



Bayesian stability analysis of the AMOC using nested time-dependent autoregressive model

Luc Hallali¹, Eirik Myrsvoll-Nilsen¹, and Christian L. E. Franzke^{2,3}

¹Department of Mathematics and Statistics, UiT The Arctic University of Norway, N-9037 Tromsø, Norway

²Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea

³Department of Integrated Climate System Science, Pusan National University, Busan, Republic of Korea

Correspondence: Luc Hallali (luc.hallali@uit.no)

Abstract.

The Atlantic Meridional Overturning Circulation (AMOC) is a major climate element subject to possible ongoing loss of stability. Recent studies have found evidence of a gradual weakening in circulation, including early warning signals (EWS), such as increased fluctuations and correlation time of the system, which are both known to be indicators of a possible forthcoming tipping point. To assess these changes in statistical behavior we propose a robust and general statistical model based on a second-order autoregressive process with time-dependent parameters that allow for the statistical changes from increased external variability and destabilization to be accounted for separately. We estimate the time evolution of the correlation parameters using a hierarchical Bayesian modeling framework which also yields uncertainty quantification through the posterior distribution. To assess possible changes in AMOC stability we apply the model to an AMOC fingerprint proxy based on the Sub-Polar Gyre and the global mean temperature anomaly. We find statistically significant EWS which suggests that AMOC is indeed undergoing a loss of stability and is getting closer to a tipping point. The methodology developed in this study is made publicly available as an extension of the R-package `INLA.ews`.

1 Introduction

The Atlantic Meridional Overturning Circulation (AMOC) is a key driver of Earth's climate, responsible for the large-scale transport of heat and water masses across the Atlantic Ocean (Rahmstorf, 1995; Boettner and Boers, 2022). This circulation system is widely believed to have multi-stable states (Stommel, 1961; Lenton et al., 2008), including a strong mode, which is currently dominant, and a weak mode, which has been associated with past climate disruptions. Paleoclimate records suggest that AMOC's abrupt shifts may have played a major role in past climate variability, especially for the Dansgaard-Oeschger events' temperature fluctuations that occurred during glacial periods (Henry et al., 2016; Boers et al., 2018). Some climate models indicate that the AMOC could undergo a critical transition if freshwater input into the North Atlantic reaches a certain threshold (Wood et al., 2019; Hawkins et al., 2011; Weijer et al., 2019). This behavior, known as hysteresis, implies that once the AMOC weakens beyond a tipping point, it may not recover even if initial conditions are restored. However, the exact conditions required for such a transition remain uncertain. While earlier climate models suggested that an AMOC collapse this century is very unlikely, more recent models show a wider range of possible responses to global warming projections, raising



25 concerns about underestimated risks due to model biases like underestimated freshwater input (Masson-Delmotte et al., 2021; Gong et al., 2022; Liu et al., 2017). A weakening or collapse of the AMOC would have far-reaching consequences, including disrupting weather patterns, altering precipitation systems, and potentially triggering cascading effects on other climate components (Jackson et al., 2016; Lenton et al., 2008; Stouffer et al., 2006). To anticipate such changes studies have focused on using critical slowing down theory, stating that when the system is approaching a tipping point its recovery from small perturbations becomes progressively weaker. This phenomenon, called early-warning signal (EWSs), can be characterized by an increased variance and autocorrelation which can be used as statistical indicators of approaching critical transitions.

To detect these statistical changes, a common approach is to start from the linear approximation of a dynamical system with white noise around some stable fixed point x_s , giving

$$dx(t) = -\lambda(x(t) - x_s)dt + \sigma dB(t), \quad (1)$$

35 where $\lambda = -F'(x_s)$ and $dB(t)$ is a white noise process. This Linearization is recognized as the Langevin stochastic differential equation, which has the following solution

$$x(t) = x_0 + \int_{-\infty}^t g(t-s)dB(s), \quad (2)$$

where $g(t-s)$ is a Green's function defined by

$$g(t) = \begin{cases} \exp(-\lambda t), & x \geq 0 \\ 0, & x < 0 \end{cases}. \quad (3)$$

40 This form of $x(t)$ is also referred to as an Ornstein-Uhlenbeck (OU) process. When discretized, this process yields a first-order autoregressive (AR) process.

$$x_t = \phi x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \frac{1-\phi^2}{2\lambda}\sigma^2\right) \quad (4)$$

with variance $\text{Var}(x_t) = \sigma^2/(2\lambda)$ and lag-one autocorrelation parameter $\phi = \exp(-\lambda\Delta t)$.

45 With this model, EWS are detected through an increase of the autocorrelation or variance. However, (Boers, 2021) showed that these indicators can be biased if the system is driven by external noise that itself has increasing autocorrelation or variance leading to false positive alarms. To account for such bias, Boettner and Boers (2022) and Morr and Boers (2024) suggests that the OU process of Eq. (2) should be driven by correlated noise rather than white noise. After discretization the resulting process yields an AR(1) process that is driven by another AR(1) process. Hence the discretization is similar to Eq. (4), except that the white noise process ε_t is replaced by an AR(1) process

$$50 \quad v_{t+1} = \rho v_t + \sigma_v \xi_t \quad (5)$$

with ρ representing the correlation parameter of the noise, σ_v is a scaling parameter and

$$\xi_t \sim \mathcal{N}\left(0, \frac{1-\phi^2}{2\lambda}\right) \quad (6)$$



is a white noise process. It can be shown that this model actually yields a second-order autoregressive (AR) process (Morr and Boers, 2024). This model encompasses the original AR(1) model in (4) when $\rho = 0$ and as showcased in (Boers, 2021) it comprehends cases in which external noise is also correlated $\rho \neq 0$ allowing the parameter ϕ not to be biased in this regard. Consequently, an increasing ϕ will act as a more reliable indicator for detecting EWS.

Climate systems that are prone to tipping, such as the AMOC, are often driven by some external forcing. For the AMOC, the freshwater forcing from Greenland melts acts like a bifurcation parameter as freshwater inputs can disturb the salinity and the temperature of the AMOC, potentially pushing the system closer to its tipping (Wood et al., 2019). To incorporate forcing into our model, we use a similar approach as in (Myrvoll-Nilsen et al., 2020, 2024), where the dynamical system is represented by

$$dx(t) = -\lambda x(t) + F(t)dt + U(t)dt, \quad (7)$$

where $F(t)$ represent the forcing and $U(t)$, as before, represents an OU process. The solution of this equation can be expressed as the sum of one forced component and one noise component

$$x(t) = \nu(t) + \xi(t). \quad (8)$$

Here, the noise component, $\xi(t)$, is represented by an AR(2) process described previously and the forced component, $\nu(t)$, is expressed by

$$\nu(t) = \frac{1}{\sqrt{2\lambda(t)\kappa_f}} \int_0^t F(s) e^{-\lambda(t)(t-s)} ds. \quad (9)$$

This model allows EWS to be detected while accounting for the influence of external forcing on the system's dynamics.

Most studies detect EWS using sliding windows to obtain estimates of the variance and correlation for each window. This approach requires selecting an appropriate window length, which introduces a fundamental compromise. A shorter window provides a more accurate representation of the system's momentary state, but the limited number of data points can reduce the reliability of the statistical estimates. In contrast, a longer window improves the robustness of these estimates by incorporating more data, but it does so at the cost of responsiveness, as it averages information over a broader time scale and may fail to capture short-term fluctuations effectively. Determining the optimal window length is thus a critical but challenging task, as it should ideally balance estimation accuracy with the ability to reflect rapid changes in the system's evolution. Myrvoll-Nilsen et al. (2024) propose an alternative model-based approach that eliminates the need for this choice. Instead of relying on a fixed window length, the correlation parameter is assumed to evolve over time according to a predefined linear structure. This assumption enables a hierarchical Bayesian model formulation, enabling the use of well-established computational techniques to infer the parameters of the linear structure. Furthermore, Myrvoll-Nilsen et al. (2024) adopts a Bayesian framework which offers the additional advantage of providing uncertainty quantification in the form of posterior distributions, making the analysis more robust and interpretable.

In this paper we build upon the hierarchical Bayesian framework developed by Myrvoll-Nilsen et al. (2024) to integrate the AR(2) model proposed by Morr and Boers (2024) and Boettner and Boers (2022). This extension helps mitigate false-alarms caused by correlated noise and eliminates the need for sliding time windows, while benefiting from the advantages of a



85 Bayesian modeling framework. This approach is then applied on an AMOC fingerprint in order to assess its potential loss of stability.

The paper is organized as follows. Section 2 outlines our methodology for Bayesian modeling framework, including details on how inference can be obtained efficiently. In Section 3 we first demonstrate the approach on simulated data and evaluate the accuracy and robustness. Then we apply the methodology to an AMOC fingerprint dataset. Further discussion and conclusions are provided in Section 4.

2 Bayesian modeling

We assume that the observed AMOC fingerprint, $\mathbf{y} = (y_1, \dots, y_n)^\top$ is expressed by

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{x} \quad (10)$$

where the forcing response $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$ is given by

$$95 \quad \mu_t = \sigma_f(t) \sum_{s=0}^t F(s) e^{-\lambda(t)(t-s+0.5)} ds \quad (11)$$

and the correlated time-dependent noise, $\mathbf{x} = (x_1, \dots, x_n)^\top$, is given by

$$\begin{aligned} x_{t+1} &= \phi x_t + v_{t+1} \\ v_{t+1} &= \rho v_t + \sigma_v \xi_t, \end{aligned} \quad (12)$$

which is an AR(2) process (Boers, 2021; Morr and Boers, 2024). To model the evolution of the autocorrelation parameters we assume that they change linearly in time, i.e.

$$100 \quad \begin{aligned} \phi(t) &= a_\phi + b_\phi t, & 0 \leq t \leq 1, \\ \rho(t) &= a_\rho + b_\rho t, & 0 \leq t \leq 1, \end{aligned} \quad (13)$$

expressed by unknown parameters a_ϕ, b_ϕ, a_ρ and b_ρ which are estimated by fitting the model to the observed AMOC fingerprint. Early warning signals due to critical slowing down is characterized through the evolution of $\phi(t)$, while potential changes in external variability is captured by $\mathbf{v} = (v_1, \dots, v_n)^\top$. Separating these signals prevents false alarms as discussed by Boers (2021).

105 To obtain robust uncertainty estimates we adopt a Bayesian framework for parameter estimation, similar to Myrvoll-Nilsen et al. (2024). Given the hierarchical nature of the model, where $\mathbf{y}$ is modeled in terms of $\boldsymbol{\mu}$ and $\mathbf{x}$, which are themselves governed by hyperparameters $\boldsymbol{\theta} = (a_\phi, b_\phi, a_\rho, b_\rho, \sigma_v, \sigma_f)$, a latent Gaussian model formulation provides a natural and efficient framework for Bayesian inference. Both components of the model, $\boldsymbol{\mu}$ and $\mathbf{x}$, depend on the parameters a_ϕ and b_ϕ through $\lambda(t) = -\log \phi(t)$. This dependency introduces a challenge for obtaining reliable inference, as the parameters may be difficult to estimate independently. We therefore choose to model the sum $\boldsymbol{\eta} = \boldsymbol{\mu} + \mathbf{x}$ as a single component. The latent Gaussian model formulation is defined in three stages as follows.



1. The first stage defines the likelihood of the model, which is assumed to be conditionally independent given the latent components μ and x . Since all variation of the AMOC fingerprint observations y is captured by the latent component $\eta = (\eta_1, \dots, \eta_n)^\top$, we model y as Gaussian with mean η and negligible observation noise, $\sigma_y \approx 0$, effectively setting $y \approx \eta$, i.e.

$$\pi(y | \eta, \theta) = \prod_{k=1}^n \pi(y_k | \eta_k, \theta) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_k - \eta_k)^2}{2\sigma_y^2}\right). \quad (14)$$

2. The second stage defines the prior distribution for the latent field η , given parameters θ . This component is assigned a multivariate Gaussian prior distribution with mean vector μ and covariance matrix corresponding to the AR(2) process above with time-dependent $\phi(t)$ and $\rho(t)$,

$$\pi(\eta | \theta) = \mathcal{N}_n(\mu, \Sigma). \quad (15)$$

Since η follows an AR(2) process, its precision matrix, $Q = \Sigma^{-1}$, is a sparse matrix of bandwidth 2. This property enables the use of computationally efficient algorithms that substantially reduce the overall computational cost.

3. The final stage defines the prior distributions for the model parameters

$$\pi(\theta) = \pi(b_\phi) \pi(a_\phi | b_\phi) \pi(b_\rho) \pi(a_\rho | b_\rho) \pi(\sigma_v) \pi(\sigma_f). \quad (16)$$

We assign uniform prior distributions $b_\phi, (a_\phi | b_\phi), b_\rho$ and $(a_\rho | b_\rho)$, and gamma distributions on $\kappa_v = 1/\sigma_v^2$ and $\kappa_f = 1/\sigma_f^2$. Note that since we assume that both $0 < \phi(t) < 1$ and $0 < \rho(t) < 1$ then the parameter space of a_ϕ and a_ρ depend on the current state of b_ϕ and b_ρ , respectively.

The joint posterior distribution for the parameters is given by

$$\pi(v, \theta | y) = \frac{\pi(y | v, \theta) \pi(v | \theta) \pi(\theta)}{\pi(y)}, \quad (17)$$

where $\pi(v)$ is the marginal likelihood, or evidence, of y . In particular, we are interested in the marginal posterior distribution of b_ϕ , which can be obtained by integrating out the other parameters, θ_{-b_ϕ} , and latent variables

$$\pi(b_\phi | y) = \int \pi(\theta, v | y) d\theta_{-b_\phi} dv. \quad (18)$$

Since solving this integral analytically is often not possible in practice, the common approach is to instead approximate it using sampling based approaches like Markov chain Monte Carlo (MCMC) methods (Robert et al., 1999). However, since the precision matrix of the latent Gaussian field is sparse, we can employ a number of computationally efficient algorithms for fast Bayesian inference. Specifically, we evaluate all marginal posterior distributions using the framework of integrated nested Laplace approximations (INLA) (Rue et al., 2009, 2017), which is particularly suited for these types of models. INLA is available as an R package at www.r-inla.org and presents a computationally superior alternative to MCMC. Since our model requires specific implementation using the custom modeling framework of R-INLA we have decided to make the code available as a new feature in the user friendly R-package `INLA.ews` originally developed for the model described in Myrvoll-Nilsen et al. (2024). The AR(2) model can be fitted by prompting:



```
results <- inla.ews(data=y, forcing=z, model = "ar2")
```

A more extensive demonstration of the package can be found in Myrvoll-Nilsen et al. (2024).

3 Results

145 3.1 Accuracy and robustness tests

To evaluate the accuracy and robustness of the proposed time-dependent nested AR(1) model, we perform two simulation-based tests. All estimations are made using R-INLA with the prior distributions described in the previous section. Each test is based on 500 independent simulated time series.

For the first test, we assess whether the model can recover known parameter values when fitted to simulated data generated from the same process (i.e time-dependent nested AR(1)). We use this process to generate 500 time series of length 150, matching the length of the AMOC fingerprint time series used in this paper. For each simulation, the slope parameters b_ϕ and b_ρ are independently drawn from a uniform distribution $\mathcal{U}(-0.9, 0.9)$. Thereafter, the intercepts a_ϕ and a_ρ are drawn from uniform distributions with boundaries that depend on the simulated slope parameters, ensuring that the resulting $\phi(t)$ and $\rho(t)$ remain within the interval $(0, 1)$ for all time steps.

We evaluate the estimation accuracy in two ways. First, we computed the root mean square error (RMSE) between the true slope values and their marginal posterior means $\hat{b}_\phi$ and $\hat{b}_\rho$. The RMSE is 0.145 for b_ϕ and 0.278 for b_ρ . Second, we assess whether the model reliably infers the sign of the slopes by comparing the marginal posterior probabilities $P(b_\phi > 0 | \mathbf{y})$ and $P(b_\rho > 0 | \mathbf{y})$ to the true value of the slopes. We consider the slope for $\phi(t)$ and $\rho(t)$ to be significantly positive if the posterior probabilities exceed 0.95, i.e. $P(b_\phi > 0 | \mathbf{y}) > 0.95$ and $P(b_\rho > 0 | \mathbf{y}) > 0.95$, respectively. If an estimated $\hat{b}_\phi$ is classified as positive, given the $P(b_\phi > 0 | \mathbf{y}) > 0.95$ threshold, we count it as a true positive if the true slope is positive, i.e. $b > 0$. If, however, $b \leq 0$ we count the estimate as a false positive. Conversely, if $P(b_\phi > 0 | \mathbf{y}) \leq 0.95$ we count it as a true negative if $b \leq 0$ and as a false negative if $b > 0$. We also count the classifications for the estimated $\hat{b}_\rho$. The sensitivity and specificity is computed as follows,

$$\text{Sensitivity} = \frac{\# \text{True Positives}}{\# \text{True Positives} + \# \text{False Negatives}}, \quad \text{Specificity} = \frac{\# \text{True Negatives}}{\# \text{True Negatives} + \# \text{False Positives}}. \quad (19)$$

For b_ϕ , the model achieves a sensitivity of 87.7% and a specificity of 99.8%. For b_ρ , the sensitivity is 72.2% and the specificity 99.4%. The results from this test are summarized in Table 1 and illustrated in Fig. 1. Repeating the test with different prior distributions similar to Myrvoll-Nilsen et al. (2024) did not show significant changes, suggesting that the model is robust to the choice of prior distributions.

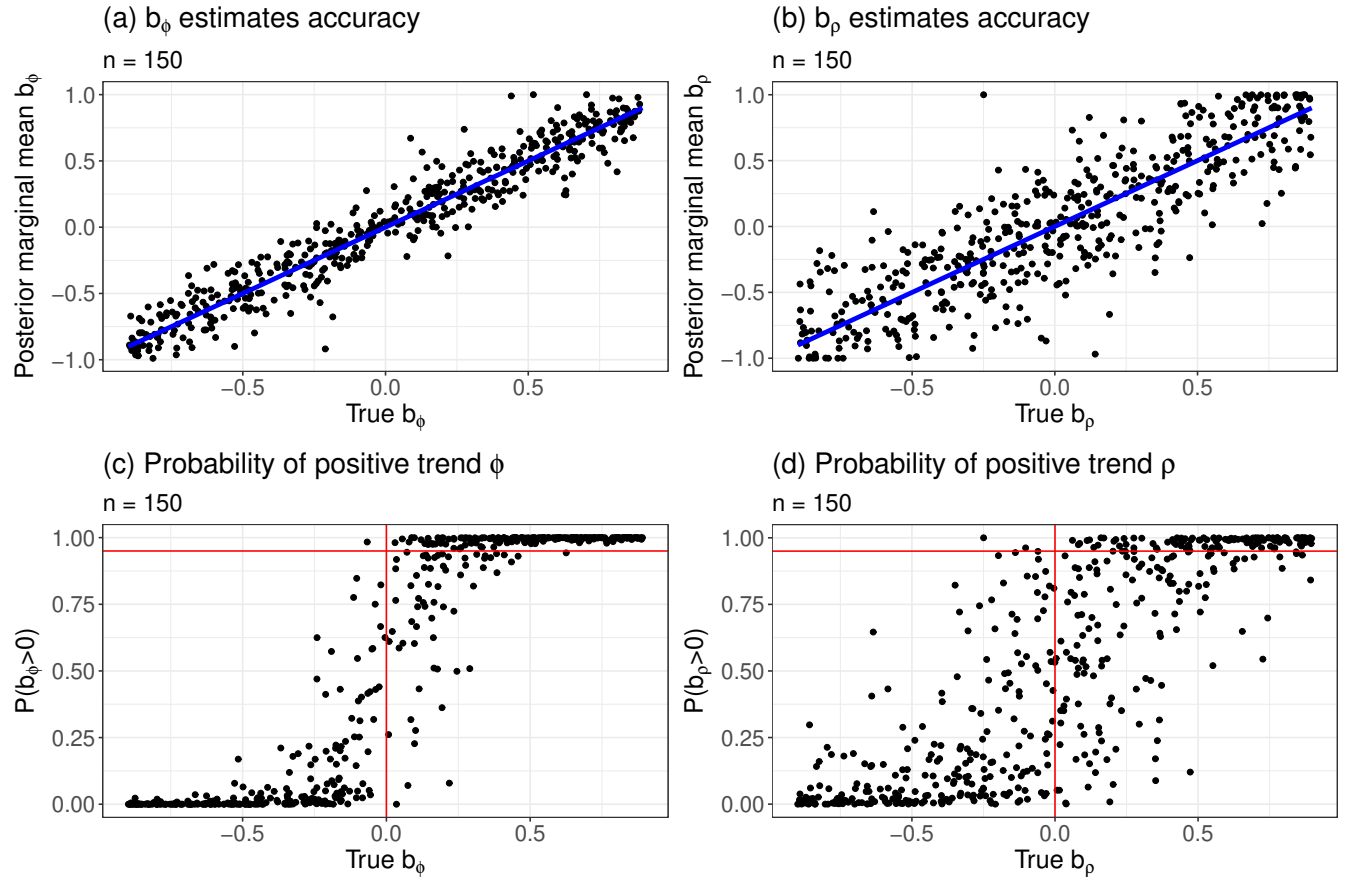


Figure 1. Results of the accuracy test for $n_r = 500$ simulated time series of length $n = 150$. Panels (a) and (b) show posterior marginal mean estimated by INLA for ϕ and ρ , respectively. The blue line shows the true b used in the simulation. Panels (c) and (d) show the estimated posterior probability of the slope being positive against true values of ϕ and ρ respectively. The horizontal red lines separates the true positive and negative values while the horizontal one indicates the probability threshold 0.95 used here to determine statistical significance

In the second test, we evaluate the ability of the model to reliably distinguish genuine early warning signals from changes driven solely by correlated external variability. To do so, we simulate data from two stochastic differential equations. The first one represents a system approaching a tipping point, and the second one remains stable but is influenced by a time-dependent autocorrelated noise. This setup follows the example in Boers (2021). The tipping process is expressed by

$$\dot{x}(t) = -x^3 + x - T + v(t), \quad (20)$$

where T increases linearly from -1 to 1 , and $v(t)$ is a time-dependent AR(1) process with parameters drawn in the same way as in the first test. The non-tipping process is generated by

$$\dot{x}(t) = -5x + v(t), \quad (21)$$



with the same structure for $v(t)$ as in Eq. (20). Each simulation is run until the tipping point is reached (for the tipping system) or for 150 time units (for the non-tipping system), resulting in time series of approximately 150 points. The same inference methodology and classification thresholds are used here, with the distinction that an early warning signal is said to be detected when $P(b_\phi > 0 | \mathbf{y}) > 0.95$. For the tipping processes the model correctly detected an EWS signal in 471 out of 500 simulations, corresponding to a sensitivity of 94.2%. For the non-tipping processes, 23 out of 500 simulations were incorrectly classified as EWS, resulting in a specificity of 95.4%. These results, presented in Table 1 and Fig. 2, indicate that the model effectively identifies true loss of stability while maintaining a low false positive ratio, even in the presence of strongly autocorrelated noise.

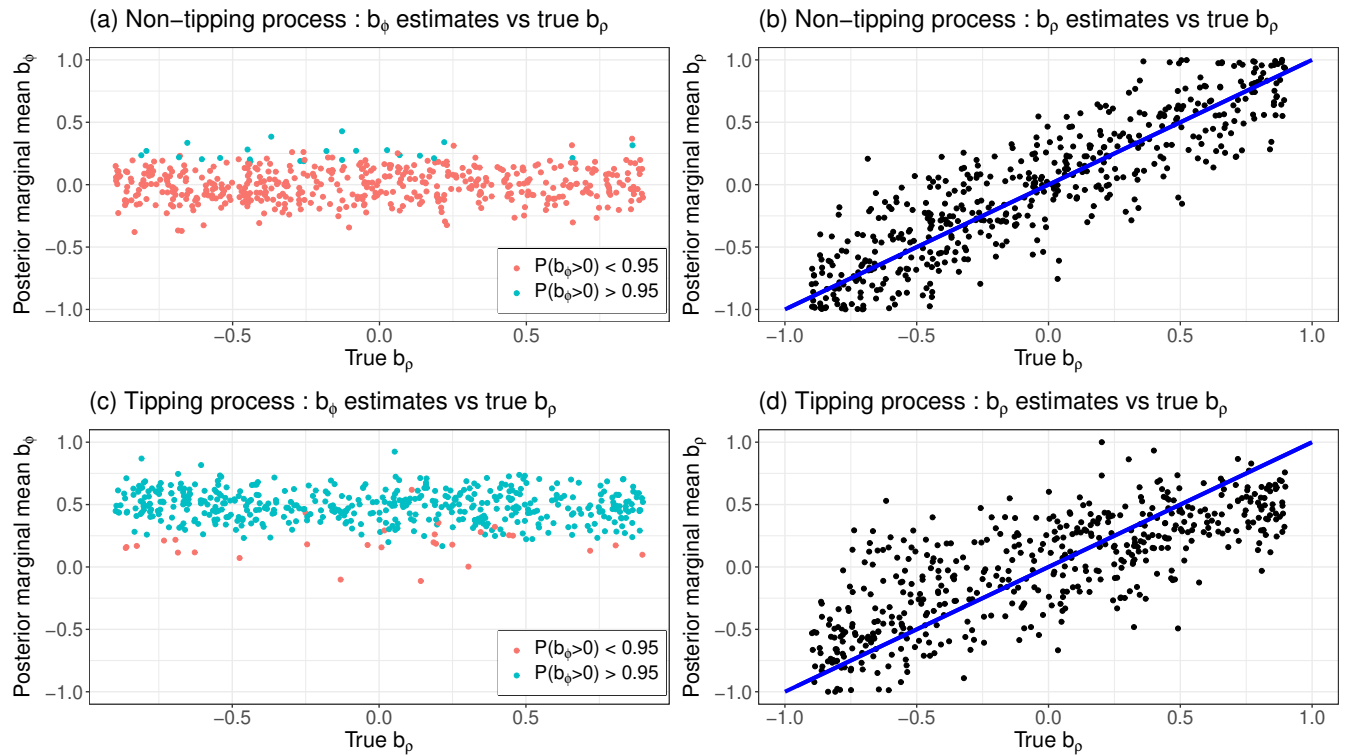


Figure 2. Results of the robustness test for $n_r = 500$ simulated time series of length $n \approx 150$. Panels (a) and (b) show posterior marginal mean estimated by INLA from the non-tipping simulations for b_ϕ and b_ρ respectively plotted against the true value of b_ρ from the correlated noise. The black line in panel (a) and blue line in (b) are showing the true b_ρ used in the simulation. Panel (c) and (d) are similar plots for the tipping simulations. In panels (a) and (c) blue dots are associated with a statistical significance for the EWS indicator b_ϕ to be positive $P(b_\phi > 0 | \mathbf{y}) > 0.95$ while red dots means no statistical significance $P(b_\phi > 0 | \mathbf{y}) < 0.95$

Overall, these two tests demonstrate that the proposed methodology reliably recovers the evolution of autocorrelation parameters, performs well in detecting EWS and is robust to prior assumptions and to structured stochastic external variability not linked to a loss of stability.



Accuracy test				
Estimates	RMSE	Sensitivity(%)	Specificity(%)	
$\hat{b}_\phi$	0.145	87.8	99.8	
$\hat{b}_\rho$	0.278	72.2	99.4	
Robustness test				
Process	$\hat{b}_\rho$ RMSE	Sensitivity(%)	Specificity(%)	$\langle \hat{b}_\phi \rangle$
Tipping	0.34	94.2	-	0.47
Non-tipping	0.26	-	95.4	0

Table 1. Summary statistics from Fig. 1 (top) and Fig. 2 (bottom). Results from Accuracy tests on simulated time-dependent AR(2) processes showing Root Mean Square Error (RMSE) of the estimates of b_ϕ and b_ρ given true values of simulations (blue lines in panels (a)-(b) of Fig. 1). We also show the sensitivity and specificity expressed in percentages for both parameters. (bottom) Results from Robustness tests on simulated tipping and non-tipping processes. We show here the RMSE of the estimates of b_ρ given true simulated values (blue lines in panels (b) and (d)). Sensitivity and specificity are presented in percentages for each process.

Finally, we evaluate the performance of our model by applying it to known real-world critical transitions such as Dansgaard-Oeschger (DO) events. These events are known to be abrupt warmings of the North Atlantic region that occurred during the last glacial period. The presence of EWS before the onset of these events is currently debated; however, several studies report a detection of EWS before some of the first 17 DO events (Myrvoll-Nilsen et al., 2024) (Rypdal, 2016) (Boers, 2018) (Hummel et al., 2024). Therefore, we compare the results of our model with these studies using a setup similar to Myrvoll-Nilsen et al. (2024) by using a second-order polynomial detrending of the data and considering $P(b_\phi > 0 | \mathbf{y}) > 0.95$ as a detection of EWS. This comparison is illustrated in Table 2 and shows that our model suggests, similarly to Myrvoll-Nilsen et al. (2024), that some specific event shows signs of critical slowing down in line with the results of Boers (2018) and Rypdal (2016). Specifically, 2 shows that our results corroborate the 5 EWS found by Myrvoll-Nilsen et al. (2024) while identifying one more EWS for the event 13. Moreover, these results corroborate the EWS found for the event 11 by Boers (2018) and the events 5, 9 by Rypdal (2016), our results also show EWS for the events 2 and 13 similarly to these two studies.



Event	Nested AR(1)	Myrvoll-Nilsen	Rypdal	Boers
1	0.893	0.9146	$p = 0.02$	—
2	0.992	0.9728	$p = 0.008$	$p < 0.05$
3	0.29	0.4893	—	—
4	0.053	0.084	—	$p < 0.05$
5	0.99	0.9959	$p = 0.13$	—
6	0.163	0.2123	—	$p < 0.05$
7	0.444	0.7132	—	—
8	0.817	0.8878	—	—
9	0.994	0.953	$p = 0.16$	—
10	0.115	0.0732	—	—
11	0.977	0.9643	—	$p < 0.05$
12	0.056	0.1662	—	—
13	0.978	0.8912	$p = 0.39$	$p < 0.05$
14	0.722	0.6629	—	$p < 0.05$
15	0.061	0.0637	—	$p < 0.05$
16	0.99	0.9935	—	—
17	0.609	0.6043	—	—

Table 2. Table comparing the posterior probability of positive slope $P(b_\phi > 0 | \mathbf{y})$ from fitting the nested AR(1) model to the different Dansgaard–Oeschger events using a second-order polynomial detrending approach. These results are compared with the probability of positive slope $P(b > 0 | \mathbf{y})$ found by Myrvoll-Nilsen et al. (2024) and p -values obtained from Boers (2018) and Rypdal (2016).



3.2 Detecting early warning signals in AMOC fingerprint

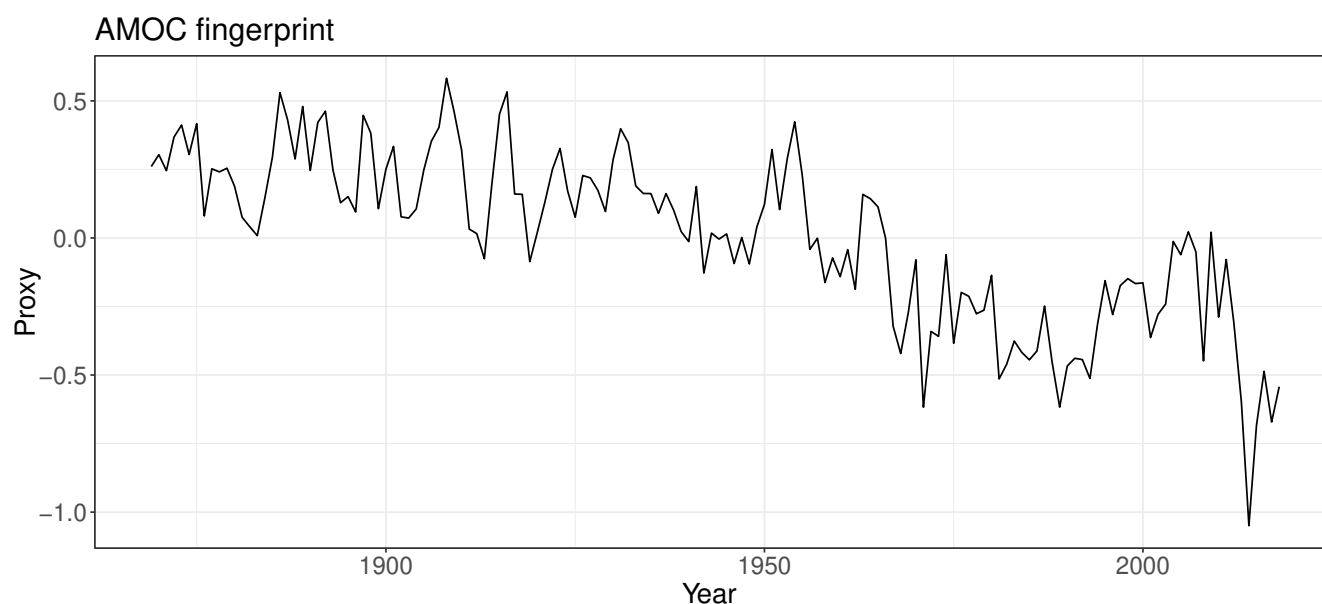


Figure 3. AMOC fingerprint proxy from 1870 to 2020, similar as (Ditlevsen and Ditlevsen, 2023) using yearly averaged subpolar gyre sea-surface temperature anomaly minus twice the global mean anomaly obtained from the Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) (Rayner et al., 2003)

We now apply the time-dependent nested AR(1) model to an AMOC fingerprint similar to the one used by Ditlevsen and Ditlevsen (2023) shown in Fig. 3. This fingerprint is constructed as the sea-surface temperature (SST) anomaly in the subpolar gyre region, averaged annually, minus twice the global mean SST anomaly to compensate for the polar amplification effects under global warming. Several studies have suggested that this proxy is a suitable indicator of AMOC strength (Caesar et al., 2018; Jackson and Wood, 2020; Latif et al., 2019), especially since direct observation of the AMOC is only available from 2004 onward. The use of such a proxy is therefore necessary to examine longer-term trends and detect potential early warning signals.

As the fingerprint exhibits significant drift, it must first be detrended to satisfy the zero-mean assumption of the model. In principle, this trend could be extracted using knowledge of the system's underlying physical processes, but such information may be unavailable, incomplete or inaccurate. To address this, we consider two different detrending strategies. In the first, we rely solely on statistical assumptions and remove the trend using either a linear or second-order polynomial fit. In the second approach, we incorporate physical information by including an explanatory variable in the model, following the structure described in (9). Specifically, we use integrated Central-West Greenland (iCWG) surface melt shown in Fig. 4 as a covariate. The iCWG represents the cumulative surface melt across years, based on the CWG melt stack from Trusel et al. (2018), and is used to capture the influence of freshwater forcing on AMOC stability.

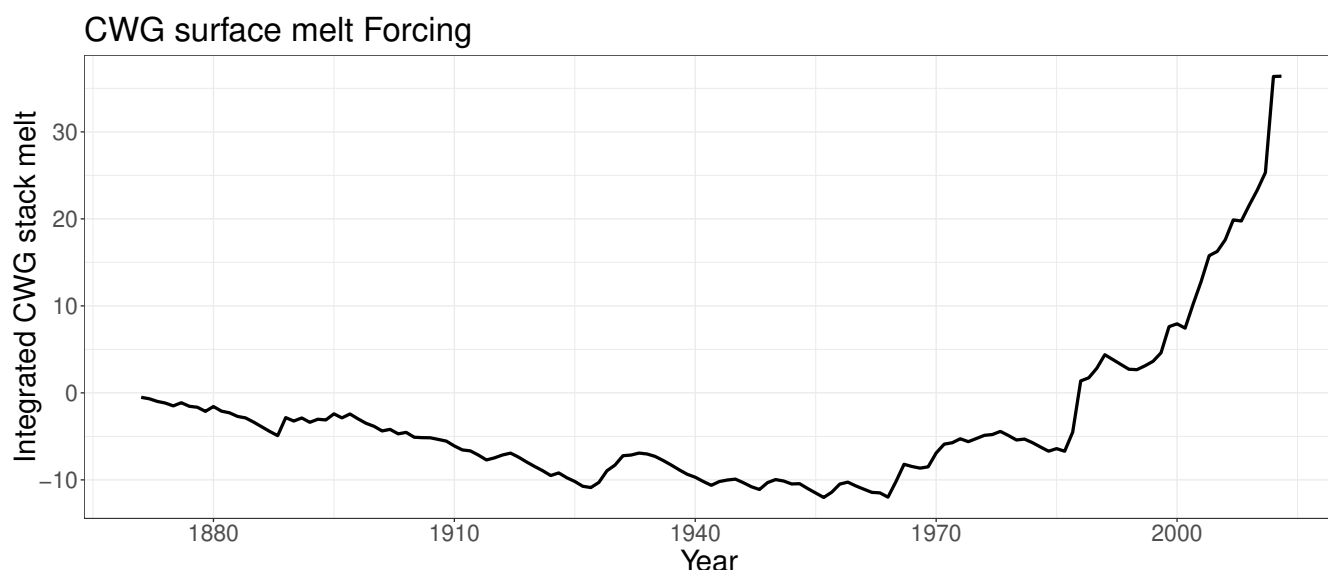


Figure 4. Cumulative Central West Greenland runoff from 1971 to 2013

215 For each model, we compare the posterior marginal mean estimate of the slope parameter b_ϕ along with the posterior probability that the slope is positive. Model fit is assessed using marginal log-likelihood. The full set of results is presented in Table 3. The fitted trends and time evolutions of $\phi(t)$ for the linear and polynomial detrending approaches are shown in Fig. 5, while the estimated response function to the iCWG forcing and associated $\phi(t)$ evolution are shown in Fig. 6. Among the different model configurations, the version incorporating iCWG forcing provides the best fit to the data as measured by model

220 likelihood. In all three detrending strategies, the model identifies statistically significant EWS. These results provide further evidence for the presence of EWS for the AMOC, consistent with the findings of Boers (2021) and Ditlevsen and Ditlevsen (2023), who reported similar signals using different methods and data configurations.

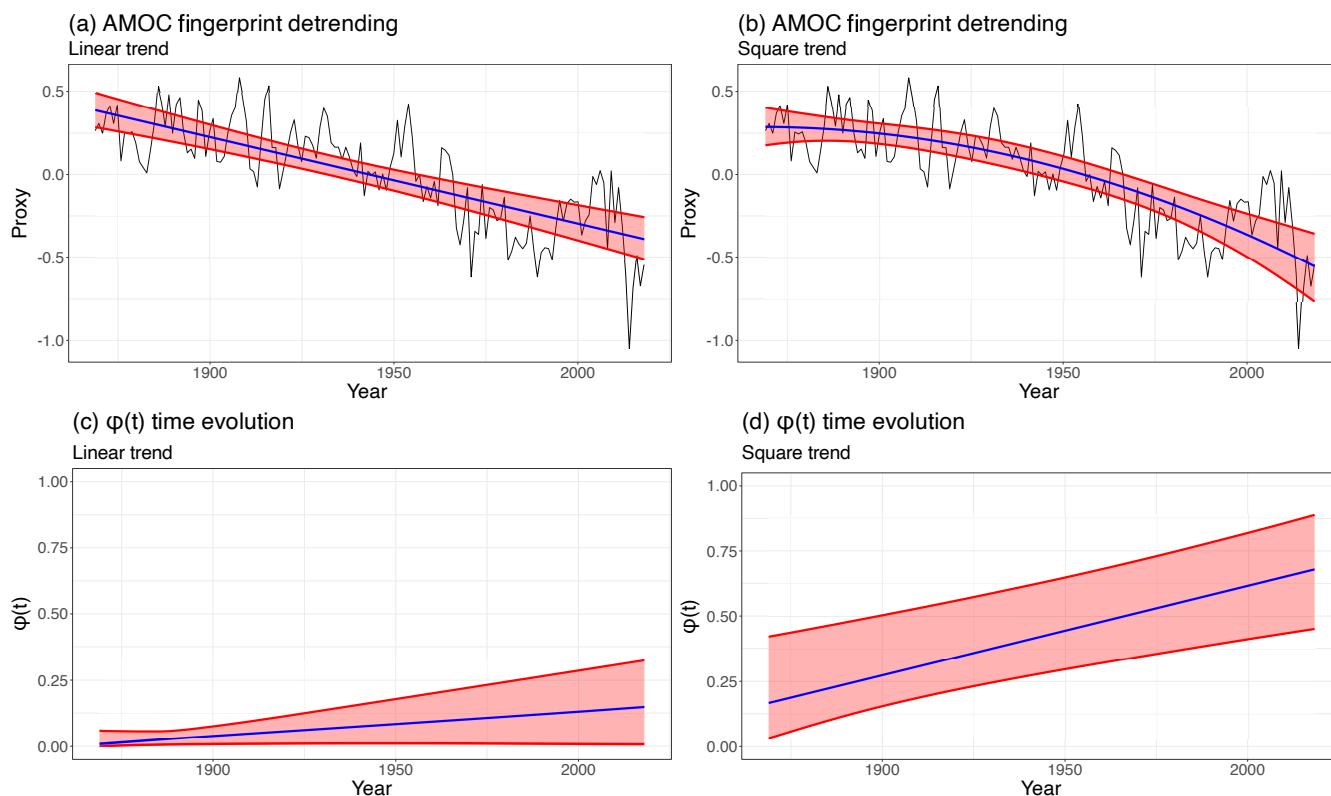


Figure 5. Panel (a) and (b) show AMOC fingerprint (black) with posterior marginal mean (blue) and 95% credible intervals (red) of the fitted trends. Panel (c) and (d) show the evolution in time of the correlation parameter $\phi(t)$ (blue) used as indicator of EWS and 95% credible intervals (red) with an estimated probability of positive slope $P(b_\phi > 0 \mid \mathbf{y})$

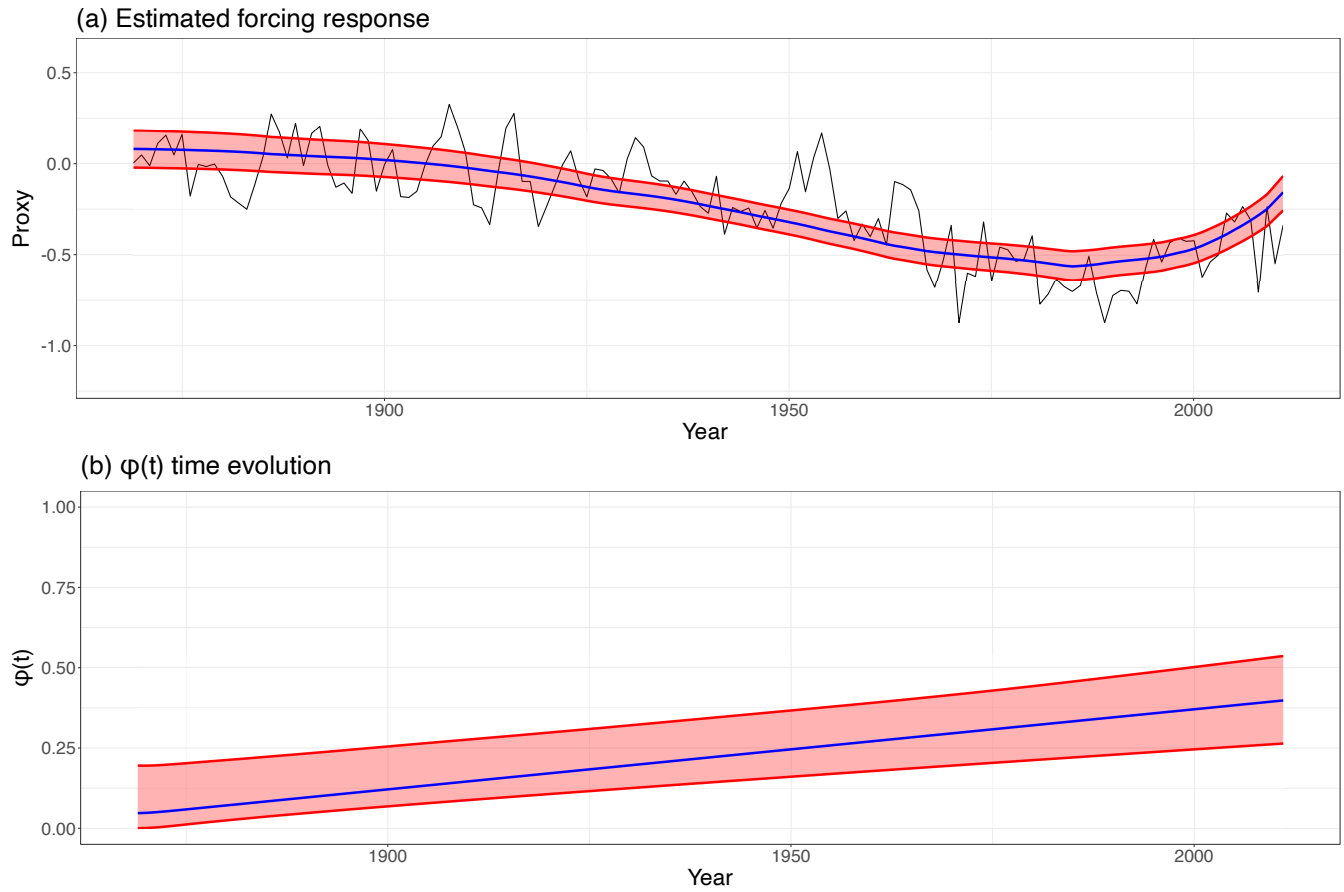


Figure 6. Panel (a) shows AMOC fingerprint (black) from 1870 to 2013 to match the time-span of the forcing data with posterior marginal mean (blue) and 95% credible intervals (red) of the estimated system’s response function to forcing. Panel (b) is a plot of the evolution in time of the correlation parameter $\phi(t)$ (blue) and 95% credible intervals (red) with an estimated probability of positive slope $P(b_\phi > 0 | \mathbf{y})$

Model	$\hat{b}_\phi$	$P(\hat{b}_\phi > 0)$	$\hat{b}_\rho$	Marg. log-likelihood
AR(2) Linear detrending	0.2	0.98	−0.35	56.49
AR(2) Square detrending	0.41	1	−0.33	54.72
AR(2) Forcing response	0.34	1	−0.99	61.97
AR(1) Linear detrending	0.145	0.98	-	53.46
AR(1) Square detrending	0.278	0.99	-	51.68
AR(1) Forcing response	0.19	0.93	-	54.13

Table 3. Summary statistics from Fig. 5 and Fig. 6 showing posterior marginal means of $\hat{b}_\phi$, probability of $\hat{b}_\phi$ positive, posterior marginal means of $\hat{b}_\rho$ and marginal log-likelihood for the three models used here. Results from the models introduced in Myrvoll-Nilsen et al. (2024) are also shown for comparison purposes.



4 Conclusions

This study investigates the stability of the Atlantic Meridional Overturning Circulation (AMOC) by proposing a time-dependent extension of the autoregressive AR(2) model introduced by Morr and Boers (2024) and Boers (2021). The primary objective of this model is to enhance the reliability of early warning signals (EWS) by minimizing false positives. This is achieved through the decomposition of the observed signal into two distinct components: $\rho(t)$, which captures time-dependent external variability, and $\phi(t)$, which reflects changes in the internal dynamics associated with system stability. By isolating these effects, the model aims to identify more accurately early signs of destabilization. Following the approach of Myrvoll-Nilsen et al. (2024), we assume a linear temporal dependence for both $\rho(t)$ and $\phi(t)$, estimating their respective slope parameters within a hierarchical Bayesian framework. This statistical approach allows us to incorporate prior information and quantify the uncertainty of the EWS through the posterior distributions of the parameters. The performance of the model is first evaluated using simulated data, demonstrating both high estimation accuracy and robustness against false detections of ongoing destabilization.

The methodology is applied to a proxy for the AMOC fingerprint. In order to meet stationarity assumptions, we consider various detrending techniques, including linear and second-order polynomial detrending, as well as incorporating a forcing component based on the integrated meltwater runoff from Central-West Greenland. Across all model configurations, we find statistically significant early warning signals. This is consistent with prior findings in the literature and supports the hypothesis of a possible ongoing destabilization of the AMOC.

While assuming a linear structure for $\phi(t)$ has proven effective for detecting EWS, we emphasize that the model proposed here should not be interpreted as a comprehensive or mechanistic representation of the underlying physical processes governing the AMOC. Despite its success in identifying early signs of destabilization, the model is limited in its ability to forecast the future trajectory of the system or predict the timing of a potential tipping point. Addressing these limitations would require a more flexible modeling approach, potentially involving a nonlinear or nonparametric structure for the correlation parameters, which lies beyond the scope of the present work.

Although our analysis has focused on a specific proxy of the AMOC fingerprint, the proposed methodology is generalizable and can be adapted to study the stability of other critical climate components, such as the Greenland Ice Sheet, Arctic sea ice, or the Amazon rainforest. To facilitate wider use and reproducibility, we have extended the existing R package `INLA.ews` to incorporate our methodological advancements. This software provides a user-friendly interface for implementing our approach, leveraging the computational efficiency of the INLA framework for Bayesian inference.

Code and data availability. The code and data sets used for this paper is available through the R-package, `INLA.ews`, which can be downloaded from: github.com/eirikmn/INLA.ews (last access May 26, 2025).

Author contributions. All authors conceived and designed the study. LH adopted the model for a Bayesian framework and wrote the code. EMN and LH carried out the examples and analysis. All authors discussed the results and drew conclusions. All authors wrote the paper.



Competing interests. CF is a member of the editorial board of the NPG journal. LH and EMN declare that they have no conflict of interest.

255 *Acknowledgements.* CF was supported by the Institute for Basic Science (IBS), Republic of Korea, under IBS-R028-D1 and by the National Research Foundation of Korea (NRF-2022M3K3A1097082 and RS-2024-00416848). EMN has also received funding from the Norwegian Research Council (IKTPLUS-IKT og digital innovasjon, project no. 332901). LH has also received funding from the Norwegian Research Council (Stability of the Arctic climate, project no. 314570)



References

- 260 Boers, N.: Early-warning signals for Dansgaard-Oeschger events in a high-resolution ice core record, *Nature communications*, 9, 2556, 2018.
- Boers, N.: Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation, *Nature Climate Change*, 11, 680–688, 2021.
- Boers, N., Ghil, M., and Rousseau, D.-D.: Ocean circulation, ice shelf, and sea ice interactions explain Dansgaard–Oeschger cycles, *Proceedings of the National Academy of Sciences*, 115, E11 005–E11 014, 2018.
- 265 Boettner, C. and Boers, N.: Critical slowing down in dynamical systems driven by nonstationary correlated noise, *Physical Review Research*, 4, 013 230, 2022.
- Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation, *Nature*, 556, 191–196, 2018.
- Ditlevsen, P. and Ditlevsen, S.: Warning of a forthcoming collapse of the Atlantic meridional overturning circulation, *Nature Communica-*
- 270 *tions*, 14, 1–12, 2023.
- Gong, X., Liu, H., Wang, F., and Heuzé, C.: Of Atlantic meridional overturning circulation in the CMIP6 project, *Deep Sea Research Part II: Topical Studies in Oceanography*, 206, 105 193, 2022.
- Hawkins, E., Smith, R. S., Allison, L. C., Gregory, J. M., Woollings, T. J., Pohlmann, H., and De Cuevas, B.: Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport, *Geophysical Research Letters*, 38, 2011.
- 275 Henry, L., McManus, J., Curry, W., Roberts, N., Piotrowski, A., and Keigwin, L.: North Atlantic ocean circulation and abrupt climate change during the last glaciation, *Science*, 353, 470–474, 2016.
- Hummel, C., Boers, N., and Rypdal, M.: Inconclusive Early warning signals for Dansgaard-Oeschger events across Greenland ice cores, *EGUsphere*, 2024, 1–44, 2024.
- Jackson, L. and Wood, R.: Fingerprints for early detection of changes in the AMOC, *Journal of Climate*, 33, 7027–7044, 2020.
- 280 Jackson, L. C., Peterson, K. A., Roberts, C. D., and Wood, R. A.: Recent slowing of Atlantic overturning circulation as a recovery from earlier strengthening, *Nature Geoscience*, 9, 518–522, 2016.
- Latif, M., Park, T., and Park, W.: Decadal Atlantic Meridional Overturning Circulation slowing events in a climate model, *Climate Dynamics*, 53, 1111–1124, 2019.
- Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth’s climate
- 285 *system*, *Proceedings of the national Academy of Sciences*, 105, 1786–1793, 2008.
- Liu, W., Xie, S.-P., Liu, Z., and Zhu, J.: Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate, *Science Advances*, 3, e1601 666, 2017.
- Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al.: Climate change 2021: the physical science basis, Contribution of working group I to the sixth assessment report of the intergovernmental panel on
- 290 *climate change*, 2, 2391, 2021.
- Morr, A. and Boers, N.: Detection of approaching critical transitions in natural systems driven by red noise, *Physical Review X*, 14, 021 037, 2024.
- Myrvoll-Nilsen, E., Sørbye, S. H., Fredriksen, H.-B., Rue, H., and Rypdal, M.: Statistical estimation of global surface temperature response to forcing under the assumption of temporal scaling, *Earth System Dynamics*, 11, 329–345, 2020.



- 295 Myrvoll-Nilsen, E., Hallali, L., and Rypdal, M.: Bayesian analysis of early warning signals using a time-dependent model, *EGUsphere*, 2024, 1–26, 2024.
- Rahmstorf, S.: Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle, *Nature*, 378, 145–149, 1995.
- Rayner, N. A., Parker, D. E., Horton, E., Folland, C. K., Alexander, L. V., Rowell, D., Kent, E. C., and Kaplan, A.: Global analyses of
300 sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Journal of Geophysical Research: Atmospheres*, 108, 2003.
- Robert, C. P., Casella, G., and Casella, G.: *Monte Carlo statistical methods*, vol. 2, Springer, 1999.
- Rue, H., Martino, S., and Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion), *J R Stat Soc Series B*, 71, 319–392, 2009.
- 305 Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K.: Bayesian Computing with INLA: A Review, *Annu. Rev. Stat. Appl.*, 4, 395–421, 2017.
- Rypdal, M.: Early-warning signals for the onsets of Greenland interstadials and the Younger Dryas–Preboreal transition, *Journal of Climate*, 29, 4047–4056, 2016.
- Stommel, H.: Thermohaline convection with two stable regimes of flow, *Tellus*, 13, 224–230, 1961.
- 310 Stouffer, R. J., Yin, J.-j., Gregory, J., Dixon, K., Spelman, M., Hurlin, W., Weaver, A., Eby, M., Flato, G., Hasumi, H., et al.: Investigating the causes of the response of the thermohaline circulation to past and future climate changes, *Journal of climate*, 19, 1365–1387, 2006.
- Trusel, L. D., Das, S. B., Osman, M. B., Evans, M. J., Smith, B. E., Fettweis, X., McConnell, J. R., Noël, B. P., and van den Broeke, M. R.: Nonlinear rise in Greenland runoff in response to post-industrial Arctic warming, *Nature*, 564, 104–108, 2018.
- Weijer, W., Cheng, W., Drijfhout, S. S., Fedorov, A. V., Hu, A., Jackson, L. C., Liu, W., McDonagh, E. L., Mecking, J. V., and Zhang, J.: Stability of the Atlantic Meridional Overturning Circulation: A review and synthesis, *Journal of Geophysical Research: Oceans*, 124, 5336–5375, 2019.
- 315 Wood, R. A., Rodríguez, J. M., Smith, R. S., Jackson, L. C., and Hawkins, E.: Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation, *Climate Dynamics*, 53, 6815–6834, 2019.