Reviewer 2

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. Please refer to the response to the reviewer 1 for their details.

1. Sections 2-3: I found Sections 2 and 3 challenging due to the abundance of technical content combined with the lack of a preliminary illustration of the methods. In other words, I think it would be easier for readers to digest the details were they provided with a broad picture of the whole methodological framework beforehand. The Abstract partially addresses this (L9-17), but I would suggest that a dedicated subsection be added at the beginning of Section 2. Including a description of the OSSE framework (similarly to Section 2.1 in Kubo and Sawada, 2025) and a brief mention of the model scenario would, in my opinion, make it easier to link with the building blocks which make up subsequent sections.

We sincerely thank the reviewer for this constructive feedback regarding the methodological complexity of our study. We recognize that integrating a diverse set of computational frameworks, including the LOVECLIM Earth system model, Gaussian Process Regression (GPR), Quasi-Monte Carlo sampling, and Replica Exchange Monte Carlo (REMC), presents an overarching workflow that requires clear structural guidance for the reader. To improve readability and logical coherence, we have thoroughly reorganized Sections 2 and 3 around a unified Observing System Simulation Experiment (OSSE) framework, following the general design of Kubo and Sawada (2025). Specifically, we have incorporated a new overview section (Section 2.1) that explicitly defines the experimental pipeline, accompanied by an illustrative schematic diagram (Figure 1 / Figure R1) to provide a visual roadmap of the sequential steps. Additionally, as detailed in our response to Reviewer 1, we have updated and clarified the validation procedure for posterior full-model simulations. We believe these structural additions make the paper's methodology far more straightforward to navigate.

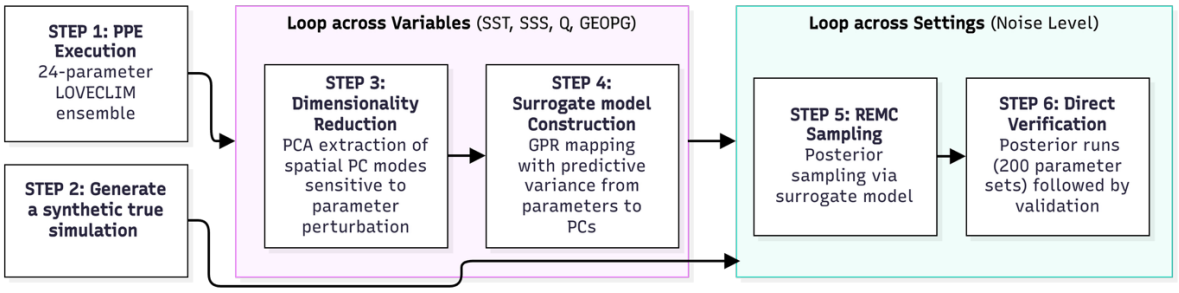


Figure R1. Schematic flowchart of the proposed Observing System Simulation Experiment (OSSE) framework for estimating 24 LOVECLIM parameters using Gaussian Process Regression (GPR) surrogates.

L137-155: The rationale for applying Principal Component Analysis (PCA) is clear and well explained. Nevertheless, I think the discussion may benefit from an improved explanation of the following points:

> *Dimensionality: in my understanding, C is a symmetric, real-valued, (NxN) matrix. Thus, its eigenvector decomposition is such that the matrix V is square (NxN), and not (NxM) as stated in the manuscript. Similarly, there are N eigenvalues ω, and not M as stated. If this comment is accepted, please make sure dimensions are consistent throughout, including whether the "x"s are row or column vectors.*
>
> *Standardisation: the rationale for standardising the data upon division by σ, as defined by the authors, is not explained in the manuscript. To my eye, the standardisation proposed is not equivalent to converting the covariance matrix C to a correlation matrix R, because in that case each variable x_n (with n=1, ..., N) is divided by the square root of its own variance, and not by that of a globally averaged variance. The point of performing PCA based on the correlation matrix is to weigh variables endowed with different variances equally (Wilks, Section 11.1.2), and I am not sure this is what the authors' transformation achieves. Here, this choice may be important as x contains information about multiple variables (L143), each potentially endowed with a different scale of variability. Finally, considering again that x contains information about multiple variables, I wonder if the formula for σ is dimensionally correct.*
>
> *It is not clear to me why σ is described as the "global mean of amplitude of annual fluctuations" (L147), considering that the authors stated earlier that x (and hence X) has been averaged over a period of 50 years (L131-132). The reference to annual mean variability seems to allude to the amplitude of random fluctuations imposed on nature run observations (L271): please clarify.*
>
> *The paragraph would benefit from a reference if available.*

We sincerely thank the reviewer for this detailed and insightful comment regarding the formulation of our Principal Component Analysis (PCA). We fully agree with the reviewer's observation regarding matrix dimensionality. The covariance matrix C is indeed a real symmetric $N \times N$ matrix, yielding N eigenvalues and an $N \times N$ eigenvector matrix V prior to truncating the system to the d leading modes. We have thoroughly revised Section 2.2 (in the new manuscript, section 2.3) to correct these dimensional representations, ensuring strict consistency in matrix dimensions (defining state vectors $\boldsymbol{x}$ as row vectors throughout).

Regarding data standardization, we appreciate the opportunity to clarify the physical and mathematical rationale for our normalization strategy. The primary reason we avoided grid-by-grid variance normalization, equivalent to constructing a correlation matrix, is that standardizing each grid point independently removes the spatial heterogeneity of climate

response amplitudes driven by parameter perturbations. In a parameter perturbation framework, grid points exhibiting larger variance carry critical dynamical signals regarding model sensitivity, normalizing by local variance would artificially inflate the influence of regions with low signal variance purely based on spatial extent. Instead, dividing by a global scaling factor σ (or conducting covariance-based PCA) preserves the relative energetic intensity and spatial structure of parameter-induced variability across the domain. Mathematically, scaling the field by a uniform scalar σ leaves the spatial eigenvectors (EOF patterns) entirely unchanged relative to unnormalized PCA, affecting only the overall magnitude of the eigenvalues. This approach aligns with established emulation methodologies for climate model ensembles, and we have added references to Dunbar et al. (2021) and Yarger et al. (2024) to support this design choice.

Finally, we acknowledge that describing σ in L147 with reference to "annual fluctuations" introduced unintended confusion with the observational noise definition in Section 2.3. To resolve this ambiguity, we removed the sentences where we define σ in the paragraph about PCA because the standardization process does not affect the results and we believe there is no need to show it. We have also verified the dimensional consistency of all equations in Section 2.2 (Section 2.3 in the revised manuscript) when multi-variable state vectors are constructed.

L156-164: Please add a reference for Gaussian Process Regression and explain why this specific model was chosen.

The selection of gaussian process regression is supported by Dunbar et al., 2021. One of the most prominent parts of GPR is that it returns prediction variance which reflects the uncertainty existing in the surrogate model itself. This uncertainty is necessary because in our setting, sparsity of training samples is one of the biggest bottlenecks controlling the accuracy of surrogate model. We should somehow consider uncertainty of surrogate model into the sampling process. We will revise the manuscript to explicitly say why we adopted GPR as a machine learning model for surrogate model. We will add the following sentences into the manuscript:

As a surrogate model, we employed Gaussian Process Regression (GPR) (Rasmussen and Williams., 2026), a non-parametric method that combines ridge regularization with the kernel trick to capture a non-linear relationship. GPR is suitable for this experiment where sparsity of training samples should affect the results substantially, because GPR-estimated prediction uncertainty can be directly used for the uncertainty of the surrogate model. GPR has been adopted as a surrogate model for ESM parameter uncertainty quantification in the previous studies (Dunbar et al., 2021, Yarger et al., 2024). We constructed a separate GPR model for each dimension j of the reduced subspace. The GPR model learns the mapping function F_j by minimizing the regularized squared error in the Reproducing Kernel Hilbert Space (RKHS). To evaluate the efficacy of the emulation, we computed the coefficient of determination (R^2) for each dimension. We optimized kernels and their hyperparameter on each latent dimension surrogate model (see section 3.3 for its details).

Section 2.2.3: This is perhaps the most challenging part of the manuscript: the main thrust of the argument comes across, but several aspects of the illustration could, in my opinion, be improved:

References would be most welcome for the Mahalanobis distance (L168; check spelling) and the Cholesky decomposition (L187). Also, the claim behind Equation 3 ("the following equation is proved by Bayes theory") could be backed up by a reference.

Since Equation 2 stems from Equation 3, I would consider presenting them in the reverse order. Furthermore, it is unclear why Equation 2 is introduced by "the likelihood function p(uobs|x) was modelled as the following equation" since it is in fact derived from Equation 3: my understanding is that the authors' modelling choice consists in the representation of the observation error, and that it is from its Normality that Equations 3 and 2 follow (please clarify if this is the case). Finally, I would move the mention to the Mahalanobis distance (currently L168) to where the likelihood is actually discussed.

For ease of reading, I would consider converting the in-line equation at L176 into a numbered equation, clearly identifying prior, likelihood, and posterior.

Please provide more details on the notation used for the right-hand side of Equation 1 and the dimensionality of the objects appearing therein.

Notation: the truncated matrix of eigenvectors is named Vd in Section 2.2.3, but is named V here. It is unclear why "x" appears at L177 and in Equation 4. Please check the dimensionality of matrices is consistent with Section 2.2.1.

We sincerely thank the reviewer for these thorough and constructive suggestions regarding Section 2.2.3. We fully agree that reorganizing the mathematical presentation, adding appropriate citations, and clarifying the underlying statistical assumptions will substantially improve the clarity and rigor of our methodological description.

To address the request for additional references and notation clarity, we will incorporate standard reference citations for the Mahalanobis distance (Mahalanobis, 1936), the Cholesky decomposition (Golub and Van Loan, 2013), and the Bayesian formulation underlying Equation 3 (Gelman et al., 2013), while also correcting the spelling of Mahalanobis at line 168. In addition, we will expand the description following Equation 1 to explicitly define all mathematical notations on its right-hand side and state the exact dimensionality of the state vectors, parameter vectors, and principal component coefficient matrices.

Regarding the mathematical flow of the likelihood derivation, we will reverse the order of Equations 2 and 3 in the revised manuscript as the reviewer's suggestion. We now explicitly introduce the formulation by stating that our fundamental modeling choice is the representation of the observational and surrogate model errors as zero-mean Gaussian random variables. From this Gaussian assumption, the formulation of the likelihood function follows naturally, rather than being introduced as an arbitrary independent model.

Finally, to enhance readability, we will convert the in-line equation at line 176 into a standalone, numbered display equation. In the text surrounding this updated equation, we will clearly label and describe each constituent term, explicitly identifying the prior distribution, the likelihood function, and the resulting posterior distribution to provide a transparent mathematical framework for the reader.

L265-270: How many synthetic observations from the nature run have been used to estimate the likelihood and derive the posterior parameter distribution? My reading from L265-266 and L269-270 is that 51 annual-mean observations were used, one per year for model years 1950–2000. Please clarify if this is indeed the case. If it is, u_obs represents (the latent-space representation of) an annual-mean variable in Equation 2, whereas m represents an estimate of the mean of the distribution of a 50 years-mean variable (L131-132). I wonder if the authors could comment on the pros and cons of combining variables derived with different choices of time aggregation.

We thank the reviewer for raising this point, which allows us to clarify a crucial aspect of our experimental design. We would like to clarify that there is no temporal mismatch between the synthetic observations (u_{obs})and the surrogate model predictions (m). In our formulation, the synthetic observation vector $u_{obs}$is not constructed from individual annual-mean values; rather, it is calculated directly as the 50-year temporal mean of the nature run over the 1950–1999 period. Consequently, both the synthetic observations and the model output m are defined on the identical 50-year time aggregation scale, ensuring complete temporal consistency throughout the likelihood evaluation in Equation 2. We acknowledge that the original wording in L265–270 created unintended ambiguity regarding how $u_{obs}$was derived. To prevent any future misunderstanding, we have thoroughly reworded lines 265–270 in the revised manuscript to state explicitly that u_{obs}represents the 50-year time-averaged state from the 1950–1999 nature run simulation.

L269-272: The authors opt to represent the amplitude of observation errors as a fraction of the "mean annual variability of all the grids".

The wording is unclear: is this the standard deviation of the 51 annual-mean samples along the time dimension, spatially averaged over all grid points?

The same observation error is applied to all grid points, implying that errors wield a larger impact at grid points where the variability is smaller. Figure S13 shows that the spatial patterns of signal-to-noise ratio are different for SSS and SST, especially at high latitudes: the authors argue this explains the superior performance of SSS compared to SST (L417). In my view, this challenges the remarks made in the conclusions (L502-503) and Short Summary (L28-29): the fact that SSS observations better constrain AMOC predictions is a direct consequence of the authors' modelling choices and does not straightforwardly apply to real-world data. Have the authors considered generalizing their definition to accommodate a spatially varying error matrix?

We thank the reviewer for this perceptive comment regarding our observational error formulation and its potential impact on the comparison between SSS and SST. To address the initial clarification request, the reviewer's interpretation is correct. The globally uniform observation error amplitude was calculated as the standard deviation of the 50 annual-mean nature run samples (1950–1999) along the time dimension using the synthetic true time series data, spatially averaged over all grid points. We will revise L269–272 in the updated manuscript to state this definition explicitly. We receive the similar comments from Reviewer 1. Please refer to the answer to comment 4 for its details.

We will revise the manuscript referring to the following technical corrections.

Technical corrections

L36 “European and global”

L46-50: This paragraph may give the impression that CESM data will be used for subsequent analysis, which is not the case. Please clarify.

L51-60: I suggest adding a 1-line explanation of how Emergent Constraint works at the beginning of the paragraph.

L95 Keetz et al., 2025 is not in the References.

L104-106: Please clarify which opinion Carslaw et al., 2026 expresses.

L118-120: Please clarify how the active core components considered in this study were selected, if relevant in relation to Kim et al. 2022. I would also consider adding a brief explanation of why the terrestrial vegetation component is necessary.

Table 1: it is mentioned at L238 that “All the parameters are standardised into [0, 1] in our experiments”, but that is not the case in Table 1.

L302-304: Please avoid conflicting notation with previous equations

L325: “coefficients of determination” instead of “determinant coefficients”

L325: “5-fold cross-validation” instead of “5-folded Cross validation”

Figure 3: This plot is very hard to read because the details are small and get blurred if one zooms over the figure (or prints it). I recommend that the authors find an alternative visualisation strategy. Since the two-dimensional marginal distributions are not explicitly referenced in the text, perhaps the authors may consider showing the one-dimensional distributions only.

Figure 4: This figure contains elements to which I struggle to attribute an obvious meaning, namely “(w/ IQR)” and “(All)” in the titles and “Combo” in the legend. Also, I wonder if the title is necessary at all given it is the same for all panels, and the information it delivers could be moved to the caption avoiding repetition. This last point applies to other figures, e.g. Figure 6.

Figure 4 (and anywhere an envelope is shown): Please explain what the envelope is.

Figure 4: it is not explained why bias decreases with noise for SSS. Also, panel 4c is not discussed.

L586: I wonder if the comment about figures having been updated is necessary here.

Grammar: L42-43; Syntax: L12, L42-43, L43-44, L48, L72, L112, L399-400, L454-455; Wording: L93 (“a lot”), L159 (“kernel trick”), L471 (“not so much”)