https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

New classes of climate model emulators to improve paleoclimate reconstructions

Auguste Gaudin[1], Myriam Khodri[1]

[1]LOCEAN-IPSL, Sorbonne-Université, CNRS/IRD/UPMC/MNHN, Paris, France

Correspondence to: Auguste Gaudin (auguste.gaudin@locean.ipsl.fr)

Abstract. Reconstructing spatial climate variability from proxy records requires forward models "emulators" that capture the dynamical structure of the climate system while remaining computationally efficient. Traditional emulators based on Empirical Orthogonal Functions and Linear Inverse Models (LIMs) face inherent limitations due to linearity and variance-based dimensionality reduction. Here we develop and evaluate a hierarchy of CMIP-class climate model emulators that integrate autoencoder-based dimensionality reduction with nonlinear prediction architectures, including Reservoir Computing (RC) and Recurrent Neural Networks (RNNs). Using a comprehensive experimental protocol applied to the IPSL-CM6A-LR model and 52 CMIP6s models, we show that combining autoencoder and RC (AERCn) provides the most robust performance across time scales and dynamical regimes when data are plentiful. The AERCn configuration captures nonlinear features of ENSO and Atlantic Multidecadal Variability, maintains high spatial reconstruction skill, and generalizes across distinct climate model structures. When training data are scarce, a multimodel pre-trained AERNN provides a data-efficient and competitive alternative. These properties make the proposed architectures particularly well suited for integration into Particle Filters and Ensemble Kalman Filter PDA frameworks. Our results highlight the importance of predictability-oriented dimensionality reduction and nonlinear dynamical memory for emulator design, and they provide a scalable path toward improved reconstructions of climate variability over the Common Era.

1 Introduction

Reconstructing spatially resolved climate variability over the Common Era (CE) is essential for advancing our understanding of natural climate dynamics and quantifying the extent to which anthropogenic forcing has altered the climate system. Instrumental observations, though precise, span only the past 150-170 years and thus fail to capture the full amplitude of internally generated multidecadal to centennial fluctuations (PAGES 2k Consortium, 2019; Cobb et al., 2003). Moreover, these observations are influenced by anthropogenic forcing, complicating efforts to disentangle natural and internally generated variability. Paleoclimate proxies (such as tree rings, corals, ice cores, and marine sediments) extend much further back in time, yet they are noisy, spatially heterogeneous, and require statistical calibration against modern observations and the dynamical constraints provided by physical climate models (Mann et al., 2008; Neukom et al., 2014; Smerdon and Pollack, 2016). To obtain dynamically consistent reconstructions of past climate states, it is therefore necessary to combine

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

these observational constraints with physically based information from climate models through paleoclimate data assimilation (PDA).

Sequential ensemble-based data assimilation frameworks, including Ensemble Kalman Filters (EnKFs) and Particle Filters (PFs), provide natural approaches for integrating proxies and model dynamics because they propagate uncertainty and explicitly approximate the evolving conditional distribution of the climate state (Houtekamer and Zhang, 2016; van Leeuwen, 2009; van Leeuwen et al., 2019). PFs are theoretically appealing because they impose no linear or Gaussian assumptions and can represent multimodal distributions arising from nonlinear climate dynamics. However, their practical application to high-dimensional geophysical systems faces two substantial limitations.

The first challenge is the curse of dimensionality. The number of particles required for accurate filtering increases exponentially with the dimension of the state vector (Snyder et al., 2008). Even reduced global fields such as surface air temperature involve thousands to tens of thousands of degrees of freedom, far beyond what is feasible for PFs without some form of dimensionality reduction. Empirical Orthogonal Functions (EOFs), or Principal Component Analysis (PCA), have therefore become standard tools for reducing dimensionality (Berliner & Wikle, 2007). While effective in many contexts, EOFs are fundamentally variance-maximizing linear decompositions. They are not optimized for prediction and can filter out low-variance but dynamically important features that contribute to climate predictability (Hannachi et al., 2007).

The second challenge concerns the computational cost of the forward model. Ideally, ensemble members would be propagated using General Circulation Models (GCMs), which offer the most physically comprehensive representation of the climate system (Eyring et al., 2016). Yet state-of-the-art CMIP6 GCMs while subject to structural biases (Zhang et al., 2023; Richter & Tokinaga, 2020; Zhu et al., 2020), they are far too computationally expensive for online use within sequential filtering, particularly for PFs which require thousands of ensemble trajectories per assimilation step. Attempts to circumvent this bottleneck include the use of intermediate-complexity models (Goosse, 2010), but these models come at the cost of limited ability to represent coupled nonlinear processes.

These limitations have motivated growing interest in climate model emulators, corresponding to statistical dynamical models trained on GCM output that approximate its behavior at a fraction of the computational cost. Among these, the Linear Inverse Model (LIM) has gained prominence because of its conceptual simplicity, low computational burden, and relatively strong performance in capturing predictable components of climate variability (Penland & Matrosova, 1998; Newman, 2013). LIM represents the evolution of the climate state using a linear deterministic operator plus additive stochastic noise that approximates unresolved nonlinear processes. When combined with EOF-based dimensionality reduction, LIM has been shown to successfully emulate interannual to decadal climate variability and to capture key predictable patterns such as ENSO and AMV (Penland and Sardeshmukh, 1995; Farrell and Ioannou, 1996; Penland and Matrosova, 1998; Newman et al., 2003; Newman, 2007; Newman, 2013; Alexander et al., 2008; Vimont, 2010). Recently, Jebri and Khodri (2023) demonstrated the feasibility of using a LIM–EOF emulator as the forward model in an online PF to reconstruct CE surface temperature fields, marking a significant advance for PDA.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

However, two structural limitations remain. First, LIM is inherently linear and memoryless. Although the additive-noise formulation can approximate some nonlinearities, key dynamical processes, such as ENSO asymmetry, episodic regime shifts, and strong atmosphere-ocean feedbacks, cannot be reproduced faithfully by linear dynamics (An & Jin, 2004; Jin et al., 2003; Timmermann et al., 2018). This deficiency limits the emulator's skill in predicting extreme climate events, which have disproportionate climatic impacts (IPCC AR6, 2021) and are central to PDA skill. Furthermore, climate predictability arises from nonlinear ocean-atmosphere interactions and long-term memory stored in oceanic heat content (Rial et al., 2004; Namias & Born, 1970), processes that cannot be represented adequately in linear Markov models.

Second, EOFs have long served as general-purpose tools that are not specifically optimized for predictive performance. Despite their utility, EOFs implicitly assume that the dominant modes of climate variability reside on a linear subspace and that the directions of largest variance are the most relevant for prediction (Hannachi et al., 2007). This assumption is often violated in the climate system, where variability emerges from a network of nonlinear feedbacks (Rial et al., 2004), asymmetric responses such as ENSO skewness (An & Jin, 2004), and regime transitions involving multi-stable dynamics (Ghil, 2002) which are essential to understanding variability and predictability. As a consequence, EOFs do not necessarily retain those structures most relevant for prediction or for emulating nonlinear processes (Ross, 2009). Indeed, modes of relatively small variance may nonetheless carry substantial predictive information (Newman, 2013), and linear decompositions frequently misrepresent regions with strongly asymmetric variability, such as the equatorial Pacific (Timmermann et al., 2018). These limitations motivate the search for emulators with richer dynamical capacity.

Recent advances in machine learning offer promising avenues to address the limitations of climate emulation tasks (Watson-Parris, 2022). Neural network architectures have shown the ability to capture spatial coherence and learn temporal dependencies from climate data, sometimes outperforming traditional dynamical models (Zhang et al., 2017; Ham et al., 2019; Maulik et al., 2020a). Deep autoencoders, in particular, provide a nonlinear alternative to EOF-based dimensionality reduction, capable of learning low-dimensional latent structure that better captures the spatial and dynamical structure of climate fields (Maulik et al., 2020b; Behrens et al. 2022). Taken together, these studies provide strong evidence that latent representations learned by autoencoders (AEs) can outperform traditional EOFs in dimensionality reduction, while offering increased flexibility and improved suitability for climate prediction and model emulation tasks.

Parallel developments in Reservoir Computing (RC) provide an attractive substitute for LIM. RC is a kind of recurrent neural-network in which only the output layer is trained, while the nonlinear reservoir dynamics are fixed (Jaeger & Haas, 2004). This yields a powerful dynamical emulation that retains LIM's computational efficiency while adding memory and the capacity to represent nonlinear interactions. RC has recently demonstrated superior performance to LIM in emulating ocean-atmosphere dynamics, including North Atlantic climate variability and ENSO-like phenomena (Nadiga, 2021; Guardamagna et al., 2025). These results indicate that RC can emulate transitions between different dynamical regimes and reproduce the skewness, intermittency, and nonlinear feedbacks characteristic of real-world climate variability.

Integrating these innovations, dimensionality reduction via AEs and dynamical emulation via RC, offers an opportunity to construct next-generation emulators tailored to the needs of PDA. Such emulators must be (i) computationally efficient, (ii)

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

capable of ensemble generation, and (iii) able to represent extreme climate events and nonlinear climate dynamics. Addressing these requirements would overcome the main obstacles that currently limit PF-based paleoclimate reconstructions: high dimensionality, non-linearity and the computational cost of physically based models.

In this study, we propose an improved emulator of CMIP-class climate models that directly addresses the structural constraints of the LIM–EOF approach. Building on previous work, we explore three main axes of improvement: (1) Replacing the LIM with RC to incorporate nonlinear dynamics and intrinsic memory; (2) Simplifying the emulator architecture by merging the prediction and reconstruction components, thereby reducing error accumulation; (3) Replacing EOFs with an AE specifically trained to retain predictable components of the climate system.

We evaluate whether these innovations improve emulation of large-scale climate variability, enhance prediction skill, and reduce errors in extreme-event emulation, particularly for ENSO whose structure makes it a stringent test of emulator performance. The overarching objective is to build an emulator that is sufficiently accurate, efficient, and robust to be used within PF-based PDA frameworks for reconstructing climate variability over the CE.

The remainder of the paper is organized as follows. Section 2 introduces the dimensionality-reduction and prediction models. Section 3 describes the emulator architectures and their implementation. Section 4 presents the datasets, experimental protocols, and evaluation framework used to assess emulator performance. Section 5 reports the results, including comparisons between EOF and AE-based dimensionality reduction, evaluation of emulator skill in a perfect-model setting, and dedicated analyses of ENSO dynamics, extreme events, and probabilistic ensemble predictions. Finally, Sect. 6 summarizes the main conclusions and outlines implications for future PDA applications and climate predictability research.

2 Emulator building blocks

This section introduces the two approaches employed for dimensionality reduction and the three individual algorithmic components used to construct the emulators. Their combination into full emulator architectures is described in Sect. 3.

2.1 Dimension Reduction

2.1.1. Empirical Orthogonal Functions

When applied to gridded climate anomaly fields, EOF analysis transforms spatially distributed variables into an orthogonal basis that maximises explained variance. Let X' denote the matrix of standardized and latitude-weighted annual near-surface temperature (tas) anomaly fields arranged as a sequence of spatial maps on a 143.144 grid. The EOF decomposition is obtained through eigen analysis of the covariance matrix

$$\mathrm{C} = cov(X'), \tag{1}$$

which characterises the spatial co-variability structure of the dataset. Diagonalisation of C yields a set of orthogonal eigenvectors ($\mathrm{EOF}_1 \perp \mathrm{EOF}_2 \perp \ldots \perp \mathrm{EOF}_n$) and their associated eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_n$), where each eigenvalue quantifies the fraction of total variance captured by the corresponding mode. The EOFs therefore define statistically independent

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

spatial patterns ordered by decreasing variance contribution. Projection of X' onto this orthogonal basis produces the Principal Components (PCs), which represent the temporal evolution of each mode and provide a compact linear description of the dominant variability. In practice, truncation to the leading modes enables efficient representation and prediction while filtering small-scale noise.

However, EOF analysis is intrinsically limited by its linear and unsupervised nature: the retained modes optimise variance representation rather than predictability. Consequently, directions associated with high variance do not necessarily correspond to those that maximise forecast skill or dynamical relevance. This limitation motivates the development of alternative dimensionality-reduction strategies capable of capturing nonlinear and predictability-relevant modes of climate variability.

2.1.2. Autoencoders

AEs are neural networks including two components: an encoder, which maps the input data into a lower-dimensional latent representation, and a decoder, which reconstructs the original field from this latent vector. During training, the AE optimizes its internal weights to minimize a reconstruction loss, typically the mean squared error between the input and its reconstruction (Bank et al., 2021). Owing to their ability to capture nonlinear structure, AEs have been successfully applied in numerous domains, including denoising, anomaly detection, data generation, and classification (Gondara, 2016; Vahdat & Kautz, 2021). Their usefulness has also been demonstrated in climate applications, where convolutional AEs have yielded promising results for dimensionality reduction of surface temperature fields (Saenz et al., 2018).

In the present work, the AE is used as a predictability-constrained dimensionality reduction method. Unlike EOFs, which retain modes of maximal variance, the AE is designed to preserve information that is most relevant for prediction. To achieve this, we construct an AE that learns a joint latent representation of the fields at time *t* and at time *t+1*. This flexibility is a key advantage of AEs and cannot be reproduced within a purely EOF framework. Aside from this task-specific adaptation, the architecture remains simple and can be viewed as an improved PCA: the encoder and decoder each consist of a single dense layer with linear activations.

The AE is applied to the standardized and latitude-weighted fields X', reshaped into vectors of dimension (143.144). Let n denote the size of the latent space. The encoder maps the input field at time *t* to a latent vector:

$$LV_t = X_t' W_1 + b_1 \quad (2)$$

where W_1 and b_1 matrix are trainable matrices of size $(143.144, n)$ and $(1, n)$, respectively. The decoder reconstructs both the current and next fields from the latent representation:

$$\left[\hat{X}_t', \hat{X}_{t+1}'\right] = AE(X_t') = decoder(LV_t) = (X_t' W_1 + b_1) * W_2 + b_2 \quad (3)$$

where *W2* and *b2* map the latent vector back to a vector of dimension 2x(143.144). The latent vector LV thus forms the reduced representation used for prediction.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Training is performed using the Adam optimizer with a learning rate of 0.001. The network is trained for up to 200 epochs with a batch size of 500. Early stopping is applied with a patience of 10 epochs, and the learning rate is reduced by a factor of 0.1 if the validation loss does not improve for five consecutive epochs.

2.2 Prediction models

After introducing the two dimensionality reduction approaches, we now turn to the prediction models. These models take as input a latent state vector. In the following, we denote this latent state vector by z of dimension n, which may correspond either to PCs or to LVs.

2.2.1. Linear Inverse Model

We begin with the LIM, a method widely used to reproduce the dynamics of climate models. The LIM is based on the assumption that the evolution of a system can be decomposed into a linear deterministic component and a stochastic component represented by Gaussian white noise. It corresponds to a multivariate linear Markov process describing the dynamics of the state vector z as:

$$\frac{dz(t)}{dt} = Bz + \varepsilon \tag{4}$$

where B is the linear operator representing the deterministic dynamics, and ε is a Gaussian white-noise term accounting for stochastic variability and assumed to be temporally uncorrelated. The matrix B can be estimated from the lag τ_0 and 0 covariance matrices, $C(\tau_0)$ and $C(0)$, as follows:

$$B = {\tau_0}^{-1} ln\left(\frac{C(\tau_0)}{C(0)}\right) \tag{5}$$

The covariance matrix Q of the noise term ε is derived from the fluctuation–dissipation relation (Penland & Matrosova, 1994):

$$Q + BC(0) + C(0)B^T = 0 \tag{6}$$

The corresponding continuous Markov process is simulated by integrating Eq. (4) with a second-order Runge–Kutta scheme (Penland & Matrosova, 1994).

2.2.2. Reservoir Computing

RC is a recurrent neural architecture designed to emulate nonlinear and chaotic dynamics (Jaeger, 2001; Jaeger and Haas, 2004). Training is limited to the output layer, via ridge regression, making RC computationally efficient and stable compared to traditional RNNs. RC has recently demonstrated strong skill in climate prediction tasks (Nadiga, 2021). We use a standard RC architecture, namely the Echo State Network (ESN). The reservoir state satisfies:

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

$$r_t = (1-\alpha)r_{t-1} + \alpha\, tanh(\, W\, r_{t-1} + W_{in} z_t\,) \tag{7}$$

and predictions follow:

$$\hat{z}_{t+1} = W_{out} r_t \tag{8}$$

To assess the respective contributions of memory and nonlinearity, we consider two simplified variants:

- RC_{nm}, a memoryless RC (α=1, W=0);
- RC_{ℓ}, a linear RC using the identity as activation function.

Finally, we adapt the classical RC to predict the full field X'_{t+1} directly from the reduced representation z_t. This modified version is referred to as RCn.

2.2.3. Recurrent Neural Networks

RNNs process sequential data by updating a hidden state that carries information over time (Rumelhart et al., 1986). Their training relies on backpropagation through time (Werbos, 1990), but this procedure is often unstable due to vanishing or exploding gradients (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997; Pascanu et al., 2013).

To mitigate these limitations, more advanced architectures such as the Gated Recurrent Unit (GRU) have been introduced (Cho et al., 2014). GRUs use gating mechanisms that regulate how past information is retained or discarded, enabling the modelling of longer temporal dependencies. The GRU update equations are given by:

$$u_t = \sigma(\, W_u\, z_t + \, U_u\, h_{t-1} + b_u\,) \qquad update\ gate \tag{9}$$

$$r_t = \sigma(\, W_r\, z_t + \, U_r\, h_{t-1} + \, b_r\,) \qquad reset\ gate \tag{10}$$

$$\widetilde{h_t} = tanh(\, W_h\, z_t + \, U_h(\, r_t \odot h_{t-1}) + \, b_h) \quad candidate\ hidden\ state \tag{11}$$

$$h_t = (1-u_t) \odot h_{t-1} + u_t \odot h_t \tag{12}$$

A GRU-based RNN is used in this study. The network takes as input a sequence of the 15 preceding latent state vectors and is trained to predict the subsequent state. The temporal evolution is learned through a GRU layer with a hidden state dimension of 200. A Dropout layer (20%) follows to reduce overfitting by randomly deactivating units during training (Srivastava et al., 2014). Finally, a dense layer with linear activation maps the GRU output to the target field.

3 Emulators Architectures

The emulators used in this study are constructed by combining the dimensionality reduction methods and prediction models introduced in Sect. 2. Each emulator combines a dimensionality-reduction method and a prediction model, organized either in a three-step architecture, i.e. projection onto a latent space, prediction in latent space, and reconstruction to the full spatial domain, or in a two-step architecture, where prediction is performed directly from the latent space to the full spatial domain.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

We begin with a baseline emulator combining EOF-based dimensionality reduction and a LIM for prediction, referred to hereafter as LIM. Building upon this reference configuration, we investigate three complementary axes of improvement.

First, we replace the linear dynamical core with RC, which, unlike the LIM, provides dynamical memory and can reproduce nonlinear behavior characteristic of climate variability. This leads to the RC-EOF configuration, hereafter RC.

Second, we simplify the overall model architecture. Both LIM-EOF and RC-EOF employ a three-step pipeline in which dimensionality reduction, prediction in latent space, and reconstruction to full space are performed sequentially. The LIM requires this structure because it operates on a square operator, imposing identical input and output dimensions. RC does not share this constraint. Thus, RC can be configured in a more efficient two-step architecture that predicts the full field directly from the reduced state. This yields the RCn-EOF configuration, denoted RCn.

Third, we improve dimensionality reduction itself by replacing EOFs with an AE, which is explicitly trained to preserve the components of the system most relevant for prediction. The AE may be used alone, performing both dimensionality reduction and prediction. This configuration is hereafter referred to as AE. It can also be combined with RC, yielding a configuration that incorporates all three improvement axes: nonlinear prediction, simplified two-step architecture, and predictability-constrained dimension reduction. This combined configuration is referred to as AERCn.

Finally, because using a RNN instead of an RC can be beneficial when training data are scarce, we also introduce the AERNN configuration, which pairs an RNN predictor with an AE for dimensionality reduction and uses the two-step architecture. All emulator configurations examined in this study are summarized in Table 1. The variants RC_{ℓ} and RC_{nm}, introduced in Sect. 2.2.2., correspond to simplified forms of the RC–EOF model that selectively omit nonlinearity or dynamical memory, respectively.

	Model							
Characteristics	LIM	RC	RC_{ℓ}	RC_{nm}	RCn	AE	AERCn	AERNN
Linearity & non linearity	No	Yes	No	Yes	Yes	No	Yes	Yes
Dynamical memory	No	Yes	Yes	No	Yes	No	Yes	Yes
Dimension reduction	EOF	EOF	EOF	EOF	EOF	AE	AE	AE
Multimodel	No	No	No	No	No	Yes	No	Yes
Architecture	3-step	3-step	3-step	3-step	2-step	2-step	2-step	2-step

3-step Architecture: Projection EOF/AE → Prediction → Reconstruction

2-step Architecture: Projection EOF/AE → Prediction into the full space

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Table 1: Summary of emulator configurations considered in this study and their main characteristics, including the presence of nonlinear dynamics, dynamical memory, dimensionality reduction method, multimodel capability, and prediction architecture. "Multimodel" indicates architectures that allow pretraining on the full CMIP6 ensemble.

4 Datasets and experimental protocol

4.1 CMIP6 model simulations

This study is conducted in a perfect-model framework in which the goal is to emulate annual surface air temperature (tas) anomalies simulated by a 1515-year-long (500–2015 CE; labelled hereafter *past2k)* 10-members ensemble generated with the IPSL-CM6A-LR climate model. IPSL-CM6A-LR was developed at the Institut Pierre-Simon Laplace as part of CMIP6 (Boucher et al., 2020; Eyring et al., 2016). The model couples the LMDZ atmospheric component (Hourdin et al., 2020) with the Nucleus for European Modelling of the Ocean (NEMO) version 3.6 (Madec and the NEMO Team, 2016). NEMO includes parameterizations of large-scale ocean dynamics, sea-ice thermodynamics and dynamics (Rousset et al., 2015), and marine biogeochemistry (Aumont et al., 2015; Séférian et al., 2019). The ocean grid is quasi-isotropic and tripolar, with a nominal 1° resolution refined to 1/3° in the equatorial Pacific and 75 vertical levels. Vertical spacing ranges from ~1 m near the surface to ~200 m at depth (Boucher et al., 2020).

IPSL-CM6A-LR reproduces ENSO seasonality reasonably well, though El Niño warm anomalies extend too far west. Its Pacific expression of the AMV teleconnection is realistic, yet tropical Atlantic variability is somewhat weaker than observed. A prominent feature of this model is a multi-centennial AMV mode with peaks spaced by roughly 200 years (Boucher et al., 2020).

The *past2k* ensemble members begin from distinct ocean states in a long pre-industrial equilibrium simulation dominated by natural forcings of the year 500 CE. The 10 ensemble members are forced with both natural and anthropogenic drivers following PMIP4 and CMIP6 protocols (Jungclaus et al., 2017; Eyring, et al, 2016) thereby providing a seamless baseline for evaluating modern variability. For sensitivity tests, a five-member single-forcing ensemble is also employed with fixed pre-industrial forcing to 1850AD and driven solely by reconstructed spectral solar irradiance (SSI) over the 1330–2010 CE (labelled hereafter *past2k-SSI)*.

To evaluate emulator robustness across different model physics, surface temperature fields from 52 CMIP6 models are used (Table A1). For each model, both piControl simulations (fixed pre-industrial forcing) and historical simulations (1850–2012) are included. All datasets are interpolated onto a 143.144 latitude-longitude grid. Annual and December–January–February (DJF) seasonal mean anomalies are constructed by removing each experiment's mean seasonal cycle and linear trends are also removed from historical simulations to isolate internal low-frequency variability.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

4.2 Experimental protocols

To evaluate emulator performance, we design a suite of experiments that systematically examines dimension reduction skill, predictive skill, robustness across models, sensitivity to training data availability, and applicability to ensemble forecasting. We begin by assessing skill within the IPSL-CM6A-LR framework, and then extend the analysis to a broad set of CMIP6-class models. Particular emphasis is placed on understanding the behavior of RC under varying dynamical regimes and training constraints. The experimental protocol consists of three components described below.

4.2.1. Training protocol

Each emulator is first tested in a deterministic prediction setting using the *past2k* simulations (500–2015 CE) from the IPSL-CM6A-LR model (Sect. 4.1). For each emulator, nine ensemble members are used for training and the remaining member for testing. This configuration provides a controlled environment with abundant training data, enabling a systematic comparison and ranking of emulator skill.

To assess generalization, all emulators are evaluated across model experiments from 52 CMIP6-class models (Table A1). Training is performed on each model's available historical ensemble members, and evaluation is carried out on the corresponding *piControl* simulation. Emulator sensitivity to training-data availability is examined by repeating the training with progressively larger subsets of the historical ensemble members, up to the maximum available for each model.

4.2.2. Transfer learning for limited-data scenarios

To enhance emulation skill when training data are scarce, we investigate a multi-model transfer learning strategy. The AE is first pre-trained on all available historical ensemble members from the full CMIP6-model ensemble dataset, and then fine-tuned on the target model. RC-based architectures (e.g., AERCn) cannot directly benefit from transfer learning because the reservoir is non-trainable. To overcome this limitation, we replace RCn with a RNN, enabling both the AE and the RNN predictor to exploit multi-model pre-training. The resulting configurations, referred to as AE_MME and AERNN_MME, are evaluated for their ability to improve emulation skill under limited-data conditions.

4.2.3. Ensemble forecasts

Finally, we convert the deterministic emulators into ensemble emulators to enable ensemble forecasts evaluation. The LIM naturally yields ensemble generation through its stochastic component (Sect. 2.2.1.). In contrast, AE-based and RC-based emulators do not possess intrinsic stochasticity; we therefore developed and tested ensembles forecasts generated by perturbing the initial state with additive Gaussian white noise.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

4.3 Evaluation framework and skill metrics

4.3.1. Climate metrics

Emulator performance is evaluated both on annual timeseries of spatially gridded surface air temperature anomaly fields and on a set of five key climate indices designed to capture large-scale and dynamical aspects of climate variability of each CMIP6 model simulations considered in this study. These indices include annual timeseries of:

- the global mean surface temperature (GMST),
- the hemispheric mean temperatures over the Northern (NH; 0°–60° N) and Southern Hemisphere (SH; 60°–0° S),
- two indices representing major modes of variability: the Atlantic Multidecadal Variability (AMV; sea surface temperature averaged over 0°-60° N, 75° W-7.5° W; Ting et al., 2009) and the Niño 3.4 indices (sea surface temperature averaged over 5° S-5° N, 170° W-120° W; McPhaden et al., 2006).

More specifically, in order to quantify the degree to which RC adapts to model-specific nonlinear behavior, we use an ENSO nonlinearity indicator denoted α, defined as the quadratic coefficient in the relationship between the first two principal components (PC1, PC2) of DJF SST anomalies over the tropical Pacific (10° S–10° N, 80° W–120° E) (Karamperidou et al., 2016):

$$PC_2 = \alpha \cdot PC_1^2 + b \cdot PC_1 + c \tag{13}$$

PC1 characterizes ENSO intensity and closely corresponds to the Niño3.4 index, whereas PC2 describes the zonal structure of events (Eastern vs. Central Pacific ENSO). Larger values of α indicate stronger ENSO asymmetry and more pronounced nonlinear behavior.

4.3.2. Dimensionality reduction evaluation

EOFs and AE are first evaluated as dimension-reduction by quantifying the information loss induced by projection into latent space. This is done by comparing a field X' with its reconstruction from the dimensionality reduction $\hat{X}'$. For EOFs, reconstruction is obtained by projecting the PCs back onto the EOF basis. For the AE, reconstruction is obtained by applying the decoder to the LVs. Their reconstruction performance is assessed in terms of explained variance between the original spatio-temporal fields and the reconstructed ones obtained after reduction to n dimensions. The fraction of variance explained collectively by the n latent dimensions is quantified using the coefficient of determination. Locally, at each grid point, it is defined as:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \tag{14}$$

Where $SSE = \sum\left(x' - \widehat{x'}\right)^2$ denotes the sum of squared reconstruction errors and $SST = \sum(x' - \overline{x'})^2$ the total variance of the reference time series x' at the considered grid point.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

The explained variance of the full spatial map is then obtained by aggregating reconstruction errors and total variance over all grid points using:

$$R^2_{global} = 1 - \frac{\sum SSE}{\sum SST} \quad (15)$$

In addition, beyond the collective contribution of the n-dimensional latent space, the contribution of individual latent dimensions is obtained by reconstructing the field while fixing all other latent coordinates to their temporal mean. The explained variance of this reconstruction, computed using Eq. (14), is the percentage of variance explained by that latent dimension.

Finally, spatial patterns associated with one individual latent dimension are obtained by introducing a spatial reconstruction, in which all latent coordinates are fixed to their temporal mean, expect one which is fixed to its mean plus one standard deviation. By differencing this reconstruction from a baseline in which all latent coordinates are fixed to their mean, allows identifying the spatial pattern associated to a specific dimension of the latent space.

4.3.3. Attribution of emulator predictive skill

Emulator performance in terms of deterministic prediction at one-year lead-time is then assessed based on two metrics: the Root Mean Square Error (RMSE) and the Pearson correlation coefficient (hereafter correlation or *r*). RMSE quantifies the average amplitude of the prediction error and is expressed in the physical units of the target variable, while the correlation measures phase agreement between predicted and reference anomalies, independent of any systematic bias or scaling differences. High predictive skill is therefore characterized by low RMSE and high correlation. These metrics are computed both for climate indices and at each grid point, enabling the construction of spatial skill maps that reveal the geographical distribution of emulator performance. Finally, spatial averages of the skill maps are used to obtain domain-mean RMSE and correlation scores, which provide concise measures of overall emulator performance. Spatial averaging can also be restricted to specific regions of interest, such as the tropical Pacific (10° S–10° N, 80° W–120° E), to assess regional predictive skill.

To contextualize emulator performance, we also include a baseline benchmark based on Persistence. The Persistence experiment is a conservative benchmark used to evaluate predictive skill. It consists of predicting for the future the same value as the present, and allows quantification of the added value of the emulator-based forecast (Jebri and Khodri, 2023).

To disentangle the respective contributions of memory and nonlinearity within RC prediction skill, we compare the full model skill with two simplified variants: a linear version (RC_ℓ) and a memoryless version (RC_{nm}). Their relative contributions are quantified as:

$$RC_{\%\,nonlinearity} = \frac{(RC - RC_l)}{RC_l}, \quad RC_{\%\,memory} = \frac{(RC - RC_{nm})}{RC_{nm}} \quad (16)$$

Comparing these metrics with α provides insight into how effectively RC captures the nonlinear and dynamical characteristics of each climate model.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

For ensemble predictions, we use the Continuous Ranked Probability Score (CRPS), a metric widely used to evaluate probabilistic forecasts (Hersbach 2000; Gneiting and Raftery, 2007; Berrocal et al., 2007). The CRPS measures the difference between the predicted cumulative distribution function (CDF) from the ensemble and the observed value, generalizing the mean absolute error to probabilistic forecasts. Lower CRPS values indicate better probabilistic skill and better-calibrated ensembles. Let F denote the predictive CDF of the ensemble for a given climate index, and let I denote the corresponding observed index value. The CRPS is defined as:

$$CRPS(\mathrm{F},\mathrm{I}) = \int_{-\infty}^{\infty}(F(z) - \mathbf{1}\{z \geq I\})^2 dz \quad (17)$$

5 Results

Section 5.1 compares the two dimensionality-reduction approaches according to the protocol defined in Sect. 4.3.2. Section 5.2 then evaluates the predictive skill of the different emulator configurations under the evaluation protocols defined in Sects. 4.2-4.3. Finally, Sect. 5.3 examines emulator performance in dedicated case studies, focusing on deterministic and ensemble-based predictions of extreme ENSO events.

5.1 Dimension reduction

PF-based data assimilation methods tend to degenerate when the dimensionality of the state space becomes too high. The latent space must be small enough to mitigate the curse of dimensionality, but sufficiently large to preserve the information necessary for accurate prediction.

As shown in Fig. 1, truncating the representation to 30 dimensions captures a large fraction of the interannual variability, with about 70 % of the total variance retained in both EOF and AE cases with similar spatial patterns (Fig. 1a-b). In both cases, the fraction of explained local variance is generally higher in the tropics than in the extratropics, and higher over the Pacific basin than over the Atlantic. EOFs explain slightly more variance on average (by about 1 %), particularly at high latitudes in the Pacific.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

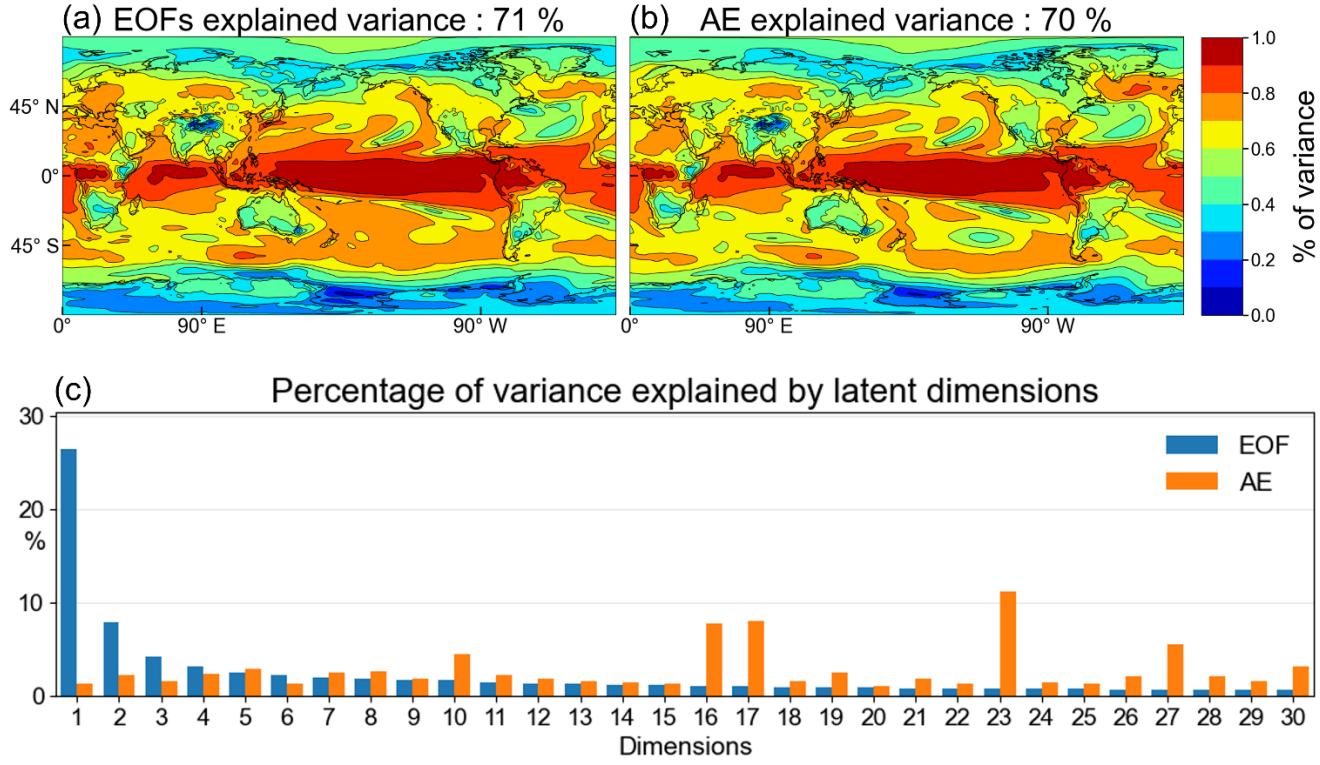


Figure 1: Spatial distribution of variance explained by latent representations of the IPSL-CM6A-LR past2k ensemble. Panels (a) and (b) show the variance explained by a 30-dimensional latent space constructed using EOFs and AE latent vectors, respectively. Panel (c) presents the fractional variance contribution associated with each latent dimension.

Although their overall variance-based skill appears comparable, both approaches differ fundamentally in how information is encoded and distributed across latent dimensions. As illustrated in Fig. 1c, EOFs yield a strongly hierarchical representation, with most of the variance concentrated in the leading modes. In contrast, the AE distributes information more evenly across dimensions, resulting in a more homogeneous latent representation with a few AE components standing out. The first three EOF modes together explain 38 % of the total variance, whereas the three most important AE dimensions, LVs 23, 17, and 16 account for only 27 %, highlighting the more diffuse and less hierarchical nature of the AE encoding. This value should in fact be regarded as an upper bound, since, unlike EOFs, AE latent dimensions are not orthogonal and their individual variance contributions cannot be summed straightforwardly. Consistently, the sum of the marginal variance contributions across all 30 AE latent dimensions reaches 83 %, exceeding by 13 % the variance actually explained by the latent space as a whole.

Although they differ substantially in the way information is decomposed, several AE latent vectors are strongly correlated with several EOFs (Fig. A1), mostly the leading ones (Fig. A2 and A3). In particular, the three AE dimensions explaining the largest fraction of variance are also those most strongly correlated with PC 1, with correlation coefficients exceeding 0.8. Yet, the leading EOFs are known to often exhibit physically interpretable spatial structures, with the first and second EOFs

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

primarily associated with externally forced variability and ENSO-related variability respectively (Fig. A2a; Fig. A3a), whereas the spatial patterns learned by the AE are less directly interpretable. Nevertheless, the time series of LVs 23, 17, and 16 that concentrate most of the variance exhibit nearly identical temporal variations than PC 1 (Fig. A2c, e, g), despite markedly different spatial patterns (Fig. A2b, d, f). The same observation applies for LVs 2 and 30, which significantly correlate with PC 2, although their contribution to the total explained variance is much weaker (Fig. A3). This indicates that the AE partly recovers the dominant EOF patterns while combining them with additional sources of variability and distributing the information across multiple latent dimensions, consistent with a more distributed representation of the dominant variability within the latent space.

The less hierarchical organization of information in the AE latent space is not, in itself, a source of improved skill, especially when the AE operates in a linear regime. Rather, the strength of autoencoders lies in the flexibility of their architecture. Unlike EOFs, which are based solely on variance maximization and are inherently unsupervised, AEs can be designed to explicitly serve the prediction task. In this study, the AE is trained to learn a joint latent space from both times t and $t + 1$. This design promotes the preservation of components that are most relevant for forecasting, thus improving predictive performance.

5.2 GCM emulator deterministic prediction skills

Section 5.2 focuses on evaluating the predictive skill of the emulators. Section 5.2.1. examines their performance on the *past2k* simulations of the IPSL-CM6A-LR model, providing a controlled setting with abundant training data. Section 5.2.2. broadens the analysis to the full ensemble of CMIP6-class models, allowing for an assessment of emulator robustness and sensitivity across a diverse set of model dynamics and simulations.

5.2.1. IPSL-CM6A-LR

As discussed in Sect. 5.1, emulator skills involve a trade-off between predictive capability and dimensionality reduction. As shown on Fig. 2, the 1-year lead-time prediction skill increases rapidly with latent-space dimension before reaching a saturation regime. For the global mean of the local scores, the GMST, the Nino3.4 and NH climate indices, all emulators reach near-asymptotic correlation values with approximately 9 latent dimensions (Fig. 2a–d). For the Southern Hemisphere mean temperature, the AE and AERCn emulators begin to level off at around 11 latent dimensions, whereas the LIM stabilizes only at higher dimensions (around 19) with a correlation skill only slightly higher than Persistence (Fig. 2e). Finally, for the AMV index, emulator skill saturates much more gradually, with a plateau reached only at relatively large latent dimensions, around 20–25 (Fig. 2f). Consistently across all considered metrics, at least one emulator outperforms the Persistence benchmark once the latent space reaches a dimension of 10. To balance predictive skill with the need to reduce the size of the problem, a latent dimension of 16 is selected for the remainder of the study, as it provides a good compromise across all considered climate metrics with correlation values of at least 0.5. This choice ensures robust forecast skill while retaining approximately 60 % of the total variance (Fig A4). Compared to a 30-dimensional representation, this corresponds

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

to a reduction of only about 10 % in explained variance, while maintaining similar predictive performance in a latent space that is nearly twice as compact. Repeating the diagnostics presented in Sect. 5.1 using a 16-dimensional latent space yields qualitatively similar results with a subset of AE LVs highly correlated (>0.7) with PC1, albeit for weaker correlations with the second EOF (Figs. A5–A6).

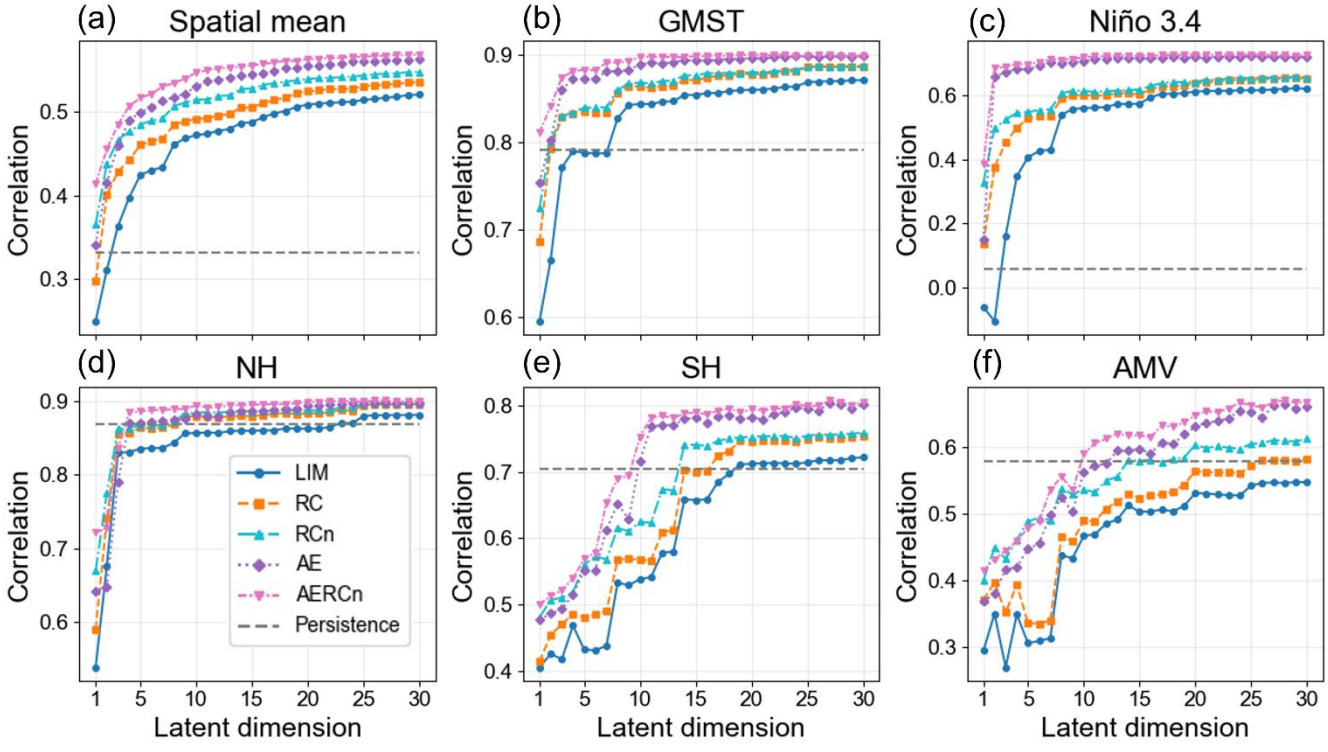


Figure 2: One-year lead forecast skill as a function of latent-space dimension. Correlation between emulator predictions and target surface temperature anomalies (500–1800 CE) is shown for (a) spatially averaged local scores, (b) GMST, (c) Niño3.4, (d) NH, (e) SH, and (f) AMV indices. The persistence benchmark is indicated by the dashed line.

With the latent-space dimension fixed, Fig. 3 shows the local skill in predicting the surface temperatures simulated by the IPSL-CM6A-LR model at a lead-time of 1 year for all emulators and for the Persistence benchmark. Table 2 summarizes the corresponding skill metrics averaged over the globe, over key regions and for specific indices. Results are discussed by progressively increasing model complexity along the three axes of improvement identified above, starting from Persistence and the LIM, and then considering RC, RCn, AE, and finally AERCn.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

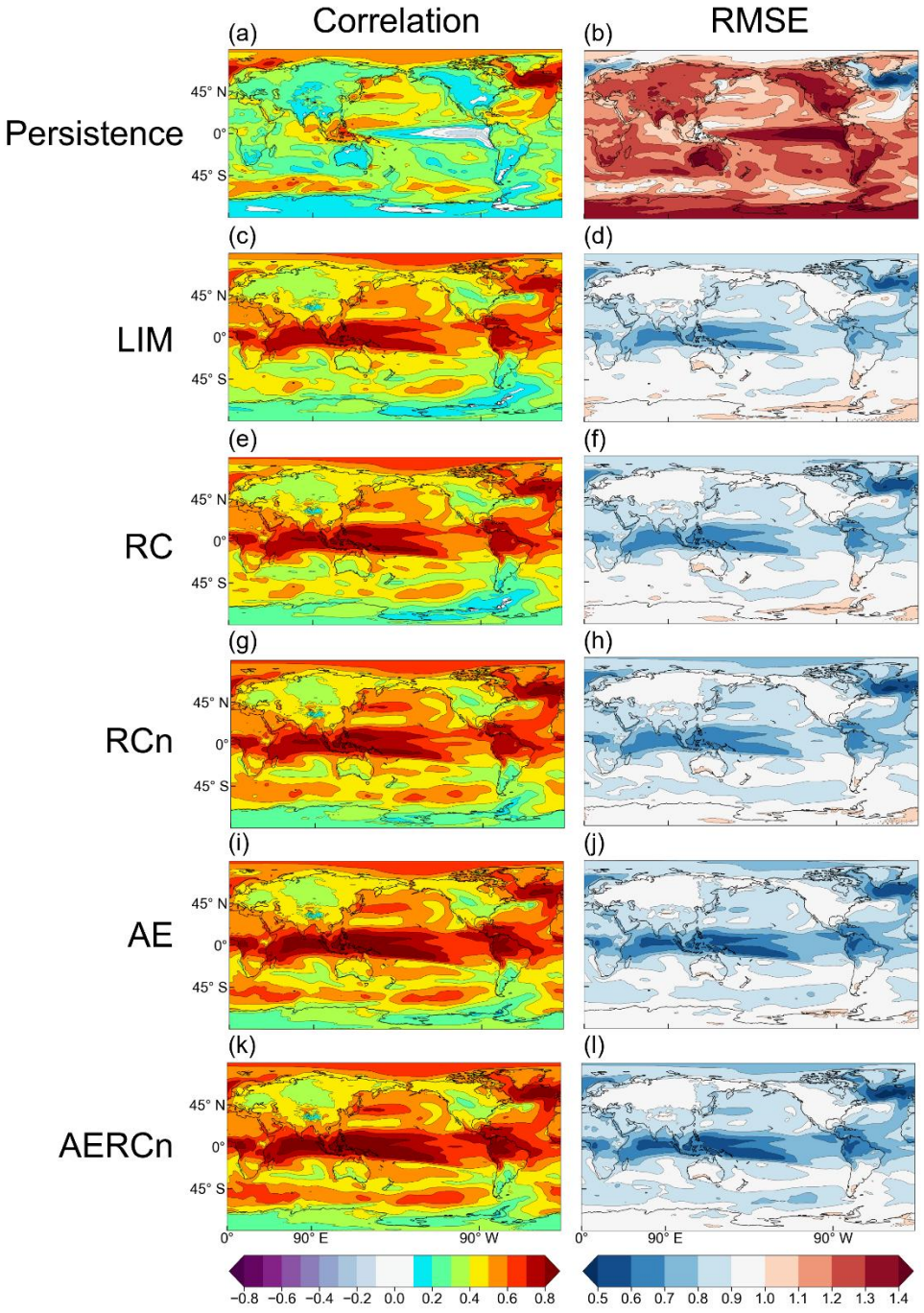


Figure 3: Spatial patterns of forecast skill at one-year lead time for IPSL-CM6A-LR. Correlation (left column) and RMSE (right column) are shown for Persistence (a–b), LIM (c–d), RC (e–f), RCn (g–h), AE (i–j), and AERCn (k–l).

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Variable	Metric	Model					
		Pers	LIM	RC	RCn	AE	AERCn
Spatial mean	**Correlation**	0.331	0.493	0.509	0.531	0.547	**0.557**
	RMSE	1.175	0.873	0.862	0.849	0.836	**0.829**
Niño 3.4	**Correlation**	0.059	0.593	0.618	0.629	0.716	**0.724**
	RMSE	0.882	0.518	0.506	0.500	0.449	**0.443**
AMV	**Correlation**	0.759	0.817	0.833	0.839	0.845	**0.854**
	RMSE	0.225	0.187	0.180	0.177	0.173	**0.169**
GMST	**Correlation**	0.791	0.856	0.873	0.879	0.894	**0.898**
	RMSE	0.150	0.121	0.114	0.111	0.104	**0.102**
NH	**Correlation**	0.869	0.860	0.880	0.887	0.886	**0.895**
	RMSE	0.230	0.229	0.214	0.208	0.208	**0.201**
SH	**Correlation**	0.704	0.658	0.699	0.738	0.786	**0.791**
	RMSE	0.151	0.150	0.144	0.134	0.122	**0.121**

Table 2: 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.

Persistence generally exhibits low overall skill, yet it can match or even exceed LIM performance in specific regions. As shown on Fig. 3a–d, Persistence outperforms LIM over the North Atlantic, a result consistent with the particularly strong autocorrelation of surface air temperature in this region in the IPSL-CM6A-LR model (Fig. A7). LIM reveals pronounced spatial disparities in the predictive skill. It performs particularly well in the tropics, especially over the western Pacific, where correlation scores exceed 0.7 (Fig. 3c; Table 2). The observed zonal asymmetry in skill across the tropical Pacific may be explained by the fact that variability in the eastern Pacific is more nonlinear (Sun et al., 2015; Zhao and Sun, 2022), which can limit the effectiveness of linear models in capturing complex dynamics. Moreover, predictive skill is generally higher over the oceans than over land, reflecting the longer memory and greater persistence of oceanic processes, whereas variability over continents is more strongly influenced by short-lived atmospheric fluctuations (Boer, 2000; Bellucci et al., 2015).

Replacing LIM with a RC model leads to consistent improvements across all performance metrics (Table 2). Unlike LIM, RC outperforms Persistence for the NH index. Figure 3e-f illustrates a general, albeit moderate increase in the spatial distribution of predictive skill, indicating that RC's memory and nonlinear processing capabilities contribute to enhanced forecast performance. Along the second axis of improvement, the architectural simplification introduced in RCn yields additional gains, with skill levels clearly exceeding those of LIM (Table 2). RCn outperforms both LIM and Persistence over

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

the North Atlantic and, more notably, over the Southern Hemisphere (Fig. 3g–h), where correlation scores exceed 0.6 (Table 2). Further gains are also apparent in the tropics, particularly in the correlation patterns (Fig. 3g).

The first two axes of improvement already lead to substantial progress in emulation skill. The third development concerns dimensionality reduction. Using the AE alone for both dimensionality reduction and prediction results in further enhancements in model effectiveness. In particular, a clear increase in Niño 3.4 correlation and a marked reduction in Niño 3.4 RMSE are observed (Table 2). Spatially, better scores are especially pronounced over central and eastern tropical Pacific and over Southern Ocean (Fig. 3i-j). Although the AE used here is linear, it nevertheless outperforms the RCn configuration, indicating that the choice of dimensionality reduction method can have a stronger impact on emulation performance than the choice of the prediction algorithm itself.

Finally, combining all three axes of improvement into the AERCn configuration yields the best overall emulator. AERCn consistently outperforms LIM, Persistence, and all other emulators across all metrics and regions (Table 2). The global-mean of local correlation increases from 0.493 for LIM to 0.557 for AERCn. For the Niño 3.4 index, correlation improves from 0.593 to 0.724, and for the SH index, from 0.658 to 0.791, substantially exceeding Persistence (0.704). Spatially, many regions that previously exhibited only moderate skill with the LIM (Fig. 3c) now exceed correlation values of 0.5, as illustrated by the expanded yellow and orange areas in Fig. 3k. Overall, these results demonstrate that progressively increasing model complexity along the three identified axes leads to systematic and robust improvements in emulation skill, both spatially and across key climate indices, with AERCn providing the best overall performance.

Although global correlation and RMSE scores indicate strong overall performance, a closer examination reveals substantial temporal variability in emulation skill. The running RMSE of the 1-year lead-time predictions obtained by the LIM and AERCn emulators, for the GMST and Niño 3.4 indices exhibits notable fluctuations revealing distinct phases of persistent increased or decreased prediction error at decadal to secural timescale (Fig. 4a-b). In particular, increases in the RMSE of the prediction for the GMST index mostly coincide with peaks in stratospheric volcanic aerosol concentration as depicted by the global mean stratospheric aerosol optical depth, suggesting that these variations are largely attributable to episodes of major volcanic eruptions (Fig 4a-b). Consistently, the *past2k* simulations including variations of the Solar Irandiance only (*past2k-SSI*), display a significant reduction in the amplitude of GMST multi-decadal to secular prediction skills variations (Fig. A8a). In contrast, for the Niño 3.4 index, temporal fluctuations in RMSE do not appear to be related to volcanic activity (Fig 4b), as similar multi-decadal prediction skill variations are observed even in the absence of eruptions (Fig. A8b). These fluctuations are instead primarily attributable to the intrinsic changes in the amplitude of ENSO events over time as shown by the significant correlation between, the 30-year running RMSE of the Niño 3.4 index prediction and the running standard deviation of the Niño 3.4 index of the *past2k* simulation itself (Fig. 4c). This relationship is particularly strong for the LIM (correlation of 0.67), which tends to underestimate or fails to predict extreme ENSO events and thus explains less variance during periods of intense ENSO activity. By contrast, this relationship is much less pronounced for the AERCn (correlation of 0.38), illustrating overall better skills at predicting the full range of amplitude of ENSO variability.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

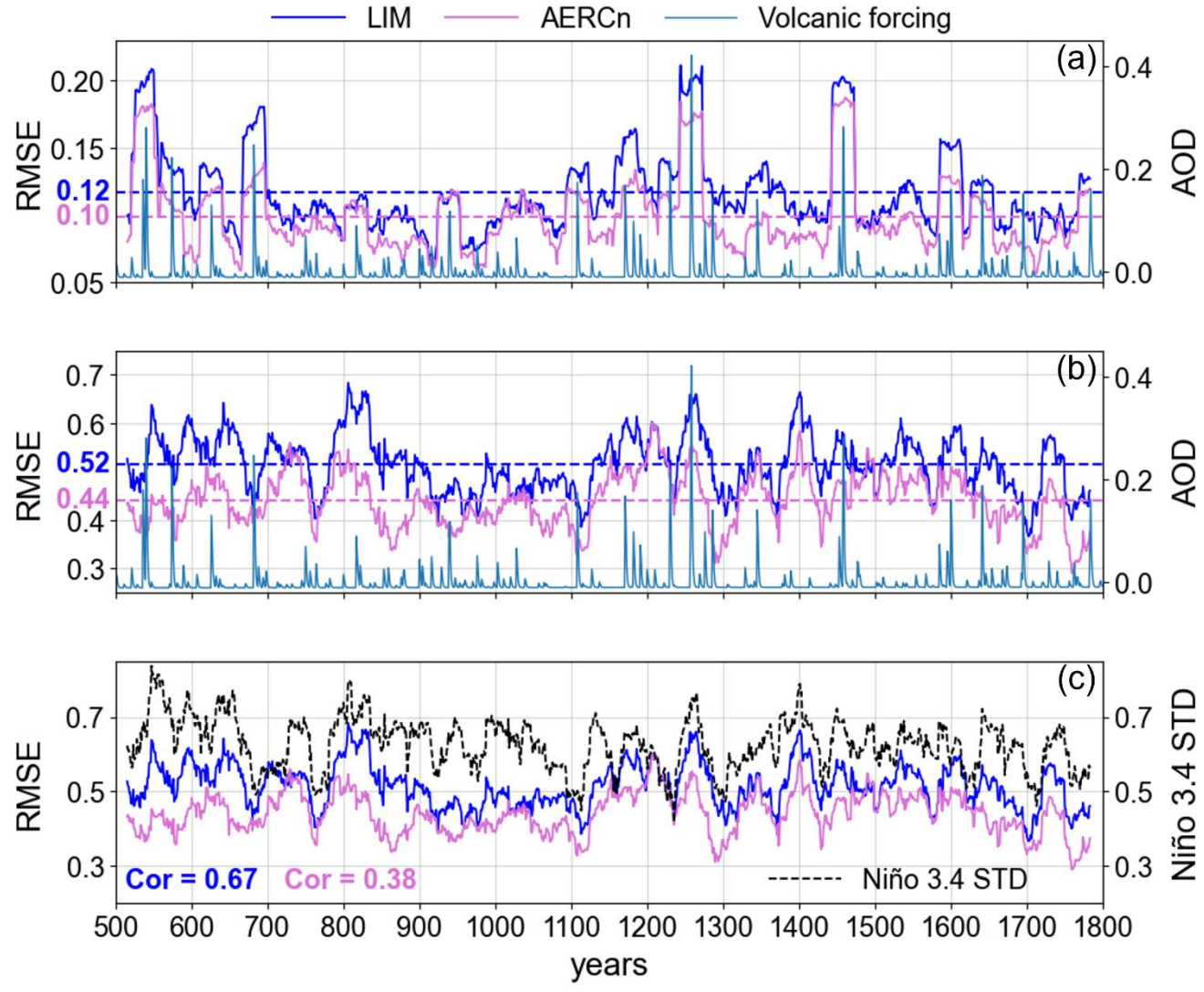


Figure 4: Temporal evolution of prediction error. Thirty-year rolling RMSE of one-year lead forecasts for LIM (blue) and AERCn (pink) is shown for (a) GMST and (b) Niño3.4 in the past2k experiment. Dotted lines denote period-mean RMSE. Volcanic AOD is overlaid (cyan) in panels (a–b). Panel (c) same as (b) adding a comparison to the thirty-year rolling standard deviation of the Niño3.4 target shown with the black dotted line. Correlations between LIM (blue) and AERCn (pink) RMSE and the Niño3.4 index are indicated.

5.2.2. CMIP6-class models

The results from the previous section, based on simulations of the IPSL-CM6A-LR over the CE, demonstrate the ability of our emulators to well capture the dynamics of a CMIP6-class model. In this section, we generalized this approach to a full set of CMIP6 models (Table A1) under the evaluation protocol defined in Sect. 4.2-4.3.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

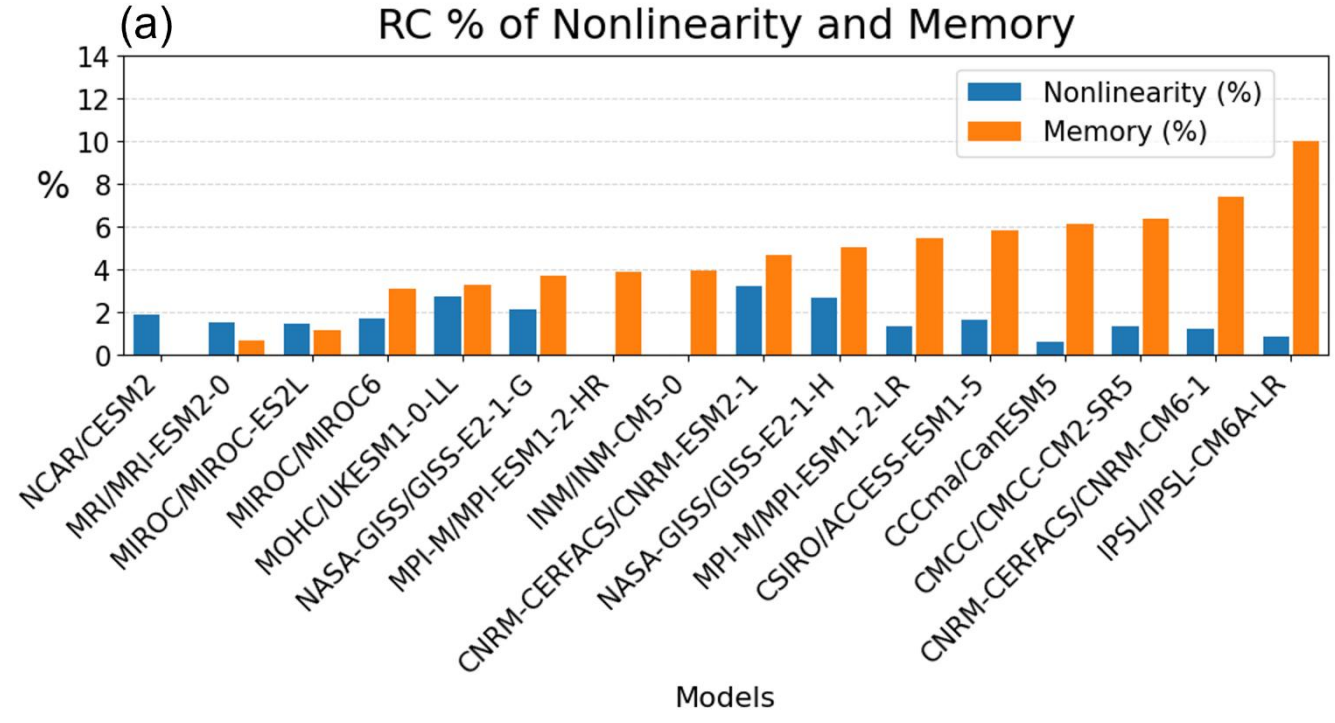
Each climate model is governed by distinct physical parameterizations, and exhibits specific biases and dynamical behaviors (Zhang et al. 2023; Richter & Tokinaga, 2020; Zhu et al. 2020). For instance, the IPSL-CM6A-LR model is known to exhibit a strong multidecadal variability in the North Atlantic (Boucher et al., 2020). Moreover, combining simulations from multiple models often yields performance metrics that better align with observed climate variability, due to a partial compensation of individual model errors and structural uncertainties (Ahmed et al., 2020; Parsons et al., 2021) stressing the importance of emulating the full diversity of physical processes represented across the model ensemble. Based on these considerations, the comparative analysis across CMIP6-class models is guided by several key questions:

- Can RC-based emulators adapt to models with varying degrees of nonlinearity?
- To what extent does the size of the training set influence predictive skill, and is the hierarchy of emulator performance observed with the IPSL-CM6A-LR model preserved across the full set of CMIP6-class models?

Sect. 5.2.1. showed that replacing LIM with RC improves emulator performance, suggesting a beneficial role for RC's memory and nonlinear processing. Since the IPSL model is relatively linear, it is not clear whether these gains arise mainly from memory or from nonlinear effects. We expect the balance between the two to vary with the degree of nonlinearity in the considered model emulated. This raises the following question: can the RC adapt to the characteristics of each model, selectively leveraging its memory and nonlinear features when needed? To investigate this, we rely on the RC variants introduced in Sect. 2.2.2., and follow the experimental protocol described in Sect 4.2.1.

For this first experiment, we only consider 16 CMIP6 models holding at least 10 historical ensemble members in order to ensure sufficient training data for RC. The impact of training set size on emulator performance will be explicitly assessed in a subsequent analysis. Moreover, because we focus on nonlinear dynamics in the tropical Pacific, as quantified by the nonlinearity coefficient α (Sect. 4.3.3.), we first evaluate RC skill in this region. Figure A9 shows that the spatially averaged correlation exceeds 0.6 between the obtained 1-year lead time prediction and the piControl target over the tropical Pacific for most of the 16 CMIP6 models including IPSL-CM6A-LR, consistent with the results obtained with the *past2k* simulations (Table 2). We then proceed to a skill attribution analysis with the percentage of skill gain attributed to RC dynamical memory and nonlinear processing over the tropical Pacific region (Fig. 5a). These contributions vary substantially across models. In the case of IPSL-CM6A-LR, approximately 10 % of the RC skill can be attributed to memory effects. By contrast, for models such as CESM2, nonlinear processing contributes more to predictive skill than memory. On average across models, memory accounts for a larger fraction of the skill gain (around 5 %) than nonlinearities (around 3 %). Figure 5b presents a linear regression between the fraction of skill gain attributed to RC memory and the coefficient α, used here as a proxy for model nonlinearity associated with ENSO. The contribution of memory decreases with increasing α, indicating an inverse relationship between memory-driven skill and the degree of nonlinearity in the tropical Pacific. In other words, the more linear the model, the larger the contribution of memory to RC performance. This result confirms that RC adapts to the dynamical characteristics of the model it emulates, relying preferentially on dynamical memory or nonlinear processes depending on the degree of linearity of the underlying system.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026



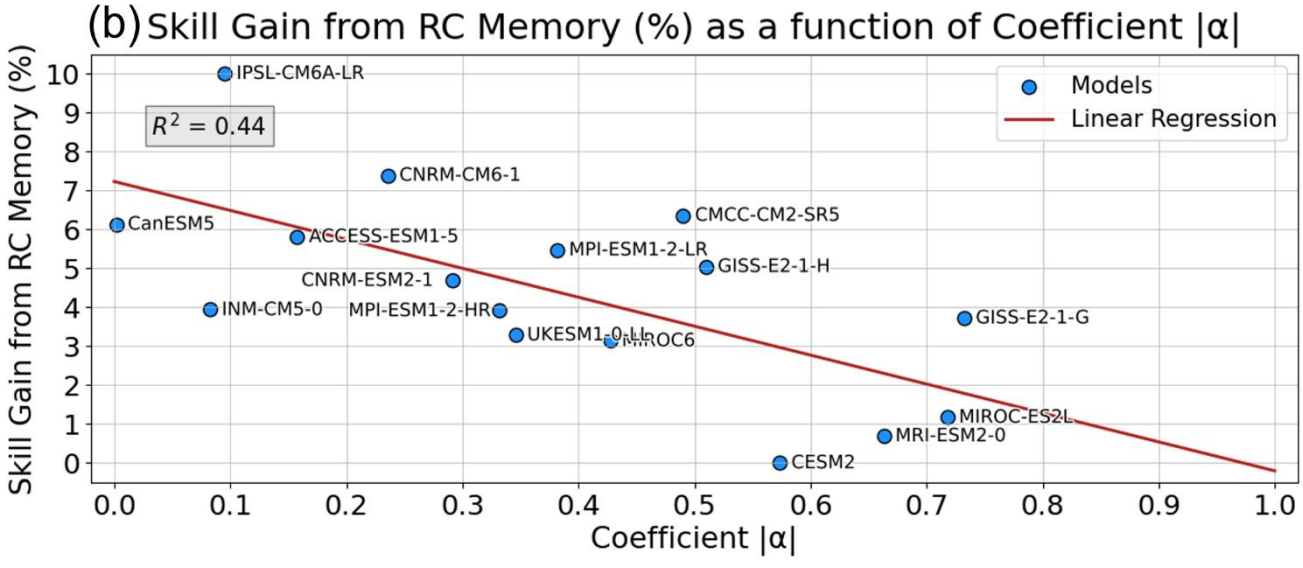


Figure 5: Skill gains attributed to reservoir computing (RC) memory and nonlinearities. Panel (a) shows the percentage improvement in spatially averaged tropical Pacific correlation relative to LIM across 16 CMIP6 models. Panel (b) presents the linear regression between memory-related skill gain and the nonlinearity indicator |α|.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

As previously noted, this analysis was restricted to models with at least 10 historical ensemble members to ensure sufficient training data. This naturally raises the second question: to what extent does the size of the training set influence predictive skill, and is the hierarchy of emulators performance observed with the IPSL-CM6A-LR model preserved across the full set of CMIP6-class models?

In this second experiment, we extend the analysis to all 52 CMIP6 models, training the emulator with progressively increasing subsets of historical ensemble members, following the protocol of Sect. 4.2.1. Figure 6 shows the 1-year lead-time correlation between the surface temperature prediction and the target as a function of the length of the training period across the full set of CMIP6 models. The LIM reaches its optimal performance relatively quickly regardless of the model (Fig. 6a-b), requiring only 495 years of training data (equivalent to 3 historical members). In contrast, for all RC-based methods, predictive skill strongly depends on the size of the training set (Fig. 6a, c, d, f). On average, 1320 years of training (about 8 members) are needed for RC-based emulators to perform as well as the LIM. Similarly, the AE skill stabilizes around 1320 years, but it requires only about 3 members (495 years) to reach performance levels comparable to the LIM (Fig. 6a, e). Overall, the best-performing emulator depends on the amount of training data available: the LIM performs best when the training set is very limited (fewer than 495 years), the AE is superior for intermediate training sizes (up to approximately 2310 years), and the AERCn becomes the most effective option when training data are abundant (Fig. 6a). In this last case, the hierarchy among emulators is consistent with the results presented in Sect. 5.2.1. (Fig 3; Table 2).

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

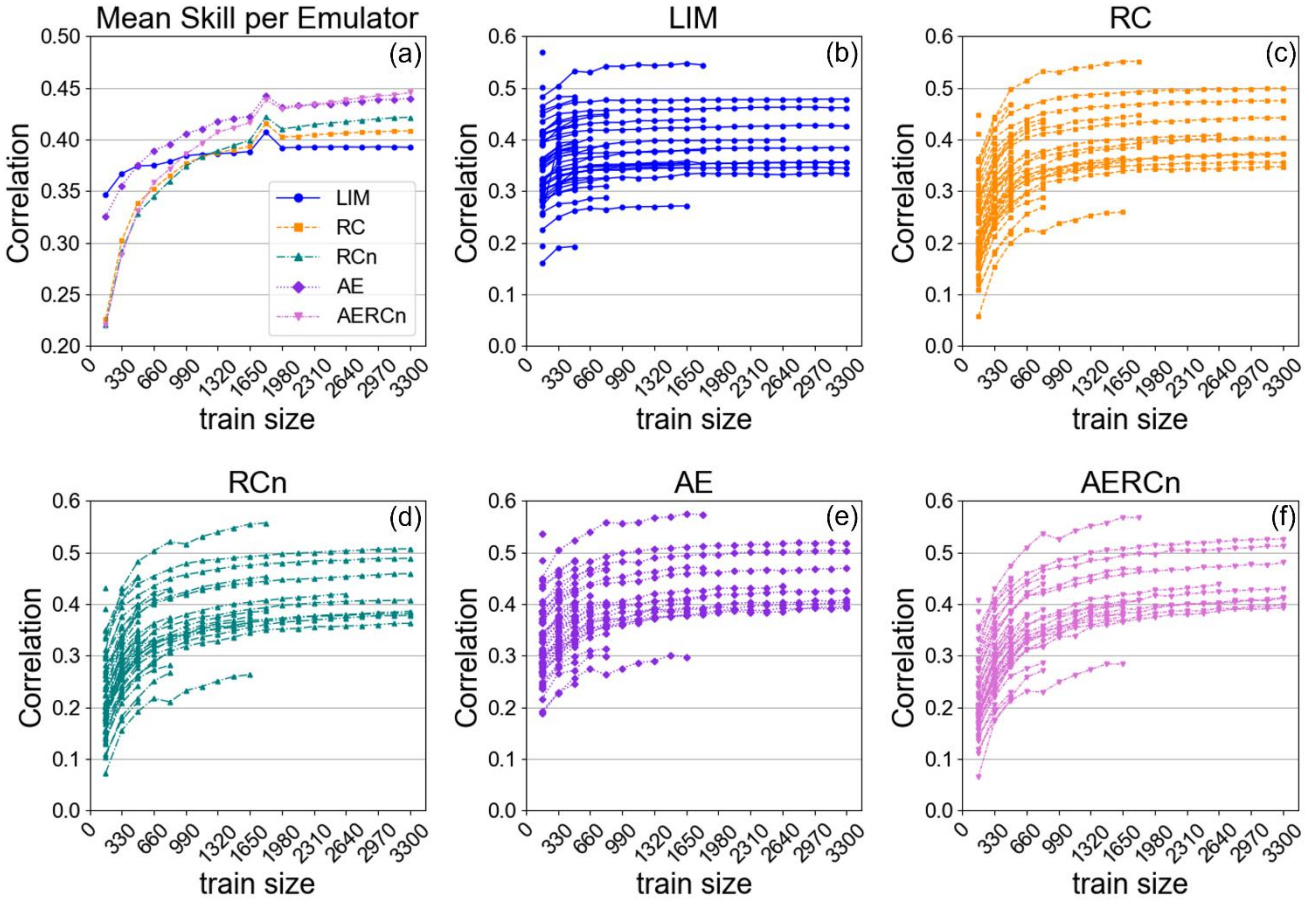


Figure 6: Sensitivity of forecast skill to training data length. Panel (a) shows the spatially averaged correlation across CMIP6 models as a function of training size (years). Panels (b–f) display model-specific results for LIM, RC, RCn, AE, and AERCn.

However, these results also reveal the limitations of the emulators in outperforming the LIM when training data are scarce. To address this issue, we investigate a transfer learning approach designed to mitigate the impact of limited training sets, following the experimental protocol described in Sect. 4.2.2. Since RC cannot be used in a CMIP-multi-model ensemble (MME) setting due to its non-trainable nature, we replace it with an RNN (Sect. 2.2.3.), enabling the construction of a CMIP6-multi-model equivalent of AERCn, referred to as AERNN. This results in two multi-model emulators, denoted AE_MME and AERNN_MME. They are first pre-trained using one historical ensemble member from each of the 52 CMIP6-class models. This pre-training phase enables the emulator to learn a shared physical basis across a wide range of climates. The models are then fine-tuned individually for each target model, starting from the pre-trained weights.

Figure 7 highlights the clear advantage of multi-model and transfer learning approaches under data-limited conditions. It shows the correlation of the 1-year lead-time prediction with the target as a function of the length of the training period across the full set of CMIP6 models. The AE_MME and AERNN_MME curves start at significantly higher skill levels compared to their their single-model counterparts (Fig. 7a) regardless of the CMIP6 host-model (Fig. 7c-f). Overall, the

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

AERNN_MME emerges as the best-performing emulator when training data are very limited (<660 years; Fig. A10). However, when training data are abundant, it is noteworthy that the multi-model version offers no advantage over the classic AERNN (Fig. 7a; Fig. A10).

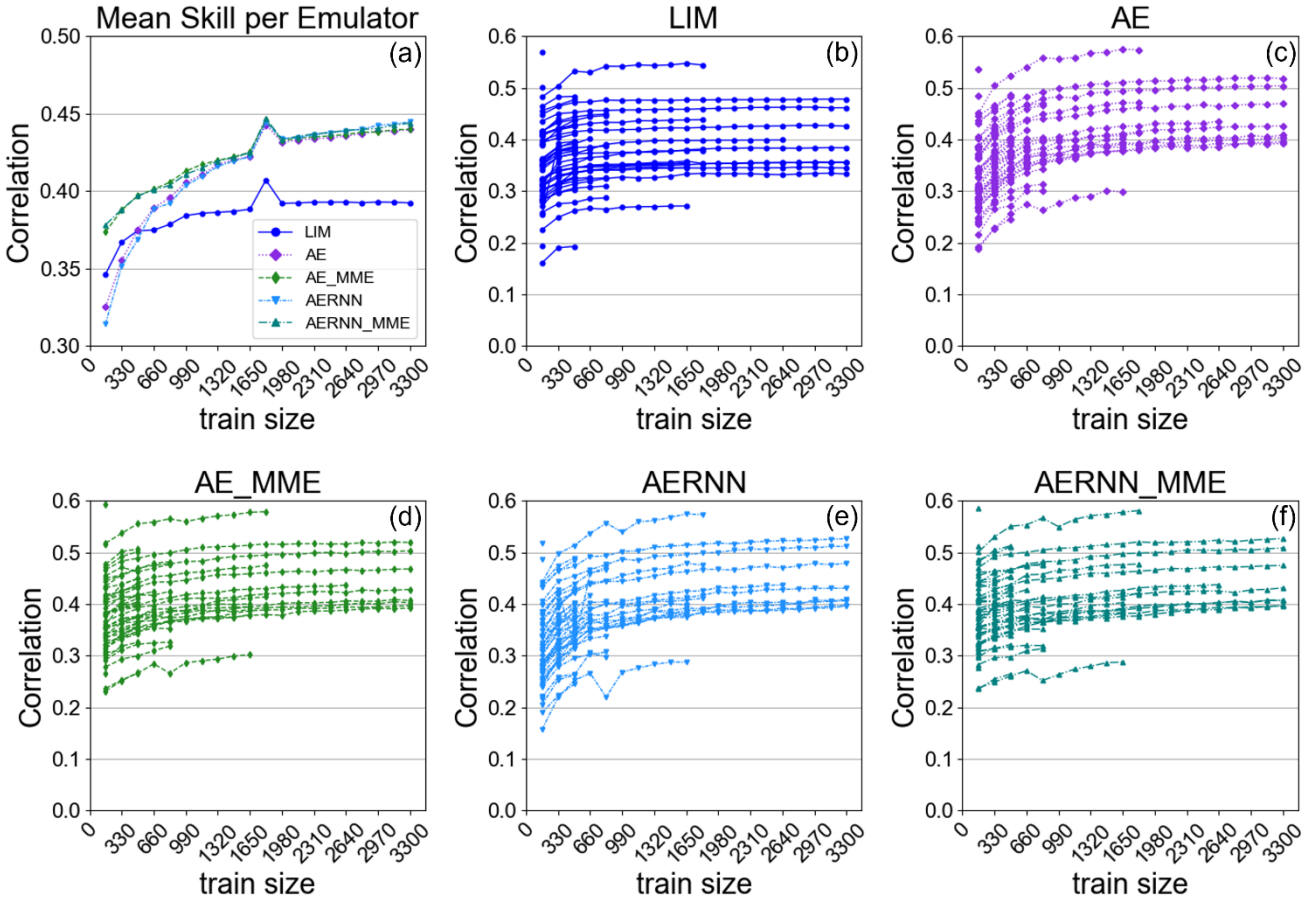


Figure 7: Impact of transfer learning on forecast skill. (a) Spatially averaged correlation is shown for LIM, AE, AE_MME, AERNN, and AERNN_MME configurations across CMIP6 models as a function of training size (years). Panels (b–f) show model-specific results.

5.3 Case study on extreme ENSO events

Sections 5.1–5.2 established that the AERCn configuration achieves the highest predictive skill among all emulators when sufficient training data are available, consistently outperforming the LIM across the suite of metrics considered. However, a key aspect remains unexplored: the ability of the emulators to reproduce extreme ENSO events, which constitute one of the primary motivations of this study and a known limitation of the LIM framework. In addition, their capacity to generate skillful ensemble predictions, rather than relying solely on deterministic forecasts, requires dedicated examination. Section 5.3 therefore focuses exclusively on extreme ENSO conditions and evaluates emulator performance under these high-

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

amplitude regimes. Section 5.3.1. examines deterministic forecast skill during extreme events, while Section 5.3.2. extends the analysis to probabilistic performance based on ensemble predictions. The analysis follows the same experimental framework as in Sect. 4.2.1.: emulators are trained on the historical ensemble members and evaluated in a perfect-model setting using the corresponding piControl simulations. We focus on two CMIP6 models with markedly different degrees of tropical Pacific nonlinearity: IPSL-CM6A-LR, which exhibits relatively linear behavior and a low α coefficient, and MIROC6, which displays a substantially larger α coefficient, indicative of stronger nonlinear dynamics and closer agreement with observations (Figure 5 and Fig. A11; Zhao and Sun, 2022). Both models provide 20 historical ensemble members, ensuring ample training data and allowing nonlinear configurations to be fully exploited. Section 5.3.1. first analyses spatial RMSE patterns for LIM and AERCn (the best-performing emulator in the data-rich regime) during extreme El Niño and La Niña years in both models. It then assesses Niño 3.4 forecast errors for all emulators (given the abundance of training data, multi-model configurations are not considered here) stratified by ENSO event type. Section 5.3.2. then extends the analysis to probabilistic forecasts by comparing ensemble predictions produced by the LIM and by AERCn following the protocol defined in Sect. 4.2.3.

5.3.1. Deterministic approach

Figure 8 shows, for the LIM and the AERCn, the RMSE of the one-year lead forecast of surface air temperatures of the IPSL-CM6A-LR and MIROC6 piControl simulations, computed exclusively during extreme El Niño years and Extreme La Niña years. Extreme warm and cold events are defined as Niño 3.4 anomalies exceeding one standard deviation (Santoso et al., 2017). For IPSL-CM6A-LR, the LIM exhibits relatively large errors in the eastern tropical Pacific during extreme events (Panels a,c), particularly for extreme La Niña years. AERCn substantially reduces the overall large-scale errors (Panels b,d), including the tropical Pacific although this region remains the most challenging. With MIROC6, LIM errors are globally comparatively smaller and in particular over the tropics (Panels e,g). Over the western Pacific, the skill during extreme El Niño years is consistent with the stronger persistence simulated in this region by the MIROC6 model (Fig. A12). AERCn nevertheless yields a systematic reduction in RMSE broadly and over the eastern tropical Pacific (Panels f,h), indicating a robust gain in skill under both extreme ENSO phases.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

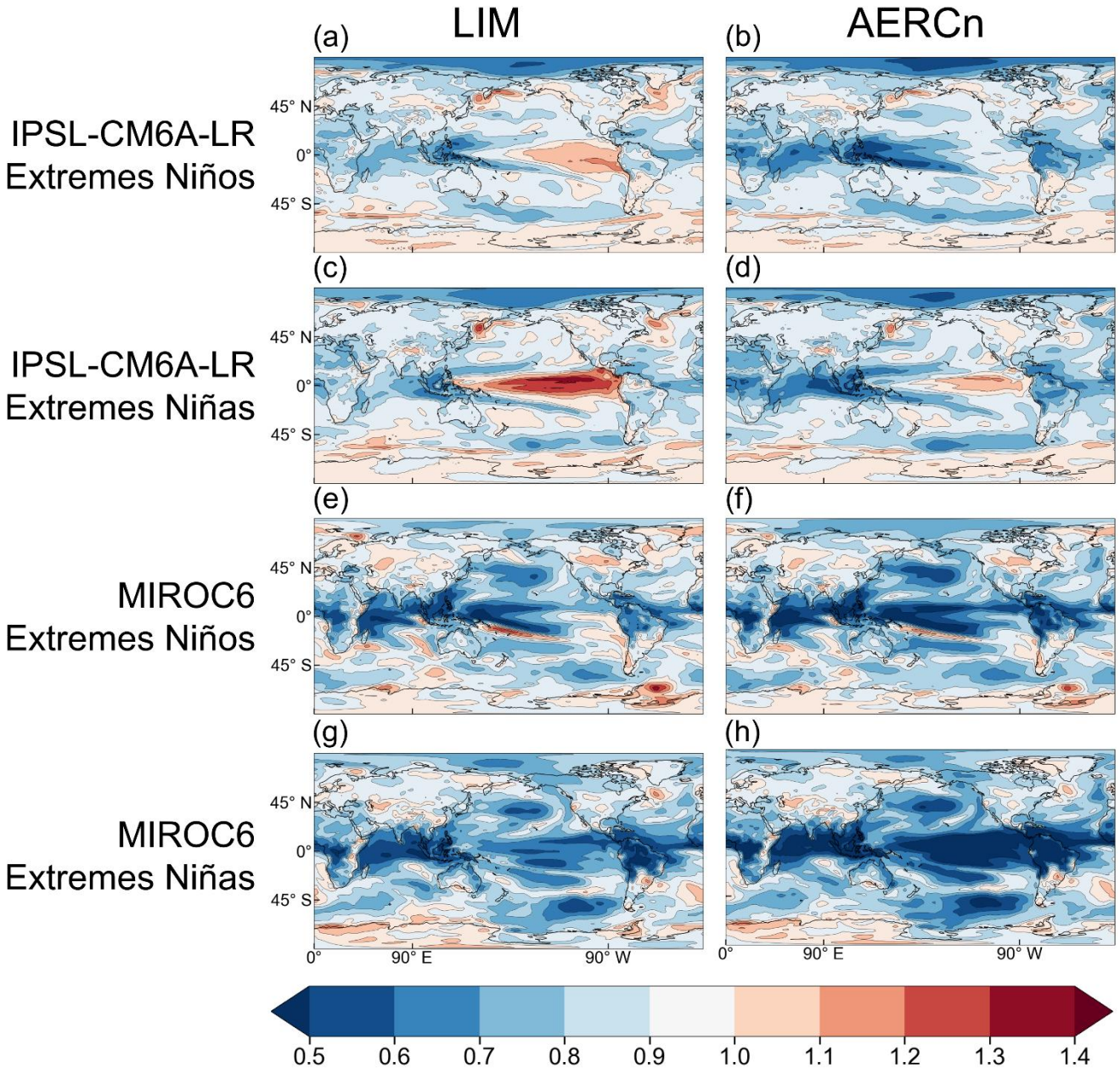


Figure 8: RMSE of one-year lead forecasts with target piControl simulations during extreme El Niño (a,b,e,f) and extreme La Niña (c,d,g,h) events. Results are shown for IPSL-CM6A-LR (a-d) and MIROC6 (e-f) using LIM (a,c,e,g) and AERCn (b,d,f,h).

To further characterise emulator performance at forecasting extreme ENSO events, we focus specifically on the Niño 3.4 index and categorize Niño 3.4 anomalies into three types: extreme El Niño, extreme La Niña, and Moderate events (defined as Niño 3.4 anomalies below ±1 standard deviation). Figure 9 presents the RMSE of the 1-year lead-time prediction obtained by the LIM, RC, AE, and AERCn, stratified over the three ENSO event types. As expected, prediction errors are larger for

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

extreme events and in both cases, replacing the LIM with the AERCn emulator substantially reduces RMSE, yielding an average improvement of approximately 20 % relative to the LIM. However, a notable contrast emerges between the two CMIP6-class models: errors are largest during extreme La Niña events for IPSL-CM6A-LR, whereas they peak during extreme El Niño episodes for MIROC6 (Fig. 9). This indicates a general asymmetry in skill between warm and cold events that depends on the host-model characteristics. In particular, these asymmetries reflect the distribution of extremes in each data-set: IPSL-CM6A-LR tends to produce stronger La Niña than El Niño events, leading to higher RMSE values when emulators attempt to reproduce these large-amplitude cold events. MIROC6 exhibits the opposite pattern. This result is consistent with broader findings for CMIP6-class models. Many models simulate ENSO dynamics that are too linear, often generating La Niña events that are excessively strong compared with observations (Zhao and Sun, 2022). Only a limited subset of CMIP6 models, including MIROC6, realistically captures ENSO asymmetry. The more linear ENSO dynamics of IPSL-CM6A-LR may therefore lead to an under-representation of ENSO asymmetry and to amplified cold anomalies, which in turn likely increases emulator forecast errors during extreme La Niña events.

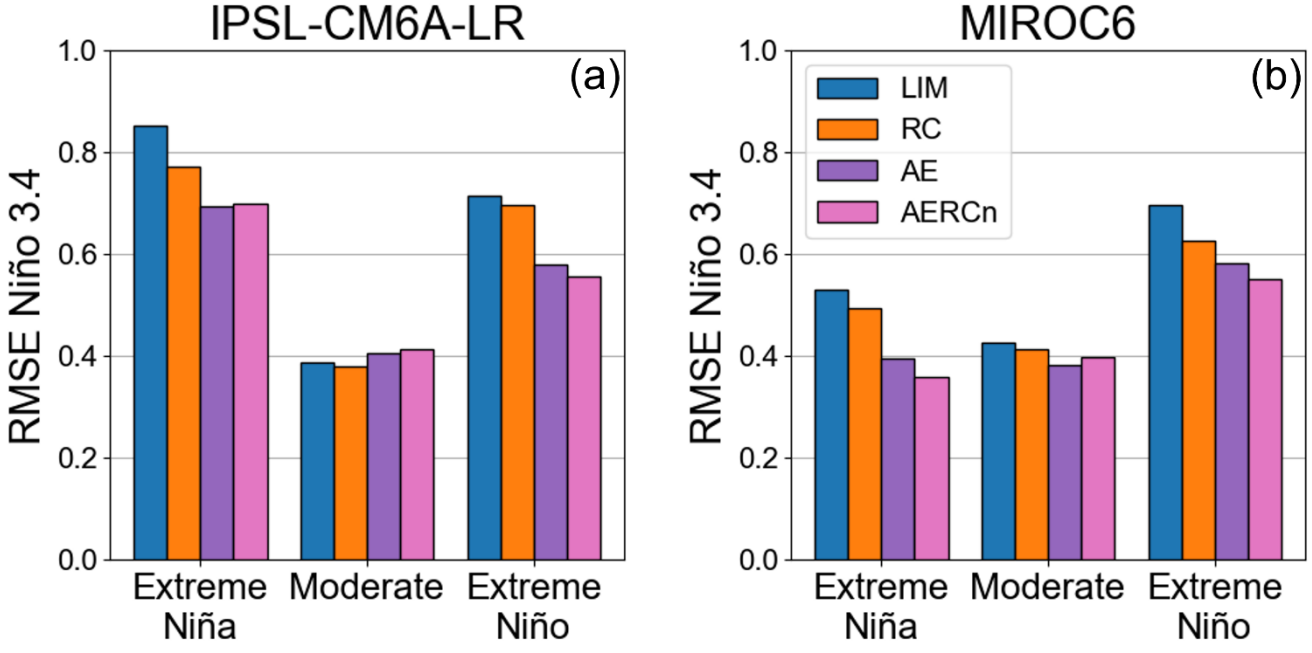


Figure 9: RMSE of one-year lead Niño3.4 forecasts for ENSO regimes. Errors are shown for extreme La Niña, neutral, and extreme El Niño conditions for (a) IPSL-CM6A-LR and (b) MIROC6 using LIM, RC, AE, and AERCn.

5.3.2. Ensemble approach

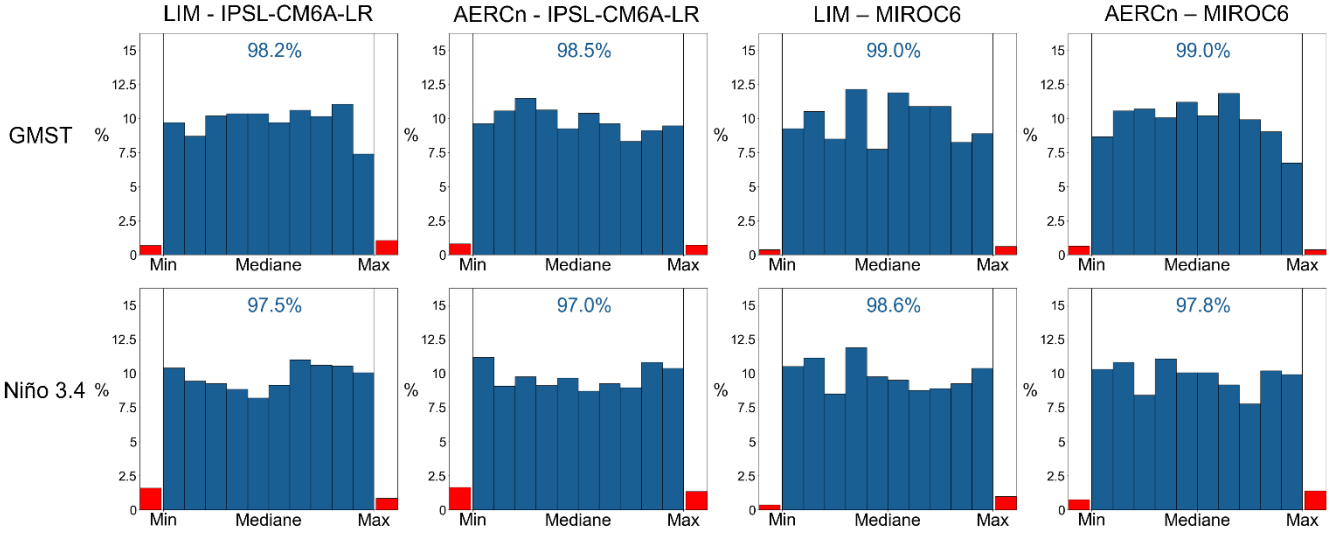
Accurately predicting the evolution of the climate system requires not only a reliable representation of its underlying dynamics, but also a robust quantification of uncertainties. Uncertainties arise from imperfect knowledge of initial conditions, unresolved nonlinear processes, and external forcing. In this context, confidence in purely deterministic forecasts

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

is limited. Relying solely on such forecasts may produce an overly narrow and potentially misleading representation of the range of possible future climate states.

To overcome this limitation, probabilistic approaches based on ensemble forecasting are required. Monte Carlo methods, such as Particle Filters, approximate the probability density of the climate system using an ensemble of particles (Jebri & Khodri, 2023). The predictive components and the associated ensemble spread are essential for representing the range of possible future states and for quantifying forecast uncertainty. The LIM framework naturally supports ensemble forecasting through the stochastic term included in its governing equations (Sect. 2.2.1.), allowing a direct transition from deterministic to probabilistic predictions. By contrast, AERCn does not include an intrinsic stochastic component. To generate ensembles with AERCn, we therefore perturb the initial conditions by adding spatially uncorrelated white noise to the input field, with an amplitude set to 2.5 times the grid point standard deviation of the anomalies.

To evaluate the ability of the LIM and AERCn models to span the plausible range of outcomes of the emulated system and to capture extreme events, we analyse ensemble rank histograms derived from probabilistic forecasts. Figure 10 presents rank histograms computed from 100-member ensemble forecasts at a 1-year lead time for the GMST and the Niño 3.4 index, using both LIM and AERCn on the IPSL-CM6A-LR and MIROC6 models. A flat rank histogram indicates a well-calibrated ensemble, meaning that the target trajectory is statistically indistinguishable from any ensemble member and behaves like one random realization of the system. Overall, both LIM and AERCn produce well-calibrated ensembles for the GMST and Niño 3.4 indices. The rank histograms are generally flat, and the target lies within the ensemble spread 97–99% of the time (Fig. 10a-h). MIROC6-based rank histograms show a slight concentration in the central ranks for the GMST index (Fig. 10c-d), suggesting a tendency toward over-dispersion, whereas the Niño-3.4 histograms display the opposite pattern, with somewhat elevated frequencies in the non-central in-spread bins, particularly for the LIM emulator (Fig. 10e, g).



https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Figure 10: Reliability of ensemble forecasts. Rank histograms for GMST (a–d) and Niño3.4 (e–h) using LIM and AERCn ensembles for IPSL-CM6A-LR and MIROC6. Blue bars indicate target frequency within ensemble deciles; red bars denote occurrences outside the ensemble spread. Percentages within the spread are indicated.

While rank histograms provide valuable insight into ensemble shape and calibration, they are insufficient for a comprehensive assessment of probabilistic forecast quality. To obtain a quantitative and statistically robust measure of ensemble skill, we therefore use the Continuous Ranked Probability Score (CRPS), a proper scoring rule that jointly evaluates forecast accuracy and calibration (Sect. 4.3.4.). Figure 11 displays the distribution of CRPS values computed during extreme ENSO years, defined as years in which the Niño 3.4 index exceeds ±1 standard deviation. For each extreme year, CRPS is calculated for the Niño-3.4 index (panels a–b) and for GMST (panels c–d). Results are shown separately for IPSL-CM6A-LR (a, c) and MIROC6 (b, d). Each violin represents the full distribution of CRPS values across extreme-event years. The width of the violin indicates the density of values: broader sections correspond to more frequent CRPS values, while narrow upper tails indicate rare but large forecast errors. Accordingly, AERCn consistently yields lower CRPS values than the LIM across all cases (including the upper tails), indicating superior probabilistic skill. This result demonstrates that AERCn produces more reliable and better-calibrated ensemble forecasts during extreme ENSO conditions, in agreement with the deterministic performance gains reported in Sect. 5.3.1.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

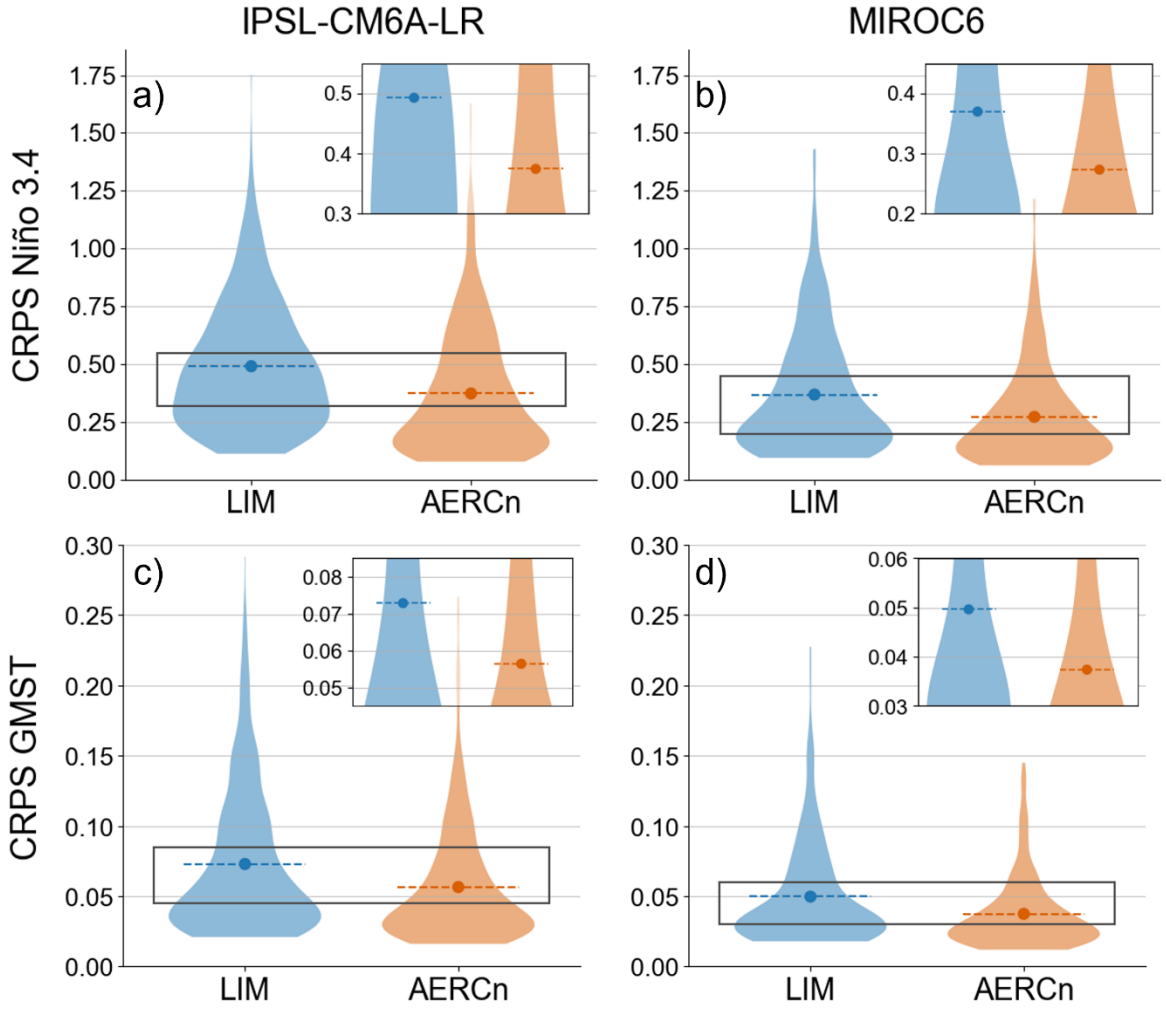


Figure 11: Distribution of Continuous Ranked Probability Score (CRPS) values during ENSO extremes. Panels show CRPS distributions for Niño3.4 (a, b) and GMST (c, d) during ±1σ Niño3.4 years for IPSL-CM6A-LR and MIROC6. Black rectangles mark the lower CRPS interval, which are magnified in the inset panels for clearer comparison of the bulk distributions. Lower CRPS values indicate improved probabilistic forecast performance.

6 Summary and Conclusions

Ensemble-based approaches such as PFs offer a powerful framework to reconstruct past climate dynamics over the Common Era. Such approaches necessitate the development of lightweight emulators capable of producing large ensembles over long periods. This study aims to design an improved forward model by addressing the key limitations of the baseline LIM-EOF configuration, which approximates climate variability using a Linear Inverse Model (LIM) applied to a latent space derived from Empirical Orthogonal Functions (EOFs). While prior studies have shown that ensemble-based LIM methods can reasonably emulate climate model dynamics, this approach remains inherently limited: its linearity prevents it from capturing asymmetric responses and nonlinear extremes (such as ENSO events), it lacks memory, and the EOF-based dimensionality

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

reduction prioritizes variance rather than predictability. In this study, we developed and evaluated a hierarchy of emulators designed to overcome these limitations. We explore three axes of improvement over LIM-EOF: (1) replacing the LIM dynamic model with a RC; (2) merging dimensionality reduction and prediction into a unified architecture; (3) using an AE instead of EOFs to reduce the dimension while retaining the predictable component of the system. We also introduce a fourth strategy to address limited training data scenarios, which involves pre-training a new AERNN_MME emulator on all CMIP6 models before fine-tuning it on the short training set of the targeted host-model.

Evaluated in a perfect-model framework across CMIP6-class model simulations, these innovations yield substantial gains over the classical LIM–EOF configuration. By explicitly capturing both dynamical memory and nonlinear interactions, RC outperforms the LIM, most notably in the tropical Pacific, where convection and nonlinear processes are prominent. It is possible to isolate the respective contributions of memory and nonlinearity to RC performance, and these contributions vary markedly across models. In particular, the importance of memory declines as the degree of nonlinearity increases. In the tropical Pacific, RC therefore relies more heavily on its dynamical memory when the host model behaves more linearly. For IPSL-CM6A-LR, memory accounts for up to 10 % of the total prediction skill in the tropical Pacific.

In addition to the dynamical improvements, the structural flexibility of RC permits a reorganization of the emulator that enhances predictive performance. Specifically, integrating dimensionality reduction and prediction into a single step mitigates error propagation across processing stages. The high latitudes of the Southern Hemisphere illustrate this effect particularly well: this region experience substantial information loss during dimensionality reduction, yet show the strongest skill gains when moving from RC to RCn. Taken together, these first two improvements yield a marked increase in predictive performance, with RCn consistently outperforming both LIM and Persistence across all evaluation metrics.

The third axis of improvement leads to an even larger gain in predictive skill. It highlights the central role of the latent space: the quality of the dimensionality reduction strongly shapes the performance of the downstream prediction models. In this respect, the standard EOF-based reduction is too restrictive for forecasting, as it prioritizes variance rather than predictability. In contrast, the AE introduced here is explicitly trained to retain the predictable component of the system. Combining all three improvements produces the AERCn model, which consistently outperforms the other architectures when sufficient training data are available. AERCn also delivers well-calibrated one-year ensemble forecasts, making it a strong candidate for use as a forward model in PF data-assimilation frameworks.

The ENSO case study confirms the ability of AERCn to better capture extreme climate events. The ensemble forecasts produce well-calibrated ensemble, with flat rank histograms and target inclusion rates between 97% and 98%, indicating a faithful reproduction of ENSO variability. Furthermore, AERCn achieves substantially lower CRPS values than the LIM during extreme episodes, pointing to clearer improvements in probabilistic forecast skill.

Although AERCn delivers the strongest performance in data-rich settings, its reliance on large training sets highlights the need of complementary strategies. When data are limited, replacing the RC with an RNN becomes advantageous, as it enables pretraining across multiple CMIP6 models. This allows the AERNN to learn dynamical features shared across models before being fine-tuned on a specific target simulation. Such transfer-learning capabilities, inherent to neural

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

architectures, offer a clear advantage over classical approaches such as LIM. They also open the possibility of training a meta-AERNN on long multi-model simulations and subsequently adapting it to observational datasets, potentially yielding emulators that more closely reflect real-world climate dynamics. For applications focused on a single host model with sufficient training data, however, AERCn remains sufficient.

While our results were obtained within a perfect-model framework, they provide a strong foundation for advancing emulator-based paleoclimate reconstruction. Emulator performance naturally reflects the dynamical properties and structural biases of the underlying CMIP6 simulations, for example, the high skill in the North Atlantic for IPSL-CM6A-LR arises from its pronounced autocorrelation in that region. Future work will integrate AERCn and AERNN within PF data-assimilation systems, extend the predictor set to additional climate variables, and rigorously assess the influence of model and training-data biases. The fine-tuning capacity of AERNN on observational records, in particular, offers a promising path toward improving emulator realism and strengthening future PDA applications

7. Appendix A

Model	piControl years	Historical members	Total historical years
AS-RCEC/TaiESM1	500	2	330
AWI/AWI-CM-1-1-MR	500	5	825
AWI/AWI-ESM-1-1-LR	100	1	165
BCC/BCC-CSM2-MR	600	3	495
BCC/BCC-ESM1	451	3	495
CAMS/CAMS-CSM1-0	500	3	495
CAS/CAS-ESM2-0	550	4	660
CAS/FGOALS-f3-L	561	3	495
CCCR-IITM/IITM-ESM	249	1	165
CCCma/CanESM5	1051	20	3300
CCCma/CanESM5-CanOE	501	3	495
CMCC/CMCC-CM2-SR5	500	11	1815
CMCC/CMCC-ESM2	500	1	165
CNRM-CERFACS/CNRM-CM6-1	500	20	3300
CNRM-CERFACS/CNRM-CM6-1-HR	300	1	165
CNRM-CERFACS/CNRM-ESM2-1	500	11	1815
CSIRO-ARCCSS/ACCESS-CM2	500	3	495
CSIRO/ACCESS-ESM1-5	1000	20	3300

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

E3SM-Project/E3SM-1-0	500	5	825
E3SM-Project/E3SM-1-1	251	1	165
EC-Earth-Consortium/EC-Earth3-AerChem	311	3	495
EC-Earth-Consortium/EC-Earth3-CC	505	4	660
EC-Earth-Consortium/EC-Earth3-Veg-LR	501	3	495
FIO-QLNM/FIO-ESM-2-0	575	3	495
HAMMOZ-Consortium/MPI-ESM-1-2-HAM	1000	3	495
INM/INM-CM4-8	531	1	165
INM/INM-CM5-0	1201	10	1650
IPSL/IPSL-CM6A-LR	2000	20	3300
MIROC/MIROC-ES2L	500	20	3300
MIROC/MIROC6	800	20	3300
MOHC/HadGEM3-GC31-LL	2000	5	825
MOHC/HadGEM3-GC31-MM	500	4	660
MOHC/UKESM1-0-LL	1880	16	2640
MPI-M/ICON-ESM-LR	500	5	825
MPI-M/MPI-ESM1-2-HR	500	10	1650
MPI-M/MPI-ESM1-2-LR	1000	10	1650
MRI/MRI-ESM2-0	701	11	1815
NASA-GISS/GISS-E2-1-G	851	20	3300
NASA-GISS/GISS-E2-1-G-CC	165	1	165
NASA-GISS/GISS-E2-1-H	801	20	3300
NASA-GISS/GISS-E2-2-H	251	5	825
NCAR/CESM2	1200	11	1815
NCAR/CESM2-FV2	500	3	495
NCAR/CESM2-WACCM	499	3	495
NCAR/CESM2-WACCM-FV2	500	3	495
NCC/NorESM2-LM	501	3	495
NCC/NorESM2-MM	500	1	165
NIMS-KMA/KACE-1-0-G	450	3	495
NOAA-GFDL/GFDL-CM4	500	1	165
NOAA-GFDL/GFDL-ESM4	500	3	495
NUIST/NESM3	500	5	825

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

SNU/SAM0-UNICON	700	1	165
THU/CIESM	500	3	495
UA/MCM-UA-1-0	500	2	330
AS-RCEC/TaiESM1	500	2	330
AWI/AWI-CM-1-1-MR	500	5	825
AWI/AWI-ESM-1-1-LR	100	1	165
BCC/BCC-CSM2-MR	600	3	495
BCC/BCC-ESM1	451	3	495
CAMS/CAMS-CSM1-0	500	3	495
CAS/CAS-ESM2-0	550	4	660

Table A1: List of CMIP6 models used in this study, including the available simulation lengths (years) for piControl and historical experiments.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

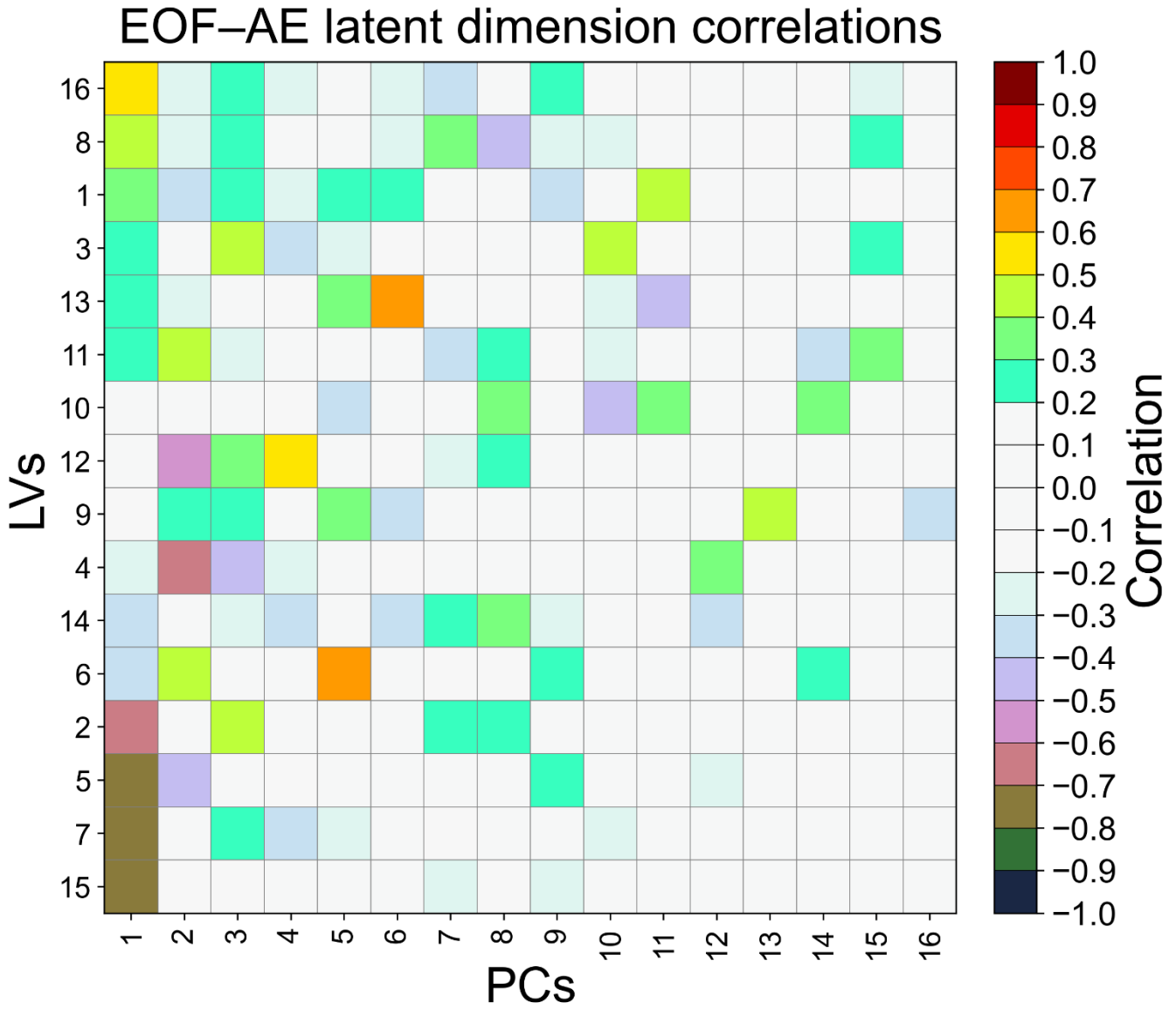


Figure A1: Correlation matrix between the AE latent vectors and EOF principal components using a 30-dimensional latent space. AE latent vectors (vertical axis) are ordered by decreasing correlation with EOF1.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

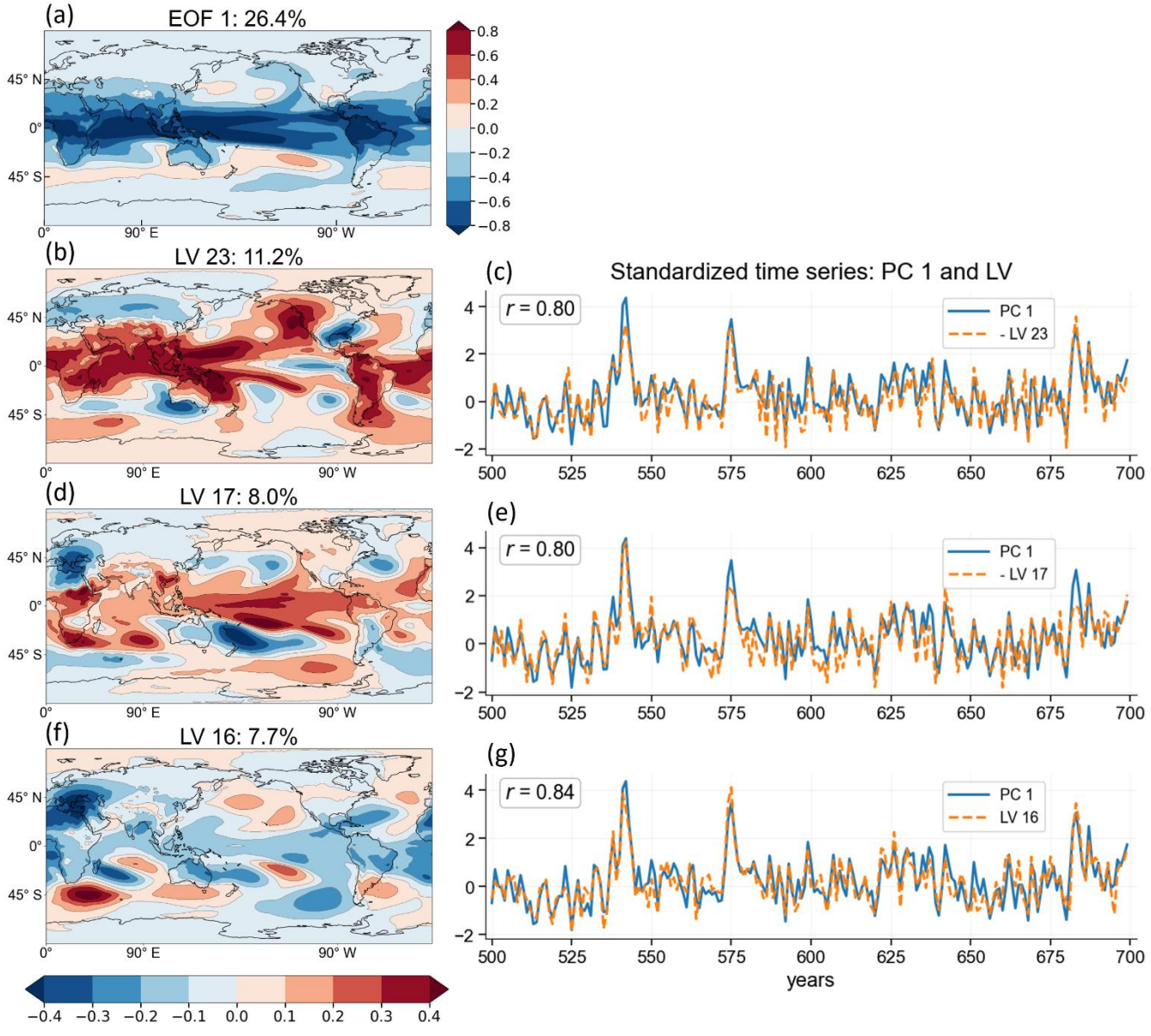


Figure A2: Variance structures associated with EOF1 and AE latent vectors using a 30-dimensional latent space. Panel (a) shows the local variance explained by EOF1. Panels (b, d, f) display variance explained by AE latent vectors selected based on correlation ≥ 0.7 with EOF1. Panels (c, e, g) show the corresponding standardized time series over the first 200 years. Correlation coefficients are indicated.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

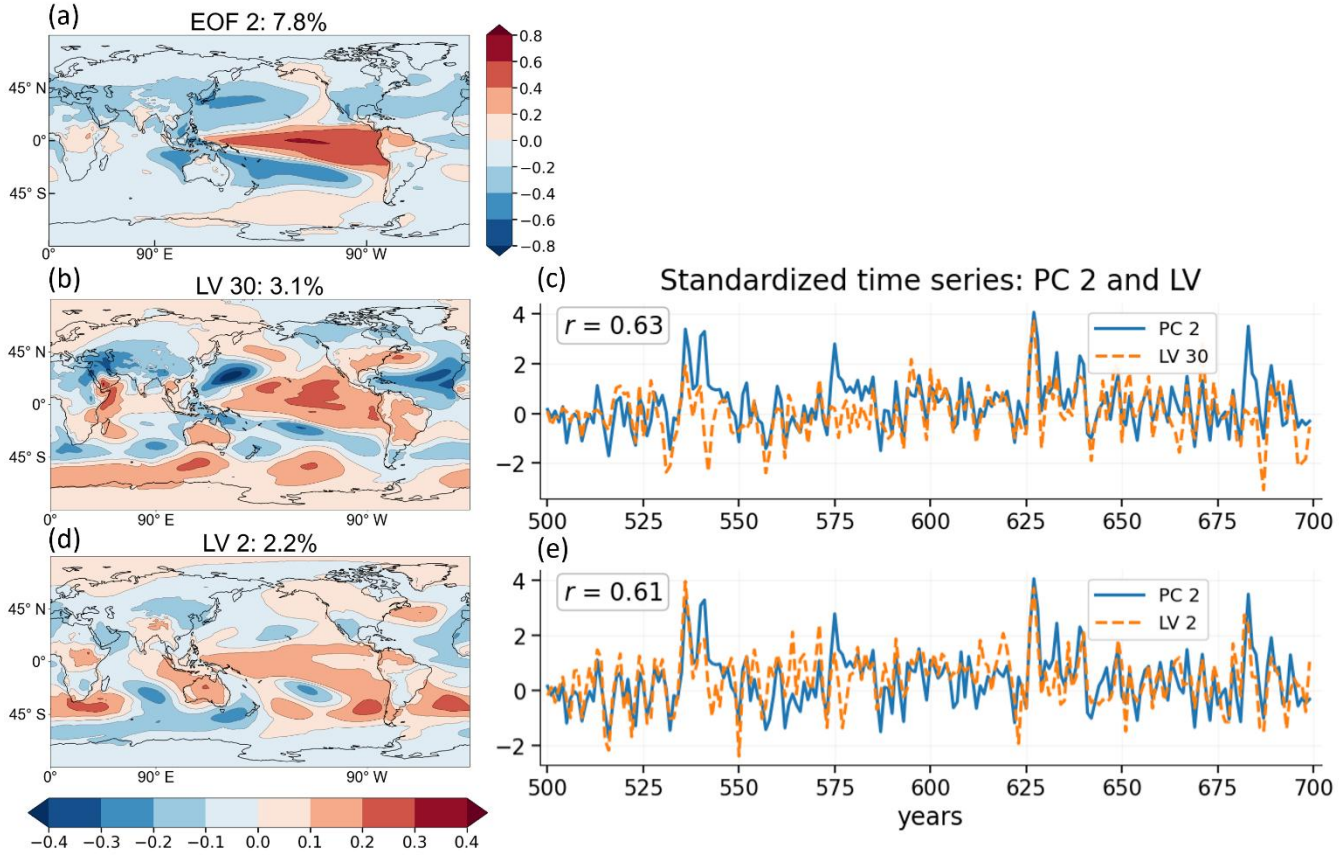


Figure A3: Variance structures associated with EOF2 and AE latent vectors using a 30-dimensional latent space. Panel (a) shows the local variance explained by EOF2. Panels (b, d) display variance explained by AE latent vectors selected based on correlation ≥ 0.6 with EOF2. Panels (c, e) show the corresponding standardized time series over the first 200 years. Correlation coefficients are indicated.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

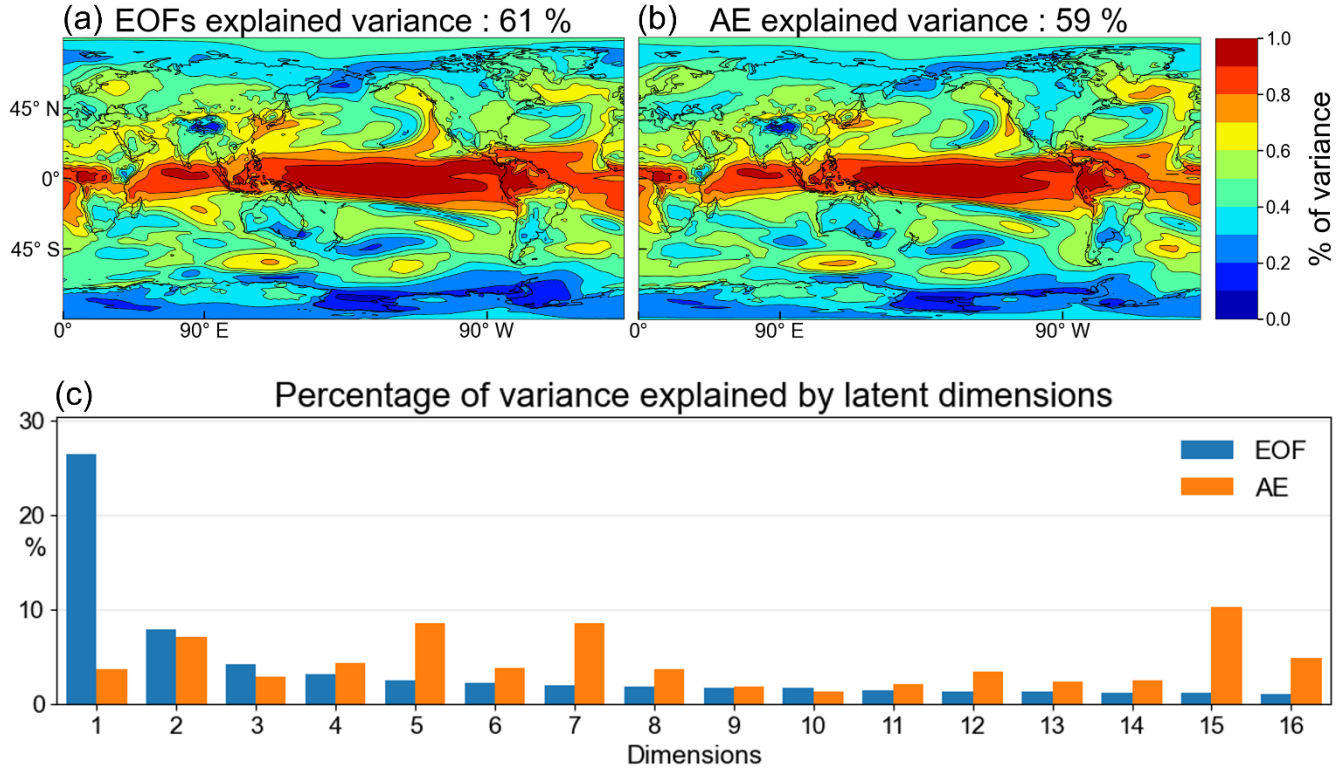


Figure A4: Spatial distribution of locally explained variance using a 16-dimensional latent space. Panels (a) and (b) show variance explained by EOF and AE representations, respectively. Panel (c) presents the variance contribution of each latent dimension.

775

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

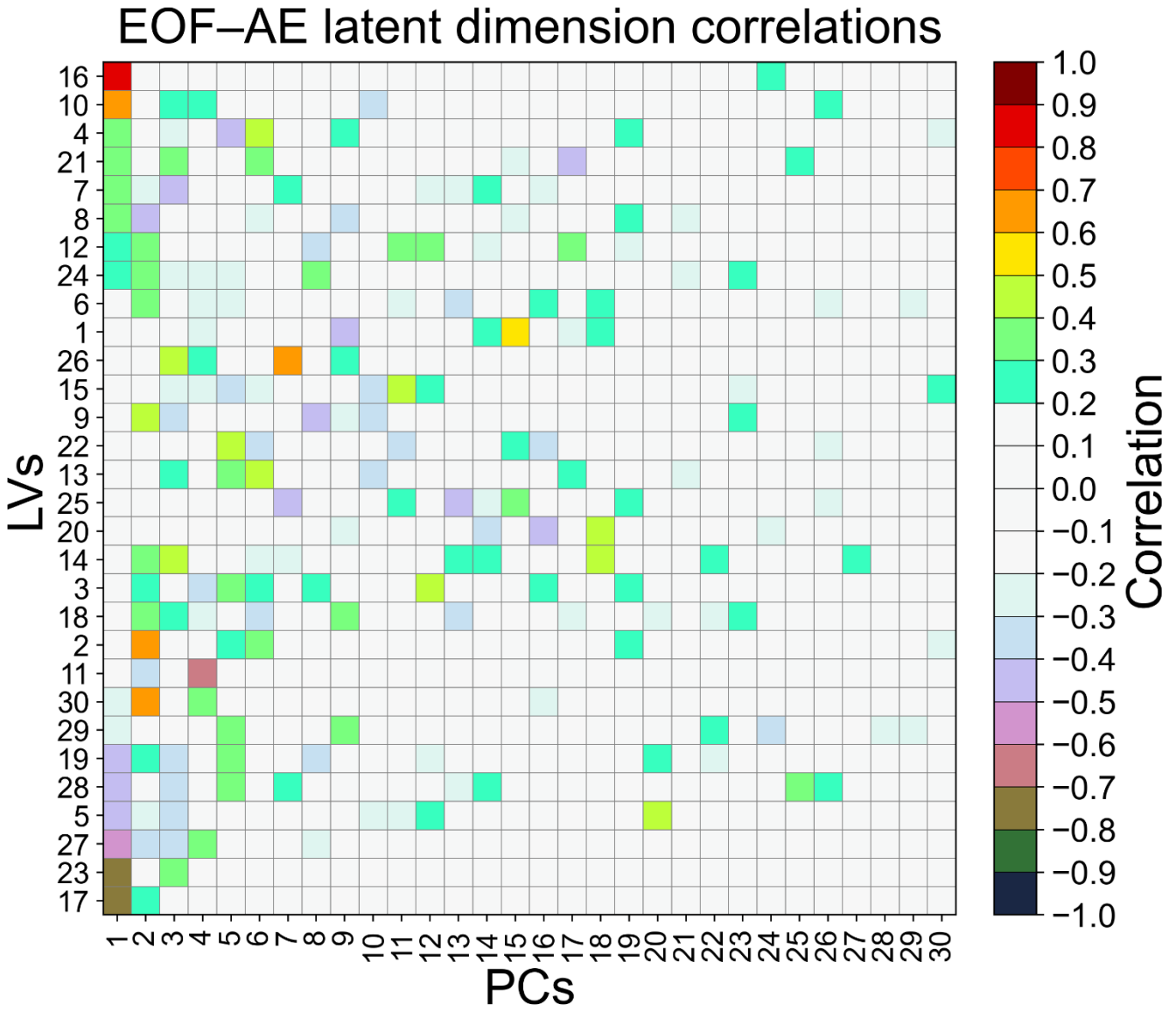


Figure A5: Correlation matrix between AE latent vectors and EOF principal components using a 16-dimensional latent space. AE latent vectors are ordered by decreasing correlation with EOF1.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

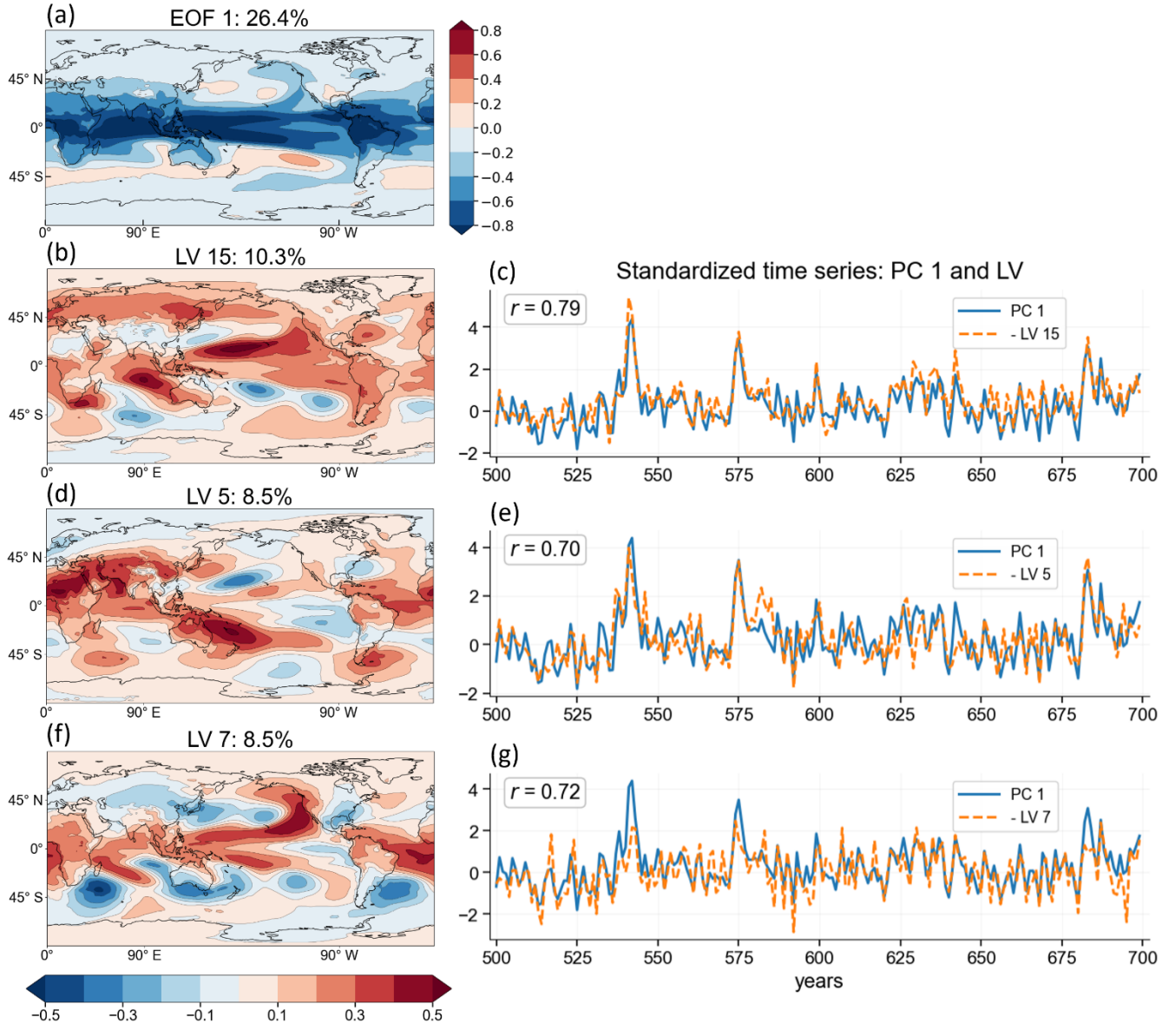


Figure A6: Variance structures associated with EOF1 within the 16-dimensional latent space. Panel (a) shows local variance explained by EOF1. Panels (b, d, f) display variance explained by selected AE latent vectors based on correlation ≥ 0.7 with EOF1. Panels (c, e, g) show standardized time series over the first 200 years. Correlation coefficients are indicated.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

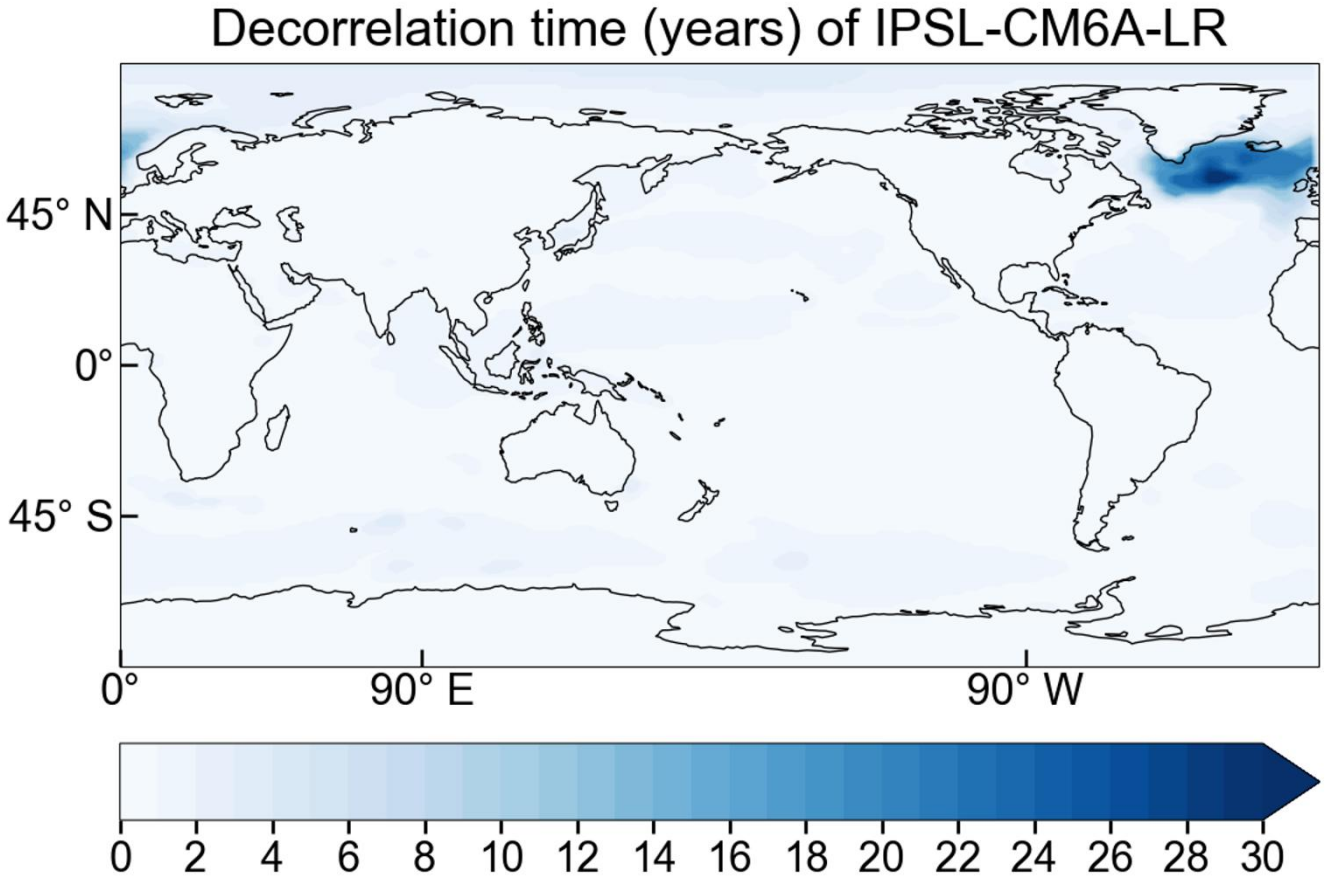
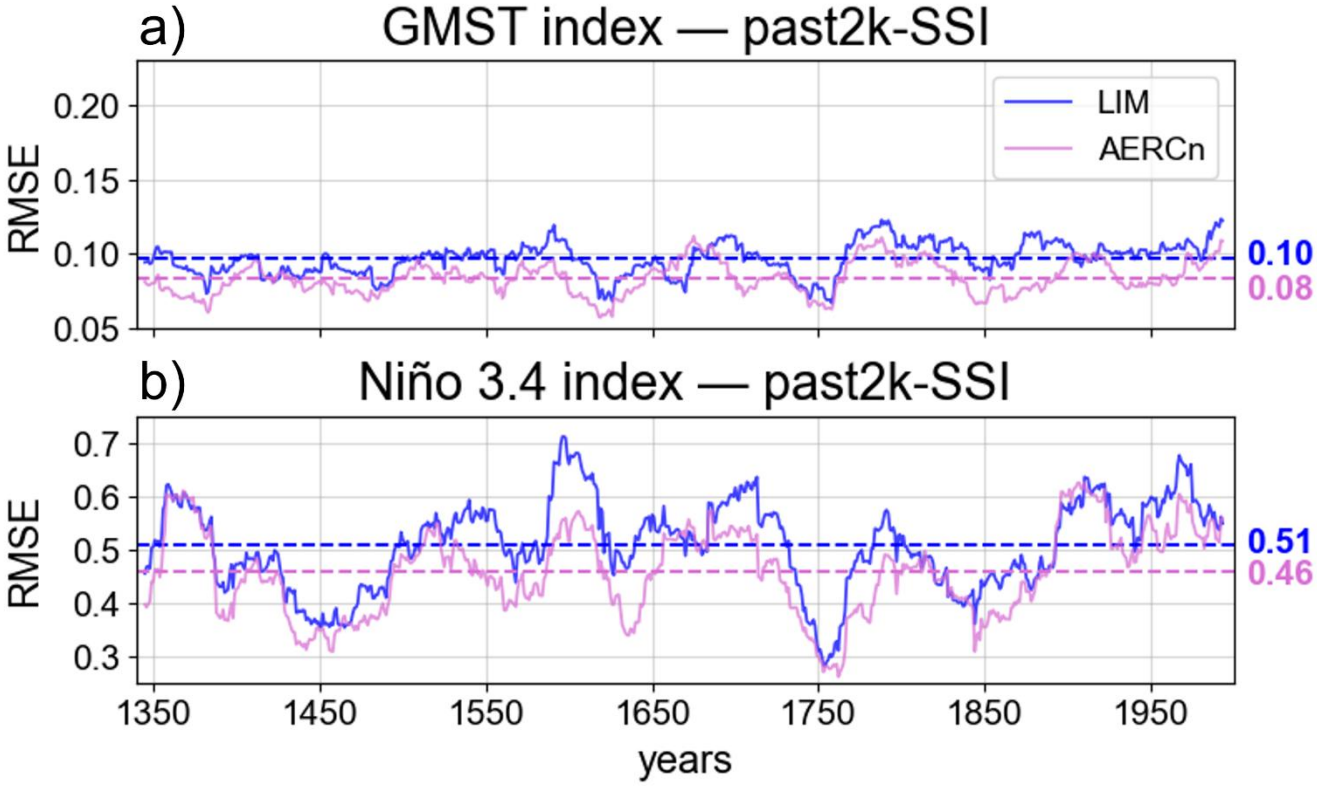


Figure A7: Decorrelation time of surface air temperature anomalies for IPSL-CM6A-LR, computed as the pointwise e-folding time of the autocorrelation function. Values are shown in years.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026



790 **Figure A8: Temporal evolution of prediction error. Thirty-year rolling RMSE of one-year lead forecasts for LIM (blue) and AERCn (pink) is shown for (a) GMST and (b) Niño3.4 in the past2k-SSI experiment.**

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

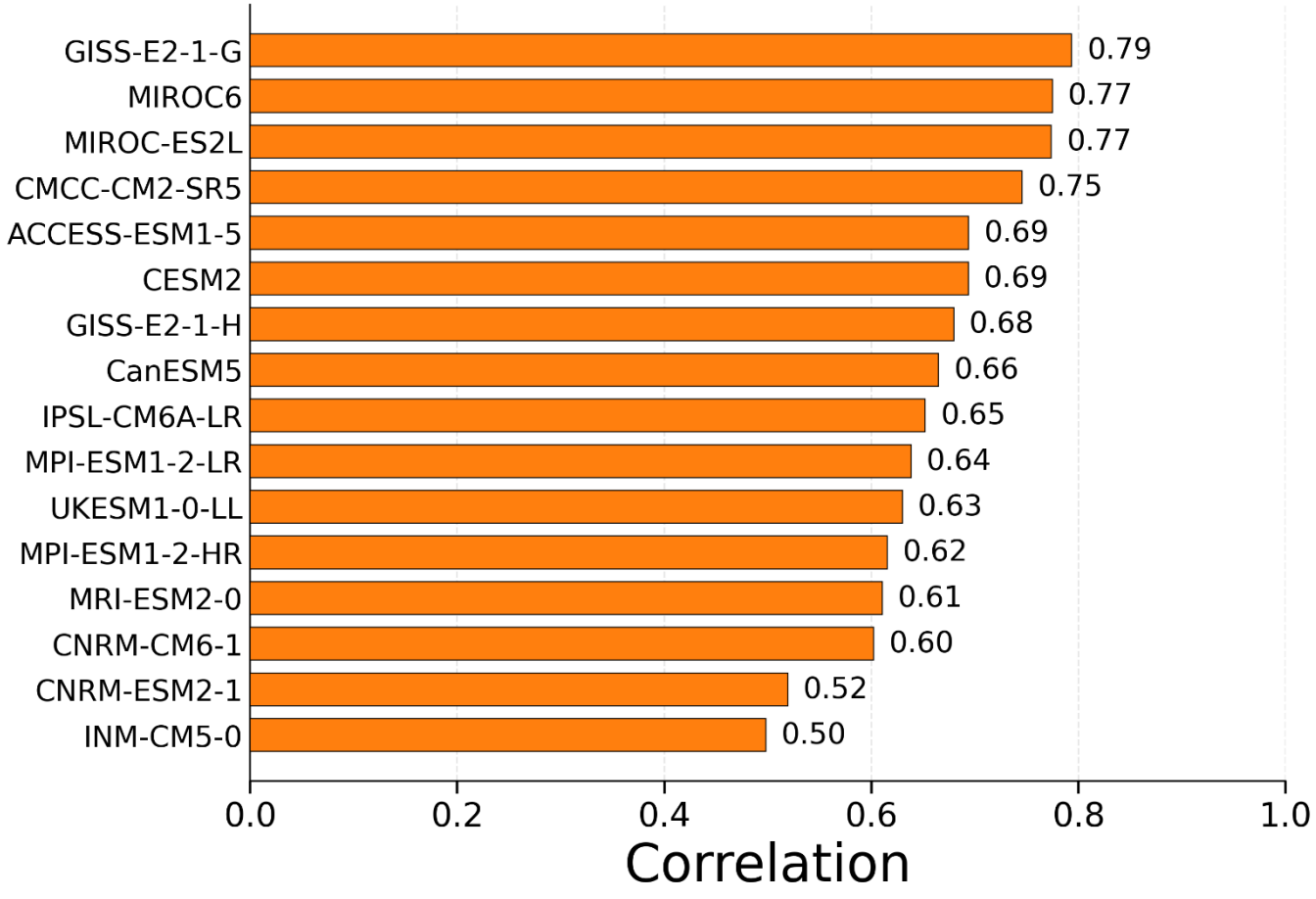


Figure A9: Spatially averaged tropical Pacific correlation at one-year lead time obtained using the RC emulator across 16 CMIP6 models. Models are ordered by decreasing skill.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

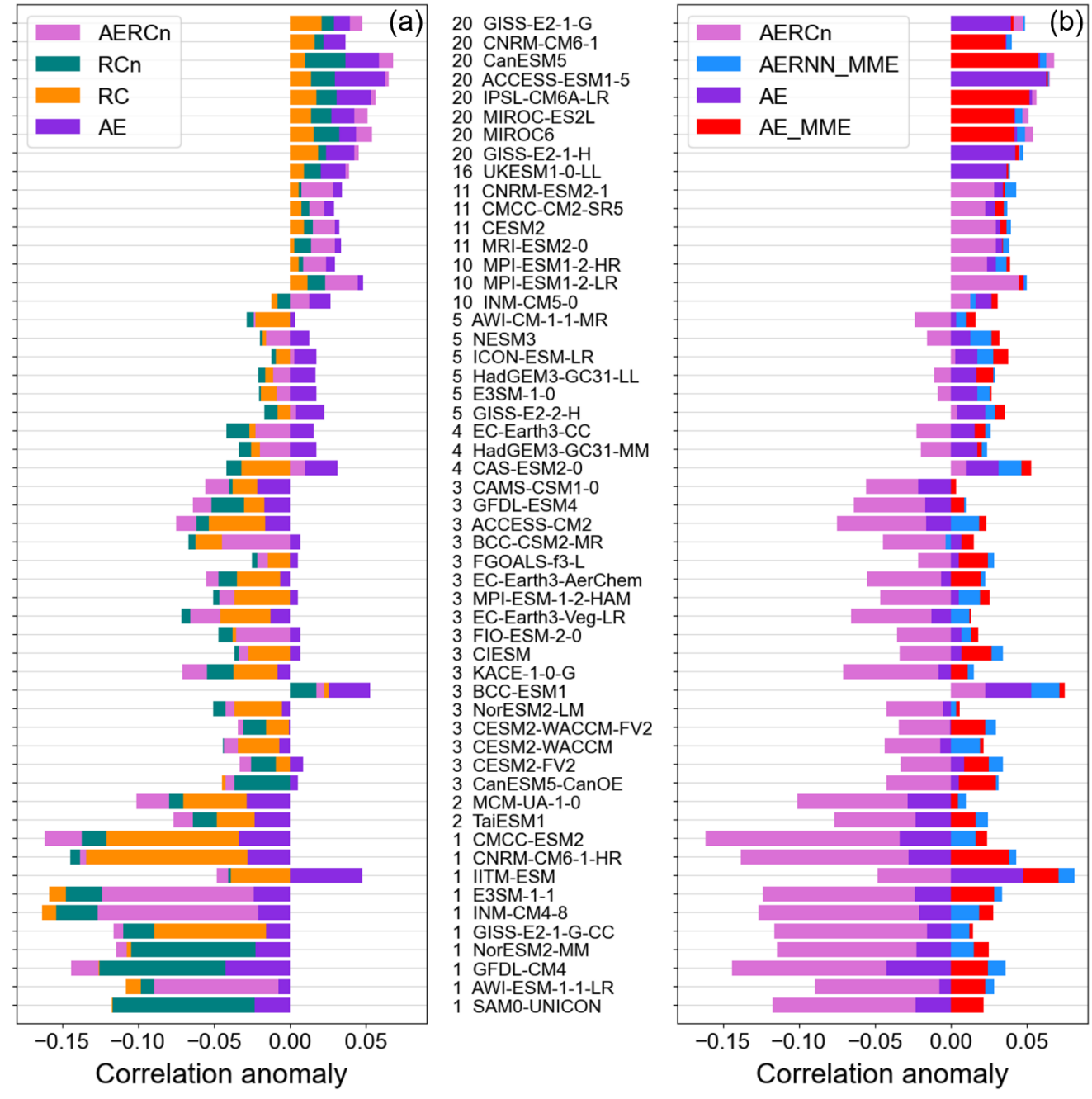


Figure A10: Emulator skill gain relative to LIM across 52 CMIP6 models. Panel (a) shows anomalies in spatially averaged correlation for AERCn, RCn, RC, and AE relative to LIM. Panel (b) shows skill differences for AE, AERCn, and their multimodel pretrained variants. Models are ordered according to training-set size, defined as the number of available historical ensemble members, which is indicated in the central panel.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

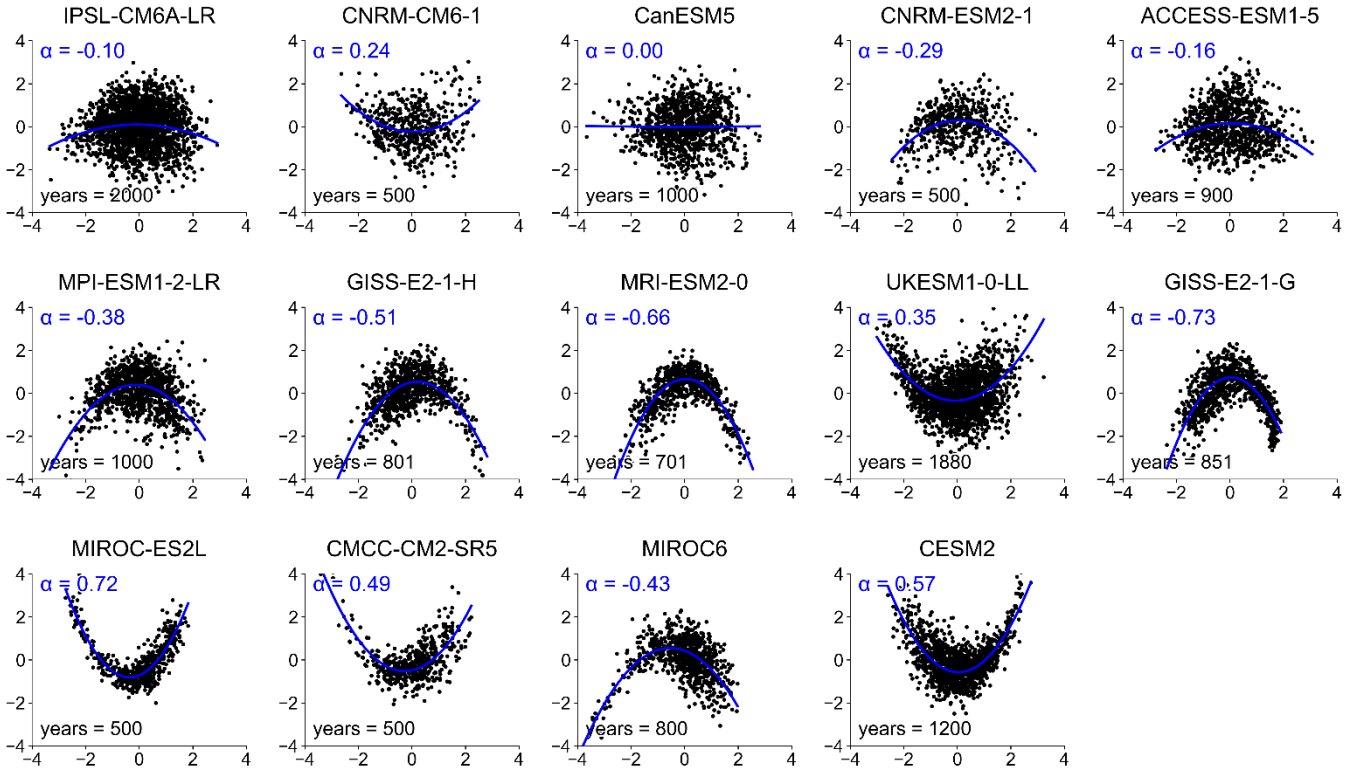


Figure A11: Phase-space representation of ENSO variability. Scatter plots of PC1 versus PC2 of DJF SST anomalies from CMIP6 piControl simulations are shown. The fitted quadratic curve provides the nonlinearity coefficient α.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

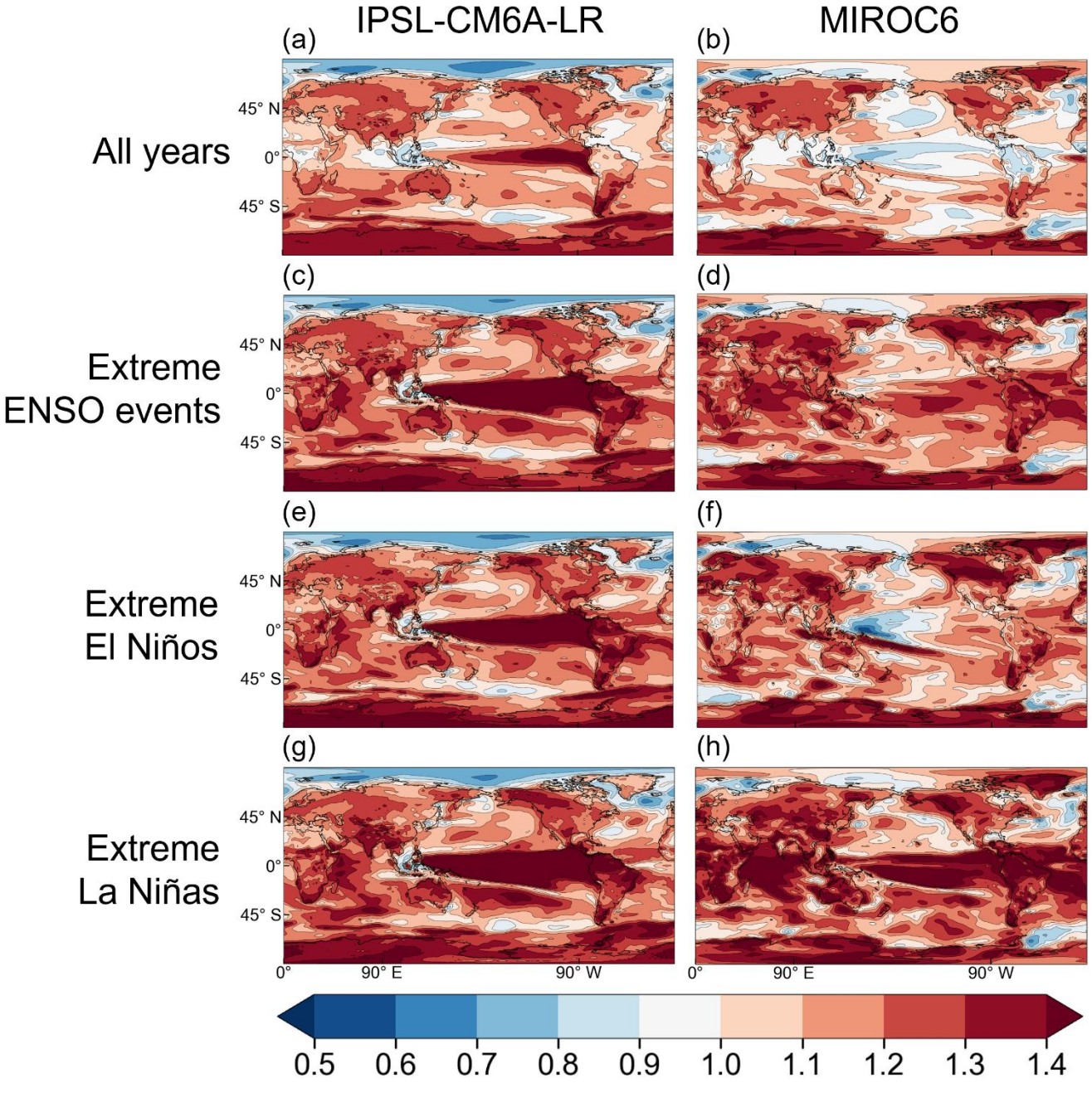


Figure A12: RMSE of one-year lead persistence forecasts with target piControl simulations during all years (a, b) extreme ENSO events (c,d), extreme El Niños (e,f) and extreme La Niñas (g,h). Results are shown for IPSL-CM6A-LR (right column) and MIROC6 (right column).

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Code availability

The code used to reproduce the results presented in this study is archived on Zenodo and publicly available at https://doi.org/10.5281/zenodo.18926500 (Gaudin, A. and Khodri, M., 2026) under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

Data availability

Data supporting the findings of this study are publicly available at https://doi.org/10.5281/zenodo.18926500 (Gaudin, A. and Khodri, M., 2026) under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.
All climate model simulations used in this study are from the CMIP6 archive and are publicly available through the ESGF data portals https://esgf-node.llnl.gov/projects/cmip6/. The list of models used is provided in Table 1. The IPSL-CM6A-LR past2k ensemble members are available via the IPSL ESGF node https://esgf-node.ipsl.upmc.fr/projects/cmip6-ipsl/.

Author contribution

MK designed the study and method and AG carried them out. AG developed the emulators and ran the experiments workflow. MK developed and ran the IPSL-past2k ensemble simulations. AG performed the analysis, made the figures, and wrote the manuscript. M.K. helped with the conceptualization of the manuscript, supervised the work and secured financial support for fellowships. Both authors contributed with ideas, discussed the results, and reviewed the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

Authors acknowledge support from the IPSL Climate Graduate School EUR and from the HPC resources of TGCC under grant no. A0150113826, A0170113826 and A0190113826 provided by GENCI (Grand Equipment National de Calcul Intensif). This study benefited from the ESPRI (Ensemble de Services Pour la Recherche l'IPSL) computing and data centre (https://mesocentre.ipsl.fr) which is supported by CNRS, Sorbonne Université, École Polytechnique and CNES.

Financial Support

This work received government funding managed by the French National Research Agency under France 2030, reference ANR-25-EXTR-0001.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

References

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., and Chung, E.-S.: Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics, Hydrology and Earth System Sciences, 23, 4803–4824, https://doi.org/10.5194/hess-23-4803-2019, 2019.

Alexander, M. A., Matrosova, L., Penland, C., Scott, J. D., and Chang, P.: Forecasting Pacific SSTs: Linear Inverse Model Predictions of the PDO, Journal of Climate, 21, 385–402, https://doi.org/10.1175/2007JCLI1849.1, 2008.

An, S.-I. and Jin, F.-F.: Nonlinearity and Asymmetry of ENSO, Journal of Climate, 17, 2399–2412, https://doi.org/10.1175/1520-0442(2004)017<2399:NAAOE>2.0.CO;2, 2004.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M.: PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies, Geoscientific Model Development, 8, 2465–2513, https://doi.org/10.5194/gmd-8-2465-2015, 2015.

Bank, D., Koenigstein, N., and Giryes, R.: Autoencoders, Mach. learn. data Sci. Handb., Data Mining. Knwl. Discov. Handb. Cham: Springer, 353-374, http://arxiv.org/abs/2003.05991, 3 April 2021.

Behrens, G., Beucler, T., Gentine, P., Iglesias-Suarez, F., Pritchard, M., and Eyring, V.: Non-Linear Dimensionality Reduction With a Variational Encoder Decoder to Understand Convective Processes in Climate Models, Journal of Advances in Modeling Earth Systems, 14, e2022MS003130, https://doi.org/10.1029/2022MS003130, 2022.

Bellucci, A., Haarsma, R., Bellouin, N., Booth, B., Cagnazzo, C., van den Hurk, B., Keenlyside, N., Koenigk, T., Massonnet, F., Materia, S., and Weiss, M.: Advancements in decadal climate predictability: The role of nonoceanic drivers, Reviews of Geophysics, 53, 165–202, https://doi.org/10.1002/2014RG000473, 2015.

Bengio, Y., Simard, P., and Frasconi, P.: Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks, 5, 157–166, https://doi.org/10.1109/72.279181, 1994.

Berliner, L. M. and Wikle, C. K.: Approximate importance sampling Monte Carlo for data assimilation, Physica D: Nonlinear Phenomena, 230, 37–49, https://doi.org/10.1016/j.physd.2006.07.031, 2007.

Berrocal, V. J., Raftery, A. E., and Gneiting, T.: Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts, Monthly Weather Review, 135, 1386–1402, https://doi.org/10.1175/MWR3341.1, 2007.

Boer, G. J.: A study of atmosphere-ocean predictability on long time scales, Climate Dynamics, 16, 469–477, https://doi.org/10.1007/s003820050340, 2000.

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,

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

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.

Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, https://doi.org/10.48550/arXiv.1406.1078, 3 September 2014.

Cobb, K. M., Charles, C. D., Cheng, H., and Edwards, R. L.: El Niño/Southern Oscillation and tropical Pacific climate during the last millennium, Nature, 424, 271–276, https://doi.org/10.1038/nature01779, 2003.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Farrell, B. F. and Ioannou, P. J.: Generalized Stability Theory. Part I: Autonomous Operators, Journal of the Atmospheric Sciences, 53, 2025–2040, https://doi.org/10.1175/1520-0469(1996)053<2025:GSTPIA>2.0.CO;2, 1996.

Gaudin, A. and Khodri, M.: New classes of climate model emulators to improve paleoclimate reconstructions, https://doi.org/10.5281/zenodo.18926500, 2026.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P.: Advanced Spectral Methods for Climatic Time Series, Reviews of Geophysics, 40, 3-1-3–41, https://doi.org/10.1029/2000RG000092, 2002.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gondara, L.: Medical Image Denoising Using Convolutional Denoising Autoencoders, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 241–246, https://doi.org/10.1109/ICDMW.2016.0041, 2016.

Goosse, H., Brovkin, V., Fichefet, T., Haarsma, R., Huybrechts, P., Jongma, J., Mouchet, A., Selten, F., Barriat, P.-Y., Campin, J.-M., Deleersnijder, E., Driesschaert, E., Goelzer, H., Janssens, I., Loutre, M.-F., Morales Maqueda, M. A., Opsteegh, T., Mathieu, P.-P., Munhoven, G., Pettersson, E. J., Renssen, H., Roche, D. M., Schaeffer, M., Tartinville, B., Timmermann, A., and Weber, S. L.: Description of the Earth system model of intermediate complexity LOVECLIM version 1.2, Geoscientific Model Development, 3, 603–633, https://doi.org/10.5194/gmd-3-603-2010, 2010.

Guardamagna, F., Wieners, C., and Dijkstra, H. A.: Explaining the high skill of reservoir computing methods in El Niño prediction, Nonlinear Processes in Geophysics, 32, 201–224, https://doi.org/10.5194/npg-32-201-2025, 2025.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B.: Empirical orthogonal functions and related techniques in atmospheric science: A review, International Journal of Climatology, 27, 1119–1152, https://doi.org/10.1002/joc.1499, 2007.

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.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Santoso, A., Mcphaden, M. J., and Cai, W.: The Defining Characteristics of ENSO Extremes and the Strong 2015/2016 El Niño, Reviews of Geophysics, 55, 1079–1129, https://doi.org/10.1002/2017RG000560, 2017.

Hourdin, F., Rio, C., Grandpeix, J.-Y., Madeleine, J.-B., Cheruy, F., Rochetin, N., Jam, A., Musat, I., Idelkadi, A., Fairhead, L., Foujols, M.-A., Mellul, L., Traore, A.-K., Dufresne, J.-L., Boucher, O., Lefebvre, M.-P., Millour, E., Vignon, E., Jouhaud, J., Diallo, F. B., Lott, F., Gastineau, G., Caubel, A., Meurdesoif, Y., and Ghattas, J.: LMDZ6A: The Atmospheric Component of the IPSL Climate Model With Improved and Better Tuned Physics, Journal of Advances in Modeling Earth Systems, 12, e2019MS001892, https://doi.org/10.1029/2019MS001892, 2020.

Houtekamer & Zhang: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489-4532, https://doi.org/10.1175/MWR-D-15-0440.1, 2016

Jaeger, The" echo state" approach to analysing and training recurrent neural networks-with an erratum note', German National Research Center for Information Technology, 2001.

Jaeger, H. and Haas, H.: Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication, Science, 304, 78–80, https://doi.org/10.1126/science.1091277, 2004.

Jebri, B. and Khodri, M.: Large Ensemble Particle Filter for Spatial Climate Reconstructions Using a Linear Inverse Model, Journal of Advances in Modeling Earth Systems, 15, e2022MS003094, https://doi.org/10.1029/2022MS003094, 2023.

Jin, F.-F., An, S.-I., Timmermann, A., and Zhao, J.: Strong El Niño events and nonlinear dynamical heating, Geophysical Research Letters, 30, 20-1-20–1, https://doi.org/10.1029/2002GL016356, 2003.

Jungclaus, J. H., Bard, E., Baroni, M., Braconnot, P., Cao, J., Chini, L. P., Egorova, T., Evans, M., González-Rouco, J. F., Goosse, H., Hurtt, G. C., Joos, F., Kaplan, J. O., Khodri, M., Klein Goldewijk, K., Krivova, N., LeGrande, A. N., Lorenz, S. J., Luterbacher, J., Man, W., Maycock, A. C., Meinshausen, M., Moberg, A., Muscheler, R., Nehrbass-Ahles, C., Otto-Bliesner, B. I., Phipps, S. J., Pongratz, J., Rozanov, E., Schmidt, G. A., Schmidt, H., Schmutz, W., Schurer, A., Shapiro, A. I., Sigl, M., Smerdon, J. E., Solanki, S. K., Timmreck, C., Toohey, M., Usoskin, I. G., Wagner, S., Wu, C.-J., Yeo, K. L., Zanchettin, D., Zhang, Q., and Zorita, E.: The PMIP4 contribution to CMIP6 – Part 3: The last millennium, scientific objective, and experimental design for the PMIP4 past1000 simulations, Geoscientific Model Development, 10, 4005–4033, https://doi.org/10.5194/gmd-10-4005-2017, 2017.

Karamperidou, C., Jin, F.-F., and Conroy, J. L.: The importance of ENSO nonlinearities in tropical pacific response to external forcing, Clim Dyn, 49, 2695–2704, https://doi.org/10.1007/s00382-016-3475-y, 2017.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Leeuwen, P. J. van: Particle Filtering in Geophysical Systems, Mon. Weather Rev. 137, 4089–4114, https://doi.org/10.1175/2009MWR2835.1, 2009.

Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/10.1002/qj.3551, 2019.

Madec, G. and NEMO Team: NEMO ocean engine – version 3.6 stable, Pôle de modélisation de l'IPSL, 2016.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proceedings of the National Academy of Sciences, 105, 13252–13257, https://doi.org/10.1073/pnas.0805721105, 2008.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., and Zhou, B. (Eds.): Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896, 2021.

Maulik, R., Egele, R., Lusch, B., and Balaprakash, P.: Recurrent Neural Network Architecture Search for Geophysical Emulation, SC20: International conference for high performance computing, networking, storage and analysis (2020), pp. 1-14, http://arxiv.org/abs/2004.10928, 13 August 2020a.

Maulik, R., Lusch, B., and Balaprakash, P.: Reduced-order modeling of advection-dominated systems with recurrent neural networks and convolutional autoencoders, Physics of Fluids, 33 (3) (2021), Article 037106, http://arxiv.org/abs/2002.00470, 9 December 2020b.

McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an integrating concept in earth science, Science, 314, 1740–1745, https://doi.org/10.1126/science.1132588, 2006.

Nadiga, B. T.: Reservoir Computing as a Tool for Climate Predictability Studies, Journal of Advances in Modeling Earth Systems, 13, e2020MS002290, https://doi.org/10.1029/2020MS002290, 2021.

Namias, J. and Born, R. M.: Temporal coherence in North Pacific sea-surface temperature patterns, Journal of Geophysical Research (1896-1977), 75, 5952–5955, https://doi.org/10.1029/JC075i030p05952, 1970.

Neukom, R., Gergis, J., Karoly, D. J., Wanner, H., Curran, M., Elbert, J., González-Rouco, F., Linsley, B. K., Moy, A. D., Mundo, I., Raible, C. C., Steig, E. J., van Ommen, T., Vance, T., Villalba, R., Zinke, J., and Frank, D.: Inter-hemispheric temperature variability over the past millennium, Nature Clim Change, 4, 362–367, https://doi.org/10.1038/nclimate2174, 2014.

Newman, M.: Interannual to Decadal Predictability of Tropical and North Pacific Sea Surface Temperatures, Journal of Climate, 20, 2333–2356, https://doi.org/10.1175/JCLI4165.1, 2007.

Newman, M.: An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies, Journal of Climate, 26, 5260–5269, https://doi.org/10.1175/JCLI-D-12-00590.1, 2013.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Newman, M., Compo, G. P., and Alexander, M. A.: ENSO-Forced Variability of the Pacific Decadal Oscillation, Journal of Climate, 16, 3853–3857, https://doi.org/10.1175/1520-0442(2003)016<3853:EVOTPD>2.0.CO;2, 2003.

PAGES 2K Consortium, Barboza, L. A., Erb, M. P., Shi, F., Emile-Geay, J., Evans, M. N., Franke, J., Kaufman, D. S., Lücke, L., Rehfeld, K., Schurer, A., Zhu, F., Brönnimann, S., Hakim, G. J., Henley, B. J., Ljungqvist, F. C., McKay, N., Valler, V., von Gunten, L., and PAGES 2k Consortium: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nat. Geosci., 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

Parsons, L. A., Amrhein, D. E., Sanchez, S. C., Tardif, R., Brennan, M. K., and Hakim, G. J.: Do Multi-Model Ensembles Improve Reconstruction Skill in Paleoclimate Data Assimilation?, Earth and Space Science, 8, e2020EA001467, https://doi.org/10.1029/2020EA001467, 2021.

Pascanu, R., Mikolov, T., and Bengio, Y.: On the difficulty of training Recurrent Neural Networks, International conference on machine learning, 1310–1318, https://doi.org/10.48550/arXiv.1211.5063, 16 February 2013.

Penland, C. and Matrosova, L.: A Balance Condition for Stochastic Numerical Models with Application to the El Niño-Southern Oscillation, Journal of Climate, 7, 1352–1372, https://doi.org/10.1175/1520-0442(1994)007<1352:ABCFSN>2.0.CO;2, 1994.

Penland, C. and Matrosova, L.: Prediction of Tropical Atlantic Sea Surface Temperatures Using Linear Inverse Modeling, 1998.

Penland, C. and Sardeshmukh, P. D.: The Optimal Growth of Tropical Sea Surface Temperature Anomalies, Journal of Climate, 8, 1999–2024, https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2, 1995.

Rial et al.: Nonlinearities, Feedbacks and Critical Thresholds within the Earth's Climate System, Climatic Change, 65, 11–38, https://doi.org/10.1023/B:CLIM.0000037493.89489.3f, 2004.

Richter, I. and Tokinaga, H.: An overview of the performance of CMIP6 models in the tropical Atlantic: mean state, variability, and remote impacts, Clim Dyn, 55, 2579–2601, https://doi.org/10.1007/s00382-020-05409-w, 2020.

Ross, I.: Nonlinear Dimensionality Reduction Methods in Climate Data Analysis, PhD thesis, University of Bristol, United Kingdom, http://arxiv.org/abs/0901.0537, 2 January 2009.

Rousset, C., Vancoppenolle, M., Madec, G., Fichefet, T., Flavoni, S., Barthélemy, A., Benshila, R., Chanut, J., Levy, C., Masson, S., and Vivier, F.: The Louvain-La-Neuve sea ice model LIM3.6: global and regional capabilities, Geoscientific Model Development, 8, 2991–3005, https://doi.org/10.5194/gmd-8-2991-2015, 2015.

Rumelhart, D. E., Hintont, G. E., and Williams, R. J.: Learning representations by back-propagating errors, Nature 323, 533–536 (1986), https://doi-org.insu.bib.cnrs.fr/10.1038/323533a0, 1986.

Saenz, J. A., Lubbers, N., and Urban, N. M.: Dimensionality-Reduction of Climate Data using Deep Autoencoders, 6th international workshop on climate informatics, http://arxiv.org/abs/1809.00027, 27 August 2018.

Santoso, A., Mcphaden, M. J., and Cai, W.: The Defining Characteristics of ENSO Extremes and the Strong 2015/2016 El Niño, Reviews of Geophysics, 55, 1079–1129, https://doi.org/10.1002/2017RG000560, 2017.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y-Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, Journal of Advances in Modeling Earth Systems, 11, 4182–4227, https://doi.org/10.1029/2019MS001791, 2019.

Smerdon, J. E. and Pollack, H. N.: Reconstructing Earth's surface temperature over the past 2000 years: the science behind the headlines, WIREs Climate Change, 7, 746–771, https://doi.org/10.1002/wcc.418, 2016.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to High-Dimensional Particle Filtering, https://doi.org/10.1175/2008MWR2529.1, 2008.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

Sun, Y., Wang, F., and Sun, D.-Z.: Weak ENSO asymmetry due to weak nonlinear air-sea interaction in CMIP5 climate models, Advances in Atmospheric Sciences, 33, 352–364, https://doi.org/10.1007/s00376-015-5018-6, 2016.

Timmermann, A., An, S.-I., Kug, J.-S., Jin, F.-F., Cai, W., Capotondi, A., Cobb, K. M., Lengaigne, M., McPhaden, M. J., Stuecker, M. F., Stein, K., Wittenberg, A. T., Yun, K.-S., Bayr, T., Chen, H.-C., Chikamoto, Y., Dewitte, B., Dommenget, D., Grothe, P., Guilyardi, E., Ham, Y.-G., Hayashi, M., Ineson, S., Kang, D., Kim, S., Kim, W., Lee, J.-Y., Li, T., Luo, J.-J., McGregor, S., Planton, Y., Power, S., Rashid, H., Ren, H.-L., Santoso, A., Takahashi, K., Todd, A., Wang, G., Wang, G., Xie, R., Yang, W.-H., Yeh, S.-W., Yoon, J., Zeller, E., and Zhang, X.: El Niño–Southern Oscillation complexity, Nature, 559, 535–545, https://doi.org/10.1038/s41586-018-0252-6, 2018.

Ting, M., Kushnir, Y., Seager, R., and Li, C.: Forced and Internal Twentieth-Century SST Trends in the North Atlantic, Journal of Climate, 22, 1469–1481, https://doi.org/10.1175/2008JCLI2561.1, 2009.

Vahdat, A. and Kautz, J.: NVAE: A Deep Hierarchical Variational Autoencoder, Adv Neural Inf Process Syst, 33, 19667–19679, http://arxiv.org/abs/2007.03898, 8 January 2021.

Vimont, D. J.: Analysis of the Atlantic Meridional Mode Using Linear Inverse Modeling: Seasonality and Regional Influences, Journal of Climate, 25, 1194–1212, https://doi.org/10.1175/JCLI-D-11-00012.1, 2012.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., and Roesch, C.: ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections, Journal of Advances in Modeling Earth Systems, 14, e2021MS002954, https://doi.org/10.1029/2021MS002954, 2022.

Werbos, P. J.: Backpropagation through time: what it does and how to do it, Proc. IEEE, 78, 1550–1560, https://doi.org/10.1109/5.58337, 1990.

Zhang, Q., Wang, H., Dong, J., Zhong, G., and Sun, X.: Prediction of Sea Surface Temperature using Long Short-Term Memory, http://arxiv.org/abs/1705.06861, 19 May 2017.

https://doi.org/10.5194/egusphere-2026-1337
Preprint. Discussion started: 25 March 2026

Zhang, Q., Liu, B., Li, S., and Zhou, T.: Understanding Models' Global Sea Surface Temperature Bias in Mean State: From CMIP5 to CMIP6, Geophysical Research Letters, 50, e2022GL100888, https://doi.org/10.1029/2022GL100888, 2023.

Zhao, Y. and Sun, D.-Z.: ENSO Asymmetry in CMIP6 Models, Journal of Climate, 35, 5555–5572, https://doi.org/10.1175/JCLI-D-21-0835.1, 2022.

Zhu, Y., Zhang, R.-H., and Sun, J.: North Pacific Upper-Ocean Cold Temperature Biases in CMIP6 Simulations and the Role of Regional Vertical Mixing, Journal of Climate, 33, 7523–7538, https://doi.org/10.1175/JCLI-D-19-0654.1, 2020.