Supporting Information for "pypfda v1.0: An open-source, model-agnostic particle-filter engine for online sequential paleo data assimilation in Earth system models"

Bijan Fallah[1], Hugues Goosse[2], Willem Huiskamp[1]

[1]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany

[2]Université catholique de Louvain, Earth and Life Institute, Louvain-la-Neuve, Belgium

Correspondence: Bijan Fallah (fallah@pik-potsdam.de)

Contents

1. S1. AMOC–SST coupling check against SODA reanalysis
2. S2. AMOC autocorrelation and e-folding timescale
3. S3. Effective degrees of freedom and significance testing
4. S4. Idealized fast–slow Lorenz benchmark
5. S5. Covariance and causal analysis of the SST–AMOC coupling
6. S6. Additional main-text figures and tables
7. S7. Computational cost of the OSSE campaign
8. S8. Initial-ensemble perturbation and proxy-site mapping
9. S9. Particle-filter degeneracy: reliability and genealogy diagnostics
10. S10. Filter operators: likelihood tempering, the weight cap and adaptive resampling

S1. AMOC–SST coupling check against SODA reanalysis

As a consistency check we compared the CM2Mc-BLING AMOC–SST relationship with the Simple Ocean Data Assimilation (SODA) 2.2.4 ocean reanalysis (Carton and Giese, 2008; Giese and Ray, 2011). Both show a positive subpolar-gyre AMOC–SST correlation (model AMOC at 26.5°N, $r = 0.44$; SODA AMOC at 45°N, $r = 0.47$; subpolar-gyre-averaged spatial $r = 0.45$ in both; Figures S1 and S2). Both also show the expected fingerprint, subpolar warming under stronger overturning (Zhang et al., 2019), together with a localized negative correlation in the Labrador Sea. SODA is itself a model-based reconstruction, so this tests coupling structure, not observational truth.

S2. AMOC autocorrelation and e-folding timescale

The e-folding timescale of the AMOC autocorrelation function is approximately 3 years, about a quarter of the dominant 13-year period (Figure S3).

S3. Effective degrees of freedom and significance testing

Table S1 reports effective degrees of freedom (N_{eff}), critical correlations (r_{crit}), and significance for all AMOC and SST correlations, following Bretherton et al. (1999):

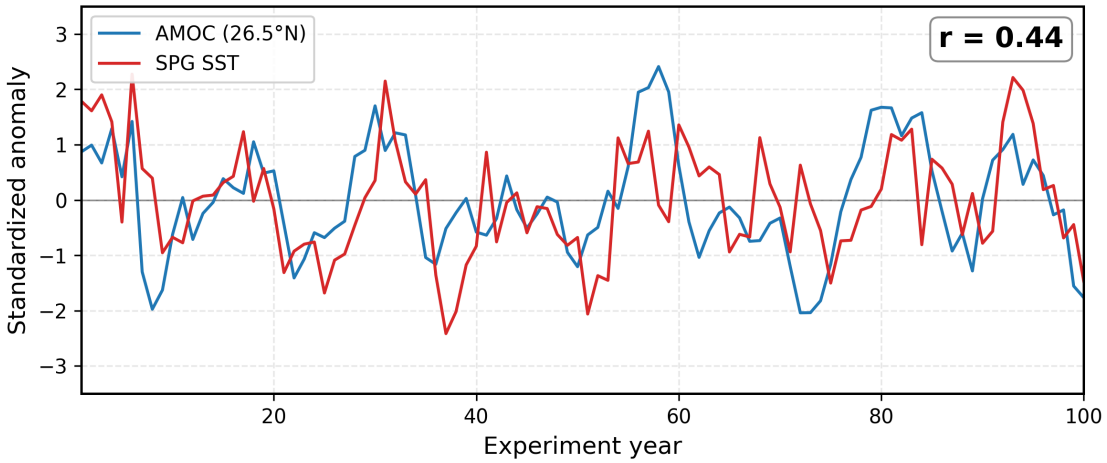
$$N_{\text{eff}} = N_t \cdot \frac{1 - \rho_1 \rho_1'}{1 + \rho_1 \rho_1'} \tag{S1}$$

where N_t is the record length in annual samples and ρ_1, ρ_1' are the lag-1 autocorrelations of the two series being correlated. A correlation between two strongly red series (large $\rho_1 \rho_1'$) therefore carries few effective degrees of freedom. Here $N_t = 100$ for the century experiments and $N_t = 300$ for the long integration of the main text. This two-series form, rather than the single-series $N(1 - \rho_1)/(1 + \rho_1)$, is the appropriate estimator for the significance of a correlation, because the latter is the effective sample size for the mean of one series.

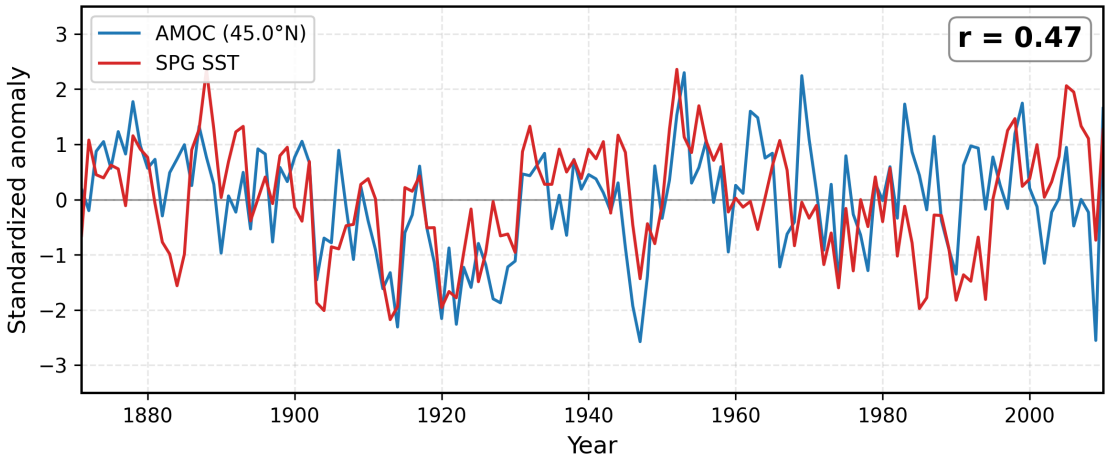
Pattern-level significance. Beyond the individual correlations of Table S1, we also test the systematic FREE→DA improvement as a pattern, using the correlation shift $\Delta r = r_{\text{DA,5yr}} - r_{\text{FREE}}$ as the statistic. A counted surrogate p-value is bounded below by the reciprocal of the number of *distinct* realizations the null admits, not by the number of draws taken from it, so we state that support for each null.

(a) *Exhaustive circular shift.* The truth trajectory is shifted circularly through all $n-1 = 99$ non-trivial lags, which preserves its autocorrelation exactly; the null is enumerated rather than sampled, and its smallest attainable value is $p = 1/n$. One of the 99 surrogates reaches the observed $\Delta r = +0.68$ (r shifts from -0.14 to $+0.54$), giving $p = (1 + 1)/(99 + 1) = 0.02$ (Phipson and Smyth, 2010).

(b) *Fourier phase randomization.* The truth's power spectrum, and hence its full autocorrelation function including the decadal (12–14-year) spectral peak, is preserved while its phases are randomized. The support is continuous, and 1,516 of 100,000 surrogates reach the observed shift, giving $p = (1516 + 1)/(100000 + 1) = 0.015$, in agreement with (a).

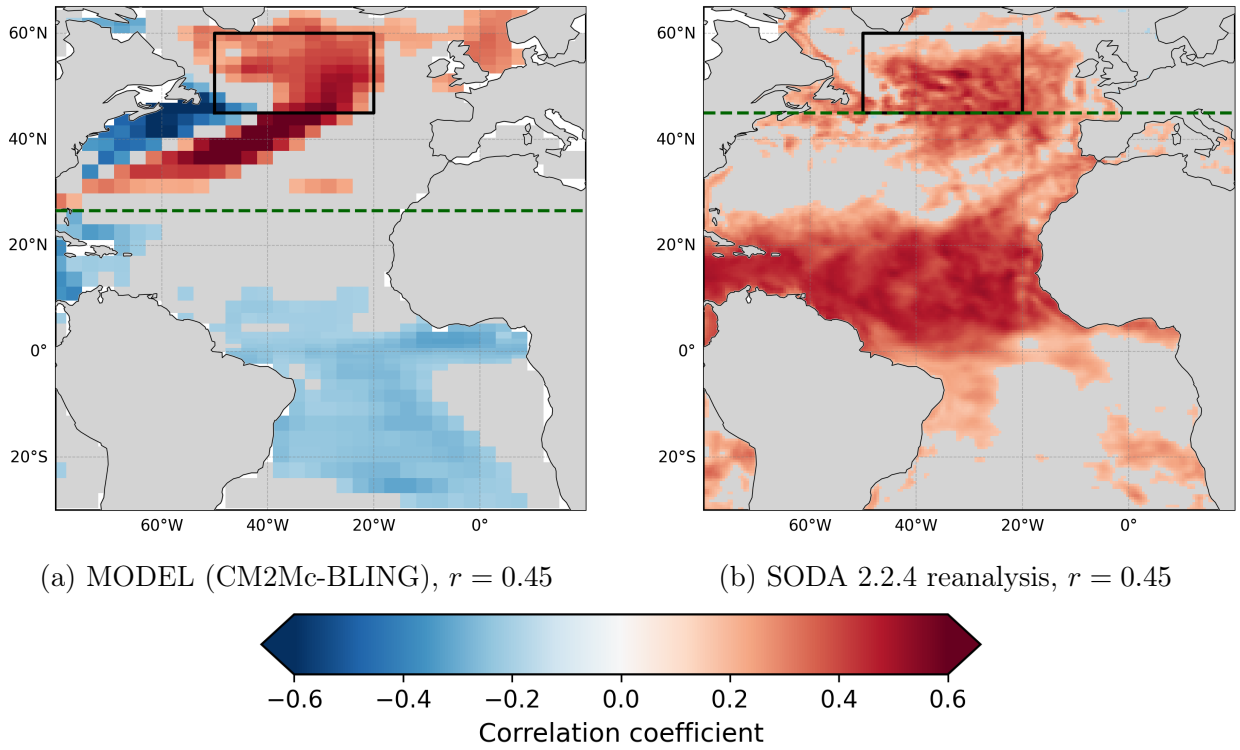


(a) MODEL (CM2Mc-BLING)



(b) SODA 2.2.4 reanalysis

Figure S1: Standardized time series of AMOC strength (blue) and subpolar gyre SST (red) for (a) the CM2Mc-BLING model (experiment years 1–100, AMOC at 26.5°N) and (b) SODA 2.2.4 reanalysis (1871–2010, AMOC at 45°N). Both time series are linearly detrended and normalized to unit variance. Correlation: $r = 0.44$ (model) and $r = 0.47$ (reanalysis). Note that SODA reconstructs AMOC through its ocean model dynamics and is not a direct observational product.



(a) MODEL (CM2Mc-BLING), $r = 0.45$

(b) SODA 2.2.4 reanalysis, $r = 0.45$

Figure S2: Spatial correlation between AMOC and Atlantic SST for (a) the CM2Mc-BLING model (experiment years 1–100, AMOC at 26.5°N) and (b) SODA 2.2.4 reanalysis (1871–2010, AMOC at 45°N). Only statistically significant correlations ($p < 0.05$) are shown; gray areas are non-significant. The black box marks the subpolar gyre region (45–60°N, 50–20°W, a box distinct from the 50–65°N, 60–10°W subpolar-gyre index of the main text) and the green dashed line the latitude at which the AMOC index is taken (26.5°N for the model, 45°N for SODA). The box-averaged spatial correlation shown here ($r = 0.45$ for both) is a different metric from the time-series correlation in Figure S1 ($r = 0.44$ model, 0.47 SODA): the former averages the local AMOC–SST correlation over the subpolar-gyre box, whereas the latter correlates the two area-mean index time series.

Table S1: Effective degrees of freedom and significance testing for the principal reported correlations. $\rho_1\rho_1'$: product of the two series' lag-1 autocorrelations; N_{eff}: effective sample size (Bretherton et al., 1999); r_{crit}: two-tailed critical correlation at $p = 0.05$. Under similar initial conditions every experiment, including the free-running ensemble, is significantly correlated with the truth, because the members begin near the truth trajectory; that is why the similar-IC skill *gain* over FREE is negligible (Table 2 of the main text). The diverse-IC configuration is the more demanding test: there the free ensemble carries no skill, and every assimilated correlation except the ten-year one reaches significance. The AMOC entries are computed in the forecast convention (Sect. 2.3 of the main text).

Experiment	r	$\rho_1\rho_1'$	N_{eff}	r_{crit}	p	**Sig.?**
AMOC at 26.5°N vs truth						
Similar IC: DA (1-yr)	0.52	0.67	19.8	0.45	0.019	Yes
Similar IC: DA (5-yr)	0.69	0.68	19.2	0.45	0.001	Yes
Similar IC: DA (10-yr)	0.53	0.70	17.8	0.47	0.024	Yes
Similar IC: FREE	0.52	0.69	18.2	0.47	0.026	Yes
Diverse IC: DA (1-yr)	0.50	0.67	19.6	0.45	0.025	Yes
Diverse IC: DA (5-yr)	0.54	0.67	20.0	0.44	0.014	Yes
Diverse IC: DA (5-yr, Atlantic-34)	0.51	0.67	19.9	0.44	0.021	Yes
Diverse IC: DA (5-yr, +ocean)	0.62	0.68	19.1	0.45	0.004	Yes
Diverse IC: DA (10-yr)	0.21	0.71	17.0	0.48	0.425	No
Diverse IC: FREE	−0.14	0.71	16.7	0.49	0.587	No
SST_{SPG} *vs truth*						
Similar IC: DA (1-yr)	0.65	0.44	39.2	0.32	<0.001	Yes
Similar IC: DA (5-yr)	0.67	0.44	38.5	0.32	<0.001	Yes
Similar IC: FREE	0.56	0.45	37.7	0.32	<0.001	Yes
Diverse IC: DA (1-yr)	0.52	0.42	40.9	0.31	0.001	Yes
Diverse IC: DA (5-yr)	0.46	0.41	41.4	0.31	0.002	Yes
Diverse IC: FREE	0.18	0.44	39.2	0.32	0.272	No

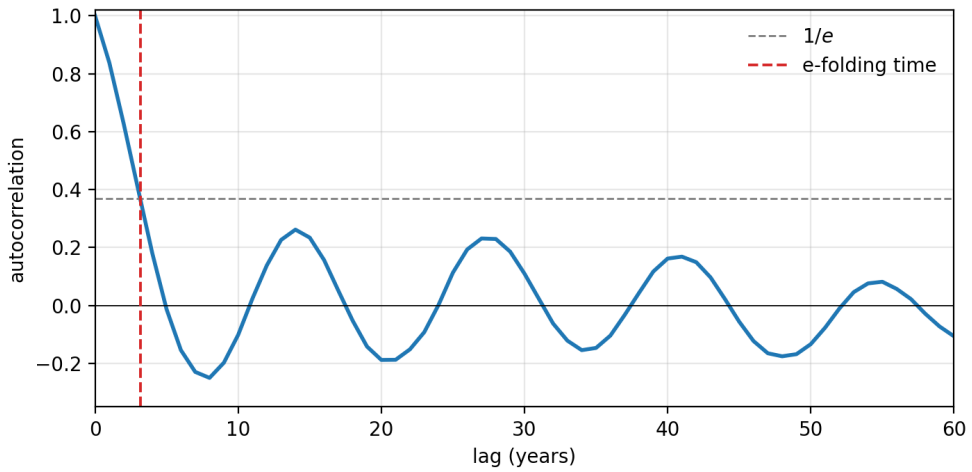


Figure S3: Autocorrelation function of AMOC at 26.5°N from the last 1500 years of the equilibrated control simulation (detrended). The horizontal dashed line marks $1/e$ and the vertical dashed line the e-folding timescale of approximately 3 years, about a quarter of the dominant AMOC period.

(c) *Block permutation* (sensitivity only). Randomly reordering contiguous 10-year segments (support $10! = 3.6 \times 10^6$) gives $p = 0.0016$. This null does not preserve structure longer than the block length and is therefore anti-conservative for a series with a decadal spectral peak, so we do not report it as the headline value.

A moving-block bootstrap places the same gain at $\Delta r = +0.68$ with a 95% confidence interval of $[+0.40, +0.94]$, excluding zero. Carrying the per-experiment p-values from (b) into the Benjamini–Hochberg family across the eight campaign experiments (main text), three survive at $q < 0.05$: the ocean-inflation experiment ($q = 0.009$, $\Delta r = +0.77$), the Atlantic-34 experiment ($q = 0.02$, $+0.66$) and the global-network headline ($q = 0.04$, $+0.68$). Five of the eight experiments have bootstrap intervals excluding zero: those three, the diverse-IC annual experiment ($[+0.35, +0.91]$) and, marginally, the similar-IC five-year experiment ($[+0.02, +0.26]$). The 100-year campaign therefore ranks the configurations consistently but is not powered to certify any one of them. That is instead the role of the 300-year integration (Sect. 3.4 of the main text), which tests the selected configuration on a second truth, with ensembles re-branched from the same control years as the campaign.

The 300-year integration. In the 300-year integration of the main text (Sect. 3.4), assimilation recovers the overturning at $r = +0.78$ (95% CI $[+0.72, +0.83]$; $\Delta r = +0.74$ computed from the unrounded correlations, 95% CI $[+0.59, +0.90]$; moving-block bootstrap, 20,000 resamples with 10-year blocks). The result does not depend on the effective-sample-size estimate. The lag-1 form of the two-series estimator gives $N_{\text{eff}} \approx 52$ here; the full-autocorrelation form of the same estimator, summed to the first zero crossing of the autocorrelation product, gives $N_{\text{eff}} \approx 96$, so the lag-1 value is the conservative one and is the one quoted in the main text. The surrogate nulls of Sect. 2.8 of the main text give $p = 1 \times 10^{-5}$ (phase randomization, the smallest value 100,000 surrogates can attain) and $p = 0.003$ (exhaustive circular

shift) for the DA correlation over the full record, and $p = 1 \times 10^{-5}$ and 0.005 over years 101–300. For the gain over the free ensemble, $\Delta r = +0.74$ over the full record and +0.81 over years 101–300, the same nulls give $p = 0.002$ and $p = 0.0004$, so the overlap-free gain is resolved by the surrogate test as well as by its bootstrap interval. MSSS stays below the mean-square ceiling of $r = +0.78$, which is $r^2 = 0.61$, because the ensemble mean's amplitude is uncalibrated. The mean retains 64% of the truth's interannual standard deviation where the mean-square-optimal value is 78% (44% versus 54% at 100 years), so it is damped rather than over-amplified. Rescaling its anomalies by the regression slope of the truth on it (1.22 at both record lengths) would therefore recover the full r^2 without changing r. We report the unscaled mean throughout, because the filter is never shown the target it would be calibrated against. This amplitude property is distinct from the dispersion of the ensemble spread about the mean (Section S9). Three cycle-end skill values of this integration differ only by convention. The value $r = +0.82$ compares the forecast ensemble mean averaged over each five-year cycle with the five-year-mean truth at the 60 observation times, which is the like-for-like quantity of the fingerprint benchmark. The value $r = +0.77$ compares the forecast ensemble mean in the single cycle-end year with the annual truth in that year. The value $r = +0.80$ is the post-resampling analysis in that same year. All headline values in this paper use the forecast convention over the full annual record. The three conventions agree to within 0.05 in r, and the five-year-mean convention, the one the fingerprint benchmark requires, is the most favorable of them.

S4. Idealized fast–slow Lorenz benchmark

Because the engine is model-agnostic, the cleanest test is whether the method, rather than some feature of a specific ocean model, recovers a slow variable from fast observations. That test runs the identical `pypfda` cycle on a chaotic system with an exact fast–slow structure that is cheap to run on a laptop. We use the two-level Lorenz-96 model (Lorenz, 1996), the two-scale extension of the single-level system of Lorenz and Emanuel (1998). The system consists of $K = 8$ slow, large-scale variables X_k, each coupled to $J = 10$ fast variables $Y_{j,k}$. The evolution equations are

$$\dot{X}_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \tfrac{hc}{b} \sum_j Y_{j,k}, \tag{S2}$$

$$\dot{Y}_{j,k} = -cb\, Y_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - c\, Y_{j,k} + \tfrac{hc}{b} X_k, \tag{S3}$$

with parameters $F = 10$, $h = 1$, a fast-subsystem time-scale factor $c = 10$, and amplitude reduction factor $b = 10$. The fast variables thus evolve ten times faster and are ten times smaller in amplitude than the slow variables.

We set up a twin OSSE that observes every second fast variable (40 of the 80, with `obs_stride`=2) and reconstructs the slow large-scale index $\bar{X} = K^{-1} \sum_k X_k$. The experiment uses a 100-member ensemble, diverse initial conditions, and the same operators as the GCM. Those operators are the tempered likelihood with parameter η, the max-weight cap with $w_{\max} = 0.3$, and the post-resample inflation kernel described in the main text. The forecast convention is the same as for the ocean model. Figure S4 shows the outcome.

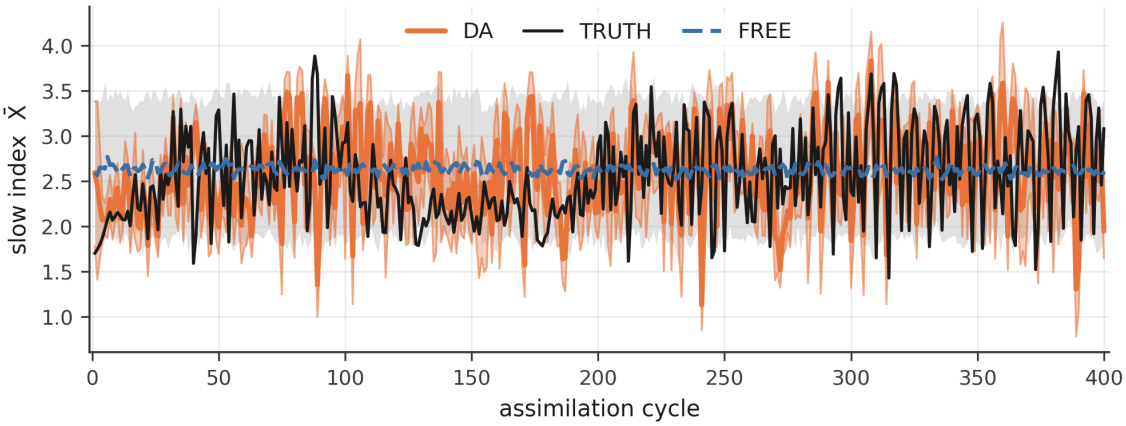


Figure S4: The same `pypfda` engine on the two-scale Lorenz-96 system, run as a twin OSSE that observes only the fast ring and reconstructs the slow large-scale index $\bar{X}$. The truth is shown in black (thin line), the FREE ensemble mean as a blue dashed line, and the DA ensemble mean as an orange line, each accompanied by its 5–95% ensemble band.

This benchmark shows that the engine and its operators transfer unchanged to an idealized model. On the two-scale Lorenz-96 system, the filter raises the slow-index correlation from $r = -0.11$ (FREE) to $r = +0.22$ (DA), a gain of $\Delta r = +0.33$. The genealogy, however, collapses to a single surviving founder in nearly every cycle. Because the FREE baseline is by construction near zero, the absolute assimilated skill is modest. The value of this benchmark is therefore not high skill but that identical operators transfer to a system with exact fast–slow structure and reproduce the genealogical collapse. The proximate causes of that low skill are in turn the sparse fast-ring sampling and the resampling threshold ($\tau_{\text{ess}} = 0.5$). That degeneracy is the problem the ocean-inflation experiment of the main text is designed to mitigate, here in its starkest, perfect-model form. The package also ships a two-timescale coupled Lorenz-63 example in the spirit of Peña and Kalnay (2004), shown in Figure S5. There, observing only the fast "atmosphere" and reconstructing the persistent slow "ocean" variable, the identical filter identifies the occupied attractor lobe and tracks it. The filter cuts the slow-variable reconstruction RMSE by 97% (8.63 to 0.29) and raises the correlation from $r = +0.52$ (FREE) to $+0.83$ (DA). Because the slow variable is strongly persistent, the RMSE reduction is the more meaningful skill measure here. Both examples run on a laptop as the package scripts `08_two_scale_l96_fast_slow.py` and `09_coupled_l63_fast_slow.py`.

Sensitivity to the tempering factor and the resampling threshold. Because a full 100-member coupled Lorenz-63 run costs seconds, we swept the filter's two free parameters, the observation-error tempering factor η and the ESS resampling threshold τ_{ess}, across the grid of Table S2. Reconstruction skill is insensitive to both parameters wherever the filter actually resamples ($\Delta r \approx +1.7$, RMSE reduced by 94–96%), and the choice of η governs how much genealogical diversity survives. The ocean model's working values ($\eta = 4$, $\tau_{\text{ess}} = 0.85$) therefore do not sit at a sharp optimum; they lie inside a flat region.

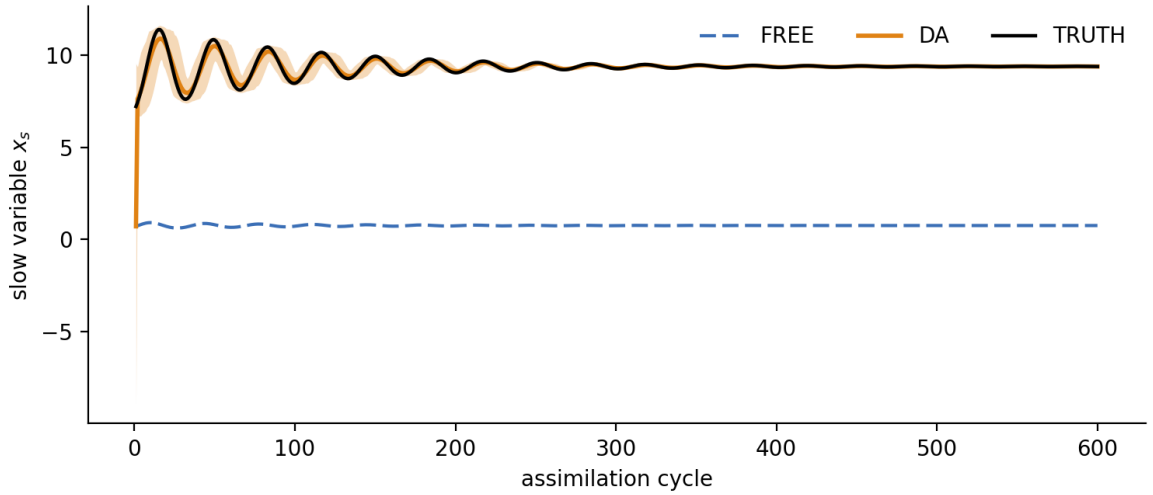


Figure S5: The same `pypfda` engine on the coupled fast–slow Lorenz-63 system (Peña and Kalnay, 2004) (package example `09_coupled_l63_fast_slow.py`): a twin OSSE observing only the fast "atmosphere" and reconstructing the slow, persistent "ocean" variable x_s. The truth is black, the FREE ensemble mean blue dashed, and the DA ensemble mean orange with its 5–95% band. The diverse-IC free ensemble averages the attractor lobes toward zero, while the filter identifies and tracks the occupied lobe. This is the well-performing idealized configuration that complements the deliberately hard two-scale Lorenz-96 case of Figure S4.

Table S2: Sensitivity of the coupled Lorenz-63 twin OSSE to the observation-error tempering factor η (rows) and the ESS resampling threshold τ_{ess} (columns). Each cell gives $\Delta r = r_{\text{DA}} - r_{\text{FREE}}$ and, in parentheses, the number of surviving founders (the effective ancestor size) at the end of the 400-cycle, 100-member run. The free-running baseline of this realization is $r_{\text{FREE}} = -0.95$, and the assimilated correlation itself ranges from $r_{\text{DA}} = +0.71$ to $+0.88$ wherever the filter resamples; the FREE correlation of a single idealized Lorenz realization is strongly seed-dependent, so each experiment in this section reports its own baseline. Skill is high and nearly flat wherever the filter resamples. Two failure modes bracket this interior. With strong tempering and a low threshold ($\eta \geq 4$, $\tau_{\text{ess}} = 0.5$) the weights stay so uniform that the effective sample size never drops below the threshold, the filter never resamples, and, because weights are not carried between cycles, the analysis equals the free run ($\Delta r = 0$, no genealogical contraction). With weak tempering ($\eta = 1$) the genealogy collapses to a single lineage. Between the limits, a larger η retains markedly more diversity at a small skill cost.

	$\tau_{\text{ess}} = 0.50$	0.70	0.85	0.95
$\eta = 1$	+1.83 (6)	+1.74 (1)	+1.75 (1)	+1.75 (1)
$\eta = 2$	+1.76 (45)	+1.76 (45)	+1.76 (45)	+1.79 (1)
$\eta = 4$	0.00 (100)	+1.70 (50)	+1.70 (50)	+1.70 (50)
$\eta = 8$	0.00 (100)	+1.66 (52)	+1.66 (52)	+1.66 (52)

A nonlinear observation operator. The margin by which the online filter exceeds a linear SST–AMOC fingerprint on the AMOC index (main text) is itself measured under a *linear* observation operator, which is the case most favorable to the fingerprint. Real proxies respond nonlinearly to their environment (saturation, thresholds), and in that regime a fixed linear map cannot represent the observation–target relationship, whereas the particle filter can, because it forward-models the operator. We test this on the coupled Lorenz-63 twin OSSE (Section S4) by replacing the identity observation with nonlinear operators h applied to the observed fast subsystem. Both the filter and a *fair* linear regression, calibrated on the free ensemble alone and never on the truth, then reconstruct the slow target from the *same* standardized observation series (Table S3). The free ensemble supplies only the regression's mapping and carries no target information of its own (its baseline is $r = -0.41$), so both estimators draw their skill from the observations alone. We report two symmetric configurations. The noise-free observations isolate the effect of h, because it is then the only thing that differs between rows. Given those observations, the filter recovers the target at $r = +0.81$ under a saturating operator against the regression's $r = +0.40$, and it leads under the identity ($+0.81$ versus $+0.53$) and quadratic ($+0.83$ versus $+0.73$) operators. Under a smoothly monotonic cubic operator, by contrast, the two are indistinguishable ($+0.82$ versus $+0.85$), because a multivariate linear fit approximates that operator well. Adding observation noise to both estimators widens every one of those margins. The reason is that the static linear reconstruction falls to $r \leq +0.13$ under all five operators, even when granted ridge regularization with a blocked cross-validated penalty and three cycles of lagged predictors. That collapse occurs because a fixed map cannot average noise over time, whereas the filter propagates each analysis through the dynamics and is nearly unchanged ($+0.82$ under saturation). The one operator that defeats the filter is the non-invertible sign-folding $h = |v|$. That is because two opposite fast states yield the same observation, so the likelihood is bimodal and the filter selects the wrong sign about as often as the right one. The filter therefore ends near its own free-running baseline ($r = -0.09$ noise-free, -0.18 noisy), while a linear fit still reads the amplitude ($+0.40$ noise-free). With noise this operator defeats both (-0.18 and -0.08). What the filter requires is therefore that h be *invertible*, not that it be linear. The saturating case widens the particle filter's advantage where a linear fingerprint breaks down.

Table S4 lists the online particle-filter configuration used for CM2Mc-BLING and the idealized testbeds. The observation-error tempering factor η is the main parameter retuned per model. CM2Mc-BLING inflates the likelihood width fourfold ($\eta = 4.0$) to delay high-dimensional weight collapse. Two further settings differ, the maximum-weight cap and the post-resampling inflation, which perturbs atmospheric temperature in CM2Mc-BLING. All remaining settings are shared defaults.

Table S3: Reconstruction skill (correlation with the truth slow variable) of the particle filter (PF) versus a fair linear regression on the coupled Lorenz-63 twin OSSE, under different observation operators h applied to the fast subsystem, in two symmetric configurations. In the first both estimators are given the identical noise-free observations, in the second the identical noisy observations. The regression is calibrated on the free ensemble alone (never the truth); in the noisy configuration it is a ridge fit whose penalty is chosen by blocked cross-validation and whose predictors include three cycles of lags, so that its loss cannot be attributed to the estimator. Values are means over three seeds. The free-running baseline is $r = -0.41$. Bold marks the saturating operator, the most physically realistic proxy nonlinearity.

	noise-free observations		noisy observations	
Observation operator $h(v)$	PF r	Regression r	PF r	Regression r
identity, v	$+0.81$	$+0.53$	$+0.80$	$+0.08$
saturating, $8\tanh(v/8)$	$\mathbf{+0.81}$	$\mathbf{+0.40}$	$\mathbf{+0.82}$	$\mathbf{+0.06}$
cubic, $\mathrm{sgn}(v)\,\lvert v/6\rvert^3 \cdot 6$	$+0.82$	$+0.85$	$+0.80$	$+0.13$
sign-folding, $\lvert v\rvert$	-0.09	$+0.40$	-0.18	-0.08
quadratic, $v\,\lvert v\rvert/8$	$+0.83$	$+0.73$	$+0.80$	$+0.10$

Table S4: Per-model online particle-filter configuration. η is the observation-error tempering factor ($\sigma_{\mathrm{eff}} = \eta\,\sigma_{\mathrm{obs}}$) and is the main parameter retuned per model, alongside the maximum-weight cap and the post-resampling inflation, whose perturbed field and amplitude are also per model. The ocean model uses $\tau_{\mathrm{ess}} = 0.85$ and $\sigma_{\mathrm{obs}} = 0.3$°C, while the idealized testbeds use values matched to their own scaling; the ensemble size $N = 100$ is common to all three. τ_{ess} is the ESS resampling threshold (fraction of N) and $w_{\max}$ the maximum-weight cap. The CM2Mc-BLING cycle is five model years (the headline configuration; the campaign also ran annual and ten-year cycles); the Lorenz cycles are in model time units (MTU). CM2Mc-BLING assimilates a global marine-SST pseudo-proxy network by snapping each site to the nearest ocean grid cell; the Lorenz testbeds observe the fast subsystem directly.

Model / testbed	η	σ_{obs}	τ_{ess}	$w_{\max}$	**Cycle**	N	**Post-resampling inflation**
CM2Mc-BLING	4.0	0.3°C	0.85	—	5 yr	100	atmospheric temperature ($\sigma_T = 0.2$ K, levels 6–18); +ocean experiment 0.01°C / 0.003 psu
Two-scale L96	2.0	0.1	0.5	0.3	0.2 MTU	100	0.3
Coupled L63	2.0	0.5	0.5	0.3	0.15 MTU	100	0.4

S5. Covariance and causal analysis of the SST–AMOC coupling

Methods for the covariance and causal-inference analyses

Data assimilation modifies the dynamical relationships between variables, not just the reconstructed states. We use three causal analysis approaches to characterize the SST–AMOC coupling and test whether it is preserved after assimilation. A particle filter is in principle well suited to preserving such couplings, because its update never forms linear combinations of states, so it cannot create the dynamically inconsistent states that a Gaussian update can (van Leeuwen et al., 2019).

Correlation-Based Physical Mechanism Analysis

We establish the expected coupling strength under unforced variability using the last 1500 years (model years 1973–3472) of the corrected control simulation. The hypothesized causal chain is

$$\mathrm{AMOC} \xrightarrow{\text{heat transport}} \mathrm{OHC_{SPG}} \xrightarrow{\text{surface flux}} \mathrm{SST_{SPG}} \tag{S4}$$

where SPG denotes the subpolar gyre region (50–65°N, 60–10°W), $\mathrm{OHC_{SPG}}$ is ocean heat content integrated over the upper 700 m in the SPG, and $\mathrm{SST_{SPG}}$ is the SPG-averaged sea surface temperature. The correlation above measures the forward (AMOC-to-surface) direction; the reverse density feedback, in which cooler SST deepens the mixed layer and convection and strengthens the overturning, is the complementary direction quantified in Section S5 (Directional Causal Coupling).

We quantify coupling strength using the Pearson correlation coefficient computed over non-overlapping 50-year segments:

$$r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{t=1}^{50}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{50}(X_t - \bar{X})^2}\sqrt{\sum_{t=1}^{50}(Y_t - \bar{Y})^2}} \tag{S5}$$

where X and Y are detrended annual time series of the relevant variables, $\bar{X}$ and $\bar{Y}$ are their 50-year means, and σ_X and σ_Y are their standard deviations. This 1500-year window provides 30 independent 50-year segments, which gives a reference distribution of coupling strengths for comparison with the DA and FREE ensembles. These segment correlations are computed at zero lag; lags between the variables are treated separately by the directional methods (Section S5, Directional Causal Coupling).

Maximum Covariance Analysis (MCA)

To identify the dominant spatial patterns of coupled SST-AMOC variability, we perform Maximum Covariance Analysis (MCA), also known as Singular Value Decomposition (SVD)

of the cross-covariance matrix (Björnsson and Venegas, 1997).

Let $\mathbf{X} \in \mathbb{R}^{n_t \times n_s}$ be the detrended, standardized SST anomaly matrix (rows = time, columns = spatial grid points). Let $\mathbf{Y} \in \mathbb{R}^{n_t \times n_{s2}}$ be the detrended, standardized anomaly matrix of the Atlantic meridional overturning streamfunction (columns = latitude–depth grid points of the overturning cell). The cross-covariance matrix is

$$\mathbf{C}_{XY} = \frac{1}{n_t - 1}\mathbf{X}^T\mathbf{Y} \tag{S6}$$

We decompose this matrix using SVD:

$$\mathbf{C}_{XY} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{S7}$$

where:

- $\mathbf{U} \in \mathbb{R}^{n_s \times r}$: Left singular vectors representing spatial patterns of SST variability
- $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$: Diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$
- $\mathbf{V}^T \in \mathbb{R}^{r \times n_{s2}}$: Right singular vectors representing spatial patterns of AMOC streamfunction variability
- $r = \min(n_s, n_{s2})$: Rank of the decomposition

The AMOC target is the two-dimensional (latitude–depth) overturning streamfunction rather than a scalar index, so the decomposition has rank > 1 and admits more than one mode. The squared-covariance fractions below (75.3% for Mode 1, 16.8% for Mode 2) are therefore well defined.

The squared covariance fraction (SCF) explained by mode k is:

$$\text{SCF}_k = \frac{\sigma_k^2}{\sum_{j=1}^{r} \sigma_j^2} \times 100\% \tag{S8}$$

For our analysis, Mode 1 captures 75.3% of the total squared covariance, representing the dominant "cold blob" pattern associated with AMOC variability. In that pattern, a strong AMOC transports heat northward, warming the subpolar gyre, while a weak AMOC produces cooling (Rahmstorf et al., 2015). This mode provides a fingerprint for validating that DA preserves the correct physical relationship between AMOC and North Atlantic SST. The 75.3% explained variance is computed from a single 100-year realization (∼7.5 AMOC cycles) and is likely inflated by the short record. A more stable estimate from the 1500-year control simulation is therefore deferred to future work.

Directional Causal Analysis

Correlations alone cannot distinguish the direction of causality. We therefore apply four complementary methods, ranging from classical bivariate tests to modern causal discovery, to characterize the SST–AMOC coupling. The cross-correlation and Granger causality analyses use the last 1500 years of the corrected control, split into 15 non-overlapping 100-year

segments (matching the OSSE length). The PCMCI and Liang–Kleeman analyses use the separate legacy 3000-year control, split into 30 such segments (see the main text). The segments give the uncertainty estimates. In addition, we compute the Liang–Kleeman information flow member-by-member on the OSSE truth, FREE and DA trajectories themselves, to test whether assimilation alters the inferred SST–AMOC causal relationship relative to the free-running ensemble (Section S5, Directional Causal Coupling).

The cross-correlation function provides the lag-dependent Pearson correlation between SPG SST and the AMOC index, computed over the full 1500-year record. The correlation at lag ℓ years is

$$r(\ell) = \frac{\sum_t (X_t - \bar{X})(Y_{t+\ell} - \bar{Y})}{\sigma_X \sigma_Y}, \tag{S9}$$

where positive lags indicate that SST leads AMOC and negative lags that AMOC leads SST. Because both series are strongly autocorrelated, the 95% significance threshold is set from the autocorrelation-corrected effective sample size (Section S3), $|r| = 1.96/\sqrt{N_{\text{eff}}}$ rather than $1.96/\sqrt{N_t}$ with N_t the raw record length.

Granger causality (Granger, 1969) tests whether past values of a source variable X improve the prediction of a target variable Y beyond what can be predicted from Y's own past alone. For each direction (SST $\to$ AMOC and AMOC $\to$ SST) and for each lag order $p = 1, \ldots, 15$, we fit two models. The restricted model uses only lagged values of Y, while the unrestricted model additionally includes lagged values of X. The null hypothesis that all coefficients of X are zero is tested with an F-statistic

$$F = \frac{(\mathrm{RSS}_r - \mathrm{RSS}_u)/p}{\mathrm{RSS}_u/(n - 2p - 1)}, \tag{S10}$$

where RSS_r and RSS_u are the residual sums of squares of the restricted and unrestricted models, p the maximum lag, and n the sample size. We report the percentage of the 100-year chunks for which the null hypothesis is rejected at the 5% level.

The Liang–Kleeman information flow (LKIF) (Liang, 2014, 2021) quantifies the instantaneous rate of entropy transfer from a source X to a target Y. From a bivariate linear stochastic model, the information flow is

$$T_{X \to Y} = \frac{C_{YY} \cdot C_{X,dY} - C_{XY} \cdot C_{Y,dY}}{C_{XX} C_{YY} - C_{XY}^2} \cdot \frac{C_{XY}}{C_{YY}}, \tag{S11}$$

where $C_{XY} = \mathrm{Cov}(X, Y)$ and $C_{X,dY} = \mathrm{Cov}(X, \mathrm{d}Y/\mathrm{d}t)$, with derivatives approximated by finite differences. The normalized measure

$$\tau_{X \to Y} = \frac{|T_{X \to Y}|}{|T_{X \to Y}| + |dH_Y - T_{X \to Y}|} \tag{S12}$$

lies in $[0, 1]$ and represents the fraction of Y's total entropy change rate that is attributable to X. Here $dH_Y = a_{YY} + Q_Y/(2C_{YY})$ is the total entropy change rate of Y, where a_{YY} is the self-feedback coefficient and Q_Y the noise variance. Both are estimated from the bivariate linear stochastic model $dY/dt = a_{YY}Y + g_Y X + \text{noise}$. To examine lagged causality, we shift one time series by ℓ years before computing τ; this lag shift is our own extension of the instantaneous estimator of Liang (2014).

PCMCI (Runge et al., 2019a,b) overcomes the limitations of bivariate methods, in particular their inability to account for common drivers and autocorrelation. It combines constraint-based structure learning (the PC algorithm) to identify potential parents for each variable at each lag, followed by a momentary conditional independence (MCI) test that conditions on all identified parents. This procedure removes spurious links that arise from indirect pathways or from autocorrelation.

We apply PCMCI with three independence tests of increasing generality. The first is partial correlation (ParCorr), which tests linear conditional independence analytically and serves as the linear baseline. The second is Gaussian process distance correlation (GPDC) (Runge et al., 2019b), which captures nonlinear dependencies by regressing out the conditioning set with Gaussian processes and then testing the residuals with distance correlation; significance is obtained from a numerically pre-computed null distribution. The third is conditional mutual information via k-nearest neighbors (CMIknn) (Runge et al., 2019b), a fully non-parametric, model-free test that estimates conditional mutual information using k-NN statistics ($k = 0.1N$) with significance assessed, in our configuration, via 50 surrogate shuffle tests.

For all tests, the maximum lag is set to 15 years for ParCorr and GPDC, and to 10 years for CMIknn for computational tractability; the PC selection criterion is the Akaike information criterion (AIC). Each test is applied independently to the same 30 non-overlapping 100-year chunks, and we report both the median link strength with 90% confidence intervals and the percentage of chunks yielding significant links ($p < 0.05$). Convergence of the three tests would indicate that the coupling is predominantly linear; divergence would suggest nonlinear causal pathways that a linear cost function would miss.

Physical Coupling Strength

The DA system uses surface SST to constrain the subsurface AMOC, so its success depends on the model reproducing the physical covariances between these variables. In the expected coupling, AMOC modulates subpolar gyre heat content, which in turn affects SST (Zhang et al., 2019). We test whether this coupling is preserved after assimilation.

Figure S6 shows Pearson correlations for three links in this chain, (a) AMOC leading SPG ocean heat content, (b) SPG OHC driving SST, and (c) the net AMOC–SST relationship. The DA ensemble (blue) reproduces the correlations of the truth (red dashed line) and the long control (gray). The FREE ensemble (orange) shows a broader, weaker coupling distribution, which is the expected consequence of phase scatter among independently evolving members.

The diverse ocean IC DA ensemble (green) shows coupling distributions comparable to the similar-IC case, confirming that the AMOC $\rightarrow$ OHC $\rightarrow$ SST chain is a property of the model physics, not of the IC strategy.

These correlations quantify coupling strength but do not establish causal direction. The assumed direction (AMOC $\rightarrow$ OHC $\rightarrow$ SST) follows from ocean dynamics rather than from the correlation analysis, because AMOC transports heat poleward and warms the subpolar gyre from below. We test directionality with information-theoretic methods in the next section.

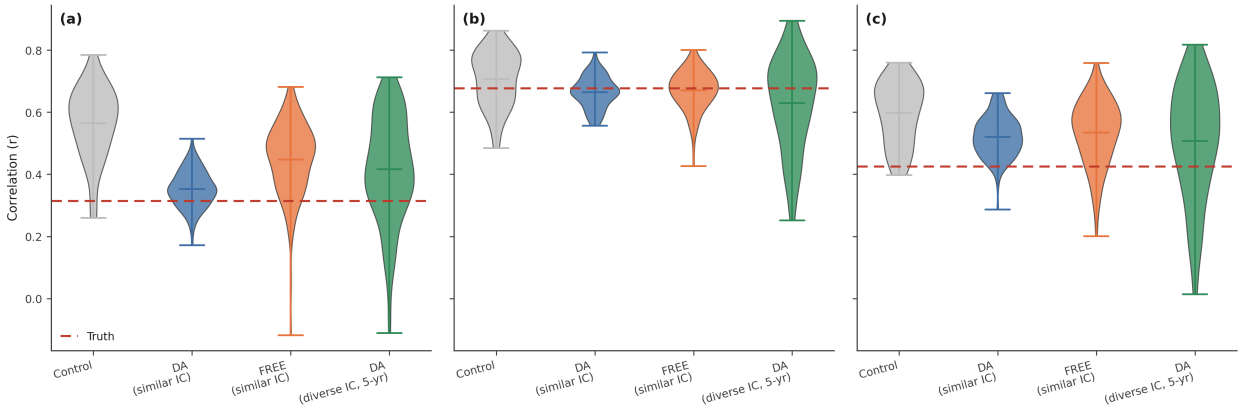


Figure S6: Distribution of Pearson correlation coefficients for the three key links in the proposed mechanism: (a) AMOC strength leading subpolar gyre (SPG) ocean heat content (OHC), (b) SPG OHC and sea surface temperature (SST), and (c) the net AMOC–SST relationship. Distributions (Violins) represent the internal variability within the corrected control simulation (Gray, $n = 30$ non-overlapping 50-year segments over the last 1500 years), the DA ensemble with similar ocean IC (Blue, $n = 100$ members), the FREE ensemble with similar ocean IC (Orange, $n = 100$), and the DA ensemble with diverse ocean IC (Green, $n = 100$). The red dashed line indicates the single realization from the truth simulation.

Directional Causal Coupling

We apply cross-correlation and Granger causality (Granger, 1969) to the last 1500 years of the equilibrated control simulation, split into 15 non-overlapping 100-year segments to obtain uncertainty estimates. The PCMCI causal discovery framework (Runge et al., 2019a,b) and Liang-Kleeman information flow (Liang, 2014, 2021) were applied to a longer (3000-year) control simulation with a different background diffusivity (1.0×10^{-5} m^2 s^{-1}). Those results are therefore qualitatively consistent, but the quantitative values may differ from the corrected configuration.

The causal process graph (Figure S7a) summarizes the bidirectional coupling structure. A *fast heat transport pathway* operates at 1–2 year lags, in which AMOC variability drives SST through meridional heat advection. This pathway is detected with a cross-correlation of $r = +0.53$ at lag 1 and confirmed by Granger causality in 100% of 15 chunks ($p < 0.05$). A *slow density feedback* operates at 5–7 year lags, in which SST anomalies modulate surface buoyancy and deep convection to influence the overturning ($r = -0.44$ at lag 6, Granger significance 87% at lag 5).

Granger causality identifies AMOC → SST in 100% of 15 chunks at lag 1, while SST → AMOC peaks at 87% (lag 5). Both directions are detected, but the AMOC → SST pathway is more consistent (significant at all lags 1–3). PCMCI analysis on a separate 3000-year control run (with background diffusivity 1.0×10^{-5} m^2 s^{-1}, twice the OSSE value) indicated that after conditioning on autocorrelation, direct links are weaker. These PCMCI percentages are approximate; the coupling strengths may differ quantitatively under the corrected diffusivity.

Three points emerge from the multi-method comparison:

1. **Bidirectional coupling on distinct timescales.** All methods identify a fast AMOC

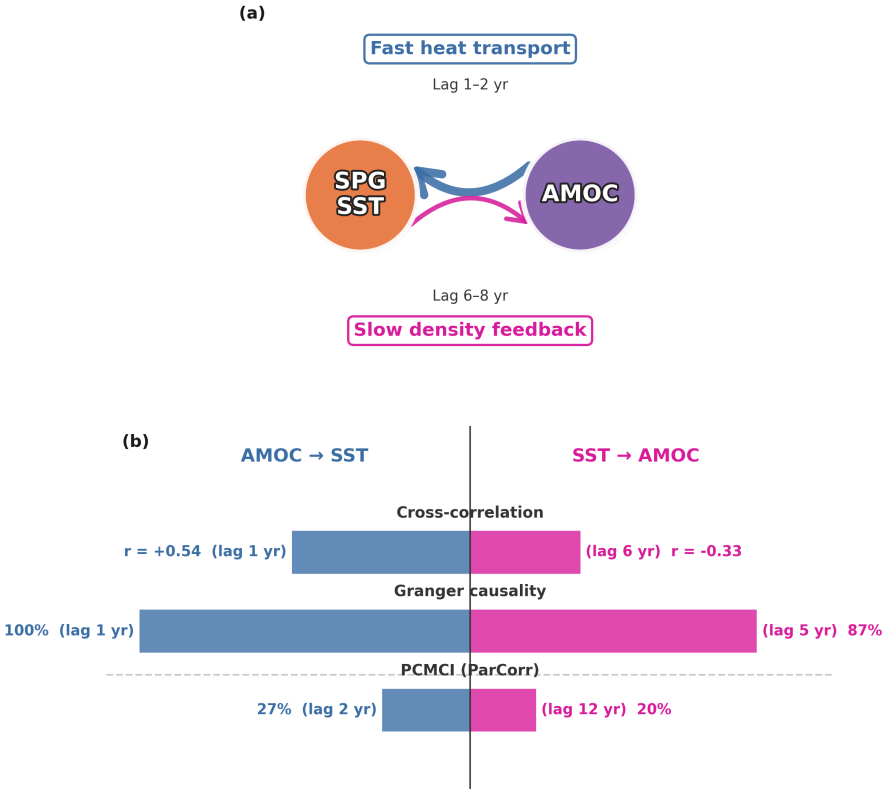


Figure S7: Causal analysis of SST-AMOC coupling. (a) Bidirectional coupling from the 1500-year corrected control: a fast heat transport pathway (AMOC → SST, 1–2 yr lag, thick blue arrow) and a slow density feedback (SST → AMOC, 5–7 yr lag, thinner red arrow). Arrow thickness reflects coupling strength. (b) Detection strength across methods: cross-correlation ($|r|$) and Granger causality (% of 15 chunks significant at $p < 0.05$) from the corrected control; PCMCI-ParCorr (% of 30 chunks significant) from a separate 3000-year control with different diffusivity.

→ SST pathway (∼1–2 years) and a slower SST → AMOC feedback (∼5–7 years).

2. **Conditioning on autocorrelation removes a substantial part of the apparent coupling.** The strong Granger signal (100% of chunks significant at lag 1) is inflated by persistent autocorrelation. After PCMCI conditioning, direct links remain detectable but modest (∼20–27% of chunks). Because the PCMCI test was run on the separate legacy 3000-year control, whereas the Granger test used the corrected control, this reduction is indicative rather than an exact partition. AMOC predictability from SST therefore rests substantially on shared memory as well as on direct forcing.

3. **The SST–AMOC coupling is close to linear in this model.** ParCorr, GPDC, and CMIknn give consistent results, so nonlinear causal pathways between SST and AMOC are weak in the control integrations analyzed here. The Gaussian likelihood used in the particle filter is correspondingly well matched to the identity observation operator of the coupled OSSEs. This concerns the internal dynamics, not the proxy forward model: real proxies respond nonlinearly to their environment, so the observation operator is nonlinear even where the underlying coupling is not, and that is the regime tested in Table S3.

Liang-Kleeman information flow applied to all trajectories (truth, FREE, DA) gives normalized values τ of 0.2–0.8 across members, with FREE and DA distributions that are statistically indistinguishable. The same model physics produces the same causal relationships with or without assimilation. LKIF and Granger causality report different dominant directions because they measure different physical processes. LKIF quantifies the instantaneous rate of entropy transfer. At zero lag, SST anomalies force surface density changes that immediately modulate deep convection, producing a strong SST → AMOC information flow. Granger causality, by contrast, tests lagged predictive power. Over 1–3 years the accumulated AMOC heat transport predicts SST, giving AMOC → SST 100% significance at lag 1 (vs 87% for SST → AMOC at lag 5). The two results are complementary, not contradictory, because the SST-to-AMOC information flow appears at zero lag, while AMOC forces SST with a lag through advective heat transport.

Coupled SST–AMOC fingerprint and AMOC spectrum

The particle filter constrains the AMOC through the model's SST–AMOC fingerprint and must resolve its dominant timescale; both are summarized here.

The leading maximum-covariance mode, the "AMOC fingerprint" (Figure S10), explains 75% of the coupled SST–AMOC squared covariance. Physically, a stronger AMOC transports more heat northward, warming the subpolar gyre, while a weaker AMOC contributes to the "cold blob" south of Greenland (Caesar et al., 2018; Rahmstorf et al., 2015), alongside other mechanisms (Keil et al., 2020). In the mode's loading pattern, however, this contrast appears as a west–east dipole across the northwest Atlantic rather than a basin-uniform subpolar warming.

Spectral analysis of the control AMOC (Figure S8; Welch's method, 200-year windows, 50% overlap) gives a dominant variability period of approximately 13.3 years (Sect. 3.5 of the main text).

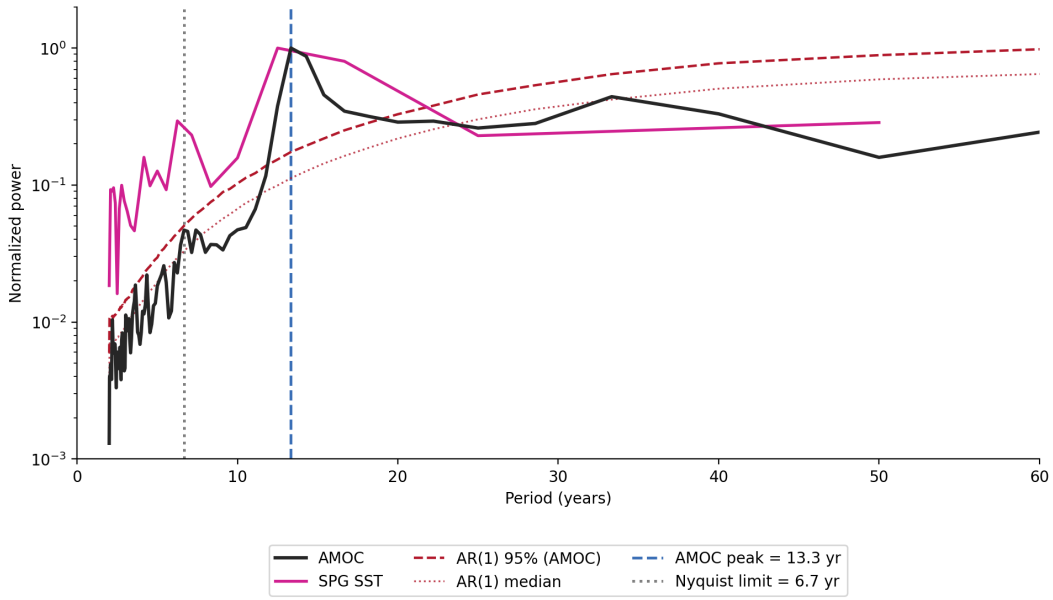


Figure S8: Power spectra of AMOC (black) and subpolar-gyre SST (magenta) from the last 1500 years of the equilibrated control simulation, computed using Welch's method with 200-year windows. Both variables show a dominant period near 13.3 years (blue dashed vertical line, AMOC spectral peak). The red dashed and dotted curves are the 95th-percentile and median spectra of an AR(1) red-noise null fitted to the control AMOC (lag-1 autocorrelation 0.84; 2000 Monte-Carlo realizations, identical Welch estimator): the decadal peak exceeds the 95% red-noise level across the 12.5–18.2-year band. Half the dominant period (6.7 years, gray dotted vertical line) marks the Nyquist limit.

S6. Additional main-text figures and tables

The following figures and tables support quantitative results reported in the main text. They cover the diverse-IC surface response, the evaluation-window sensitivity, the ocean-state-inflation comparison, the SST–AMOC maximum-covariance fingerprint, the two initial-condition designs, the free-running AMOC under each, and the realism of the synthetic truth against the RAPID array, and are collected here to keep the main text concise.

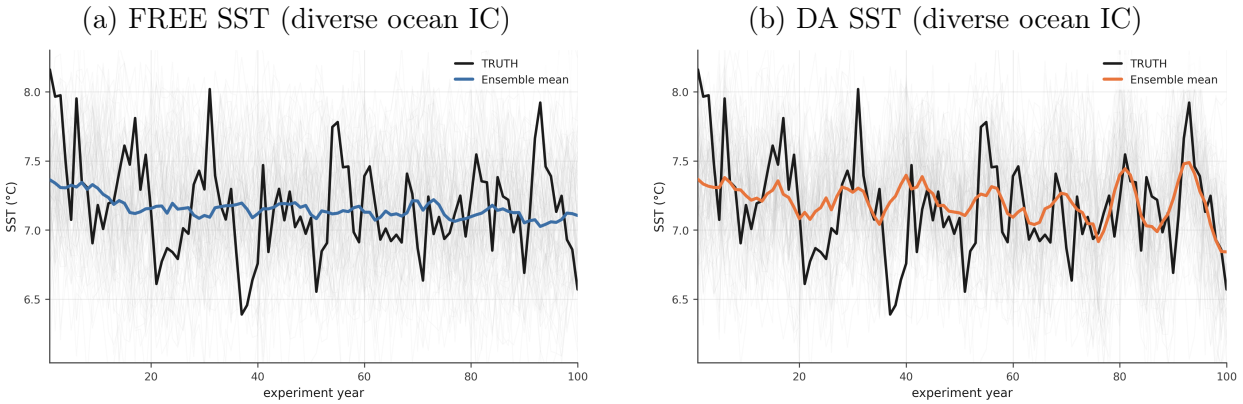


Figure S9: Subpolar-gyre SST ensemble trajectories for the diverse ocean initial-condition experiment (five-year cycles, the headline configuration): (a) FREE and (b) DA. Each ensemble member originates from a different control year (50-year spacing); SST is averaged over 50°N–65°N and 60°W–10°W. The Pearson correlation of each ensemble mean against the truth over all 100 experiment years is $r = +0.18$ for FREE and $r = +0.46$ for DA. The corresponding AMOC reconstruction is the headline reconstruction of the main text.

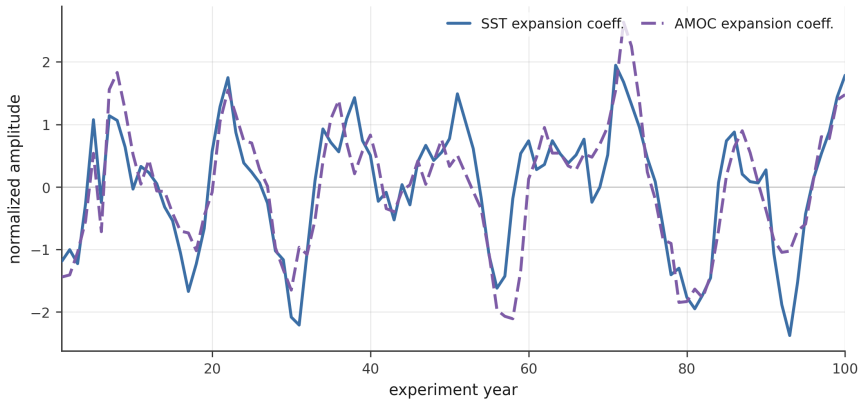
Table S5: Evaluation-window sensitivity of the assimilation gain (diverse-IC, 5-year cycles). Gains are DA − FREE except the final row, which is the DA correlation itself. "Spin-up excluded" means years 31–100 for the 100-year experiment (Sects. 3.1 and 3.3 of the main text) and years 31–300 for the 300-year integration (Sect. 3.4 of the main text).

Metric	**Full record**	**Spin-up excluded**
Δr, subpolar-gyre SST (gridded)	+0.21	+0.36
Δr, subpolar-gyre surface air temperature, SAT (gridded)	+0.14	+0.30
Δr, AMOC index	+0.68	+0.90
r_{DA}, AMOC (300-year run)	+0.78	+0.80

Only one ocean-inflation amplitude was run at a length sufficient to resolve the AMOC correlation, so Table S6 compares that experiment with its no-inflation reference on the same truth, the same initial ensemble and the same 100-year window. The dependence on amplitude is untested.

In the diverse-IC design (Figure S11), the synthetic truth branches from control year 5363, so the nearest member branch lies 13 years away, about one dominant AMOC period. That

(a) MCA PC1 (75.3% of squared covariance; the SST and AMOC expansion coefficients correlate at $r = 0.84$)



(b) Mode 1 SST pattern

(c) Mode 1 AMOC pattern

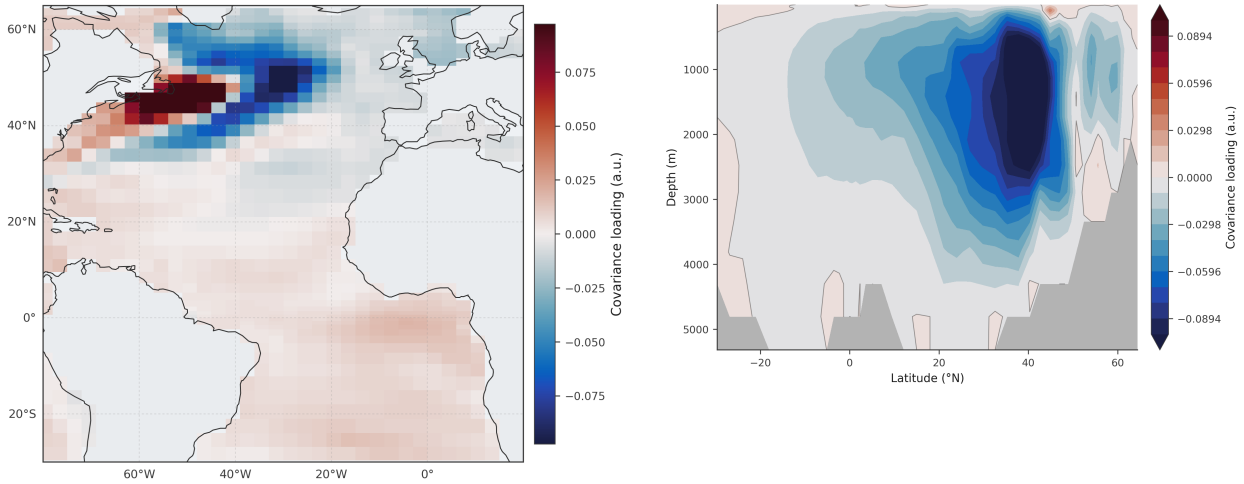


Figure S10: Maximum Covariance Analysis (MCA) of SST and AMOC streamfunction from the truth simulation. (a) Principal component time series for Mode 1, which captures 75.3% of the total squared covariance. (b, c) Spatial loading patterns (covariance loadings, arbitrary units; each panel carries its own colorbar) for SST and the Atlantic overturning streamfunction in the depth–latitude plane (gray: ocean floor), respectively. The SST loading is a northwest-Atlantic dipole rather than a basin-uniform warming, with positive loadings along the North American coast and Gulf Stream extension and negative loadings in the central subpolar basin south of Iceland, while the AMOC pattern shows the deep overturning cell structure.

Table S6: Ocean-state inflation against its no-inflation reference (100 members, five-year cycles, diverse ICs, the same truth and the same initial ensemble), forecast-convention AMOC correlation over the full 100-year record. The amplitude is the configured base value; the ESS scaling applied its 3× cap at every event, so the perturbation that ran was three times that value. "Distinct founders" is the number of surviving lineages after 100 years and "Mean ESS" the per-cycle effective sample size, both read from the per-cycle resampling draw. Only the 0.01°C amplitude was run at scoring length; the dependence on amplitude is untested.

Amplitude (°C)	**Length**	**AMOC r**	**Distinct founders**	**Mean ESS (%)**
0 (Diverse-5yr ref.)	100 yr	+0.54	3	53
0.01 (+ocean)	100 yr	+0.62	2	56

proximity does not survive into the trajectories. Over the 300-year integration (Sect. 3.4 of the main text) that member's free-running correlation with the truth is $r = +0.03$, ranking 42nd of 100, and it contributes nothing to the surviving genealogy, which descends entirely from a member branched 3487 years from the truth.

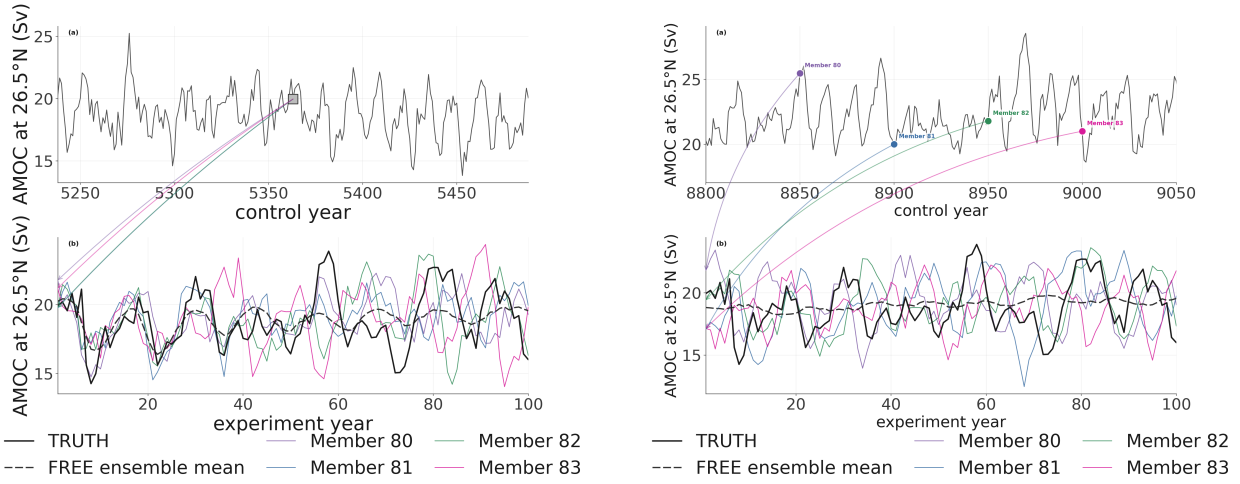
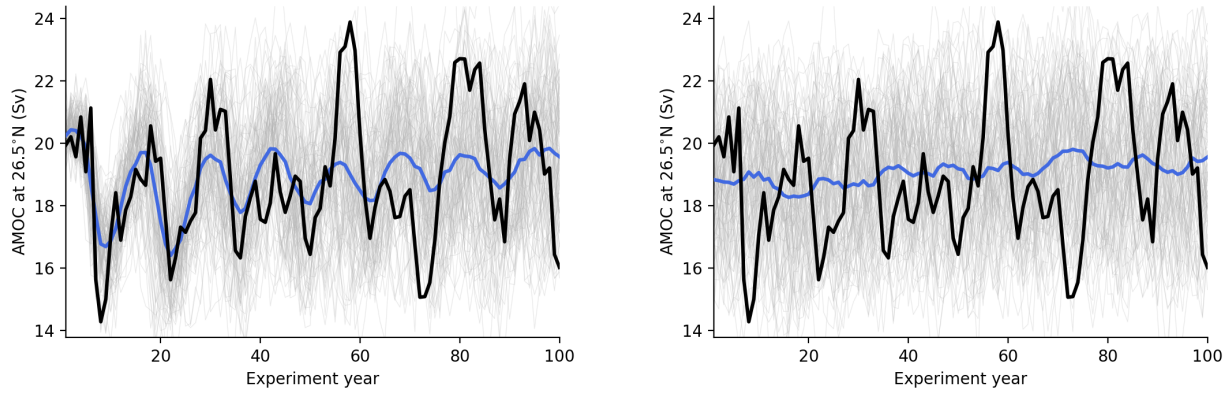


Figure S11: Initial condition strategies for the OSSE. Left: similar ocean IC; right: diverse ocean IC. In both images the top panel (a) shows the control-run AMOC at 26.5°N over the branch window. The similar-IC panel marks the single branch year in gray; the diverse-IC panel marks the branch years, one every 50 years, in the member colors. The bottom panel (b) shows selected ensemble members and the FREE ensemble mean (black dashed) against the truth (black solid) over the 100-year experiment. The branch years and ocean-state ranges are given in the Methods of the main text.

S7. Computational cost of the OSSE campaign

Table S9 summarizes the cost of the OSSE skill campaign. Every CM2Mc-BLING member integration in this study ran on 28 cores and advanced five model years per job, at a measured mean of 2310 s per job. That is 3.6 CPU core-hours per member-year, so a 100-member,



(a) Similar ocean IC

(b) Diverse ocean IC

Figure S12: Free-running (unassimilated) AMOC under two initial-condition strategies. Annual AMOC at 26.5°N (maximum below 500 m) for the truth (black), the 100 FREE ensemble members (gray), and the FREE ensemble mean (blue); both panels share the same axes. (a) Similar ocean IC: all members branch from the same control year as the truth; the ensemble mean (blue) follows the truth's overturning swings ($r = +0.52$ over the full record; +0.70 over years 1–50; +0.32 over years 51–100). (b) Diverse ocean IC: each member branches from a different control year; the ensemble mean (blue) remains nearly flat ($r = -0.14$).

Table S7: Simulated time-mean AMOC and variability at 26.5°N for the CM2Mc-BLING truth run, compared with the RAPID–MOCHA observational array (McCarthy et al., 2015; Smeed et al., 2018). For the model, Std is the raw annual-mean interannual standard deviation; the RAPID row lists the full (monthly) and annual-mean values for comparison on a matched timescale. The '—' mark indicates a not-applicable cell (RAPID compared with itself). The "vs RAPID" column expresses the model's mean departure from the RAPID observed mean (∼17.0 Sv) in units of RAPID's monthly standard deviation ($\sigma \approx 3.3$ Sv), and its Std relative to RAPID's annual-mean value and as a fraction of the 3.3 Sv monthly value.

Model	**Mean AMOC (Sv)**	**Std (Sv)**	**vs RAPID**
RAPID–MOCHA (obs.)	∼17.0	3–4 (monthly); ∼1.5–2 (ann.)	—
CM2Mc-BLING	18.9	2.1	+1.9 Sv (11%) above the RAPID mean; ≈ann.-mean obs., ∼64% of monthly

100-year experiment costs $\sim 3.6 \times 10^4$ core-hours. The full campaign comprises on the order of twenty century-scale ensemble integrations. That is roughly 2000 century-long member simulations and over 1.5×10^5 assimilated model-years, excluding spin-up and the long control runs. Scheduler accounting over the whole project records 1.7×10^6 CPU core-hours across 3.8×10^5 CM2Mc-BLING member jobs, which exceeds the productive-integration total because it also carries spin-up, the multi-millennial controls, the tier sweeps, and integrations that were restarted after numerical blow-up.

Cost of the assimilation step itself. The figures above are the cost of advancing the model. The cost of *assimilating* is a separate quantity, and it is governed by how a model is coupled rather than by the filter. Table S8 decomposes it for the 300-year integration, using the 58 consecutive five-year cycles for which scheduler accounting resolves both the forecast and the following analysis. The particle-filter algebra proper, that is weighting, the effective-sample-size check and systematic resampling for 100 members against 59 observations, takes $\sim$0.5 ms of single-core Python and is negligible in every configuration we ran. The likelihood evaluation that precedes it, which stages 100 bundled five-year ocean history files to node-local memory and forms the cost vector, takes 8 s on one core. Nearly all of the remaining cost is file movement and scheduling. A mean of 90 of the 100 restart sets were overwritten per resampling (range 71–99), and each restart set is 132 MB, so a typical analysis clones $\sim$12 GB of restart state across the shared filesystem. Summed, the analysis step averaged 14.9 minutes against a 50.5-minute ensemble forecast, so about 22% of each cycle was spent not integrating the model. We report means rather than medians because the stages then sum to the full cycle; the delays are right-skewed and forecast and analysis delays co-occur under filesystem contention, so medians would not add. On the median cycle the analysis share is 20%, and it ranges from 14% to 32% between the tenth and ninetieth percentiles, because it tracks filesystem and queue contention rather than anything in the algorithm. The two consequences for adopters, a fixed per-cycle latency that longer windows amortize and a cost that scales with restart size rather than with the number of observations, are drawn in Sect. 2.2 of the main text. The in-process coupling mode avoids the scheduler term entirely.

S8. Initial-ensemble perturbation and proxy-site mapping

The initial spread of the similar-IC ensemble (Sect. 2.5.1 of the main text) is generated by perturbing the atmospheric pressure thickness field (DELP) in the four steps below. Throughout, $m \in \{1, 2, \ldots, 100\}$ indexes the ensemble members and (i, j, k) the grid points in longitude, latitude, and vertical level.

1. **Generate a random noise field**: For each member of the ensemble m, we generate a three-dimensional white noise field:

$$\zeta^{(m)}(i, j, k) \sim \mathcal{N}(0, 1) \tag{S13}$$

   The random number generator uses a reproducible seed equal to $2000 + m$.

Table S8: Per-cycle cost decomposition of the online particle filter in the cluster-side (SLURM) coupling mode, measured over the 58 scored five-year cycles of the 300-year CM2Mc-BLING integration (the pilot cycle is excluded, and the one cycle that the scheduler forced into twelve chunks is counted once). Values are means over those cycles, so that the stages sum to the full cycle (each stage mean is rounded independently; the analysis share is the mean of the per-cycle ratios, which is not the ratio of the two means). The forecast row is the wall-clock span from the first member starting to the last member finishing; it exceeds the single-member time because the scheduler does not start all 100 tasks at once. The analysis row is the interval between the last member of one cycle finishing and the next forecast array being submitted, and therefore contains every assimilation activity. The indented entries under each stage are magnitudes of individual components rather than an exhaustive partition of it; the unlisted balance of the analysis stage is history archiving, restart preparation and orchestrator polling. The restart-cloning entry is a data volume rather than a duration.

Stage	**What it comprises**	**Mean per cycle**
Forecast (ensemble)	100 members × 5 model years, 28 cores each	50.5 min
single member	one member, 5 model years	38.5 min
queue wait	job-array submission to first task start	401 s
Analysis (total)	archive, likelihood, weights, resample, prepare	14.9 min
likelihood	100 history files → cost vector, 1 core	8 s
filter algebra	weights, ESS, systematic resampling ($N = 100$, 59 obs)	0.5 ms
restart cloning	90 of 100 members × 132 MB	∼12 GB
Full cycle	forecast + analysis	65.6 min
Analysis share	fraction of cycle wall clock (median 20%)	22.0%

Table S9: Computational cost of the OSSE skill campaign on the PIK high-performance computing (HPC) cluster. Standard nodes are dual-socket 128-core CPU nodes. The final column reports the per-run cost in CPU core-hours per 100-year experiment. FREE and DA ensembles comprise 100 members each.

Model	**Model + ocean**	**Grid**	**Hardware**	**Wall-clock (resource cost)**
CM2Mc-BLING	MOM5 p^*-level	3° (0.6° eq.)	standard CPU	$\sim 3.6 \times 10^4$ core-h / 100-yr expt.

2. **Apply spatial smoothing**: We apply Gaussian smoothing to ensure spatial coherence:

$$\zeta'^{(m)}(i,j,k) = \left(G_{\sigma_h} *_h \left(G_{\sigma_v} *_v \zeta^{(m)}\right)\right)(i,j,k) \tag{S14}$$

where G_σ denotes a Gaussian kernel with standard deviation σ, and $*_h$ and $*_v$ denote horizontal and vertical convolution, respectively. We use $\sigma_h = 2.5$ grid cells (approximately 500 km at mid-latitudes) for horizontal smoothing and $\sigma_v = 1.5$ vertical levels for vertical smoothing.

3. **Scale and apply perturbation**: The smoothed noise field is scaled to 0.1% of the global mean DELP and added to the base state:

$$\mathrm{DELP}'^{(m)}(i,j,k) = \mathrm{DELP}_0(i,j,k) + \alpha \cdot \overline{\mathrm{DELP}} \cdot \zeta'^{(m)}(i,j,k) \tag{S15}$$

where DELP_0 is the unperturbed pressure thickness field from the control state, $\overline{\mathrm{DELP}}$ is the global volume-weighted mean pressure thickness (approximately 10,000 Pa across all levels), and $\alpha = 0.001$ is the perturbation amplitude. This scaling yields perturbations of approximately 10 Pa root mean square, small enough to remain within the model's stable regime but sufficient to generate ensemble spread through chaotic amplification during integration.

4. **Enforce physical bounds**: To prevent numerical instabilities at the model top where pressure layers are thin, we enforce a minimum threshold during initial ensemble generation:

$$\mathrm{DELP}'^{(m)}(i,j,k) = \max\left(\mathrm{DELP}'^{(m)}(i,j,k),\ 100\ \mathrm{Pa}\right) \tag{S16}$$

Proxy-site mapping. The nearest grid point for a given proxy location is determined by minimizing the approximate spherical distance over all ocean grid points:

$$(i_p, j_p) = \underset{(i,j)\in\Omega_{\mathrm{ocean}}}{\arg\min} \sqrt{(\phi_{\mathrm{grid}}(i,j) - \phi_p)^2 + (\lambda_{\mathrm{grid}}(i,j) - \lambda_p)^2 \cos^2(\phi_p)}, \tag{S17}$$

where (ϕ_p, λ_p) are the latitude and longitude of proxy p, $(\phi_{\mathrm{grid}}, \lambda_{\mathrm{grid}})$ are the model grid coordinates, and Ω_{ocean} restricts the search to ocean points.

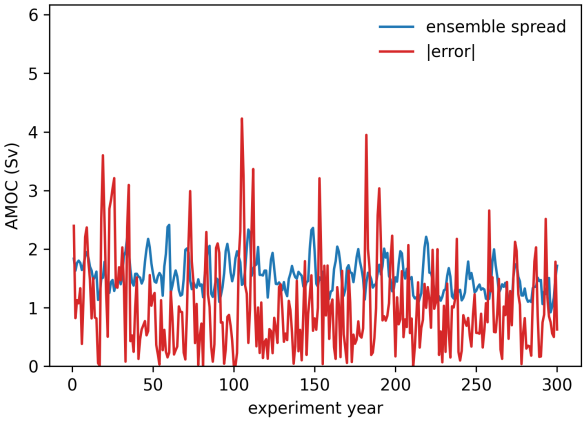
S9. Particle-filter degeneracy: reliability and genealogy diagnostics

Path degeneracy has to be shown directly, and it has to be kept separate from a reduction in ensemble spread and from a lack of reliability. This section quantifies the three separately for the 300-year CM2Mc-BLING integration (Sect. 3.4 of the main text).

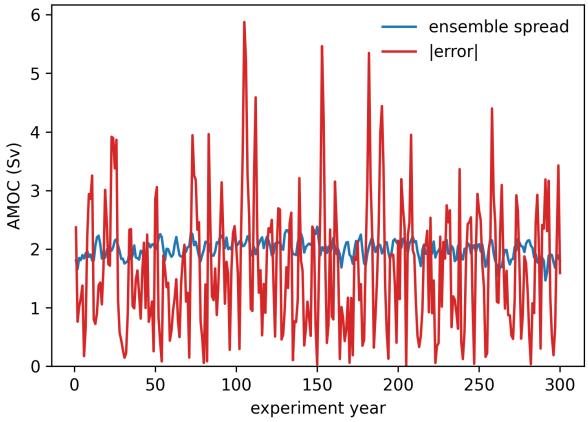
Reliability. We compute the spread–skill ratio mean(σ_{ens})/RMSE($\overline{x} - x_{\text{truth}}$), for which a value near 1 is calibrated and a value < 1 signals overconfidence (the spread underestimates the actual error). For the 300-year diverse-IC CM2Mc-BLING integration (Sect. 3.4 of the main text) the DA AMOC ratio is 1.20 over the full record, compared with 1.00 for the unassimilated FREE ensemble; the subpolar-gyre SST ratio is 1.17 (DA) versus 1.02 (FREE). The assimilated ensemble is therefore, if anything, mildly over-dispersed rather than overconfident. Rank histograms (probability integral transform) agree. The truth falls outside the DA ensemble in 1.7% of years, against the 2.0% a flat histogram implies for 100 members, and inside the central half of the ensemble in 58% of years, against the 49% a flat rank distribution implies for the interpolated quartiles of 100 members. The residual departure from flatness is modest ($\chi^2 = 16.9$ on 9 nominal degrees of freedom for DA, 6.2 for FREE). These χ^2 statistics assume independent verification times. Because the AMOC decorrelates over $\sim$10 years, the 300-year record supplies only $\sim$30 independent verification times, so the nominal degrees of freedom overstate significance. The single-series estimate here is $N(1-\rho_1)/(1+\rho_1) \approx 33$ at $\rho_1 = 0.80$, which is distinct from the two-series $N_{\text{eff}} \approx 52$ that governs the correlation test of Sect. 3.4 of the main text. We therefore read the rank histograms descriptively and treat the spread–skill ratio as the main quantitative reliability metric. Both diagnostics for DA and FREE are shown in Figure S13. Among the 100-year experiments the same holds at the five- and ten-year cycles (headline spread–skill 0.96, rank $\chi^2 = 9.8$). The two annual experiments are the exception. They compress the spread to a ratio of 0.44, put the truth outside the ensemble in 24–29% of years, and give strongly U-shaped histograms ($\chi^2 > 118$). Overconfidence in this system is therefore a symptom of weight degeneracy at short cycles, not a general property of the filter. The spread itself is near-stationary over the 300 years (Figure S14). The assimilated ensemble holds a mean σ_{ens} of 1.53 Sv against 2.00 Sv for FREE, a ratio of 0.77, so resampling reduces the spread by roughly a quarter and then keeps it there rather than collapsing it further.

Genealogy and re-diversification timescale. Path degeneracy is distinct from weight degeneracy: the effective sample size can be adequate at any single analysis, while the ensemble nonetheless loses all memory of independent trajectories. The per-cycle genealogy of the 300-year CM2Mc-BLING experiment (Sect. 3.4 of the main text) is saved at every cycle. When we trace the ancestry of every member through the resampling maps, the number of distinct surviving founders collapses monotonically (Figure S16b). The ensemble has 3 founders left at year 100 and a single lineage at year 155. Throughout that run the per-cycle effective sample size stays high (mean 54%, minimum 32%), and a typical analysis draws on 66 distinct immediate parents out of 100. The Atlantic-network experiment (Sect. 3.7 of the main text), which is less degenerate, stays further from collapse, retaining 72 of 100 distinct *immediate* parents in a typical analysis (66% mean ESS) and 7 founders after 100 years, against 3 for the global network; neither reaches a single lineage within the century.

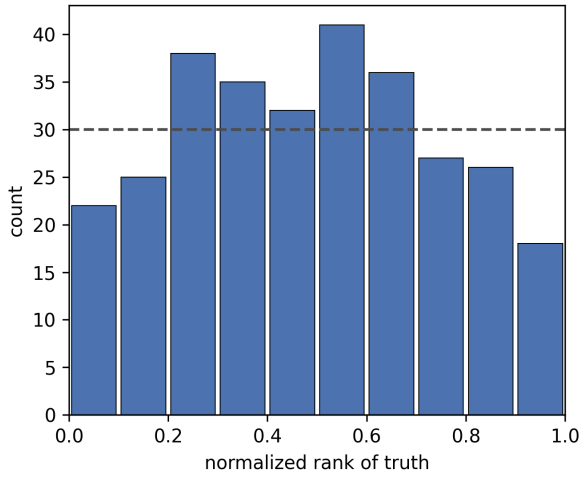
To quantify how quickly cloned members re-separate, we measure the mean pairwise AMOC distance between members *sharing* an ancestor versus members with *different* ancestors, as a function of lead time since the resampling event, and report their ratio (Figure S16a). This re-diversification timescale is short. At the first post-resampling forecast, one five-year window after cloning, the ratio is 0.48. Same-ancestor pairs then recover to



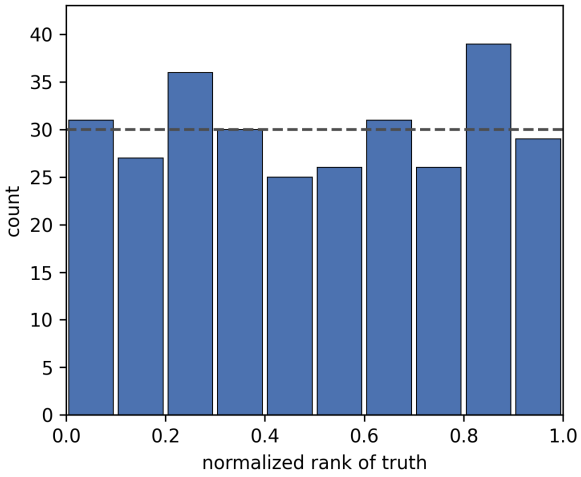
(a) DA: ensemble spread vs. error



(b) FREE: ensemble spread vs. error



(c) DA: AMOC rank histogram



(d) FREE: AMOC rank histogram

Figure S13: Reliability of the AMOC ensemble over the 300-year CM2Mc-BLING integration (diverse-IC, five-year cycles; Sect. 3.4 of the main text). (a,b) Ensemble spread $\sigma_{\text{ens}}(t)$ against the absolute error of the ensemble mean $|\overline{x} - x_{\text{truth}}|(t)$, for DA and for FREE respectively. (c,d) Rank (probability-integral-transform) histograms sharing a common vertical axis: each year the truth's rank among the 100 members is normalized to $[0, 1]$ (0: truth below every member; 1: above every member), and the dashed line is the calibrated reference count ($N_t/10 = 30$ per bin). Spread–skill ratios and rank-histogram statistics are given in Section S9.

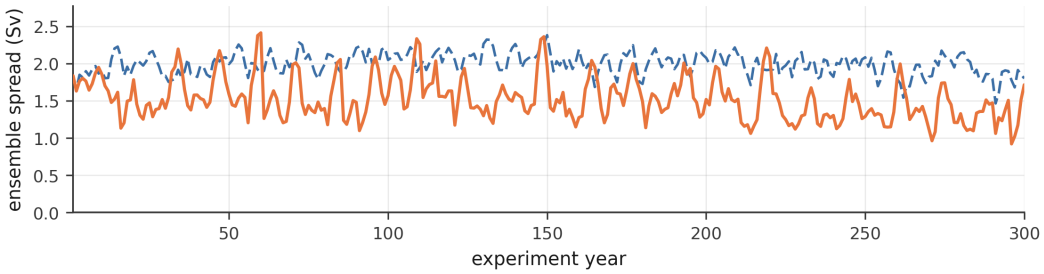
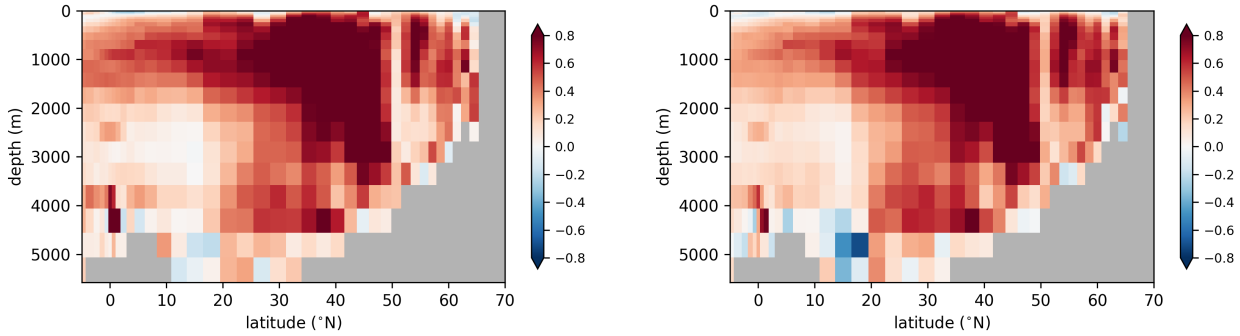


Figure S14: Ensemble AMOC spread $\sigma_{\mathrm{ens}}(t)$, the per-year standard deviation across members, over the 300-year CM2Mc-BLING integration (diverse initial conditions, five-year cycles; Sect. 3.4 of the main text). DA in orange, FREE in blue dashed. Mean values and their ratio are given in the text above.

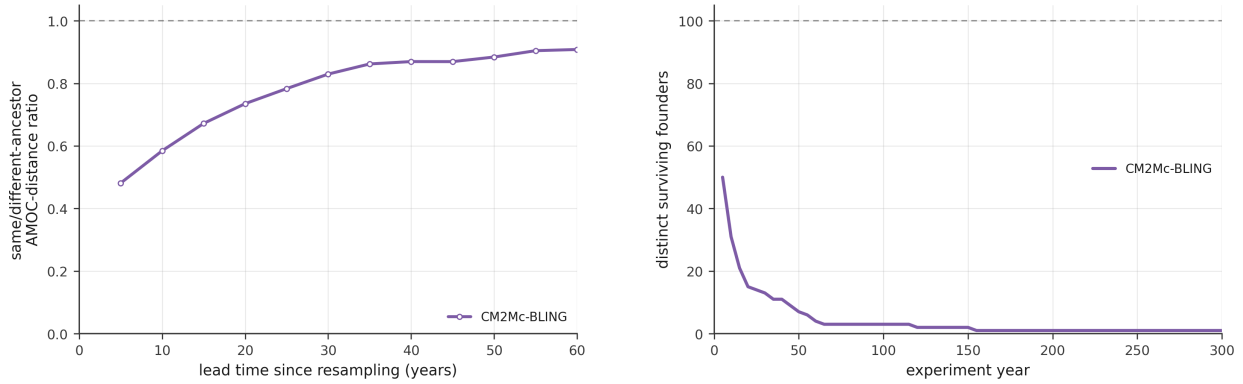


(a) Global 59-proxy network

(b) Atlantic-only 34-proxy network

Figure S15: Per-cell difference in time correlation with the truth, $\Delta r = r(\mathrm{DA}, \mathrm{truth}) - r(\mathrm{FREE}, \mathrm{truth})$, for the Atlantic overturning streamfunction ψ in the depth–latitude plane, over years 31–100 of the 100-year campaign (Sect. 3.1 of the main text). Red: DA more correlated with the truth than FREE; blue: less. Gray: seafloor. The deep North Atlantic Deep Water cell referred to in the main text spans ∼30–55°N, 0.5–3 km. These are the overturning-plane counterparts of the surface maps in Figure 3 of the main text; the 300-year run is scored on the same plane in Figure 6 of the main text.

80% of the contemporaneous different-ancestor distance within $\tau \approx 30$ years, or six assimilation cycles, and then plateau near 0.91. Even *different*-ancestor pairs do not reach the fully independent FREE reference: their distance is about 1.25 times smaller than in FREE. The resampled ensemble therefore retains much of its natural spread among unrelated lineages. This is expected in CM2Mc-BLING, where the post-resampling atmospheric perturbation re-diversifies surface weather within a cycle. At the same time, the deep density structure that sets the AMOC retains the founder's state for decades.



(a) Same- vs different-ancestor divergence

(b) Surviving founders vs. time

Figure S16: Genealogical degeneracy in the 300-year CM2Mc-BLING experiment of Sect. 3.4 of the main text, from its per-cycle resampling maps. (a) Ratio of the mean pairwise AMOC distance for same-ancestor versus different-ancestor member pairs, as a function of lead time since a resampling event (dashed line = ratio = 1, equal same- and different-ancestor distances). (b) Number of distinct surviving founders traced to the start of the run (dashed line = FREE's unresampled 100). Re-diversification ratios and collapse timescales are given in Section S9.

S10. Filter operators: likelihood tempering, the weight cap and adaptive resampling

These three operators are part of the engine and are used with fixed settings in every CM2Mc-BLING experiment of the main text (Sect. 2.3).

Tempered likelihood

The effective error $\sigma_{\text{eff}} = \eta\,\sigma_{\text{obs}}$ of Equation 4 of the main text is a likelihood *tempering*, not only a widening of the error. Inflating the Gaussian error standard deviation by η is the same as raising the likelihood to a power,

$$p_\beta(\mathbf{y}_t \mid \mathbf{x}_m) \;\propto\; p(\mathbf{y}_t \mid \mathbf{x}_m)^\beta, \qquad \beta = \eta^{-2} \in (0, 1], \tag{S18}$$

Since the exponent of the diagonal Gaussian scales as $\sigma_{\text{eff}}^{-2} = \eta^{-2}\sigma_{\text{obs}}^{-2}$, the inverse temperature β flattens the weight distribution toward uniform. It counteracts the weight collapse that

afflicts particle filters as the observation dimension grows (Snyder et al., 2008). In the two limiting cases, $\beta \to 1$ ($\eta = 1$) recovers the untempered filter, while $\beta \to 0$ ($\eta \to \infty$) ignores the observations. It is the main control for keeping the effective sample size (Equation 6 of the main text) from collapsing, and the main parameter to retune for a new model. Because β is fixed, the analysis targets a tempered posterior, and the ESS values reported are those of the tempered weights. The untempered weights from the same cost vectors would give an ESS near 2 (mean 1.8, range 1.0–4.6 over the 300-year run).

Max-weight cap

As an optional degeneracy safeguard, the engine can bound the single largest normalized weight. Let $w_{(1)} = \max_m w_m$ be the largest weight, attained at index $m^\star$. The cap operator with ceiling $w_{\max} \in (1/N, 1]$ leaves the weights untouched when $w_{(1)} \le w_{\max}$ and otherwise sets

$$w_{m^\star} \leftarrow w_{\max}, \qquad w_m \leftarrow w_m \left(1 + \frac{w_{(1)} - w_{\max}}{1 - w_{(1)}}\right) \quad (m \neq m^\star), \tag{S19}$$

redistributing the surplus $w_{(1)} - w_{\max}$ over the remaining members in proportion to their current weights and renormalizing. This bounds the dominant lineage, which drives genealogical collapse (Sect. 3.8 of the main text). It does not bound every weight. Because the cap is applied once to the largest weight and is not iterated, a member that was previously second can end up above $w_{\max}$ once the surplus is redistributed. The cap is inactive for CM2Mc-BLING, where the fourfold likelihood tempering ($\eta = 4.0$) alone keeps $w_{(1)}$ bounded. By contrast, it is applied with $w_{\max} = 0.3$ in the idealized benchmarks of Section S4 in the Supplement (Table S4).

Adaptive resampling

Resampling is triggered only when the effective sample size of Equation 6 of the main text drops below a fraction τ_{ess} of N ($\tau_{\text{ess}} = 0.85$ here). The filter therefore incurs the genealogical cost of resampling only once the weights have actually degenerated. Across the eight 100-year CM2Mc-BLING experiments reported here, though, the ESS fell below this fraction in all but four of the 300 analyses (one in the similar-IC five-year experiment and three in the similar-IC ten-year one), so resampling was in practice unconditional. All four exceptions are similar-IC; none of the 170 diverse-IC analyses skipped resampling. The idealized Lorenz $\eta \times \tau_{\text{ess}}$ sweep of Table S2 deliberately includes a dead zone in which the filter never resamples. The post-resampling inflation of Sect. 2.3.4 of the main text then perturbs each member to restore the ensemble diversity needed to keep the ESS (Equation 6 of the main text) from re-collapsing on the next cycle. This trade-off is the diversity-versus-memory tension we examine in this paper.

References

Halldór Björnsson and Silvia A. Venegas. A manual for EOF and SVD analyses of climatic data. *CCGCR Report*, 97(1):1–52, 1997. McGill University, Montreal.

Christopher S Bretherton, Martin Widmann, Valentin P. Dymnikov, John M Wallace, and Ileana Bladé. The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, 12(7):1990–2009, 1999. doi: 10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2.

L. Caesar, S. Rahmstorf, A. Robinson, G. Feulner, and V. Saba. Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature*, 556(7700):191–196, 2018. doi: 10.1038/s41586-018-0006-5.

James A. Carton and Benjamin S. Giese. A reanalysis of ocean climate using simple ocean data assimilation (SODA). *Monthly Weather Review*, 136(8):2999–3017, 2008. doi: 10.1175/2007MWR1978.1.

Benjamin S. Giese and Sulagna Ray. El Niño variability in simple ocean data assimilation (SODA), 1871–2008. *Journal of Geophysical Research: Oceans*, 116(C2):C02024, 2011. doi: 10.1029/2010JC006695.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Paul Keil, Thorsten Mauritsen, Johann Jungclaus, Christopher Hedemann, Dirk Olonscheck, and Rohit Ghosh. Multiple drivers of the North Atlantic warming hole. *Nature Climate Change*, 10:667–671, 2020. doi: 10.1038/s41558-020-0819-8.

X. San Liang. Unraveling the cause-effect relation between time series. *Physical Review E*, 90(5):052150, 2014. doi: 10.1103/PhysRevE.90.052150. URL `https://link.aps.org/doi/10.1103/PhysRevE.90.052150`.

X. San Liang. Normalized multivariate time series causality analysis and causal graph reconstruction. *Entropy*, 23(6):679, 2021. doi: 10.3390/e23060679.

Edward N. Lorenz. Predictability: A problem partly solved. In *Proceedings of the Seminar on Predictability*, volume 1, pages 1–18. ECMWF, Reading, UK, 1996.

Edward N. Lorenz and Kerry A. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55(3): 399–414, 1998. doi: 10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2.

Gerard D. McCarthy, David A. Smeed, William E. Johns, Eleanor Frajka-Williams, Ben I. Moat, Darren Rayner, Molly O. Baringer, Christopher S. Meinen, Jed Collins, and Harry L. Bryden. Measuring the Atlantic Meridional Overturning Circulation at 26°N. *Progress in Oceanography*, 130:91–111, 2015. doi: 10.1016/j.pocean.2014.10.006.

Malaquías Peña and Eugenia Kalnay. Separating fast and slow modes in coupled chaotic systems. *Nonlinear Processes in Geophysics*, 11(3):319–327, 2004. doi: 10.5194/npg-11-319-2004.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Stefan Rahmstorf, Jason E. Box, Georg Feulner, Michael E. Mann, Alexander Robinson, Scott Rutherford, and Erik J. Schaffernicht. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nature Climate Change*, 5(5):475–480, 2015. doi: 10.1038/nclimate2554. URL `https://www.nature.com/articles/nclimate2554`.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10:2553, 2019a. doi: 10.1038/s41467-019-10105-3. URL `https://www.nature.com/articles/s41467-019-10105-3`.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019b. doi: 10.1126/sciadv.aau4996. URL `https://www.science.org/doi/10.1126/sciadv.aau4996`.

David A. Smeed, Simon A. Josey, Claudie Beaulieu, William E. Johns, Ben I. Moat, Eleanor Frajka-Williams, Darren Rayner, Christopher S. Meinen, Molly O. Baringer, Harry L. Bryden, and Gerard D. McCarthy. The North Atlantic Ocean is in a state of reduced overturning. *Geophysical Research Letters*, 45(3):1527–1533, 2018. doi: 10.1002/2017GL076350.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeffrey Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008. doi: 10.1175/2008MWR2529.1.

Peter Jan van Leeuwen, Hans R Künsch, Lars Nerger, Roland Potthast, and Sebastian Reich. Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, 145:2335–2365, 2019. doi: 10.1002/qj.3551.

Rong Zhang, Rowan Sutton, Gokhan Danabasoglu, Young-Oh Kwon, Robert Marsh, Stephen G. Yeager, Daniel E. Amrhein, and Christopher M. Little. A review of the role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and associated climate impacts. *Reviews of Geophysics*, 57(2):316–375, 2019. doi: 10.1029/2019RG000644.