## Author’s response

We would like to thank the editor and the two reviewers for their careful reading of our manuscript and for their constructive comments. We have revised the manuscript accordingly.

The responses below are organized by reviewer. For each reviewer, the comments are reproduced in bold text and followed by our point-by-point responses. Additional figures supporting some of our responses are provided at the end of the corresponding response section.

## Response to RC1: ‘Comment on egusphere-2026-1337’, Anonymous Referee #1, 25 Apr 2026

### General

**This study analyzes the limitations of the LIM-EOF model in traditional paleoclimate reconstruction methods. To address these issues, an improved climate model simulator is proposed, targeting the structural deficiencies of the traditional method with three specific improvements. The study evaluates whether these innovative measures enhance the simulation accuracy of large-scale climate variability, improve predictive capability, and reduce simulation errors in extreme events. The research is detailed, thoroughly elaborating on the dimensionality reduction algorithm and prediction model, describing the architecture and implementation of the simulator, and using extensive datasets and climate models to evaluate its multifaceted performance. Simulations based on the CMIP6 model suite demonstrate that the proposed improvements significantly outperform the traditional LIM-EOF model in terms of prediction accuracy, dynamic performance, avoidance of error accumulation, etc. In addition to a detailed analysis of the advantages of these improvements, the study also discusses certain limitations of the simulator under specific conditions, while proposing directions for further research. The work contributes notably to enhancing the accuracy of paleoclimate reconstruction and holds strong exploratory significance for advancing various research methods in the field.**

### RC1: Reply

We would like to thank the reviewer for their helpful comments and feedback on our manuscript. Below are the reviewer comments in bold text followed by our response. Additional figures supporting our responses are provided at the end of this document.

**R1C1: When analyzing the correlation matrix and variance structure of latent vectors from EOF and AE, why are both 30-dimensional and 16-dimensional analyses conducted?**

**ANS:** The dimensionality of the latent space used in a particle filter must be carefully calibrated: it should retain a sufficiently large fraction of the climate variability to limit particle degeneracy while remaining compact enough to avoid computational costs associated with high dimension problems. In line with this requirement, Jebri and Khodri (2023) used a truncated 30-dimensional space, which explains approximately 70 % of the variance. We therefore first conducted the analysis using 30 latent vectors, in order to remain consistent with this previous reference configuration.

However, as shown in Fig. 2, performance across most metrics tends to plateau relatively quickly. A 16-dimensional representation already provides good overall skill, suggesting that it offers a suitable compromise: it preserves substantial predictive ability while keeping the latent space sufficiently low-dimensional for future integration into particle-filter framework.

**R1C2: To assess the impact of differences in physical mechanisms among models on simulation results, the study uses surface temperature data from 52 CMIP6 models. What aspects were primarily considered in selecting these data, and what characteristics make these data suitable for the evaluation work in this study?**

**ANS:** To test the robustness of our emulators and to perform the sensitivity experiments, we applied them to a broad CMIP6 multi-model ensemble. We selected models from all CMIP6 modelling institutions represented in the available ESGF archive, using practical availability criteria: each retained model had to provide monthly near-surface air temperature (tas) for both a piControl simulation and at least one 165-year historical run. This selection provides substantial physical diversity, as illustrated in Fig. A11 by the large inter-model differences in the tropical Pacific nonlinearity coefficient α. A few models had to be excluded because of missing data or inconsistencies in the available files. Nevertheless, the final dataset preserves institutional coverage, with at least one model retained from each CMIP6 modelling institution initially considered. We also thank the reviewer for this question, as it led us to identify duplicated entries at the end of Table A1. This issue has been corrected in the revised manuscript.

We also clarify that a more restricted subset of models was used for the specific analysis addressing the question: "Can RC-based emulators adapt to models with varying degrees of nonlinearity?" This subset was not selected on the basis of predefined physical characteristics, but according to a practical requirement: the models had to provide a sufficiently large training dataset for the RC-based emulators. As shown in Fig. 6, RC skill is strongly sensitive to training-set size, and RC-based emulators require substantially more training data than LIM to reach stable performance. We therefore retained models with at least 10 historical ensemble members for this analysis, in order to avoid interpreting undertrained RC behavior as a physical model-dependent effect.

The subsequent comparison between IPSL-CM6A-LR and MIROC6 was used as an illustrative contrast between a model with relatively low ENSO nonlinearity, as measured by the α coefficient, and a model with higher ENSO nonlinearity.

**R1C3: Could you summarize which specific CMIP6 models show significant improvements in prediction accuracy and strong adaptability due to the proposed improvements in this study, and provide the relevant constraints?**

**ANS:** The first point is that, regardless of the CMIP6 model considered, AERCn provides significant skill improvements when the training dataset is sufficiently large. For example, among the eight models with 20 historical runs available for training, all show improved skill with AERCn compared with LIM (Fig R1.1). The spatial patterns of improvement, however, differ across models and appear to be related to model-specific predictability. For instance, IPSL-CM6A-LR, ACCESS-ESM1-5, and CanESM5 show broad AERCn skill gains over the tropical Pacific, especially in its eastern part, whereas in MIROC6, MIROC-ES2L, and the GISS models, the improvement is more localized over the western tropical Pacific.

Nevertheless, we do not conclude that AERCn is only applicable to a specific subset of CMIP6 models. Rather, we consider that AERCn can in principle be applied to any climate model, but that the construction of the training dataset is critical. Beyond having enough training data, it is preferable to train on past2k simulations, as done for IPSL-CM6A-LR, so that the emulator learns under forcing conditions closer to the target reconstruction period.

Finally, our results suggest that, across models, the largest source of predictability comes from the AE-based dimensionality reduction. This confirms that the way the climate state is represented is a key component of the proposed improvement over the classical LIM-EOF framework.

**Supplementary figures for Reviewer 1**

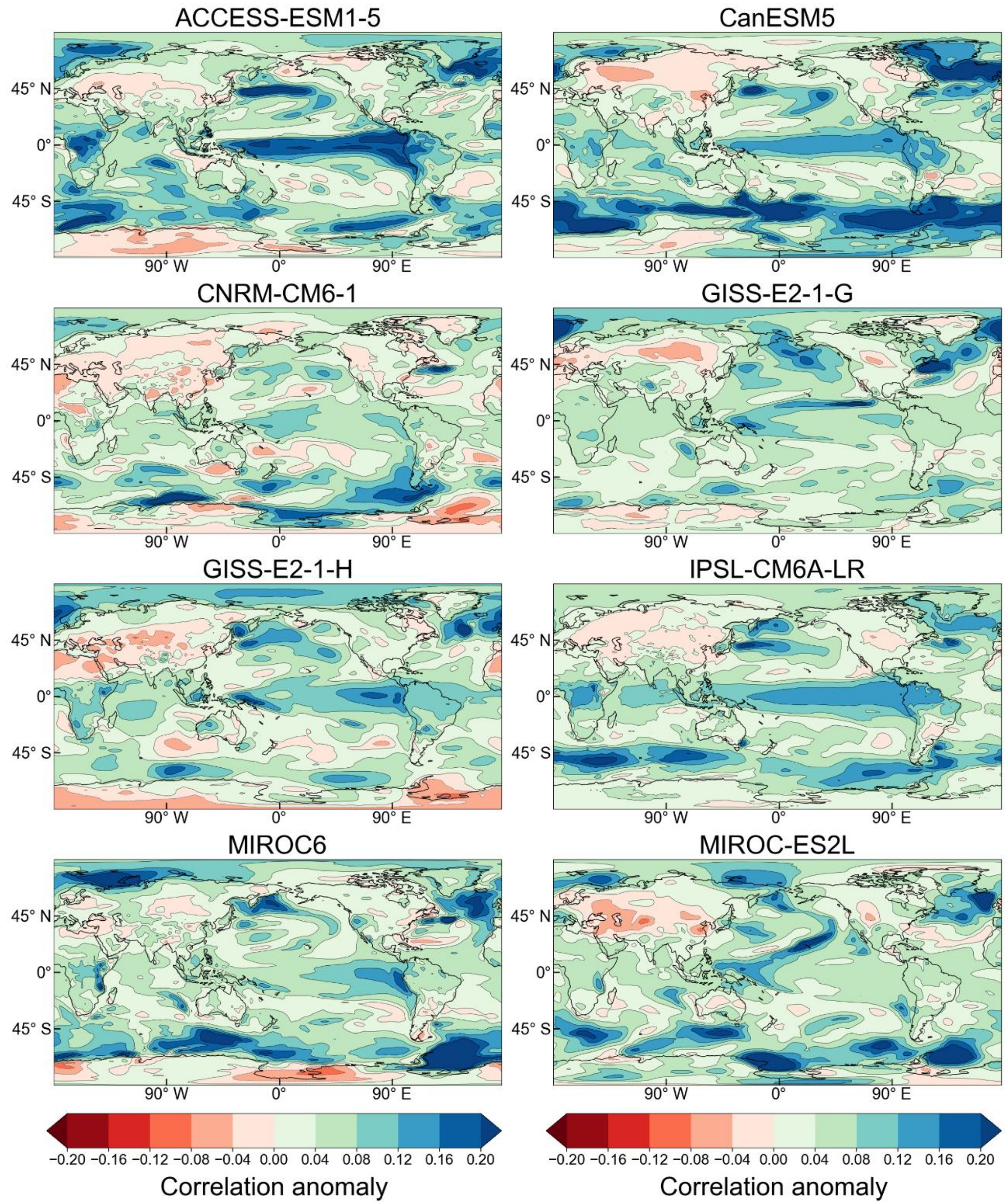


**Figure R1.1: Spatial differences in one-year lead correlation skill between AERCn and LIM for the eight CMIP6 models with 20 training members. Positive values indicate regions where AERCn outperforms LIM.**

## Response to RC2: 'Comment on egusphere-2026-1337', Anonymous Referee #2, 27 May 2026

**General**

**Authors use different setups with classical and novel techniques related to emulators for reconstruction of climate indices for past periods, specifically related to ENSO. For the setup of the emulators the CMIP6 suite of Earth System models covering the past two millennia is used with a specific focus on the IPSL model. The manuscript is very clear and concisely written and the methodological steps are described in detail, including formal descriptions for reproduction of results.**

**I suggest publication of the manuscript after some modifications and clarifications listed further below are addressed.**

**RC2: Reply**

We would like to thank the reviewer for their helpful comments and feedback on our manuscript. Below are the reviewer comments in bold text followed by our response. Additional figures supporting our responses are provided at the end of this document.

**R2C1: A conceptual question is, whether the study presents a full climate model emulator. In the present form statistical models are presented for key parameters or indices of climate modes of variability. In itself this is very important but should somehow be reflected in the title.**

**ANS:** We thank the reviewer for this useful clarification. We agree that the scope of the emulator should be stated more explicitly. In this first application we do not emulate the full multivariate coupled climate system. However, they are not restricted to a few climate indices either: they predict annual spatial tas fields, while indices such as Niño3.4, AMV and GMST are used as diagnostics of the predicted fields. We have clarified this point in the revised manuscript by explicitly presenting the study as a proof of concept focused on surface air temperature fields, with possible future extensions to multivariate emulation. The corresponding changes were made in the Abstract (l. 10 and l. 19 ff.) and in the Conclusion (l. 781 ff.).

**R2C2: The abstract is well written concerning the main methodological issues. Still it would be helpful to provide more concrete information on how the emulator outperforms traditional concepts and some critical comments of the new emulators.**

**ANS:** We have revised the Abstract to specify more explicitly how AERCn outperforms traditional concepts (l. 11 ff). In addition, the changes introduced in response to R2C1 clarify the main limitation and perspective of the present study, namely that it should be viewed as a proof of concept focused on tas fields, with future extensions toward multivariate climate-field emulation (l. 19 ff.).

**R2C3: ll 33ff: The authors should already mention here the two conceptually different approaches within the data assimilation process. The online approach that is typically used for present-day applications and the offline used for paleo applications.**

**ANS:** We thank the reviewer for this suggestion. We have added a short clarification in the introduction distinguishing offline and online PDA approaches (l. 33ff).

**R2C4: l. 75: It is true that the individual EOFs do not necessarily represent physical meaningful processes. However, in the combined use the EOFs (together with their principal components) still represent a very large part of the full state-space vector, also including potential non linear effects. Therefore one should point to the fact that it is important to know how the EOF concept is introduced and used within a prediction (or reconstruction) approach.**

**ANS:** We thank the reviewer for this clarification. We agree that EOFs, used together with their PCs, can represent a large part of the state space, and that the ability to reproduce nonlinear behaviour also depends on the emulator applied afterward in this reduced space. This is indeed illustrated in our study by the improvement obtained when replacing the LIM with RC, without changing the dimensioning reduction method. Nevertheless, the EOF basis itself remains constrained, as it relies on a stationarity assumption and on a linear, orthogonal, variance-based decomposition. Because of these constraints, EOFs have known limitations in identifying true physical modes (Dommenget and Latif, 2002). More importantly for our purpose, explained variance is not necessarily aligned with predictability: low-variance components may contain important predictive information, as shown in the principal-component regression literature (Jolliffe, 1982). We have therefore revised the paragraph to clarify this point (l. 75ff).

**R2C5: l. 131 The term “filtering” is misleading. In fact the Eigenvectors are just a re-ordering of the original covariance matrix related to variance. The truncation – and hence what is defined as “noise” – is a somewhat subjective decision. For instances, in cases where it s important that original fields can be re-constructed just based on the EOFs, the addition of higher indexed (and more “noisy”) EOFs is very important. This allows a more realistic representation of the original field, being important for the generation and prediction of extreme events.**

**ANS:** We agree with the reviewer that EOF truncation is partly subjective and that discarded higher-order modes should not be interpreted directly as noise. Although criteria such as North et al. (1982) can be used to assess whether successive EOFs are statistically distinguishable, our choice of latent dimension is primarily guided by the trade-off between compactness and predictive skill. We have therefore replaced the sentence “In practice, truncation to the leading modes enables efficient representation and prediction while filtering small-scale noise” by “In practice, truncation to the leading modes provides a compact and efficient representation of the data, while retaining enough information for prediction.” (l. 137).

**R2C6: l. 121: Usually the Eigenvectors are not based on standardized variables. In the standard approach area-weighted anomaly fields are used. Using standardized fields is only be applied when different variables/units are used in a joint EOF analysis. Otherwise the spatial EOF patterns might be substantially different, because all grid points show the same amount of variance by definition in the normalized case. In the case standardized variables are used for the calculation of the covariance matrix an explanation should be provided why variables are standardized prior to the calculation of EOFs.**

**ANS:** Indeed, EOFs are usually computed from area-weighted anomaly fields without prior standardization. In our case, the total fraction of explained variance is very similar with and without standardization, although the resulting spatial patterns differ slightly (Fig. R2.1). However, when the EOF space is used for prediction, the LIM skill is lower without standardization, especially over the tropics and the Southern Hemisphere, as shown in Fig. R2.2. This is also confirmed by the full skill comparison reported in Table R2.1: the LIM based on standardized EOFs, as used in the manuscript, gives better overall performance. The only slight loss of skill occurs for the North Atlantic/AMV region, which is also a region where the IPSL model exhibits strong autocorrelation. Overall, these diagnostics support our choice of using standardized input fields for the EOF-based LIM benchmark. We have therefore added a sentence in Sect. 2.1.1 to justify the use of standardized fields before EOF computation (l. 127ff).

**R2C7: l. 132: Here it should be noted whether a spatial or temporal prediction is meant. Again, using EOFs still can also reproduce non-linear effects, depending on the algorithm that is eventually used for prediction (e.g. Analog method (Zorita and von Storch, 1998) or any other non-linear method). The real challenge is how the information contained within the EOFs is used for setting up such a model and if there is any predictive skill included in the temporal structure (i.e. red noise or low-frequency variability).**

**This difference should be better elaborated: For the temporal prediction it is not the (spatial) EOF or eigenvectors that are preventing a better prediction per se. It is rather the method that is used for temporal training/validation and eventual prediction of the process or variable under consideration. EOFs in this context are usually used for spatial dimensionality reduction rather than for temporal.**

**ANS:** We thank the reviewer for this clarification. EOFs are indeed used here to reduce the spatial dimension, and a second algorithm then performs the prediction from this reduced space. A nonlinear prediction algorithm can therefore be applied to EOFs, as we do with the RC configuration. Our point is that the way the dimensionality reduction is performed already conditions the skill of the emulator used afterward. We show that, using the same prediction algorithm (RCn), predictive skill increases substantially simply by changing the dimensionality-reduction method. This is why, in our case, the AE is trained to encode information from both t and t+1, whereas EOFs are designed to provide an optimal variance-based representation of the field at time t. We have therefore revised the paragraph to clarify this point (l. 140ff).

**R2C8: l. 145: The authors state that for AE the temporal information is already implicitly used for setting up a prediction model using temporal information. This is not the case in the standard EOF. A way EOFs could be used in this way is so called "Extended EOFs" where a sequence of EOF patterns might be used to setup an improved prediction scheme.**

**ANS:** We thank the reviewer for this useful clarification. We agree that the EOF reference used in our study is a standard spatial EOF decomposition, whereas the AE formulation introduces temporal information in the learning objective. Extended EOFs can indeed include temporal information by applying EOF analysis to sequences of fields, typically combining current and past states (Weare and Nasstrom, 1982). However, this remains different from our t/t+1 AE formulation: Extended EOFs constrain the reduced representation using a sequence of fields at current and previous times, whereas our AE constrains the representation of the field at time t using the future field at t+1 as a training target. This encourages the latent representation to retain predictability-relevant information. Our approach focuses on retaining the source of predictability.

**R2C9: l. 169: Using white noise as stochastic component is of course an option, especially for atmospheric variables with only little or no memory. Is it possible also to include other noise terms and when would it be advisable ? Maybe the algorithm can use some a-priori information based on the training data (in this case tas ?)**

**ANS:** We thank the reviewer for this suggestion. In the present study, we use the classical LIM–EOF formulation as a benchmark, in which the stochastic component is represented as Gaussian white noise. Other stochastic parameterizations could indeed be considered. For instance, recent extensions such as Colored-LIM replace white noise with temporally correlated noise in order to represent additional memory effects (Lien et al., 2025). In the present study, however, our objective is to compare the proposed emulators against the classical and widely used LIM–EOF benchmark. We therefore retain the standard Gaussian white-noise formulation, while noting that colored-noise extensions would be a relevant direction for future work, especially for Particle Filter applications.

**R2C10: l 262: Since ENSO is a target variable of the manuscript I suggest to include a figure with the frequency spectrum of observed ENSO (Based on ERA/NCEP or similar) together with the one of IPSL including the projection of the Nino3 index on the tropical SSTs.**

**ANS:** We thank the reviewer for this suggestion. We computed additional ENSO diagnostics comparing IPSL-CM6A-LR with ERA5 and have added them to the Appendix. First, we compared the frequency spectrum of the annual Niño3.4 index in ERA5 and in the IPSL-CM6A-LR past2k ensemble (Fig. R2.3). The IPSL ensemble shows spectral peaks in a similar interannual range to ERA5, with most of the power concentrated within the canonical ENSO band. Second, we regressed annual surface temperature anomalies onto the Niño3.4 index for ERA5 and IPSL-CM6A-LR (Fig. R2.4). This diagnostic shows that IPSL-CM6A-LR produces a warm anomaly that extends slightly too far westward over the tropical Pacific, consistent with the model bias already identified by Boucher et al., (2020). We have also added a short discussion of these diagnostics in Sect. 4.1 of the revised manuscript.

**R2C11: l. 304: Could you state how results might deviate using other sources of noise (e.g red Noise with different memory ?)**

**ANS:** We thank the reviewer for this useful comment. In this study, ensemble predictions were generated by perturbing the initial state. Other formulations would indeed be possible, and we tested several alternatives, including the addition of random perturbations within the latent space, as well as output-space noise whose covariance was estimated from prediction errors, in a spirit similar to the stochastic component of the LIM.

The latent-space perturbations give probabilistic diagnostics that remain close to those obtained with input-space perturbations for Niño3.4, but lead to a more peaked rank histogram for GMST. The output-noise formulation performs less well for Niño3.4, with a rank histogram that is too depleted in the central bins and substantially higher CRPS values. Overall, the input-space perturbations provide the best compromise across both indices (Fig. R2.5). We therefore suggest to add this sensitivity diagnostic to the revised manuscript as Fig. A15. We also revised the method paragraph in Sect. 4.2.3 and the discussion of the ensemble approach in Sect. 5.3.2 to clarify why input-space perturbations were retained.

Furthermore, we agree that introducing red-noise perturbations would be an interesting extension. However, the emulators are evaluated here outside the full particle-filter framework. In the intended application, particles would be resampled at each time step, and in some cases only a very small fraction of the propagated trajectories would be retained. Consequently, diagnostics computed from free-running ensemble trajectories without resampling would not necessarily represent the effective ensemble distribution obtained within the filter.

We therefore retained the classical initial-state perturbation approach to generate prediction ensembles with our emulators, as it provides a simple and robust baseline and leads to satisfactory probabilistic diagnostics. We nevertheless agree that the choice of stochastic forcing is an important direction for future improvement, in particular through perturbations that depend more explicitly on the current state of the system and on the associated uncertainties.

**R2C12: ll 345 ff: Authors could also use the Brier skill score (Wilks, 2010) including both correlation and variance information**

**ANS:** We thank the reviewer for this suggestion. The Brier skill score is indeed useful metric for evaluating probabilistic forecasts including both correlation and variance information (Wilks, 2010). In our case, however, the ensemble forecasts are issued for continuous climate variables. We therefore use the CRPS, which is explicitly presented as a generalization of the Brier score to continuous variables (Hersbach, 2000).

**R2C13: ll. 380 ff: The fact that the variance patterns show this specific structure might be related to the fact that the variables are normalized to unit variance before entering the dimension reduction. Patterns with original units might show a substantial different structure.**

**The comparison between EOF and AE could of course be carried out. I just wonder what is really learned given the (fundamental) different construction and interpretation of the individual results or patterns. I suggest to shorten the entire section and summarize the most important conclusion without being too speculative on potential (physical) connections.**

**ANS:** We agree that the standardization affects the spatial patterns of explained variance, as shown in Fig. R2.1. Without standardization, the spatial patterns associated with both EOFs and AE latent vectors are also modified. However, the comparison of EOF/PC and AE/LV patterns and time series without standardization shows similar correlations between the leading PCs and the AE dimensions explaining the largest fraction of variance (Fig. R2.6). This suggests that the main conclusion is not driven by standardization: the dominant AE dimensions remain strongly connected to the leading EOF/PC structure. We have therefore shortened the paragraph and refocused it on the main conclusion (l. 427ff).

**R2C14: ll 425 ff: The authors should also take into account potential effect of over fitting and the deterioration of performance skill when using a large amount of predictors. In this case it is advisable to use e.g. metrics related to the concept of the adjusted R2 (Heinzl and Mittlböck, 2002). The adjusted R2 takes into account the number of predictors used. The difference is between the original and the adjusted R2 will be especially large when predictors are co-linear, i.e. not statistically independent. On top, standard performance metrics related to stat. significance (e.g. using Monte Carlo or Bootstrap methods) should be provided.**

**ANS:** We thank the reviewer for pointing out this issue. We agree that adjusted $R^2$ can be useful to account for the fact that the number of predictors increases with latent-space dimension. We therefore repeated the skill-versus-dimension analysis shown in Fig. 2 using adjusted $R^2$. This diagnostic is shown for GMST prediction in Fig. R2.7. The adjusted $R^2$ curves are almost indistinguishable from the original $R^2$ curves for both LIM and AERCn, indicating that the adjustment has a negligible effect in our case and does not modify the interpretation of the skill saturation with increasing latent dimension.

In addition, we estimated statistical uncertainty using a block-bootstrap procedure. We generated 1000 pseudo-series by resampling 10-year blocks with replacement from the original time series, and recomputed correlation and RMSE for each pseudo-series. The 95 % confidence intervals, defined as the 2.5th and 97.5th percentiles of the bootstrap distribution, are shown in Fig. R2.8. However, because adding these intervals makes Fig. 2 difficult to read and does not change the interpretation of the saturation behaviour, we do not modify this figure in the revised manuscript. We nevertheless agree that bootstrap confidence intervals are useful, and we have added them in the revised manuscript for Fig. 4 and Table 2, as discussed in our responses to R2C16 and R2C22.

**R2C15: Ll 487 ff: Again, authors should address the impact of increasing number of predictors on the real and potentially spurious increase in the performance metric.**

**ANS:** In the passage referred to here, the comparison between emulators is performed at a fixed latent dimension of 16 for all configurations. Thus, LIM, RC, RCn, AE and AERCn all rely on reduced spaces of the same dimension. The reported improvement of AERCn over LIM and the other emulators therefore cannot be attributed to an increase in the number of predictors, but rather to the quality of the emulator architecture itself.

**R2C16: ll 490 ff: I suggest to include also running (Pearson) correlations r with global level of statistical significance (Based on Monte Carlo or Bootstrap).**

**ANS:** We thank the reviewer for this suggestion. We tested the same diagnostic using 30-year running Pearson correlations for GMST and Niño3.4. However, as shown in Fig. R2.9, this metric highlights the relationship with volcanic forcing less clearly than the running RMSE. The temporal variations discussed in this section mainly concern changes in the amplitude of the prediction error rather than the phasing, especially during volcanic episodes, and RMSE is therefore better suited to capture this effect. In addition, the RMSE diagnostic is directly comparable to the running standard deviation of the Niño3.4 index shown in panel (c), since both quantities are expressed on the same amplitude scale. The running standard deviation provides a reference for the typical magnitude of Niño3.4 fluctuations, whereas RMSE measures the magnitude of prediction errors. We therefore retained the RMSE-based diagnostic in Fig. 4.

Regarding the request for statistical significance, we assessed uncertainty using a 10-year block-bootstrap procedure. We generated 1000 pseudo-series by resampling 10-year blocks with replacement from the original time series. For each pseudo-series, correlation and RMSE were recomputed, and 95 % confidence intervals were defined as the 2.5th and 97.5th percentiles of the resulting bootstrap distribution. We propose to revise Fig. 4 by adding 95 % confidence intervals around the period-mean RMSE values shown by the dotted lines in panels (a) and (b). This bootstrap procedure has also been added to Sect. 4.3.3 of the revised manuscript (l. 375ff).

**R2C17: ll 539ff: A hint here is to just estimate the lag(1) autocorrelation of the individual PC1/PC2 which also should be a good indication of the differences in between the different CMIP6 models over the tropical Pacific.**

**ANS:** We thank the reviewer for this suggestion. Indeed, the lag-1 autocorrelation of the leading tropical Pacific PCs could potentially explain the varying contribution of the RC memory component across CMIP6 models. We therefore computed the lag-1 autocorrelation of PC1 and PC2 for each model and compared it with the estimated RC memory contribution. The results, shown in Fig. R2.10, do not reveal any clear relationship: the corresponding correlations are close to zero. This suggests that, in our analysis, the memory contribution of RC is not simply controlled by the lag-1 autocorrelation of the leading tropical Pacific PCs.

**R2C18: Ll 545ff: I was wondering if non-linearity and memory are orthogonal in a statistical sense or whether also overlaps exist in terms of commonalities.**

**ANS:** We thank the reviewer for this interesting point. We agree that it is difficult to claim that nonlinearity and memory are strictly orthogonal in a formal statistical sense. Our initial diagnostic was based on the full RC model, from which we separately removed either the nonlinear activation or the recurrent memory component. To further assess whether these two effects can be meaningfully decomposed, we also performed the reverse comparison, starting from a simple Ridge regression baseline and adding either nonlinearity or memory. The results are shown in Fig. R2.11. The percentages obtained from the Ridge-based decomposition are very close to those obtained from the original RC-based ablation. This suggests that, in our experiments, the skill improvement from Ridge to RC can be consistently interpreted as the combination of a linear baseline plus memory and nonlinearity contributions. While this does not constitute a formal proof of statistical orthogonality, it supports the robustness of our decomposition.

**R2C19: ll 704 ff: The statement that the EOF based dimensionality reduction "prioritize variance rather than predictability" is in my opinion questionable. There is no contradiction between the two, because it s the setup of the statistical/emulator model using temporal evolution that renders the capabilities for prediction.**

**ANS:** We thank the reviewer for this clarification. We agree that the prediction itself is performed by the temporal emulator, not by the EOF decomposition. However, the emulator operates on a latent space that is already constrained by the dimensionality-reduction step. Standard EOFs provide a compact representation that maximizes the variance of the field at time t and is therefore optimal for reconstructing the current state in a variance sense. This does not necessarily guarantee that all retained directions are the most relevant for predicting the next state. Conversely, some lower-variance components may contain useful predictive information. In our AE formulation, the decoder is trained to reconstruct both the current and next states, which encourages the latent representation to retain information relevant to the prediction task. The skill difference between RCn and AERCn, in favor of AERCn, illustrates that. However, we understand the reviewer's comment and agree that LIM-EOFs remain valuable predictive tools, particularly because EOFs can retain substantial predictive skill. We have clarified this point in the revised manuscript where appropriate: our aim is to explore an alternative dimensionality-reduction approach that explicitly maximizes predictability rather than explained variance.

**Figures and Tables:**

**R2C20: Table1: I suggest to be more specific which type of "model" is referred too. The study uses both, numerical CMIP6 type of Earth System Models together with statistical models/emulators for index reconstruction.**

**ANS:** We thank the reviewer for this suggestion. Throughout the manuscript, we use "model" to refer to CMIP6 climate models and "emulator" to refer to the statistical architectures developed and evaluated in this study. To avoid any ambiguity in the tables, we have replaced "Model" with "Emulator" in Tables 1 and 2.

**R2C21: Figure 1: Which variable is used as a basis? – this should be mentioned in the Figure titles and Figure captions.**

**ANS:** We thank the reviewer for this comment. We have revised the panel titles and caption of Fig. 1 (and Fig. A6) to explicitly state that the latent representations are constructed from annual surface air temperature (tas) anomaly fields. The caption now starts as follows: “Figure 1: Spatial distribution of variance explained by latent representations of annual surface air temperature (tas) anomaly fields from the IPSL-CM6A-LR past2k ensemble.”. The panel titles were also revised to: “(a) EOFs tas variance explained: 71 %” and “(b) AE tas variance explained: 70 %”.

**R2C22: Table 2: Values for upper and lower confidence/statistical significance should be provided for correlations using bootstrap or Monte-Carlo methods. (I assume that with 1300 degrees of freedom the effect of serial correlations on the nominal level of significance can be ignored in this context).**

**ANS:** We thank the reviewer for this suggestion. We agree that providing uncertainty estimates for the correlation scores is useful. We therefore computed confidence intervals for the correlations reported in Table 2 using 1000 block-bootstrap resamples with 10-year blocks, in order to account for possible temporal dependence. To keep the main table readable, we retain the central correlation and RMSE values in Table 2 and provide the corresponding lower and upper confidence bounds for the correlations in Table A2. We have revised the manuscript accordingly to refer to Table A2 (ll. 470 and 516).

**R2C23: Figure 4: Please also include confidence intervals for correlations for the running indices and I suggest to include a companion Figure for running correlations in the Appendix.**

**ANS:** This comment is closely related to R2C16. As explained above, we tested the running-correlation diagnostic but found that the RMSE-based version of Fig. 4 was more appropriate for the point discussed here, namely temporal changes in prediction-error amplitude. We therefore propose to retain Fig. 4 in its RMSE form, while adding 95 % confidence intervals around the period-mean RMSE values shown by the dotted lines.

**R2C24: Figure 6 and 7: A short note in the caption helps to know what the lines using same color and with different train sizes represent (I assume they relate to the different CMIP6 models ???).**

**ANS:** We thank the reviewer for this suggestion. We have clarified the captions of Figs. 6 and 7. In panel (a), each curve corresponds to one emulator and represents the spatially averaged correlation averaged over all CMIP6 models. In panels (b–f), each curve corresponds to one CMIP6 model for the emulator indicated in the panel title.

**R2C25: Figure 8. please re-scale size of the colorbar.**

**ANS:** We thank the reviewer for this comment. We have corrected the size and positioning of the colorbar in Fig. 8.

*Additional References:*

*Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D’Andrea, F., Davini,* *P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, E., Lionel, Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, Journal of Advances in Modeling Earth Systems, 12, e2019MS002010, https://doi.org/10.1029/2019MS002010, 2020.*

*Dommenget, D. and Latif, M.: A Cautionary Note on the Interpretation of EOFs, Journal of Climate, 15, 216–225, https://doi.org/10.1175/1520-0442(2002)015<0216:ACNOTI>2.0.CO;2, 2002.*

*Heinzl, H. and M. Mittlböck (2002): Adjusted R2 Measures for the Inverse Gaussian Regression Model. Computational Statistics, 17, 525–544. https://doi.org/10.1007/s001800200125.*

*Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.*

*Jolliffe, I. T.: A Note on the Use of Principal Components in Regression, Journal of the Royal Statistical Society. Series C (Applied Statistics), 31, 300–303, https://doi.org/10.2307/2348005, 1982.*

*Lien, J., Kuo, Y.-N., Ando, H., and Kido, S.: Colored linear inverse model: A data-driven method for studying dynamical systems with temporally correlated stochasticity, Phys. Rev. Res., 7, 023042, https://doi.org/10.1103/PhysRevResearch.7.023042, 2025.*

*North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J.: Sampling Errors in the Estimation of Empirical Orthogonal Functions, Monthly Weather Review, 110, 699–706, https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2, 1982.*

*Weare, B. C. and Nasstrom, J. S.: Examples of Extended Empirical Orthogonal Function Analyses, Monthly Weather Review, 110, 481–485, https://doi.org/10.1175/1520-0493(1982)110<0481:EOEEOF>2.0.CO;2, 1982.*

*Wilks, D.S. (2010): Sampling distributions of the Brier score and Brier skill score under serial dependence. Q.J.R. Meteorol. Soc., 136: 2109-2118. https://doi.org/10.1002/qj.709*

*Zorita, E. and H. v. Storch (1999): The analog method as a simple statistical downscaling technique: comparison with more complicated methods. Journal of Climate 12, 2474-2489.*

**Supplementary figures for Reviewer 2**

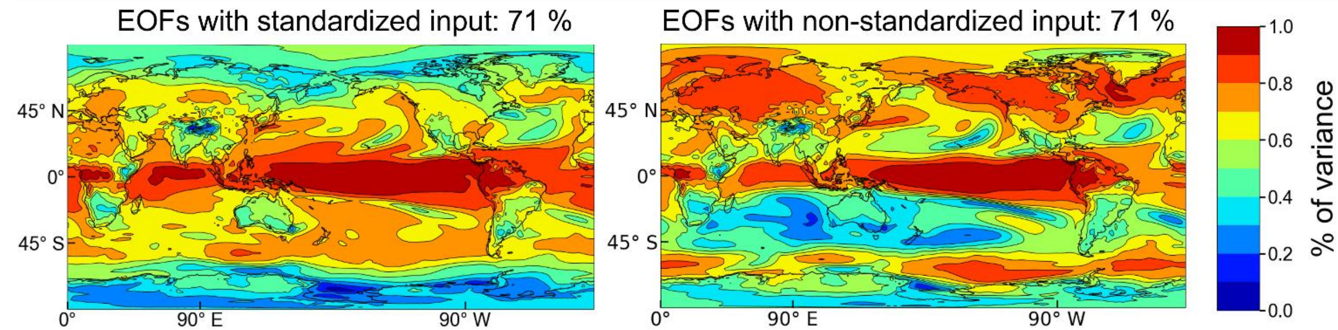


**Figure R2.1: Spatial distribution of variance explained by latent representations of annual surface air temperature (tas) anomaly fields from the IPSL-CM6A-LR past2k ensemble. Panels (a) and (b) show the variance explained by a 30-dimensional EOF latent space constructed from standardized input fields and non-standardized input fields, respectively.**

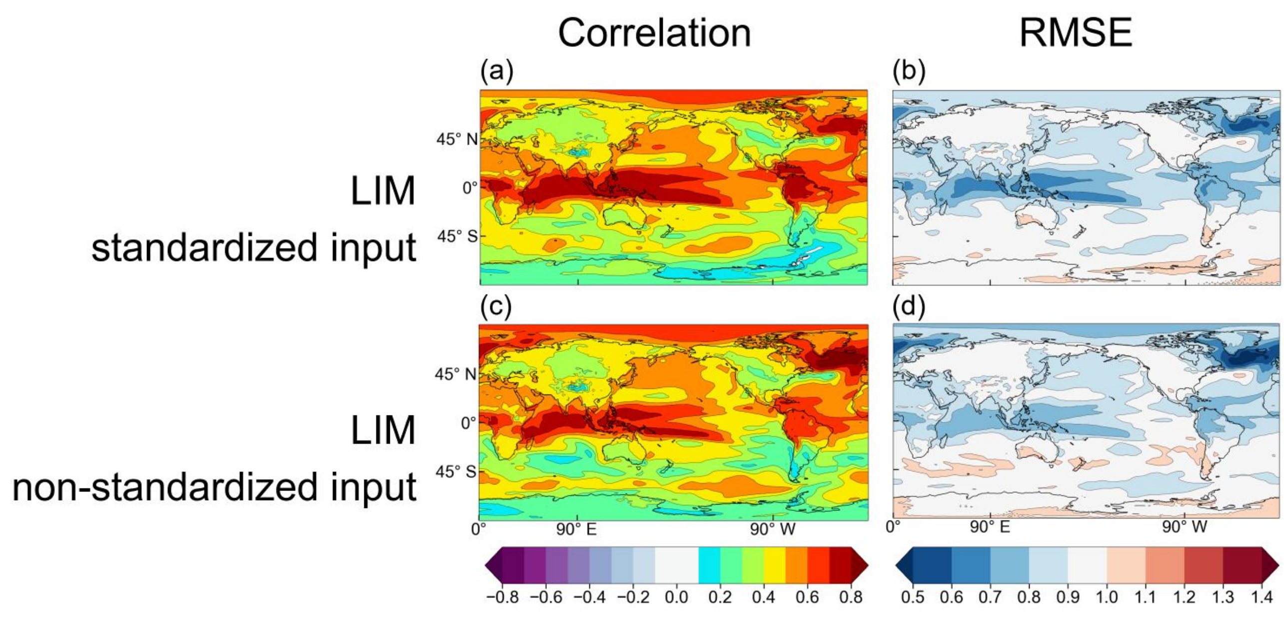


**Figure R2.2: Forecast skill at a lead time of 1 year for IPSL-CM6A-LR. Correlation (left column) and RMSE (right column) maps are shown for the LIM–EOF emulator using standardized input fields, in panels (a, b), and non-standardized input fields in panels (c, d).**

| | | Model | |
|---|---|---|---|
| **Variable** | **Metric** | **LIM with standardized input** | **LIM with non-standardized input** |
| **Spatial mean** | **Correlation** | **0.493** | 0.473 |
| | **RMSE** | **0.873** | 0.887 |
| **Niño 3.4** | **Correlation** | **0.593** | 0.480 |
| | **RMSE** | **0.518** | 0.564 |
| **AMV** | **Correlation** | 0.817 | **0.828** |
| | **RMSE** | 0.187 | **0.182** |
| **GMST** | **Correlation** | **0.856** | 0.844 |
| | **RMSE** | **0.121** | 0.125 |
| **NH** | **Correlation** | 0.860 | **0.889** |
| | **RMSE** | 0.229 | **0.206** |
| **SH** | **Correlation** | **0.658** | 0.637 |
| | **RMSE** | **0.150** | 0.153 |

**Table R2.1: Forecast skill at a lead time of one year over IPSL-CM6A-LR past2k simulations (500–1800 CE). Correlation and RMSE between emulator predictions and target anomalies are reported for spatially averaged local scores and selected climate indices (Niño3.4, AMV, GMST, NH, SH). Best scores for each metric are highlighted in bold.**

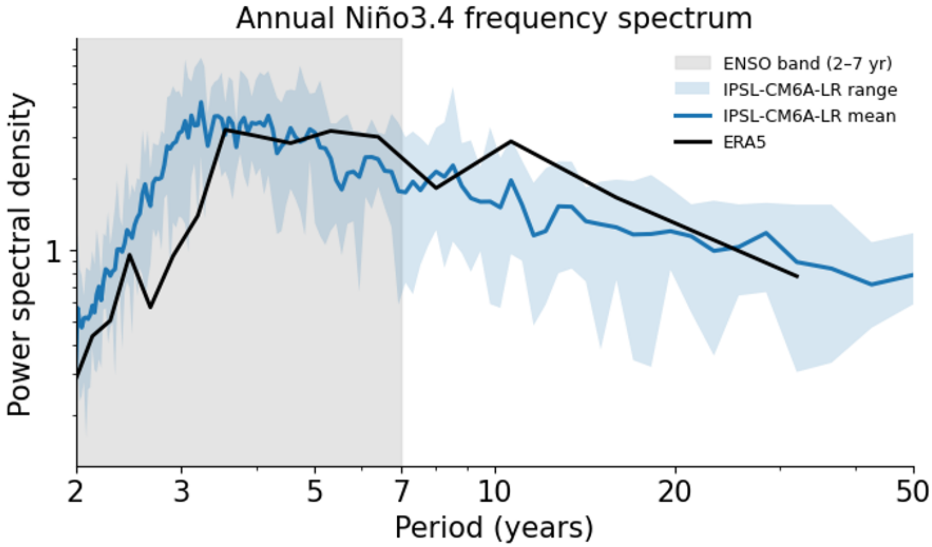


**Figure R2.3: Power spectrum of the annual Niño3.4 index in ERA5 (black curve) and in the IPSL-CM6A-LR past2k simulations. The IPSL-CM6A-LR ensemble mean is shown by the blue curve, and the light-blue shading indicates the full range across ensemble members. The grey shading indicates the 2–7-year ENSO band.**

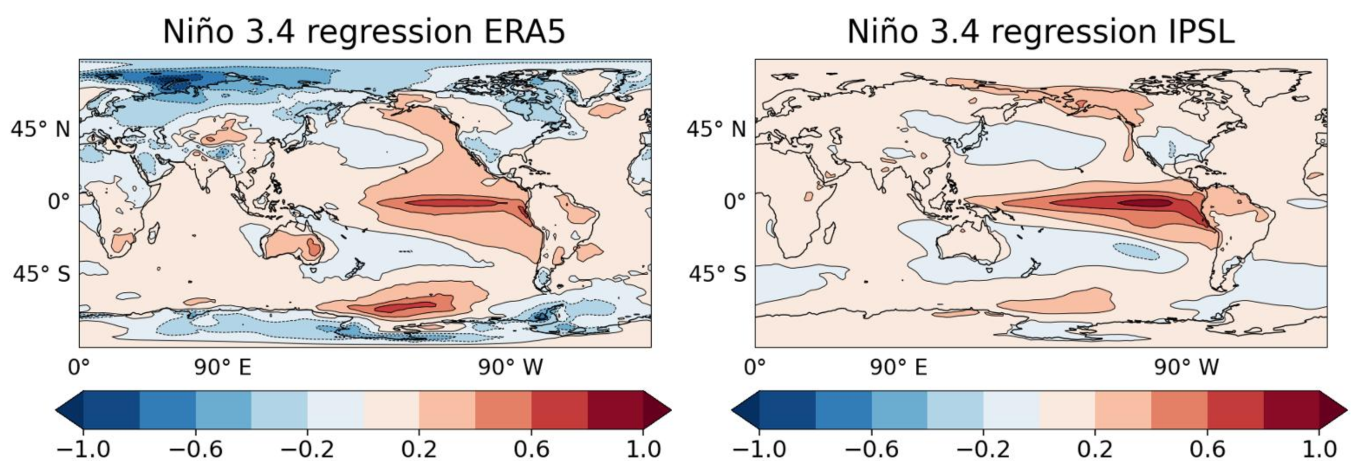


**Figure R2.4: Regression maps of annual surface temperature anomalies onto the Niño3.4 index in ERA5 (left) and in the IPSL-CM6A-LR past2k simulations (right).**

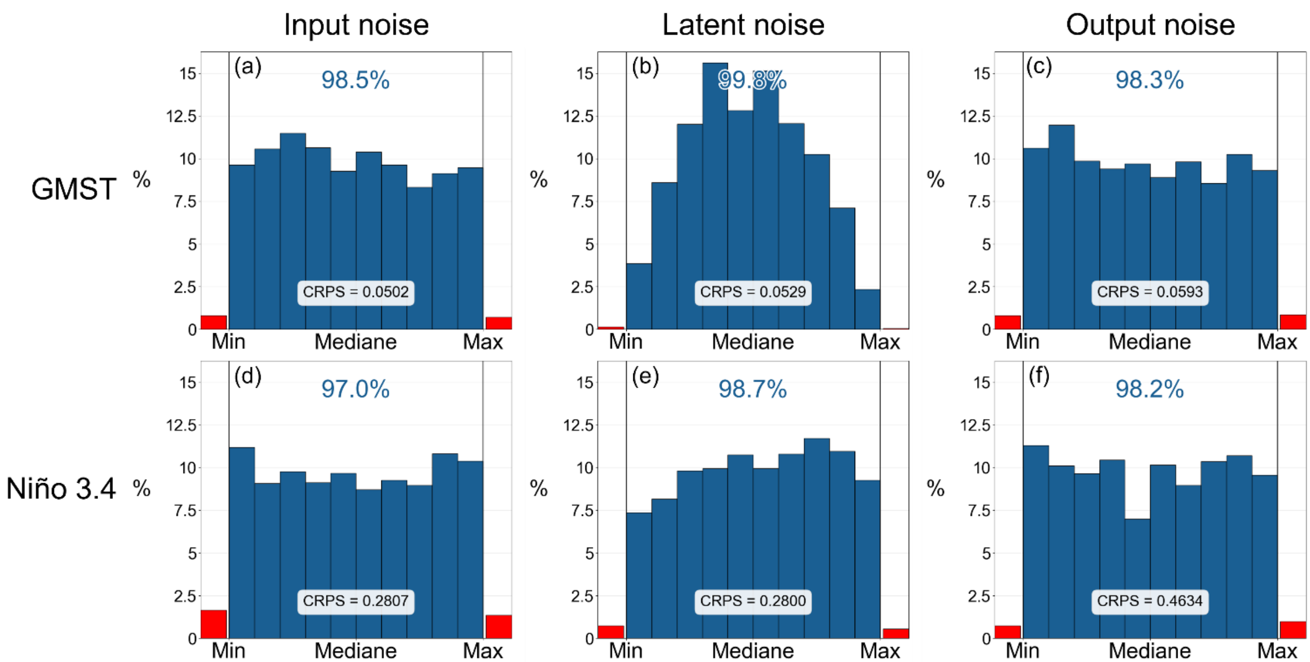


**Figure R2.5: Reliability of AERCn ensemble forecasts generated with different perturbation strategies for IPSL-CM6A-LR. Rank histograms are shown for GMST (a–c) and Niño3.4 (d–f) using 100-member one-year lead ensembles. Columns correspond to input-space white-noise perturbations, latent-space white-noise perturbations, and output-space noise based on the covariance of prediction errors. Blue bars indicate target frequency within ensemble deciles; red bars denote occurrences outside the ensemble spread. Percentages within the spread and mean CRPS values are indicated in each panel.**

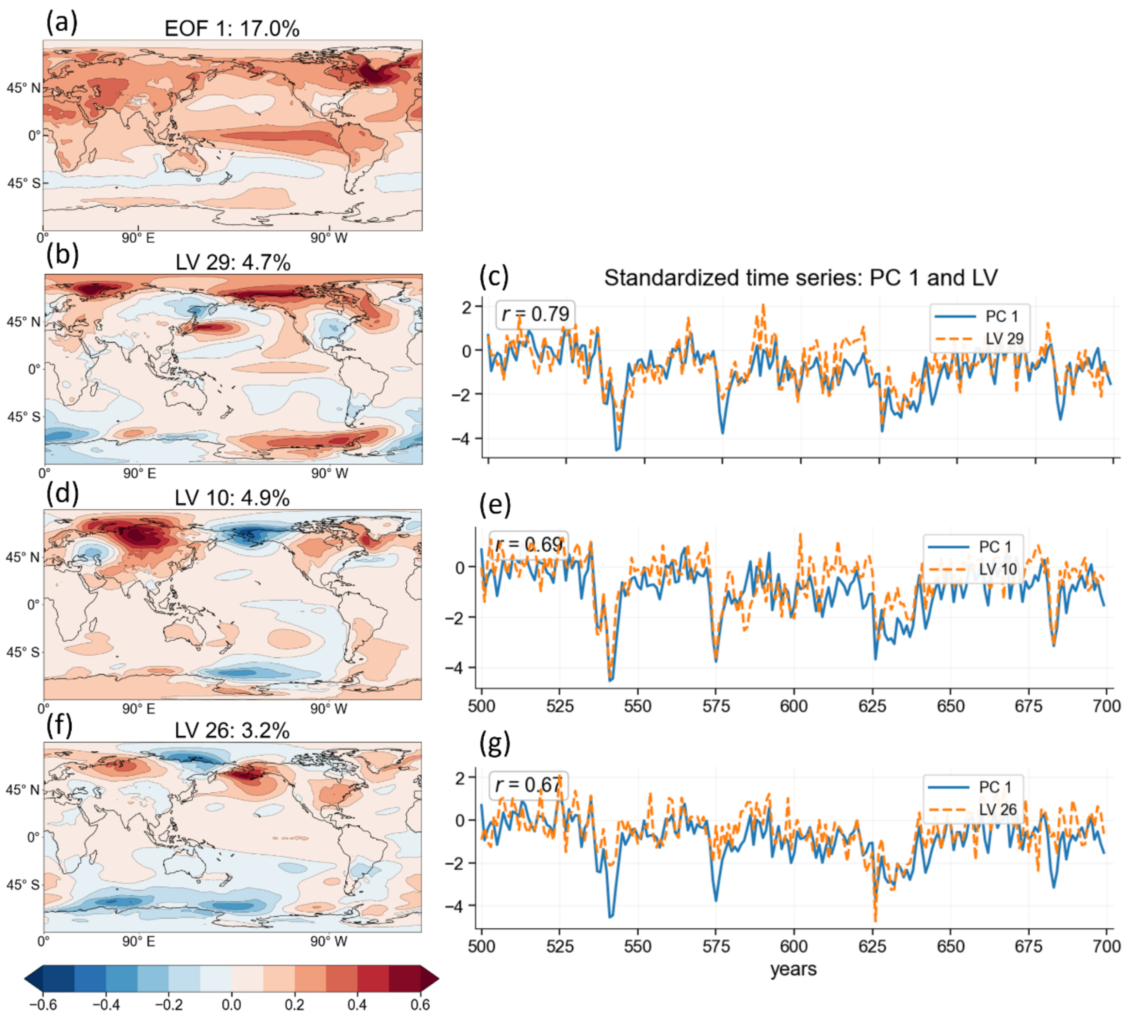


**Figure R2.6: Variance structures associated with EOF1 and AE latent vectors using a 30-dimensional latent space. Panel (a) shows the local variance explained by EOF1. Panels (b, d, f) display variance explained by AE latent vectors selected based on correlation with EOF1. Panels (c, e, g) show the corresponding standardized time series over the first 200 years. Correlation coefficients are indicated.**

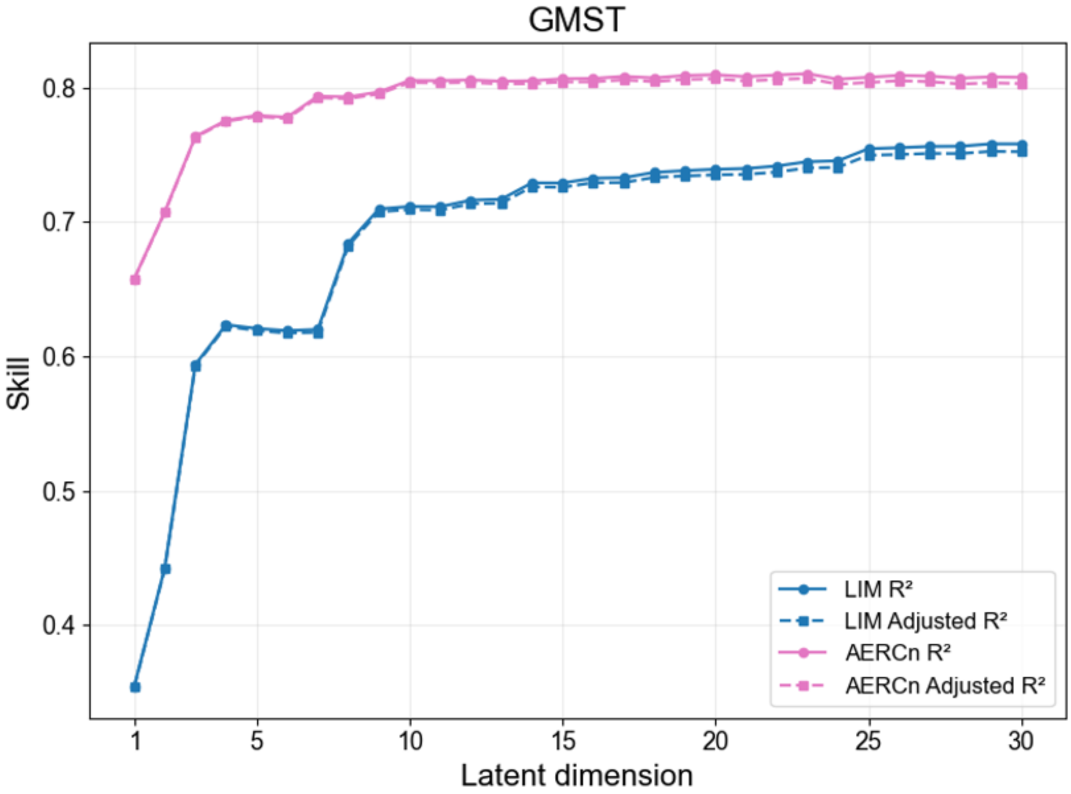


**Figure R2.7: GMST one-year lead prediction skill as a function of latent-space dimension for LIM and AERCn. Solid lines show the original $R^2$, while dashed lines show the adjusted $R^2$.**

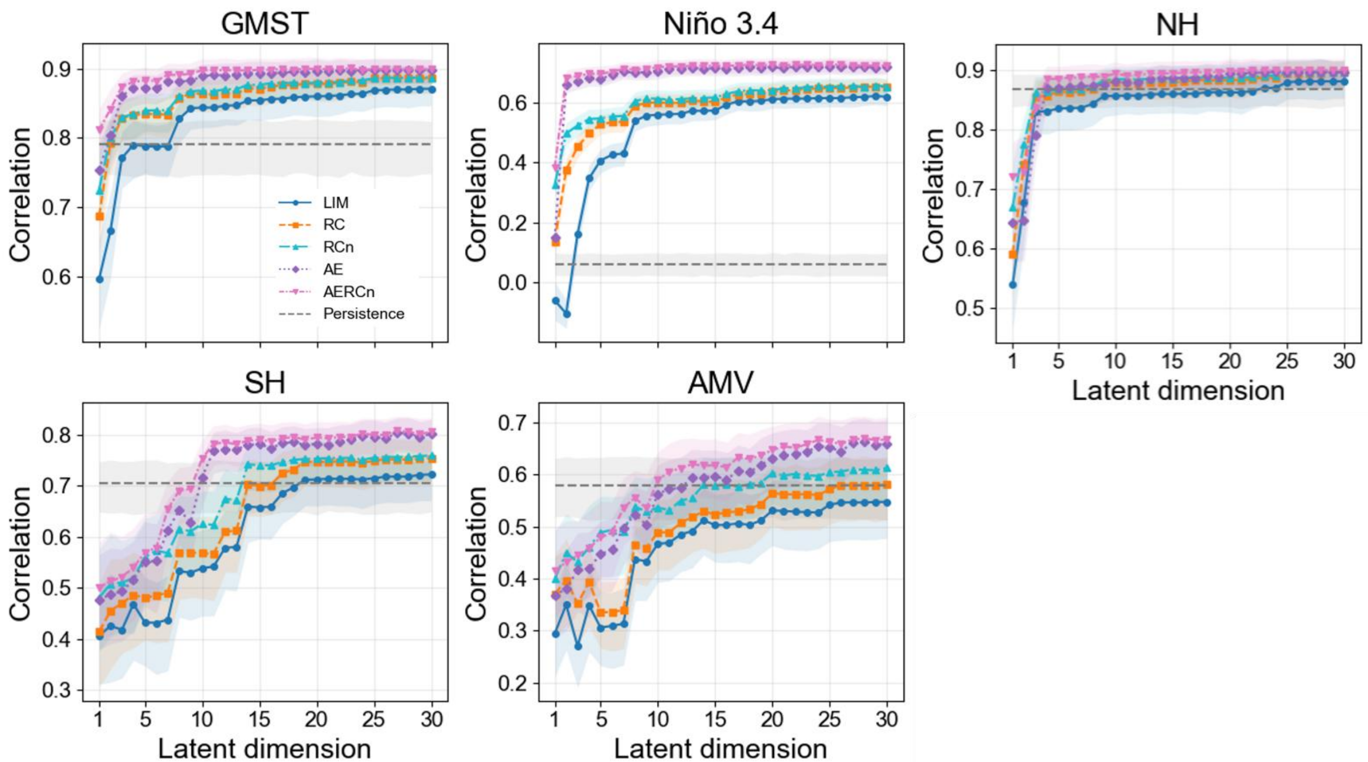


**Figure R2.8: One-year lead forecast skill as a function of latent-space dimension, with bootstrap uncertainty estimates. Correlation between emulator predictions and target surface temperature anomalies (500–1800 CE) is shown for GMST, Niño3.4, NH, SH, and AMV indices. Shaded areas indicate 95 % confidence intervals obtained from 1000 block-bootstrap resamples using 10-year blocks. The persistence benchmark is indicated by the dashed line.**

| | Emulator | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **Pers** | **LIM** | **RC** | **RCn** | **AE** | **AERCn** |
| **Spatial mean** | 0.302–0.360 | 0.467–0.516 | 0.484–0.532 | 0.508–0.553 | 0.524–0.568 | **0.534–0.578** |
| **Niño 3.4** | 0.024–0.098 | 0.563–0.622 | 0.590–0.647 | 0.600–0.656 | 0.692–0.739 | **0.700–0.747** |
| **AMV** | 0.708–0.799 | 0.781–0.845 | 0.800–0.859 | 0.807–0.864 | 0.817–0.869 | **0.826–0.876** |
| **GMST** | 0.745–0.828 | 0.826–0.877 | 0.848–0.892 | 0.855–0.898 | 0.872–0.911 | **0.878–0.914** |
| **NH** | 0.837–0.893 | 0.827–0.887 | 0.853–0.904 | 0.861–0.909 | 0.858–0.909 | **0.869–0.915** |
| **SH** | 0.646–0.747 | 0.594–0.713 | 0.639–0.749 | 0.684–0.783 | 0.745–0.817 | **0.748–0.823** |

**Table R2.2: 95 % confidence intervals for one-year lead correlation skill over IPSL-CM6A-LR past2k simulations (500–1800 CE). Intervals are reported for spatially averaged local correlations and selected climate indices (Niño3.4, AMV, GMST, NH, SH). Confidence intervals were estimated from 1000 block-bootstrap resamples using 10-year blocks. Best scores for each variable are highlighted in bold.**

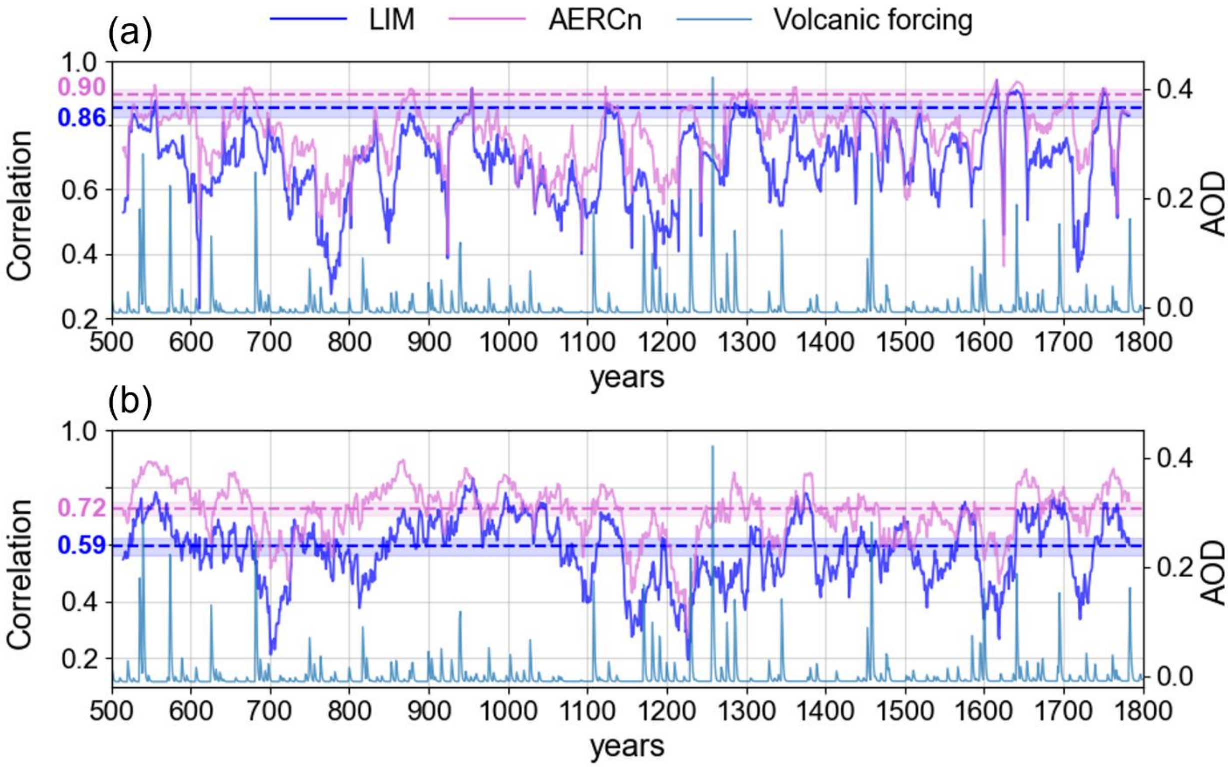


**Figure R2.9: Temporal evolution of forecast skill. Thirty-year rolling correlations of one-year lead forecasts for LIM (blue) and AERCn (pink) are shown for (a) GMST and (b) Niño3.4 in the past2k experiment. Dashed horizontal lines indicate full-period correlations, and shaded bands show the corresponding 95 % confidence intervals estimated with a block-bootstrap procedure. Volcanic AOD is overlaid (cyan) in panels (a–b).**

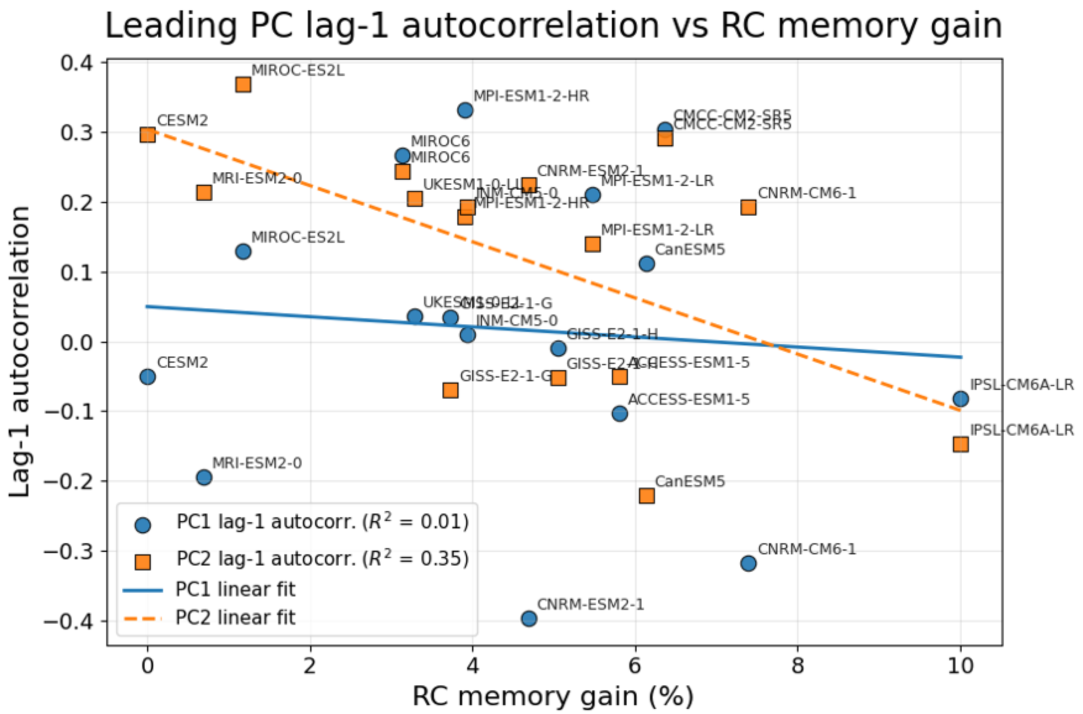


**Figure R2.10: Relationship between the lag-1 autocorrelation of the leading tropical Pacific PCs and the estimated RC memory contribution across CMIP6 models. Blue circles show the lag-1 autocorrelation of PC1 and orange squares show the lag-1 autocorrelation of PC2. Solid and dashed lines indicate the corresponding linear fits for PC1 and PC2, respectively. The very low (R^2) values indicate that the RC memory contribution is not simply explained by the lag-1 autocorrelation of the leading tropical Pacific PCs.**

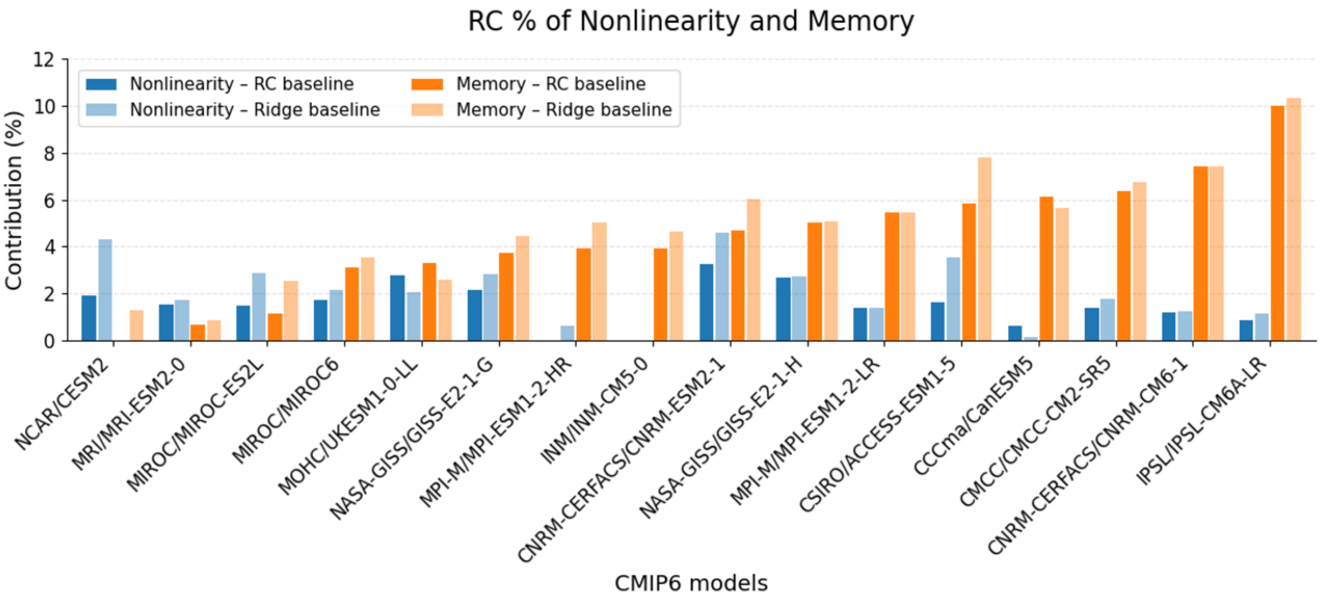


**Figure R2.11: Comparison of estimated nonlinearity and memory contributions across CMIP6 models using two reference baselines. Dark blue and dark orange bars show the contributions estimated relative to the RC baseline, while light blue and light orange bars show the corresponding estimates relative to a Ridge baseline. Models are ordered by increasing memory contribution estimated with the RC baseline.**