

Extrapolation from historical data cannot reliably predict the time of a potential AMOC collapse

Andreas Morr^{1,2*}, Maya Ben-Yami^{2,3}, Brian Groenke²,
Christof Schötz^{2,3}, Alessandro Cotronei⁴, Eirik Myrvoll-Nilsen⁴,
Sebastian Bathiany^{2,3}, Martin Rypdal⁴, Niklas Boers^{2,3,5*}

¹Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Boltzmannstraße 3, Garching, 85748, Bavaria, Germany.

²Research Domain IV – Complexity Science, Potsdam Institute for Climate Impact Research, Telegrafenberg A 31, Potsdam, 14473, Brandenburg, Germany.

³Munich Climate Center and Earth System Modelling Group, Department of Aerospace and Geodesy, TUM School of Engineering and Design, Technical University of Munich, Lise-Meitner-Straße 9, Ottobrunn, 85521, Bavaria, Germany.

⁴Department of Mathematics and Statistics, Faculty of Science and Technology, UiT The Arctic University of Norway, Hansine Hansens veg 18, Tromsø, 9019, Troms, Norway.

⁵Department of Mathematics and Statistics, University of Exeter, North Park Road, Exeter, UK EX4 4QE, Devon, United Kingdom.

*Corresponding author(s). E-mail(s): andreas.morr@tum.de;
n.boers@tum.de;

Ditlevsen and Ditlevsen [1] (DD23 hereafter) propose a statistical framework to estimate the timing of a potential collapse of the Atlantic Meridional Overturning Circulation (AMOC) based on extrapolating information from observed sea-surface temperature (SST) variability. By fitting a stochastic one-dimensional fold-bifurcation model to an SST-based fingerprint of the AMOC using Maximum Likelihood Estimation (MLE), they conclude that a collapse is most likely to occur in the middle of the 21st century, with a reported 95% confidence interval covering the time span

from 2037 to 2109. Given the profound implications of such a claim for both climate and society, it is essential to thoroughly test the robustness of this result, to critically assess the underlying assumptions and uncertainties, and to estimate the extent to which the reported confidence interval reflects the true limits of current knowledge.

Here we examine the sensitivity of DD23’s results and argue that four types of uncertainty are insufficiently explored in their analysis: (i) structural uncertainty associated with the assumed low-order bifurcation model, (ii) statistical uncertainty in their model fit, (iii) uncertainty in the representativeness of SST-based fingerprints as proxies for the high-dimensional AMOC dynamics, and (iv) uncertainty in the underlying data, arising from non-stationary observational coverage and dataset pre-processing. Our results do not dispute the possibility that the AMOC might abruptly or irreversibly transition to an alternative state in response to anthropogenic forcing, nor that the stability of AMOC has already declined. However, using synthetic experiments and a systematic analysis of alternative fingerprints and observational products, we show that the tipping times estimated by DD23 are highly sensitive to the uncertainties listed above, and extend several millennia into the future when these uncertainties are thoroughly propagated.

Modelling Assumptions and Structural Uncertainty

The predictive framework introduced in DD23 is based on extrapolation of information from spatially averaged historical sea-surface temperature data, assuming an identical linear trend in forcing for past and future. It relies on a specific stochastic differential equation model:

$$\begin{aligned} dX_t &= -(A(X_t - m)^2 + \lambda(t))dt + \sigma dB_t \\ \lambda(t) &= \lambda_0(1 - \Thetat - t_0/\tau_r), \end{aligned} \tag{1}$$

where X_t represents the AMOC strength and $\lambda(t)$ denotes a linearly evolving external forcing that induces a tipping transition at time $t_c = t_0 + \tau_r$. DD23 devise a Maximum Likelihood Estimation (MLE) routine for the time of tipping t_c and other model parameters. Although DD23 demonstrated that this approach provides a good statistical fit to their choice of SST-based AMOC fingerprint, its validity as a predictive tool hinges on the adequacy of the underlying physical assumptions. While we focus on the MLE routine, most points of criticism raised throughout this comment also apply to the running window approach alternatively used in DD23, which relies on the extrapolation of variance and autocorrelation trends.

The representation of the AMOC, a highly complex and spatially extended basin-scale ocean circulation system, as a one-dimensional stochastic equation with time-scale-independent noise, is problematic for predicting the time of a potential future AMOC tipping event. The mathematical justification provided by DD23 for using such a model as a conceptual representation of the AMOC is that all fold bifurcations are topologically equivalent to Eq. (1). However, this is a local result valid only in an infinitesimal neighborhood of the bifurcation point [2]. It may not account for AMOC dynamics during the historical time period, which can deviate significantly from the normal form.

To probe how strongly the DD23 inference depends on the assumed structure of the model, we extend their maximum-likelihood framework by replacing the quadratic fold normal form with a more general cubic approximation of the dynamics (see Methods and Cotronei et al. [3] for details on this adaptation). A cubic drift, i.e. a third-order deterministic term in the model equation (1), is the minimal polynomial extension that can represent both a true fold bifurcation and a reversible, smooth shift of the stable state, without imposing the existence of a tipping point a priori as in the model of DD23. The model structure with a cubic drift also arises naturally from the reduced Stommel-Cessi box model of the thermohaline circulation [4] and, importantly, permits multiple stable states when a tipping point is present. The model with quadratic drift, as proposed in DD23, on the other hand, necessarily has a bifurcation point for all parameter choices and permits only one stable state. When the alternative third-order model is used to construct an MLE for the slope of the ramping parameter λ , in the same vein as DD23, the observed AMOC fingerprint is fitted equally well, if not better (measured by residual squared errors, a visual inspection of the residual quantile-quantile plot, and model comparisons via AIC and BIC, which favor the cubic model with a difference of 10 and 5 in the two criteria, respectively). However, the extrapolated future trajectory changes qualitatively: instead of an imminent fold-type collapse, the best-fit cubic model yields a smooth continuation of the weakening trend (see Fig. 1a). The divergence between the quadratic and cubic projections reflects structural model uncertainty; different, equally plausible low-order representations of the drift lead to fundamentally different assessments of future tipping risks. This demonstrates that the DD23 tipping estimate relies on imposing a normal-form fold-bifurcation geometry a priori, rather than that geometry being uniquely determined from the data.

The consequences of such a structural mismatch are further illustrated in Fig. 1b. Here we apply the original DD23 MLE framework to synthetic time series generated from the best-fit cubic model inferred in Fig. 1a (see Methods for details). By construction, these realizations approximately reproduce the statistical properties of the observed AMOC fingerprint (such as the evolution of its mean state, variance, and autocorrelation, with 75% of realizations passing the significance test of DD23 in terms of their increases above baseline) but do not exhibit a true tipping point when evolved forward in time. Nevertheless, when the DD23 fold-bifurcation model is fitted to the observed period of each realization of the model with third-order drift, the likelihood optimization systematically yields finite tipping times clustered around the mid-21st century. This behavior demonstrates that the DD23 framework can produce spurious tipping-time estimates when applied to data generated by a model not exhibiting a bifurcation in the present parameter configuration, and hence not exhibiting tipping dynamics. While the model with third-order drift deviates from the normal-form structure assumed by DD23, its sample time series still pass goodness-of-fit metrics with respect to the normal-form structure, like the residual quantile-quantile analysis proposed by DD23.

In addition to the specific points regarding structural uncertainty discussed above, several other possible model deviations should be discussed. For example, alternative choices of the stochastic term in Eq. (1) can lead to non-stationary time-series statistics

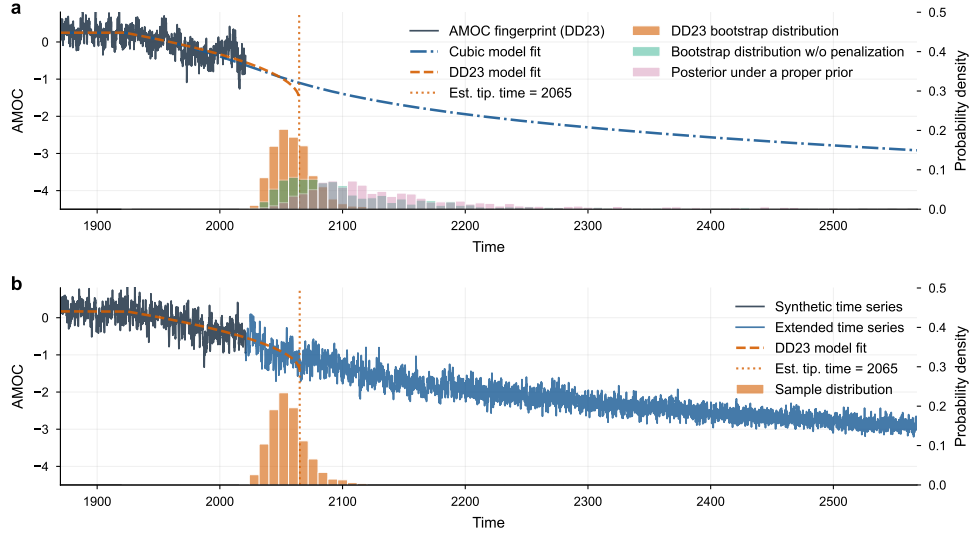


Fig. 1 Structural uncertainty in tipping-time estimation. (a) Analysis of the AMOC fingerprint proposed by DD23 using both their MLE method and an adaptation with a third-order drift term. If the bifurcation model is chosen such that tipping is not mechanically prescribed but optional (switching from a second-order to a third-order drift term, see Eqs. (2) and (3) in the Methods below), the MLE fitting routine suggests no tipping in finite time at all. In contrast to the originally proposed drift by DD23, the third-order drift term has the additional advantage that it in principle allows for two alternative AMOC stable states. Distributions are obtained by (i) the bootstrap routine proposed by DD23 for uncertainty quantification, (ii) the same bootstrap routine without using a penalization on the parameter A (an accumulation of estimates around the year 2800 is not depicted here), and (iii) a Bayesian MCMC routine using a statistically proper and less informative prior on A . (b) Synthetic AMOC time series generated from the model with third-order drift inferred in (a). Applying the DD23 method, assuming the second-order drift (Eq. (2)), to the black portion of these time series produces spurious tipping-time estimates around the mid-21st century, despite the absence of a true transition. The histogram shows the sample distribution of this estimation over 1000 independent realizations of the model time series.

independently of any AMOC weakening [5, 6]. We point the reader to Figure 1 in Ben-Yami et al. [7], which shows that DD23’s method predicts a tipping time even for a linear system subjected to non-linear forcing and non-stationary correlated noise. It is also plausible that the recent downward curvature in the AMOC SST fingerprint time series used by DD23 just reflects the impact of a strongly non-linear external forcing, for example, the global warming trend that has accelerated during the 20th century, or non-linear increases in freshwater inflow to the North Atlantic. This trend in forcing may not extrapolate well into the future, thus biasing estimates of the tipping time under their model.

Failures of the Confidence Assessment

One of the main results of DD23 is the prediction that AMOC tipping will occur “with high confidence ... as soon as mid-century” [1, p.7], supported by a 95% confidence

interval (CI) for the tipping time of 2037–2109 (revised from 2025–2095 in a correction). We identify two reasons why this statement of confidence is misleading: First, we show via a simulation study that their CI procedure, based on a penalized MLE and parametric bootstrap, does not achieve the nominal 95% uniform coverage across the full parameter space, in fact yielding 0% coverage under the original implementation and 45% under our improved implementation for some parameter configurations. Second, we argue that, even if nominal coverage were to be achieved, DD23’s interpretation of their CI as providing a likely range of future tipping times (see Figure 7 caption in [1]) is not justified under their chosen statistical framework. All numerical experiments in this section build on the original stochastic differential equation model of DD23 in Eq. (1)

Regarding the first reason, we note that a 95% CI procedure for the tipping time parameter t_c is defined for a family of data generating processes $\mathbb{P}_{\phi, t_c}$ indexed by unknown parameters $(\phi, t_c) \in \Phi$ as any method of computing two numbers L and R from the generated data such that the probability of L being smaller and R being larger than t_c is at least 95% uniformly over all parameter vectors $(\phi, t_c) \in \Phi$. This can be stated formally as

$$\inf_{(\phi, t_c) \in \Phi} \mathbb{P}_{\phi, t_c}(L \leq t_c \leq R) \geq 95\%.$$

In DD23, $\phi = (A, m, \lambda_0, \sigma, t_0)$ and $\mathbb{P}_{\phi, t_c}$ is defined by Equation (1) using $\tau_r = t_c - t_0$. No non-trivial restrictions are given for the parameter space Φ . The CI procedure (L, R) is based on a parametric bootstrap using an MLE with a data-driven penalty on the parameter A .

Using Monte-Carlo integration, we exactly follow this bootstrapping procedure and approximate $\mathbb{P}_{\phi, t_c}(L \leq t_c \leq R)$ for selected parameter vectors (ϕ, t_c) , see Figure 2 and Methods. Specifically, we choose $m, \lambda_0, \sigma, t_0$ to fit the observational time series in DD23 and vary A and t_c . For $t_c = 3924$ (equivalently $\tau_r = 2000$) and $A = 0.05$, we obtain $\mathbb{P}_{\phi, t_c}(L \leq t_c \leq R) \approx 0$ using the original code for the CI procedure. Since the original implementation directly imposes $A \geq 0.1$ through an ad hoc constraint, we additionally run a modified implementation of the CI procedure that still enforces the requirement $A > 0$ while allowing values below 0.1 (see Methods) and obtain $\mathbb{P}_{\phi, t_c}(L \leq t_c \leq R) \approx 45\%$. Hence, the method used in DD23 is not a 95% CI procedure and can severely underestimate sampling uncertainty. Moreover, for most parameter configurations considered, the left limit L is lower than the true parameter t_c with probability $\approx 100\%$. Thus, any non-conservative CI procedure (L^*, R^*) for a centered 95% confidence interval, $\mathbb{P}_{\phi, t_c}(L^* \leq t_c) = \mathbb{P}_{\phi, t_c}(t_c \leq R^*) = 97.5\%$, would necessarily yield larger lower (and upper) bounds for these parameters.

While the apparent bias of the CI procedure depends on the parameters A and t_c and appears to diminish when evaluated at the parameter estimates reported in DD23 ($A = 0.86$; $t_c = 2065$, equivalently $\tau_r = 141$), the failure of the CI procedure is not confined to extreme cases (Figure 2b,c,e,f). For example, with $A = 0.4$ and $t_c = 2174$ (equivalently $\tau_r = 250$) the procedure yields intervals whose endpoints are at most as large as the estimated ones, 2037 and 2109, respectively, with $\approx 15\%$ probability. This occurs despite the true value being 2174, in clear violation of the 95% confidence

requirement. We therefore find no principled restriction of the parameter space for the observed data under which the procedure achieves valid coverage. Applying the parametric bootstrap with the unpenalized MLE produces coverage much closer to the nominal 95% in our simulations. In the considerations further below and in the Methods section, we show that this penalization introduces an ad hoc but highly influential hyperparameter (Fig. 2g). For the observational time series used by DD23, the unpenalized MLE 95% CI for t_c is 2042–2834 using the original implementation (see green distribution in Fig. 1) and 2042–5108 using our improved implementation. These CIs should be regarded as more accurately capturing the sampling uncertainty of the MLE method.

Regarding the second reason, we argue that, even if the CI were to achieve proper coverage, DD23’s “high confidence” predictions that tipping will occur within the reported interval would still be misleading. The 95% coverage probability of a CI only describes the long-run performance of the procedure used to generate the CI under repeated sampling from the data generating process. This is generally not the same as estimating the probability (or plausibility) that the true value of the parameter (here tipping time) is contained within an interval computed from a single sample [8, 9]. Therefore, colloquial claims of “confidence”, “likelihood”, or otherwise about the future tipping time based on a specific CI, such as DD23’s statement that “transition of the AMOC is most likely to occur around 2037–2109 (95% confidence interval)” [1, p.2], are not supported by the underlying statistical theory.

This could be partially addressed by instead adopting a Bayesian interpretation of the model and estimating the posterior distribution over the tipping time, given the observed fingerprint. However, this would in turn require formal specification of a prior distribution over all parameters of the model, including the tipping time itself. While confidence intervals generated by the parametric bootstrap may, under certain conditions, approximate the Bayesian posterior [10], this correspondence cannot generally be assumed without mathematical or empirical verification. Furthermore, under the Bayesian interpretation, a robust probabilistic forecast of future tipping time would need to include a sensitivity analysis of the predictions to both the choice of model and prior [11].

This is here particularly relevant since the aforementioned penalty term in the likelihood effectively acts as an improper but highly informative prior on the parameter A in Eq. (1). We show that a Bayesian interpretation confirms that DD23’s inferences are sensitive to this choice by using Markov Chain Monte Carlo to obtain an estimate of the posterior distribution over tipping times under an alternative choice of prior for A (see Methods). The alternative prior fulfills the same basic function as DD23’s penalty by down-weighting small values of A while being statistically proper and only weakly informative. We use noninformative uniform priors for all other parameters, including the tipping time. The resulting 95% credible range of tipping times is then 2055–2512 with median $t_c = 2122$ (Fig. 1). This demonstrates that, even under a generous Bayesian interpretation, DD23 still considerably underestimate statistical uncertainty within their own model.

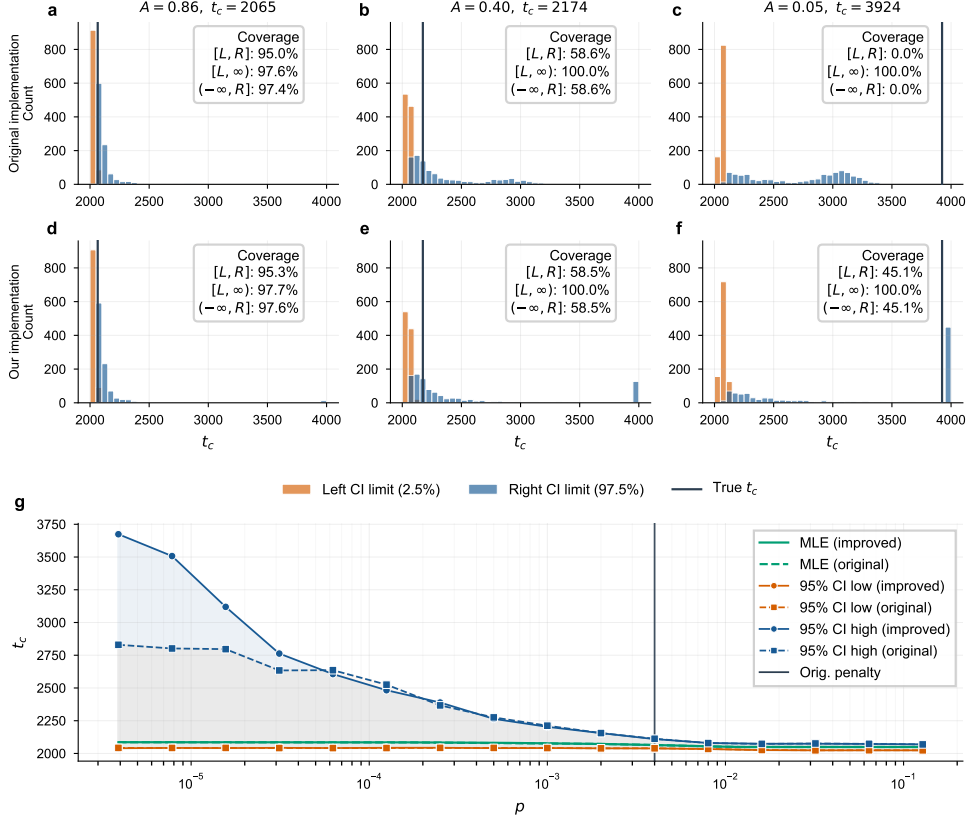


Fig. 2 Evaluation of the confidence interval procedure of DD23. (a–f) Each panel shows histograms of the lower (orange) and upper (blue) confidence limits across 1000 simulation repetitions calculated using the original implementation (top row) or our implementation (bottom row). The panel columns correspond to different ground truth parameter vectors of the simulation, with the true critical time t_c marked by a vertical black line. All values > 4000 are collected in the rightmost bin of each panel. The text box in each panel reports the empirical coverage for the corresponding parameter vector, broken down into the two-sided centered interval and the two one-sided intervals. (g) MLE and bootstrap confidence intervals as a function of penalty weight. We compute the penalized MLE (green) for the critical time t_c and the parametric bootstrap 95% confidence intervals (orange and blue) from DD23 on the observed time series across a range of penalty weights p using the original implementation and our improved one. Both orange and green lines coincide in this plot, respectively. The penalty weight $p = 0.004$ adopted in DD23 is indicated in black. The upper confidence limit exhibits strong sensitivity to the choice of penalty weight.

Representativeness of Proxies for Underlying Dynamics

DD23 rely on a specific sea-surface temperature (SST) fingerprint, based on averaging SSTs over the subpolar gyre (SPG) region. We note that in calculating the fingerprint they omitted a step in the latitudinal area weighting of the gridded SST data,

and included grid cells with sea ice coverage where SSTs are at the freezing temperature and hence not informative of ocean heat transport (see Methods and Figure 3). When these issues are corrected, DD23’s fingerprint does not pass their initial test for significantly high variance and lag-one autocorrelation. More importantly, they subtract $2\times$ the global mean SST to obtain their specific one-dimensional proxy for the Atlantic Meridional Overturning Circulation (AMOC) strength. This differs from the established version of the SPG AMOC fingerprint introduced by [12], where only $1\times$ the global mean is subtracted. DD23 justify this different fingerprint with two arguments: First, they contend that due to polar amplification, the warming trend in the SPG region is twice that of the global mean temperature increase, and thus more must be subtracted. Second, DD23 optimize $SST_{SPG} - \beta SST_{global}$ with respect to β by fitting the fingerprint to hydrographic section data, and quoting $\beta = 1.95 \approx 2$ as the optimal value for the coefficient.

The physical basis of the AMOC SPG SST fingerprint is that a reduction in AMOC strength causes the SPG SSTs to cool relative to how much they would have warmed without AMOC weakening. Thus, the subtraction of some coefficient of the global mean SST timeseries needs to account for the amplification of the global SST warming trend at the SPG’s latitudes. To calculate a rough approximation of this coefficient, we may consider the warming at the same latitude band of SSTs in the Pacific Ocean relative to the global SST warming trend. When we average the SST trend amplification for the SST dataset used by DD23 (1870-2021 HadISST) in the area $40\text{--}65^\circ\text{N}$ and $130^\circ\text{E}\text{--}110^\circ\text{W}$, the value is $\beta = 1.02$, much smaller than the twofold amplification used by DD23 to justify the $2\times$ global mean subtraction.

Regarding the justification of a factor 2 via more direct observations, DD23’s Figure 7 [1] uses the seasonally corrected hydrographic sections from Kanzow et al. 2010 [13] (Figure 3, solid pink), which DD23 refer to as the “ MOC_z ” data (citing [14]). It should be noted that there exists no obvious way to convert monthly AMOC streamfunction data to an SST value. DD23 use the 3.8 Sv/K slope from Figure 5 of [12], which was introduced as a conversion rate between centennial AMOC streamfunction trends to centennial SST trends. Furthermore, Figure 3 shows two more possible versions of the MOC_z data. The first (dashed pink) is the same converted data shifted down by 0.3 K , to make it visually similar to the data plotted in DD23’s Figure 7. The second (solid purple) is an updated seasonal correction of the hydrographic section observations, using 2004-2024 RAPID-MOCHA data [15] instead of the 2004–2008 data used by [13]. None of these versions of the MOC_z data are a good fit to any version of the SST fingerprint (Figure 3). In particular, when we minimize $SST_{SPG} - \beta SST_{global}$ with respect to β using our own SST_{SPG} and SST_{global} calculations, the best fit for the centred MOC_z data from [13] is $\beta = 0.29$, and the best fit for the 0.3K shifted data is $\beta = 0.92$ (Figure 3). None of these calculations should be taken to mean that subtracting one time the global mean SST from the SPG SSTs results in an optimal AMOC fingerprint. They demonstrate, however, that we find no compelling reason to modify the traditional SPG AMOC fingerprint by subtracting a different coefficient of the global mean SST.

More generally, the AMOC is a complex, spatially extended system with many degrees of freedom, and it is an oversimplification to assume its future tipping time is

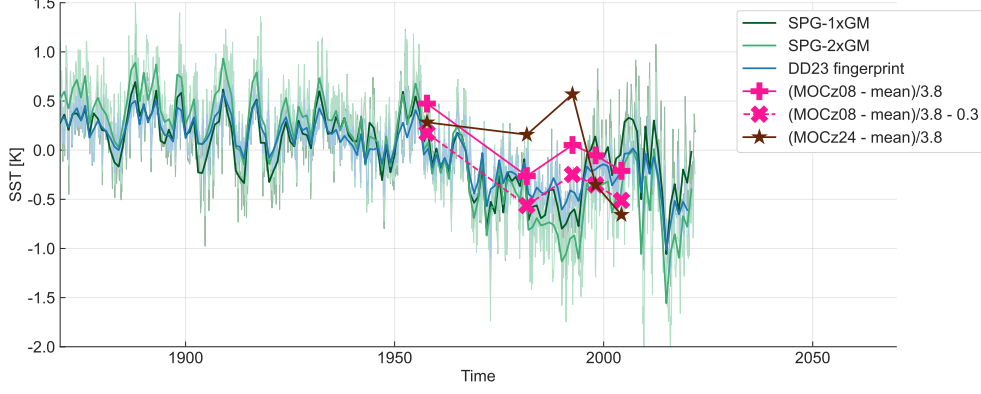


Fig. 3 Comparison of SST fingerprints and MOC_z data The MOC_z data [16] comprises five values for the AMOC streamfunction at $26^\circ N$ calculated from hydrographic sections. Kanzow et al. 2010[13] corrected the seasonal aliasing of the observations using RAPID data from 2004-2008 (MOC_z08), which DD23 compare to the SST fingerprints by converting to K using a $3.8Sv/K$ factor [12] (pink pluses, solid line). To visually reproduce DD23’s Figure 7, we shifted the data down by approximately $0.3K$ (pink exes, dotted line). We additionally correct the seasonal aliasing with RAPID data from 2004-2024 (MOC_z24 , brown star, solid line). While DD23 report that the best fit to the shifted MOC_z08 data is approximately their SPG-2xGMT fingerprint (blue), when the correct data processing method (see Methods) is used, the best fit is approximately SPG-1xGMT (dark green), not SPG-2xGMT (light green). The annual mean (thick line) is plotted over the monthly fingerprint values (thin line).

fully encoded in a single one-dimensional temperature time series. While the “warming hole” in the North Atlantic has been linked to AMOC weakening [17–21], the correlation between the SPG-based index and the actual AMOC streamfunction is non-stationary and sensitive to the time period and forcing scenario considered [22, 23]. This non-stationarity is plausible because the North Atlantic warming hole is not solely caused by ocean circulation; it likely results from a complex interplay of ocean heat transport, atmospheric forcing, and aerosol-induced radiative anomalies [24–28]. This is further complicated by recent observations from the OSNAP program, which suggest that the SPG region plays a different role in deep water formation than previously assumed [29, 30], casting doubt on the reliability of this fingerprint as a predictive tool. Additionally, evidence suggests that the SPG itself may possess an independent tipping point separate from the AMOC [31–33]. Consequently, as DD23 themselves concede, the early warning signals (EWS) detected in the SPG-based index may represent a localized transition rather than a collapse of the global overturning circulation [34].

When the MLE method of DD23 is applied to the dipole index [35, 36], an alternative physically motivated AMOC fingerprint, the resulting tipping-time estimates span centuries to millennia (see Fig. 4). When calculated from the same SST dataset as DD23’s data, the conventionally used fingerprint constructed by subtracting $1\times$ rather than $2\times$ the global mean SST from the SPG SST [12] yields an estimated tipping time around 2100, as noted by DD23 themselves. We emphasize that many other fingerprints could be constructed [36], and no single choice is uniquely justified.

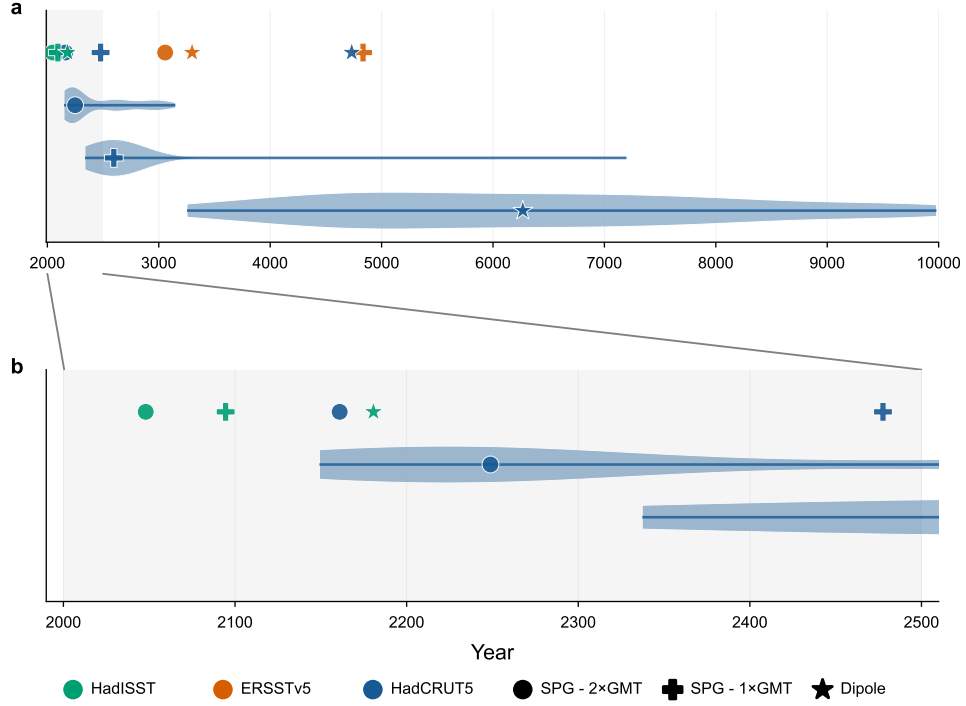


Fig. 4 Range of tipping times in alternative fingerprints and data. Tipping times estimated using DD23’s MLE method. The best estimate of the tipping time is calculated for the classical SPG-based AMOC index (circle), the fingerprint used by DD23 (plus), and the Dipole index (star). We use three different observational SST datasets for this analysis: HadISST1 (turquoise), ERSSTv5 (orange), and HadCRUT5 (blue). Additionally, the blue violins display the tipping times for each of the 200-member uncertainty ensembles of HadCRUT5-derived fingerprints. Note that the modified fingerprint by DD23 consistently shows the earliest tipping time, partly due to subtracting $2\times$ the global mean SST rather than only one time. This Figure is similar to Fig. 4 in Ben-Yami et al. [7], accounting for the error-corrected version of DD23’s MLE method and now also determining the optimal penalization term for each individual time series.

Dataset Preprocessing and Non-stationary Coverage

The reliability of tipping time prediction is further undermined by potential biases in the historical climate data used. In particular, methods for predicting tipping times, including the MLE approach used by DD23, build on higher-order statistics like variance and autocorrelation, which are often unreliable due to the non-stationarity of observational coverage and the preprocessing methods used to fill data gaps [34, 37, 38].

Specifically, historical SST records have seen an exponential growth in the number of observations [39, 40]. In earlier years, measurements were sparse and concentrated in specific areas, requiring sophisticated infilling procedures to produce globally complete datasets. The HadISST1 dataset used by DD23 employs Reduced Space Optimal Interpolation (RSOI) [41], which regularizes the fit toward the mean in regions and

times where less data is available. This process has the effect of damping the variance in the earlier part of the record where observations are more sparse. As observational density increases toward the mid-20th century, the captured variability grows, resulting in an artificial trend in variance (and potentially other higher-order statistics) that can be misinterpreted as a signal of critical slowing down and, hence, of stability loss [34, 41, 42].

The sensitivity of tipping time estimates to these artifacts becomes clear when comparing different observational products. Alternative datasets, such as ERSSTv5 [43] and HadCRUT5 [44], employ different infilling and bias-correction procedures, leading to different trends in variance and autocorrelation [7]. Applying the DD23 methodology to fingerprints constructed from the ERSSTv5 or HadCRUT5 datasets yields a range of tipping times spanning thousands of years (see Fig. 4). Given that the variance increase in HadISST1 is known to be at least partially artificial, any prediction relying on the precise evolution of this statistic lacks the necessary robustness to support a definitive warning of collapse in the mid-21st century.

Furthermore, applying the tipping-time estimation to the HadCRUT5 uncertainty ensemble yields a similarly wide spread of inferred collapse times (violins in Fig. 4). Because this ensemble is designed to reflect fundamental limitations in historical observational accuracy, the resulting dispersion in tipping-time estimates represents an irreducible source of uncertainty. Such uncertainty cannot be curtailed by improved statistical fitting alone and should therefore be explicitly incorporated into any comprehensive uncertainty assessment.

Concluding Remarks

The results presented here highlight that the tipping-time estimates reported in DD23 strongly depend on specific modelling assumptions and data choices. The uncertainty ranges associated with these assumptions and choices are only marginally assessed by DD23, leading to reported confidence intervals in their study that are considerably biased and too narrow. When structural uncertainty in the dynamical model, parametric and statistical uncertainties, ambiguity in the physical meaning of SST-based AMOC fingerprints, and non-stationary observational artifacts and measurement uncertainties are taken into account, the inferred collapse times span several millennia. In addition, we have shown that using an equally plausible model leads to a configuration without tipping, for which DD23’s method would still falsely predict a tipping time in the near future. The narrow confidence intervals reported in DD23 reflect uncertainty conditional on all the choices discussed above, rather than the full uncertainty relevant for real-world prediction.

Our analysis does not preclude the possibility of a future AMOC collapse in any way, nor does it diminish the importance of continued monitoring of AMOC stability and immediate action to mitigate the risks posed by anthropogenic climate change. However, it demonstrates that current observational records and simplified statistical models do not support a meaningful quantitative prediction of a tipping time. Claims of a specific time frame of collapse should therefore not be communicated to the public or policymakers, who could mistake them for reliable or actionable knowledge.

Competing Interests: The authors declare no competing interests.

Data Availability: All code for the discussed data analyses and simulation studies are available upon request.

Author Contributions: AM, MBY, BG, AC, CS, EMN, SB, MR, and NB conceived the study. AM, MBY, BG, AC, CS, and EMN developed the methods. AM, MBY, BG, and CS carried out the analyses. All authors contributed to the writing and revision of the manuscript.

Acknowledgements: This is ClimTip contribution #146; the ClimTip project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101137601. This project is funded by the Advanced Research + Invention Agency (ARIA). NB acknowledges additional funding by the VolkswagenFoundation. B.G. acknowledges funding from the Past to Future (P2F) project under the European Union’s Horizon Europe research and innovation programme grant agreement No. 101184070: views and opinions are solely that of the authors and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). The authors gratefully acknowledge the Ministry of Research, Science and Culture (MWFK) of Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research.

References

- [1] Ditlevsen, P.D., Ditlevsen, S.: Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications* **14**(1), 4254 (2023) <https://doi.org/10.1038/s41467-023-39810-w>
- [2] Kuznetsov, Y.A.: *Elements of Applied Bifurcation Theory*, 4th edn. Applied Mathematical Sciences, vol. 112. Springer, New York (2004). <https://doi.org/10.1007/978-1-4757-3978-7>
- [3] Cotronei, A., Myrvoll-Nilsen, E., Rypdal, M.: Evaluating model uncertainty in critical threshold estimations from time series data: Application to the Atlantic meridional Overturning Circulation. *Frontiers in Climate* **8**, 1761461 (2026) <https://doi.org/10.3389/fclim.2026.1761461>
- [4] Cessi, P.: A simple box model of stochastically forced thermohaline flow. *Journal of Physical Oceanography* **24**(9), 1911–1920 (1994) [https://doi.org/10.1175/1520-0485\(1994\)024\(1911:ASBMOS\)2.0.CO;2](https://doi.org/10.1175/1520-0485(1994)024(1911:ASBMOS)2.0.CO;2)
- [5] Morr, A., Boers, N.: Detection of Approaching Critical Transitions in Natural Systems Driven by Red Noise. *Physical Review X* **14**(2) (2024) <https://doi.org/10.1103/PhysRevX.14.021037>
- [6] Morr, A., Riechers, K., Gorjão, L.R., Boers, N.: Anticipating critical transitions

- in multidimensional systems driven by time- and state-dependent noise. *Physical Review Research* **6**(3), 033251 (2024) <https://doi.org/10.1103/PhysRevResearch.6.033251>
- [7] Ben-Yami, M., Morr, A., Bathiany, S., Boers, N.: Uncertainties too large to predict tipping times of major Earth system components from historical data. *Science Advances* **10**(31), 4841 (2024) <https://doi.org/10.1126/sciadv.adl4841>
 - [8] Neyman, J.: Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences* **236**(767), 333–380 (1937) <https://doi.org/10.1098/rsta.1937.0005>
 - [9] Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.-J.: The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* **23**(1), 103–123 (2015) <https://doi.org/10.3758/s13423-015-0947-8>
 - [10] Efron, B.: Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics* **6**(4) (2012) <https://doi.org/10.1214/12-AOAS571>
 - [11] Berger, J.O., Moreno, E., Pericchi, L.R., Bayarri, M.J., Bernardo, J.M., Cano, J.A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., Dasgupta, A., Gustafson, P., Wasserman, L., Kadane, J.B., Srinivasan, C., Lavine, M., O’Hagan, A., Polasek, W., Robert, C.P., Goutis, C., Ruggeri, F., Salinetti, G., Sivaganesan, S.: An overview of robust Bayesian analysis. *Test* **3**(1), 5–124 (1994) <https://doi.org/10.1007/BF02562676>
 - [12] Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature* **556**(7700), 191–196 (2018) <https://doi.org/10.1038/s41586-018-0006-5>
 - [13] Kanzow, T., Cunningham, S.A., Johns, W.E., Hirschi, J.J.-M., Marotzke, J., Baringer, M.O., Meinen, C.S., Chidichimo, M.P., Atkinson, C., Beal, L.M., Bryden, H.L., Collins, J.: Seasonal variability of the atlantic meridional overturning circulation at 26.5°N. *Journal of Climate* **23**(21), 5678–5698 (2010) <https://doi.org/10.1175/2010JCLI3389.1>
 - [14] Frajka-Williams, E., Ansorge, I.J., Baehr, J., Bryden, H.L., Chidichimo, M.P., Cunningham, S.A., Danabasoglu, G., Dong, S., Donohue, K.A., Elipot, S., Heimbach, P., Holliday, N.P., Hummels, R., Jackson, L.C., Karstensen, J., Lankhorst, M., Le Bras, I.A., Susan Lozier, M., McDonagh, E.L., Meinen, C.S., Mercier, H., Moat, B.I., Perez, R.C., Piecuch, C.G., Rhein, M., Srokosz, M.A., Trenberth, K.E., Bacon, S., Forget, G., Goni, G., Kieke, D., Koelling, J., Lamont, T., McCarthy, G.D., Mertens, C., Send, U., Smeed, D.A., Speich, S., Berg, M., Volkov, D., Wilson, C.: Atlantic meridional overturning circulation: Observed transport and variability. *Frontiers in Marine Science* **6**(JUN), 1–18 (2019) <https://doi.org/10.3389/fmars.2019.00260>

- [15] Frajka-Williams, E., Moat, B.I., Smeed, D.A., Rayner, D., Johns, W.E., Baringer, M.O., Volkov, D., Collins, J.: Atlantic meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heatflux Array-Western Boundary Time Series) array at 26N from 2004 to 2020 (v2020.1) (2021). <https://doi.org/10.5285/cc1e34b3-3385-662b-e053-6c86abc03444>
- [16] Bryden, H.L., Longworth, H.R., Cunningham, S.A.: Slowing of the atlantic meridional overturning circulation at 25° n. *Nature* **438**, 655–657 (2005) <https://doi.org/10.1038/nature04385>
- [17] Rahmstorf, S., Box, J.E., Feulner, G., Mann, M.E., Robinson, A., Rutherford, S., Schaffernicht, E.J.: Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nature Climate Change* **5**(5), 475–480 (2015) <https://doi.org/10.1038/nclimate2554>
- [18] Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature* **556**(7700), 191–196 (2018) <https://doi.org/10.1038/s41586-018-0006-5>
- [19] Drijfhout, S., Oldenborgh, G.J., Cimatoribus, A.: Is a decline of AMOC causing the warming hole above the North Atlantic in observed and modeled warming patterns? *Journal of Climate* **25**(24), 8373–8379 (2012) <https://doi.org/10.1175/JCLI-D-12-00490.1>
- [20] Menary, M.B., Wood, R.A.: An anatomy of the projected North Atlantic warming hole in CMIP5 models. *Climate Dynamics* **50**(7-8), 3063–3080 (2018) <https://doi.org/10.1007/s00382-017-3793-8>
- [21] Liu, W., Fedorov, A.V., Xie, S.P., Hu, S.: Climate impacts of a weakened Atlantic meridional overturning circulation in a warming climate. *Science Advances* **6**(26), 2–10 (2020) <https://doi.org/10.1126/sciadv.aaz4876>
- [22] Little, C.M., Zhao, M., Buckley, M.W.: Do Surface Temperature Indices Reflect Centennial-Timescale Trends in Atlantic Meridional Overturning Circulation Strength? *Geophysical Research Letters* **47**(22), 1–10 (2020) <https://doi.org/10.1029/2020GL090888>
- [23] Jackson, L.C., Wood, R.A.: Hysteresis and Resilience of the AMOC in an Eddy-Permitting GCM. *Geophysical Research Letters* **45**(16), 8547–8556 (2018) <https://doi.org/10.1029/2018GL078104>
- [24] Li, L., Lozier, M.S., Li, F.: Century-long cooling trend in subpolar North Atlantic forced by atmosphere: an alternative explanation. *Climate Dynamics* **58**(9-10), 2249–2267 (2022) <https://doi.org/10.1007/s00382-021-06003-4>
- [25] He, C., Clement, A.C., Cane, M.A., Murphy, L.N., Klavans, J.M., Fenske, T.M.: A

- North Atlantic Warming Hole Without Ocean Circulation. *Geophysical Research Letters* **49**(19) (2022) <https://doi.org/10.1029/2022GL100420>
- [26] Ferster, B.S., Simon, A., Fedorov, A., Mignot, J., Guilyardi, E.: Slowdown and Recovery of the Atlantic Meridional Overturning Circulation and a Persistent North Atlantic Warming Hole Induced by Arctic Sea Ice Decline. *Geophysical Research Letters* **49**(16) (2022) <https://doi.org/10.1029/2022GL097967>
- [27] Ghosh, R., Putrasahan, D., Manzini, E., Lohmann, K., Keil, P., Hand, R., Bader, J., Matei, D., Jungclaus, J.H.: Two distinct phases of North Atlantic Eastern Subpolar Gyre and Warming Hole evolution under Global Warming. *Journal of Climate* **36**(6), 1–31 (2022) <https://doi.org/10.1175/JCLI-D-22-0222.1>
- [28] Keil, P., Mauritsen, T., Jungclaus, J., Hedemann, C., Olonscheck, D., Ghosh, R.: Multiple drivers of the North Atlantic warming hole. *Nature Climate Change* **10**(7), 667–671 (2020) <https://doi.org/10.1038/s41558-020-0819-8>
- [29] Lozier, M.S., Li, F., Bacon, S., Bahr, F., Bower, A.S., Cunningham, S.A., De Jong, M.F., De Steur, L., DeYoung, B., Fischer, J., Gary, S.F., Greenan, B.J.W., Holliday, N.P., Houk, A., Houpert, L., Inall, M.E., Johns, W.E., Johnson, H.L., Johnson, C., Karstensen, J., Koman, G., Le Bras, I.A., Lin, X., Mackay, N., Marshall, D.P., Mercier, H., Oltmanns, M., Pickart, R.S., Ramsey, A.L., Rayner, D., Straneo, F., Thierry, V., Torres, D.J., Williams, R.G., Wilson, C., Yang, J., Yashayaev, I., Zhao, J.: A sea change in our view of overturning in the subpolar North Atlantic. *Science* **363**(6426), 516–521 (2019) <https://doi.org/10.1126/science.aau6592>
- [30] Chafik, L., Holliday, N.P., Bacon, S., Rossby, T.: Irminger Sea Is the Center of Action for Subpolar AMOC Variability. *Geophysical Research Letters* **49**(17), 1–11 (2022) <https://doi.org/10.1029/2022GL099133>
- [31] Sgubin, G., Swingedouw, D., Drijfhout, S., Mary, Y., Bennabi, A.: Abrupt cooling over the North Atlantic in modern climate models. *Nature Communications* **8** (2017) <https://doi.org/10.1038/ncomms14375>
- [32] Swingedouw, D., Bily, A., Esquerdo, C., Borchert, L.F., Sgubin, G., Mignot, J., Menary, M.: On the risk of abrupt changes in the North Atlantic subpolar gyre in CMIP6 models. *Annals of the New York Academy of Sciences* **1504**(1), 187–201 (2021) <https://doi.org/10.1111/nyas.14659>
- [33] McKay, D.I.A., Staal, A., Abrams, J.F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S.E., Rockström, J., Lenton, T.M.: Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science* **377**(6611) (2022) <https://doi.org/10.1126/science.abn7950>
- [34] Ben-Yami, M., Skiba, V., Bathiany, S., Boers, N.: Uncertainties in critical slowing

- down indicators of observation-based fingerprints of the Atlantic Overturning Circulation. *Nature Communications* **14**(1), 8344 (2023) <https://doi.org/10.1038/s41467-023-44046-9>
- [35] Roberts, C.D., Garry, F.K., Jackson, L.C.: A multimodel study of sea surface temperature and subsurface density fingerprints of the Atlantic meridional overturning circulation. *Journal of Climate* **26**(22), 9155–9174 (2013) <https://doi.org/10.1175/JCLI-D-12-00762.1>
 - [36] Jackson, L.C., Wood, R.A.: Fingerprints for Early Detection of Changes in the AMOC. *Journal of Climate* **33**(16), 7027–7044 (2020) <https://doi.org/10.1175/JCLI-D-20-0034.1>
 - [37] Smith, T., Zotta, R.M., Boulton, C.A., Lenton, T.M., Dorigo, W., Boers, N.: Reliability of resilience estimation based on multi-instrument time series. *Earth System Dynamics* **14**(1), 173–183 (2023) <https://doi.org/10.5194/esd-14-173-2023>
 - [38] Boers, N.: Reply to: Evidence lacking for a pending collapse of the Atlantic Meridional Overturning Circulation. *Nature Climate Change* **14**(January), 43–48 (2023) <https://doi.org/10.1038/s41558-023-01878-z>
 - [39] Lundstad, E., Brugnara, Y., Pappert, D., Kopp, J., Samakinwa, E., Hürzeler, A., Andersson, A., Chimani, B., Cornes, R., Demarée, G., Filipiak, J., Gates, L., Ives, G.L., Jones, J.M., Jourdain, S., Kiss, A., Nicholson, S.E., Przybylak, R., Jones, P., Rousseau, D., Tinz, B., Rodrigo, F.S., Grab, S., Domínguez-Castro, F., Slonosky, V., Cooper, J., Brunet, M., Brönnimann, S.: The global historical climate database HCLIM. *Scientific Data* **10**(1), 1–16 (2023) <https://doi.org/10.1038/s41597-022-01919-w>
 - [40] Kennedy, J.J., Rayner, N.A., Atkinson, C.P., Killick, R.E.: An Ensemble Data Set of Sea Surface Temperature Change From 1850: The Met Office Hadley Centre HadSST.4.0.0.0 Data Set. *Journal of Geophysical Research: Atmospheres* **124**(14), 7719–7763 (2019) <https://doi.org/10.1029/2018JD029867>
 - [41] Rayner, N.A., Parker, D.E., Horton, E.B., Folland, C.K., Alexander, L.V., Rowell, D.P., Kent, E.C., Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres* **108**(14) (2003) <https://doi.org/10.1029/2002jd002670>
 - [42] Smith, T., Zotta, R.-M., Boulton, C.A., Lenton, T.M., Dorigo, W., Boers, N.: Reliability of resilience estimation based on multi-instrument time series. *Earth System Dynamics* **14**(1), 173–183 (2023) <https://doi.org/10.5194/esd-14-173-2023>
 - [43] Huang, B., Thorne, P.W., Banzon, V.F., Boyer, T., Chepurin, G., Lawrimore, J.H., Menne, M.J., Smith, T.M., Vose, R.S., Zhang, H.M.: Extended reconstructed

- Sea surface temperature, Version 5 (ERSSTv5): Upgrades, validations, and inter-comparisons. *Journal of Climate* **30**(20), 8179–8205 (2017) <https://doi.org/10.1175/JCLI-D-16-0836.1>
- [44] Morice, C.P., Kennedy, J.J., Rayner, N.A., Winn, J.P., Hogan, E., Killick, R.E., Dunn, R.J.H., Osborn, T.J., Jones, P.D., Simpson, I.R.: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres* **126**(3), 1–28 (2021) <https://doi.org/10.1029/2019JD032361>
 - [45] Pilipovic, P., Samson, A., Ditlevsen, S.: Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics* **52**(2) (2024) <https://doi.org/10.1214/24-AOS2371>
 - [46] ter Braak, C.J.F., Vrugt, J.A.: Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing* **18**(4), 435–446 (2008) <https://doi.org/10.1007/s11222-008-9104-9>
 - [47] Hartig, F., Minunno, F., Paul, S.: BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics. Comprehensive R Archive Network (2017). <https://doi.org/10.32614/CRAN.package.BayesianTools>
 - [48] Pérez-Hernández, M.D., Hernández-Guerra, A., Cana-Cascallar, L., Arumí-Planas, C., Caínzos, V., González-Santana, A.J., Gutiérrez-Guerra, M.A., Martínez-Marrero, A., Mosquera Giménez, A., Presas Navarro, C., Santana-Toscano, D., Vélez-Belchí, P.: The seasonal cycle of the eastern boundary currents of the north atlantic subtropical gyre. *Journal of Geophysical Research: Oceans* **128**(4), 2022–019487 (2023) <https://doi.org/10.1029/2022JC019487> <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022JC019487>. e2022JC019487 2022JC019487

Methods

To assess the robustness of the tipping time predictions presented in DD23, we applied their Maximum Likelihood Estimation (MLE) method to a suite of synthetically generated time series, and combinations of alternative SST-based AMOC fingerprints and observational datasets. We additionally extend the likelihood formulation by replacing the quadratic drift term in Eq. (2) with a more general third-order polynomial representation of the deterministic dynamics, introduced in Eq. (3) below.

Third-order drift extension of the maximum likelihood method

We implemented an alternative statistical model within the DD23 maximum-likelihood framework by replacing the quadratic fold normal form drift with a cubic polynomial, following the formulation introduced in Cotronei et al. [3]. Instead of the DD23

stochastic differential equation (SDE)

$$dX_t = -(A(X_t - m)^2 + \lambda(t)) dt + \sigma dB_t, \quad (2)$$

we used a model with a cubic deterministic drift

$$dX_t = -(A(X_t - m)^3 + C(X_t - m) + \lambda(t)) dt + \sigma dB_t, \quad (3)$$

which corresponds to Eq. (12) in [3]. Here X_t denotes the AMOC fingerprint, m a constant reference state, A and C shape the deterministic equilibrium branch, $\lambda_t = \lambda_0(t - t_0)$ is the externally forced control parameter, and σ is the noise amplitude. In contrast to the original drift used by DD23, this formulation does not strictly impose the existence of a fold bifurcation: depending on the fitted parameters, the equilibrium branch may exhibit either a fold or a smooth inflection.

Parameter estimation was performed by maximum likelihood using analogous discretized transition densities and the same numerical optimization procedure as in DD23. The likelihood was computed analogously to the ansatz of DD23 using Strang splitting of the SDE [45], and the free parameters ($A = 0.29, C = 1.68, \tau_r = 137.49$) were estimated by minimizing the negative log-likelihood over the observed SPG-SST-based AMOC time series. For more information on the exact routine, see the Supplementary Material of Cotronei et al. [3]. The time of onset of the external forcing, $t_0 = 1924$, was kept identical to DD23.

Synthetic time series from the model with third-order drift

To test the sensitivity of the DD23 maximum-likelihood framework to the model choice, we generated synthetic time series from the stochastic model with cubic drift defined in Eq. (3), using the parameter values inferred from the application of the MLE method with cubic drift to the observed AMOC fingerprint. This setup ensures that the synthetic data are statistically consistent with the observations, while originating from dynamics that do not necessarily contain a tipping point.

We generated 1000 independent realizations using an Euler-Maruyama discretization with twenty integration time steps per month. Each realization was subsequently coarse-grained to observational resolution and analyzed using the original DD23 MLE procedure, without modification. This constitutes a controlled experiment in which the data-generating process is known, but the inference model is structurally misspecified. Following DD23's approach, we also tested whether the synthetic time series showed statistically significant differences in variance and lag-1 autocorrelation from their baseline values at the 5% level. This ensures that the time series would not be rejected based on non-significant critical slowing down indicators (for more details, see the middle of page 4 in DD23). Approximately 75% of model-generated realizations passed this criterion, identical to the rate seen for the model proposed by DD23.

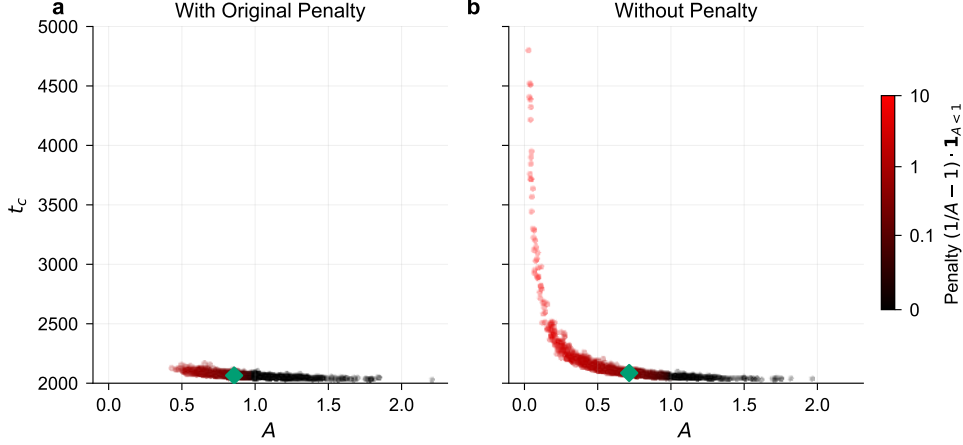


Fig. 5 Maximum likelihood estimates of the parameters A and t_c obtained during the parametric bootstrap confidence interval procedure of DD23, applied to their AMOC fingerprint. Each dot corresponds to one of the 1000 synthetic bootstrap samples; the green diamond marks the MLE on the original observations. Panel (a) shows results with the original penalty on A , and panel (b) without it. Dot colour indicates the penalty value $(1/A - 1) \cdot \mathbf{1}_{A < 1}$ at the estimated A . The penalty effectively suppresses large t_c estimates by discouraging values of A far below unity. High estimates with $t_c > 5000$ (reaching up to 30,000) are not depicted here.

Analysis of the bootstrap confidence interval procedure

A plausible explanation of the failure of the CI procedure used in DD23 with respect to the claimed 95% coverage probability is the use of an additive penalty $p \cdot (\frac{1}{A} - 1) \cdot \mathbf{1}_{A < 1}$ on the approximate likelihood used in the MLE. The penalty is only mentioned in the supplementary information, where the authors justify its use by stating: *Since division by A enters the calculations of λ_0 and m and thus the pseudo-likelihood, estimates are sensitive to small values of A . We therefore regularize the optimization problem by adding a penalization term on small values of A .* We do not see why this would justify the use of a penalty as the MLE is invariant under transformations such as $B = \frac{1}{A}$ or $B = \log A$ and the MLE based on B does not seem to justify the equivalent penalty $p \cdot (B - 1) \cdot \mathbf{1}_{B > 1}$. As there is empirically a negative dependence between A and t_c (Figure 5), the penalty introduces a bias towards smaller values of t_c in the MLE. The reported CIs are thus doubly biased by the penalty: the biased MLE estimates are used to specify the data-generating process from which bootstrap samples are drawn, and the penalized MLE is then applied to each bootstrap sample to obtain the bootstrap distribution of t_c , thereby amplifying the bias.

Moreover, the data-driven choices of t_0 and p are not accounted for in the bootstrap procedure that produces the CI. The penalty weight p in particular is highly influential on the results, as shown in Figure 2g. Specifically, p is chosen as follows: first, apply the unpenalized MLE; then simulate many (we use 1000) trajectories from the SDE Equation (1) using the resulting parameter estimates; finally, optimize p over a grid of values to minimize the mean squared error between the true ramping time τ_r and its

penalized-MLE estimate. This selection procedure is not documented in DD23 or in the published code, where p is hardcoded to 0.004. The supplementary information only states that p was *determined by cross-validation on simulated data sets by minimizing the mean squared distance between the estimated ramping time on each data set and the value of the ramping time used in the simulation, with optimal value $p = 0.004$* . We obtained a description of the algorithm from the authors via personal communication. Our implementation reproduces the penalty weight $p = 0.004$ reported in DD23 when applied to DD23’s fingerprint time series data, confirming a faithful replication of their method.

To define a data generating process from (1), we need to set the parameters $t_0, \alpha_0, \mu_0, \sigma^2, A, \tau_r$, where $\tau_r = t_c - t_0$, $\lambda_0 = -\alpha_0^2/(4A)$, $m = \mu_0 - \alpha_0/(2A)$ following the SI and provided code of DD23. For $t_0, \alpha_0, \mu_0, \sigma^2$ we use the values estimated in DD23 from the observational data, i.e., $t_0 = 1924$, $\alpha_0 \approx 3.06$, $\mu_0 \approx 0.248$, $\sigma^2 \approx 0.296$. The other estimated parameters in DD23 are $A \approx 0.856$ and $\tau_r \approx 141$. Aside from this parameter combination, we evaluate $(A, \tau_r) \in \{(0.4, 250), (0.05, 2000)\}$. For a given parameter vector, we sample a time series of simulated observations from Equation (1). Then we fix the estimate of t_0 to the simulation’s true value, 1924 (in DD23 this value is determined by their first method, the moment estimator), and follow the second method of DD23: We use the MLE without penalty for an initial parameter estimate, and cross-validation on 1000 time series simulated with this parameter vector to determine the penalty weight p . We then run the penalized MLE and the subsequent parametric bootstrap procedure as described in DD23, yielding one realization of the CI. We repeat the entire procedure 1000 times for each of the three parameter vectors described above. From these 1000 realizations of the CI, we approximate its distribution and arrive at the numbers reported above and in Figure 2.

In the publicly available part of the code used in DD23, a general purpose optimizer is used to obtain the MLE. To ensure $A > 0$ in the likelihood, the authors transform the input argument of the implemented likelihood function via $A \leftarrow \max(0.1, A)$, i.e., all inputs $A < 0.1$ are treated as 0.1. If the likelihood is indeed maximized at a value $A < 0.1$, the optimizer therefore returns $A = 0.1$ (to be more precise, the return value is the first value of A that is tested and fulfills $A \leq 0.1$, which typically is a value just below 0.1). The reason behind this functionality seems to be the requirement that $A > 0$. But effectively truncating the range of possible values of A from below by 0.1 can introduce a severe bias that also strongly influences the estimation of t_c as shown in Figure 2. We therefore provide an alternative implementation that enforces $A > 0$ by reparametrizing it as $A' = \log(A)$, optimizing the likelihood over $A' \in \mathbb{R}$, and transforming back via $A = \exp(A')$. Since the MLE is invariant under monotone reparametrization, this yields identical results (up to numerical precision) when the artificial barrier $A \geq 0.1$ is not triggered and is more faithful to the theoretical estimator in DD23 otherwise. See Figure 2 for a comparison of the bootstrapped CIs from both implementations.

We verified with the authors of DD23 that our interpretation of the CI procedure aligns with their intended meaning—in particular, that the coverage requirements for the CI are independent of the EWS test (significant increase of variance and autocorrelation). Additional checks that incorporate a dependence on the EWS test also reveal

a failure in coverage (not shown here). Similarly, repeating the above procedure with the penalty weight fixed at the value used in DD23, namely $p = 0.004$, also results in a failure in coverage. Finally, one could also subject the data-driven choice of t_0 to evaluation; however, fixing it to its true simulation value, as we do in our simulation study, should only improve the precision of the MLE and the CI.

Bayesian estimation of tipping times

We adopt a Bayesian interpretation of DD23’s model by assigning explicit priors to all free parameters in Eq. (1) as well as the Ornstein-Uhlenbeck (OU) process used to model the stationary period. In order to fully propagate the uncertainty from all model parameters, we combine their two-step inference procedure into a single joint probability model over the stationary $X_{t \leq t_0}$ and nonstationary $X_{t > t_0}$ transitions,

$$\begin{aligned} X_{t > t_0} &\sim \text{TP}(A, \tau_r, \alpha, \mu_t, \sigma^2), \\ X_{t \leq t_0} &\sim \mathcal{N}(X_{t-1}\rho + \mu_0(1 - \rho), \gamma^2(1 - \rho^2)), \\ \alpha &\sim \mathcal{U}(0, 10), \\ \mu_0 &\sim \mathcal{U}(-10, 10), \\ \sigma_0 &\sim \mathcal{U}(0, 10), \\ \tau_r &\sim \mathcal{U}(0, 10^4), \end{aligned}$$

where $\gamma^2 = \sigma_0^2/2\alpha$ is the conditional variance, $\rho = \exp(-\alpha\Delta)$ is the autocorrelation, and Δ is the time step size ($\frac{1}{12}$ years). As noted by DD23, the transition density for the non-stationary part, TP, is not available in closed form and is thus approximated using the Strang splitting scheme as implemented by DD23 (see corrected supplement of [1]). The combination of the stationary and non-stationary segments into a single model is justified by observing that using only point estimates of α , σ_0 , and μ_0 , as done by DD23, neglects uncertainty in the estimate of the OU parameters. We assign non-informative uniform priors to all other model parameters except for A . The parameter ranges for the uniform priors are chosen to provide numerical stability while avoiding influence on the resulting posterior. Note that the use of a double-bounded uniform prior for the ramping parameter assumes that tipping will occur within the next 10,000 years; this reflects an assumption of DD23’s model, i.e. that tipping time must occur within a finite time span. For the parameter, A , we consider three cases: first, we apply the same improper penalty term described by DD23, $\log p(A) = -0.004n(1/A - 1)$ with n equal to the number of time steps in the nonstationary period. Second, we consider the case where the prior is set to be $A \sim \mathcal{U}(0.1, 10)$ and exclude the penalty term. In both of these cases, we use the original implementation provided by DD23 which arbitrarily clips values of $A < 0.1$. Finally, we consider an alternative prior $\log A \sim \mathcal{N}(0, 2)$, which applies a similar but more modest penalty to small (and large) values of A and assigns zero density to $A = 0$ using the corrected implementation of the likelihood described in the previous section. For each model configuration, we draw 10,000 approximate samples from the posterior distributions conditioned on the original fingerprint published by DD23 using the Differential Evolution Markov Chain Monte Carlo (DE-MCMC) sampler of [46] as implemented by the `BayesianTools`

R library [47]. An additional 90,000 initial samples are simulated and discarded as burn-in to ensure convergence of the Markov chains, which is assessed via standard diagnostics (Gelman-Rubin $\hat{R} < 1.1$ and effective sample size > 1000).

When using DD23’s original choice of improper prior, we obtain a posterior credible interval for t_c of 2039-2109 with median $t_c = 2065$, which is similar to their bootstrap confidence interval. This should be expected due to the strong influence of the penalty term. Under the uniform prior with $A > 0.1$, we obtain a posterior credible interval of 2050-2350 with median $t_c = 2110$. Results for the alternative prior using the corrected implementation of the likelihood (which should be regarded as the most reliable) are reported in the main text. We also tried running MCMC on DD23’s original code without clipping (i.e. $A \sim \mathcal{U}(0, 10)$); however, the Markov chains failed to converge within 100,000 iterations ($\hat{R} > 1.1$).

Corrections to DD23’s fingerprint calculation

An important practical point, in addition to their specific choice of fingerprint, is that DD23 make two numerical errors in the construction of the fingerprint time series. First, in their latitudinal weighting during area-averaging of SST in the SPG region, they multiply the SST values by the gridcell weights but do not divide by the sum of the weights (i.e. $\sum wx$ instead of $\sum wx / \sum w$). Additionally, instead of removing sea ice gridcells from the HadISST1 data, they set the value of all sea ice gridcells to -1.8 °C, and include these datapoints in the SPG and global SST means. As the relationship between ocean heat transport (OHT) and sea ice is qualitatively different from that between OHT and SST, there is no fixed temperature value that can include sea ice in an AMOC fingerprint based on SSTs. Thus, when constructing our fingerprints from HadISST1 data we do not include sea ice cells. While correcting for these inconsistencies does not substantially alter the estimated tipping times, the increases in variance and autocorrelation become statistically insignificant according to the test in DD23.

MOC_z data

Bryden et al. 2005 [16] estimated the strength of the overturning at 26N using five hydrographic sections from October 1957, August-Sep 1981, July-August 1992, February 1998, and April 2004. Following DD23, we call this set of datapoints the “MOC_z” data (as opposed to MOC_z generally representing the strength of the AMOC in depth space). Since the months of these observations introduce a seasonal aliasing to the data [14], Kanzow et al. 2010 [13] corrected the original MOC_z data by subtracting the seasonal anomalies of the upper mid-ocean transport (T_{UMO}), which they calculated from 2004-2008 RAPID-MOCHA [15] data. However, updated observations of T_{UMO} seasonal cycle reveal a more uneven cycle than that used by [13], with a pronounced decrease in transport in Oct-Dec [48]. Following [13], we calculate the T_{UMO} seasonal anomalies from 2004-2024 RAPID-MOCHA data, and subtract the corresponding values from the original MOC_z, resulting in a different AMOC reproduction.

As DD23 have not published the code for calculating their fingerprint optimization or for plotting their Figure 7, we have approximated their process as described in the

main text, by (i) subtracting the mean of the MOC_z data, (ii) dividing the data by 3.8 Sv/K [12], (iii) shifting the result down by around 0.3 K. While this shifted data is closer to the SST fingerprints, when those fingerprints are correctly calculated, both versions of the MOC_z data are a better fit to the SPG-1xGMT fingerprint than to SPG-2xGMT.

Alternative fingerprints and datasets

We extended the analysis beyond the HadISST1 dataset used in DD23 to include two additional observational products: ERSSTv5 [43] and HadCRUT5 [44]. For each dataset, we computed three distinct fingerprints:

1. The index defined in DD23 (SPG SSTs minus twice the global mean SST).
2. The classical SPG-based AMOC index (SPG SSTs minus the global mean SST) [18].
3. The dipole fingerprint, calculated as the difference between averaged SSTs in the North Atlantic ($45^\circ\text{--}80^\circ\text{N}$, $70^\circ\text{W--}30^\circ\text{E}$) and the South Atlantic ($0^\circ\text{--}45^\circ\text{S}$, $70^\circ\text{W--}30^\circ\text{E}$) [35, 36].

Additionally, we investigated the propagation of observational uncertainty by considering a 200-member uncertainty ensemble of HadCRUT5 [44]. The tipping time t_c was estimated by maximizing the likelihood of the observed discrete-time increments under the statistical fold-bifurcation model of DD23. For the HadCRUT5 dataset, observational uncertainty was propagated by applying the MLE procedure independently to each of the 200 ensemble members.

To ensure comparability with DD23, we used the same time-interval settings (starting in 1870, with forcing initiated in 1924). We determined an optimal penalization term for each of the observation time series as described by DD23 in personal correspondence. This uniform treatment enables a direct assessment of how sensitive the tipping estimate is to choices in data processing and fingerprint construction. For transparency, we did not pre-select time series based on significant trends in conventional early-warning indicators. Indeed, none of the observational fingerprints considered here display a statistically significant increase in variance and autocorrelation at the 5% level according to DD23’s criterion for avoiding spurious tipping time predictions mentioned above. While such a signal was reported in DD23 for their preferred fingerprint, we find that it disappears once the errors in the surface-area weighting and sea-ice handling are corrected (see above).