https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

A Prototype Probability Based Warning System for a Strong AMOC Decline

Nireas Vasileiadis[1], Oliver Mehling[1], René M. van Westen[1], and Henk A. Dijkstra[1,2]

[1]Institute for Marine and Atmospheric research Utrecht, Department of Physics, Utrecht University, Utrecht, the Netherlands.
[2]Center for Complex Systems Studies, Department of Physics, Utrecht University, Utrecht, the Netherlands.

Correspondence: Henk A. Dijkstra (h.a.dijkstra@uu.nl)

Abstract. With a future strong, possibly accelerated, weakening of the Atlantic Meridional Overturning Circulation (AMOC) becoming a plausible scenario, there is a need to develop a reliable warning system for the onset of such an event. While many statistical and physical early warning signals have been suggested for AMOC weakening, there is no warning system yet that demonstrates skill even in a perfect model framework. Here, we present a prototype of a probability based warning system for a strong AMOC decline and illustrate its performance through two examples. In the first, we use a conceptual stochastic AMOC box model where the update rule is equation based. In the second example, we use data from an Earth System Model, where the update rule is based on a machine learning approach. In both examples, a warning probability that the AMOC decreases below a specific threshold within a certain future time horizon is obtained. Using an Observing System Simulation Experiment (OSSE), we show that the warning system has skill compared to persistence as measured by the Brier skill score. This prototype is hence a useful starting point for the development of warning systems that can eventually be useful for policy makers and long-term planning decisions.

1 Introduction

There is substantial interest at the moment in the projection of the behavior of the Atlantic Meridional Overturning Circulation (AMOC) over the next century (Bellomo et al., 2021; Lenton et al., 2025; Bonan et al., 2025; Baker et al., 2025; Dijkstra and van Westen, 2026). While observations at the OSNAP (Fu et al., 2023) and RAPID-MOCHA (Moat et al., 2025) arrays in the North Atlantic do not show a statistically significant AMOC decline over their operational intervals, a meridionally consistent decline in the AMOC's western boundary contribution over the past two decades has been observed (Xing et al., 2026). Proxy-based (temperature) reconstructions of the AMOC indicate that its strength might have already decreased by 15% since 1950 (Caesar et al., 2018; Li and Liu, 2025; Michel et al., 2025).

Projections based on models participating in the Coupled Model Intercomparison Project (CMIP) phase 6, hereafter referred to as CMIP6 models, indicate that the AMOC will decline over the next century under increased greenhouse gas forcing (Weijer et al., 2020). While much of the decline is caused by the changes in forcing (Drijfhout et al., 2025), the AMOC may also undergo a strong, accelerated decline over the next century due to positive feedbacks in a so-called tipping event (Lenton et al., 2025). Although both types of strong AMOC decline are often referred to as an AMOC collapse (Dijkstra et al., 2026),

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

the plausibility of a tipping event has been increased by the analysis of CMIP6 model simulations where some show an AMOC collapse under extended climate forcings, even at low emission scenarios (Drijfhout et al., 2025).

The underlying 'mental image' for the AMOC tipping phenomenon is that of a so-called back-to-back saddle-node bifurcation diagram (Stommel, 1961), where an equilibrium AMOC-ON state and an AMOC-OFF state can exist under the same forcing conditions. Forcing variations are viewed as a slow change in a control parameter which changes the equilibrium state of the climate model (Lenton, 2011). The existence of multiple equilibrium AMOC states has been demonstrated in the Community Earth System Model (CESM) and has been connected to the export of freshwater by the AMOC at 34.5°S (van Westen et al., 2024). Observations indicate that the AMOC exports freshwater at this latitude (Arumí-Planas et al., 2024). This provides support for the notion that the present-day AMOC state is in a multistable regime (Dijkstra et al., 2026), where the salt-advection feedback is destabilizing the AMOC (Weijer et al., 2019; Vanderborght et al., 2025).

Because the climate impacts of a strong AMOC weakening are potentially severe, in particular over the northern European countries (van Westen and Baatsen, 2025), there is a clear need for a reliable warning system for AMOC tipping behavior. In many studies early warning indicators have been based on the analysis of single time series, in particular searching for critical slowing down. Since the early work of Held and Kleinen (2004), an increase in both variance and autocorrelation has been identified in several AMOC fingerprints, consistent with a destabilization of the AMOC (Boers, 2021; Ditlevsen and Ditlevsen, 2023) under quasi-equilibrium conditions, but having large uncertainties (Ben-Yami et al., 2024). Several other early warning indicators have been suggested, from properties of the underlying dynamical system (Ashwin et al., 2012) to complex network constructed ones (Feng et al., 2014), to fully data based ones (Bury et al., 2021).

In the CESM, the collapse of the AMOC is dependent on the background AMOC state and also forcing rate dependent (Van Westen et al., 2026). As a physics-based indicator for the onset of the strong AMOC weakening under climate change forcing a sign change (from negative to positive) of the surface buoyancy flux over the 40°N-65°N Atlantic region has been suggested (van Westen et al., 2025). Application to CMIP6 models indicates that, based on this indicator, the onset of the collapse will occur around the year 2060 ($\pm \sim 30$ years). While this is around the same time as projected by critical slowing down using AMOC fingerprints (Ditlevsen and Ditlevsen, 2023), the result is expected to be model dependent and hence also has a large uncertainty.

Recently, an Observing System Simulation Experiment (OSSE) framework was developed for assessing how observations can improve model-based warnings of an AMOC collapse (Kubo and Sawada, 2025). They show that accurately resolving AMOC internal variability is more important for tipping prediction than observation frequency and establish when observations are sufficiently informative to improve model-based prediction of AMOC tipping. However, they do not translate this information into a warning signal or evaluate the skill of such warnings. The present study addresses this gap by developing and testing an end-to-end framework that assimilates observations, estimates the probability of an AMOC collapse within a specified time horizon, and converts this probability into actionable warning levels.

In section 2, we describe the components of such a prototype probability based warning system, with details provided on the data-assimilation method used and the computation of probabilities of strong AMOC weakening, either forcing or feedback induced, over a certain time horizon. The results in section 3 show applications to (i) a conceptual box model and (ii) to data

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

from the CESM, where the model update rules are equation based and data-based, respectively. A summary and discussion follows in section 4.

2 Methods

2.1 Components of a warning system

A typical warning system has at least three components (Fig. 1). One component deals with sensor placement (the yellow box in Fig. 1), i.e., where the observations are located in space and what temporal sampling frequencies are used. The second component converts these observations into a warning signal (green boxes in Fig. 1), which can be physics-based or statistical (or a combination of both). The third component (purple boxes in Fig. 1) consists of a test environment where the skill of the warning signal can be evaluated in a perfect model setup. The resulting warning signal can be used to issue alarms (orange-red boxes in Fig. 1).

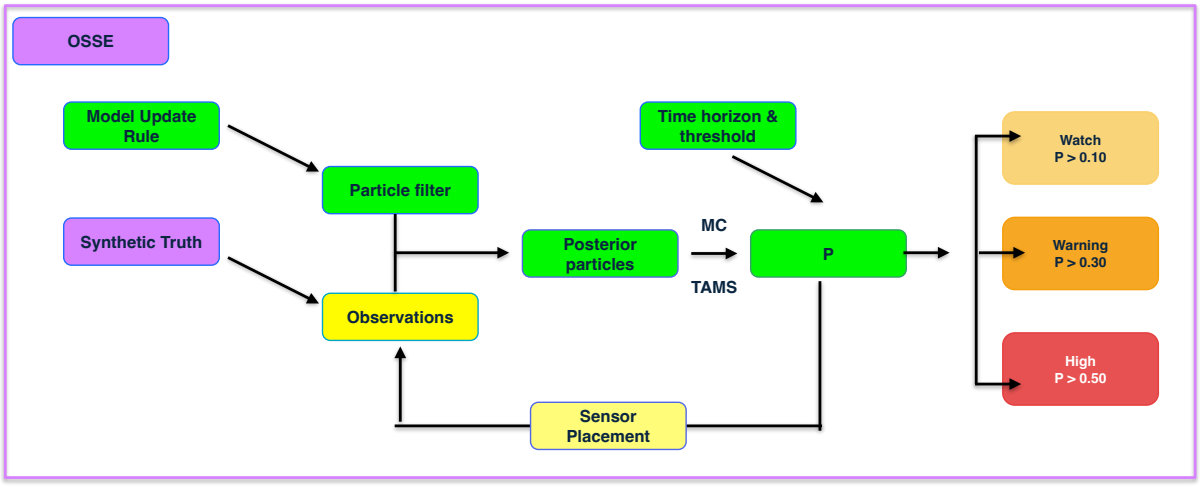


Figure 1. *Sketch of a prototype probability based warning system, consisting of the sensor configuration (yellow), the warning signal P (green) and the OSSE test setup (purple), together with possible alarms (orange-red). Here MC and TAMS refer to Monte Carlo and Trajectory Adaptive Multilevel Splitting, respectively.*

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

In this paper, we will not consider the sensor placement problem, as observation locations relevant for the AMOC are not expected to change much in the near future. Hence, with the given observational configuration, our goal is to build a probability based warning system for a strong AMOC decline. Let the AMOC strength at 26°N in a model be indicated by Ψ, then we use a threshold value Ψ_c to define the collapsed set $\mathcal{C}$ by

$$\mathcal{C} = \{\mathbf{x} : \Psi \leq \Psi_c\}. \tag{1}$$

where $\mathbf{x}$ is the state vector of the model and t indicates time. Furthermore, the collapse time t_c, given the time t_k up to which observations are available, is

$$t_c = \inf\{t > t_k : \mathbf{x}(t) \in \mathcal{C}\}. \tag{2}$$

We now define the warning probability as the conditional probability P_τ (given the available observations) that the AMOC strength decreases below the chosen threshold within a time horizon τ beyond t_k. Hence,

$$P_\tau(t_k) = p(t_c \leq t_k + \tau \mid \mathbf{y}_{1:k}), \tag{3}$$

where $\mathbf{y}_{1:k}$, with $\mathbf{y} = H(\mathbf{x})$ for a specific observation operator H, denotes the observation vector up to the time t_k. For a chosen warning level $p_\star$, the warning time t_w based on the probability $P_\tau(t_k)$ is given by

$$t_w = \inf\{t_k : P_\tau(t_k) > p_\star\}. \tag{4}$$

If t_c^{true} is the time the collapse occurs (so the time where $\Psi \leq \Psi_c$), the lead time t_l is then given by $t_l = t_c^{\text{true}} - t_w$.

The choice of warning level $p_\star$ affects the frequency of false alarms and missed events. At a given time t_k, a warning is a false positive (false alarm) if the AMOC strength does not reach Ψ_c within the time horizon τ. Conversely, if the AMOC strength reaches Ψ_c within this horizon and no warning is issued, the outcome is a false negative (missed event). The sensitivity is the fraction of cases with a decline within τ for which a warning is issued. The specificity is the fraction of cases without such a decline for which no warning is issued. Lowering $p_\star$ can increase sensitivity at the expense of specificity, whereas raising it can increase specificity at the expense of sensitivity. The choice of $p_\star$ therefore depends on the relative consequences of false alarms and missed events.

2.2 Warning probability computation

Clearly the main challenge in this warning system is the computation of the warning probability (3). Here, we will use a particle filtering technique (Doucet et al., 2001; van Leeuwen, 2020), which can be applied when only a model update rule is available (e.g., no adjoints are needed) and that does not assume Gaussian statistics for the state variables. This update rule on the state vector $\mathbf{x}$ of the model is written here in discrete time as

$$\boldsymbol{x}_{k+1} = M_k(\boldsymbol{x}_k) \tag{5}$$

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

where M_k also includes the forcing of the model. Given an initial state $\mathbf{x_0}$, the evolution of the state vector, $\mathbf{x}_k$, is described by (5). In 'training mode', observations $\mathbf{y}_k$ become available at discrete times. The aim of a particle filtering technique is to obtain a conditional probability density function $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ of the current state $\mathbf{x}_k$ at $t = t_k$ given all available observations.

The Monte Carlo approximation to the filtering density $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ uses a weighted ensemble of N model realizations, called particles, with states $\mathbf{x}_k^{(i)}$, $i = 1, \ldots, N$. The discrete distribution is

$$p_N(\mathbf{x}_k|\mathbf{y}_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}), \tag{6}$$

Here, $w_k^{(i)}$ denotes the weight of particle i and δ is the Dirac delta function. We draw the initial particle states independently from $p(\mathbf{x}_0)$ and assign equal weights ($w_0^{(i)} = \frac{1}{N}$).

The particle filter computes the distribution recursively through prediction and update steps. Given the states $\mathbf{x}_{k-1}^{(i)}$ and weights $w_{k-1}^{(i)}$ representing $p_N(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})$, we integrate each particle forward to obtain $\mathbf{x}_k^{(i)}$. The weights carry forward through the prediction step, giving the prior

$$p_N(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \sum_{i=1}^{N} w_{k-1}^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}). \tag{7}$$

At $t = t_k$, the observation $\mathbf{y}_k$ becomes available. We then apply Bayes' theorem to update the prior:

$$p_N(\mathbf{x}_k|\mathbf{y}_{1:k}) = \frac{p(\mathbf{y}_k|\mathbf{x}_k) p_N(\mathbf{x}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k)}. \tag{8}$$

Substituting Eqs. (6) and (7) into Eq. (8) gives the updated weights:

$$w_k^{(i)} = \frac{p(\mathbf{y}_k|\mathbf{x}_k^{(i)})}{p(\mathbf{y}_k)} w_{k-1}^{(i)}. \tag{9}$$

The denominator $p(\mathbf{y}_k)$ serves as a common normalization factor. In practice, we normalize the updated weights so that $\sum_{i=1}^{N} w_k^{(i)} = 1$.

The assumed observation error distribution determines the likelihood $p(\mathbf{y}_k|\mathbf{x}_k^{(i)})$. For a scalar measurement y_k with Gaussian measurement error of standard deviation σ_{filter}, the likelihood has the form

$$p(y_k|\mathbf{x}_k^{(i)}) \sim \exp\left(-\frac{(H(\mathbf{x}_k^{(i)}) - y_k)^2}{2\sigma_{\text{filter}}^2}\right). \tag{10}$$

The observation operator H maps the particle state to the model equivalent of the measurement, $H(\mathbf{x}_k^{(i)})$.

In 'prediction mode', no new observation $\mathbf{y}_k$ is assimilated, so the algorithm proceeds with fixing the weights w_k^i. By conditioning the warning probability $P_\tau(t_k)$ given by (3) on the current state $\mathbf{x}$, we find

$$P_\tau(t_k) = \int p(t_c \le t_k + \tau \mid \boldsymbol{x}_k = \boldsymbol{x}) p(\boldsymbol{x}_k = \boldsymbol{x} \mid \boldsymbol{y}_{1:k})\, \mathrm{d}\boldsymbol{x} = \int p_\tau(\boldsymbol{x}) p(\boldsymbol{x}_k = \boldsymbol{x} \mid \boldsymbol{y}_{1:k})\, \mathrm{d}\boldsymbol{x}, \tag{11}$$

where $p_\tau(\boldsymbol{x})$ is the τ-year transition probability (i.e., the probability that $\Psi \le \Psi_c$ within a time τ) starting from $\boldsymbol{x}$. The integral in (11) is approximated using the ensemble used for the particle filter. When a sufficiently large ensemble can be propagated

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

in time from the state $\boldsymbol{x}$, or the transition probability $p_\tau(\boldsymbol{x})$ is relatively large, a Monte Carlo approach will be feasible (Singh et al., 2018). However, when the ensemble members are limited or the transition probability is very low, it is better to use the Trajectory Adaptive Multilevel Splitting (TAMS) method (Jacques-Dumas et al., 2024), which is explained for convenience in Appendix A.

2.3 Performance

To test the warning system, Observing System Simulation Experiments (OSSEs) are used. These are numerical model experiments in which a high-fidelity 'truth' provides synthetic observations that are assimilated into analysis/forecast systems to test candidate observing networks. They are widely used in climate science (Prive and Errico, 2014; Hoffman and Grassotti, 2016; Li and Born, 2019), mostly addressing the sensor placement problem, and increasingly applied to climate tipping phenomena where early detection depends critically on observing system design.

The OSSE truth $\boldsymbol{x}_k^{\text{true}}$ is generated from (5) and the binary truth label $E_k \in \{0,1\}$ at time t_k is given by

$$E_k = \mathbf{1}\{t_c \le t_k + \tau\}. \tag{12}$$

where $\mathbf{1}$ is the indicator function. In an OSSE, an ensemble simulation is available and each of these simulations can be used as the 'truth'. Hence, for every OSSE realization (or seed) r, $r = 1, \ldots, N_r$ (possibly with different truth trajectories), there are truth labels E_k^r. Since the forecast has a probabilistic outcome, its skill can be assessed with the Brier score (Brier, 1950). Pooling over N_r OSSE realizations and all valid snapshots t_k, the Brier score B_S is the mean-squared error of the probabilistic forecast against the event labels,

$$\mathrm{B_S} \;=\; \frac{1}{N_{eval}} \sum_{r,k} \left(P_k^r - E_k^r\right)^2, \tag{13}$$

where N_{eval} is the total number of pooled (P, E) pairs. The sum above for r is from 1 to N_r and k is summed over the set of valid evaluation times t_k for OSSE realization r. Skill is usually expressed relative to a reference $\bar{B}_S$, and gives the Brier Skill Score B_{SS}

$$\mathrm{B_{SS}} \;=\; 1 - \frac{\mathrm{B_S}}{\bar{B}_S}, \tag{14}$$

so that $\mathrm{B_{SS}} = 1$ is a perfect forecast and $\mathrm{B_{SS}} = 0$ is no better than the reference. Below, we will use different references, e.g., a constant base-rate forecast $\bar{P}_k^r = \bar{E}$, a constant forecast $\bar{P}_k^r = 1/2$, and a persistence forecast $\bar{P}_k = E_{k-1}$ (initialized at $\bar{E}$). A positive $\mathrm{B_{SS}}$ against all three references confirms that the forecast carries information beyond these naive predictors.

To separate calibration from informativeness, the pooled (P, E) pairs are sorted into n_b equal-width probability bins indexed by j, each containing n_j forecasts with mean forecast probability $\bar{p}_j$ and observed event frequency $\bar{f}_j$. The Brier score can then be expressed using the decomposition (Murphy, 1973)

$$\mathrm{B_S} \;\sim\; \underbrace{\frac{1}{N_{eval}} \sum_j n_j \left(\bar{p}_j - \bar{f}_j\right)^2}_{\mathrm{B_{REL}}} \;-\; \underbrace{\frac{1}{N_{eval}} \sum_j n_j \left(\bar{f}_j - \bar{E}\right)^2}_{\mathrm{B_{RES}}} \;+\; \underbrace{\bar{E}\left(1 - \bar{E}\right)}_{\mathrm{B_{UNC}}}. \tag{15}$$

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

The reliability term $\mathrm{B_{REL}}$ penalizes miscalibration, that is the gap between forecast probabilities and the frequencies that actually materialize, and is small for a well-calibrated forecast. The resolution term $\mathrm{B_{RES}}$ rewards the forecast's ability to discriminate events from non-events by issuing probabilities that depart from the base rate, and is large for an informative forecast. The uncertainty term $\mathrm{B_{UNC}}$ is a measure of the climatological uncertainty and depends only on the base rate, not on the forecast. A near-zero $\mathrm{B_{REL}}$ together with a $\mathrm{B_{RES}}$ approaching $\mathrm{B_{UNC}}$ therefore characterizes a forecast that is simultaneously well calibrated and sharp.

3 Results

To illustrate the warning system and its skill, we will use two models at both ends of the AMOC model hierarchy. In subsection 3.1, the stochastic Stommel-Cessi model is used, with an equation-based update rule. Then in subsection 3.2, we use data from the Community Earth System Model (CESM) with a data-based (i.e., using machine learning techniques) update rule (Bracco et al., 2025).

3.1 Stochastic Stommel-Cessi model

The Stommel-Cessi model (Stommel, 1961; Cessi, 1994) is the cornerstone model for AMOC tipping and here we will use the non-dimensional version of a stochastic version, given by

$$dx_t = (-\alpha(x_t - 1) - x_t\left[1 + \mu^2(x_t - y_t)^2\right])dt + \sigma_x dW_t^x, \tag{16}$$

$$dy_t = (F - y_t\left[1 + \mu^2(x_t - y_t)^2\right])dt + \sigma_y dW_t^y, \tag{17}$$

where x is the non-dimensional equatorial-pole temperature contrast, y is the non-dimensional salinity contrast, α is the ratio of a diffusive to a thermal relaxation time scale, μ^2 measures the relative strength of advective exchange, and F is the non-dimensional freshwater forcing. As standard parameters, we choose $\alpha = 100$, $\mu^2 = 4.0$ and noise parameters $\sigma_x = \sigma_y = 2.0 \times 10^{-2}$. As forcing, we choose $F(t) = F_0 + F_H t$, with $F_0 = 0.9$ and $F_H = 2.0 \times 10^{-3}$. Using for example the Euler-Maruyama scheme, the model update rule can be written in the form (5). The noise represents in a schematic way here the internal variability as well as the forced variability of the AMOC.

The exchange factor $1 + \mu^2(x - y)^2$ is always positive, so it is not a signed overturning index. For a warning problem, a convenient signed AMOC-like index is therefore $z = x - y$. A transition from a strong AMOC to a weak AMOC is defined by crossing a threshold in z, usually chosen near the unstable equilibrium separating the two attracting basins. Let z_c be the chosen critical value, then the collapsed set is $\mathcal{C} = \{\mathbf{x} = (x, y)^T : z \leq z_c\}$. For the deterministic case with $F_H = 0$ (hence constant $F = F_0 = 0.9$), the saddle-node bifurcation points of the model are located at $F_1 = 0.964$ and $F_2 = 0.914$. Hence under the transient forcing, the saddle-node F_1 will be reached at approximately $t = 0.064/0.002 = 32$. Hence, we use a dimensionless time interval $t \in [0, 70]$ as the OSSE length and take observations every integer time step.

In Fig. 2a, the state variables x and y of the truth trajectory are plotted under the freshwater forcing $F(t)$ (shown in Fig. 2b), indicating that a transition occurs in the interval $t \in [30, 50]$. The AMOC variable $z = x - y$ of this synthetic truth is shown

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

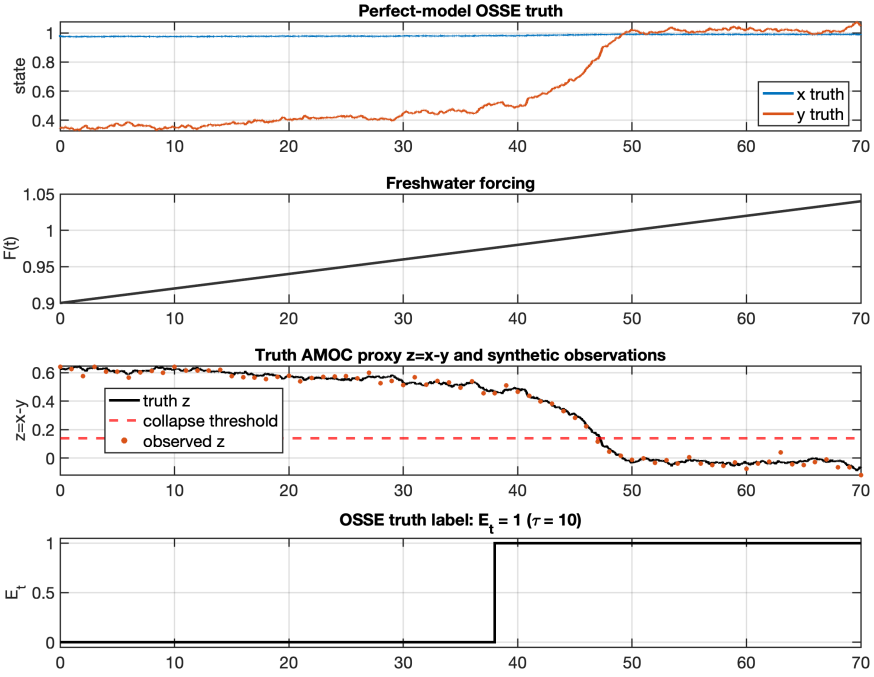


Figure 2. *(a) Trajectory of one realization (the 'truth' trajectory) of the Stommel-Cessi model, where both x_t and y_t are plotted versus time. (b) Freshwater flux forcing versus time. (c) Truth trajectory of the AMOC strength $z = x - y$; the collapse threshold z_c is indicated as the red dashed line. Observations generated from the truth are plotted as red dots. (d) Truth label E_t of the truth trajectory for $\tau = 10$.*

in Fig. 2c, and observations are generated with Gaussian zero mean noise having standard deviation $\sigma_{\text{filter}} = 0.025$. For the weakening of the AMOC, we choose a critical value of $z_c = 0.14$ (shown as the dashed curve in Fig. 2c) which represents a 75% reduction in the AMOC strength. The truth label E_k at every time t_k for a chosen time horizon $\tau = 10$ is shown in Fig. 2d.

Next, we test the particle filter using an ensemble of $N = 500$ members. Initial conditions are generated near the equilibrium state of $F_0 = 0.9$ with a standard deviation $\sigma = 0.03$ in both x and y. Figure 3a shows the particle filter posterior (mean and 95% band) z^a given by

$$z_k^a = \sum_{i=1}^{N} w_k^{(i)} z(\boldsymbol{x}_k^{(i)}), \tag{18}$$

over the simulation period, while assimilating observations every integer time. The absolute value of the posterior error (mean - truth) on z is overall smaller than 0.04 (Fig. 3b). The weight degeneracy at each step k is monitored by the effective number of particles

$$N_{\text{eff}}(t_k) = \Big(\sum_{i=1}^{N} \Big(w_k^{(i)}\Big)^2\Big)^{-1}. \tag{19}$$

When the fraction of effective number of particles N_{eff}/N drops below 0.5 (Fig. 3c), then the weights are resampled. Particles are drawn with replacement according to weights $w_k^{(i)}$, and the new weights are set to $w_k^{(i)} = \frac{1}{N}$. The posterior uncertainty,

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

given by the square root of the variance

$$\mathrm{Var}(z_k \mid \boldsymbol{y}_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \left[z(\boldsymbol{x}_k^{(i)}) - z_k^a \right]^2, \tag{20}$$

is shown in Fig. 3d . It is largest at the beginning and near the transitions but is overall smaller than 0.02.

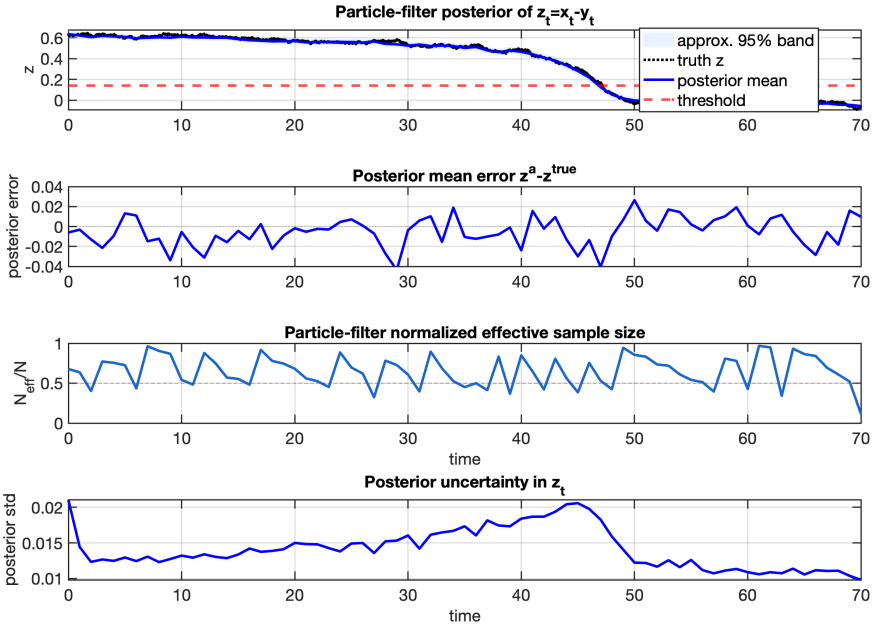


Figure 3. *(a) Particle filter posterior distribution of the AMOC strength z, where the black dotted curve is the 'truth' trajectory. The horizontal red dashed line is the value of $z_c = 0.14$. (b) Posteriori mean error ($z^a - z^{true}$). (c) Relative effective sampling size N_{eff}/N and threshold for resampling (red line) and (d) the posterior uncertainty in z, determined by (20).*

For the computation of the warning probability $P_\tau(t_k)$ for a particular time t_k over the time horizon $\tau = 10$ i.e., the probability that z becomes smaller than z_c within this time, we use TAMS (see appendix A with a number of trajectories $N_T = 100$) conditioned on the posterior particle distribution. This method is a bit overkill here compared to Monte Carlo, but it is used here for illustration that it can be applied (when really needed). Histograms of these posterior distributions (upper panels) and the estimates of the warning probability (lower panels) are shown in Fig. 4 for four different times. As expected, at small times we find $P_\tau = 0$ and at large times we find $P_\tau = 1$. At intermediate times, where the transition happens, there is substantial variance in the TAMS estimate of P_τ.

The posterior weighted TAMS probability P_{10} is shown in Fig. 5a and also in Fig. 5b (in color) overlaid with the truth trajectory and the observations. Clearly, P_{10} increases from $t = 30$ and can provide a warning sign for the upcoming transition. From this figure, one can determine the warning time t_w, and the lead time t_l, when a tolerance $p^\star$ has been chosen. For example, for $p^\star = 0.5$ a lead time $t_l \sim 10$ is found. This is indeed compatible with the time where the AMOC strength begins to weaken more rapidly (around $t = 37$).

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

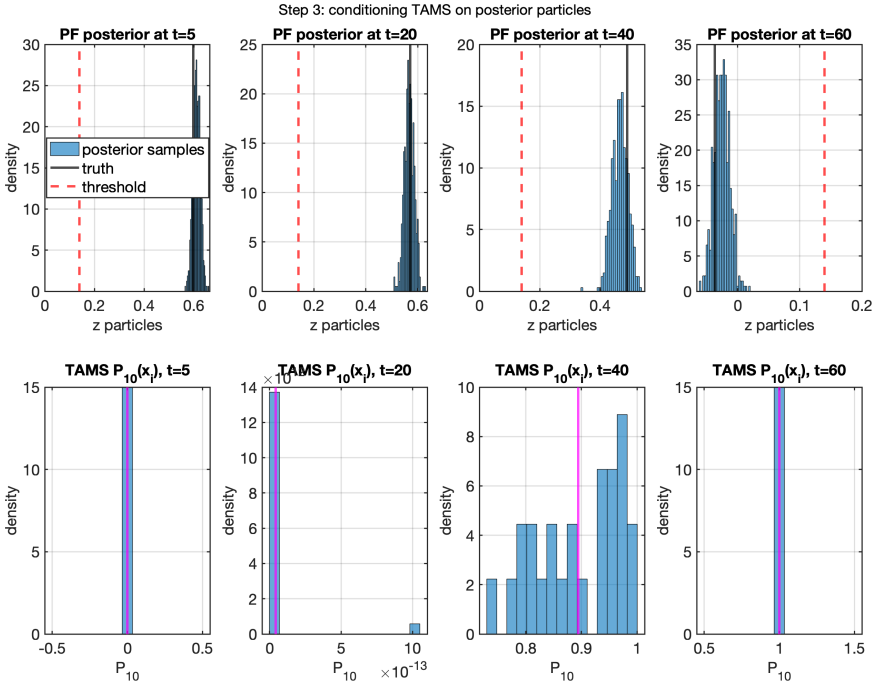


Figure 4. *Upper row: particle filter posterior distributions at different times $t_k = 5, 20, 40, 50$. Lower row: Transition probability $P_\tau(t_k)$ using TAMS at the same times for $\tau = 10$. The purple curve in the lower panels indicates the mean.*

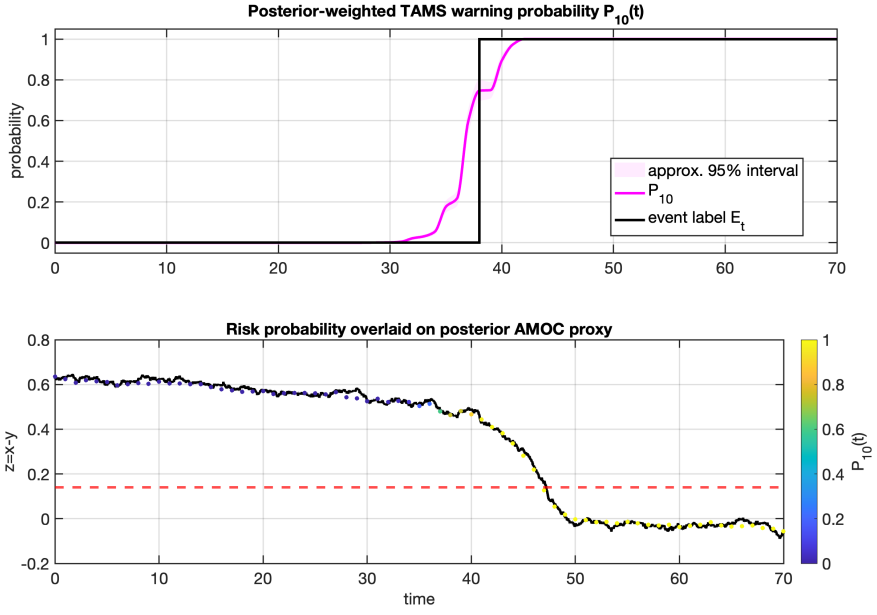


Figure 5. *(a) Posterior-weighted TAMS transition probability with $\tau = 10$, so $P_{10}(t)$, versus time, with the event label of the 'truth' plotted as the drawn black curve. (b) Same $P_{10}(t)$ but now plotted along the 'truth' trajectory.*

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

To quantify the skill of the warning system we repeat the full OSSE five times ($N_r = 5$) by generating independent truth trajectories, and synthetic observations (Fig. 6e). The forecast obtains a Brier score $B_S = 0.0140$ (Fig. 6a). Its Murphy decomposition $B_S \sim B_{\mathrm{REL}} - B_{\mathrm{RES}} + B_{\mathrm{UNC}}$ pairs a small reliability term ($B_{\mathrm{REL}} = 0.0022$) with a resolution ($B_{\mathrm{RES}} = 0.2231$) that recovers most of the climatological uncertainty ($B_{\mathrm{UNC}} = 0.2349$). Against trivial baselines, it obtains a Brier skill score $B_{SS} \simeq +0.94$ relative to both the climatology and constant-0.5 references, and $B_{SS} \simeq +0.25$ relative to persistence (Fig. 6b). Hence, for this set of realizations, the probabilistic warning forecast carries skill beyond all three reference forecasts. The reliability diagram (Fig. 6c) is consistent with the small value of B_{REL}, although the intermediate-probability bins contain relatively few forecasts because most probabilities are concentrated near zero or one. The forecast is well calibrated and sharp (Fig. 6d), assigning confident probabilities to 88% of snapshots. The lead-time audit indicates a median warning lead of approximately 10.2 dimensionless time units at the threshold $p^\star = 0.3$, decreasing to approximately 9.7, 8.4, and 6.4 for $p^\star = 0.5$, 0.7, and 0.9, respectively (Fig. 6f). Thus, the forecast retains substantial probabilistic skill, while the most stringent warning levels are reached closer to the actual transition.

This example was only to illustrate the framework of the AMOC warning system. A systematic study of the warning system parameters was undertaken, but because the model is so simple, we do not show these results and instead turn to a much more challenging application in the next subsection.

3.2 CESM

Our second example is the Community Earth System Model (CESM), in particular version CESM1.0.5 in which an AMOC tipping event has been found under quasi-equilibrium freshwater flux forcing (van Westen et al., 2024). In this case, the methods to determine the warning probability cannot be applied to the model directly because the state vector is too high dimensional. In this case, we construct a model update rule using a machine learning approach.

3.2.1 Data

We use output of the CESM AMOC hysteresis simulation presented in van Westen and Dijkstra (2023). The freshwater flux forcing F_H ramps linearly from 0 to 0.66 Sv with a rate $+3 \times 10^{-4}$ Sv/yr over years 0–2200, then reduces back to zero at the same rate over years 2200–4400. The onset of the AMOC collapse occurs at about model year 1750 or $F_H \approx 0.525$ Sv. The reduced CESM state vector used here consists of ten physically interpretable AMOC-related diagnostics,

$$\mathbf{x}_k = [\psi_{26N,1000m},\ \psi_{\max},\ \psi_{\min,deep},\ F_{ovS},\ F_{ovN},\ T_N,\ S_N,\ \Delta T,\ \Delta S,\ \Delta\rho]_k\,. \tag{21}$$

Here, $\psi_{26N,1000m}$ is the overturning strength at 25.8°N and 1041 m depth (the RAPID-MOCHA array equivalent), $\psi_{\max}$ is the maximum overturning streamfunction over the Atlantic basin from 30°S to 70°N, restricted to depths below 500 m to avoid contamination by the wind-driven shallow subtropical cell, and $\psi_{\min,deep}$ is the minimum of the AMOC strength below 2000 m, which measures the bottom-cell strength. The diagnostics F_{ovS} and F_{ovN} are the freshwater transport indicators by the overturning at the southern (34°S) and northern (60°N) Atlantic boundaries. The thermohaline group $\{T_N, S_N, \Delta T, \Delta S, \Delta\rho\}$ is computed from the Atlantic zonally averaged temperature, salinity and potential density, averaged over the upper 1000 m

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

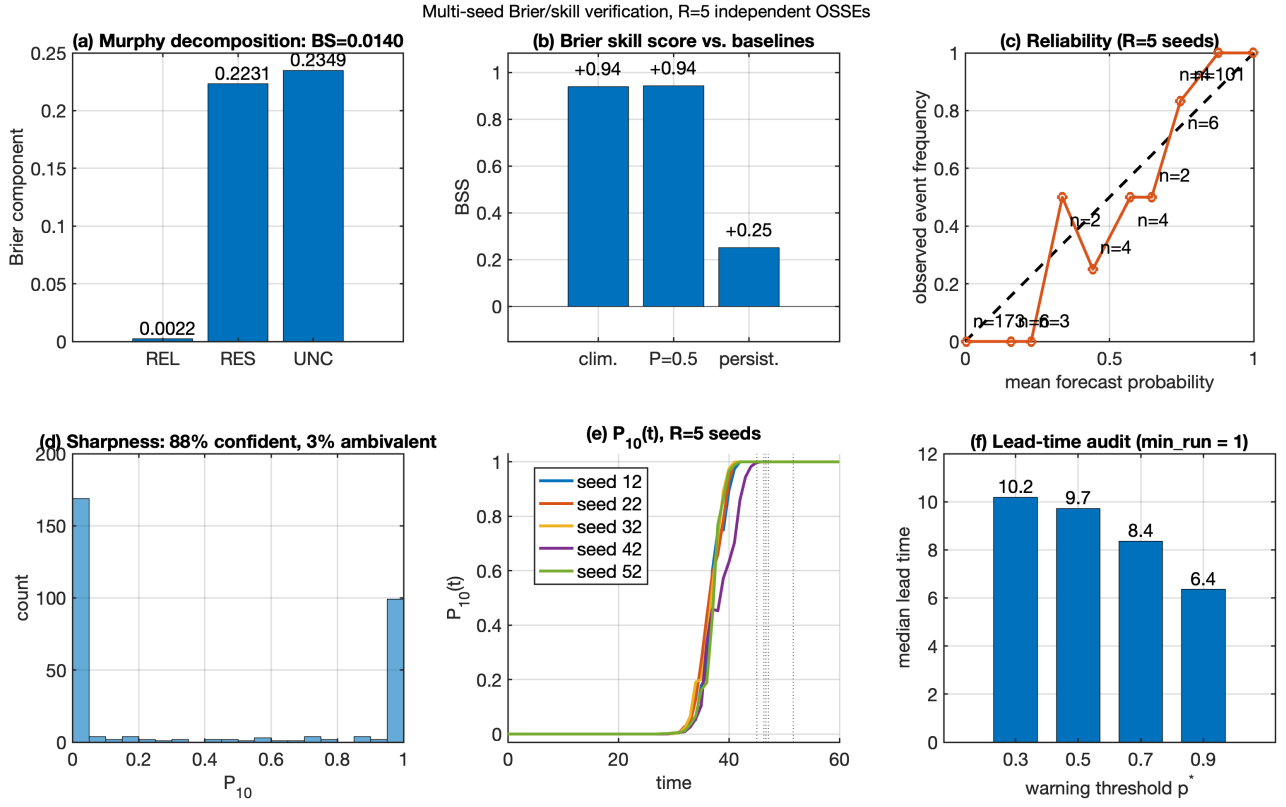


Figure 6. Brier/skill verification of the Stommel–Cessi early-warning forecast, pooled over $R = 5$ independent OSSE realizations (or seeds). (a) Murphy decomposition $B_S \sim B_{\mathrm{REL}} - B_{\mathrm{RES}} + B_{\mathrm{UNC}}$ ($B_{\mathrm{REL}} = 0.0022$, $B_{\mathrm{RES}} = 0.2231$, $B_{\mathrm{UNC}} = 0.2349$, $B_S = 0.0140$). (b) Comparison to trivial baselines ($B_{SS} \simeq +0.94$ versus the climatology and constant-0.5 references, and $B_{SS} \simeq +0.25$ versus persistence). (c) Reliability diagram across the five seeds. (d) Sharpness histogram of the forecast probability P_{10} (88% confident). (e) $P_{10}(t)$ trajectories for each of the five seeds, with thin vertical lines marking each seed's own collapse time. (f) Lead-time audit (for the first time the threshold is crossed). The median warning lead time decreases from approximately 10.2 dimensionless time units at $p^\star = 0.3$ to approximately 6.4 at $p^\star = 0.9$, with all five seeds contributing at every threshold. Here min_run indicates the number of consecutive threshold crossings required.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

(weighted by layer thickness) and over latitude within a northern box (45–60°N) and a southern box (30–34°S). T_N and S_N are the northern-box temperature and salinity, and ΔT, ΔS and $\Delta\rho$ are the north-minus-south differences; hence $\Delta T < 0$ and $\Delta\rho > 0$ in the AMOC-ON state. This diagnostic state provides a compact but physically interpretable state space on which the autoencoder, latent dynamics model, and particle filter can operate.

3.2.2 Autoencoder reconstruction

As a 10-dimensional space is still large for the application of the particle filter method, a further reduction is pursued. Many methods are available for this purpose, and we choose here a machine-learning based method (Bracco et al., 2025), an autoencoder. The first step is to test whether the 10-dimensional CESM diagnostic state can be represented accurately in a low-dimensional latent space (with dimension L) using such an autoencoder. Let $\mathbf{x}_k \in \mathbb{R}^d$, with $d = 10$, be the state vector (21) at time step t_k. We first normalize each component in the state vector by subtracting the mean and dividing by the standard deviation. We choose the latent space dimension $L = 4$. The architectures of both the encoder and decoder are multi-layer perceptrons with two hidden layers of 128 neurons and a $\tanh$ activation function. For $\mathbf{z}_k \in \mathbb{R}^L$, the encoder (Enc) and decoder (Dec) represent

$$\mathbf{z}_k = \mathrm{Enc}(\mathbf{x}_k), \quad \widehat{\mathbf{x}}_k = \mathrm{Dec}(\mathbf{z}_k). \tag{22}$$

The autoencoder pre-training loss is given by

$$\mathcal{L}_{\mathrm{AE}} = \|\mathbf{x}_k - \mathrm{Dec}(\mathrm{Enc}(\mathbf{x}_k))\|_2^2, \tag{23}$$

where $\|\|_2$ is the L_2 norm. The autoencoder is trained on years 1–3500 of the CESM data, since the AMOC-OFF state is needed but not the recovery. The continguous block of years 1500–1700 is reserved as a pre-collapse validation holdout. The training and validation losses decrease (Fig. B1 in Appendix B) rapidly during the early epochs and then approach a stable plateau. The validation curve remains close to the training curve, indicating that the autoencoder is not simply memorizing the training trajectory.

The reconstructed trajectories in Fig. 7 show that the autoencoder captures the dominant low-frequency evolution of the CESM AMOC transition. In particular, the rapid weakening of $\psi_{26N,1000m}$ and $\psi_{\max}$, the sign change and recovery structure of F_{ovS}, and the large-scale evolution of T_N, S_N, ΔT, ΔS, and $\Delta\rho$ are all represented by the decoder. This is important because the particle filter (see next subsection) never observes the full CESM model state directly; it operates by propagating particles in latent space and then decoding them back to the diagnostic state. As such, the quality of the decoder is integral as it directly controls the quality of the predicted observations used inside the likelihood function in the particle filter. However, the reconstruction is not equally strong for all variables. On the validation block, the coefficient of determination R^2 reaches 1.00 for $\psi_{\min,deep}$ and F_{ovN} and lies between 0.91 and 0.97 for F_{ovS}, T_N, S_N, ΔT and ΔS, but falls to 0.50 for $\psi_{26N,1000m}$, 0.52 for $\Delta\rho$ and 0.55 for $\psi_{\max}$. These three variables retain visibly noisier residuals because their high-frequency content is essentially unrepresentable in an $L = 4$ latent space without sacrificing reconstruction quality elsewhere.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

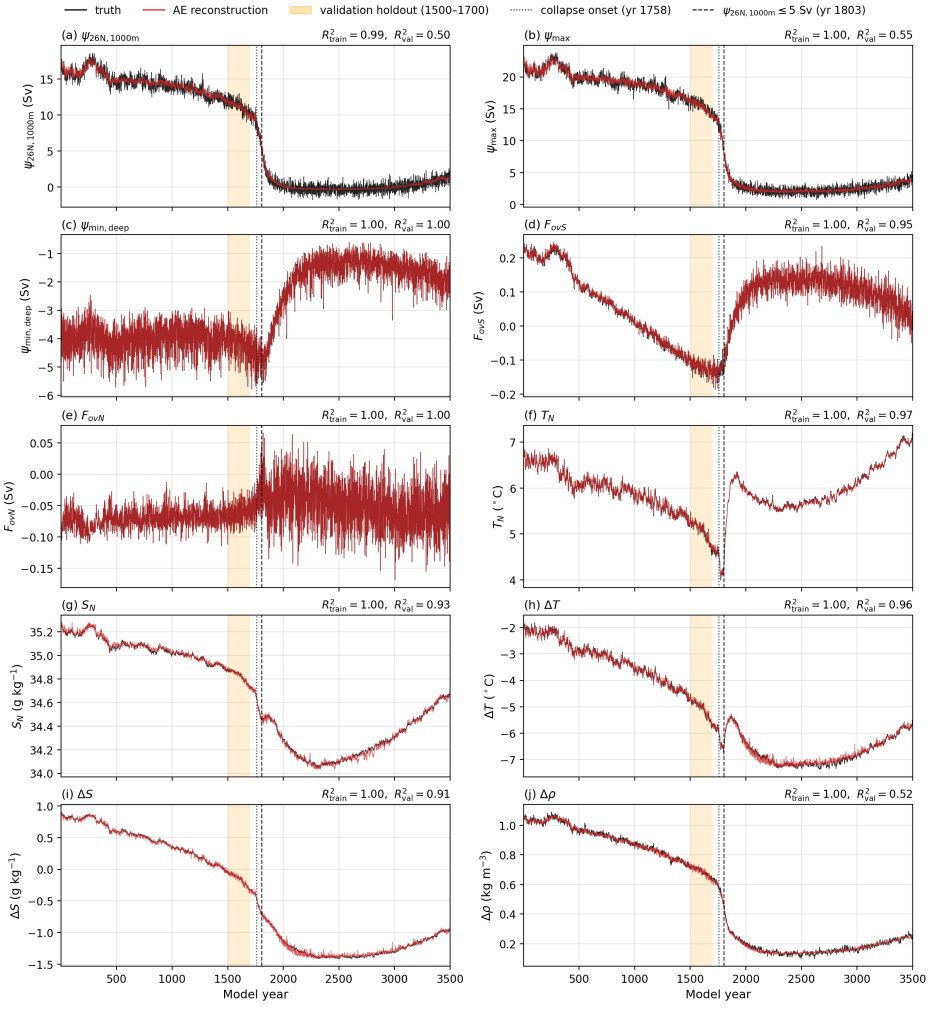


Figure 7. Truth (black) and autoencoder reconstruction (red) for all ten state-vector components over the training window 1–3500. The orange shaded band marks the validation holdout (years 1500–1700); the vertical lines mark the collapse onset (year 1758) and the crossing of $\psi_{26N,1000m} = 5$ Sv (year 1803). Panel titles give the coefficient of determination on the training set (R^2_{train}) and on the validation block (R^2_{val}). Reconstruction is near-exact for $\psi_{\mathrm{min},deep}$ and F_{ovN} ($R^2_{\mathrm{val}} = 1.00$) and good for F_{ovS}, whose sign-reversal structure is captured, and for T_N, S_N, ΔT and ΔS. It is weakest for $\psi_{26N,1000m}$ (0.50), $\Delta\rho$ (0.52) and $\psi_{\max}$ (0.55), which retain high-frequency variability that the $L = 4$ latent compression cannot represent.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.



3.2.3 Latent dynamics model

After training the autoencoder, a latent dynamics model is trained to obtain an update rule for the latent state, so we can use that in the particle filter. Here, rather than asking the neural network to predict the full next latent state $\mathbf{z}_{k+1}$ directly from the current latent state $\mathbf{z}_k$ and forcing input $\mathbf{u}_k$ (specified below), the network predicts only the increment $\Delta\mathbf{z}_k$. The update is therefore written as

$$\Delta\mathbf{z}_k = \mathbf{z}_{k+1} - \mathbf{z}_k = f_\theta(\mathbf{z}_k, \mathbf{u}_k), \tag{24}$$

where f_θ is a two hidden layer 128-neuron network with tanh activations at the two hidden layers and a linear, unactivated, output layer. In this formulation, the current state $\mathbf{z}_k$ is carried forward explicitly, and the network only has to learn the correction needed to move the system to the next time step. This choice is useful because the latent variables typically have magnitudes of order $\mathcal{O}(1)$, while the step-to-step changes are much smaller, of order $\mathcal{O}(0.1)$. Learning these smaller changes is a better-conditioned regression problem than learning the full map from $\mathbf{z}_k$ to $\mathbf{z}_{k+1}$.

The output layer of the network is initialized with its weights scaled to 1% of their default random-initialization magnitude and its bias set to zero, so at the beginning of training the model predicts only very small increments. Consequently, its initial behavior is approximately $\mathbf{z}_{k+1} \approx \mathbf{z}_k$. The external forcing input $\mathbf{u}_k \in \mathbb{R}^2$ carries two channels, $\mathbf{u}_k = [F_{H,\mathrm{norm}}(t_k), r_F(t_k)]$, where $F_{H,\mathrm{norm}} \in [0,1]$ is the normalized freshwater flux amplitude and $r_F \in \{-1, 0, +1\}$ is the sign of dF_H/dt. The reason is intrinsic to the hysteresis simulation protocol: a value such as $F_H = 0.4$ Sv (model years 1333 and 3067) occurs during both the ON-state ramp-up and the OFF-state ramp-down, so without the ramp-direction channel the network can better distinguish the two branches. Including r_F breaks this ambiguity at the input level and lets the same network f_θ represent both branches of the hysteresis simulation.

The network f_θ is trained in two stages. First, a one-step training minimizes the MSE on $\mathbf{z}_{k+1}$ prediction with the encoder and decoder frozen, using gradient clipping at norm 1.0 to stabilize early iterations. However, one-step training alone underfits the cumulative collapse dynamics: rollouts initialized at year 1500 collapse approximately 300 years late, and rollouts initialized at year 1000 do not collapse at all over an 800-year horizon. A second multi-step refinement stage therefore fine-tunes the latent dynamics on K-step free rollouts in latent space, with K following $K = 5 \to 10 \to 20 \to 30 \to 40$ growing in stages of 300–500 epochs, with the final $K = 40$ stage running for the remaining 1000 epochs. Short rollouts keep the optimization well conditioned, since error accumulation over a few steps is bounded and the loss landscape remains close to the one-step-trained initialization. K is then increased only once the network has mastered the shorter horizon, rather than asking it to fit 40-step rollouts directly from a one-step-only starting point. This forces the network to capture the cumulative dependence of the latent trajectory on F_H rather than the dominant noise term in single-step prediction.

The latent dynamics loss stabilizes without evidence of validation divergence (Fig. B2 in Appendix B). The validation loss lies below the training loss in this experiment, which is not problematic here. The holdout block lies in the early-collapse approach phase where one-step latent residuals are smaller than in the noisier multi-century stationary regime that dominates the training set. This is a property of where the validation block was placed, not of generalization. The important point is that the validation loss remains stable, indicating that the learned latent dynamics does not become unstable over the validation

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

window. The multi-step rollout experiment provides a more physically meaningful validation than the training and validation

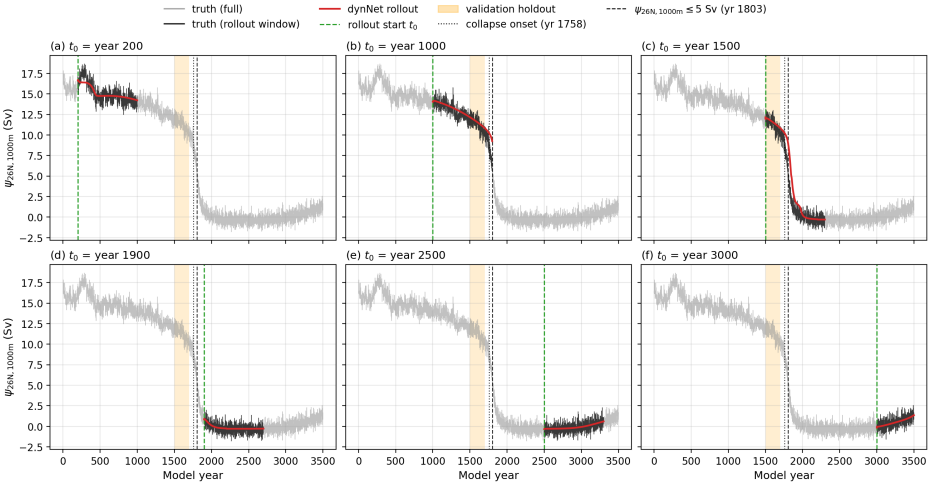


Figure 8. Multi-step rollouts of the latent dynamics network decoded to $\psi_{26N,1000m}$ over an 800-year horizon, initialized from six representative years ($t_0 \in \{200, 1000, 1500, 1900, 2500, 3000\}$). Grey shows the full CESM truth trajectory for context; black, the truth over the rollout window; red, the latent dynamics network rollout. Green dashed lines mark the rollout start year; the red dashed line marks the truth collapse year (~ 1803), where $\Psi \leq \Psi_c = 5$ Sv (so not the onset year); the orange band marks the holdout block during training. Rollouts initialized on the ON-state equilibrium ($t_0 = 200, 1000$) remain close to truth for several centuries; rollouts initialized inside the approach phase ($t_0 = 1500$) reproduce the collapse at approximately the correct year; rollouts initialized post-collapse ($t_0 = 1900, 2500, 3000$) remain in the OFF state with no spurious recovery.

loss alone. Starting from several different years (Fig. 8), the latent dynamics is integrated forward and decoded to $\psi_{26N,1000m}$. The resulting forecasts do not represent the high-frequency variability, but they capture the correct long-time scale AMOC behavior. Rollouts initialized before the transition move toward the collapse, rollouts initialized near the transition reproduce the rapid weakening of the AMOC, and rollouts initialized after the transition remain close to the collapsed OFF-state regime (Fig. 8).

3.2.4 Particle-filter

The next step is to test the particle filter on the quasi-equilibrium AMOC-ON state trajectory. This validation is necessary because a particle filter that is already inconsistent in this quasi-stationary regime cannot be interpreted reliably during the later transition. The filter represents the posterior distribution over the latent state using particles, i.e.,

$$p(\mathbf{z}_k \mid \mathbf{y}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{z}_k - \mathbf{z}_k^{(i)}), \tag{25}$$

where $\mathbf{z}_k^{(i)}$ is the state vector of the i-th latent particle and $w_k^{(i)}$ is its normalized importance weight.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

In the results below, $N = 1000$ particles are used. The synthetic observations are generated by adding Gaussian noise with standard deviation $\sigma_{\mathrm{obs},j} = \alpha\,\mathrm{std}(x_j)$ to each truth variable, where $\mathrm{std}(x_j)$ is computed over the quasi-stationary window (years 1–400) and $\alpha = 0.05$, giving a signal-to-noise ratio of 20 for every variable. In the likelihood function (10), the filter uses $\sigma_{\mathrm{filter},j} = \left(\sigma^2_{\mathrm{obs},j} + \mathrm{RMSE}^2_{\mathrm{rec},j}\right)^{1/2}$, where $\mathrm{RMSE}_{\mathrm{rec},j}$ is the autoencoder reconstruction error over the same window; 10% of the weighted ensemble spread is further added to this in quadrature (for details, see Appendix C). For the operating observation set this gives $\sigma_{\mathrm{obs}} = 0.046$ Sv and $\sigma_{\mathrm{filter}} = 0.64$ Sv for $\psi_{\mathrm{26N,1000m}}$, and $\sigma_{\mathrm{obs}} = 7.3 \times 10^{-4}$ Sv and $\sigma_{\mathrm{filter}} = 4.2 \times 10^{-3}$ Sv for F_{ovS}.

The prior–posterior PDF snapshots are shown for a representative state-vector at time step $t_k = 300$ years, within the early quasi-equilibrium window. The results for $\psi_{26N,1000m}$ and F_{ovS} are shown in Fig 9. The prior distribution corresponds to the decoded forecast ensemble before assimilating the observation; the posterior distribution shows the same ensemble after assimilation (including the likelihood-based weight update, and any resampling). For a well-calibrated filter the posterior should contract relative to the prior and shift toward the observation without reducing to an unrealistically narrow distribution. Indeed, the posterior of $\psi_{26N,1000m}$ and F_{ovS} show this behavior. The variables $\psi_{\max}$ and $\psi_{\min,deep}$ also display this behavior (not shown) at the same time step.

Detailed diagnostics of the performance of the particle filter are presented in appendix C. These results suggest that assimilating all ten reduced CESM diagnostics is not automatically optimal. Some variables are dynamically redundant, and some are inconsistent with the assumed observation-noise model. We performed an exhaustive observation-subset Pareto-feasibility analysis (not shown), but that did not deliver a clear result. We hence decided to base the operating observation set $\{F_{ovS}, \psi_{26N,1000m}\}$ on physical grounds. F_{ovS} is a measure of the salt-advection feedback, and $\psi_{26N,1000m}$ directly measures AMOC strength at the RAPID-MOCHA latitude and clearly distinguishes the non-collapsed and collapsed regimes.

3.2.5 Warning probability

The collapse threshold is defined as $\psi_{26N,1000m} \leq \Psi_c = 5$ Sv. The truth trajectory crosses this threshold at approximately year $t_{\mathrm{collapse}} \approx 1803$ (Fig. 10a). At each saved learning time t_k, the posterior particle ensemble $\mathbf{z}_k^{(i)}, w_k^{(i)}$ for $i = 1, \ldots, N$ is projected forward over the horizon $[t_k, t_k + \tau]$ under the actual forcing scenario, using the learned latent dynamics without further assimilation. Each projected particle is decoded at every step of the horizon to obtain its AMOC strength. The posterior predictive collapse probability is then

$$P_\tau(t_k) = \sum_{i=1}^{N} w_k^{(i)} \mathbf{1}\left[\min_{t_k \leq t \leq t_k+\tau} \psi^{(i)}_{26N,1000m}(t; t_k) \leq \Psi_c\right]. \tag{26}$$

where $\tau = 150$ years was chosen here. The projection uses a latent process noise $\sigma_{\mathrm{forecast}} = \sqrt{Q_{\mathrm{diag}}}$, where Q_{diag} is the diagonal of the empirical one-step latent residual covariance calibrated over the stationary window, applied per step rather than $\sigma_{\mathrm{forecast}} = 0$ as in the original Singh et al. (2018) formulation. The added process noise restores the spread of the projected ensemble around its deterministic centroid. Saved snapshots are taken every 25 years, and the warning years quoted below are

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

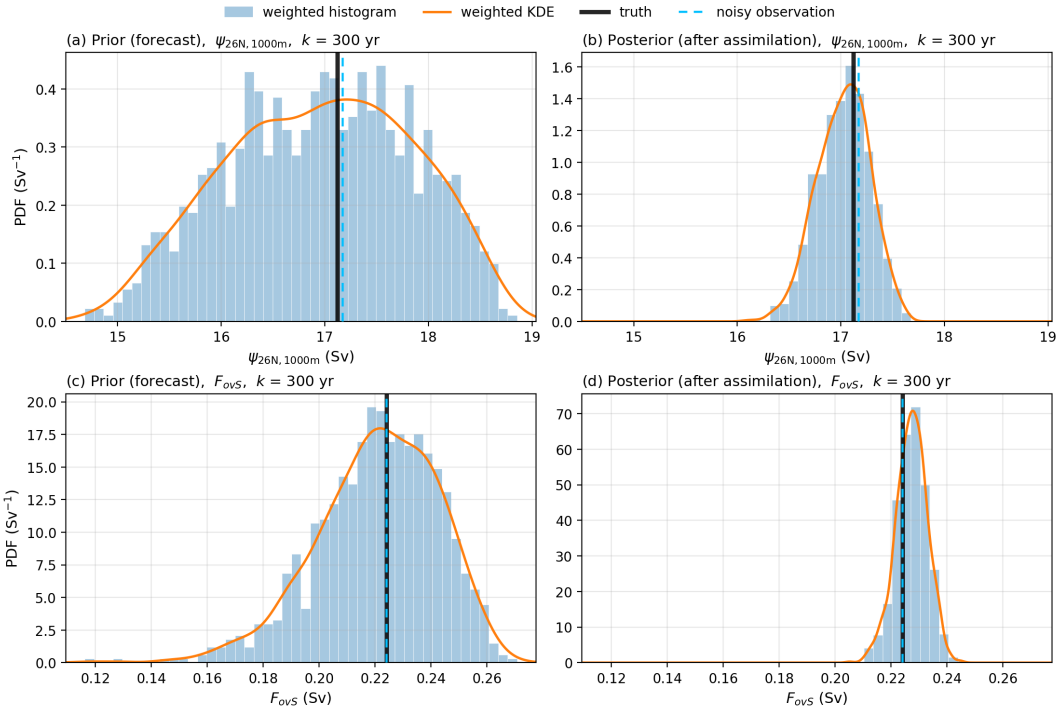


Figure 9. Prior (left) and posterior (right) probability density functions for $N = 1000$ particles, for (a,b) $\psi_{26N,1000m}$ and (c,d) F_{ovS} at $t_k = 300$ years within the stationary CESM window. Light-blue: weighted histogram of decoded particles; orange: weighted KDE overlay; black dashed: truth value; cyan dashed: noisy observation. The prior is the decoded forecast ensemble before assimilation at step k; the posterior is the same ensemble after assimilation. At this time step the posterior contracts sharply around the noisy observation, with the truth–posterior gap of order σ_{obs}, as expected from the all-variable likelihood.

defined as the first persistent threshold crossing, requiring at least two consecutive snapshots (i.e., min_run = 2) above the level to avoid registering single-snapshot blips.

The warning probability (Fig. 10b) remains close to zero during the early part of the simulation and then rises sharply (Fig. 10c) before the true collapse year. The probability exceeds 0.5 at approximately year $t(P_\tau \geq 0.5) \approx 1651$, corresponding to a lead time of about $1803 - 1651 \approx 152$ years, and exceeds 0.9 at approximately year $t(P_\tau \geq 0.9) \approx 1726$, a lead time of about $1803 - 1726 \approx 77$ years. The result indicates that the posterior particle distribution contains enough information by approximately year 1650 to assign more than 50% probability to a collapse within the following $\tau = 150$ years, and by approximately year 1725 that probability exceeds 90%.

These warning years are subject to three sources of uncertainty. First, a crossing year can only be resolved to the snapshot stride, here 25 years. Second, the autoencoder and the latent dynamics network are retrained for each realisation of the experiment, and the resulting latent coordinates are not identical between trainings; repeating the training moves each crossing by one snapshot interval. Third, the particle-filter seed contributes an interquartile spread in P_τ of at most 0.026 across the $N_r = 5$

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

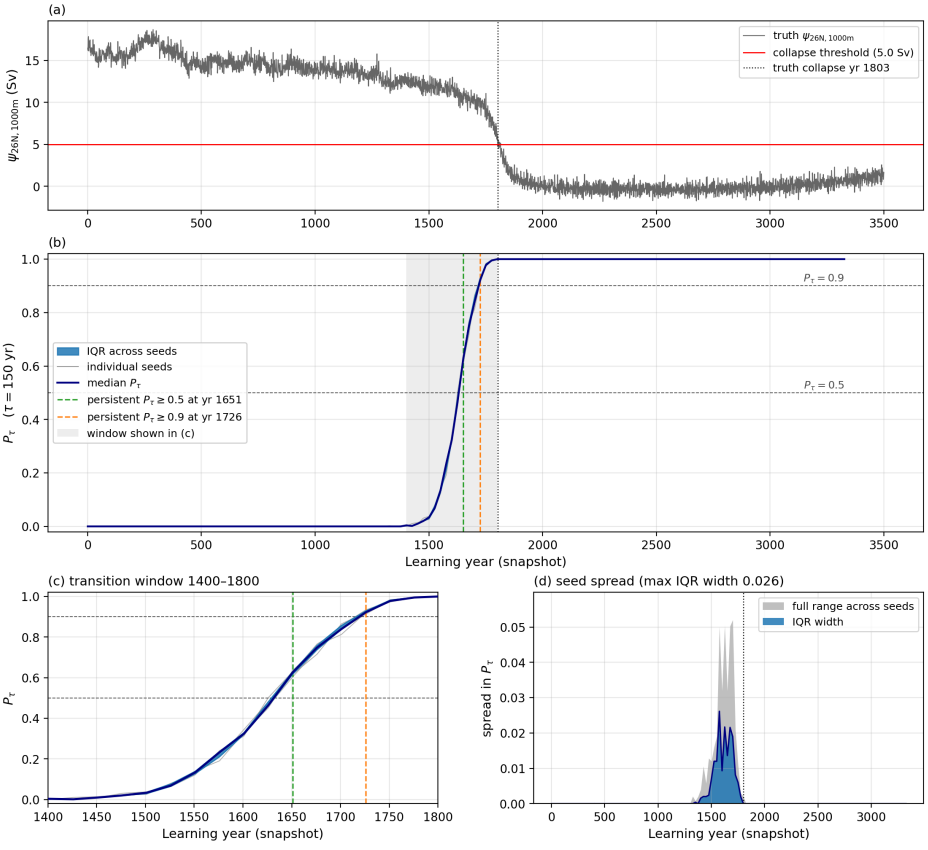


Figure 10. Warning probability P_τ for a rolling horizon of $\tau = 150$ years, as a function of learning year t_k. (a) Truth $\psi_{26N,1000m}$ (grey) with the collapse threshold $\Psi_c = 5$ Sv (red) and the truth collapse year $t_{\text{collapse}} \approx 1803$ (black dotted). (b) $P_\tau(t_k)$ over the full range of saved snapshots, computed by projecting the posterior particle ensemble at each t_k forward over $[t_k, t_k + \tau]$ under the actual forcing and testing first passage of the decoded $\psi_{26N,1000m}$ below Ψ_c. Snapshots are saved every 25 years; the dark blue line is the median over $N_r = 5$ particle-filter seeds, the blue band their interquartile range, and the thin grey lines the individual seeds. The green and orange dashed lines mark the first persistent crossings of $P_\tau = 0.5$ and $P_\tau = 0.9$, at years 1651 and 1726, requiring at least two consecutive snapshots above the level. The grey shading indicates the window shown in (c). (c) The same curves over the transition window 1400–1800. (d) Spread across the five seeds: interquartile width (blue) and full range (grey). The spread is non-zero only during the transition, reaching a maximum interquartile width of 0.026, since elsewhere every seed gives $P_\tau \approx 0$ or 1. Note that a snapshot is only scored when the full horizon τ fits inside the training window, so the probability curve ends at year 3350 while the truth in (a) extends to 3500.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

particle-filter seeds, peaking at year 1576 on the rising limb and confined to the transition window (Fig. 10d), since elsewhere every seed gives $P_\tau \approx 0$ or $P_\tau \approx 1$. The warning years are therefore determined to approximately one snapshot interval, with the training-to-training variability dominating the contribution of the filter itself.

Two limitations should be kept in mind. The projection uses the actual forcing scenario over the time horizon rather than a forecast of it, so P_τ measures the information content of the posterior state estimate under a known forcing trajectory, not the combined uncertainty of state and forcing. The transition from low to high probability also appears abrupt in the figure even though the underlying curve is continuous in t_k, because the probability is evaluated only at saved snapshots. Even with these limitations, the result demonstrates that the autoencoder–particle filter framework can be used not only for state estimation, but also for probabilistic prediction of a future tipping event.

3.2.6 Performance

Also in the CESM example, a Brier-score verification has been applied to measure the skill of the probabilistic forecast (Figure 11) using $N_r = 5$ particle-filter seeds. Unlike the box-model case, CESM provides a single truth trajectory, so these seeds sample the filter and forecast noise rather than independent realizations of the event. The forecast obtains a Brier score $\mathrm{B_S} = 0.0067$, and its Murphy decomposition (Figure 11a) $\mathrm{B_S} = \mathrm{B_{REL}} - \mathrm{B_{RES}} + \mathrm{B_{UNC}}$ pairs a small reliability term ($\mathrm{B_{REL}} = 0.0066$) with a resolution ($\mathrm{B_{RES}} = 0.25$) that recovers almost all of the climatological uncertainty ($\mathrm{B_{UNC}} = 0.25$). The forecast is sharp (Figure 11e), assigning confident probabilities near 0 or 1 to 93% of snapshots, with 1% ambivalent. Relative to the reference forecasts it scores a Brier skill score $\mathrm{B_{SS}} = +0.97$ against climatology, $+0.85$ against persistence of the observed state and $+0.82$ against linear extrapolation (Figure 11b). The score is stable under changes in the snapshot stride (save_every) and temporal offset, with $\mathrm{B_S}$ between 0.004 and 0.012 and $\mathrm{B_{SS}}$ above $+0.95$ with respect to climatology (Figures 11c,d).

These numbers should be read with the composition of the verification sample in mind. The scored snapshots span the full window, so roughly half of them lie after the collapse, where the event has already occurred and the forecast correctly returns $P_\tau \approx 1$. The base rate is therefore close to 0.5 (Fig. 11e) and the climatological uncertainty term takes its maximum value $\mathrm{B_{UNC}} = 0.25$; a substantial part of the resolution, and hence of the skill scores, is earned on snapshots that are trivially classified. The scores demonstrate that the forecast is calibrated and does not merely track the base rate, but they overstate the difficulty of the task relative to a verification restricted to the pre-collapse period. The lead-time audit (Fig. 11f) gives a median warning lead of ≈ 77 years at $p^* \geq 0.9$ and ≈ 127 years at $p^* \geq 0.7$, identical across the min_run persistence settings. At the more permissive thresholds the audit returns ≈ 152 years at $p^* \geq 0.5$ and ≈ 202 years at $p^* \geq 0.3$; the latter exceeds the horizon $\tau = 150$ years over which P_τ is defined, and the former lies at that ceiling, so neither can really be interpreted as a warning issued that far ahead. Only the leads at $p^* \geq 0.7$ and above are bounded strictly by τ and are quoted as such.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

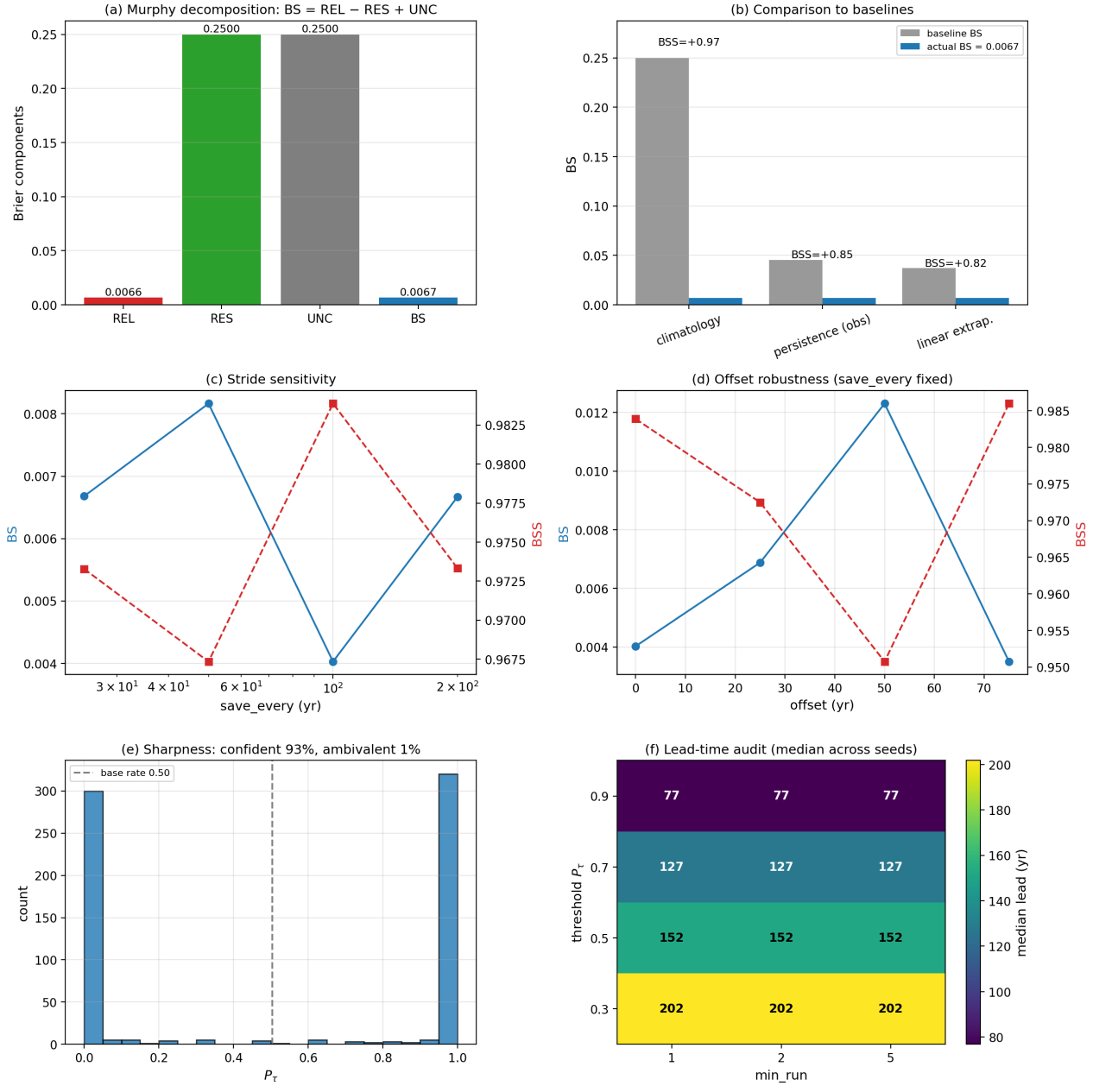


Figure 11. Brier-score verification of the CESM forecast. (a) Murphy decomposition $\mathrm{B_S} = \mathrm{B_{REL}} - \mathrm{B_{RES}} + \mathrm{B_{UNC}}$ ($\mathrm{B_{REL}} = 0.0066$, $\mathrm{B_{RES}} = 0.2500$, $\mathrm{B_{UNC}} = 0.2500$, $\mathrm{B_S} = 0.0067$). (b) Comparison to the reference forecasts ($\mathrm{B_{SS}} = +0.97$ versus climatology, $+0.85$ versus persistence of the observed state and $+0.82$ versus bootstrapped linear extrapolation). (c) Stride and (d) offset robustness of $\mathrm{B_S}$ and $\mathrm{B_{SS}}$. (e) Sharpness histogram of P_τ (93% confident, 1% ambivalent; the dashed line marks the base rate of 0.50). (f) Lead-time audit (median across seeds): median warning lead of 77 years at $p^* \geq 0.9$, 127 years at $p^* \geq 0.7$, 152 years at $p^* \geq 0.5$ and 202 years at $p^* \geq 0.3$, stable across the min_run setting.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

4 Summary and Discussion

A skillful warning system for a strong AMOC weakening at decadal-to-centennial lead times is important for disaster preparedness and for deep adaptation (Nummelin et al., 2026; Biesbroek et al., 2025). In this paper, we have provided a prototype of a probability based warning system. It consists of a model update rule (equation or data based), a particle filter method, a method to compute warning probabilities, all embedded in an OSSE testing framework. Note that this is still a prototype, as much work has to be done to generalize it to include effects of model bias, to assimilate real observations and to address the sensor location problem.

The approach was illustrated using the Stommel-Cessi box model (Stommel, 1961; Cessi, 1994), where the update rule is provided by two coupled stochastic differential equations and both the bifurcation diagram and transition behavior are well known. Because of its small dimensionality, a large ensemble can be used in the particle filter and in the computation of the transition probability. Here, we illustrated the use of Trajectory Adaptive Multilevel Splitting (TAMS) for the transition probability computation, although this method is only needed when transition probabilities are low and/or when evaluating the update rule is expensive. It was subsequently applied to output from the Community Earth System Model (CESM) under a background pre-industrial climate and quasi-equilibrium freshwater flux forcing. For this case, we build a data-based update rule using an autoencoder method from a 10-dimensional physics-informed set of quantities of the CESM state vector. In this case, both particle filtering and warning probability are carried out in the latent space of the autoencoder; a Monte Carlo method was used for determining the warning probability.

The final result obtained is an estimate of the warning probability

$$P_\tau(t_k) = \mathbb{P}(t_c \le t_k + \tau \mid \mathbf{y}_{1:k}), \tag{27}$$

where t_c is the first time that the AMOC strength Ψ at 26°N becomes smaller than a critical value Ψ_c, τ is the chosen time horizon and $\mathbf{y}_{1:k}$ denotes the observation vector up to the present analysis time t_k. Such a quantity can be used to provide alarms when specific probabilities are exceeded for given forecast horizons. Using an OSSE approach, it was shown that this warning system has high skill in providing a warning for the AMOC collapse, as found in the CESM (van Westen et al., 2024), giving for example a probability exceeding 0.7 about 127 years before the collapse threshold. The Brier Skill score is about +0.85 compared to persistence.

There are three key challenges in this approach. The first is determining a proper model update rule. In the CESM case, it is impossible to use the full model state vector evolution because it is then computationally too expensive to use a large ensemble. The second is the weight degeneracy problem (see also Appendix C) in the particle filter method (Frei and Künsch, 2013), being one of the reasons why such a method is not used operationally in weather (and for example hurricane track) forecasting. The third is the computation of transition probabilities, where also a large ensemble (when using Monte-Carlo) or an adequate score function (when using TAMS) are required. There is room for large improvements here which are left for future work. In addition, we have not considered any model errors when using a perfect-model OSSE, where the truth model, particle-filter forecast model, and TAMS/Monte Carlo model are identical for the box model; for the CESM, only emulator error is considered. This isolates only uncertainty due to observational noise, finite particle number, finite Monte Carlo or

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

TAMS sampling, and finite time stepping, but not model biases. The latter bias is an important issue as model differences are the dominant uncertainty of 21st-century AMOC weakening in CMIP6 models (Portmann et al., 2026).

We have focussed here on an AMOC warning system, but the approach can also be applied to other tipping phenomena (Lenton et al., 2025). Of particular interest are transitions in the North Atlantic Subpolar Gyre (SPG). Here, the mechanisms of transition are related to changes in convection, boundary layer current strength and mixed-layer depth and hence different observables have to be used (Born and Stocker, 2014) than for the AMOC. Because the transition time scale is considered to be shorter than for the AMOC, it may be more feasible to use large ensembles, even for Earth System Models, to obtain a skillful warning probability.

Finally, we have not addressed the sensor placement problem where some of the observations contribute much more information to the warning probability than others. However, this problem can be addressed systematically in our OSSE approach, where the observations used in the particle filter can be varied to find an optimal sensor arrangement for a warning probability. For the AMOC, this may not be so relevant as the available observation locations (from array data or satellites) may not change over the next decades, but for the SPG transitions, this may likely be useful.

Code and data availability. The codes and data to reproduce all the figures in the main text of the paper can be downloaded from https://doi.org/10.5281/zenodo.22860525.

Author contributions. N.V. and H.A.D developed the idea, N.V. carried out all the work on the CESM model, R.M.v.W. contributed with data from the CESM. All authors contributed to the writing of the paper.

Competing interests. None

Acknowledgements. The CESM simulation and the analysis of all the model output were conducted on the Dutch National Supercomputer Snellius within NWO-SURF project 2025.038. The authors thank Michael Kliphuis (IMAU, UU) for performing these simulations. R.M.v.W., O.M. and H.A.D. are funded by the European Research Council through the ERC-AdG project TAOC (project 101055096; PI: Dijkstra). O.M. and H.A.D were also partially funded by the ARIA project AdvanTip (project code SCOP-PR01-P003). This paper is part of the MSc thesis of NV at the Utrecht University. AI tools were used for coding and for improving details of the text.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix A: Trajectory Adaptive Multilevel Splitting

After the particle filter approach, the probability $P_\tau(t_k)$ is approximated by

$$P_\tau(t_k) \approx \sum_{i=1}^{N} w_k^{(i)} p_\tau(\boldsymbol{x}_k^{(i)}). \tag{A1}$$

where for using TAMS above Monte Carlo methods, it is assumed that $p_\tau(\boldsymbol{x})$ is very small. In TAMS, a score function S is required which monitors the progression towards the desired target. For a simulated trajectory $\omega = \{\boldsymbol{x}_n\}_{n=0}^{N_\tau}$ over a time τ, define its maximum progress score as

$$S(\omega) = \max_{0 \le n \le N_\tau} \xi(\boldsymbol{x}_n) \; ; \; \xi(\boldsymbol{x}_n) = \Psi_c - \Psi \tag{A2}$$

and let the transition event be indicated as

$$\mathcal{A}_\tau = \{S(\omega) \ge 0\}. \tag{A3}$$

For a fixed initial condition $\boldsymbol{x}$, TAMS begins with N_T independent trajectories $\omega^{(1)}, \ldots, \omega^{(N_T)}$. For each trajectory compute $S_j = S(\omega^{(j)})$, it finds the least advanced trajectory $j_\star = \arg\min_j S_j$ and its score $\ell = S_{j_\star}$. Next, it removes trajectory $j_\star$ and selects a survivor trajectory j with $S_j > \ell$. Finally it finds the first time $\tau_\ell^{(j)}$ at which $\xi(\boldsymbol{x}_\tau^{(j)}) \ge \ell$, clones the trajectory j up to $\tau_\ell^{(j)}$ and continues the clone with a new stochastic forcing. If K killing/cloning steps occur before termination and a number N_{hit} of the final replicas reach the event, the one-kill TAMS estimator is

$$\widehat{p}_\tau^{\text{TAMS}}(\boldsymbol{x}) = \left(1 - \frac{1}{N_T}\right)^K \frac{N_{\text{hit}}}{N_T}. \tag{A4}$$

The full observation-conditioned probability estimator is therefore

$$\widehat{P}_\tau(t_k) = \sum_{i=1}^{N} w_k^{(i)} \widehat{p}_\tau^{\text{TAMS}}(\boldsymbol{x}_k^{(i)}). \tag{A5}$$

Appendix B: Auto-encoder performance

The hyperparameters of the autoencoder and the latent dynamics network are listed in Table B1. All ten diagnostics are standardised with the mean and standard deviation of the training set (years 1–1499 and 1701–3500); years 1500–1700 are held out for validation. The encoder and decoder are fully connected networks with two hidden layers of 128 units and $\tanh$ activations, mapping the ten diagnostics to a latent state of dimension $L = 4$ and back. The autoencoder is trained on the squared reconstruction error summed over the ten standardised variables.

The latent dynamics network has the same hidden structure and predicts the increment $\Delta \mathbf{z}_k$ from $(\mathbf{z}_k, \mathbf{u}_k)$, with $\mathbf{u}_k = (F_H/0.66, r_F)$. Its output layer is initialised close to zero, so that training starts from persistence. It is first trained on one-step increments (Fig. B2) and then refined on free multi-step rollouts with a curriculum in the rollout length $K = 5, 10, 20, 30, 40,$

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table B1. Hyperparameters of the autoencoder and latent dynamics network.

	Autoencoder	Dynamics (one-step)	Dynamics (multi-step)
Architecture	$10-128-128-4$ (+ mirror)	$6-128-128-4$	as one-step
Activation	$\tanh$	$\tanh$	$\tanh$
Parameters	36 878	17 924	17 924
Optimiser	Adam	Adam	Adam
Learning rate	10^{-3}	10^{-3}	3×10^{-4}
Batch size	264	264	264
Updates	5000	5000	2500
Gradient clipping (norm)	–	1.0	1.0
Rollout length K	–	1	5, 10, 20, 30, 40
Training / validation samples	3299 / 201	3297 / 200 pairs	–

which is required for the rollouts to reproduce the timing of the collapse (Fig. 8). In both networks an "epoch" denotes one gradient update on a mini-batch of 264 samples drawn at random from the training set.

A limited sensitivity test was performed at fixed $L = 4$ with three alternative autoencoder configurations: a decoder-Jacobian penalty ($\lambda = 10^{-2}$ and 5×10^{-2}) and increased loss weights for F_{ovS}, F_{ovN}, T_N and S_N ($\lambda = 10^{-3}$). None of these reduced the ratio of decoded prediction spread to filter observation noise below a factor of two for all variables, so the unregularised configuration was retained. The latent dimension itself was not varied. Figures B1 and B2 show the performance of the autoencoder and the latent dynamics network.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

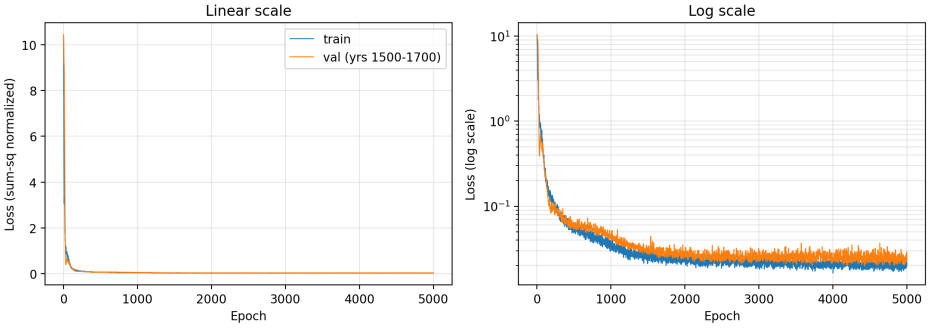


Figure B1. Training and validation loss for the autoencoder over 5000 epochs, on linear (left) and logarithmic (right) scales. The training set comprises years 1–1499 and 1701–3500, with years 1500–1700 held out as the validation block, chosen to span the pre-collapse approach into collapse onset. Loss is the mean per-sample sum-of-squared errors over the ten normalized diagnostics. Both curves plateau by epoch ~ 2500 and remain close throughout, indicating no overfit; the validation loss tracking the training loss across the regime transition is the key generalization check before the AE is fed into the latent dynamics and particle filter pipelines.

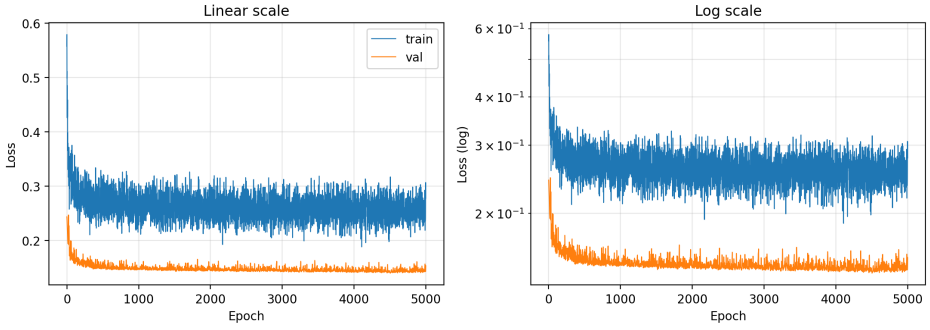


Figure B2. Training and validation loss for the latent dynamics network over 5000 epochs of one-step prediction training, on linear (left) and logarithmic (right) scales. The validation loss lies below the training loss consistently; this reflects the placement of the holdout block (years 1500–1700) in the early-collapse approach phase, where one-step latent residuals are smaller than in the noisier multi-century stationary regime, rather than any anomaly in generalization. The plateau is reached by epoch ~ 1000 and the loss remains stable thereafter, confirming that the residual dynamics of Eq. (24) with two-channel forcing (F_H, r_F) is well-posed on the training data.

Appendix C: CESM Particle filter performance

For CESM, filtering is performed in latent space, which keeps propagation cheap and lets uncertainty be controlled through a low-dimensional process noise. Observation likelihoods, by contrast, are evaluated in the original physical space, since that is where the data live and where the noise model is specified. The weight degeneracy in the particle filter is monitored by the effective sample size, $\mathrm{ESS} = \left(\sum_i (w^{(i)})^2\right)^{-1}$, and handled by an EnKF-guided proposal in latent space (Frei and Künsch, 2013). The cross-covariance between latent particles and their decoded observations produces a Kalman gain that translates

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

observation-space innovations into latent-space corrections; each particle is then nudged toward a particle-specific perturbed observation before the likelihood is evaluated. The observation-noise covariance is inflated by a factor $1/\gamma$ so that the nudge is damped instead of collapsing the ensemble onto a deterministic estimate, and the importance weights are tempered by $(1-\gamma)$, which allows the proposal to remain approximately consistent with the posterior. This single change shifts the operating regime from almost-always-degenerate to one in which the filter can run the full trajectory without continuous emergency reseeding.

It is important to note that the EnKF approach does not, by itself, guarantee that the filter uncertainty is statistically well calibrated. This calibration is assessed using the Normalized Innovation Squared (NIS), defined as

$$\mathrm{NIS} = \boldsymbol{\nu}^T S^{-1} \boldsymbol{\nu}, \tag{C1}$$

where the innovation vector is

$$\boldsymbol{\nu} = \mathbf{y}_{\mathrm{obs}} - \hat{\mathbf{y}}_{\mathrm{prior}}, \tag{C2}$$

and the innovation covariance is

$$S = \mathrm{Cov}_w(\hat{\mathbf{y}}_{\mathrm{prior}}) + R. \tag{C3}$$

Here, $\mathrm{Cov}_w(\hat{\mathbf{y}}_{\mathrm{prior}})$ is the weighted covariance of the predicted observations computed from the prior ensemble, before the EnKF nudge is applied, and R is the observation-error covariance matrix. NIS is a practical diagnostic as it distinguishes a well calibrated filter from one which produces a low posterior spread through an overconfident likelihood or, on the other hand, from one whose R has been inflated so considerably that NIS is far below the $\chi^2_{n,obs}$ expectation.

These two failure modes are interpreted in opposite ways. Values of $\mathrm{NIS} \gg \mathbb{E}[\chi^2]$ suggest that the forecast is drifting away from the observations, or that the filter is over-confident in its predicted state. In contrast, values of $\mathrm{NIS} \ll \mathbb{E}[\chi^2]$ indicate that the innovation covariance S is too large, meaning that the diagonal observation-error covariance R absorbs more uncertainty than is justified by the actual observation mismatch. This second behavior is observed in the all-variable CESM filter: $\mathrm{NIS}/n_{\mathrm{obs}} \ll 1$ for almost all candidate observation subsets. Therefore, the limiting factor is not the ability of the observations to infer the target variable, but rather the mis-specification of σ_{obs}, which controls which observation subsets are statistically feasible when assimilated together. In the CESM implementation, the observation-noise budget was separated into a truth-side observation noise and a filter-side observation noise, each calibrated per channel as a fraction of the corresponding climatological variability, following the covariance-calibration perspective of Mehra (1970) and Dee (1995). This avoids carrying a single hard-coded value of σ_{obs} across all variables.

In addition, the operating observation set was reduced from all ten state-vector elements to $\{F_{ovS}, \psi_{26N,1000m}\}$. This removes, among others, $\psi_{\mathrm{min},deep}$ and F_{ovN}, whose small filter-side observation error (a consequence of their near-perfect reconstruction) dominated the likelihood and caused the intermittent $\mathrm{ESS} = 1$ collapses of the all-variable filter. To prevent the filter from degenerating, the EnKF nudge and the weight update are applied adaptively: a step is assimilated only if at least three steps have passed since the previous assimilation and the effective sample size before the update exceeds $0.3N$. Over the stationary window this leaves 134 assimilation steps out of 400. With the operating set, the minimum ESS at assimilation steps

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

is $374/1000$ (40 resamples, $\sim 30.0\%$ over 134 assimilations; Figure C1a). The filter remains under-confident, with a mean NIS of 0.82 against $E[\chi^2_2] = 2$, and the lag-1 innovation autocorrelation exceeds the white-noise bound for both channels (0.17 for F_{ovS} and 0.44 for $\psi_{26N,1000m}$), partly because innovations are also evaluated on non-assimilation steps. This is consistent with the σ_{obs} misspecification discussed above.

The same mismatch explains the posterior-mean errors in Figure C1d. Measured in units of $\sigma_{\text{obs,filter}}$, the error reaches 16.5 for $\psi_{\text{min},deep}$, 11.8 for T_N and 10.4 for F_{ovN}, which would suggest that the filter fails badly on these channels. Measured against their own climatological variability, however, the posterior-mean error lies between 0.6 and 1.0 for all ten diagnostics, with a median of 0.74. The filter therefore estimates every channel to a comparable physical accuracy; what differs is the width of the error bar it was assigned, which for the best-reconstructed channels is only 6 to 9% of their natural variability.

For CESM, the all-variable filter is strongly under-confident (NIS$/n_{\text{obs}} \ll 1$) and suffers intermittent ESS collapses driven by the tight filter-side observation error of the best-reconstructed channels, $\psi_{\text{min,deep}}$ and F_{ovN}. Reducing the operating set to $\{F_{\text{ovS}}, \psi_{\text{26N,1000m}}\}$ and separating truth-side from filter-side observation noise removes these collapses. The resulting diagnostics on the stationary window (Figure C1) show a mean NIS of 0.82 against a reference mean of 2, falling within its confidence band 94% of the time, and a posterior on $\psi_{\text{26N,1000m}}$ that tracks the observations at the $\sigma_{\text{obs,filter}}$ level.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

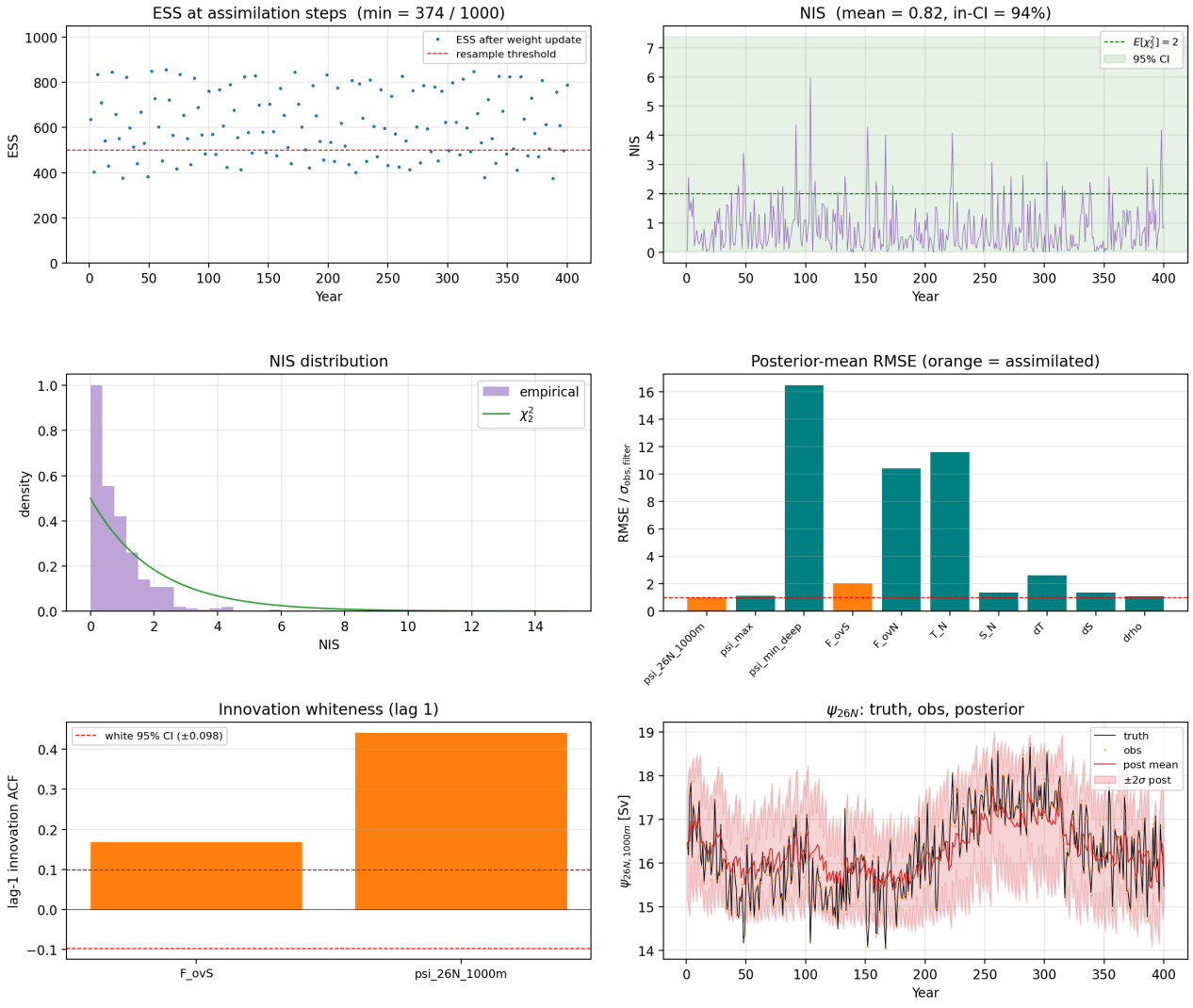


Figure C1. CESM particle-filter diagnostics on the stationary window (years 1–400) for the operating configuration $\{F_{ovS}, \psi_{26N,1000m}\}$ with EnKF-guided proposal ($\gamma = 0.5$): (a) ESS after the weight update at assimilation steps; (b) NIS time series with the χ^2_2 95% interval; (c) NIS distribution against χ^2_2; (d) posterior-mean RMSE in units of $\sigma_{\mathrm{obs,filter}}$ (orange: assimilated channels); (e) lag-1 innovation autocorrelation; (f) $\psi_{26N,1000m}$ truth, observations and posterior ($\pm 2\sigma$).

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Arumí-Planas, C., Dong, S., Perez, R., Harrison, M. J., Farneti, R., and Hernández-Guerra, A.: A multi-data set analysis of the freshwater transport by the Atlantic meridional overturning circulation at nominally 34.5 S, Journal of Geophysical Research: Oceans, 129, e2023JC020 558, 2024.

Ashwin, P., Wieczorek, S., Vitolo, R., and Cox, P.: Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system, Philosophical Transactions of the Royal Society A, 370, 1166–1184, 2012.

Baker, J. A., Bell, M. J., Jackson, L. C., Vallis, G. K., Watson, A. J., and Wood, R. A.: Continued Atlantic overturning circulation even under climate extremes, Nature, 638, 987–994, https://doi.org/10.1038/s41586-024-08544-0, 2025.

Bellomo, K., Angeloni, M., Corti, S., and von Hardenberg, J.: Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response, Nature Communications, 12, 3659, https://doi.org/10.1038/s41467-021-24015-w, 2021.

Ben-Yami, M., Morr, A., Bathiany, S., and Boers, N.: Uncertainties too large to predict tipping times of major Earth system components., Science Advances, 10, https://doi.org/10.1126/sciadv.adl4841, 2024.

Biesbroek, R., Haasnoot, M., Mach, K. J., and Petersen, A. C.: Adaptation planning in the context of a weakening and possibly collapsing Atlantic Meridional Overturning Circulation (AMOC), Regional Environmental Change, 25, 93, https://doi.org/10.1007/s10113-025-02434-5, 2025.

Boers, N.: Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation, Nature Climate Change, 11, 680–688, https://doi.org/10.1038/s41558-021-01097-4, 2021.

Bonan, D. B., Thompson, A. F., Schneider, T., Zanna, L., Armour, K. C., and Sun, S.: Observational Constraints Imply Limited Future Atlantic Meridional Overturning Circulation Weakening, Nature Geoscience, 18, 479–487, https://doi.org/10.1038/s41561-025-01709-0, 2025.

Born, A. and Stocker, T.: Two Stable Equilibria of the Atlantic Subpolar Gyre, Journal of Physical Oceanography, 44, 246–264, https://doi.org/10.1175/JPO-D-13-073.1, 2014.

Bracco, A., Brajard, J., Dijkstra, H. A., Hassanzadeh, P., Lessig, C., and Monteleoni, C.: Machine Learning for the Physics of Climate, Nature Reviews Physics, 7, 6–20, https://doi.org/10.1038/s42254-024-00776-3, 2025.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Bury, T. M., Sujith, R. I., Pavithran, I., Scheffer, M., Lenton, T. M., Anand, M., and Bauch, C. T.: Deep learning for early warning signals of tipping points, Proceedings of the National Academy of Sciences, 118, e2106140 118, https://doi.org/10.1073/pnas.2106140118, 2021.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation, Nature, 556, 191–196, https://doi.org/10.1038/s41586-018-0006-5, 2018.

Cessi, P.: A Simple Box Model of Stochastically Forced Thermohaline Flow, Journal of Physical Oceanography, 24, 1911–1920, https://doi.org/10.1175/1520-0485(1994)024<1911:ASBMOS>2.0.CO;2, 1994.

Dee, D. P.: On-line estimation of error covariance parameters for atmospheric data assimilation, Monthly Weather Review, 123, 1128–1145, https://doi.org/10.1175/1520-0493(1995)123<1128:OLEOEC>2.0.CO;2, 1995.

Dijkstra, H. A. and van Westen, R. M.: The Probability of an AMOC Collapse Onset in the Twenty-First Century, Annual Review of Marine Science, 18, 23–46, https://doi.org/10.1146/annurev-marine-040324-024822, 2026.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

Dijkstra, H. A., van Westen, R. M., Boot, A. A., Soons, J., Smolders, E., Vanderborght, E., Abe-Ouchi, A., Arumí-Planas, C., Ben-Yami, M., Boers, N., Burmeister, K., Cessi, P., Ditlevsen, P., Lohmann, J., Drijfhout, S., England, M. H., Katsman, C., Mecking, J., Rahmstorf, S., Srokosz, M., Weijer, W., and Wood, R.: Multi-Stability of the Present-Day Atlantic Meridional Overturning Circulation, WIREs Climate Change, 17, e70 049, https://doi.org/https://doi.org/10.1002/wcc.70049, e70049 WCC-1333.R2, 2026.

Ditlevsen, P. and Ditlevsen, S.: Warning of a forthcoming collapse of the Atlantic meridional overturning circulation, Nature Communications, 14, 4254, https://doi.org/10.1038/s41467-023-39810-w, 2023.

Doucet, A., de Freitas, N., and Gordon, N., eds.: Sequential Monte Carlo Methods in Practice, Springer, 2001.

Drijfhout, S., Angevaare, J. R., Mecking, J., van Westen, R. M., and Rahmstorf, S.: Shutdown of northern Atlantic overturning after 2100 following deep mixing collapse in CMIP6 projections, Environmental Research Letters, 20, 094 062, 2025.

Feng, Q. Y., Viebahn, J., and Dijkstra, H. A.: Deep Ocean Early Warning Signals of an Atlantic MOC Collapse, Geophysical Research Letters, 41, 6009–6015, https://doi.org/10.1002/2014GL061019, 2014.

Frei, M. and Künsch, H. R.: Bridging the ensemble Kalman and particle filters, Biometrika, 100, 781–800, https://doi.org/10.1093/biomet/ast020, 2013.

Fu, Y., Lozier, M. S., Biló, T. C., Bower, A. S., Cunningham, S. A., Cyr, F., Jong, M. F. d., deYoung, B., Drysdale, L., Fraser, N., Fried, N., Furey, H. H., Han, G., Handmann, P., Holliday, N. P., Holte, J., Inall, M. E., Johns, W. E., Jones, S., Karstensen, J., Li, F., Pacini, A., Pickart, R. S., Rayner, D., Straneo, F., and Yashayaev, I.: Seasonality of the Meridional Overturning Circulation in the subpolar North Atlantic, Communications Earth & Environment, 4, 181, https://doi.org/10.1038/s43247-023-00848-9, 2023.

Held, H. and Kleinen, T.: Detection of Climate System Bifurcations by Degenerate Fingerprinting, Geophysical Research Letters, 31, https://doi.org/10.1029/2004GL020972, 2004.

Hoffman, R. N. and Grassotti, C.: Future observing system simulation experiments, Bulletin of the American Meteorological Society, 97, 1601–1616, 2016.

Jacques-Dumas, V., Westen, R. M. v., and Dijkstra, H. A.: Estimation of AMOC Transition Probabilities Using a Machine Learning–Based Rare-Event Algorithm, Artificial Intelligence for the Earth Systems, 3, https://doi.org/10.1175/aies-d-24-0002.1, 2024.

Kubo, A. and Sawada, Y.: Predictability of climate tipping focusing on internal variability in the Earth system, Geophysical Research Letters, 52, e2024GL113 146, https://doi.org/10.1029/2024GL113146, 2025.

Lenton, T. M.: Early Warning of Climate Tipping Points, Nature Climate Change, 1, 201–209, https://doi.org/10.1038/nclimate1143, 2011.

Lenton, T. M., Milkoreit, M., Willcock, S., Abrams, J. F., Armstrong McKay, D. I., Buxton, J. E., Donges, J. F., Loriani, S., Wunderling, N., Alkemade, F., Barrett, M., Constantino, S., Powell, T., Smith, S. R., Boulton, C. A., Pinho, P., Dijkstra, H. A., Pearce-Kelly, P., Roman-Cuesta, R. M., Dennis, D., et al.: The Global Tipping Points Report 2025, Tech. rep., University of Exeter, Exeter, UK, https://global-tipping-points.org/, full report; edited volume synthesizing tipping points science and governance, 2025.

Li, C. and Born, A.: Coupled atmosphere-ice-ocean dynamics in Dansgaard-Oeschger events, Quaternary Science Reviews, 203, 1–20, https://doi.org/10.1016/j.quascirev.2018.10.031, 2019.

Li, K.-Y. and Liu, W.: Weakened Atlantic Meridional Overturning Circulation causes the historical North Atlantic Warming Hole, Communications Earth & Environment, 6, 416, https://doi.org/10.1038/s43247-025-02403-0, 2025.

Mehra, R. K.: On the identification of variances and adaptive Kalman filtering, IEEE Transactions on Automatic Control, 15, 175–184, https://doi.org/10.1109/TAC.1970.1099422, 1970.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.

Michel, S. L. L., Dijkstra, H. A., Guardamagna, F., Jacques-Dumas, V., Westen, R. M. v., and Heydt, A. S. v. d.: Deep learning based reconstructions of the Atlantic meridional overturning circulation confirm twenty-first century decline, Environmental Research Letters, 20, 064 036, https://doi.org/10.1088/1748-9326/add7f0, 2025.

Moat, B. I., Smeed, D. A., Rayner, D., Johns, W. E., Smith, R., Volkov, D., Elipot, S., Petit, T., Kajtar, J., Baringer, M. O., and Collins, J.: Atlantic Meridional Overturning Circulation observed by the RAPID-MOCHA-WBTS array at 26N from 2004 to 2024 (v2024.1), Tech. rep., British Oceanographic Data Centre, Natural Environment Research Council, https://doi.org/10.5285/3f24651e-2d44-dee3-e063-7086abc0395e, rAPID-MOCHA-WBTS (RAPID Meridional Overturning Circulation and Heatflux Array – Western Boundary Time Series), 2025.

Murphy, A. H.: A new vector partition of the probability score, Journal of Applied Meteorology, 12, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, 1973.

Nummelin, A., Gjermundsen, A., Guo, C., Jong, M. F. d., Karpouzoglou, T., Labastida, R. G. N., Leppänen, S., Milkoreit, M., Rost, S., Ruiz-Angulo, A., and Räisänen, H.: Preparing for a potential crossing of an AMOC tipping point, PLOS Climate, 5, e0000 881, https://doi.org/10.1371/journal.pclm.0000881, 2026.

Portmann, V., Swingedouw, D., Khattab, O., and Chavent, M.: Observational constraints project a ~ 50% AMOC weakening by the end of this century., Science advances, 12, eadx4298, https://doi.org/10.1126/sciadv.adx4298, 2026.

Prive, N. C. and Errico, R. M.: An Introduction to Observing System Simulation Experiments, Tech. rep., NASA Technical Report, 2014.

Singh, R., Quinn, J. D., Reed, P. M., and Keller, K.: Skill (or lack thereof) of data-model fusion techniques to provide an early warning signal for an approaching tipping point, PLOS ONE, 13, e0191 768, https://doi.org/10.1371/journal.pone.0191768, 2018.

Stommel, H.: Thermohaline convection with two stable regimes of flow, Tellus, 13, 224–230, https://doi.org/10.1111/j.2153-3490.1961.tb00079.x, 1961.

van Leeuwen, P. J.: Nonlinear Data Assimilation for High-Dimensional Systems, Physica D: Nonlinear Phenomena, 412, 132 620, 2020.

van Westen, R. M. and Baatsen, M. L. J.: European Temperature Extremes Under Different AMOC Scenarios in the Community Earth System Model, Geophysical Research Letters, 52, e2025GL114 611, https://doi.org/10.1029/2025GL114611, 2025.

van Westen, R. M. and Dijkstra, H. A.: Asymmetry of AMOC hysteresis in a state-of-the-art global climate model, Geophysical Research Letters, 50, e2023GL106 088, https://doi.org/10.1029/2023GL106088, 2023.

van Westen, R. M., Kliphuis, M., and Dijkstra, H. A.: Physics-Based Early Warning Signal Shows That AMOC Is on Tipping Course, Science Advances, 10, eadk1189, https://doi.org/10.1126/sciadv.adk1189, 2024.

van Westen, R. M., Vanderborght, E., Kliphuis, M., and Dijkstra, H. A.: Physics-based indicators for the onset of an AMOC collapse under climate change, Journal of Geophysical Research: Oceans, 130, e2025JC022 651, 2025.

Van Westen, R. M., Börner, R., and Dijkstra, H. A.: Failure to track a stable AMOC state under rapid climate change, Nature Climate Change, pp. https://doi.org/10.1038/s41 558–026–02 730–w, 2026.

van Westen, R. M. et al.: Physics-based early warning signals for the Atlantic Meridional Overturning Circulation, Earth System Dynamics, 2024.

Vanderborght, E., van Westen, R. M., and Dijkstra, H. A.: Feedback Processes Causing an AMOC Collapse in the Community Earth System Model, Journal of Climate, 38, 5083–5102, https://doi.org/10.1175/JCLI-D-24-0570.1, 2025.

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, https://doi.org/10.1029/2019JC015083, 2019.

https://doi.org/10.5194/egusphere-2026-5755
Preprint. Discussion started: 3 November 2026
© Author(s) 2026. CC BY 4.0 License.



Weijer, W., Cheng, W., Garuba, O. A., Hu, A., and Nadiga, B. T.: CMIP6 Models Predict Significant 21st Century Decline of the Atlantic Meridional Overturning Circulation, Geophysical Research Letters, 47, e2019GL086 075, https://doi.org/10.1029/2019GL086075, 2020.

Xing, Q., Elipot, S., Johns, W. E., Smeed, D. A., Moat, B. I., and Loder, J. W.: Meridionally consistent decline in the observed western boundary contribution to the Atlantic Meridional Overturning Circulation, Science Advances, 12, eadz7738, https://doi.org/10.1126/sciadv.adz7738, 2026.