Response for the reviewer 1

We would like to express our sincere gratitude to the Editor and the Reviewers for their thoughtful evaluation, constructive feedback, and valuable time spent reviewing our manuscript. We have carefully addressed all comments and conducted extensive additional experiments and statistical analyses to resolve the methodological concerns raised. These revisions have substantially improved the statistical rigor, clarity, and physical grounding of our manuscript. Below, we provide a detailed, point-by-point response to all comments, accompanied by a summary of the major updates incorporated into the revised paper.

Overall, we acknowledge that our central conclusion that SSS assimilation is superior to SST for reducing uncertainty in AMOC reproduction and prediction cannot be fully supported solely by surrogate-model uncertainty quantification and simple posterior simulation analyses. To decisively address this methodological limitation, we have incorporated additional rigorous statistical tests that operate directly on full model simulations, completely bypassing the surrogate model. As detailed in our responses below, these surrogate-model independent analyses confirm that our primary findings are statistically robust.

In addition, a revised version of the parameter list is provided on the following page (Table R1). Please note that Parameter 21 (P21) in the original manuscript, which represents the scaling factor for computing the Bering Strait throughflow, has been renamed to Parameter 20 (P20) in both the updated parameter list and throughout this response letter to ensure consistency across our revised analyses.

Table R1

Index	Parameter	Definition	Synthetic Truth	Range
P1	ampwir	Scaling coefficient for the longwave radiation scheme (amplw)	1	[0.5, 1.5]
P2	ampeqir	Scaling coefficient for the longwave radiation scheme (amplw)—for the equator area between 15S and 15N.	1.8	[1.0, 2.5]
P3	expir	Exponent for the longwave radiation scheme	0.4	[0.2, 0.6]
P4	relhmax	The relevant threshold pfthe total precipitable water below 500 hPa	0.83	[0.50, 0.90]
P5	cwdrag	Drag coefficient to compute wind stress	2.1×10^{-3}	$[1.0\times10^{-3}, 4.0\times10^{-3}]$
P6	cdrag	Drag coefficient to compute sensible and latent heat fluxes	1.4×10^{-3}	$[1.0\times10^{-3}, 2.0\times10^{-3}]$
P7	uv10rfx	Reduction in wind speed between 800 hPa and 10 m	0.8	[0.7, 0.9]
P8	dragan	Rotation of the wind vector in the boundary layer (unit: degree)	15	[10, 20]
P9	alphd	Albedo of snow	0.72	[0.60, 0.90]
P10	alphdi	Albedo of bare ice	0.62	[0.50, 0.80]
P11	alphs	Albedo of melting snow	0.53	[0.30, 0.60]
P12	albice	Albedo of melting ice (general)	0.44	[0.30, 0.60]
P13	albin	Albedo of melting ice (Arctic)	0.44	[0.30, 0.60]
P14	albis	Albedo of melting ice (Antarctic)	0.44	[0.30, 0.60]
P15	cgren	Increase in snow/ice albedo under cloudy conditions	0.04	[0.01, 0.10]
P16	corAN	Adjestment parameter of precipitation in Atlantic Ocean (North)	-0.085	[−0.10, −0.05]
P17	corAS	Adjestment parameter of precipitation in Atlantic Ocean (South)	-0.085	[−0.10, −0.05]
P18	corAC	Reduction in precipitation over the Arctic	-0.25	[−0.30, − 0.20]
P19	evfac	Maximum evaporation factor over land	1	[0.5, 1]
P20	bering	Scaling factor for computing the Bering Strait throughflow	0.3	[0.2, 0.5]
P21	ai	Coefficient of isopycnal diffusion (unit: m2 s−1)	300	[200, 400]
P22	aitd	Gent-McWilliams thickness diffusion coefficient (unit: m2 s−1)	300	[200, 400]
P23	ahs	Horizontal diffusivity for scalars (unit: m2 s−1)	100	[50, 150]
P24	ahu	Horizontal viscosity (unit:m2 s−1)	1.0×10^{5}	$[0.5\times10^{5}, 1.5\times10^{5}]$

Major Comments:

1. lines 175–180 and lines 550–559.

Equation (2) introduces a covariance inflation coefficient alpha, which is set to 100. This is a quite a consequential modeling choice: multiplying the covariance by 100 substantially broadens the likelihood and therefore directly determines how strongly the observations can constrain the parameter posterior. Later, the authors themselves identify covariance inflation as one of the main reasons for the remaining large posterior uncertainty.

Why is α = 100 chosen? The authors should (i) provide a clear statistical motivation for this inflation, (ii) show the results for a reasonable range of alpha, and (iii) demonstrate whether the main conclusion—in particular the much stronger constraint obtained from SSS compared with SST—is robust to this choice.

Without such a sensitivity experiment, it is difficult to interpret the reported posterior widths as meaningful estimates of parametric uncertainty.

We sincerely thank the reviewer for this comment. Addressing this point led us to thoroughly audit our posterior sampling code, resulting in an essential correction that has significantly improved the statistical rigor of our parameter estimation framework.

This comment is about inflation coefficients α of likelihood function. The likelihood function $p(u_{obs}|x)$ was modelled as the following equation:

$$\ln p(u_{obs}|\theta) = -\frac{1}{2}\big(u_{obs} - m(\theta)\big)^T(\alpha\Sigma_{total}(\theta))^{-1}\big(u_{obs} - m(\theta)\big) + \text{const} \qquad (R1)$$

where $\theta \in \mathbb{R}^n$ is a parameter vector, m: $\mathbb{R}^n \to \mathbb{R}^d$, which is prediction mean of Gaussian Process Regression, and Σ_{total}: $\mathbb{R}^n \to \mathbb{R}^d \times \mathbb{R}^d$, which is estimation variance from prediction variance of the Gaussian Process Regression Σ_{model} and observation error Σ_{obs},

$$\Sigma_{total} = \Sigma_{model} + \Sigma_{obs}.$$

u_{obs} is the observational data projected onto the subspace spanned by the leading d principal components.

1. Identification of Root Cause and Removal of Artificial Inflation

In the original manuscript, the covariance inflation factor $\alpha = 100$ was introduced as an empirical workaround. Without inflation, we observed that smaller observation errors counter-intuitively yielded larger Mean Squared Errors (MSEs) in parameter estimates, caused by sampling instability in the REMC algorithm when resolving narrow posterior regions. For instance, we observed that, without inflation, the estimated MSE decreases as observation error increases when SSS is used as an observation variable.

To address the reviewer's points (i) and (ii), we performed a sensitivity test of the inflation coefficient α. We tested $\alpha \in \{1,10,100\}$ to find a sufficient inflation to avoid invalid sampling. We found that α should be over 10 and decided selecting $\alpha = 100$.

However, while addressing the reviewer's comment, we found a critical oversight in our original implementation: the predictive variance matrix provided by the Gaussian Process Regression (GPR) was inadvertently omitted from the likelihood function. We corrected the codes and found that even without artificial inflation, the sampling process is stable (Figure R1). Therefore, we removed inflation (or set the inflation coefficient $\alpha = 1.0$).

2. Impact on the main conclusion: SSS is superior to SST to reduce uncertainty in AMOC

To address point (iii), we re-evaluated the relative utility of SSS and SST observations using the corrected setup ($\alpha = 1.0$). When comparing the overall posterior distributions averaged across all 25 parameters, the difference in total parameter MSE reduction between SSS and SST assimilation becomes smaller than reported in the original manuscript (Figure R1).

However, a parameter-by-parameter analysis reveals distinct physical sensitivities. Figure R2 showed the marginal MSEs of posterior parameters related to the salinity flux in the North Atlantic Ocean, such as precipitation adjustment parameter in the Atlantic Ocean and scaling factor for computing the Bering Strait throughflow (Parameter 16, Parameter 17 and Parameter 20), are significantly smaller when SSS is adopted as observation. On the other hand, there are several parameters whose error is reduced more effectively by SST observation than SSS observation, such as albedo of melting snow, drag coefficients to compute sensible and latent heat fluxes, and albedo of bare ice (Parameter 11, Parameter 6 and Parameter 10).

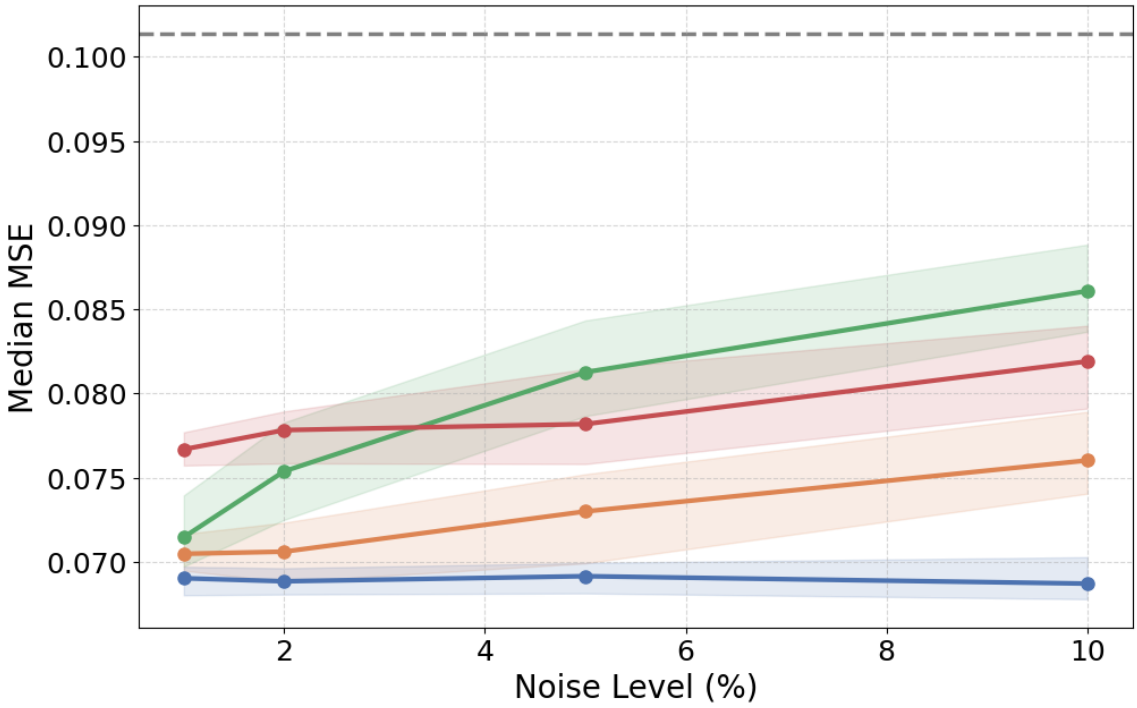


Figure R1. Sensitivity of parameter estimation accuracy and uncertainty to observation noise levels across different observation variables. The vertical axis shows the mean squared error, and the horizontal axis represents the observation noise amplitude. Solid lines show the medians across 20 independent trials initialized with different initial conditions. Orange, blue, green, and red lines correspond to assimilation results using SST, SSS, Q, and GEOPG, respectively.

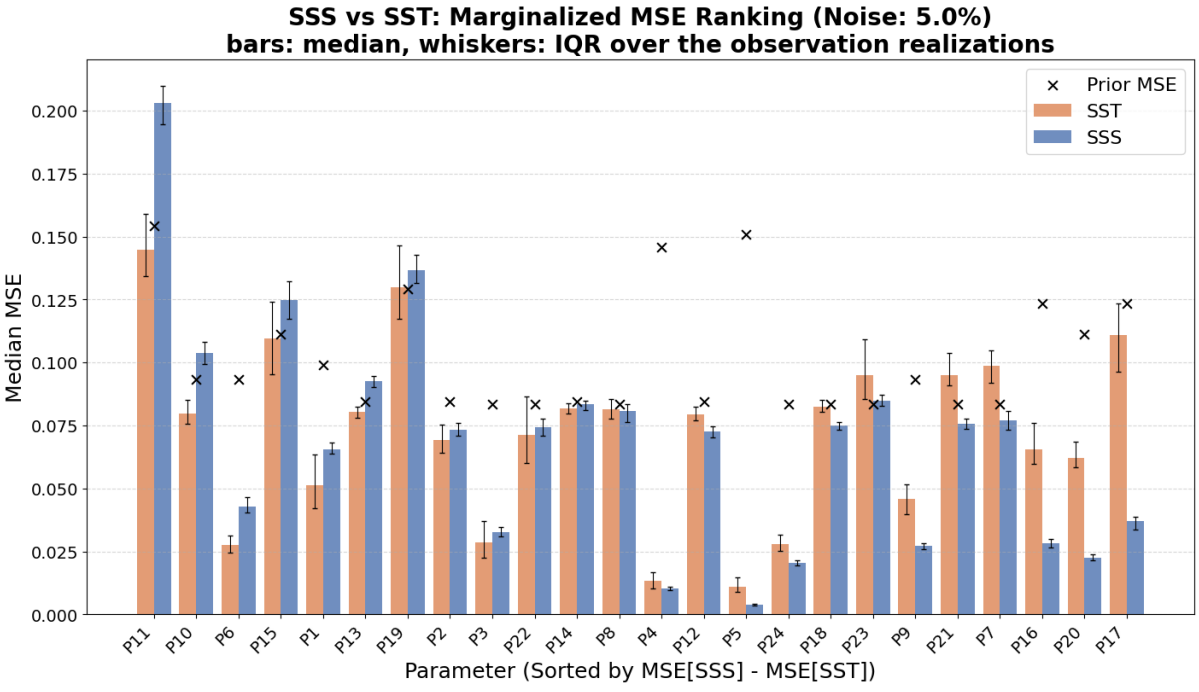


Figure R2. Comparison of the estimated parameter distributions of SSS with those with the other observation variables. Marginal MSEs of posterior parameter distributions. Orange and blue bars show the MSEs obtained using SST and SSS as an observation variable, respectively. Parameters are ordered along the horizontal axis such that those farther to the right exhibit a smaller marginal MSE under SSS assimilation than under SST. Crosses show the prior marginal MSEs. Error bars represent the range of marginal MSEs across synthetic true simulations with different initial conditions.

2. lines 157–164, lines 324–335 and lines 550–559

The manuscript evaluates the GPR mainly using R-squared values of the PCA components and argues that the first few components are reproduced sufficiently well. This demonstrates that the emulator captures a substantial fraction of the large-scale variance, but this is not necessarily sufficient for posterior inference. Posterior estimation depends on accurately resolving potentially narrow regions of parameter space, rather than only reproducing the dominant global variance of the training ensemble. Indeed, the authors later acknowledge that 1000 training simulations may be insufficient to resolve narrow, highly correlated posterior structures in the 25-dimensional parameter space.

A more direct validation of the emulator should be provided to show that it is fit for purpose. For example, one could compare emulator predictions with withheld LOVECLIM simulations specifically in high likelihood/posterior regions and evaluate whether the predictive uncertainty of the GPR is calibrated. Ideally, posterior inference based on the emulator should also be compared for a subset of cases against direct LOVECLIM evaluations. At minimum, the manuscript should demonstrate that the reported differences between SSS and SST are larger than uncertainties introduced by the emulator itself.

We appreciate the reviewer for the critical point regarding surrogate model fidelity in posterior regions. In response, we performed comprehensive direct validations using full LOVECLIM mode simulations (subsection.1) and conducted two independent statistical analyses without relying on the surrogate model (subsection.2). We confirmed that our previous conclusion does not hold, but other robust conclusions were obtained through the analysis.

1. Direct evaluation of the Surrogate model using Full Posterior simulations

We evaluated how the surrogate model reproduces the state variables such as SST and SSS using posterior simulations with 688 valid members out of 800 samples. Note that those posterior simulations are not the same as we had in the original draft, which are not sufficiently spun up. We reran the posterior simulations with sufficient spin-up periods with updated posterior samples. Refer to the answer to the comment 6 for its details.

We reconstructed the physical state fields by combining the GPR prediction (mapping parameters to latent PCA space) with the inverse PCA transformation. Comparing the emulated fields against the full LOVECLIM runs (Figures R3 and R4). Figure R3 (a,b) and R4 (a,b) showed that both GPRs of SSS and SST achieve large R^2 at low- and mid-latitude areas ($R^2 > 0.6$), while their R^2 at high-latitude areas is low. While the normalized RMSEs of the GPR of SSS standard deviation of interannual variability are also small at lower latitudes, the GPR of SST has large normalized RMSEs at the area, as shown in Figure R3(c) and R4(c).

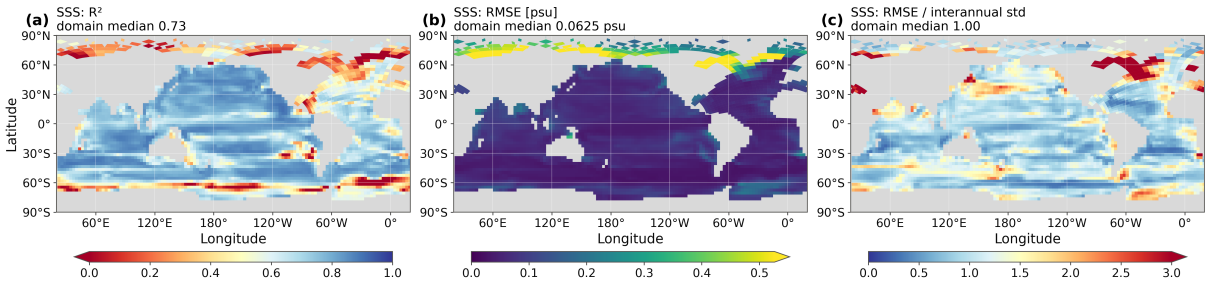


Figure R3. Skill of the surrogate model for SSS at the parameters of the 688 completed posterior simulations out of 800 posterior parameter samples calibrated on 5%-error observation data of SST, SSS, precipitable water and geopotential heights. (a) R^2 score of the surrogate model on each grid, (b) RMSE of the surrogate model prediction in psu. (c) The same RMSE divided by the interannual standard deviation of the true run on each grid.

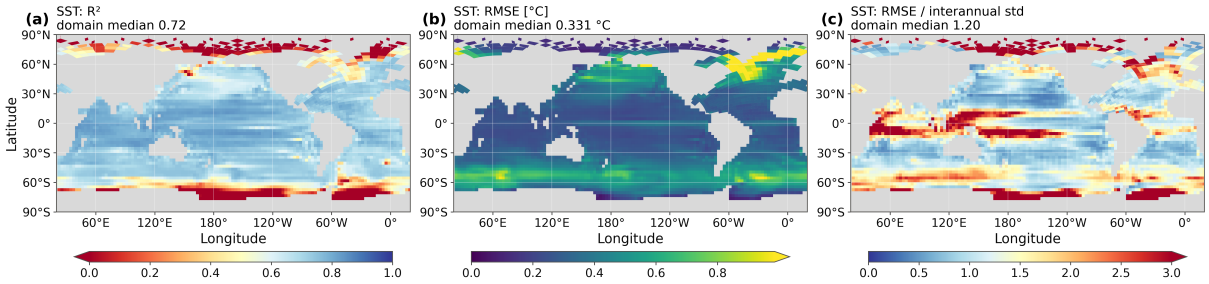


Figure R4. Skill of the surrogate model for SST at the parameters of the 688 completed posterior simulations out of 800 posterior parameter samples calibrated on 5%-error observation data of SST, SSS, precipitable water and geopotential heights. (a) R^2 score of the surrogate model on each grid, (b) RMSE of the surrogate model prediction in psu. (c) The same RMSE divided by the interannual standard deviation of the true run on each grid.

2. Surrogate-model Independent Proof of SSS Superiority via Odds-Ratio Analysis and Multi-variable regression

As for your comments, "*At minimum, the manuscript should demonstrate that the reported differences between SSS and SST are larger than uncertainties introduced by the emulator itself.*", We do not think we can prove that the reported differences between SSS and SST are statistically significant and entirely independent of surrogate-model uncertainty only using the posterior parameters and simulations. Then, we performed several additional experiments to demonstrate that the reported differences between SSS and SST are larger than uncertainties introduced by the surrogate model from the multiple perspectives. We performed odds-ratio statistical tests and multi-dimensional linear regression to show that reducing error of SSS or SST in the observation period leads to statistically significant improvement in prediction of future AMOC strength.

Odds-ratio statistical tests aimed to identify which of the observed variables is associated with a correct representation of the Atlantic meridional overturning circulation (AMOC). We used the posterior simulations directly, without the surrogate model. The analysis covered the N = 688 completed posterior simulations of the main experiment. Each simulation is scored against the true run. For each observed variable v, we computed the area-weighted root-mean-square error (RMSE) of the 1950-1999 mean field,

$$\mathrm{RMSE_v} = \left(\frac{\sum_j w_j \, \left(x_{v,j} - x^{\star}_{v,j} \right)^2}{\sum_j w_j} \right)^{1/2},$$

where $x^{\star}$ is the true run and j runs over the grid cells. The weights w_j are the ocean cell areas for SST and SSS and for the atmospheric fields. Geopotential height is evaluated at 500hPa. The AMOC is scored by the absolute error of the AMOC index, i.e. the maximum of the Atlantic overturning stream function between 20 and $70°N$ below 500m, averaged over 50 years. This is done for the present (1950-1999) and for two future periods (2550-2599 and 2950-2999). Each simulation is then classified as good or bad for each score. A simulation is good if its score is better than the criteria, such as the median, 25th quantile or 75th quantile over all simulations. This gives four binary predictors $g_v \in \{0,1\}$, one per observed variable, and two binary outcomes, y, for the present and the future AMOC and we computed the conditional probabilities $P(y = 1 \mid g_v = 1)$ and $P(y = 1 \mid g_v = 0)$. The observed variables are correlated across simulations, which means a simulation that reproduces SST well tends to reproduce SSS well too. We therefore estimated their effects jointly with the logistic regression

$$\mathrm{logit}\, P(y = 1 \mid g) = \beta_0 + \sum_v \beta_v \, g_v,$$

fitted separately for the present and the future outcome. The coefficients are maximum-likelihood estimates obtained by Newton-Raphson iteration. The odds ratio $\mathrm{OR_v} = \exp(\beta_v)$ is the factor by which the odds of a good AMOC change when variable v is good rather than bad, with the other three variables held fixed. $\mathrm{OR_v} > 1$ means that reproducing v well goes with a better AMOC. We report 95% confidence intervals $\exp\left(\widehat{\beta_v} \pm 1.96\, \mathrm{se_v}\right)$ and two-sided p values from the Wald statistic $\widehat{\beta_v}/\mathrm{se_v}$ with the standard errors taken from the inverse Fisher information. The p values are not corrected for multiple testing (four predictors and two outcomes).

We also did multi-dimensional linear regression to find which parameters control the AMOC within the posterior simulations. We regressed the AMOC index of each simulation on its 24 parameter values. Each parameter θ_i, in normalized units on [0, 1], is standardized over the simulations, $z_i = (\theta_i - \overline{\theta_i})/s_i$, with s_i its standard deviation across the posterior simulations.

A coefficient therefore gives the change of the AMOC index (in Sv) per one standard deviation of the parameter within the posterior. For each period we fit the multiple linear regression

$$A = \beta_0 + \sum_{i=1}^{24} \beta_i \, z_i + \varepsilon$$

by ordinary least squares, with all 24 parameters entered together. We give 95% confidence intervals $\widehat{\beta_i} \pm 1.96\,\widehat{\sigma_i}$ with the standard errors taken from the least-squares covariance $\widehat{\sigma^2}(Z^{\top}\boldsymbol{Z})^{-1}$ and $\widehat{\sigma^2}$ the residual variance. The skill of the regression is given by the coefficient of determination R^2 and its 5-fold cross-validated value.

From the results of odds-ratio statistical tests (Figure R5), we found that when it comes to AMOC reproduction or prediction, SSS observation improves the future AMOC prediction more than the other three observation variables, which is statistically significant (Figure R5 (d,e,f) and Figure R5 (g,h,i) independent of threshold selection of quantiles. Figure R5 (a,b,c) also suggested that improving SST reproduction leads to high accuracy of AMOC strength reproduction in the observation period. These results imply that fitting parameters to SST observation is likely to improve the current AMOC reproduction, while SSS observation leads to improvement in future AMOC prediction. We found that the posterior simulations of AMOC strength are strongly dependent on Parameter 16 and 20, which is precipitation adjustment parameter in North Atlantic Ocean and Scaling factor for controlling Bering Strait throughflow as shown in Figure R6. The parametric uncertainty reduction of these two parameters given by SSS observation is greater than those given the other three observation variables, which also supports that SSS observation contributes to improvement in the prediction of future AMOC strength. Note that this does not support that AMOC reproduction is well performed by assimilating SST observation than SSS observation, which requires further analysis by the next manuscript submission. These two supplementary results support the conclusion that assimilating SSS observation further reduces uncertainty in AMOC prediction than other three variables.

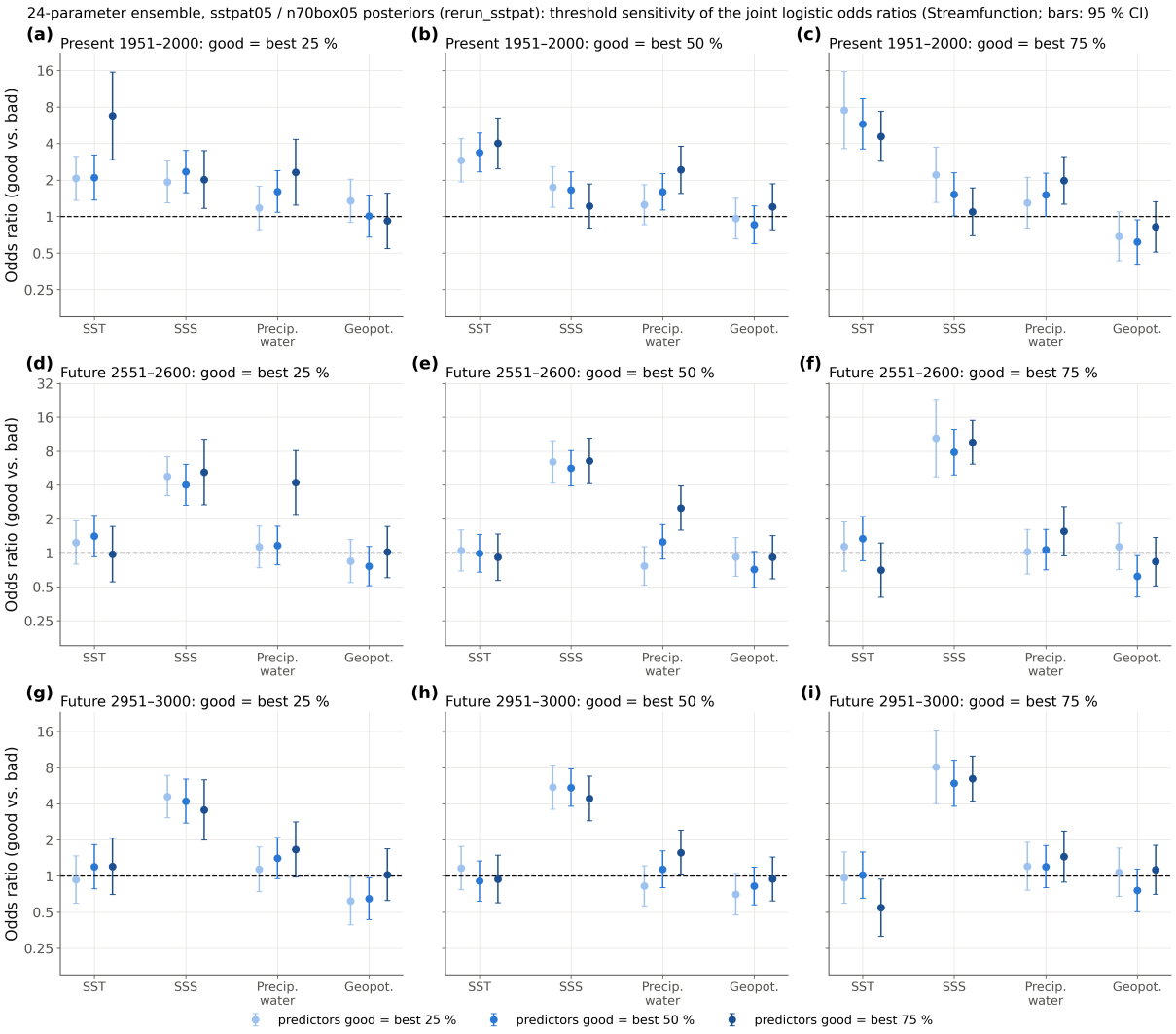


Figure R5. Sensitivity of the odds ratios to the thresholds that define good and bad runs. Joint logistic regression of the outcome on binary indicators of the 1950–1999 RMSE of SST, SSS, precipitable water and geopotential for the 688 completed posterior runs. The outcome is the absolute error of the AMOC index (maximum of the Atlantic stream function between 20 and 70°N below 500 m of the 50-year mean) in the present in (a–c) and the future (mean percentile rank of the 2550–2599 errors in (d–f) and 2950–2999 errors in (g–i). The outcome if its error is within the best 25 % (a, d), 50 % (b, e) or 75 % (c, f) of all runs. Light blue, blue and dark blue plots show the odds ratio of the four variables with different thresholds of Squared Errors, 25, 50 and 75 percentiles, respectively.

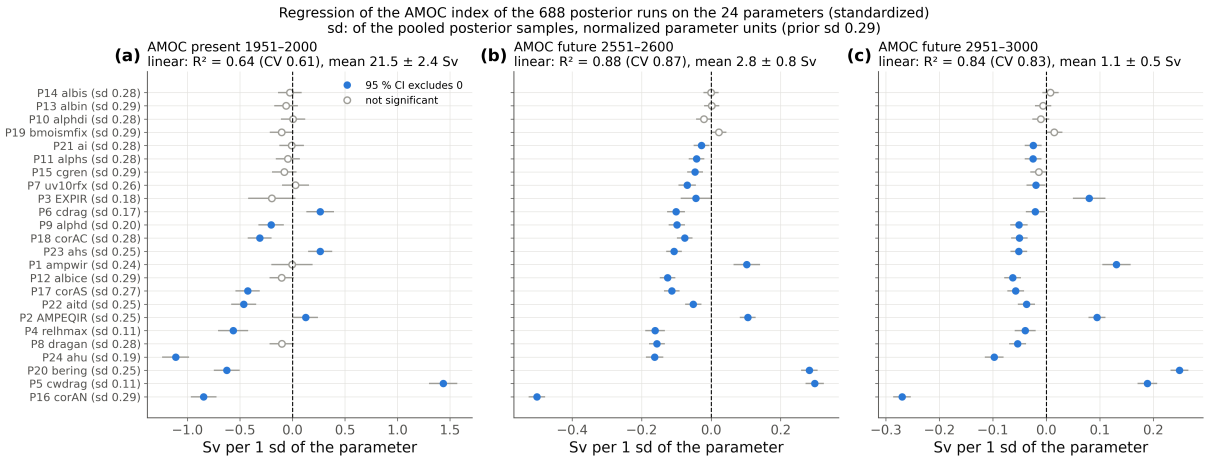


Figure R6. The parameter sensitivity to the posterior simulations of AMOC. The AMOC index, maximum of the Atlantic stream function between 20 and 70°N below 500 m of the 50-year mean of the 688 completed posterior is regressed on the 24 parameters, standardized by their mean and standard deviation. (a–c) Coefficients of a linear regression with 95 % confidence intervals; filled. The panel titles give R^2 and the 5-fold cross-validated R^2.

3. lines 137–155, in particular lines 150–155.

The manuscript states that retaining the dominant principal components 'effectively filters out internal variability that is independent of the parameters. However, PCA ranks spatial patterns according to variance; it does not distinguish parameter-induced variance from internal variability. A mode dominated by internal variability can in principle explain substantial variance, while a parameter-sensitive mode relevant for AMOC predictability could explain relatively little total variance.

The retained PCA modes determine the likelihood, and hence which parameter combinations can be constrained. It should therefore be quantified how the number of retained PCs affects the posterior and the SSS–SST comparison. Repeated simulations with identical parameters but different initial conditions would provide a much more direct way of separating internal variability from parameter-induced variability. If such simulations are unavailable, the limitations of the current PCA approach should at least be discussed explicitly.

We thank the reviewer for highlighting this important methodological aspect regarding the distinction between parameter-induced variance and internal variability in PCA. As the reviewer suggested, repeated simulations with identical parameters but different initial conditions are straightforward to reduce the effects from internal variability. However, such simulations increase the necessary simulations which are almost infeasible. To address this concern, we conducted additional sensitivity experiments testing different numbers of retained principal components ($n = 20, 30,$ and 40) and evaluated their impact on posterior parameter distributions and the comparison between SSS and SST.

Figure R7 showed increasing the number of retained principal components reduces MSEs of parameter estimation when the amplitude of observation noise is small. Under both SSS and SST observation scenarios, the overall MSEs of posterior parameter distribution are similarly reduced. When we examine improvement in the marginal MSEs along each parameter direction, as shown in Figure R8, the improvements of estimation performance differ between SSS and SST. While SSS improves the marginal MSEs of Parameter 11, 6 and 19, which represent albedo of melting snow, drag coefficients of the latent heat flux, and maximum evaporation factor over land, SST improve those of Parameter 17, precipitation adjustment parameter in the North Atlantic.

These results indicate that increasing the number of retained PCs does not change the results that SSS better reduces parameter uncertainty than SST when you evaluate the parameter estimation performance with MSEs. The conclusion that SSS observation improves more

accuracy of AMOC prediction is also supported by the answer to the comment 2. We kindly refer the reviewer to our response to Comment 2 for the detailed methodology, which establishes how our key conclusions remain statistically sound despite the inherent limitations of the surrogate model. We will revise the manuscript according to your comment to explicitly state that our method can be more robust to internal variability such as by introducing different normalization of covariance matrix before PCA, or by using ensemble simulations with identical parameters but different initial conditions.

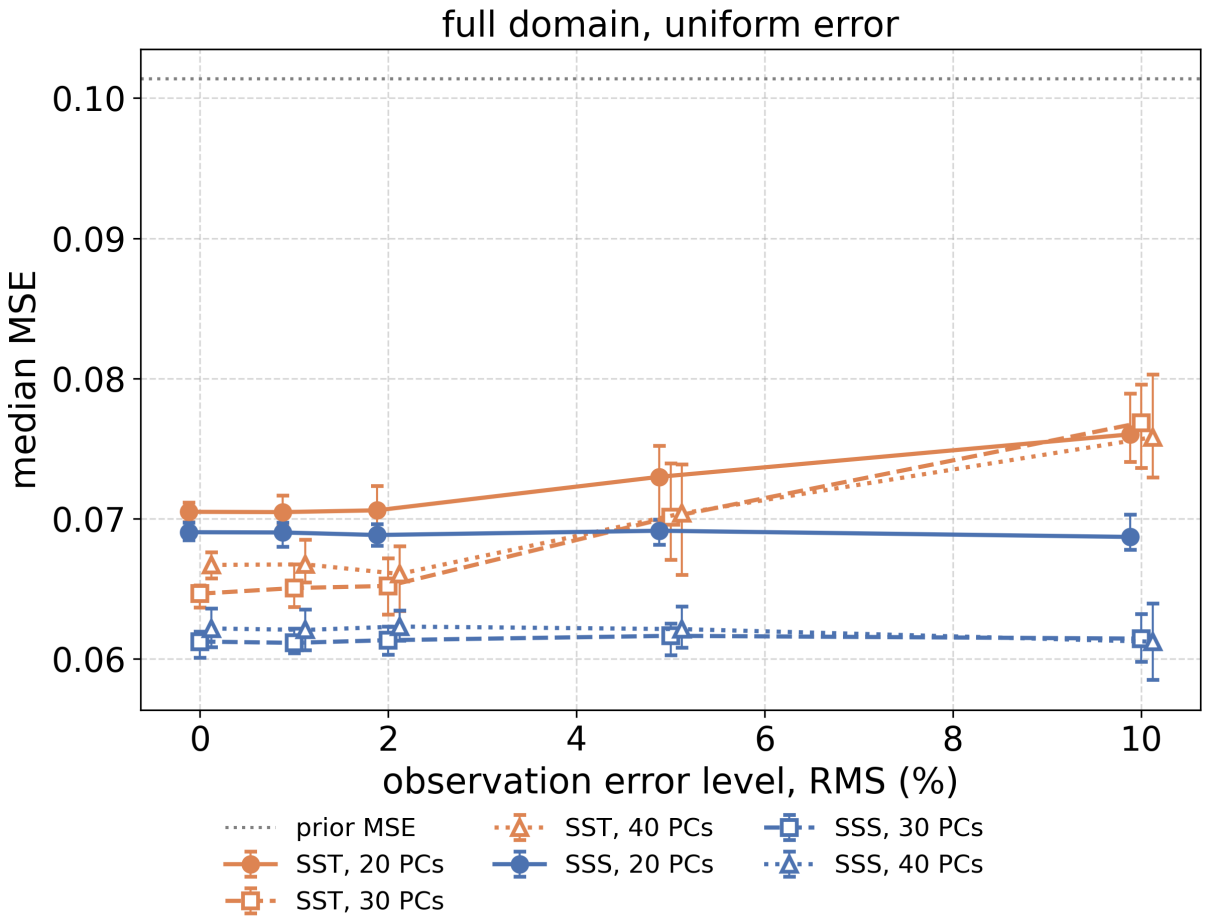


Figure R7. Sensitivity of MSEs of parameter estimation on the number of principal components. The horizontal axis is the amplitude of observation noise, while the vertical axis is median of MSE and their interquartile range.

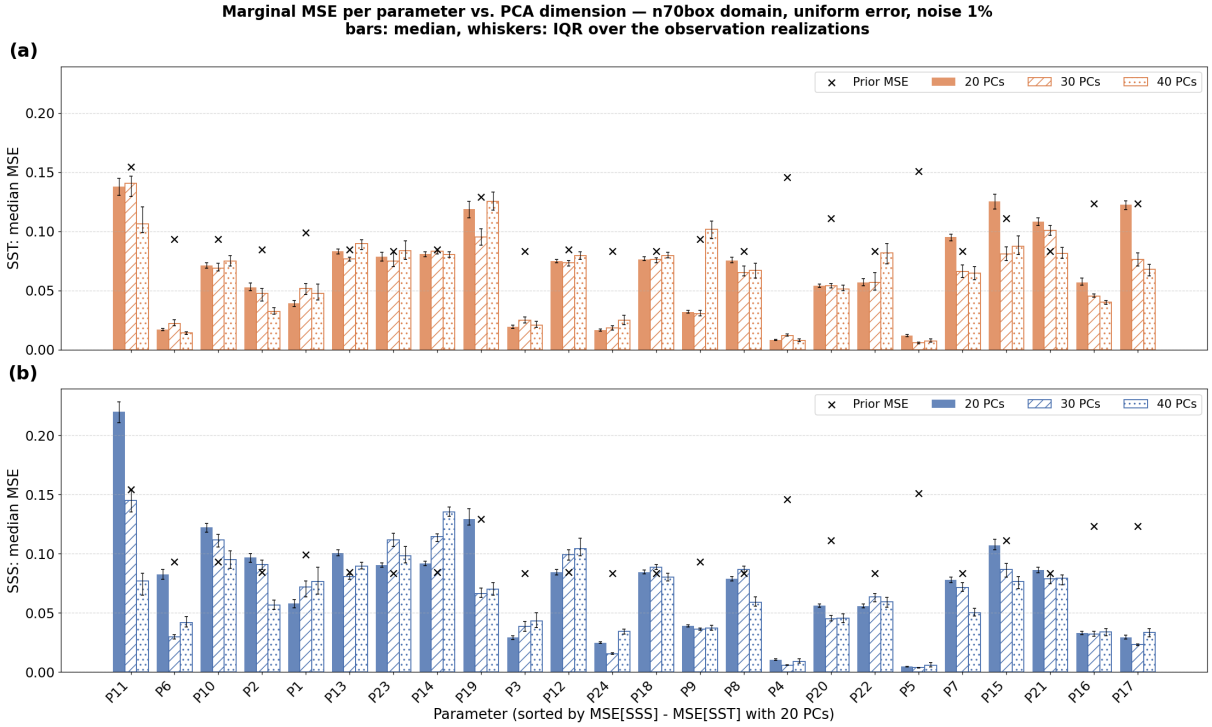


Figure R8. Comparison of the estimated parameter distributions of SSS and SST with different number of adopted principal components. (a) Marginal MSEs of posterior parameter distribution of SST observation, (b) Marginal MSEs of posterior parameter distribution of SSS observation. Orange and blue bar show the MSE with SST and SSS as an observation variable, respectively. Solid, hatched and dotted bars show the results with 20, 30 and 40 retained PCs, respectively. Parameters are ordered along the horizontal axis such that those farther to the right exhibit a smaller marginal MSE under SSS assimilation than under SST. Crosses show the prior marginal MSE. Intervals on each bar shows the range of marginal MSEs with synthetic true data with different initial conditions.

4. lines 267–274 and lines 485–503

The manuscript's explanation for why SSS outperforms SST relies partly on differences in signal-to-noise ratio across regions. This makes the assumed noise model directly relevant to the central result. If the noise normalization is changed, the relative information content of SST and SSS may also change.

The authors should either perform additional experiments with more realistic, spatially varying errors or substantially qualify the interpretation. At minimum, they should show that the ranking of SSS and SST is robust to alternative reasonable definitions of observational error and to regional rather than global assimilation.

We thank the reviewer for raising this important issue regarding the observational noise model and regional assimilation sensitivity. In response, we performed two sets of additional experiments: one evaluating the impact of spatially varying observation errors, and another assessing the effect of removing high-latitude observation data.

To evaluate whether our results depend on the assumption of spatially uniform noise, we incorporated a spatially varying observation error. As in Janjić et al. (2017), observation errors are strongly influenced by representation and sampling errors arising from coarse-graining observations onto climate model grids, which typically results in larger errors in regions with high subgrid-scale activity, such as western boundary currents (e.g., Oyashio) and the Antarctic Circumpolar Current (ACC). We adopted the temporal standard deviation map of SST as a reference spatial error pattern (Figure R9), scaling the error field so that the total observation error variance across all ocean grid cells remained identical to that of the uniform noise experiment. As shown in Figures R10(a–c), introducing spatially varying observation errors does not substantially alter the statistical performance of parameter estimation. Furthermore, Figure R11 demonstrates that the marginal parameter uncertainty reduction trends remain consistent between uniform and spatially varying noise setups. These results clearly indicate that the superior performance of SSS over SST in AMOC representation and prediction is robust and not an artifact of a simplified noise model.

To examine the spatial origin of the information provided by SSS, we conducted regional assimilation experiments by excluding observation data from the high-latitude Northern Hemisphere, where satellite SSS retrievals often suffer from higher uncertainty due to sea ice cover and low sea surface temperatures (Reul et al., 2020). As shown in Figures R12(a–c), removing high-latitude Northern Hemisphere observations degrades the bias, variance, and MSE of parameter estimation under SSS assimilation. A closer examination of individual

parameters (Figure R13) reveals that this performance degradation is primarily driven by Parameter 20, which controls the scaling factor for computing Bering Strait throughflow. Constraining Parameter 20 becomes markedly worse when high-latitude SSS data are excluded. This result implies that high-latitude SSS observations provide constraints on freshwater transport parameters governing North Atlantic salinity, highlighting mechanism through which SSS influences AMOC prediction.

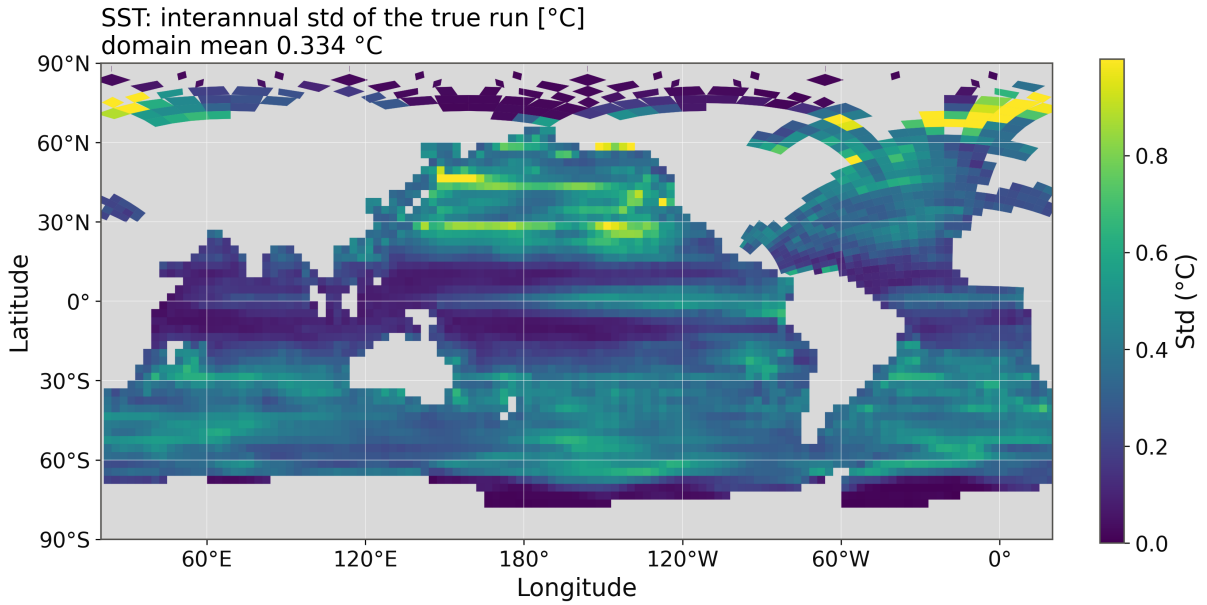


Figure R9. The distribution of standard deviations of SST interannual variabilities from 50 annual data from 1951 to 1999. Yellow means higher standard deviations on grids, while dark clue means lower standard deviations.

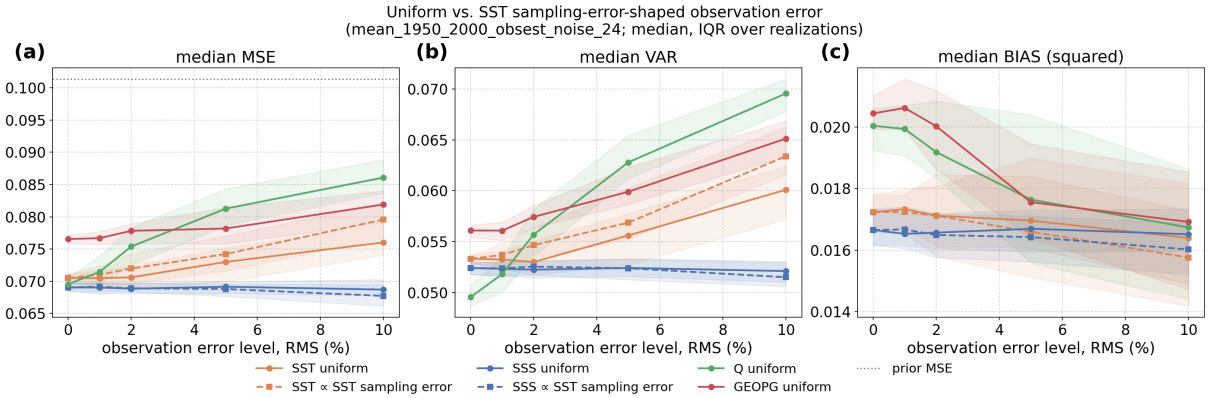


Figure R10. Sensitivity of parameter estimation accuracy and uncertainty to observation noise levels across different observation variable settings. The panel shows the mean squared error in its vertical axis. The orange, blue, green and red lines show the results with SST, SSS Q and GEOPG as observation variable, respectively. The horizonal axis is observation noise amplitude. The solid lines show the results with globally uniform observation noise, while the dashed lines show those with spatially varying observation noise. Note that we tested only SSS and SST observation with spatially varying observation noise.

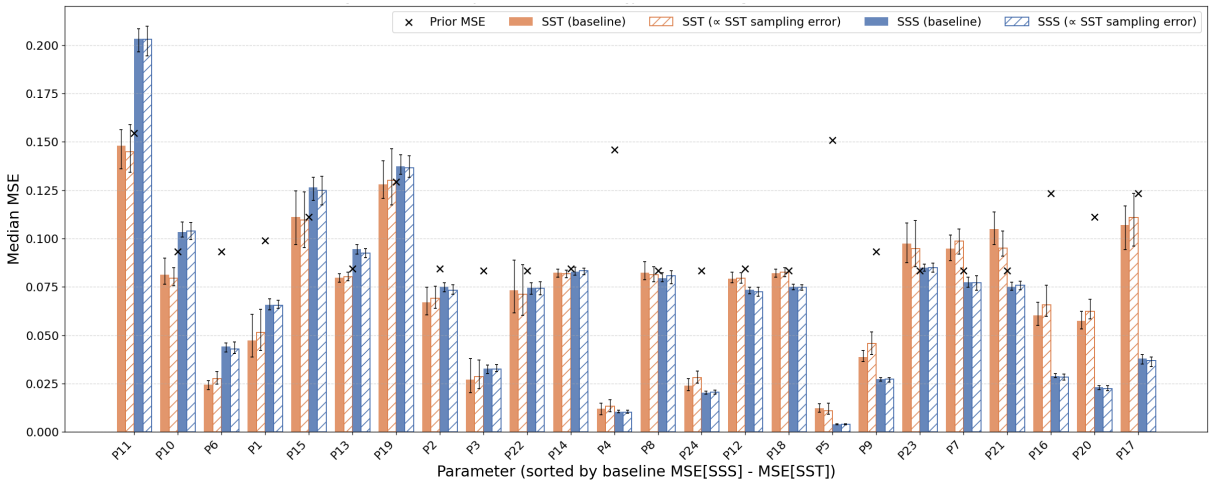


Figure R11. Comparison of the estimated parameter distributions of SSS with those with the other observation variables. Orange and blue bar show the marginal MSEs with SST and SSS as an observation variable, respectively. Parameters are ordered along the horizontal axis such that those farther to the right exhibit a smaller marginal MSE under SSS assimilation than under SST. Crosses show the prior marginal MSEs. Interval on each bar shows the range of marginal MSEs with different synthetic true data with different initial condition.

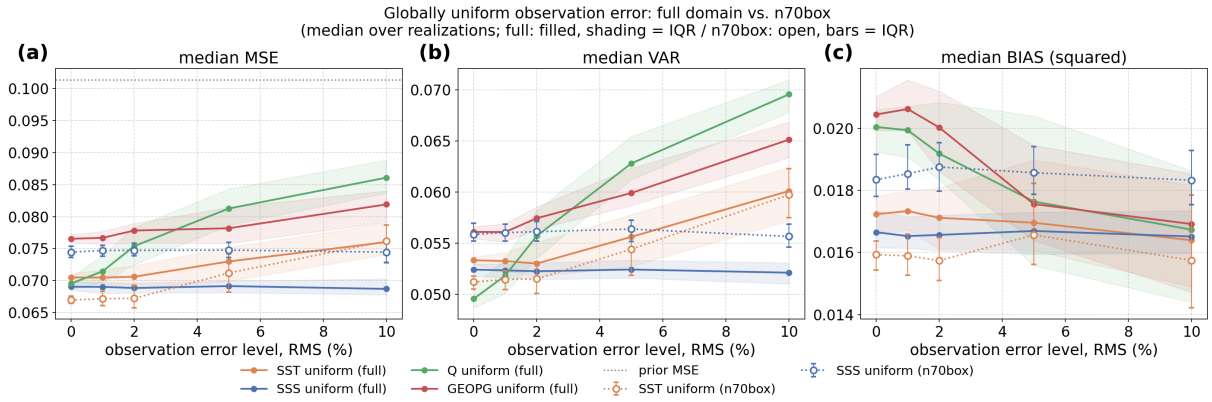


Figure R12. Sensitivity of parameter estimation accuracy and uncertainty to observation noise levels across different observation variable settings. The panel shows the mean squared error in its vertical axis. Orange, blue, green and red lines show the results with SST, SSS Q and GEOPG as observation variable, respectively. The horizonal axis is observation noise amplitude. Solid lines show the results with global observation, while dotted lines show those with observation except for the high latitude in North hemisphere. Note that we tested only SSS and SST observation with spatially varying observation noise.

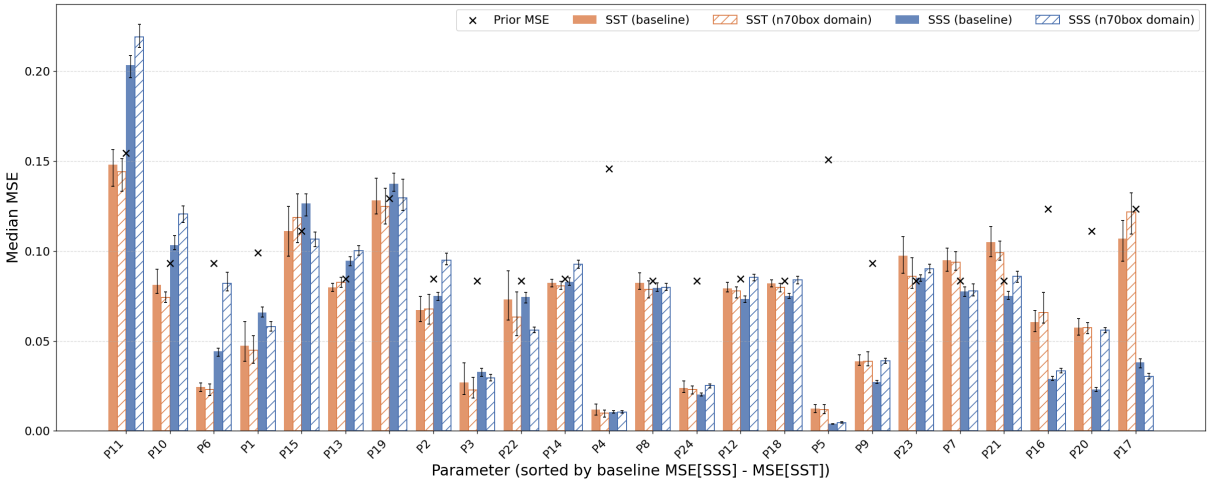


Figure R13. Comparison of the estimated parameter distributions of SSS and SST with different observation error settings. Orange and blue bars show the MSE with SST and SSS as an observation variable, respectively. Solid and hatched bars show the results with results with globally uniform error and spatially varying error, respectively. The parameters on the righter side have smaller the marginal MSE with SSS than the marginal MSE with SST. Crosses show the prior marginal MSE. Intervals on each bar show the range of marginal MSEs with different synthetic true data with different initial condition.

5. lines 467–472

The tipping time is defined as the first time at which AMOC strength falls below 2.2 Sv. However, it is not clear why 2.2 Sv represents a dynamical tipping point rather than an operational threshold for a strongly weakened AMOC. Under a prescribed, continuously increasing freshwater forcing, crossing a low AMOC threshold does not by itself demonstrate the existence of a bifurcation or an irreversible transition. This distinction is important because the manuscript repeatedly uses the terms 'tipping', 'collapse' and 'tipping predictability.

We appreciate the reviewer's comment that finds out problems in phrasing in the manuscript. As the reviewer's comments, our experiments are not dealing with "bifurcation" because of the limitation in the model. The experimental settings cause significant decline in AMOC strength by adding freshwater forcing in the North Atlantic, but never ensures that it is bifurcation, which in definition means that the system undergoes irreversible changes even after recovering the freshwater forcing to the original level. We referred to Kim et al, 2022 for their experimental settings, but they never verify that the freshwater forcing induces bifurcation in AMOC. Therefore, we will modify the expression in the main manuscript to use AMOC weakening instead of tipping or collapse.

6. lines 282–294, especially lines 288–294

For computational reasons, posterior LOVECLIM simulations are initialized from the spun-up state of the nearest prior parameter sample rather than being spun up independently. The authors assume that this state is sufficiently close to the equilibrium state associated with the posterior parameter vector. Given the long adjustment time scales of the ocean and the importance of small differences in AMOC state for the subsequent threshold-crossing time, this assumption could influence the results.

Please quantify the parameter-space distance between posterior samples and their selected initial-condition samples and validate this procedure by fully spinning up a representative subset of posterior simulations. The resulting differences in AMOC strength and 'tipping time' should be compared with the posterior uncertainty discussed in the manuscript.

We appreciate the reviewer's comment that the spin up might be insufficient in the series of experiments. We checked whether the spin up of the posterior simulations is sufficient or not. We found that the posterior simulations used in the original manuscript do not reach the steady state because of sufficient spin up period. Instead, we ran 2950-yr spin up simulations to obtain training samples and posterior simulations. We confirmed that the period of 2950 years is sufficient for the spin up since almost all of the posterior samples are successfully completed their spin up. FigureR15 (a,b,c) show that time series of ocean climatology including deep ocean states such as volume-mean ocean temperature, Deep-ocean temperature, and AMOC index are almost steady. Figure R15(e,f) show that drifts of ocean-heat uptakes and deep-ocean temperature of more than 99% of all the samples are within sufficiently small ranges. This shows the updated spin up is sufficient.

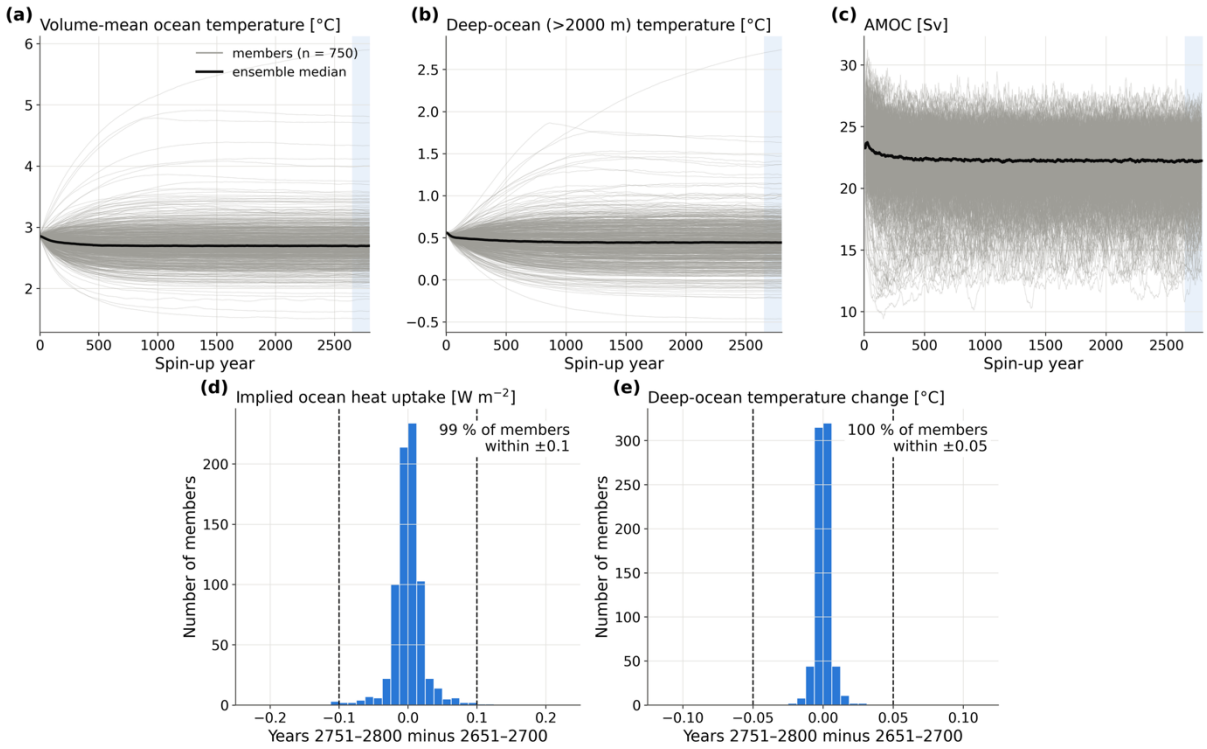


Figure R15. Spin-up of the posterior ensemble. (a) Volume-mean ocean temperature, (b) deep-ocean (>2000 m) temperature and (c) maximum Atlantic meridional overturning circulation (AMOC, 20-70°N, below 500 m) during the 2800-year spin-up of the 750 members (of 800) that completed it. Note that the number of successful posterior simulations is 688, which means 62 members collapsed after spin up. Grey lines show 10-year running means in (a, b), and 20-year in (c). Black lines show with the ensemble median. Shading marks the years used in (d, e). (d) Implied ocean heat uptake and (e) change in deep-ocean temperature between years 2651-2700 and 2751-2800. Dashed lines show the thresholds 0.1 W m-2 and 0.05 °C.

7. lines 10–20, lines 97–107, and lines 505–515

The finding that SSS is more informative than SST means that SSS is more informative within the assumed LOVECLIM parameter space and under the forcing and observation-error assumptions used here. It does not directly demonstrate that present-day SSS observations will constrain the real-world AMOC, or a CMIP-class model, to the same degree.

I suggest that the authors formulate the central conclusion explicitly in terms of potential information content in a perfect-model experiment and discuss how structural model error could alter the ranking of observational variables. This distinction is especially important because the manuscript contrasts its estimates with CMIP6 inter-model uncertainty and emergent constraints. Those uncertainty sources are fundamentally different and cannot simply be added or compared as if they represented equivalent probability distributions.

As the reviewer's comment, the experiment in the manuscript is a perfect-model experiment, and we should avoid discussion about comparison between our results and previous studies with CMIP-6 models and actual observation data ignoring the effects from structural model error. We removed the sentences about the quantitative comparison between those two studies out of the manuscript. Instead, we explicitly added that our uncertainty estimate is results with perfect-model assumption and idealized observation settings and structural model error could alter the ranking of observational variables. We also added the sentences into the manuscript that most of the previous studies of AMOC prediction uncertainty focused on structural model uncertainty, not so much on parametric uncertainty.

We will revise the 4th paragraph in section 5.1 into the following paragraph:

Bonan et al., 2025 described uncertainty in the prediction of AMOC strength in the CMIP6 model ensemble. They showed that the range of AMOC strength between the 10th and 90th percentiles in the original CMIP6 ensemble was approximately 14 Sv. They also showed that this interquartile range is reduced to approximately 3 Sv by their emergent constraint method. Although our uncertainty estimates are based on artificial assumption ignoring structural uncertainty and are not to be compared with the estimates in Bonan et al., 2025, there is a possibility that current CMIP6 and Emergent Constraint-based projection uncertainty quantification underestimates uncertainty in the future climate projection ignoring the parametric uncertainty.

8. lines 5–20 and Section 5.1, lines 505–519, especially lines 510–515

It is argued that the large parametric uncertainty found here implies that existing CMIP6-based estimates may underestimate the 'whole uncertainty' in AMOC projections. I do not think this conclusion follows directly from the presented results. The LOVECLIM prior parameter ranges are prescribed by the authors and generate a particular amount of spread. A large spread in this PPE does not demonstrate that CMIP6 models contain an additional uncertainty component of the same magnitude.

Moreover, structural and parametric uncertainties are not necessarily independent: part of what appears as structural uncertainty across CMIP models may itself arise from different parameter choices or parameterizations. The comparison with Bonan et al. (2025) is interesting but should therefore be presented as qualitative context rather than evidence that CMIP6 uncertainty has been underestimated by a particular amount. The relevant statements in Section 5.1 and the Abstract should be revised to make this distinction clear.

As in the answer to the major comment 7, we should avoid quantitative comparison between our experiments and previous studies using CMIP-6 models to assess structural error. We explicitly stated in the manuscript that the comparison between our estimate and Bonan et al., (2025) and other relevant studies is no comparable. Refer to the answer to the comment 7 for the details.

We will revise the original manuscript following the minor comments below.

Minor comments:

The manuscript would benefit substantially from careful English-language editing; I will not list any typos below.

Throughout: please use consistent capitalization for "Earth system model", "surrogate model", "Gaussian", "North Atlantic", and names of observational variables.

l. 1: Title suggestion: "Parametric uncertainty in predictions of AMOC weakening with an Earth system model of intermediate complexity."

Introduction: literature on parameter-sensitivity studies using EMICs/ESMs is missing here; for example, Boot and Dijkstra, "The effect of freshwater biases on AMOC stability across the model complexity spectrum", Clim. Dyn. 64, 409 (2026).

l. 19: "other" -> "several".

l. 20: delete "current AMOC and future".

l. 32: "deep ocean" -> "ocean current" (this is not the deep current).

l. 37: consider using the Lenton et al. (2025), Global Tipping Points Report 2025, reference; delete "current".

l. 56: "relationship" -> "constraint".

l. 62: after "projections", references are needed.

l. 63: "climate" -> "AMOC".

l. 94: "AMOC climate" -> "AMOC".

l. 99 ff.: LOVECLIM is more precisely described as an Earth system model of

intermediate complexity (EMIC); please use this terminology consistently rather

than implying equivalence with a comprehensive CMIP-class ESM.

l. 168: "maharanobis" -> "Mahalanobis".

l. 189: Please check Eq. (4). If Sigma_total = LL^T, the usual quadratic form involves solving Lz = (u_obs - m) and evaluating z^T z, rather than multiplying the residual by L. The notation as currently written appears inconsistent with the Cholesky factorization.

l. 192–215: Consider moving the detailed REMC description to an appendix, as it is quite technical and not needed to follow the main methodology. In particular, the 'temperature' terminology may be confusing in the context of this paper, so it should be explained more clearly.

l. 222: "freshwater is poured" -> "a freshwater flux is applied". Is any compensation applied to conserve global mean salinity?

l. 231–233: This is unclear; specify the CO2 ranges and/or radiative forcing.

Table 1: "Adjestment" -> "Adjustment" (for both P16 and P17).

l. 317: "k" -> "N"? Please check the notation.

Figure 3 is very difficult to interpret; please clarify substantially or omit it.

l. 381: The statement concerning P19 needs references.

l. 425–427: Unclear statement: please define what is meant by "thermohaline balance" here.

l. 662–668: duplicated reference; please check the reference list.