probability density function (PDF) (Fig. 4**a**) and the corresponding cumulative distribution function (Fig. 4**b**) are obtained. The mean is $\langle t_c \rangle = 2047$ and the 90% confidence interval is $2026 - 2071$. The small discrepancy in mean is due to the approximate model used for estimation being different from the data generating model (1), confirming that the linear model still provides valid estimates even if the true dynamics are unknown. To test the goodness-of-fit, uniform residuals (see Methods) were calculated for the data from 1920 and up to today. These are plotted in Fig. 4**d**. The model is seen to fit the data well, further supporting the obtained estimates.

5 Discussion

We have provided a novel robust statistical analysis to quantify the uncertainty in observed early warning signals for a forthcoming critical transition. The confidence depends on how rapid the system is approaching the tipping point. With this the significance of the observed EWS for the AMOC has been established and we predict with high confidence the tipping to happen as soon as 2051. This is indeed a worrisome result, which should call for fast and effective measures to reduce global greenhouse gas emissions in order to avoid the steadily change of the control parameter towards the collapse of the AMOC (i.e. reduce temperature increase and fresh water input through ice melting into the North Atlantic region). As a collapse of the AMOC has strong societal implications [26], it is important to monitor the flow and EWS from direct measurements [**?**, **?**, 27].

6 Methods

Assume the evolution of a representative variable of the system X_t is governed by eq. (1), and that the control parameter λ is constant up to time t_0 and then increases linearly as

$$\lambda(t) = \lambda_0(1 - \Theta[t - t_0](t - t_0)/\tau_r). \tag{2}$$

Since the exact dynamics eq. (1) are assumed unknown, they are locally (i.e., for given λ) approximated with a linear SDE, the Ornstein-Uhlenbeck process [28]. A Taylor expansion around the mean $\mu(\lambda)$ yields the approximation

$$dX_t \approx -\alpha(\lambda)(X_t - \mu(\lambda))dt + \sigma dB_t \tag{3}$$

where $\mu(\lambda) = m + \sqrt{|\lambda|}$ and $\alpha(\lambda) = 2\sqrt{|\lambda|}$. For fixed λ the process is stationary and the estimators of μ, $\rho = e^{-\alpha\Delta t}$ and $\gamma^2 = \sigma^2/2\alpha$ using MLE are given in the online material, see also [25]. The asymptotic variances of the estimators obtained by inverting the Fisher information are $\mathrm{Var}(\hat{\mu}) \approx \gamma^2(1+\rho)/(1-\rho)n$, $\mathrm{Var}(\hat{\rho}) \approx (1-\rho^2)/n$ and $\mathrm{Var}(\hat{\gamma}^2) \approx 2(\gamma^2)^2(1+\rho^4)/(1-\rho^2)n$ (see online material). For $\alpha\Delta t \ll 1$ we approximate $(1+\rho^4)/(1-\rho^2) \approx 1/(\alpha\Delta t)$ and $1 - \rho^2 \approx 2\alpha\Delta t$ and obtain

$$\mathrm{Var}(\hat{\gamma}^2) \approx \frac{2(\gamma^2)^2}{\alpha T_w} = \frac{\sigma^4}{2\alpha^3 T_w}; \quad \mathrm{Var}(\hat{\rho}) \approx \frac{2\alpha\Delta t^2}{T_w}, \tag{4}$$

where $T_w = n\Delta t$ is the observation window.

The uncertainties in $\hat{\gamma}^2$ and $\hat{\rho}$ can be made arbitrarily small by observing over a long time window T_w, we may therefore assume that $\rho_0 = \exp(-\alpha_0\Delta t)$ and $\gamma_0^2 = \sigma^2/(2\alpha_0)$ are known, where $\alpha_0 = 2\sqrt{|\lambda_0|}$. When $\lambda(t)$ starts increasing according to (2), the normal state disappears after $t_c = t_0 + \tau_r$ time units, whereafter the transition eventually will occur. Time t_c is denoted the tipping time, however, it can happen earlier due to a noise-induced tipping. As $\lambda(t)$ increases, α decreases, and thus variance and autocorrelation increase. The question is then how large T_w needs to be in order to detect a statistically significant increase compared to the baseline values γ_0^2 and ρ_0. For a given estimate $\hat{\gamma}^2$, the estimated difference from the baseline variance is

$$\Delta_{\gamma^2} = \hat{\gamma}^2 - \gamma_0^2 \quad = \quad \gamma_0^2(\alpha_0/\hat{\alpha} - 1), \tag{5}$$

and the estimated difference from the baseline autocorrelation is

$$\Delta_\rho = \hat{\rho} - \rho_0 \quad = \quad \rho_0(e^{(\alpha_0 - \hat{\alpha})\Delta t} - 1) \approx \rho_0(\alpha_0 - \hat{\alpha})\Delta t. \tag{6}$$

Since the two EWS, γ^2 and ρ, are treated on an equal footing, in the following we let $\hat{\psi}$ denote either of the estimators (11) or (12) and $\hat{\Delta}$ denote either of the two estimated differences (5) or (6). The estimator $\hat{\Delta}$ follows a distribution depending on T_w. The null hypothesis is that $\lambda = \lambda_0$, or equivalently $\alpha = \alpha_0$. The null distribution of $\hat{\psi}$ is obtained from simulating eq. (3) many times with $\lambda = \lambda_0$, and estimate ψ on the simulated samples. Subsequently, after suitable normalization by subtracting ψ and dividing the estimate with the standard error $s(\hat{\psi}) = \mathrm{Var}(\hat{\psi})^{1/2}$, eq. (4), we obtain a quantile q, which expresses the acceptable uncertainty in measuring the statistical quantity ψ (being variance or autocorrelation). For large T_w, it can be assumed that q approaches the quantile obtained from the standard normal distribution (confirmed by simulations). We thus get that $\hat{\Delta} < qs(\hat{\psi})$ at the q-confidence level (95%, 99% or similar) under the null hypothesis. In order to detect an EWS at the q-confidence level based on measuring ψ at time t, we require that $\hat{\Delta}(t) > q(s(\hat{\psi}(t)) + s(\psi_0))$, which, solved for T_w gives for variance:

$$T_w > 2q^2 \left(\frac{\alpha(t)/\sqrt{\alpha_0} + \alpha_0/\sqrt{\alpha(t)}}{\alpha_0 - \alpha(t)} \right)^2, \tag{7}$$

and for autocorrelation,

$$T_w > 2q^2 \left(\frac{\sqrt{\alpha_0} + \sqrt{\alpha(t)}}{\alpha_0 - \alpha(t)} \right)^2 \rho_0^{-2}. \tag{8}$$

where we assume q independent from T_w and identical for variance and autocorrelation. Substituting $\alpha(t) = 2\sqrt{|\lambda(t)|}$, provides the time window T_w needed to detect an EWS at time t. In Fig. 2 the different time scales are shown as functions of λ.

6.1 Estimation of the tipping time

To estimate the tipping time once it has been established that the variance and autocorrelation are increasing, only data after the linear ramping has started is used, $x = (x_0, x_1, \ldots, x_n)$, where $x_0 = x_{t_0}$ and $x_n = x_{today}$. The time t_0 is not known, we only know that it is before the time where the increase in EWSs are sufficiently large to be detected. To obtain robust estimates, we repeated the estimation of the tipping time for t_0 taken to be all years from the beginning of the time series and up to the time of detection of an EWS. If t_0 is chosen too early, the estimate will be off because it is contaminated by stationary data from the beginning of the record. For too late choice of t_0, the estimate will be uncertain due to limited data, and estimates will be hugely fluctuating. We expect a time interval where estimates are stable, and thus, an exact estimate of t_0 is less important, as long as it falls within this middle interval. This is indeed the case, see Fig. 4**c**.

We use approximation (3), but now with time varying λ, where $\alpha(\lambda(t - t_0)) = 2\sqrt{|\lambda_0|(1 - t/\tau_r)}$ and $\mu(\lambda(t - t_0)) = m + \sqrt{|\lambda_0|(1 - t/\tau_r)}$, where λ_0 is the value of $\lambda(t)$ up to time t_0.

Simplifying by assuming that λ is constant between observations, i.e., piecewise constant and jumping every month where new AMOC observations are available, the approximate likelihood function of the parameters $\theta = (\lambda_0, \tau_r, m, \sigma^2)$ is the product of Gaussian transition densities

$$L_x(\theta) = \prod_{i=1}^{n} \varphi((x_i - m_i)/s_i; \theta)$$

where $\varphi(\cdot)$ is the standard normal probability density. The conditional mean is $m_i := E(X_i | X_{i-1} = x_{i-1}) = x_{i-1}\rho_{i-1} + \mu_{i-1}(1 - \rho_{i-1})$ and the conditional variance is $s_i^2 = \gamma_{i-1}^2(1 - \rho_{i-1}^2)$, where $\mu_i = m + \sqrt{|\lambda_0|(1 - (t_i - t_0)/\tau_r)}$, $\rho_i = \exp(-2\sqrt{|\lambda_0|(1 - (t_i - t_0)/\tau_r)}\Delta t)$ and $\gamma_i^2 = \sigma^2/4\sqrt{|\lambda_0|(1 - (t_i - t_0)/\tau_r)}$. The parameter estimates $\hat{\theta}$ are found numerically by minimizing $-\log L_x(\theta)$. For this we use the optimizer `optim` in R, using the Nelder-Mead algorithm. Starting values for the parameters are required. We draw randomly 100 sets of starting values and chose the optimal run to avoid the risk of falling into a local minimum. Starting values for $-\lambda_0$ were drawn from a gamma distribution with mean 2 and variance 1, τ_r was drawn from a normal distribution with mean 200yrs (twice the time from y_0 to the time of estimation) and variance 50^2, m was drawn from a normal distribution with mean 0 and variance 2^2, and σ^2 was drawn from a gamma distribution with mean 0.25 and variance 0.1^2. Out of the 100 runs, 16 runs arrived at the same smallest minimum, suggesting they had reached a global minimum.

The likelihood approach provides asymptotic confidence intervals, however, these assume that the likelihood is the true likelihood. To incorporate also the uncertainty due to the data generating mechanism (1) not being equal to the Ornstein-Uhlenbeck process (3) used in the likelihood, we chose to construct parametric bootstrap confidence intervals. This was obtained by simulating 1000 trajectories from the original model with the estimated parameters, and repeat the estimation procedure on each data set. Empirical confidence intervals were then extracted from then 1000 parameter estimates. These were indeed larger than the asymptotic confidence intervals provided by the likelihood approach, however, not by much.

6.2 Model control

To test the model fit, uniform residuals, $u_i, i = 1, \ldots, n$, were calculated for the AMOC data from 1920 and up to today as follows. The model assumes that observation x_i is normally distributed with mean m_i and variance s_i^2 for the estimated parameter values. If this is true, then $u_i = F_{i,\hat{\theta}}(x_i)$ is uniformly distributed on $(0, 1)$, where $F_{i,\hat{\theta}}$ is the cumulative normal distribution function with the estimated mean and variance for the i'th observation. Transforming these residuals back to a standard normal distribution provides standard normally distributed residuals if the model is true. Thus, a normal quantile-quantile plot reveals the model fit. The points should fall close to the identity line. The reason for making the detour around the uniform residuals is twofold. First, since the data is not stationary, each observation follows its own distribution, and residuals cannot be directly combined. Second, since the model is stochastic, standard residuals are not well-defined, and observations should be evaluated according to their entire distribution, not only the distance to the mean.

References

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

[2] S. Rahmstorf. Bifurcations of the atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature*, 378:145–149, 1995.

[3] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber. Tipping elements in the earth's climate system. *Proceedings of the National Academy of Sciences*, 105:1786–1793, 2008.

[4] W. Dansgaard, S. J. Johnsen, H. B. Clausen, D. Dahl-Jensen, N. S. Gundestrup, C. U. Hammer, C. S. Hvidberg, J. P. Steffensen, A. E. Sveinbjornsdottir, J. Jouzel, and G. Bond. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, 364:218–220, 1993.

[5] G. Vettoretti, P. Ditlevsen, M. Jochum, and S. O. Rasmussen. Atmospheric co2 control of spontaneous millennial-scale ice age climate oscillations. *Nature Geoscience*, 15:300–306, 2022.

[6] A. Ganopolski and S. Rahmstorf. Rapid changes of glacial climate simulated in a coupled climate model. *Nature*, 409:153–158, 2001.

[7] R. A. Wood, J. M. Rodríguez, R. S. Smith, L. C. Jackson, and E. Hawkins. Observable low-order dynamical controls on thresholds of the atlantic meridional overturning circulation. *Clim. Dyn.*, 53:6815–6834, 2019.

[8] E. Hawkins, R. S. Smith, L. C. Allison, J. M. Gregory, T. J. Woollings, H. Pohlmann, and B. de Cuevas. Bistability of the atlantic overturning circulation in a global climate model and links to ocean freshwater transport. *Geophysical Research Letters*, 38(10), 2011.

[9] J.V. Mecking, S.S. Drijfhout, L.C. Jackson, and M.B. Andrews. The effect of model bias on atlantic freshwater transport and implications for amoc bi-stability. *Tellus A: Dynamic Meteorology and Oceanography*, 69(1):p.1299910, 2017.

[10] V. et al. Masson-Delmotte. *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[11] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1986.

[12] R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, pages 255–284, 1966.

[13] P. D. Ditlevsen and S. Johnsen. Tipping points: Early warning and wishful thinking. *Geophys. Res. Lett.*, 37:L19703, 2010.

[14] C. Boulton, L. Allison, and T. Lenton. Early warning signals of atlantic meridional overturning circulation collapse in a fully coupled climate model. *Nat Commun*, 5:5752, 2014.

[15] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara. Early-warning signals for critical transitions. *Nature*, 461:53–59, 2009.

[16] D. A. Smeed, G. D. McCarthy, S. A. Cunningham, E. Frajka-Williams, D. Rayner, W. E. Johns, C. S. Meinen, M. O. Baringer, B. I. Moat, A. Duchez, and H. L. Bryden. Observed decline of the atlantic meridional overturning circulation 2004-2012. *Ocean Science*, 10(1):29–38, 2014.

[17] L. Caesar, S. Rahmstorf, A. Robinson, G. Feulner, and V. Saba. Observed fingerprint of a weakening atlantic ocean overturning circulation. *Nature*, 552:191–196, 2018.

[18] L. C. Jackson and R. A. Wood. Fingerprints for early detection of changes in the amoc. *Journal of Climate*, 33(16):7027 – 7044, 2020.

[19] M. Latif. Reconstructing, monitoring, and predicting multidecadal-scale changes in the north atlantic thermohaline circulation with sea surface temperature. *J. Climate*, 17:1605–1614, 2004.

[20] M.M. Holland and C.M. Bitz. Polar amplification of climate change in coupled models. *Climate Dynamics*, 21:221–232, 2003.

[21] N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, 108:4407, 2003.

[22] S. Rahmstorf, J. E. Box, G. Feulner, M. E. Mann, A. Robinson, S. Rutherford, and E. J. Schaffernicht. Exceptional twentieth-century slowdown in atlantic ocean overturning circulation. *Nat. Climate Change*, 5:475–480, 2015.

[23] N Boers. Observation-based early-warning signals for a collapse of the atlantic meridional overturning circulation. *Nat. Clim. Chang.*, 11:680–688, 2021.

[24] Q. Yang, T. Dixon, P. Myers, J. Bonin, D. Chambers, M. R. van den Broeke, M. H. Ribergaard, and J. Mortensen. Recent increases in arctic freshwater flux affects labrador sea convection and atlantic overturning circulation. *Nat Commun.*, 7:10525, 2016.

[25] S. Ditlevsen, A. Cencerrado Rubio, and P. Lansky. Transient dynamics of Pearson diffusions facilitates estimation of rate parameters. *Communications in Nonlinear Science and Numerical Simulation*, 82:105034, 2020.

[26] Luke Kemp, Chi Xu, Joanna Depledge, Kristie L. Ebi, Goodwin Gibbins, Timothy A. Kohler, Johan Rockström, Marten Scheffer, Hans Joachim Schellnhuber, Will Steffen, and Timothy M. Lenton. Climate endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, 119(34):e2108146119, 2022.

[27] R. Alexander-Turner, P. Ortega, and J. I. Robson. How robust are the surface temperature fingerprints of the atlantic overturning meridional circulation on monthly time scales? *Geophysical Research Letters*, 45(8):3559–3567, 2018.

[28] K. Hasselmann. Stochastic climate models. *Tellus*, 28:473–485, 1976.

[29] N. Berglund. Kramers' Law: Validity, Derivations and Generalisations. *Markov Processes and Related Fields*, 19(3):459–490, 2013.

[30] M.I. Freidlin and A.D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1984.

Supplementary material

Maximum likelihood estimators of the Ornstein-Uhlenbeck process

The approximate model is an Ornstein-Uhlenbeck (OU) process, defined as the solution to the equation

$$dX_t = -\alpha(X_t - \mu)dt + \sigma dB_t. \tag{9}$$

The variance is $\gamma^2 = \sigma^2/2\alpha$ and the Δt-lag autocorrelation is $\rho = e^{-\alpha\Delta t}$. The likelihood function of the parameters given observations $(x_0, x_1, \ldots, x_n)$ is the product of the transition densities

$$L_n(\theta) = \prod_{i=1}^{n} p(\triangle, x_{i-1}, x_i; \theta)$$

where $\theta = (\mu, \rho, \gamma^2)$. Here, $x_i = x(t_i)$ and $\Delta t = t_i - t_{i-1}$. The transition density is normal with conditional mean $E(X_i|X_{i-1} = x_{i-1}) = x_{i-1}\rho + \mu(1-\rho)$ and conditional variance $\gamma^2(1-\rho^2)$,

$$p(\triangle, x_{i-1}, x_i; \theta) = \frac{1}{\sqrt{2\pi\gamma^2(1-\rho^2)}} \exp\left(-\frac{(x_i - x_{i-1}\rho - \mu(1-\rho))^2}{2\gamma^2(1-\rho^2)}\right).$$

The maximum likelihood estimators (MLEs) are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{\hat{\rho}}{n(1-\hat{\rho})}(x_n - x_0) \approx \frac{1}{n+1}\sum_{i=0}^{n} x_i \equiv \bar{x}, \tag{10}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})(x_{i-1} - \hat{\mu})}{\sum_{i=1}^{n}(x_{i-1} - \hat{\mu})^2}, \tag{11}$$

$$\hat{\gamma}^2 = \frac{\sum_{i=1}^{n}\left(x_i - x_{i-1}\hat{\rho} - \hat{\mu}(1-\hat{\rho})\right)^2}{n\,(1-\hat{\rho}^2)}, \tag{12}$$

the symbol $\hat{}$ indicates an estimator. These are obtained as follows. The score function is the vector of derivatives of the log-likelihood function with respect to the parameters. The MLE is given as solution to the likelihood equations $\partial_{\theta_k} \log L_n(\theta) = 0$, where θ_k is either μ, ρ or γ^2. The score function is

$$\begin{aligned}
\frac{\partial}{\partial\mu}\log L_n(\theta) &= \frac{(1-\rho)}{\gamma^2(1-\rho^2)}\sum_{i=1}^{n}(x_i - x_{i-1}\rho - \mu(1-\rho)),\\
\frac{\partial}{\partial\rho}\log L_n(\theta) &= \frac{n\rho}{1-\rho^2} + \frac{\sum_{i=1}^{n}(x_i - x_{i-1}\rho - \mu(1-\rho))(x_{i-1} - \mu)}{\gamma^2(1-\rho^2)}\\
&\quad - \frac{\rho\sum_{i=1}^{n}(x_i - x_{i-1}\rho - \mu(1-\rho))^2}{\gamma^2(1-\rho^2)^2},\\
\frac{\partial}{\partial\gamma^2}\log L_n(\theta) &= -\frac{n}{2\gamma^2} + \frac{\sum_{i=1}^{n}(x_i - x_{i-1}\rho - \mu(1-\rho))^2}{2\gamma^4(1-\rho^2)},
\end{aligned}$$

whose zeros provide the MLEs in equations (10)–(12). It requires that $\sum_{i=1}^{n}(x_i - \hat{\mu})(x_{i-1} - \hat{\mu}) > 0$, otherwise the MLE does not exist.

The Fisher Information $\mathcal{I}$ of the MLEs equals minus the expectation of the Hessian $\mathcal{H}$ of the log-likelihood

function. For the OU log-likelihood, the elements of $\mathcal{H}$ are given by

$$\begin{aligned}
\frac{\partial^2}{\partial\mu^2}\log L_n(\theta) &= -\frac{n(1-\rho)}{\gamma^2(1+\rho)},\\
\frac{\partial^2}{\partial\mu\rho}\log L_n(\theta) &= \sum_{i=1}^{n}\left(C_1(x_{i-1}-\mu)+C_2(x_i-x_{i-1}\rho-\mu(1-\rho))\right),\\
\frac{\partial^2}{\partial\mu\gamma^2}\log L_n(\theta) &= C_3\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho)),\\
\frac{\partial^2}{\partial\rho^2}\log L_n(\theta) &= \frac{n(1+\rho^2)}{(1-\rho^2)^2}+C_4\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho))(x_{i-1}-\mu)-\frac{1}{\gamma^2(1-\rho^2)}\sum_{i=1}^{n}(x_{i-1}-\mu)^2\\
&\quad -\frac{1+3\rho^2}{\gamma^2(1-\rho^2)^3}\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho))^2,\\
\frac{\partial^2}{\partial\rho\gamma^2}\log L_n(\theta) &= C_5\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho))(x_{i-1}-\mu)+\frac{\rho}{\gamma^4(1-\rho^2)^2}\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho))^2,\\
\frac{\partial^2}{\partial(\gamma^2)^2}\log L_n(\theta) &= \frac{n}{2\gamma^4}-\frac{\sum_{i=1}^{n}(x_i-x_{i-1}\rho-\mu(1-\rho))^2}{\gamma^6(1-\rho^2)},
\end{aligned}$$

where $C_i, i=1,\ldots,5$, are deterministic constants that will disappear when taking expectations. Using that $E(X_i-\mu)^2=\gamma^2$, $E(X_i-X_{i-1}\rho-\mu(1-\rho))^2=\gamma^2(1-\rho^2)$ and $E(X_i-X_{i-1}\rho-\mu(1-\rho))(Y_{i-1}-\mu)=0$, we obtain the Fisher Information

$$\mathcal{I}=-E\mathcal{H}=n\begin{bmatrix}\frac{(1-\rho)}{\gamma^2(1+\rho)} & 0 & 0\\ 0 & \frac{1+\rho^4}{(1-\rho^2)^2} & \frac{\rho}{\gamma^2(1-\rho^2)}\\ 0 & \frac{\rho}{\gamma^2(1-\rho^2)} & \frac{1}{2\gamma^4}\end{bmatrix}.$$

The inverse of the Fisher Information provides the asymptotic covariance matrix,

$$\frac{1}{n}\begin{bmatrix}\frac{\gamma^2(1+\rho)}{(1-\rho)} & 0 & 0\\ 0 & 1-\rho^2 & 2\rho\gamma^2\\ 0 & 2\rho\gamma^2 & \frac{2\gamma^4(1+\rho^4)}{1-\rho^2}\end{bmatrix}.$$

The diagonal elements provide the asymptotic variances of μ, ρ and γ^2, respectively.

Noise induced tipping

The drift term in eq. (1) is the negative gradient of a potential, $f(x,\lambda)=-\partial_x V(x,\lambda)=-((x-m)^2+\lambda)$ with $V(x,\lambda)=(x-m)^3/3+(x-m)\lambda$. For $\lambda<0$, the drift has two fixed points, $m\pm\sqrt{|\lambda|}$. The point $m+\sqrt{|\lambda|}$ is a local minimum of the potential $V(x,\lambda)$ and is stable, whereas $m-\sqrt{|\lambda|}$ is a local maximum and unstable. The system thus has two basins of attraction separated by $m-\sqrt{|\lambda|}$, with a drift towards either $m+\sqrt{|\lambda|}$ or $-\infty$ dependent on whether $X_t>m-\sqrt{|\lambda|}$ or $X_t<m-\sqrt{|\lambda|}$. We denote the two basins of attraction the normal and the tipped state, respectively. When $\lambda=0$, the normal state disappears and the system undergoes a bifurcation and X_t will be drawn towards $-\infty$.

Due to the noise, the process can escape into the tipped state by crossing over the potential barrier $\Delta(\lambda)=V(-\sqrt{|\lambda|},\lambda)-V(\sqrt{|\lambda|},\lambda)=4|\lambda|^{\frac{3}{2}}/3$. Assume X_t to be close to $\sqrt{|\lambda|}$ at some time t, i.e., in the normal state. The escape time will asymptotically (for $\sigma\to 0$) follow an exponential distribution such that

$$P(t,\lambda)=1-\exp(-t/\tau_n(\lambda)) \tag{13}$$

where $P(t,\lambda)$ is the probability of observing an escape time shorter than t for a given value of λ. The mean noise induced escape time $\tau_n(\lambda)$ is [29, 30]:

$$\tau_n(\lambda) = \frac{2\pi \exp(2\Delta(\lambda)/\sigma^2)}{\sqrt{V''(-\sqrt{|\lambda|},\lambda)|V''(\sqrt{|\lambda|},\lambda)|}} = (\pi/\sqrt{|\lambda|}) \exp(8|\lambda|^{\frac{3}{2}}/3\sigma^2). \tag{14}$$

Assume that the rate of change of $\lambda(t)$ follows eq. (2), then for $\tau_r < \tau_n(\lambda)$, the waiting time for a random crossing is so long that a crossing will not happen before a bifurcation induced transition happens (b-tipping). If $\tau_r > \tau_n(\lambda)$, a noise-induced tipping is expected before the bifurcation point is reached. Since $\tau_n(\lambda)$ decreases with increasing λ, at some point, the two time scales will end up matching.

Figures

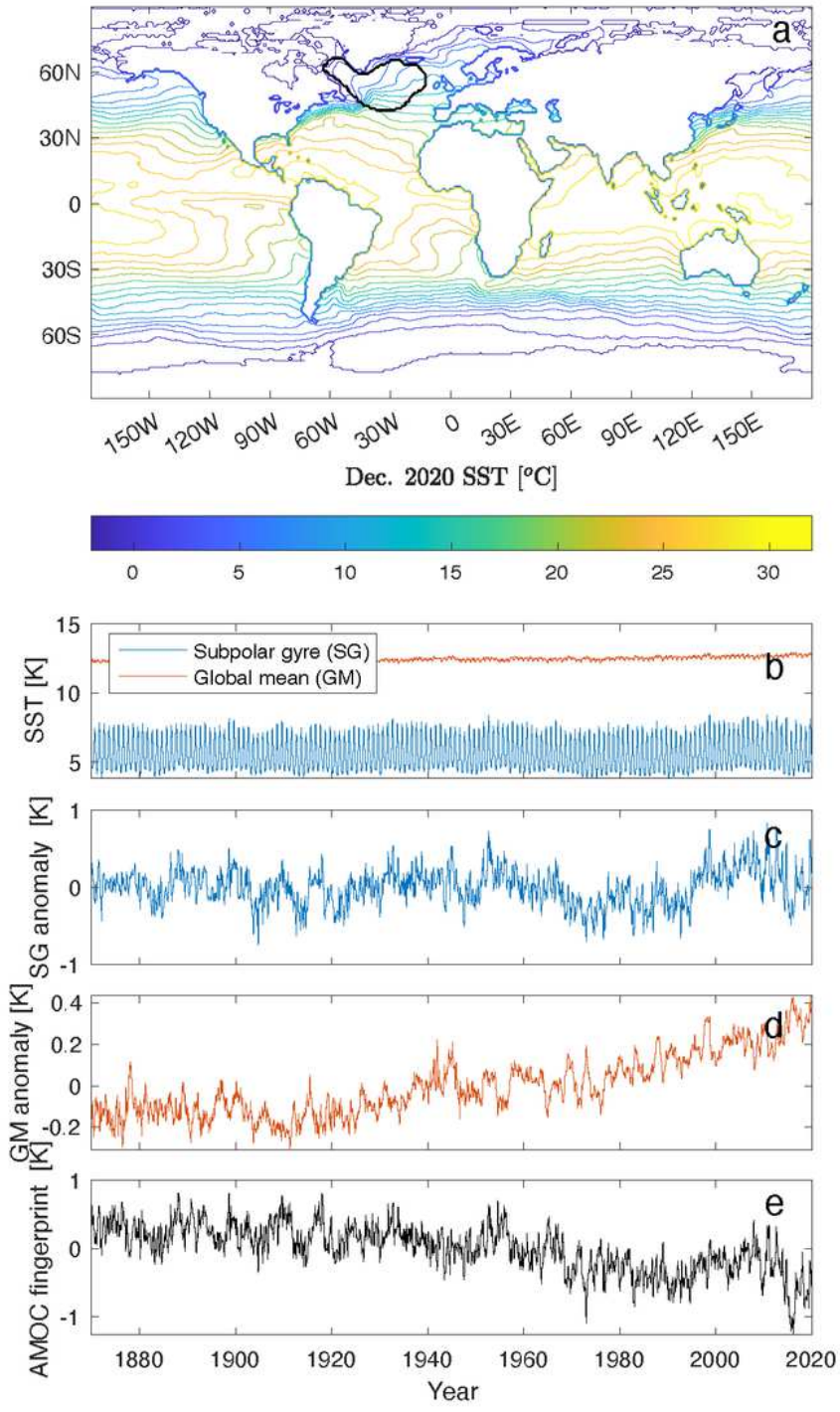

Figure 1

Panel a shows the Subpolar gyre (SG) region (black contour) on top of the HasISST SST reconstruction for Dec. 2020. The SG region SST has been identified as an AMOC fingerprint [17]. Panel b shows full monthly record of the SG SST together with the global mean (GM) SST. Panels c and d show the SG and

GM anomalies, which are the records subtracted the monthly mean over the full record. Panel e shows the AMOC fingerprint proxy, which is here defined as the SG anomaly minus twice the GM anomaly, compensating for the polar amplified global warming.

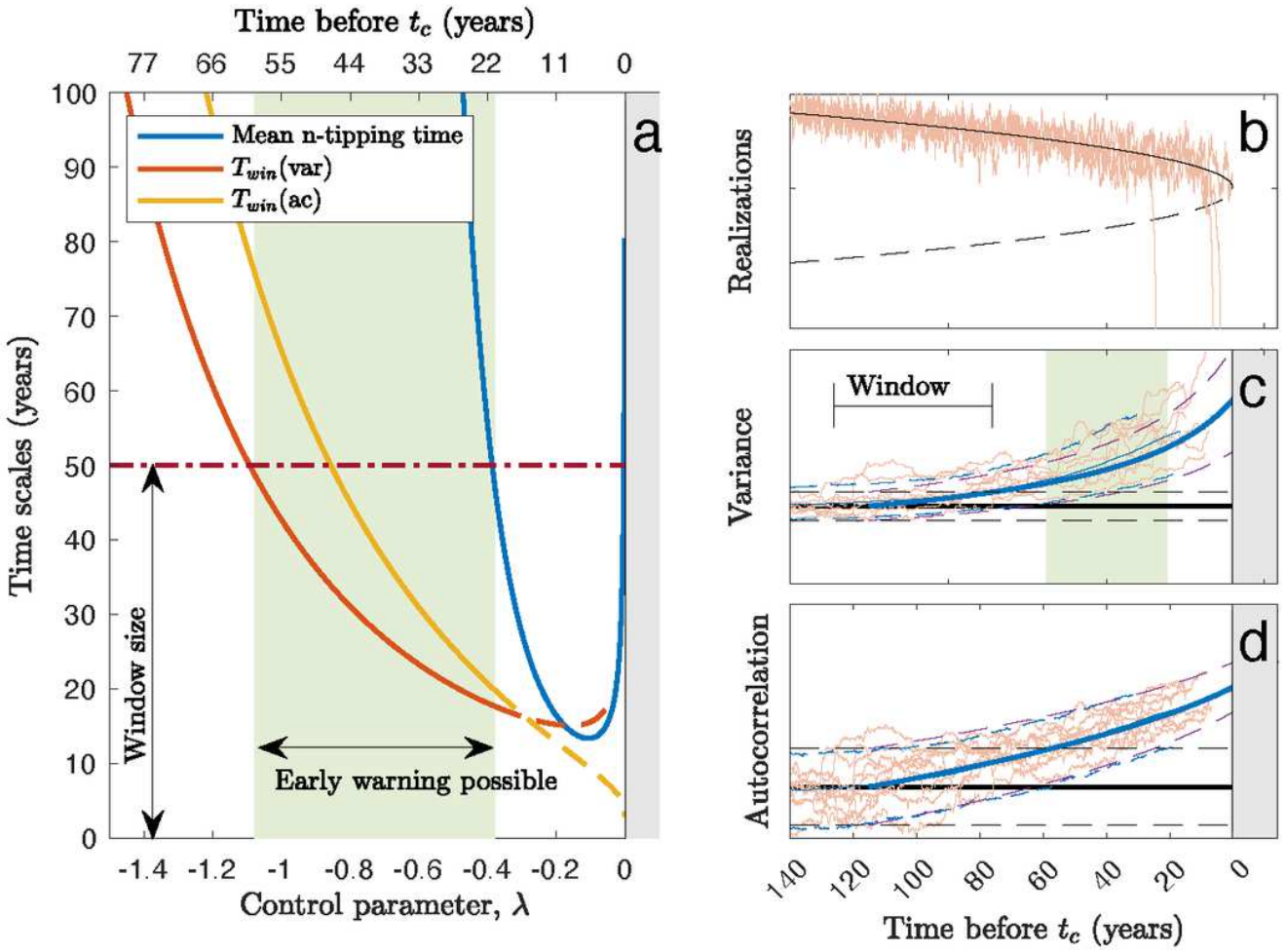

Figure 2

Panel a shows time scales involved in the critical transition ramping the control parameter λ from λ_0 = −2.56 to λ_c = 0, with a ramping time τ_r = 140yrs and σ^2 = 0.28. These parameters are obtained as best estimates from the HasISST data. The time remaining before t_c is shown on top of the plot. The red and orange curves shows the time window, T_w, needed in order to detect increase in variance (red) and autocorrelation (orange) above the pre-ramping values at the 95% confidence level. Close to the bifurcation point, the (quasi-)stationarity approximation becomes less valid, which is indicated by the dashed part of the two curves. It is seen that detecting significant increase in autocorrelation requires a longer data window than detecting a significant increase in variance. With T_w = 50yrs (red dot-dashed line) an increase in

variance can only be detected at the 95% confidence level after the red curve is below the 50yrs level. The blue curve shows the mean waiting time for a noise-induced transition, when this becomes shorter than

the 50yrs level the EWS is no longer relevant, due to n-tipping occurring before tc, thus the range of time, where an EWS can be applied is indicated by the green band (limited by the crossings of the red and blue curves with the size of the window). Panel b shows ten model realizations of the ramped approach to t_c, notice a few n-tippings prior to t_c. The black (black dashed) curve is the stable (unstable) fixed point of the model. Panel c shows the increased variance as EWS: Black line is the pre-ramping steady state value, while dashed lines are the two-sigma uncertainty range for calculating variance within the 50yr data window. The blue and dashed blue curves are the same, but for the model approaching the transition. The brown curves correspond to the ten realizations in Panel b, while the green band corresponds to the green band in Panel a. The thin blue lines are the same obtained from simulating 1000 realizations. Panel d is the same as Panel c but for the autocorrelation.

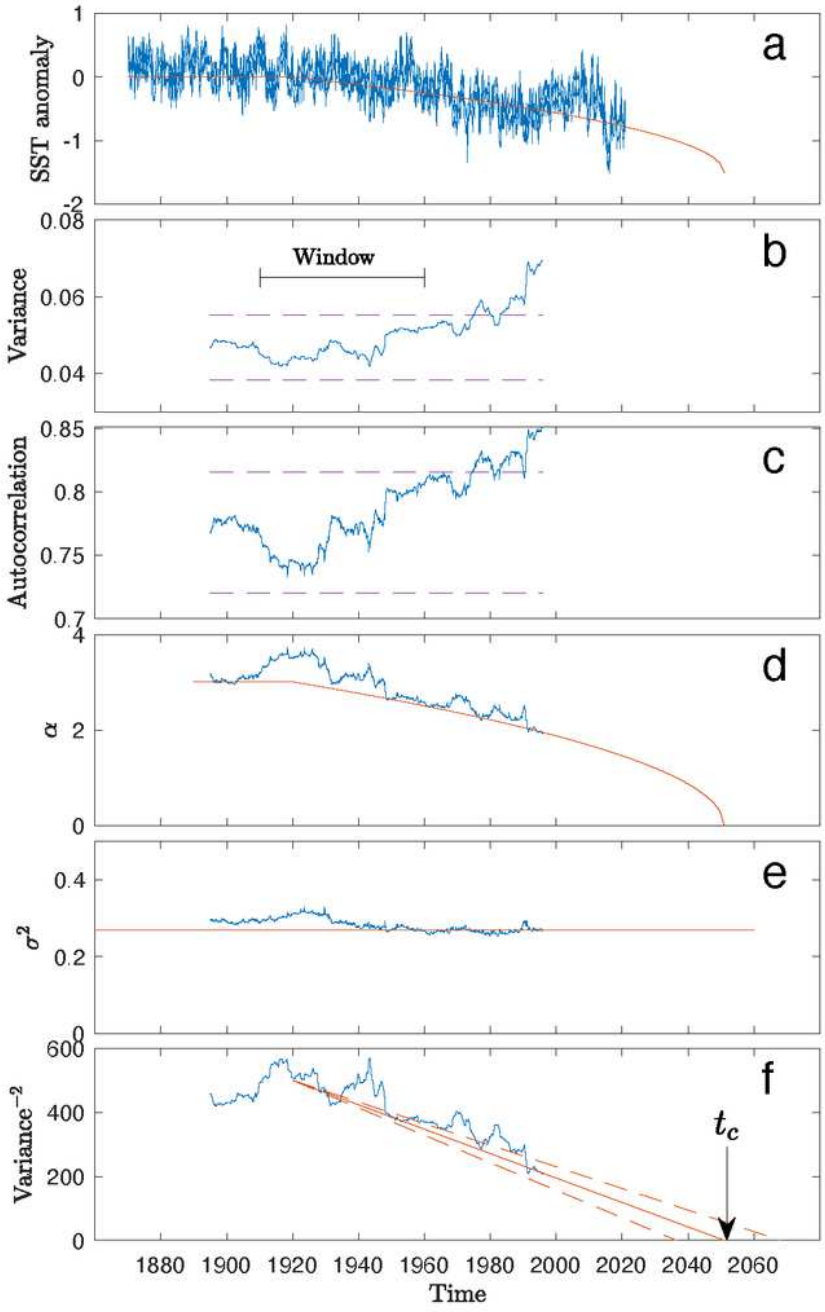

Figure 3

Panel a shows the SST anomaly (identical to Figure 1e) together with best estimate model of the steady state approaching a critical transition. Panels b and c show variance and autocorrelation calculated within running 50yr windows, similar to Figure 2c and d. The two-sigma levels (dashed purple lines) are obtained using the model to estimate the time varying α (Panel d) and σ2 (Panel e) from the data. Panel f shows the best estimate for t_c (see Methods).

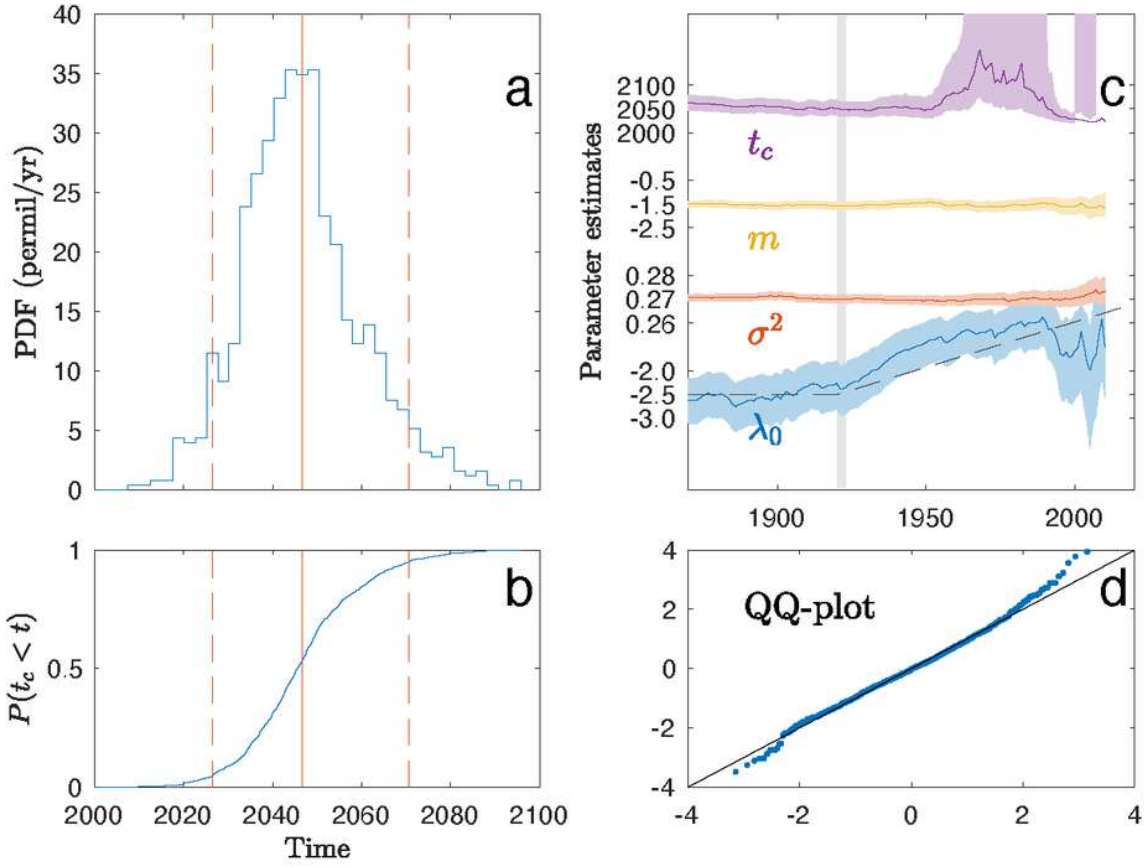

Figure 4

With parameters obtained from the data, a set of 1000 realizations of the model are used in bootstrap study to obtain a probability density for tc. Panel a is probability density, Panel b is the cumulative distribution. The mean is t_c = 2047 (vertical red line) and the 90% confidence interval is 2026-2071 (vertical dashed lines). Panel c shows the evolution of the four parameter estimates when they are estimated from data from year t_0 indicated on the x-axis up to year 2020. The 95% confidence intervals (shadow) around the full drawn lines are based on the inverse Fisher information. The gray vertical line is t_0 = 1920, used in the final estimate. The constancy of the estimates around 1920 shows the estimates are robust to the exact choice of t_0.