

Learning the coupled dynamics of global climate modes

In the format provided by the
authors and unedited

Contents

Supplementary Notes	2
1 Experimental settings	2
1.1 Baselines	2
2 Detailed ENSO forecast performance and robustness	2
2.1 Error metrics and multi-dataset validation	2
2.2 Performance across the spring predictability barrier	3
2.3 Evaluation of event-centric forecast skill	3
3 Comprehensive performance on global climate modes	4
3.1 Seasonal predictability and baseline comparison	4
3.2 Robustness of forecast skill across datasets	4
3.3 Field-significance testing	4
3.4 Comparison with field-to-field forecasting models	5
4 Analysis of learned correlations between climate modes	6
4.1 Lagged-correlation evaluation	6
4.2 Precursors and couplings for non-ENSO extreme events	6
5 Ablation studies and sensitivity analysis	7
5.1 Architectural ablation study	7
5.2 Hyperparameter sensitivity analysis	7
5.3 Quantitative assessment of training data requirements	8
5.4 Grid-resolution sensitivity	8
6 Analysis of pre-training and fine-tuning	9
Supplementary Tables	10
Supplementary Figures	13
References	50

Supplementary Notes

1 Experimental settings

1.1 Baselines

To ensure a rigorous and fair performance comparison, we implemented four representative baseline models: CNN¹, ResoNet², XRO³, and the recent physics-guided deep echo state network (DESN)⁴. Because DESN was originally proposed for long-range ENSO forecasting, we include it as a strong recent physics-guided baseline for ENSO prediction and, where applicable, as an additional nonlinear comparison in the multi-mode evaluation. Each model was configured following the architectural specifications in their respective original publications. Supplementary Table 1 summarizes the architectural and training settings of the baseline models. While architectures were kept consistent with original publications, we optimized the training phase for all DL-based baselines. We employed the AdamW optimizer and performed a grid search across learning rates $\eta \in \{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$ (Supplementary Fig. 1). Models were trained for a maximum of 200 epochs with an early-stopping patience of 10 epochs based on validation loss to prevent overfitting.

In the ENSO prediction literature, model evaluation generally follows two paradigms. In an in-distribution evaluation, a model is trained and tested on data drawn from the same observational reanalysis product. For instance, XRO is fitted to ORAS5 reanalysis over 1979–2022 and verified via hindcasts over the same period³, while DESN reports an in-sample experiment in which the model is trained and validated on the overlapping ORAS5 period of 1979–2023 to assess its fitting capacity⁴. Under this protocol, the training and test data share the same underlying data-generating process, primarily testing a model’s ability to learn and reproduce the observed dynamics. In contrast, UniCM and its baselines (except DESN) in this study are evaluated under an out-of-distribution protocol: models are pre-trained exclusively on CMIP6 climate simulations and then tested on observational reanalysis data (ORAS5, GODAS, ERA5, and SODA v2.2.4) (Supplementary Table 3). Because the training and test data originate from fundamentally different sources, numerical simulations versus observational products, this protocol constitutes a more stringent test of generalization and is more practical. It is also more practical, as climate simulations provide a far greater diversity of climatological states than the short observational record, better supporting the learning of rare climate phenomena.

Specifically, for DESN, we initially attempted to follow the same out-of-distribution protocol. However, pre-training on CMIP6 simulations led to substantially degraded performance, likely because the reservoir dynamics are sensitive to distributional differences between climate model output and observational reanalysis. We therefore trained DESN directly on ORAS5 reanalysis data from 1958 to 1980 and evaluated it on ORAS5 data after 1980, which constitutes an out-of-sample but in-distribution setting. This yielded a skillful prediction lead time of 16 months, consistent with the results reported in the original study⁴, where DESN was trained on ORAS5 from 1958 to 1999 and achieved ACC > 0.5 up to 16 months in out-of-sample evaluation (2002–2023) and approximately 20 months in in-sample evaluation (1979–2023). This gap further illustrates the performance difference between evaluation paradigms, even within the same data source.

2 Detailed ENSO forecast performance and robustness

In the main text, we demonstrated UniCM’s performance in forecasting the El Niño-Southern Oscillation (ENSO), primarily using the Anomaly Correlation Coefficient (ACC) metric (see main text, Fig. 2). Here, we provide additional results that substantiate these findings, including error-based metrics and comprehensive robustness checks across multiple reanalysis datasets.

2.1 Error metrics and multi-dataset validation

To complement the ACC results presented in the main text, we evaluated the Root Mean Square Error (RMSE) of UniCM against baseline models (ResoNet, XRO, and CNN). A systematic comparison of both ACC and RMSE, validated against the ERA5, GODAS, SODA v2.2.4, and ORAS5 datasets, is presented in Supplementary Figures 2 through 8.

Across all validation datasets, the results confirm the superior performance and robustness of UniCM. For instance, when validated against ERA5 (Supplementary Figs. 2-3), UniCM not only maintains a higher ACC for the entire 24-month forecast horizon, extending the skillful forecast lead time (ACC > 0.5) to 19 months compared to 8-14 months for baselines, but also consistently achieves the lowest RMSE. At a 12-month lead time, UniCM reduces the RMSE by approximately 16.7% relative to the next best model (CNN) and by 23.1% compared to XRO, indicating a more accurate prediction of event magnitudes. This pattern of superior performance is consistent across other datasets. On GODAS (Supplementary Figs. 4-5), UniCM extends the skillful forecast horizon by 11 months over XRO and maintains the lowest RMSE throughout the 24-month period. Similarly, on SODA v2.2.4 (Supplementary Figs. 6-7), it shows a 10-month extension in skillful ACC and a clear advantage in RMSE. This comprehensive, multi-metric, and multi-dataset validation underscores that the advantage of UniCM is not specific to a particular metric or dataset but reflects a genuine improvement in predictive capability.

2.2 Performance across the spring predictability barrier

The main text highlights UniCM’s unique ability to overcome the ENSO “spring predictability barrier” (SPB), a period in the boreal spring when forecast skill typically declines. Supplementary Figures 9, 10, 11, and 12 provide detailed visualizations of the seasonal forecast skill for UniCM and all baseline models across four different datasets.

These heatmaps, which plot ACC skill against target season and forecast lead time, offer clear quantitative evidence for UniCM’s advantage. In Supplementary Figure 9a, for example, UniCM maintains skillful predictions ($\text{ACC} > 0.5$, indicated by the black contour) for lead times up to 14 months, even for forecasts targeting the challenging March–April–May (MAM) and April–May–June (AMJ) seasons. For forecasts targeting the MAM season, UniCM’s skill at a 12-month lead is approximately 0.6, well above the threshold. In stark contrast, the baseline models exhibit a rapid decay in skill. The XRO model (Supplementary Fig. 9b) shows a rapid skill drop, with its ACC falling below 0.3 for the same target season and lead time. The 0.5 ACC contour for all baseline models, including CNN (Supplementary Fig. 9c) and ResoNet (Supplementary Fig. 9d), collapses before a 9-month lead when targeting the spring season. This pattern is consistently observed across all other validation datasets (Supplementary Figs. 10–12), confirming that UniCM provides a more reliable foundation for year-round operational forecasting.

2.3 Evaluation of event-centric forecast skill

To complement correlation-based metrics in the main text, we conducted additional event-centric evaluations to assess the ability of UniCM to detect, characterize, and temporally track discrete climate extremes. These analyses focus on (i) categorical event detection skill, (ii) intensity accuracy and distributional behavior, and (iii) temporal evolution consistency.

Categorical event detection skill. Following established practices in ENSO prediction studies⁵, we converted the continuous Niño3.4 index into categorical event labels using a threshold-based definition:

$$\begin{aligned} \text{El Niño} &: \text{Index} \geq 0.5^\circ\text{C for at least five consecutive months,} \\ \text{La Niña} &: \text{Index} \leq -0.5^\circ\text{C for at least five consecutive months,} \\ \text{Neutral} &: \text{otherwise.} \end{aligned} \tag{1}$$

For each lead time (1–24 months), we computed: (1) Overall Accuracy, defined as the fraction of correctly classified states, and (2) The Critical Success Index (CSI), also known as the Threat Score, for El Niño and La Niña events:

$$\text{CSI} = \frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False Alarms}}. \tag{2}$$

Unlike Accuracy, CSI excludes correct Neutral predictions and explicitly penalizes both missed detections and false alarms, making it more suitable for evaluating rare extreme events.

Supplementary Fig. 14 compares UniCM with the XRO baseline across all lead times. While XRO maintains relatively high Accuracy, its CSI for both El Niño and La Niña drops significantly with increasing lead time compared with UniCM, indicating a conservative bias toward Neutral predictions. In contrast, UniCM consistently achieves higher CSI scores, demonstrating enhanced sensitivity to extreme events and a reduced trade-off between missed detections and false alarms.

Intensity accuracy and distributional behavior. To examine whether UniCM exhibits systematic bias or regression-to-the-mean effects, we analyzed the joint distribution of predicted and observed Niño3.4 intensities using density-based scatter plots (hexbin maps). Evaluations were conducted at representative lead times (6, 12, and 18 months). Each hexbin plot displays the density of prediction–observation pairs, with the diagonal line indicating perfect agreement. Reference lines at $\pm 0.5^\circ\text{C}$ and $\pm 1.5^\circ\text{C}$ mark commonly used thresholds for moderate and strong events.

Supplementary Fig. 15 shows the hexbin density scatter plots of predicted versus observed Niño3.4 SSTA at representative lead times (6, 12, and 18 months). Across lead times, the highest-density regions are concentrated near 0°C , indicating that UniCM produces stable and precise predictions for neutral/climatological conditions. Moreover, the dominant density ridge closely follows the $y = x$ identity line, confirming that UniCM maintains strong intensity agreement with observations even at a 12-month lead time. At the distribution tails (strong events; e.g., $|\text{Index}| > 1.5^\circ\text{C}$), UniCM exhibits a mild but systematic amplitude attenuation: the predicted values (y-axis) tend to be slightly smaller in magnitude than the corresponding observed values (x-axis) for both positive and negative extremes. While UniCM exhibits moderate “intensity shrinkage” for rare, high-magnitude cases, this behavior is expected in data-driven climate prediction due to limited extreme-event samples and variance-penalizing loss functions. Crucially, this shrinkage does not undermine the model’s ability to detect and characterize strong events, as UniCM reliably preserves the sign and relative strength of extreme anomalies.

Temporal Evolution Consistency. To assess whether the model captures the full life cycle of climate events, we visualized the predicted and observed Niño3.4 evolution using Hovmöller-like heatmaps. Each row corresponds to a forecast initialized at a given date, and columns represent lead times up to 24 months. We compare UniCM predictions with observations and the XRO baseline using a consistent color scale centered at zero.

As shown in Supplementary Fig. 16, UniCM exhibits a high degree of temporal phase coherence with observations across the full 24-month prediction horizon. The predicted heatmap closely reproduces the observed spatiotemporal patterns both vertically (across forecast initialization dates) and horizontally (across lead times), indicating that the model consistently tracks the temporal evolution of climate events. In contrast, the XRO baseline shows pronounced signal blurring and progressive phase lag beyond 6–12 months. At longer leads, XRO predictions frequently exhibit amplitude attenuation, characterized by a fading toward neutral conditions, and often fail to correctly represent the termination of events. This comparison demonstrates that UniCM preserves the temporal structure and dynamical continuity of the climate system more effectively than the baseline, enabling more reliable long-range tracking of event evolution.

3 Comprehensive performance on global climate modes

The main text (Fig. 3) established that UniCM substantially outperforms the XRO baseline across seven key climate modes. The supplementary figures below offer a more granular view of this performance, detailing the seasonal dependency of forecast skill and demonstrating the robustness of this advantage across different data sources.

3.1 Seasonal predictability and baseline comparison

A key claim in the main text is that the linear assumptions of the XRO model limit its predictive skill for modes governed by complex, state-dependent teleconnections. Supplementary Figure 13 provides a detailed visualization of the XRO model's seasonal forecast skill for six non-ENSO modes: the North Pacific Meridional Mode (NPMM), South Pacific Meridional Mode (SPMM), Indian Ocean Basin-wide Mode (IOB), Indian Ocean Dipole (IOD), subtropical Indian Ocean Dipole (SIOD), and Tropical North Atlantic (TNA) mode. The ACC heatmaps visually confirm XRO's limitations. For modes like IOB (panel c), the skillful forecast horizon is limited to approximately 6 months, beyond which the ACC drops near zero. For SIOD (panel e) and TNA (panel f), the forecast skill of XRO collapses below the 0.5 correlation threshold within 5–6 months. Even for modes with stronger ENSO connections like NPMM (panel a), XRO's skill is highly seasonal and drops sharply after 8 months. This contrasts sharply with UniCM's performance (main text, Fig. 2b), which maintains skillful, year-round predictions for these modes well beyond 12 months. This provides strong quantitative evidence that a linear, ENSO-centric approach is insufficient for capturing the dynamics of the broader, interconnected climate system.

3.2 Robustness of forecast skill across datasets

To demonstrate that UniCM's superior performance is a robust finding, we conducted a systematic comparison against the XRO baseline for each of the non-ENSO modes across three different reanalysis datasets (ERA5, GODAS, and SODA v2.2.4). The results are presented in Supplementary Figures 17 through 22.

For every mode, UniCM (red line) consistently shows a higher ACC and a longer skillful forecast horizon than XRO (blue line). For instance, in forecasting the TNA mode (Supplementary Fig. 20), UniCM maintains an ACC above the 0.5 skill threshold for nearly the entire 24-month horizon across all three datasets. Specifically, at a 12-month lead time on the ERA5 dataset, UniCM achieves an ACC of 0.52, while XRO's skill has fallen to 0.11. By 18 months, UniCM's ACC is still 0.48, whereas XRO's is below zero. Similarly, for the IOB mode (Supplementary Fig. 17), UniCM extends the skillful forecast lead time from 8 months (XRO) to over 20 months (UniCM) on the ERA5 dataset, which is an improvement of more than 12 months. At a 15-month lead, UniCM's ACC is 0.55, while XRO's is -0.1. This comprehensive, multi-dataset validation confirms that UniCM's ability to learn nonlinear, cross-basin interactions translates into a robust and significant improvement in predictive skill for the entire global system of climate modes.

3.3 Field-significance testing

To rigorously evaluate the predictive performance of UniCM and its baselines (XRO, ResoNet, and CNN), we implement a field-significance testing framework designed to account for spatio-temporal dependencies and the risks of multiple-hypothesis testing.

Monte Carlo Permutation and Null Distributions. We use a non-parametric Monte Carlo permutation approach⁶ to estimate the null distribution of our evaluation metrics (*ACC* and *RMSE*). For each model, we perform $N = 1,000$ iterations. In each iteration, we randomly permute the sample indices of the observed time series to break the association between predictions and observations while maintaining the marginal distribution of the data. For each climate mode m and forecast lead month l , the

local p -value is determined by the empirical rank of the actual model performance against the shuffled null pool:

$$p_{ACC}(m, l) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(ACC_{null,i} \geq ACC_{actual}) \quad (3)$$

$$p_{RMSE}(m, l) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(RMSE_{null,i} \leq RMSE_{actual}) \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

False Discovery Rate (FDR) Correction. To control the inflation of Type I errors across the 7 climate modes and 24 lead months (total $M = 168$ tests), we apply the Benjamini-Hochberg False Discovery Rate (FDR) correction^{7,8}. This procedure ensures that the expected proportion of false discoveries is kept below a specified level α_{fdr} . We adopt a significance threshold of $\alpha_{fdr} = 0.05$. A cell (m, l) in the skill heatmaps is considered “field-significant” and marked with stippling only if its p -value satisfies:

$$p_{(k)} \leq \frac{k}{M} \alpha_{fdr} \quad (5)$$

where $p_{(k)}$ is the k -th smallest p -value in the ranked set of all tests. This rigorous correction confirms that UniCM’s extended predictability horizon represents a robust physical signal rather than a statistical artifact.

Results. As illustrated in the field-significance maps (Supplementary Fig. 23 and Supplementary Fig. 24), “stippling” (black dots) denotes mode-lead combinations that pass the Benjamini-Hochberg (BH) False Discovery Rate (FDR) correction at a global significance level. For the UniCM model, robust field significance is maintained across nearly the entire 24-month forecast horizon for the major climate mode, ENSO, and extends up to 12–18 months for tropical and extra-tropical modes such as the North Pacific Meridional Mode (NPM), Indian Ocean Basin (IOB), and Tropical North Atlantic (TNA). The consistency between the high-correlation regions in the ACC maps and the low-error regions in the RMSE maps confirms that UniCM captures both the phase and magnitude of climate anomalies with high confidence. In contrast, the baseline XRO model exhibits a significantly faster decay in significance, with stippled regions shrinking rapidly beyond 9-13 months for most modes. These results confirm that the performance gains of UniCM are robust against both temporal autocorrelation and the risks of multiple testing, reinforcing the claim that UniCM extends the predictability horizon of the coupled global climate system.

3.4 Comparison with field-to-field forecasting models

Recent advancements in atmospheric and oceanic science have led to the development of deep-learning-based models, such as FuXi-S2S⁹ and CAS-Canglong¹⁰, which are specifically optimized for subseasonal-to-seasonal (S2S) forecasting tasks. These models typically employ a field-to-field prediction paradigm, mapping high-dimensional physical fields at a given time to their future states. These architectures demonstrate state-of-the-art performance in weather and seasonal field forecasting.

To evaluate the performance of UniCM against state-of-the-art subseasonal-to-seasonal (S2S) foundation models, we conducted additional experiments using adapted versions of the FuXi-S2S and CAS-Canglong architectures. While these models are traditionally optimized for field-to-field mapping at synoptic scales, we modified their input layers to accommodate the five physical variables used in our study (SST , τ_x , τ_y , h_T , T_{300}). To maintain consistency, these models were trained using the same dataset and autoregressive rollout strategy as UniCM, targeting a 24-month forecast horizon. Climate mode indices were derived post-hoc from the models’ predicted sea surface temperature (SST) fields.

Prediction skill for multiple climate modes. The results of this comparison are illustrated in Supplementary Fig. 25a (line plot) and Supplementary Fig. 25b (bar chart). As shown in Supplementary Fig. 25a, UniCM maintains a correlation skill for ENSO above 0.5 for 19 months. CAS-Canglong and FuXi-S2S exhibit a significantly faster decay in predictive skill, with correlation scores falling below 0.5 at 9 months and 6 months, respectively. The bar chart in Supplementary Fig. 25b demonstrates that UniCM outperforms these S2S models across nearly all tested modes, including IOB, SIOD, TNA, and various NINO regions. The S2S models struggle particularly with modes governed by complex, large-scale couplings like SIOD and IOD. The rapid decay of skill in FuXi-S2S and CAS-Canglong at interannual lead times suggests that architectures optimized for short-term weather dynamics may prioritize high-frequency signals that effectively become unpredictable noise at longer leads. UniCM’s success lies in its ability to explicitly decouple and then re-integrate the slow-moving, large-scale climate modes with local physical fields.

Spatial prediction. While climate mode indices provide a condensed representation of large-scale variability, evaluating the underlying spatial fields is essential to confirm that a model’s predictive skill is rooted in physically realistic representations of the ocean-atmosphere system. To further validate UniCM, we conduct a global spatial benchmark against the adapted architectures of FuXi-S2S and CAS-Canglong at various forecast horizons.

We take the spatial prediction performance at a 12-month lead time (Supplementary Fig. 29) as a representative example. UniCM demonstrates superior performance across both metrics. UniCM maintains a global mean ACC of 0.379, whereas FuXi-S2S and CAS-Canglong achieve 0.262 and 0.207, respectively. Notably, UniCM retains significant correlation skill in the Equatorial Pacific, where baseline models struggle to maintain a coherent signal. Meanwhile, UniCM achieves a reduced global mean RMSE of 0.465, compared to 0.505 for FuXi and 0.515 for Canglong. The improvement is most prominent in the core ENSO region, which is critical for steering global inter-basin teleconnections. It confirms that UniCM effectively captures the dynamics of the Equatorial Pacific, the pacemaker of the global coupled system^{11–13}. By maintaining lower error and higher correlation in this key region, UniCM provides a more reliable foundation for long-lead climate forecasting than standard field-to-field S2S architectures.

We further report spatial prediction results at additional lead times, including 1 month (Supplementary Fig. 26), 3 months (Supplementary Fig. 27), 6 months (Supplementary Fig. 28), and 24 months (Supplementary Fig. 30). As the forecast lead time increases, the ACC gradually decreases and the RMSE correspondingly increases, reflecting the expected decay of predictability in a chaotic climate system and the growing influence of internal variability at longer horizons. Despite this inevitable degradation, UniCM consistently outperforms the baseline models at all lead times, demonstrating a clear advantage in capturing long-term, large-scale ocean-atmosphere dynamics.

4 Analysis of learned correlations between climate modes

4.1 Lagged-correlation evaluation

A central advance of UniCM is its ability to learn physically meaningful, temporal correlations between climate modes directly from data. The main text (Figs. 3e-f) demonstrated this for the ENSO-NPMM and ENSO-TNA teleconnections. Here, we show that this capability extends to a much wider range of global teleconnections, confirming that the model has learned a generalizable representation of the coupled climate system.

Supplementary Figures 31, 32, 33, and 34 present an extensive set of lagged cross-correlations between pairs of climate modes, validated against four different datasets. In these panels, the model’s task is to reproduce the observed ground truth correlation (black line) at various positive and negative lags. Across this wide array of teleconnections, UniCM’s forecasts (red lines) consistently and accurately replicate the timing, magnitude, and asymmetry of the observed relationships. For example, in the teleconnection between the IOD and SPM (Supplementary Fig. 31c), UniCM correctly captures the peak positive correlation of $r \approx +0.4$ at a lag of approximately +4 months, closely matching the ground truth. In contrast, the XRO forecast (blue line) overestimates the correlation magnitude to $r \approx +0.6$ and fails to capture the precise temporal structure of the peak. Similarly, for the interaction between the IOB and NPMM (Supplementary Fig. 31j), UniCM accurately reproduces the observed persistent negative correlation of $r \approx -0.2$, whereas XRO incorrectly forecasts a near-zero correlation, failing to capture this inter-basin relationship entirely. This high fidelity extends to other key teleconnections. For the SPM-TNA interaction (Supplementary Fig. 31a), UniCM correctly identifies a peak positive correlation of $r \approx +0.4$ with TNA leading SPM by approximately 7 months, a relationship that is entirely absent in the XRO forecast. This comprehensive validation across dozens of mode pairs and multiple datasets confirms that UniCM has learned a physically realistic and robust representation of the interconnected global climate system.

4.2 Precursors and couplings for non-ENSO extreme events

Event identification. To identify significant events in non-ENSO modes (IOD, NPMM, SPM, IOB, SIOD, and TNA), we adopt a relative threshold based on standard deviations (σ) rather than a fixed temperature value. This is necessitated by the fact that background sea surface temperature (SST) anomaly variance differs significantly across ocean basins¹¹. Following standard practice^{11,14} in climate science to ensure comparable event intensity, we define a significant event as the index exceeding a unified relative threshold:

$$|Index| > 1.0\sigma \text{ for at least five consecutive months.} \quad (6)$$

The identified events across different modes are illustrated in the SSTA time series analysis (Supplementary Figs. 35a-f).

Analysis of interaction patterns. By extracting the internal attention weights of the Modeformer branch during the 12 months preceding these identified events, we quantify the inter-mode interaction strengths. This provides several insights:

1. **1986 NPMM transition:** For the 1986 NPMM event (Supplementary Fig. 36a), UniCM identified a sharp transition from stochastic noise to organized coupling, with prominent interaction weights between the NPMM and other climate modes.

2. **1994 TNA-basin linkage:** During the 1994 IOD/TNA/NPMM co-occurrence (Supplementary Fig. 36b), the model revealed an active role for the TNA in modulating Indian Ocean variability. The TNA emerged as a significant driver, suggesting inter-basin teleconnections that transcend the typical ENSO-centric view.
3. **2011 multi-mode extreme:** In the multi-mode extreme of 2011 (Supplementary Fig. 36c), UniCM autonomously identified the TNA as the dominant “hub” node in the global interaction network. The high attention weights concentrated on the TNA column indicate that it coordinated the synchronization of anomalies across the Pacific (NPMM/SPMM) and Indian (SIOD) basins.

These findings demonstrate that UniCM is capable of revealing how the global climate system shifts between different mechanical regimes and highlighting the role of inter-basin “hubs” in coordinating extreme events.

5 Ablation studies and sensitivity analysis

In this section, we provide a systematic evaluation of the architectural components of UniCM and its sensitivity to different hyperparameter configurations. These experiments validate the necessity of the dual-branch design and the robustness of the model across different training regimes.

5.1 Architectural ablation study

To evaluate the contribution of the individual modules to the overall forecasting skill, we conducted two sets of ablation experiments focusing on the integration of physical inductive biases and the dual-branch Transformer architecture.

Integration strategy for mode dynamics We investigated two primary strategies for fusing the large-scale climate mode dynamics (learned by the ModeFormer) with the fine-grained global feature space:

- **Replacement strategy:** Substituting specific latent spatial representations with mode-focused embeddings.
- **Additive strategy (UniCM Default):** Dynamically adding mode-specific dynamics to the global feature maps via a learned projection layer.

As shown in Supplementary Fig. 37a, the *additive strategy* consistently outperformed the *replacement strategy*. We attribute this to the additive approach’s ability to preserve global spatial continuity while allowing the mode-specific information to act as a diagnostic constraint on the global manifold, preventing information bottlenecks.

Synergy of the dual-branch architecture To assess the necessity of capturing multi-scale interactions, we compared three architectural variants:

- **GlobalFormer-only:** A single-branch model processing only the high-resolution global spatio-temporal fields ($5^\circ \times 5^\circ$).
- **ModeFormer-only:** A model focusing exclusively on the interactions between predefined climate indices.
- **UniCM (Hybrid):** The full dual-branch architecture.

The hybrid UniCM demonstrates a performance gain over both single-branch variants (Supplementary Fig. 37b). This confirms that while the GlobalFormer captures the high-dimensional physical state, the ModeFormer provides critical teleconnection constraints. The synergy between these scales is vital for capturing the long-term oceanic memory required for skillful multi-year forecasts.

5.2 Hyperparameter sensitivity analysis

The stability of UniCM was evaluated through a series of sensitivity tests on key hyperparameters, and the results are shown in Supplementary Fig. 38.

Learning rate. A grid search was performed over $\eta \in \{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$. We observed that $\eta = 5 \times 10^{-4}$ yielded the optimal balance between convergence stability and asymptotic performance (see Supplementary Fig. 38a). While higher learning rates accelerated initial training, they led to higher variance in the validation loss during the fine-tuning phase.

Loss function balancing. According to the UniCM framework, the total loss $\mathcal{L}_{total}$ is a multi-objective weighted sum of three components:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{field} + \lambda_2 \mathcal{L}_{global-mode} + \lambda_3 \mathcal{L}_{mode-aux} \quad (7)$$

where $\mathcal{L}_{field}$ represents the MSE for global physical field forecasts from the Globalformer, $\mathcal{L}_{global-mode}$ optimizes the mode indices from the Globalformer, and $\mathcal{L}_{mode-aux}$ is the MSE for multiple climate mode indices from the Modeformer. In our sensitivity analysis, we evaluated the impact of varying the coefficients $\lambda_1, \lambda_2, \lambda_3$. As shown in Supplementary Fig. 38b, we confirmed that setting $\lambda_1 = \lambda_3 = 1.0$, and $\lambda_2 = 0.01$ ensures that all objectives, ranging from global physical consistency to structured inter-mode interactions, are meaningfully optimized without the model overfitting to a single task.

Model size. We evaluated the impact of model depth (number of blocks) and embedding dimensions (d_{model}). As shown in Supplementary Fig. 38c, increasing model capacity generally correlates with improved forecast skill at long lead times. This suggests that the UniCM architecture is not yet saturated and possesses the expressive power to benefit from larger simulation ensembles or higher-resolution data in future iterations.

5.3 Quantitative assessment of training data requirements

The observational record provides high-quality data primarily for the post-satellite era (approx. 1980–present). This duration is insufficient to cover the full spectrum of multi-decadal climate variability. To circumvent this limitation, we utilized CMIP6 historical runs spanning 165 years (1850–2014). This extended period exposes the model to various climate background states, including different phases of the Interdecadal Pacific Oscillation (IPO) and varying degrees of anthropogenic forcing, thereby enhancing the model’s ability to generalize to unseen real-world conditions. Training samples were generated using a sliding window approach applied to the 165 years of monthly data ($165 \times 12 = 1,980$ months). Each sample consists of a 12-month input sequence and a 24-month target forecast horizon. We employed a 1-month stride for the primary training phase, producing a total of $N = 1,980 - 12 - 24 + 1 = 1,945$ training samples. Notably, with a 1-month stride, consecutive samples share $\approx 91.7\%$ (11/12) of their input features. While this high frequency ensures the model learns the smooth temporal evolution of climate states, it introduces significant temporal autocorrelation and information redundancy.

To disentangle the impact of raw sample count (data density) from temporal coverage (climatological diversity), we performed a sensitivity study using two sub-sampling strategies:

- **Temporal continuous subsets:** Training on chronological segments representing the first 20%, 40%, 60%, and 80% of the 165-year span to simulate shorter historical records.
- **Randomly sampled subsets:** Training on samples randomly selected from the total pool of 1,945, maintaining the same total volume as the temporal subsets to isolate the effect of information density.

The results reveal a non-linear scaling relationship between performance and training data volume (Supplementary Fig. 39). In random subset experiments (Supplementary Fig. 39a-b), a model trained on only 40% of the samples achieves nearly full predictive skill, uncovering the temporal redundancy inherent in the 1-month stride. Because consecutive samples feature over 90% overlap, the model encounters highly autocorrelated gradients, leading to diminishing marginal returns once the underlying dynamical manifold is captured. Conversely, temporal subsets (Supplementary Fig. 39c-d) exhibit a performance decline compared to random subsets of the same size, proving that climatological diversity (temporal span) is more critical than raw data density. While random sampling across the 165-year record exposes the model to a comprehensive manifold of decadal oscillations and varying background states, continuous shorter segments lack the necessary variability for robust generalization. These findings suggest that for global climate modeling, priority should be placed on extending the historical horizon to capture complete physical cycles rather than increasing sampling frequency. Practically, increasing the training stride to 3 or 6 months can significantly reduce computational overhead with negligible loss in forecast skill.

5.4 Grid-resolution sensitivity

To test whether the use of a coarse grid ($5^\circ \times 5^\circ$) is a potential source of model bias or primarily a computational choice, we conducted an additional grid-resolution sensitivity experiment at a higher spatial resolution of $2.5^\circ \times 2.5^\circ$. We regridded the same input physical fields and climate-mode indices to the 2.5° latitude–longitude grid using the same preprocessing pipeline, temporal coverage, normalization, and train/validation/test split as in the main experiments. Unless otherwise stated, we kept the UniCM architecture and training protocol identical (loss function, optimizer, learning-rate schedule, early stopping, and evaluation metrics), and only adjusted runtime-dependent settings (e.g., batch size) when necessary to fit GPU memory constraints due to the increased token count.

As shown in Supplementary Fig. 40, increasing the spatial resolution from 5° to 2.5° does not yield a performance gain in climate-mode prediction skill. Instead, the mean ACC across climate modes decreases slightly from 0.473 (at 5°) to 0.464 (at

2.5°). This result suggests that the 5° grid does not appear to be the primary limiting factor for UniCM; rather, it provides an appropriate scale for learning the large-scale, low-frequency coupled dynamics that dominate seasonal-to-interannual climate variability. In contrast, the finer grid substantially increases the number of spatial tokens and the Transformer attention cost, and may expose the model to more small-scale variability that is largely stochastic at these lead times, thereby offering limited benefit for predictability while reducing computational efficiency.

6 Analysis of pre-training and fine-tuning

To evaluate the model's adaptability to real-world observations, we investigate whether fine-tuning UniCM on historical reanalysis data (e.g., from 1958–1980) can further enhance its predictive performance. As illustrated in Supplementary Fig. 41, the performance gap between the pre-trained-only model and the fine-tuned model is relatively modest. We interpret this as a critical indicator of the framework's robustness, attributed to several factors.

By training on 165-years of records, UniCM successfully extracted the fundamental dynamical laws governing climate teleconnections. The high correlation skill of the pre-training curve (> 0.5 at a 19-month lead for ENSO) suggests that the model's inductive bias is sufficient to bridge the “sim-to-real” gap with minimal observational correction. Because the input CMIP6 models were pre-selected for physical accuracy, the knowledge gained during pre-training was already highly aligned with the real-world dynamical manifold. The 20-year (two-decade) window of reanalysis data provides a limited number of realization cycles for multi-decadal oscillations compared to the 165-year CMIP6 record. Consequently, fine-tuning serves as a local calibration of the model's weights rather than a fundamental reorganization of the learned physics.

Supplementary Tables

To ensure the clarity and reproducibility of our study, we provide supplementary tables detailing the model’s architecture, the datasets used, and the scientific definitions of the climate modes. Supplementary Table 2 provides a comprehensive summary of the architectural and training settings for the UniCM framework. Supplementary Table 3 outlines all datasets used for model training, validation, and testing, along with their respective temporal coverages and sources. Supplementary Table 4 presents the standard scientific definitions for the seven key climate modes analyzed in this study, including their acronyms and precise geographical boundaries for calculating the Sea Surface Temperature Anomaly (SSTA) indices.

Supplementary Table. 1. Summary of architectural and training parameters for baseline models.

Deep-learning baselines		
Parameter	CNN ¹	ResoNet ²
Model category	Convolutional neural network	Hybrid CNN–Transformer
Main components	3 Conv, 2 MaxPool, 1 FC	2-stage attention blocks
Kernel / window size	$8 \times 4, 4 \times 2, 4 \times 2$	$3 \times 3, 7 \times 7$ (attention)
Feature map / hidden dim	30 or 50 filters	128 (stage 1), 256 (stage 2)
Activation function	Tanh / ReLU	GELU
Optimizer	Adam	AdamW
Learning rate (η)	5×10^{-4}	1×10^{-3}
Loss function	MSE	MSE
Physics-guided / dynamical baselines		
Parameter	XRO ³	DESN ⁴
Model category	Nonlinear dynamical model	Physics-guided deep echo state network
Main components	Extended nonlinear recharge oscillator	ReservoirPy reservoir + ridge readout
Kernel / window size	N/A	N/A
State size / hidden dimension	State variables	4000 reservoir units
Activation function / dynamics	Prescribed nonlinear coefficients	Leaky reservoir dynamics
Optimizer / solver	Multivariate regression	Ridge readout fitting (no backpropagation)
Learning rate (η)	N/A	N/A (leak rate = 1)
Loss function	MSE	MSE

Note: DESN was implemented as a reservoir computing model with one-step-ahead fitting and autoregressive multi-step rollout. In the accessible implementation, typical hyperparameters include 4000 reservoir units, spectral radius 0.95, reservoir connectivity 0.14–0.2, input connectivity 0.1, input scaling 1, reservoir/input noise around 0.01, and ridge regularization of 5×10^{-6} to 6×10^{-6} . Here, the reported leak rate for DESN refers to the reservoir dynamics rather than an optimizer learning rate.

Supplementary Table. 2. Model and training hyperparameters. This table details the architectural configurations, training settings, and reproducibility protocols for the UniCM framework.

Category	Parameter	Value / Description
Architecture	Hidden Dimension (D)	256 (for both Globalformer & Modeformer)
	Number of Layers	8 (4 encoder + 4 decoder) per branch
	Attention Heads	4
	Dropout Rate	0.2 (applied to attention and feedforward layers)
Input/Output	Spatial Resolution	$5^\circ \times 5^\circ$ (interpolated physical fields)
	Patch Size ($p \times p$)	2×2 (spatial tokenization)
	Input Window (T_h)	12 months (historical context)
	Forecast Horizon (T_p)	24 months (consistent across all evaluation)
Training	Optimizer	AdamW
	Learning Rate (LR)	5×10^{-4} with linear warm-up (3 epochs)
	LR Scheduler	Cosine annealing decay
	Weight Decay	1×10^{-6}
	Batch Size	32
	Gradient Clipping	1.0 (ℓ_2 norm threshold)
	Training Epochs	200 (with best checkpoint selection)
Reproducibility	Random Seeds	1–20 (Results are means of 20 independent runs)
	Initialization	Sliding-window protocol (1-month step)
	Hardware	Single NVIDIA A100 GPU (80 GB)

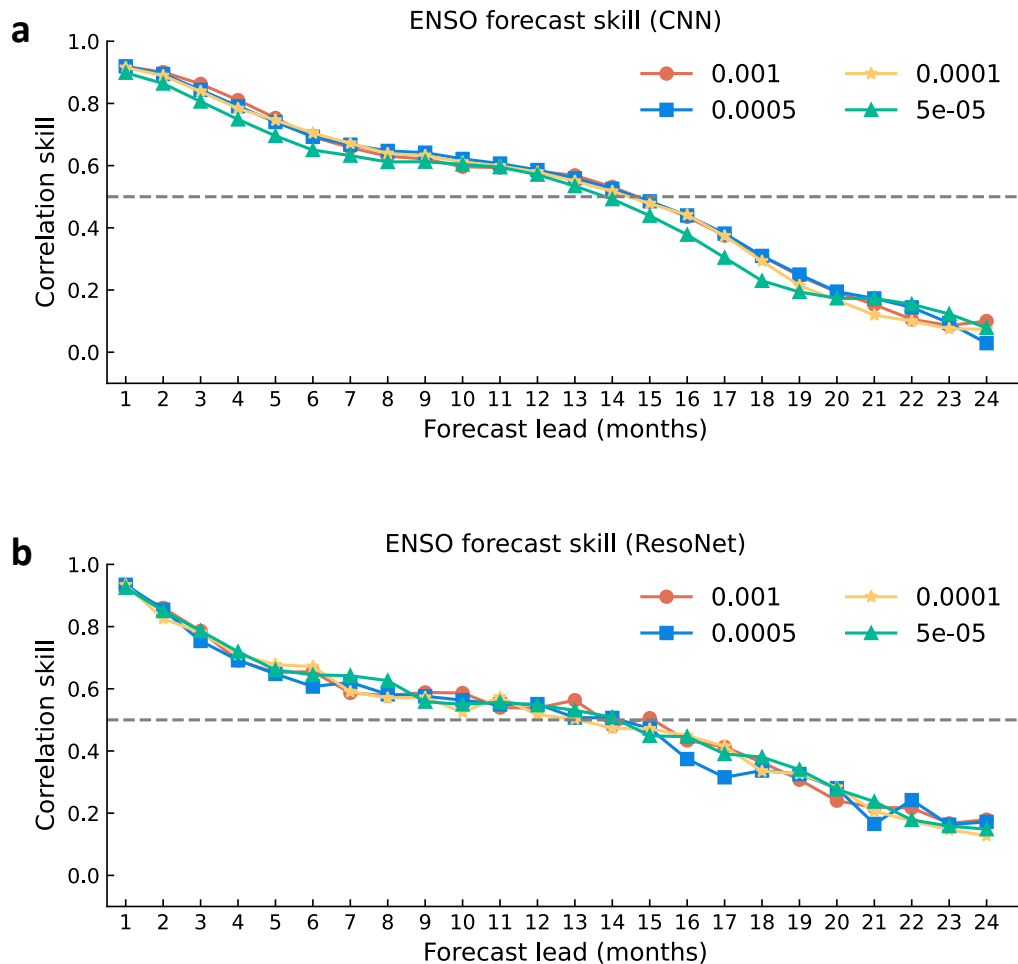
Supplementary Table. 3. Dataset specifications. Summary of the datasets used for model training, validation, and forecast performance testing.

Dataset	Purpose	Temporal Coverage	Source
CMIP6	Training	1850–2014	https://cds.climate.copernicus.eu/datasets/projections-cmip6
ORAS5	Validation	1958–1980	https://cds.climate.copernicus.eu/datasets/reanalysis-oras5
ERA5	Testing	1980–2014	https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means
GODAS	Testing	1980–2021	https://psl.noaa.gov/data/gridded/data.godas.html
ORAS5	Testing	1980–2024	https://cds.climate.copernicus.eu/datasets/reanalysis-oras5
SODA v2.2.4	Testing	1980–2011	https://soda.umd.edu/

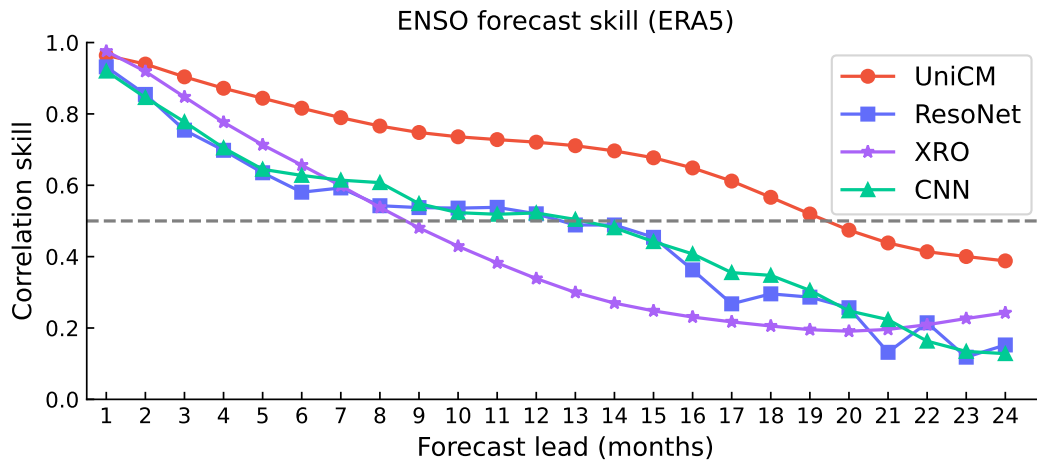
Supplementary Table. 4. Definition of the Sea Surface Temperature (SST) indices for climate modes used in this study, adapted from Zhao et al.³.

Climate Mode	Acronym	Description	References
El Niño–Southern Oscillation	ENSO	SSTAs averaged over Niño3.4 region 170°–120°W, 5°S–5°N	(Trenberth, 1997 ⁵)
North Pacific Meridional Mode	NPMM	SSTAs averaged over 160°–120°W, 10°–25°N	
South Pacific Meridional Mode	SPMM	SSTAs averaged over 110°–90°W, 25°–15°S	(Zhang et al., 2014 ¹⁵)
Indian Ocean Basin Mode	IOB	SSTAs averaged over 40°–100°E, 20°S–20°N	(Xie et al., 2009 ¹⁶)
Indian Ocean Dipole Mode	IOD	SSTAs over 50°–70°E, 10°S–10°N minus those over 90°–110°E, 10°S–0°N	((Saji et al., 1999 ¹⁷)
Southern Indian Ocean Dipole Mode	SIOD	SSTAs over 65°–85°E, 25°–10°S minus those over 90°–120°E, 30°–10°S	(Jo et al., 2022 ¹⁸)
Tropical North Atlantic variability	TNA	SSTAs averaged over 55°–15°W, 5°–25°N	(Enfield et al., 1999 ¹⁹)

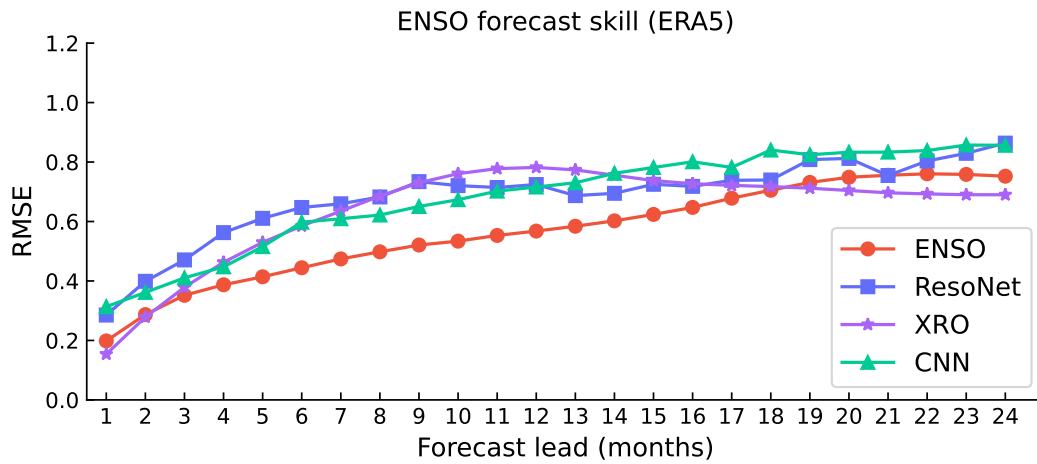
Supplementary Figures



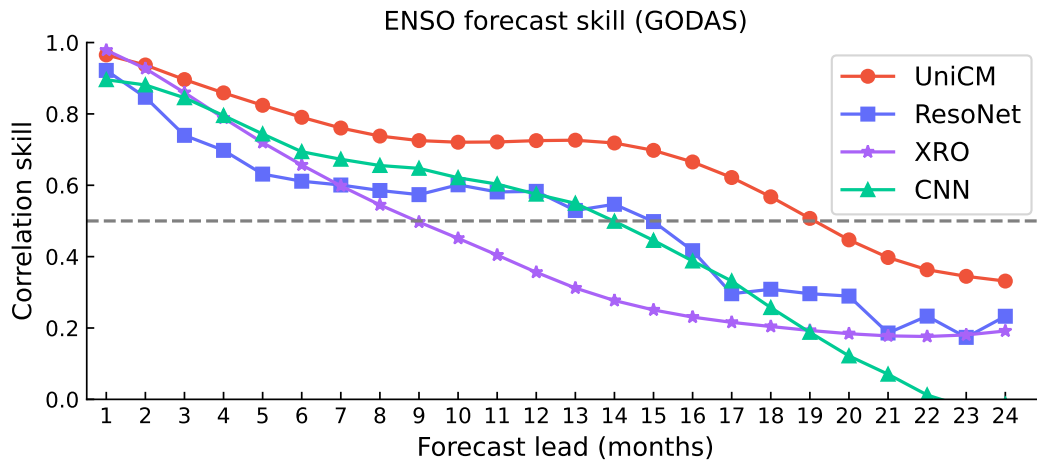
Supplementary Figure 1. Learning-rate tuning for the deep learning baselines. Correlation skill (ACC) of ENSO forecasts as a function of lead time for **a.** CNN and **b.** ResoNet under different learning rates; the best-performing learning rate on the validation set is adopted for the baseline comparisons.



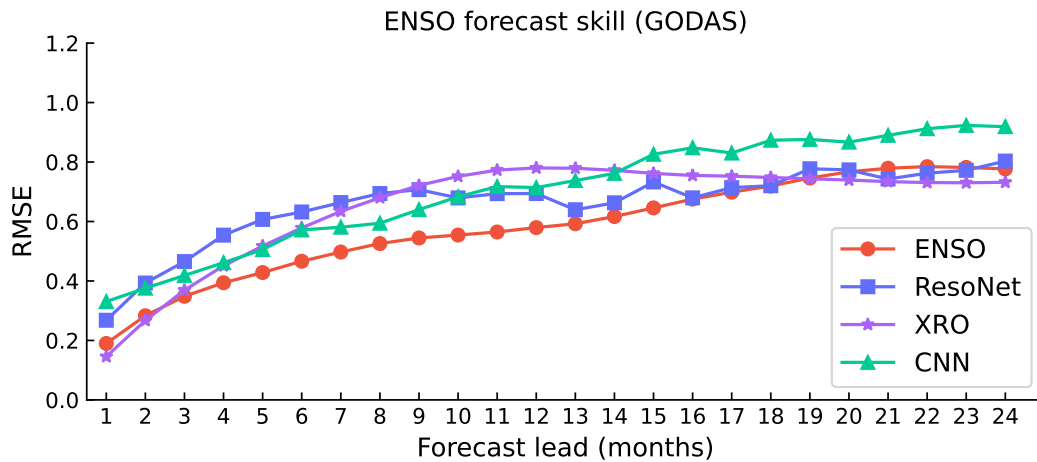
Supplementary Figure 2. ENSO forecast skill comparison. Anomaly Correlation Coefficient (ACC) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the ERA5 dataset. The dashed gray line indicates an ACC skill threshold of 0.5.



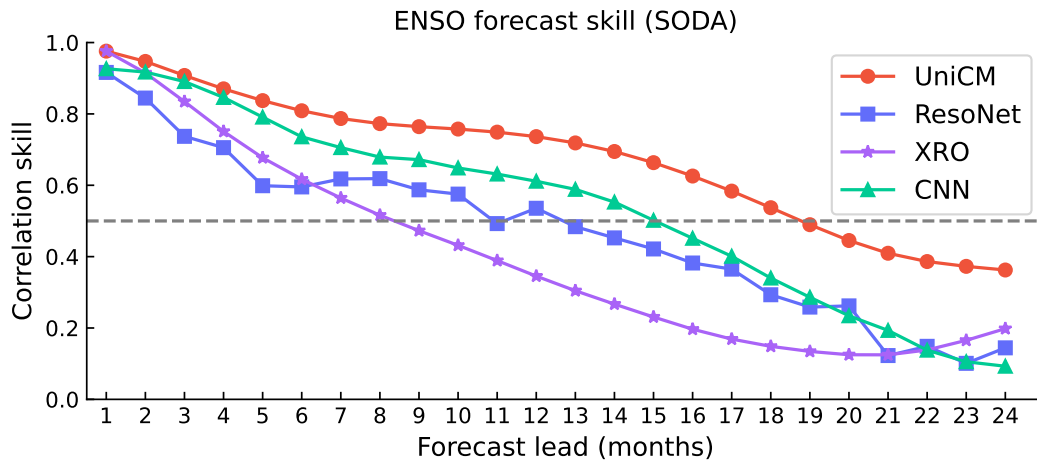
Supplementary Figure 3. ENSO forecast error comparison. Root Mean Square Error (RMSE) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the ERA5 dataset. Lower RMSE values indicate better forecast skill.



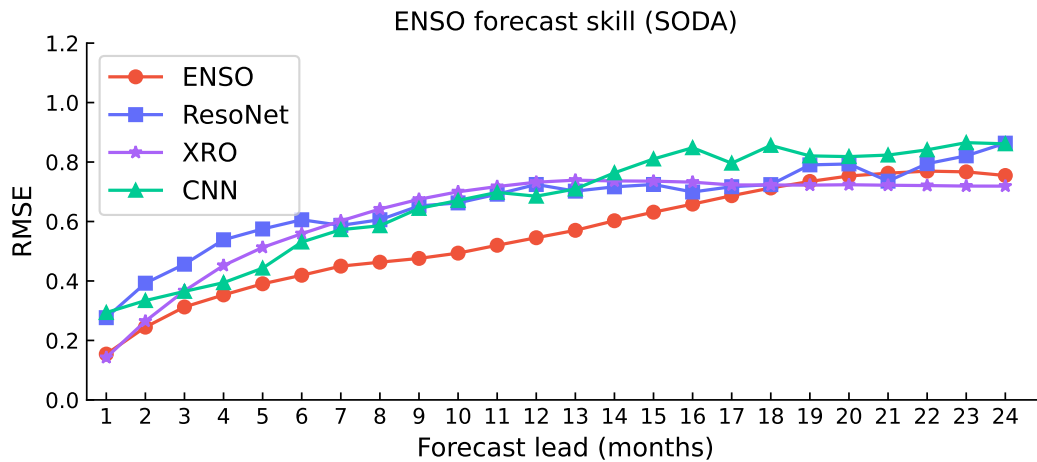
Supplementary Figure 4. ENSO forecast skill comparison. Anomaly Correlation Coefficient (ACC) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the GODAS dataset. The dashed gray line indicates an ACC skill threshold of 0.5.



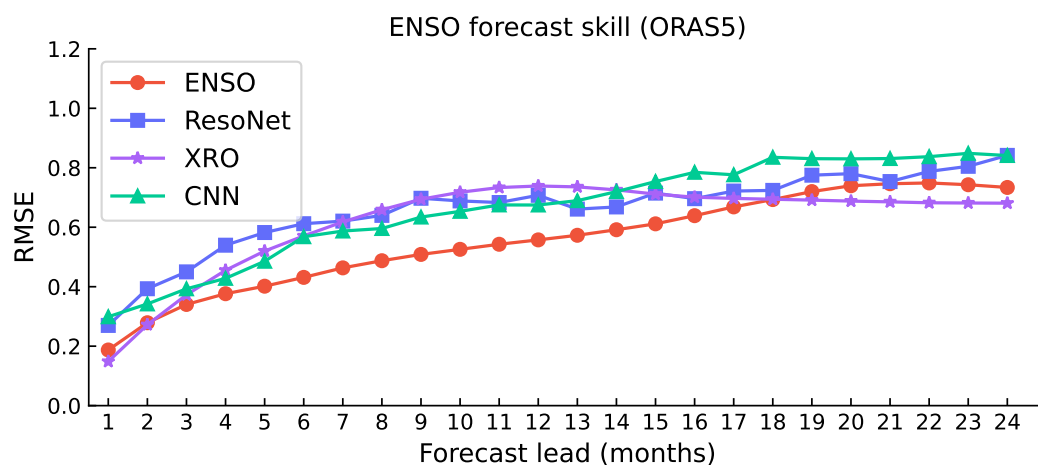
Supplementary Figure 5. ENSO forecast error comparison. Root Mean Square Error (RMSE) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the GODAS dataset. Lower RMSE values indicate better forecast skill.



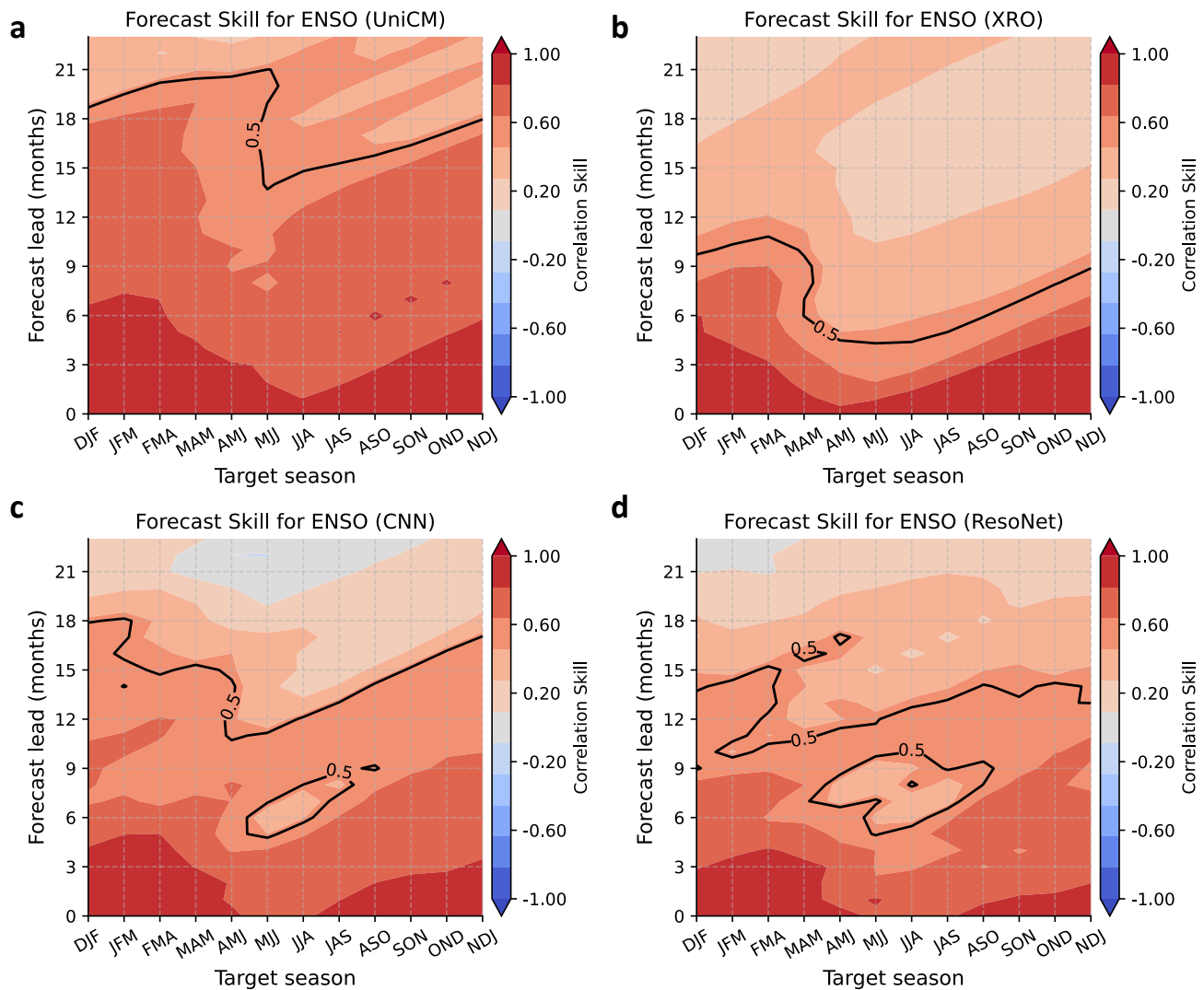
Supplementary Figure 6. ENSO forecast skill comparison. Anomaly Correlation Coefficient (ACC) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the SODA v2.2.4 dataset. The dashed gray line indicates an ACC skill threshold of 0.5.



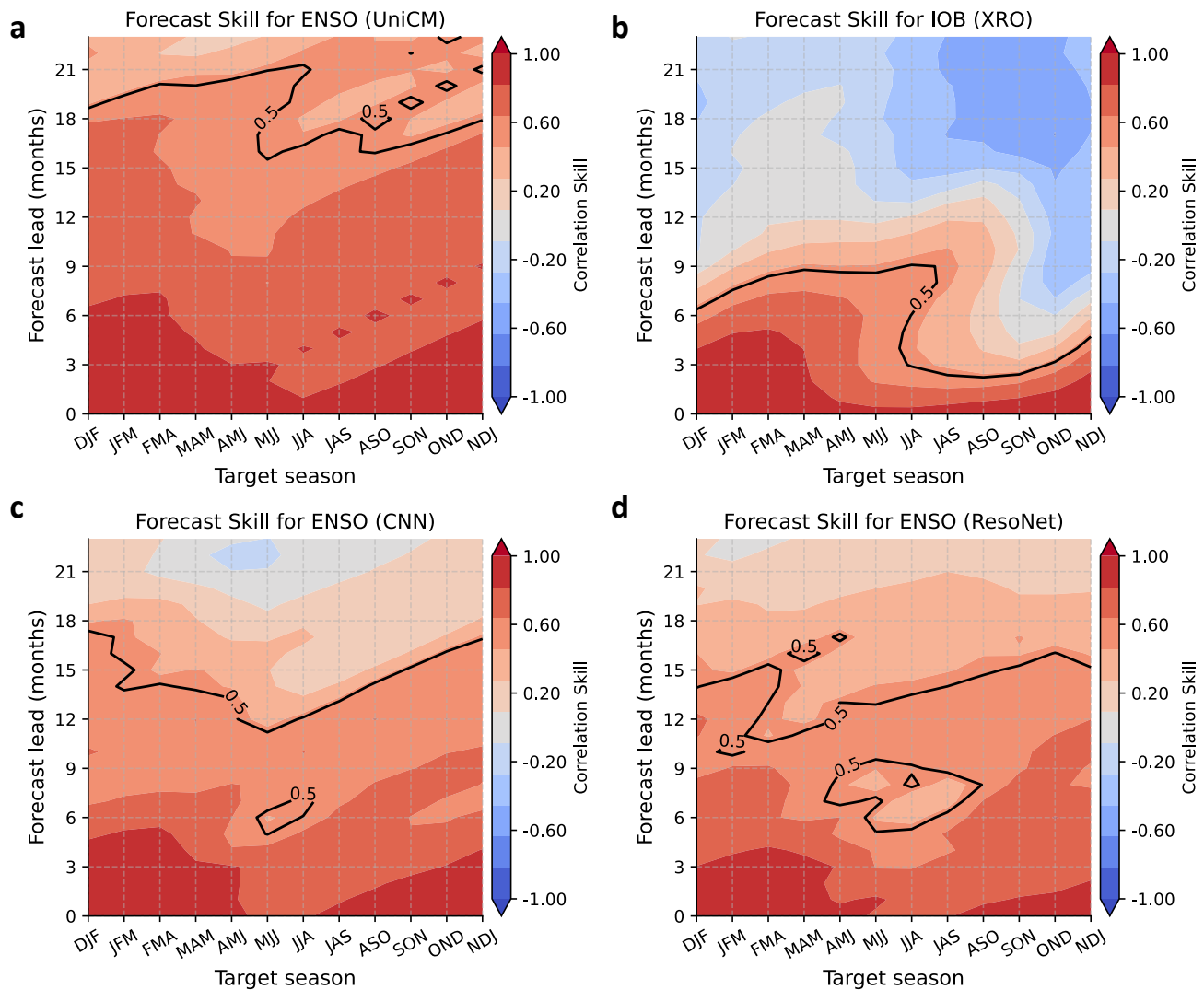
Supplementary Figure 7. ENSO forecast error comparison. Root Mean Square Error (RMSE) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the SODA v2.2.4 dataset. Lower RMSE values indicate better forecast skill.



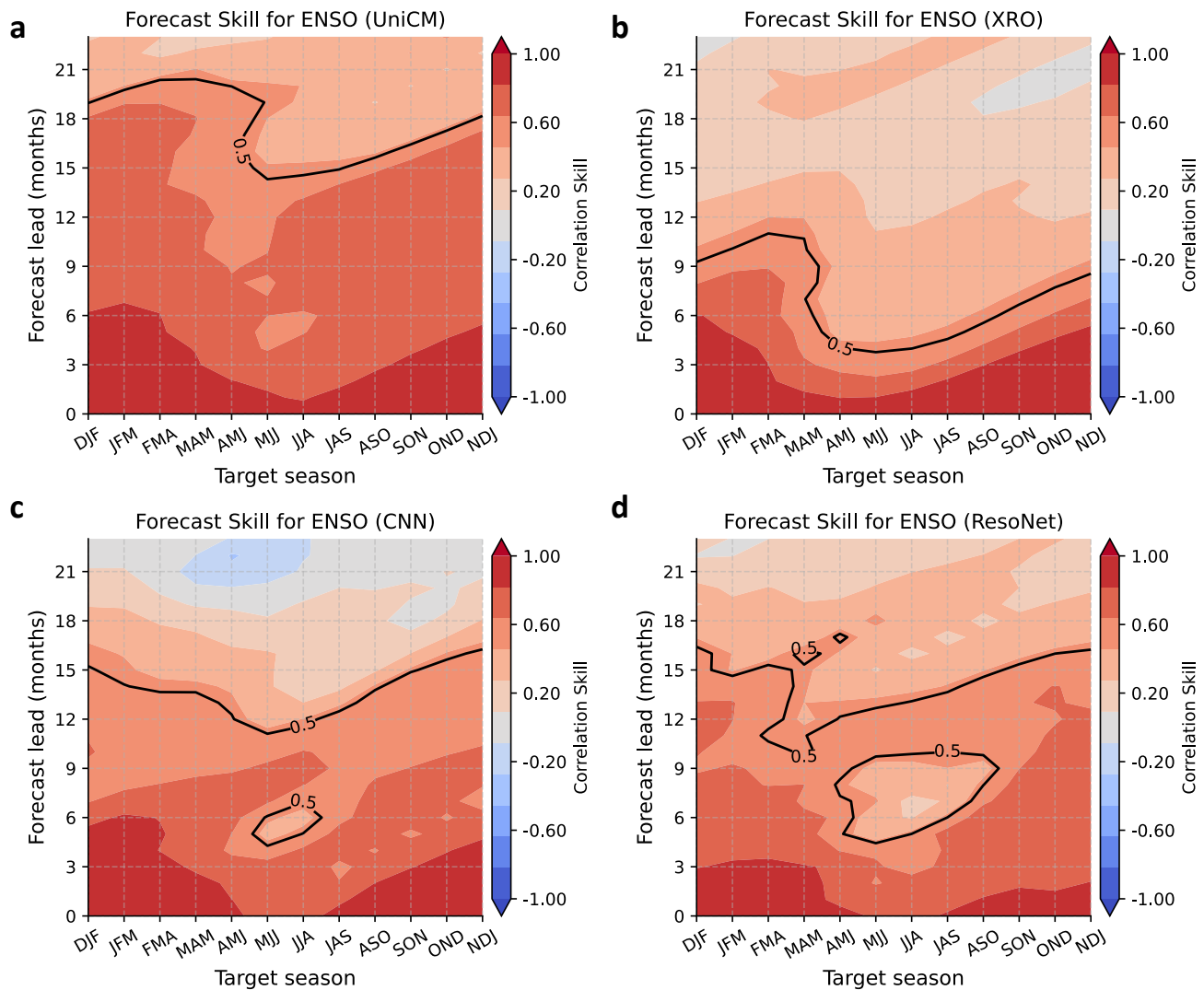
Supplementary Figure 8. ENSO forecast error comparison. Root Mean Square Error (RMSE) for El Niño-Southern Oscillation (ENSO) forecasts from the UniCM model and three baseline models (ResoNet, XRO, and CNN). Forecasts are validated against the ORAS5 dataset. Lower RMSE values indicate better forecast skill.



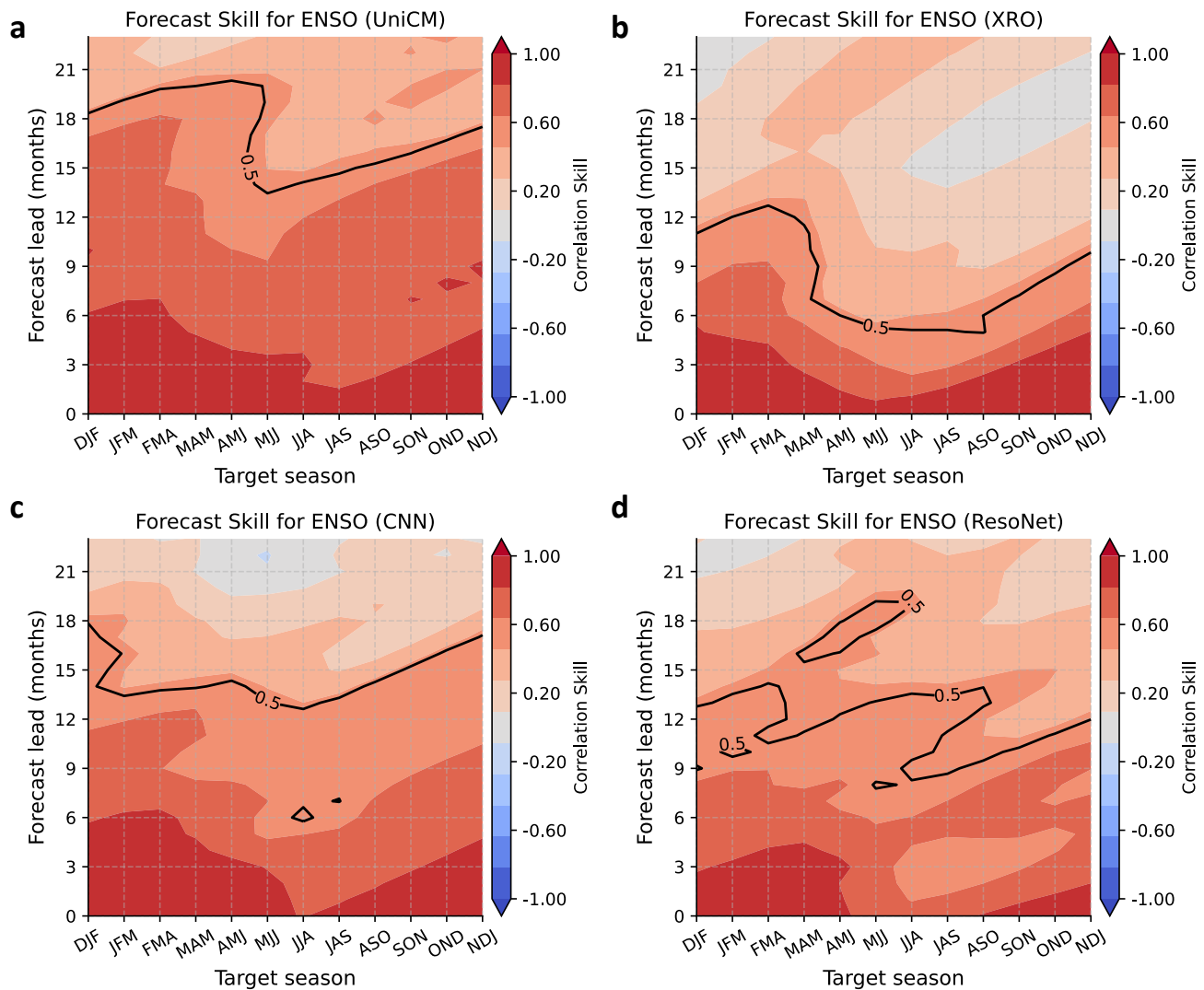
Supplementary Figure 9. Forecast skill for ENSO across lead times and target seasons. Anomaly Correlation Coefficient (ACC) skill maps for ENSO forecasts generated by **a.** UniCM, **b.** XRO, **c.** CNN, and **d.** ResoNet models, validated against the ERA5 dataset. The horizontal axis indicates the target season, and the vertical axis represents the forecast lead time (months). Color shading denotes the ACC, with values above 0.5 (black contour) generally considered skillful.



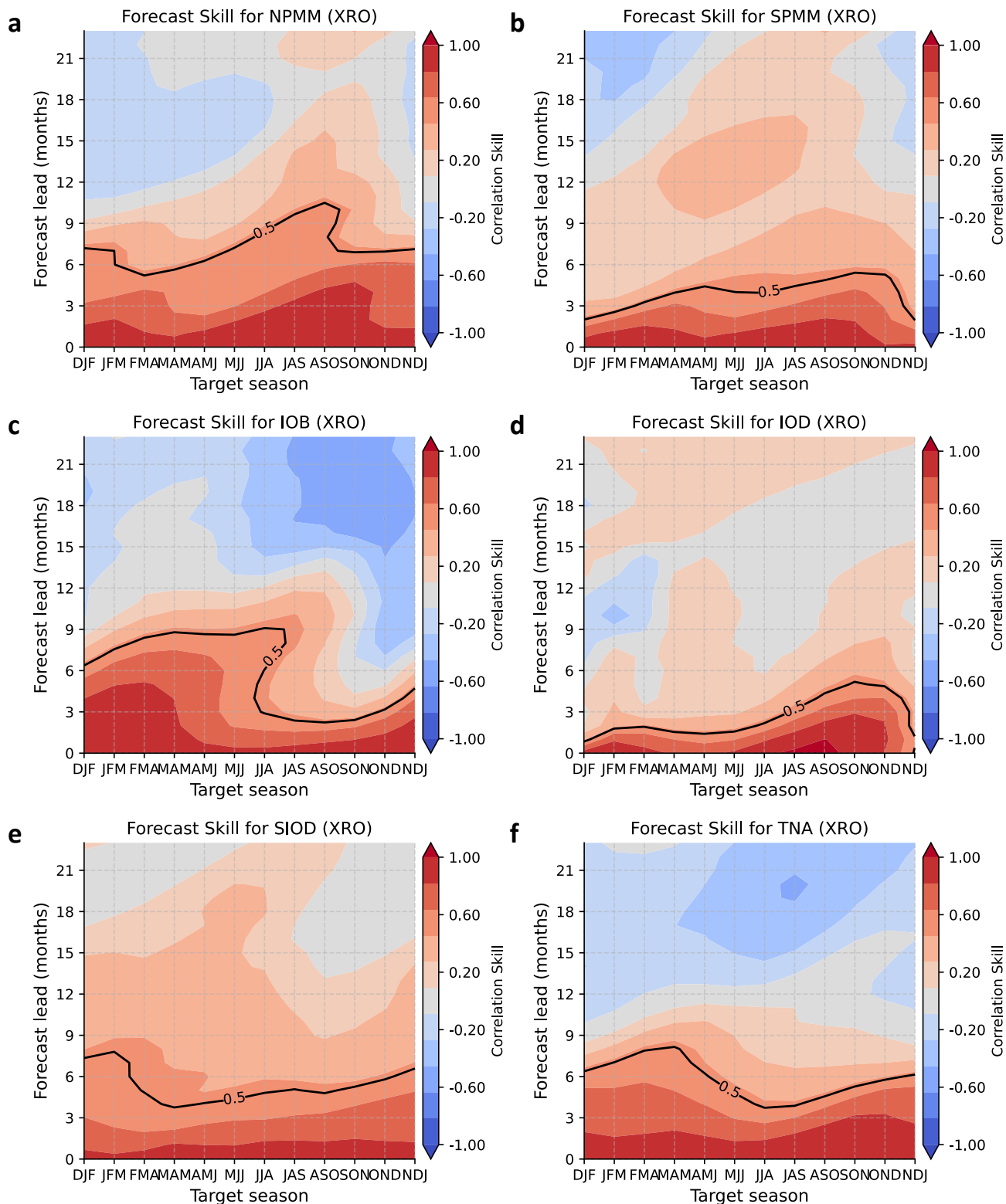
Supplementary Figure 10. Forecast skill for ENSO across lead times and target seasons. Anomaly Correlation Coefficient (ACC) skill maps for ENSO forecasts generated by **a.** UniCM, **b.** XRO, **c.** CNN, and **d.** ResoNet models, validated against the ORAS5 dataset. The horizontal axis indicates the target season, and the vertical axis represents the forecast lead time (months). Color shading denotes the ACC, with values above 0.5 (black contour) generally considered skillful.



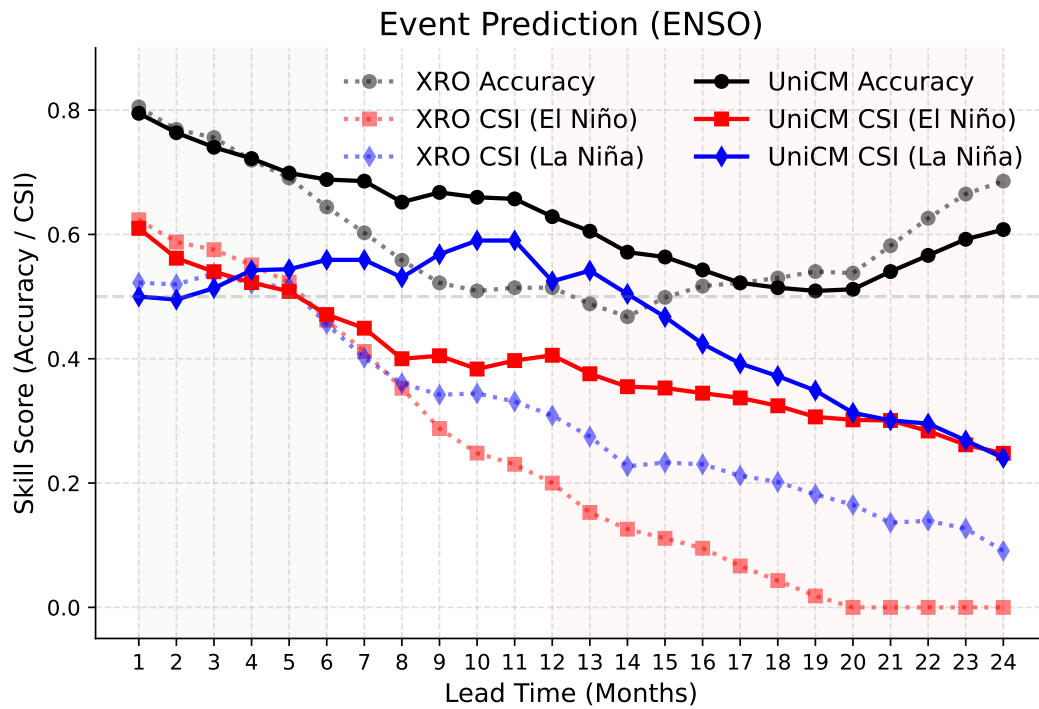
Supplementary Figure 11. Forecast skill for ENSO across lead times and target seasons. Anomaly Correlation Coefficient (ACC) skill maps for ENSO forecasts generated by **a.** UniCM, **b.** XRO, **c.** CNN, and **d.** ResoNet models, validated against the GODAS dataset. The horizontal axis indicates the target season, and the vertical axis represents the forecast lead time (months). Color shading denotes the ACC, with values above 0.5 (black contour) generally considered skillful.



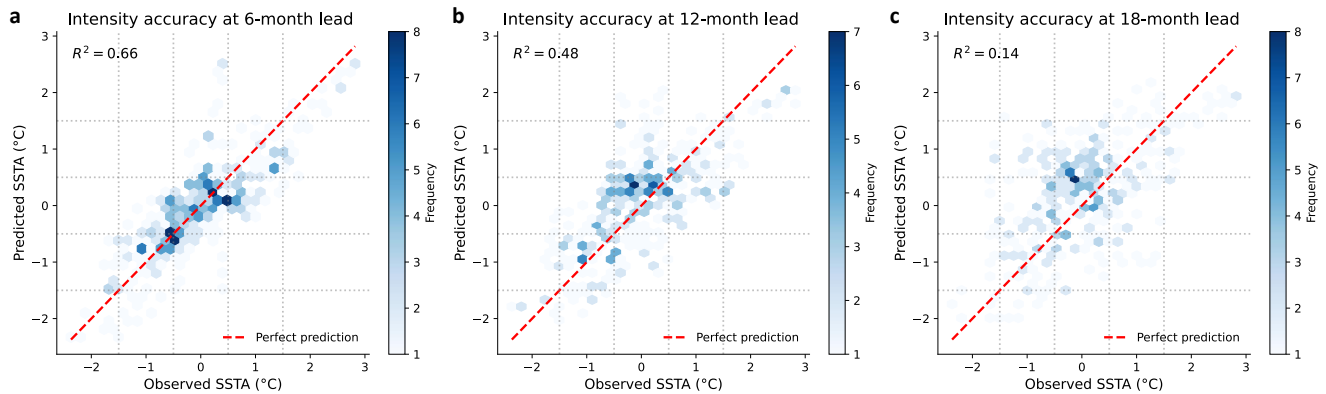
Supplementary Figure 12. Forecast skill for ENSO across lead times and target seasons. Anomaly Correlation Coefficient (ACC) skill maps for ENSO forecasts generated by **a.** UniCM, **b.** XRO, **c.** CNN, and **d.** ResoNet models, validated against the SODA v2.2.4 dataset. The horizontal axis indicates the target season, and the vertical axis represents the forecast lead time (months). Color shading denotes the ACC, with values above 0.5 (black contour) generally considered skillful.



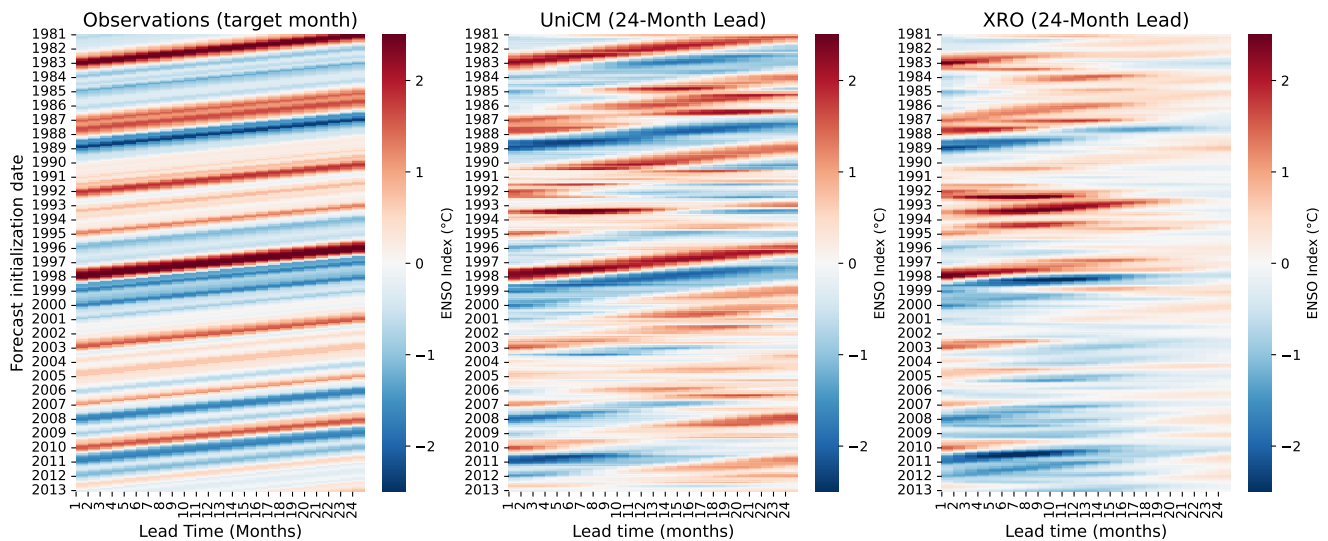
Supplementary Figure 13. Forecast skill of the XRO model for multiple climate modes. Anomaly Correlation Coefficient (ACC) skill maps for the XRO model, validated against the ORAS5 dataset. The panels show the forecast skill for the: **a.** North Pacific Meridional Mode (NPMM), **b.** South Pacific Meridional Mode (SPMM), **c.** Indian Ocean Basin-wide Mode (IOB), **d.** Indian Ocean Dipole (IOD), **e.** subtropical Indian Ocean Dipole (SIOD), and **f.** Tropical North Atlantic (TNA) mode. The horizontal axis indicates the target season, and the vertical axis represents the forecast lead time in months. Color shading denotes the ACC, with the black contour line indicating the 0.5 skill threshold.



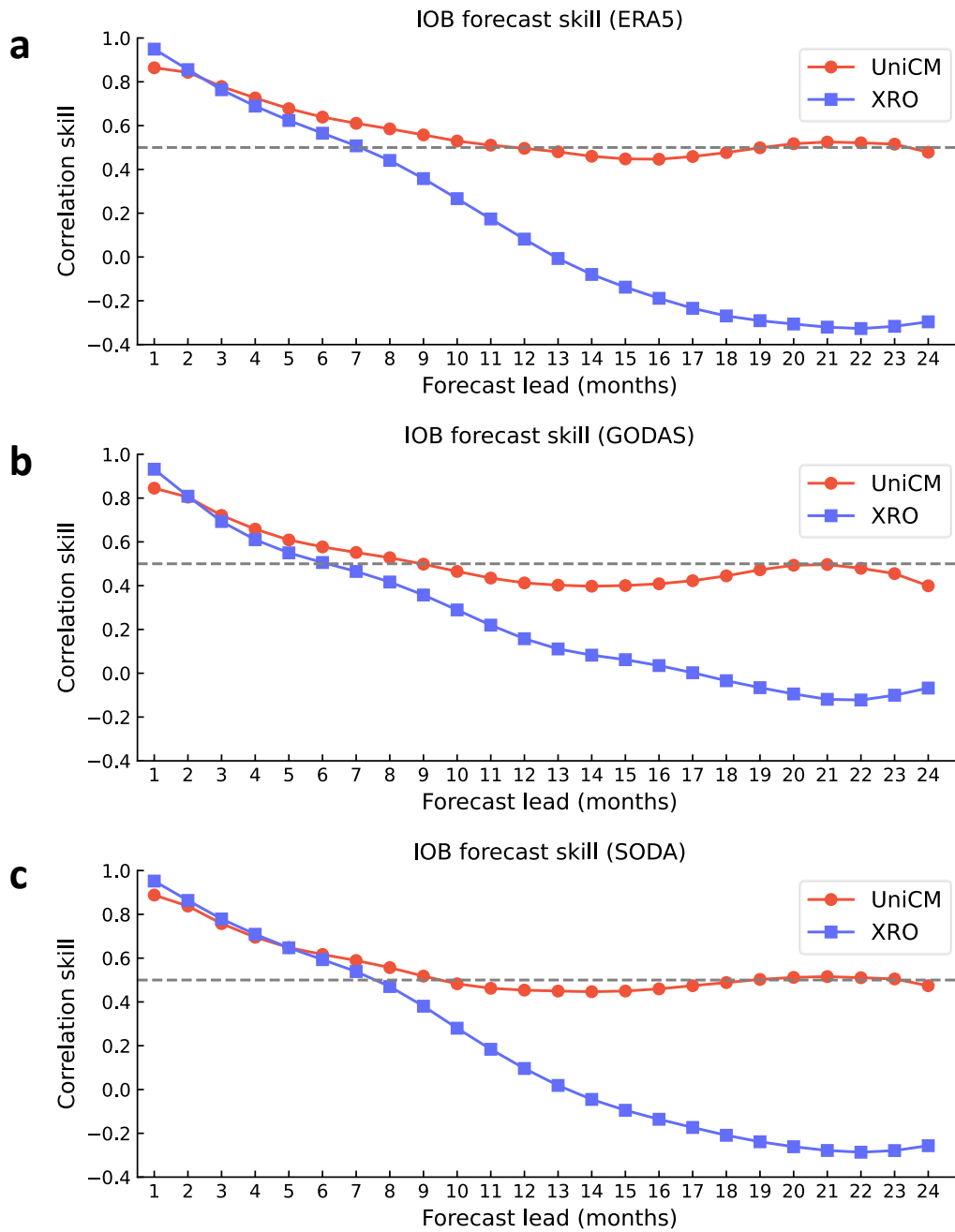
Supplementary Figure 14. Categorical event detection skill for ENSO. Comparison of categorical forecast performance between UniCM (solid lines) and the XRO baseline (dashed lines) across lead times from 1 to 24 months. **a.** Overall Accuracy, representing the fraction of correctly classified climate states (El Niño, La Niña, or Neutral). **b.** and **c.** Critical Success Index (CSI) for El Niño and La Niña events, respectively. The consistently higher CSI scores for UniCM across all leads demonstrate its enhanced sensitivity to extreme events and a significantly reduced false alarm rate compared to the more conservative XRO model.



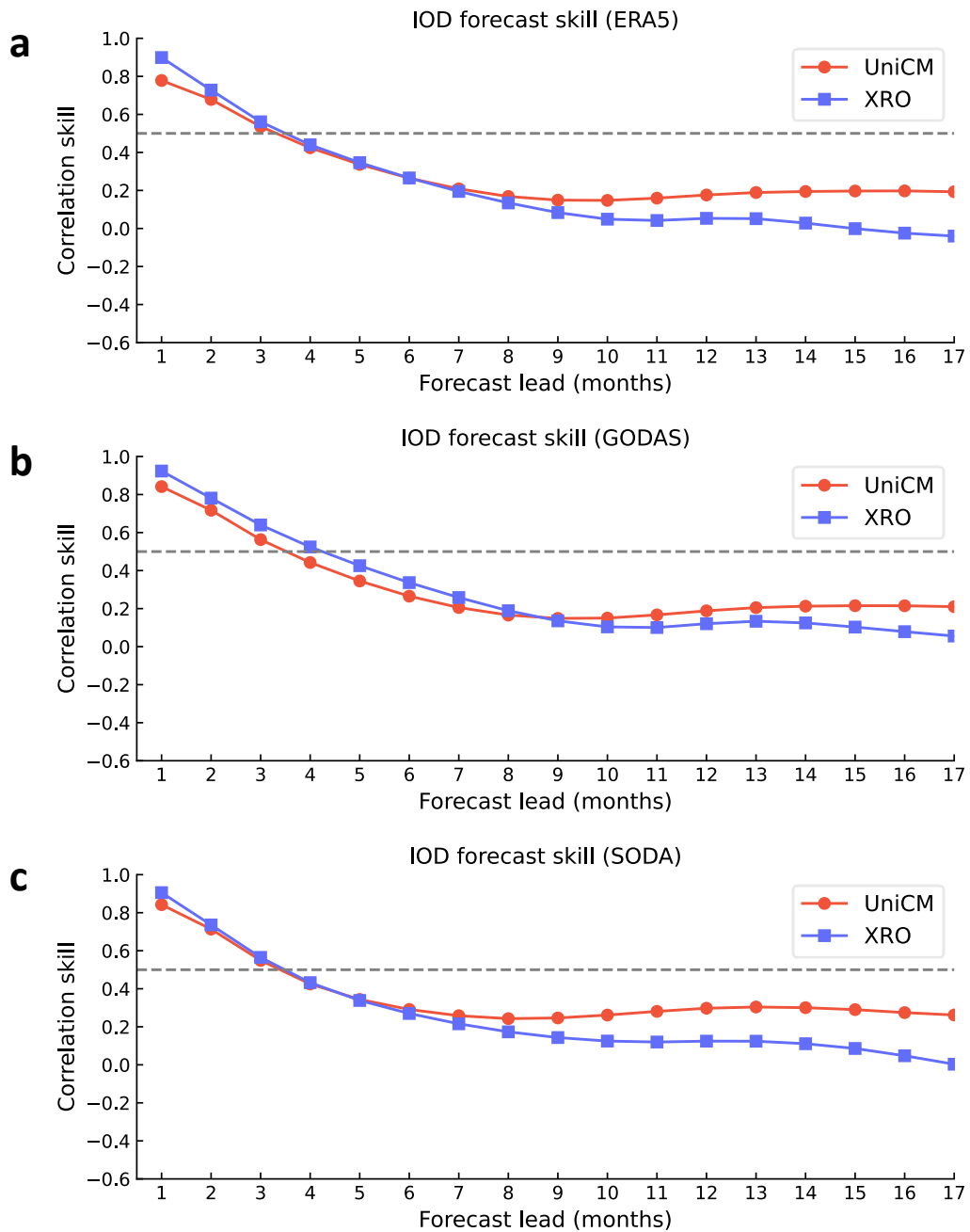
Supplementary Figure 15. Joint distribution and intensity accuracy of Niño3.4 SSTA forecasts. Hexbin density scatter plots of predicted versus observed Niño3.4 sea-surface temperature anomalies (SSTA) at **a.** 6-, **b.** 12-, and **c.** 18-month lead times. The red dashed line denotes perfect prediction ($y=x$), and the coefficient of determination (R^2) is reported in each panel. Reference lines at $\pm 0.5^\circ\text{C}$ and $\pm 1.5^\circ\text{C}$ denote thresholds for moderate and strong events. The concentration of high-density regions along the identity line confirms UniCM's ability to maintain strong intensity agreement even at long lead times.



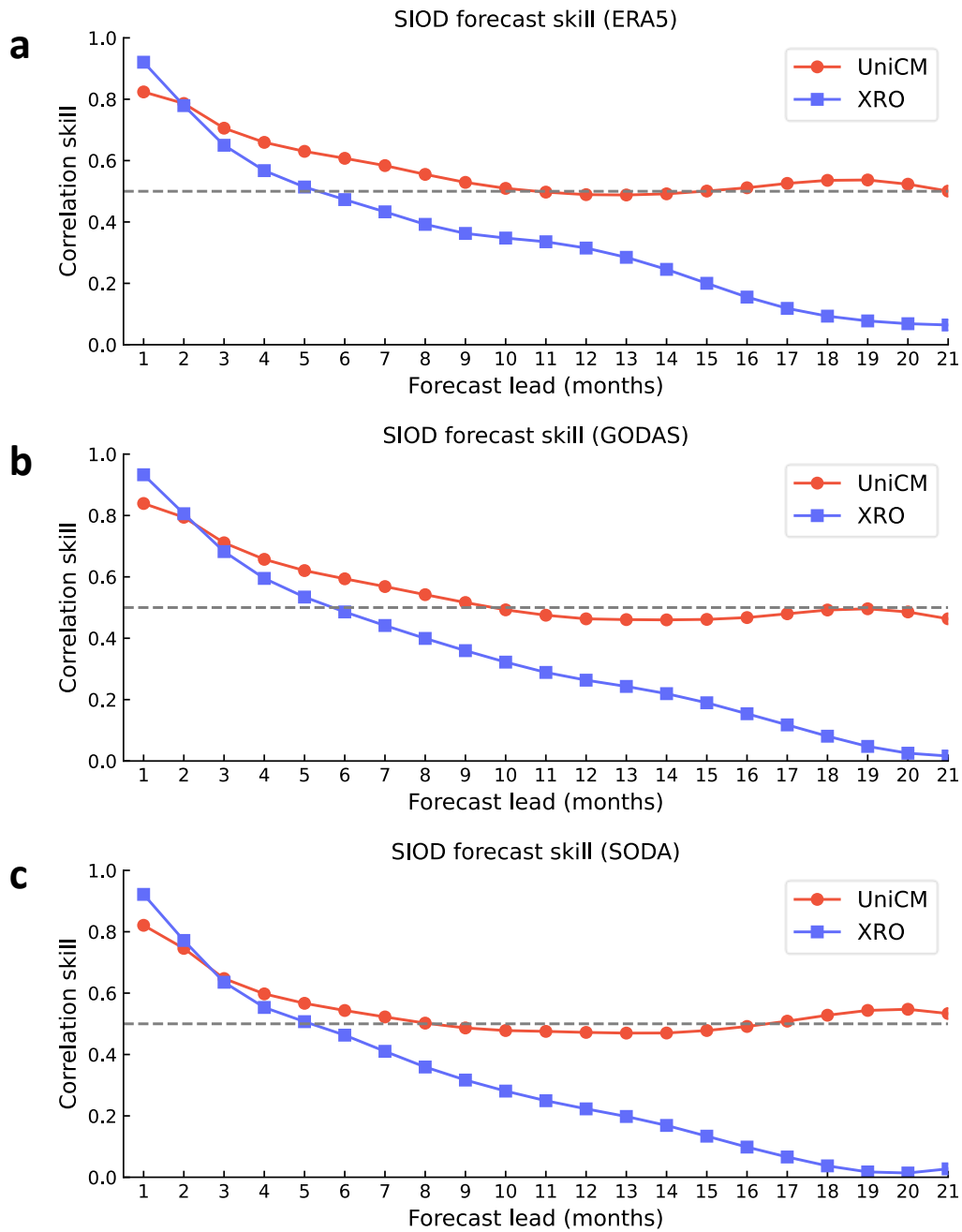
Supplementary Figure 16. Year-by-year ENSO evolution across lead times. **a.** Heatmaps of the Niño3.4 index versus forecast initialization date (rows) and lead time (columns), showing (left) observations at the target month, (middle) UniCM forecasts, and (right) XRO forecasts (24-month lead setting). This visualization highlights the temporal continuity of ENSO phase transitions and the relative ability of models to preserve event timing and amplitude across years.



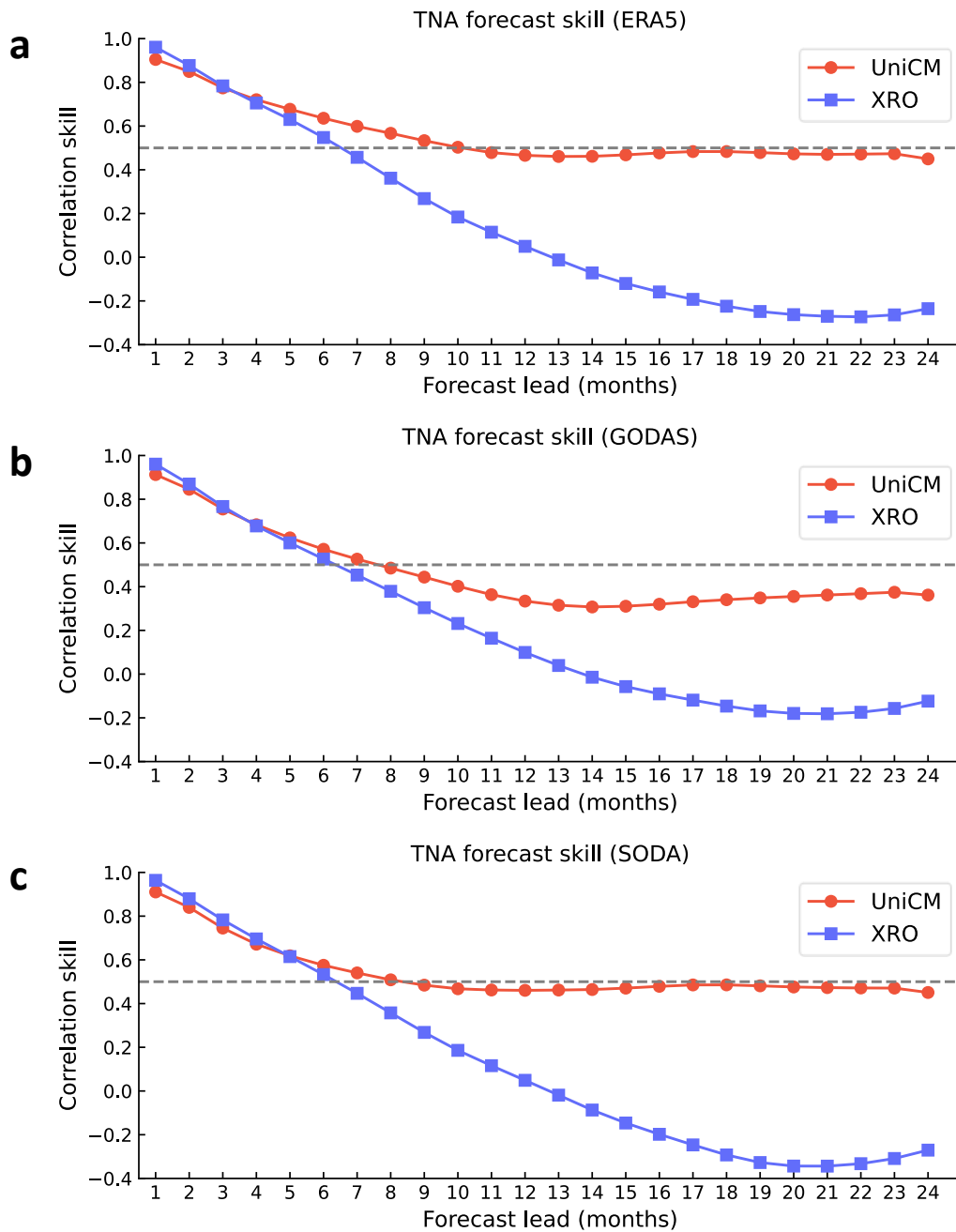
Supplementary Figure 17. Comparison of IOB forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for Indian Ocean Basin-wide Mode (IOB) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



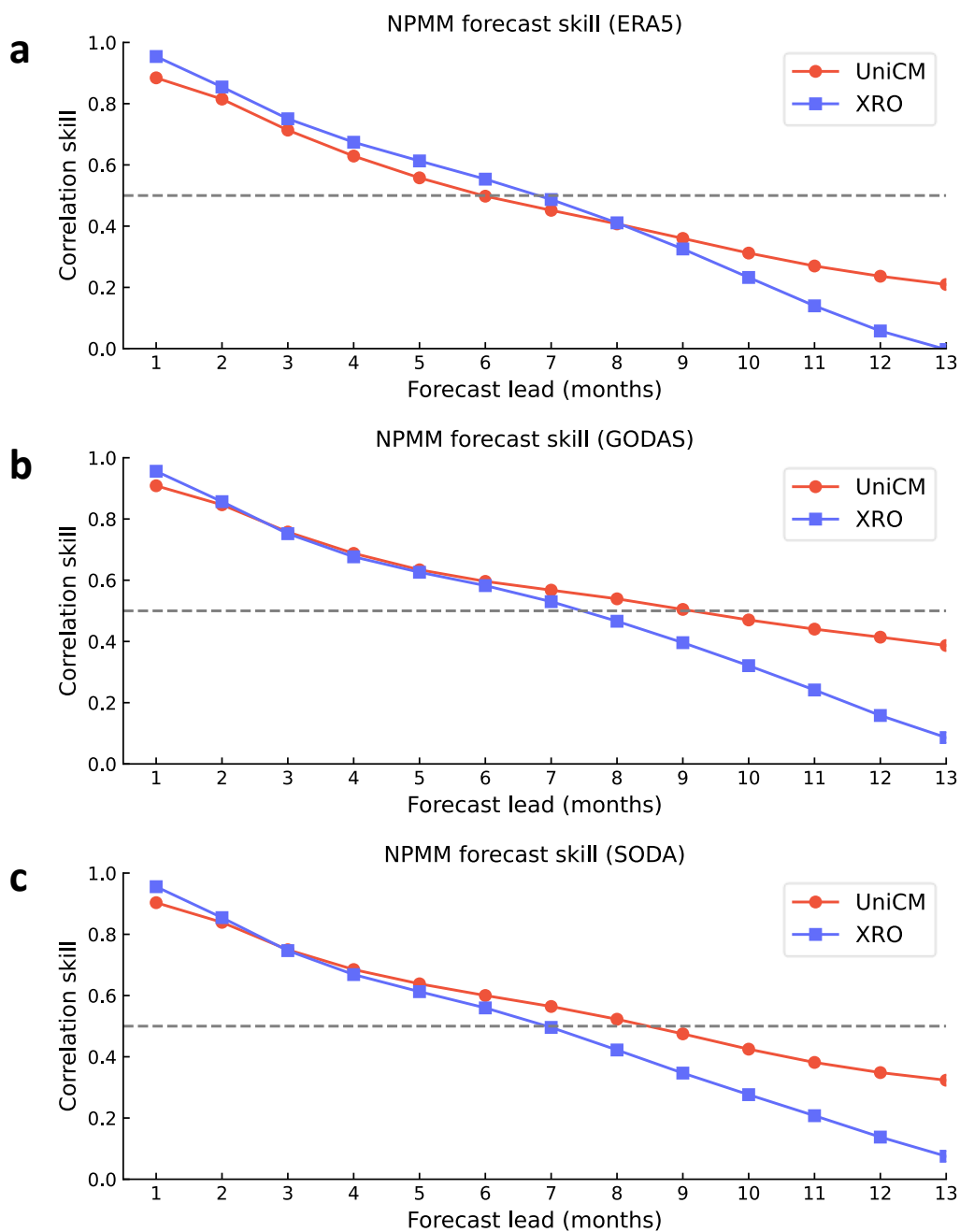
Supplementary Figure 18. Comparison of IOD forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for Indian Ocean Dipole (IOD) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



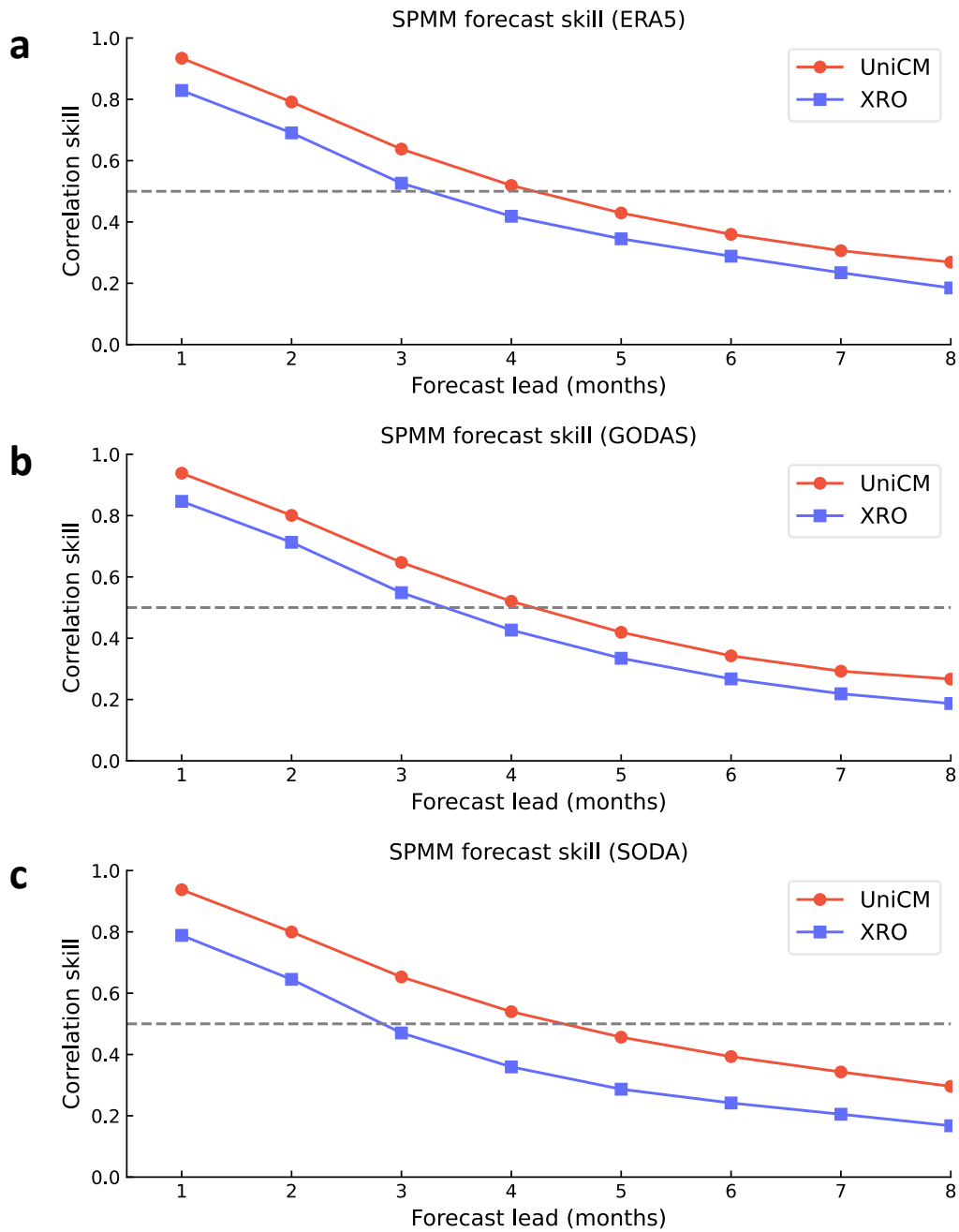
Supplementary Figure 19. Comparison of SIOD forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for Subtropical Indian Ocean Dipole (SIOD) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



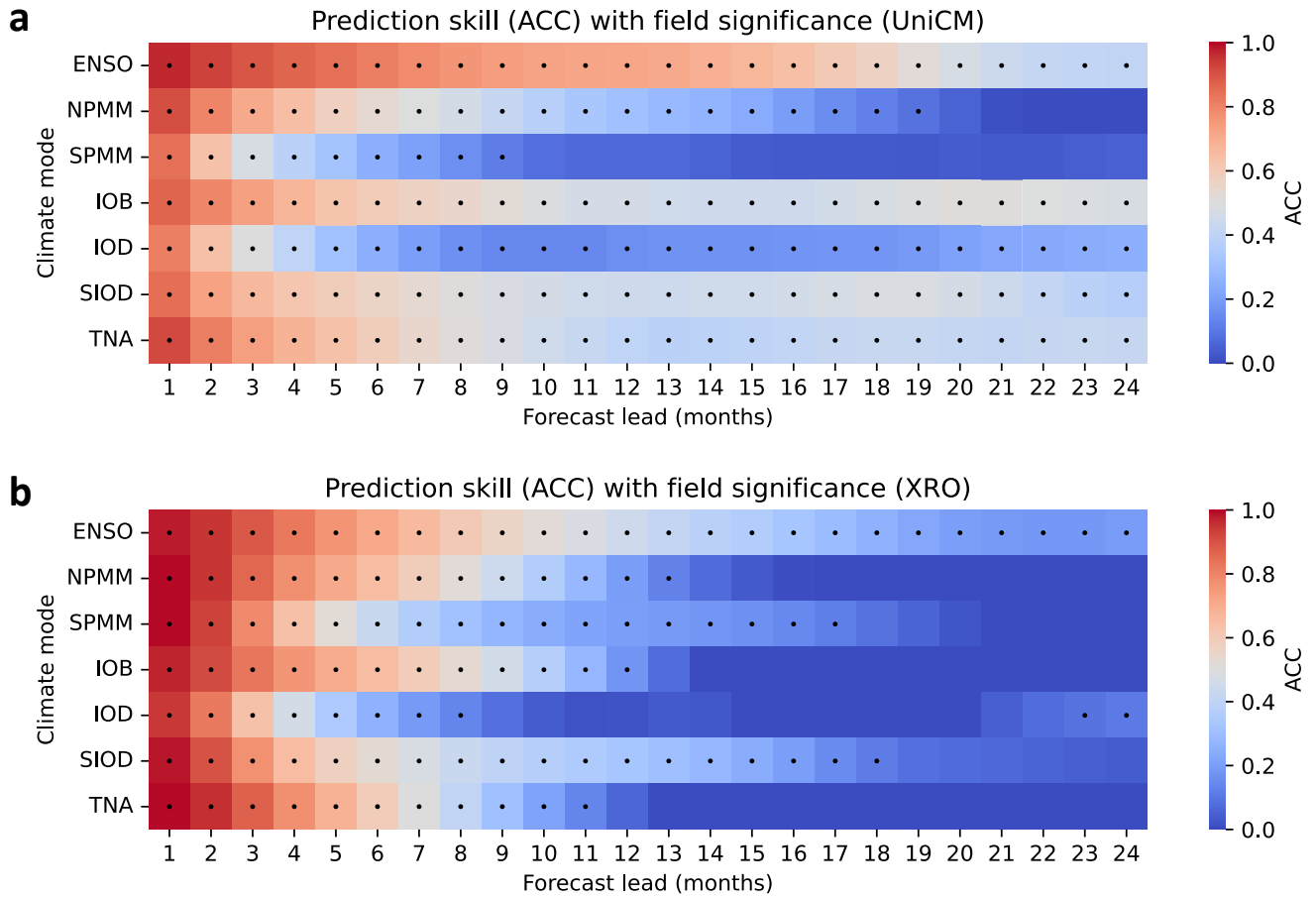
Supplementary Figure 20. Comparison of TNA forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for Tropical North Atlantic (TNA) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



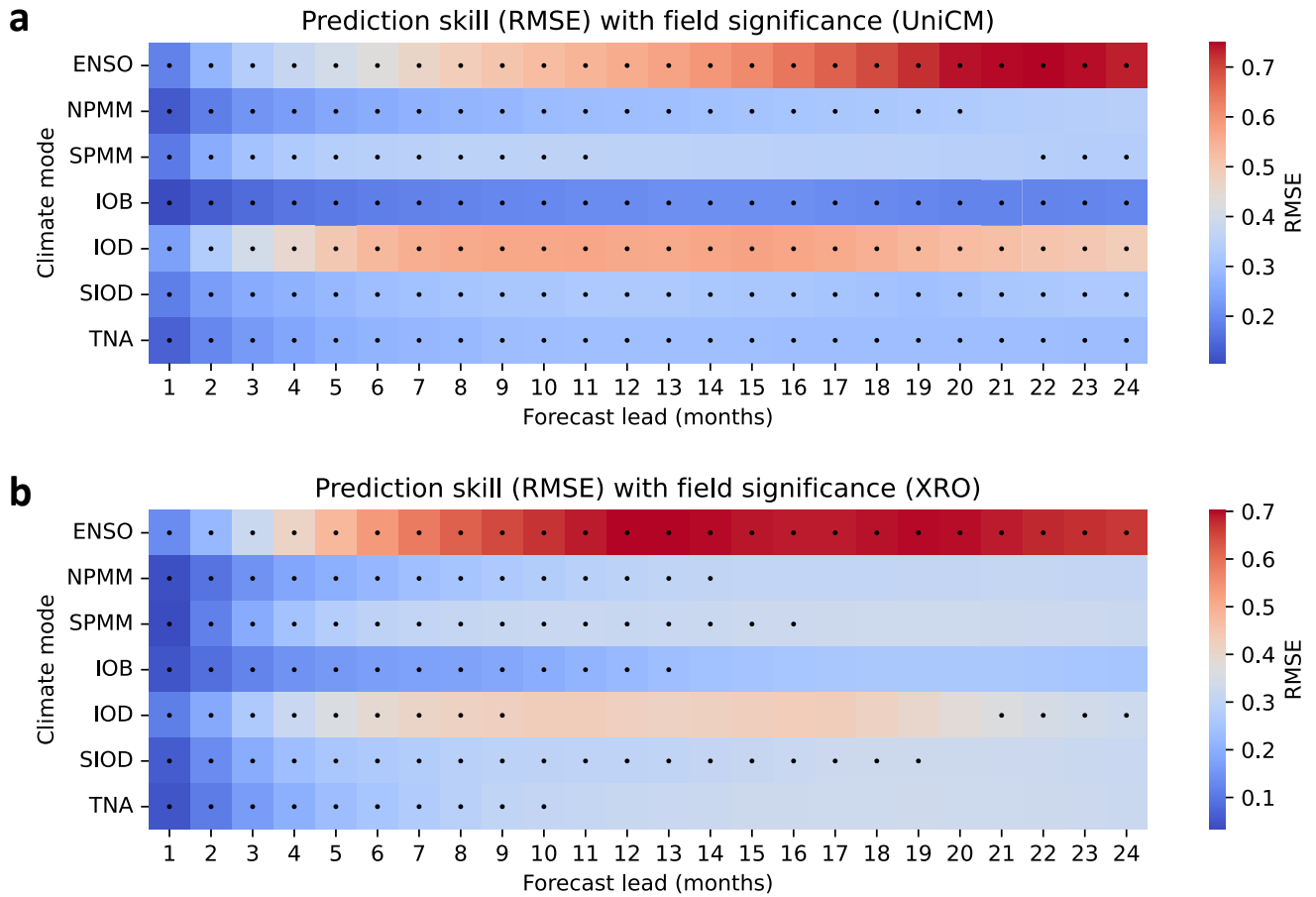
Supplementary Figure 21. Comparison of NPM forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for North Pacific Meridional Mode (NPM) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



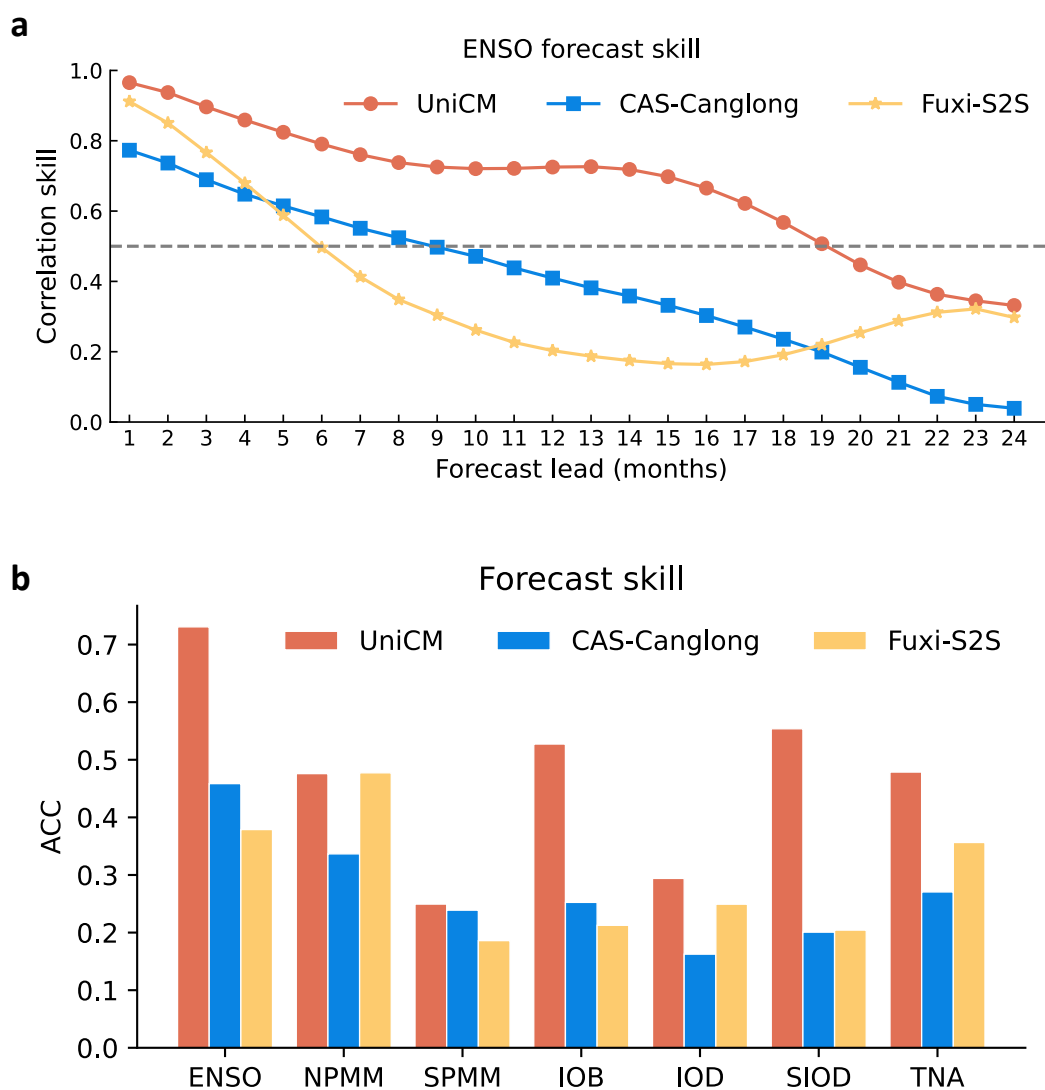
Supplementary Figure 22. Comparison of SPM forecast skill across validation datasets. Anomaly Correlation Coefficient (ACC) for South Pacific Meridional Mode (SPMM) forecasts from the UniCM (red) and XRO (blue) models, plotted against forecast lead time. The forecasts are validated against three different reanalysis datasets: **a.** ERA5, **b.** GODAS, and **c.** SODA v2.2.4. The dashed line indicates the 0.5 skill threshold.



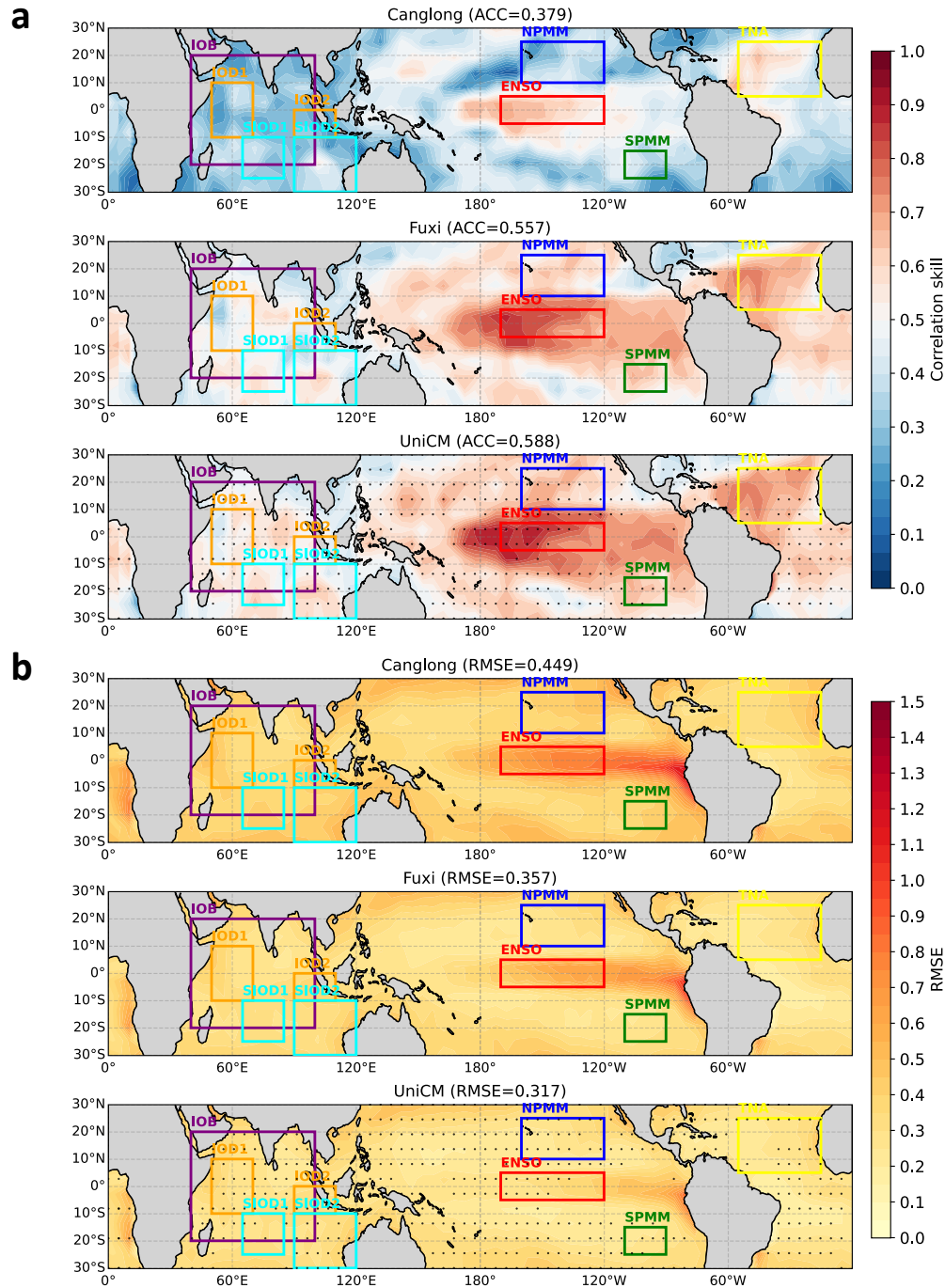
Supplementary Figure 23. Field-significance testing for correlation skill (ACC). Heatmaps of ACC for climate-mode index forecasts across lead times for **a.** UniCM and **b.** XRO. Dots mark statistically significant skill after field-significance testing with multiple-comparison control.



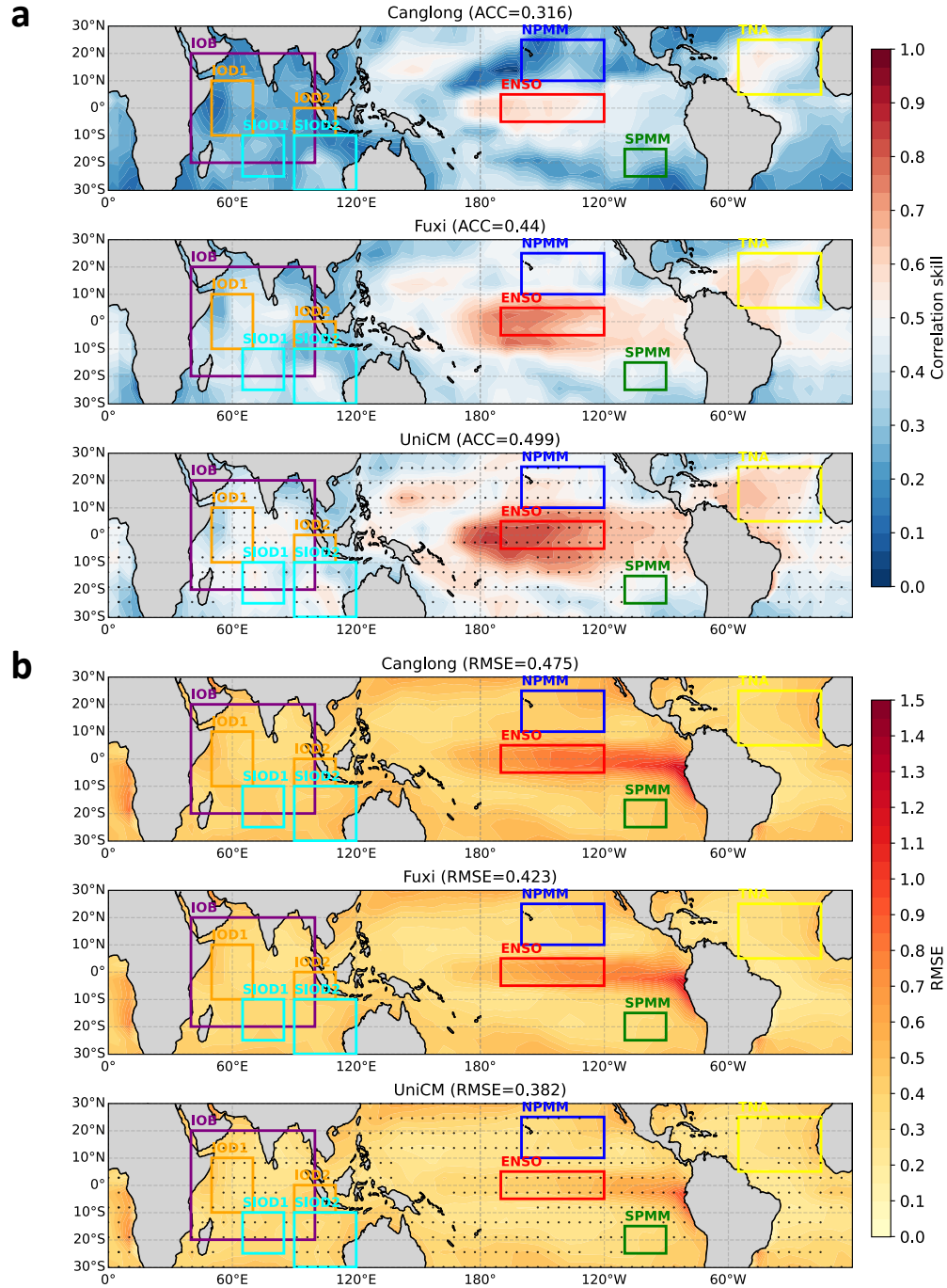
Supplementary Figure 24. Field-significance testing for RMSE. Heatmaps of RMSE for climate-mode index forecasts across lead times for **a.** UniCM and **b.** XRO. Dots mark statistically significant skill after field-significance testing with multiple-comparison control.



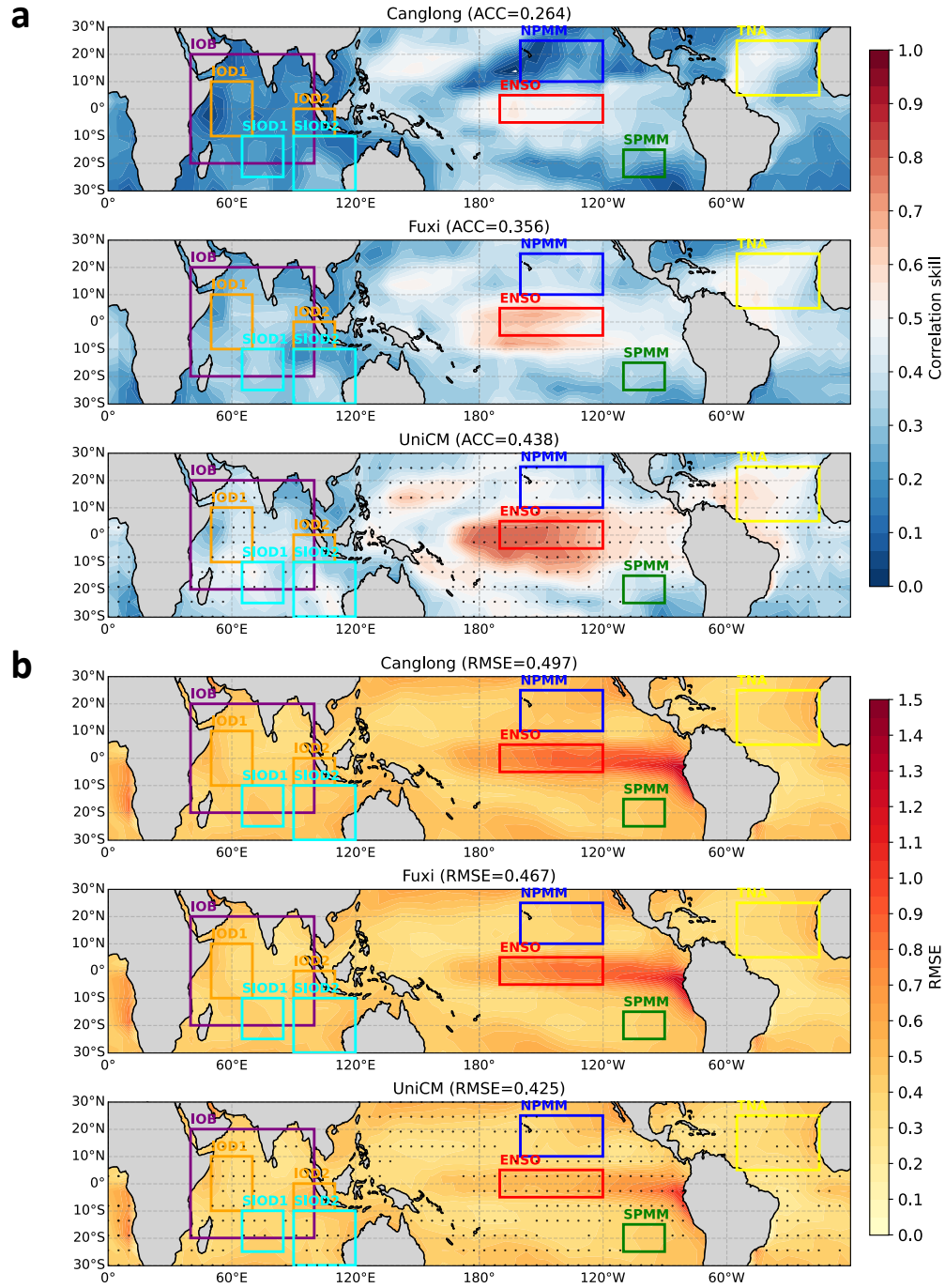
Supplementary Figure 25. Comparison with adapted field-to-field S2S foundation models. a. ENSO correlation skill (ACC) versus lead time for UniCM, CAS-Canglong, and FuXi-S2S. **b.** ACC at a representative long-lead horizon (24-month lead) across multiple climate modes, comparing UniCM with the two adapted S2S baselines.



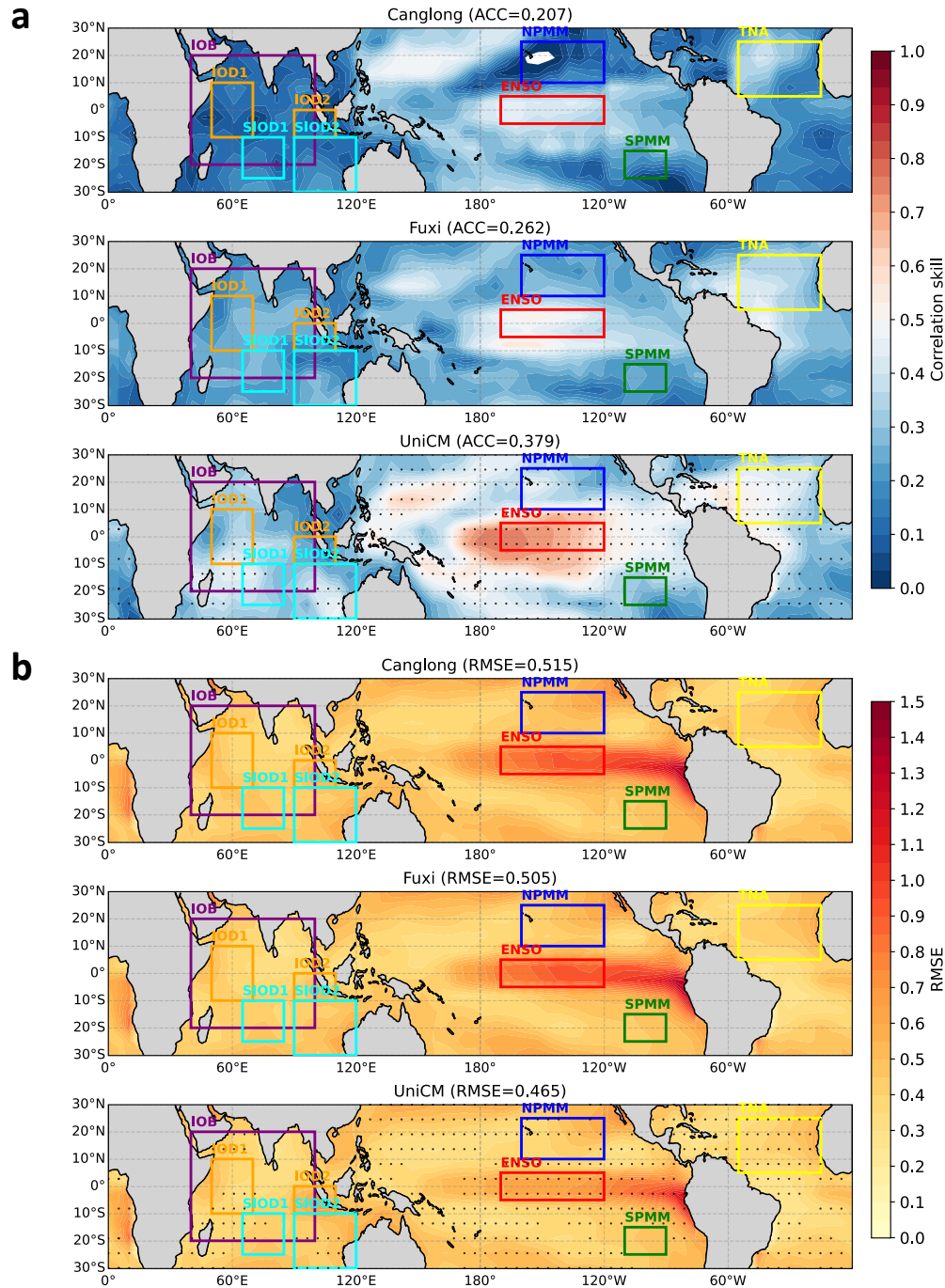
Supplementary Figure 26. Global spatial prediction skill at a 1-month lead. Spatial distributions of **a.** the Anomaly Correlation Coefficient (ACC) and **b.** Root Mean Square Error (RMSE) for SSTA forecasts. Performance is compared between UniCM and two state-of-the-art foundation models, FuXi-S2S and CAS-Canglong. Black stippling in the UniCM panel denotes grid points with statistically significant improvements ($p < 0.05$) over baselines. Significance is determined via a two-tailed paired t-test, with Effective Sample Size (ESS) and False Discovery Rate (FDR) adjustments to account for spatiotemporal dependencies.



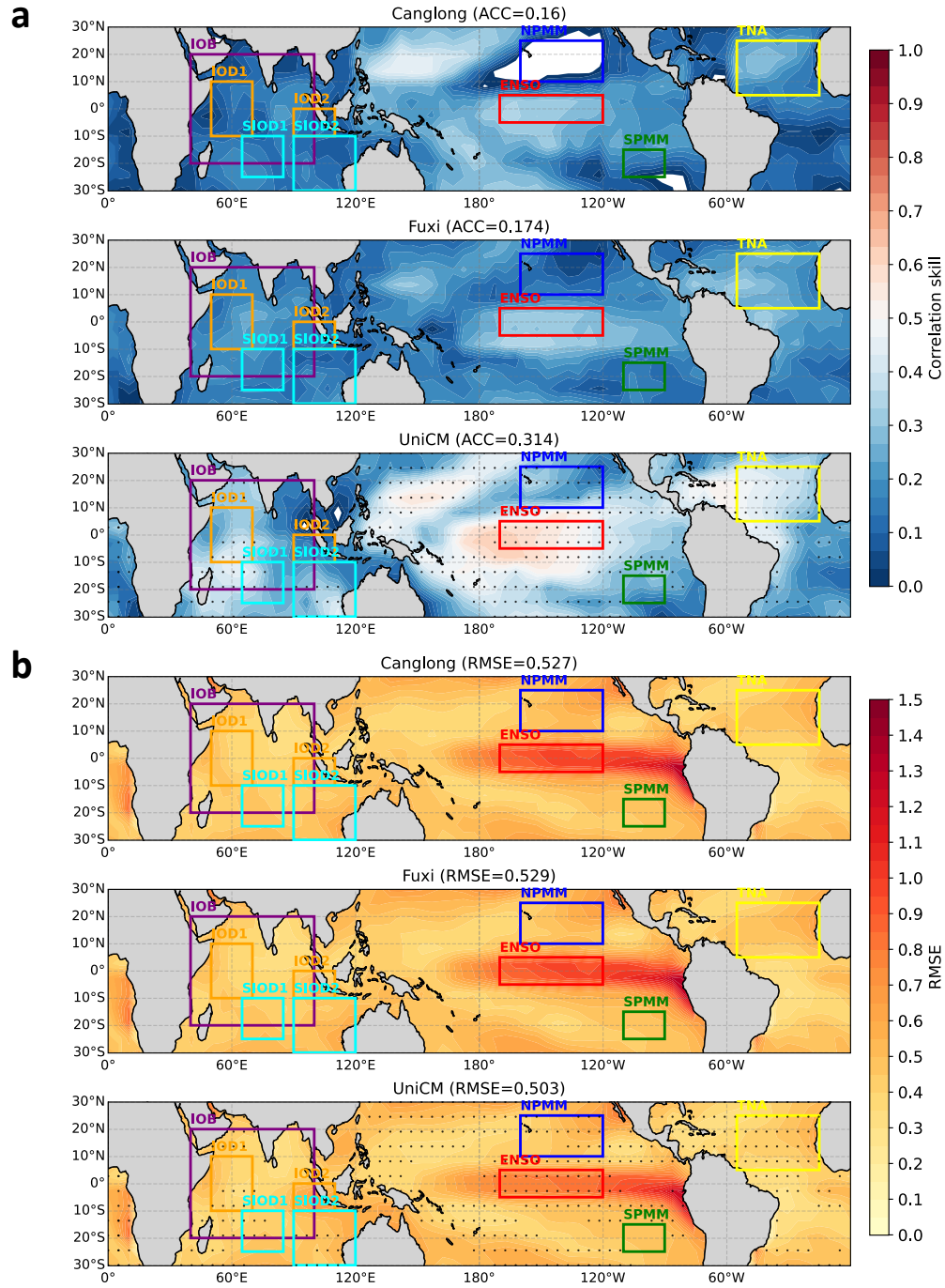
Supplementary Figure 27. Global spatial performance at 3-month lead. Comparison of spatial **a.** ACC and **b.** RMSE for SSTA forecasts at a 3-month lead time. Black stippling in the UniCM panel denotes grid points with statistically significant improvements ($p < 0.05$) over baselines. Significance is determined via a two-tailed paired t-test, with Effective Sample Size (ESS) and False Discovery Rate (FDR) adjustments to account for spatiotemporal dependencies.



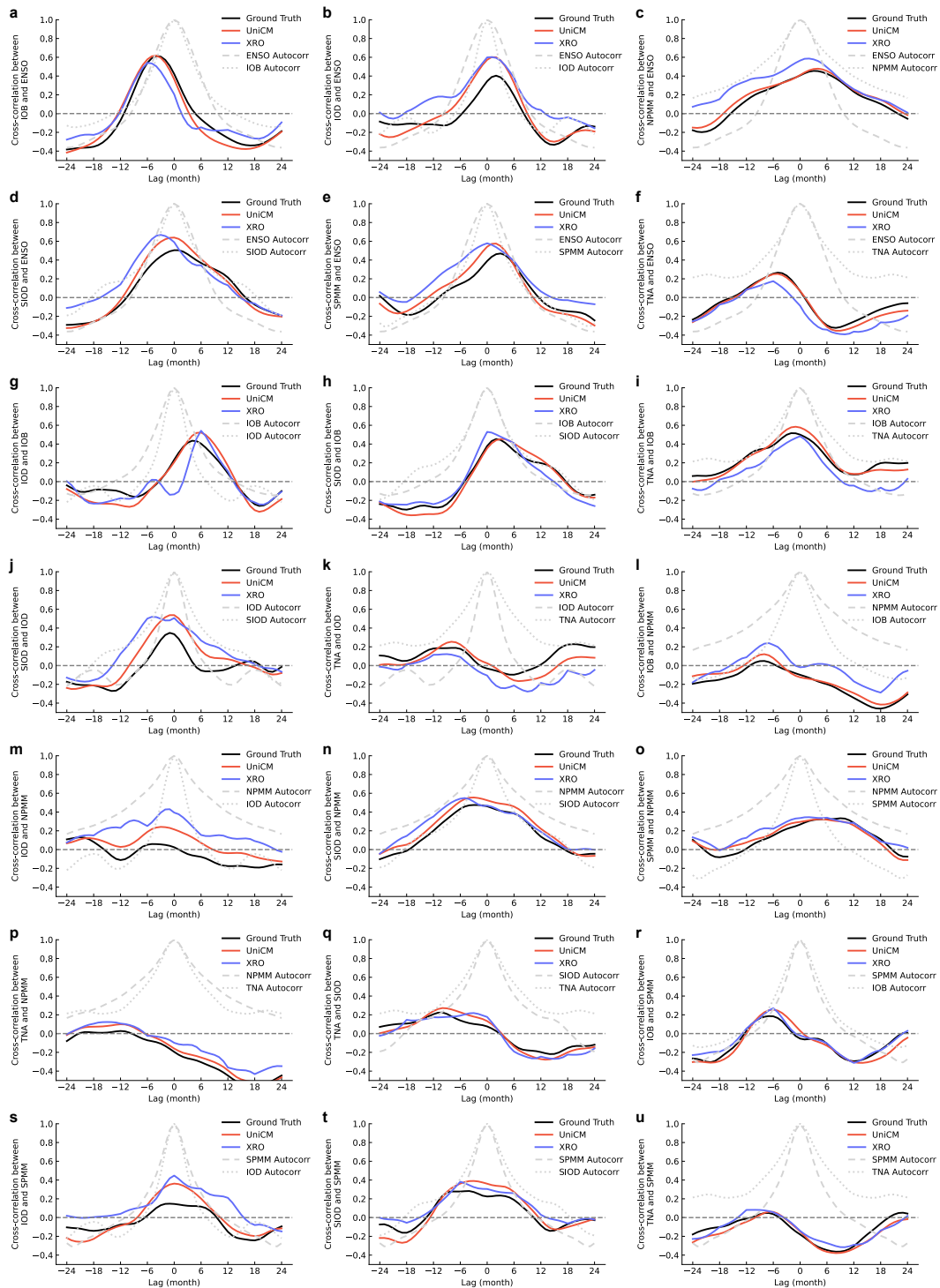
Supplementary Figure 28. Global spatial performance at 6-month lead. Comparison of spatial **a.** ACC and **b.** RMSE for SSTA forecasts at a 6-month lead time. UniCM maintains coherent spatial structures across major climate basins, whereas the field-to-field baselines begin to show fragmented skill patterns. Black stippling in the UniCM panel denotes grid points with statistically significant improvements ($p < 0.05$) over baselines. Significance is determined via a two-tailed paired t-test, with Effective Sample Size (ESS) and False Discovery Rate (FDR) adjustments to account for spatiotemporal dependencies.



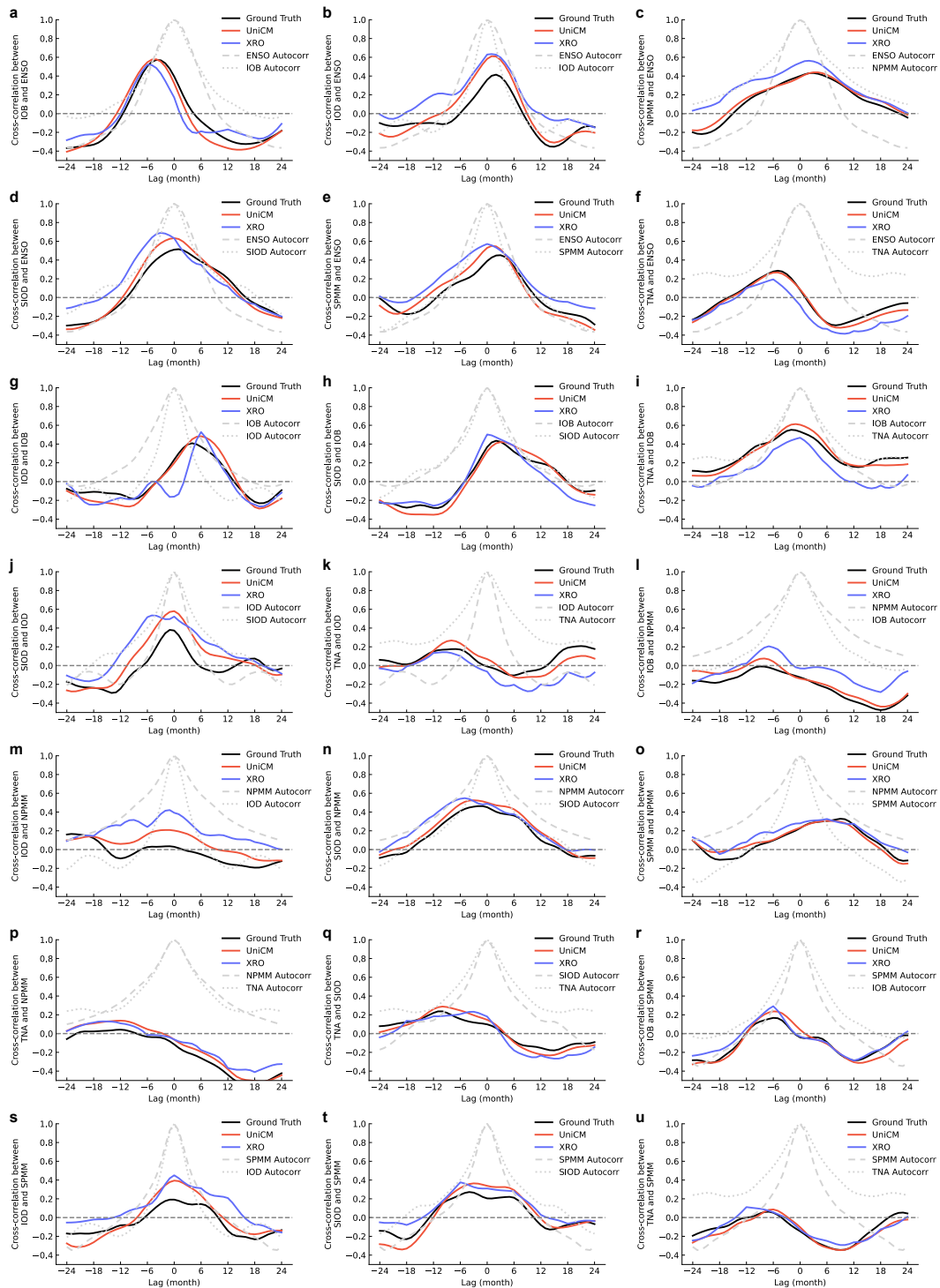
Supplementary Figure 29. Global spatial performance at 12-month lead. Benchmarking UniCM against S2S foundation models at a 12-month lead. **a.** ACC and **b.** RMSE comparisons show that UniCM retains significant predictive skill ($ACC > 0.4$) in the Niño 3.4 region and the North Atlantic, while the S2S models struggle to maintain signal consistency beyond seasonal timescales. Black stippling in the UniCM panel denotes grid points with statistically significant improvements ($p < 0.05$) over baselines. Significance is determined via a two-tailed paired t-test, with Effective Sample Size (ESS) and False Discovery Rate (FDR) adjustments to account for spatiotemporal dependencies.



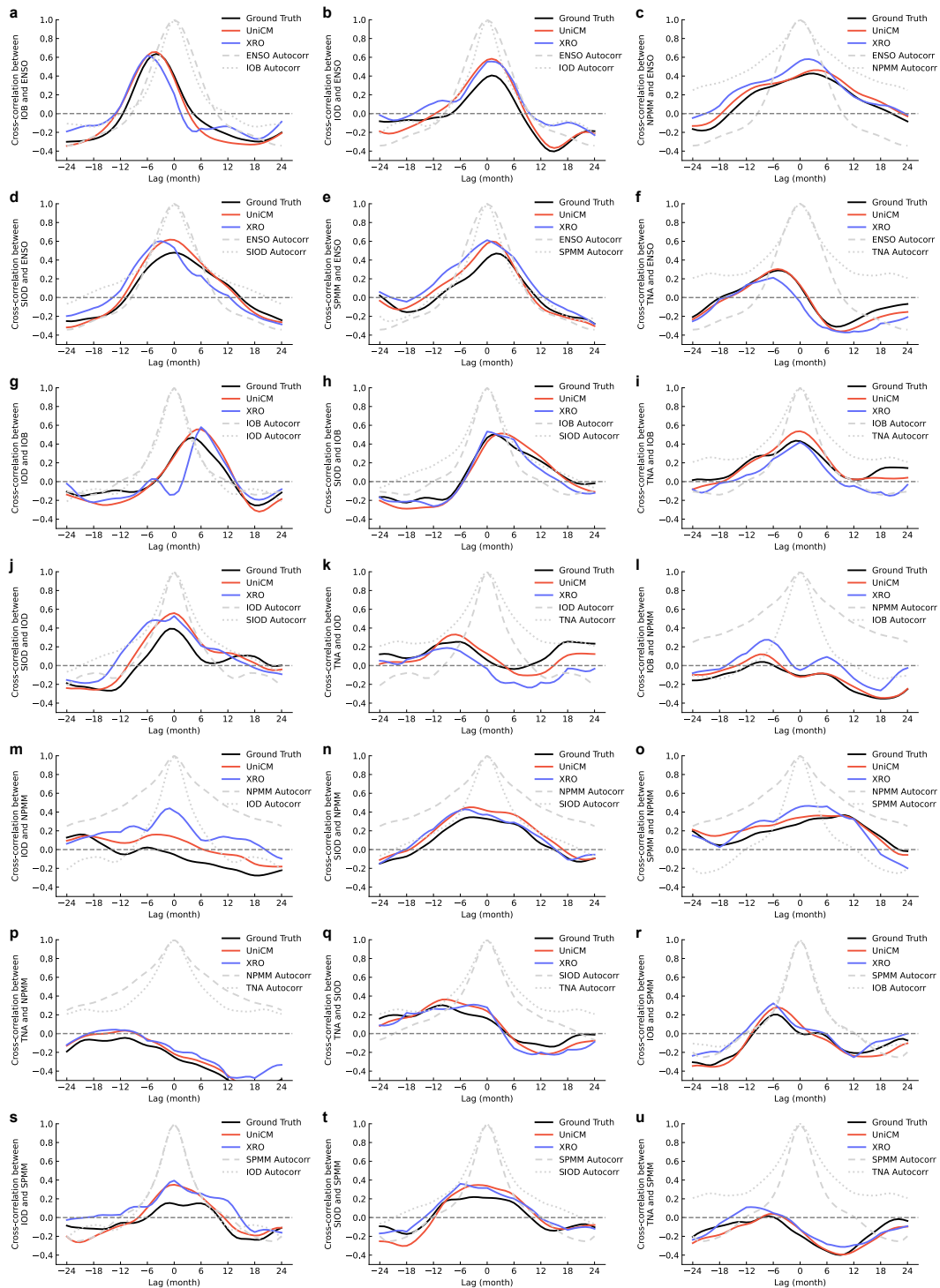
Supplementary Figure 30. Global spatial performance at 24-month lead. Performance comparison at the 24-month limit showing **a.** ACC and **b.** RMSE. Despite the inherent loss of predictability at long leads, UniCM continues to capture large-scale SSTA modes. Black stippling in the UniCM panel denotes grid points with statistically significant improvements ($p < 0.05$) over baselines. Significance is determined via a two-tailed paired t-test, with Effective Sample Size (ESS) and False Discovery Rate (FDR) adjustments to account for spatiotemporal dependencies.



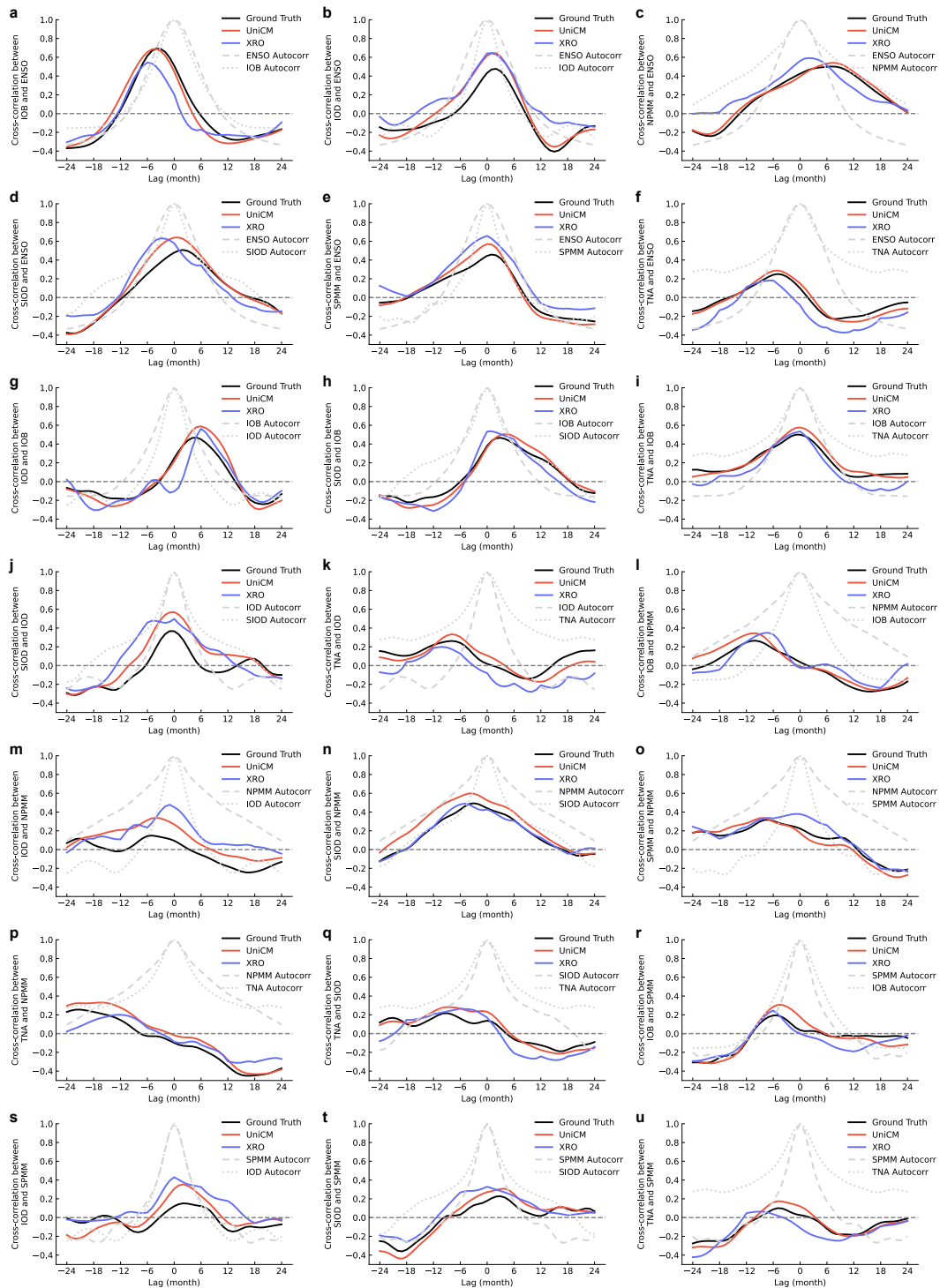
Supplementary Figure 31. Lagged cross-correlations among key climate indices. Lagged cross-correlations between pairs of climate modes for observations from the ORAS5 dataset (Ground Truth, black), UniCM forecasts (red), and XRO forecasts (blue). The x-axis indicates the lag in months, and the y-axis is the cross-correlation coefficient. Gray lines show the autocorrelation of the individual indices. The panels (a-u) assess the models' skill in capturing the strength and timing of various climate teleconnections.



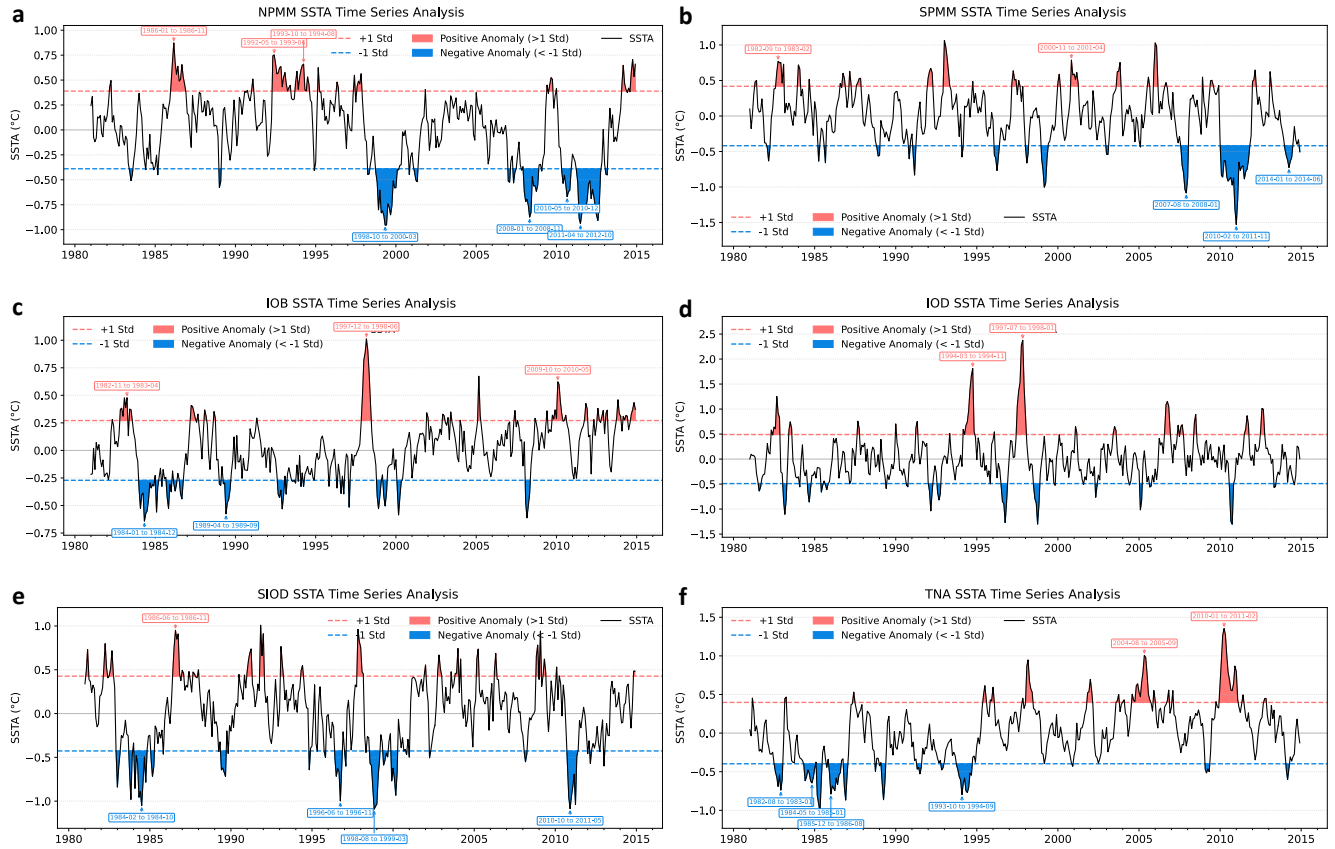
Supplementary Figure 32. Lagged cross-correlations among key climate indices. Lagged cross-correlations between pairs of climate modes for observations from the ERA5 dataset (Ground Truth, black), UniCM forecasts (red), and XRO forecasts (blue). The x-axis indicates the lag in months, and the y-axis is the cross-correlation coefficient. Gray lines show the autocorrelation of the individual indices. The panels (a-u) assess the models' skill in capturing the strength and timing of various climate teleconnections.



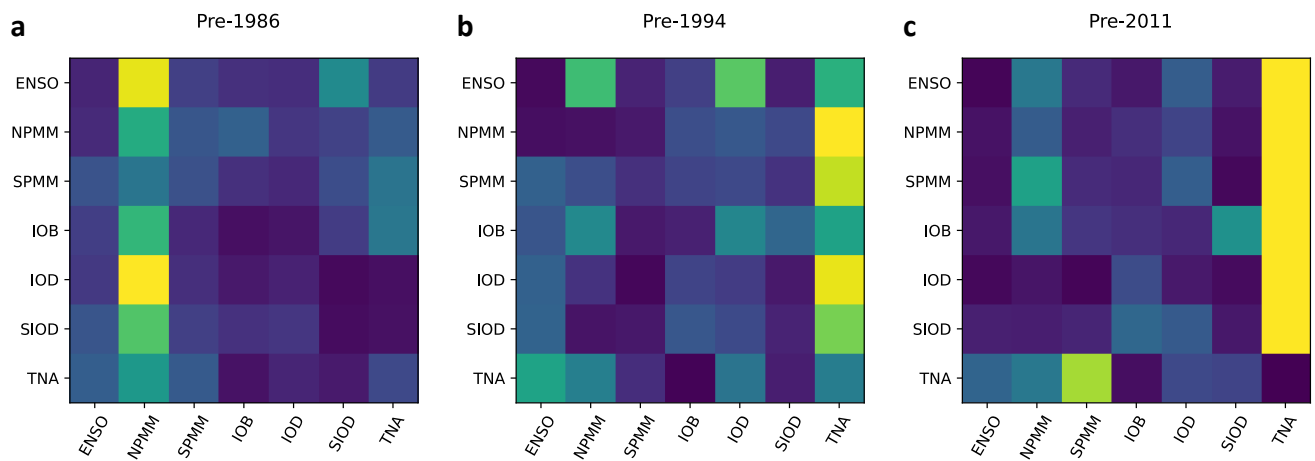
Supplementary Figure 33. Lagged cross-correlations among key climate indices. Lagged cross-correlations between pairs of climate modes for observations from the GODAS dataset (Ground Truth, black), UniCM forecasts (red), and XRO forecasts (blue). The x-axis indicates the lag in months, and the y-axis is the cross-correlation coefficient. Gray lines show the autocorrelation of the individual indices. The panels (a-u) assess the models' skill in capturing the strength and timing of various climate teleconnections.



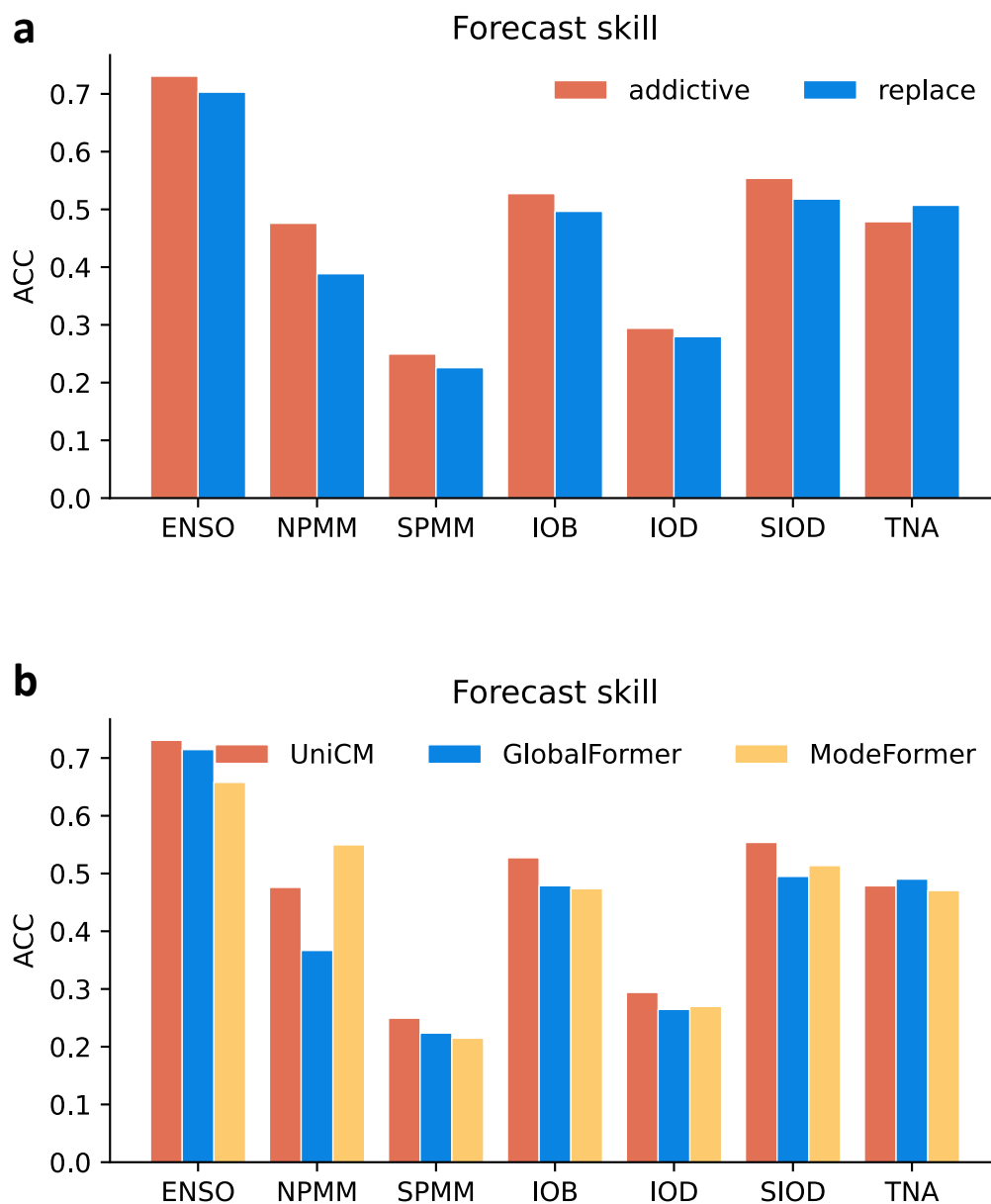
Supplementary Figure 34. Lagged cross-correlations among key climate indices. Lagged cross-correlations between pairs of climate modes for observations from the SODA v2.2.4 dataset (Ground Truth, black), UniCM forecasts (red), and XRO forecasts (blue). The x-axis indicates the lag in months, and the y-axis is the cross-correlation coefficient. Gray lines show the autocorrelation of the individual indices. The panels (a-u) assess the models' skill in capturing the strength and timing of various climate teleconnections.



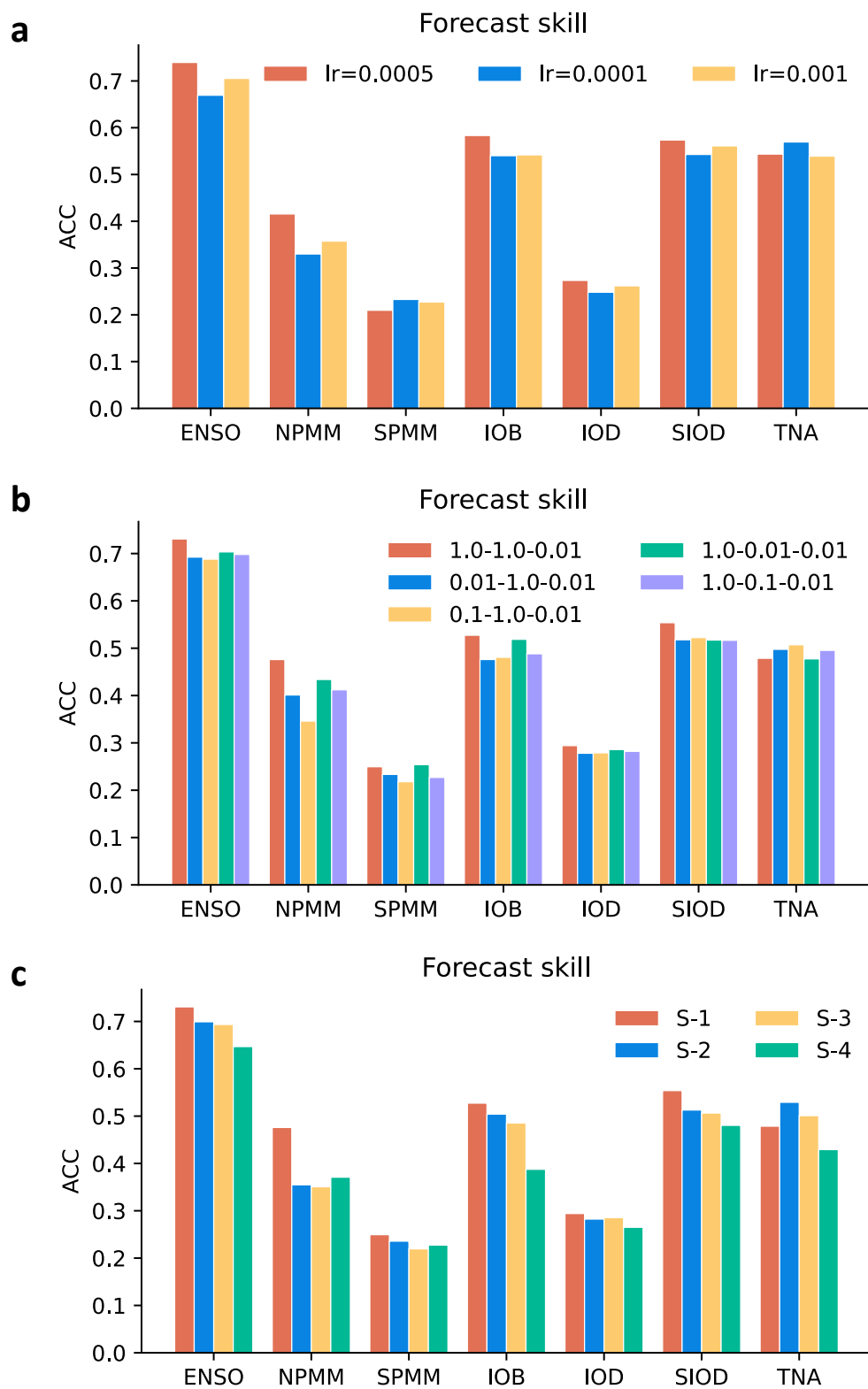
Supplementary Figure 35. SSTA time series and identified extreme events for non-ENSO modes. Time series analysis of SSTA indices for six climate modes: **a.** IOD, **b.** NPMM, **c.** SPMM, **d.** IOB, **e.** SIOD, and **f.** TNA. Shaded regions indicate significant events identified using a relative threshold of $|Index| > 1.0\sigma$ for at least five consecutive months. These identified periods serve as the basis for the inter-mode interaction analysis in Supplementary Fig. 36.



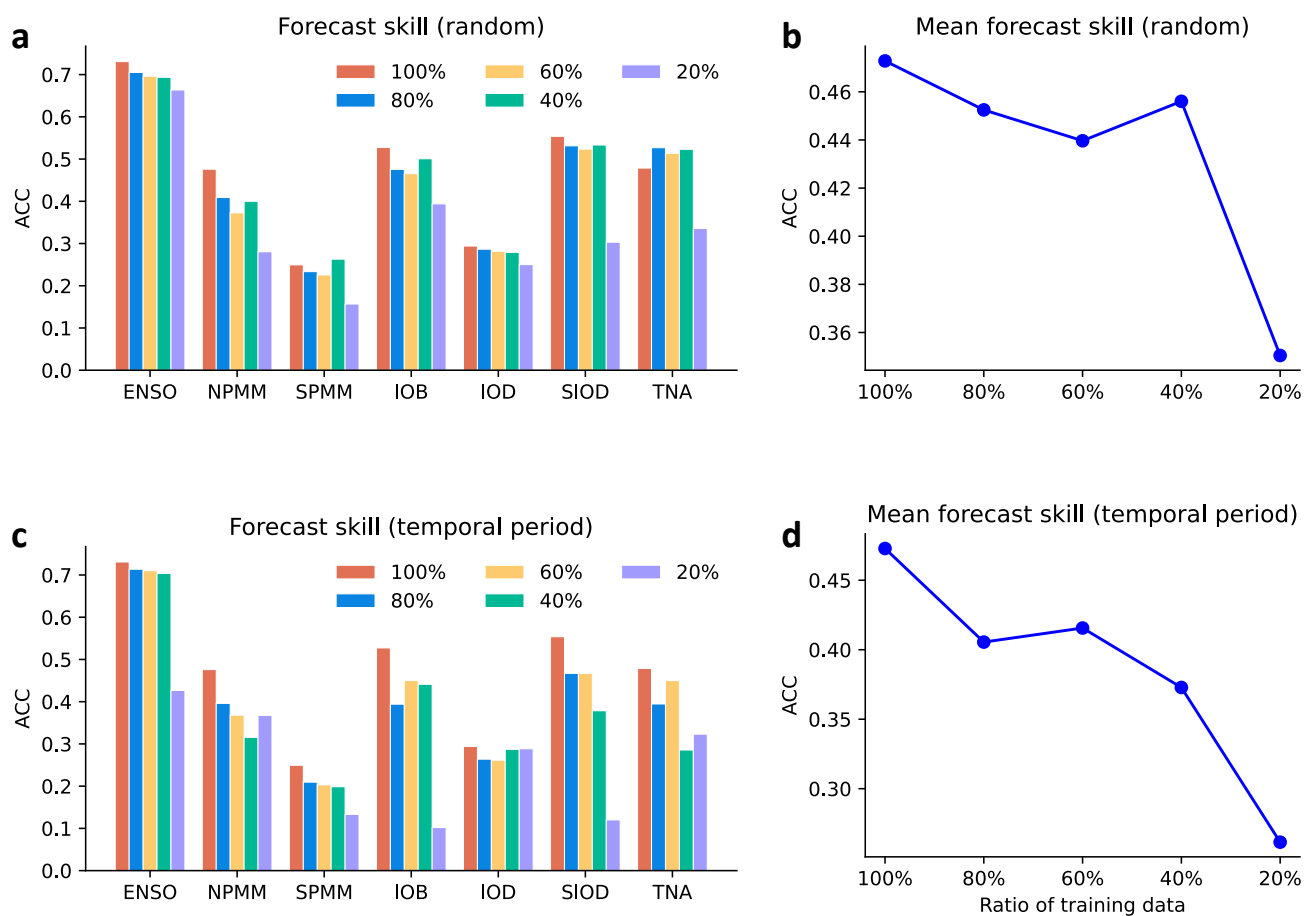
Supplementary Figure 36. Learned inter-mode interaction patterns during extreme events. Visualization of internal attention weights from the Modeformer branch during three landmark events: **a.** the 1986 NPMM transition, **b.** the 1994 TNA-basin linkage, and **c.** the 2011 multi-mode extreme.



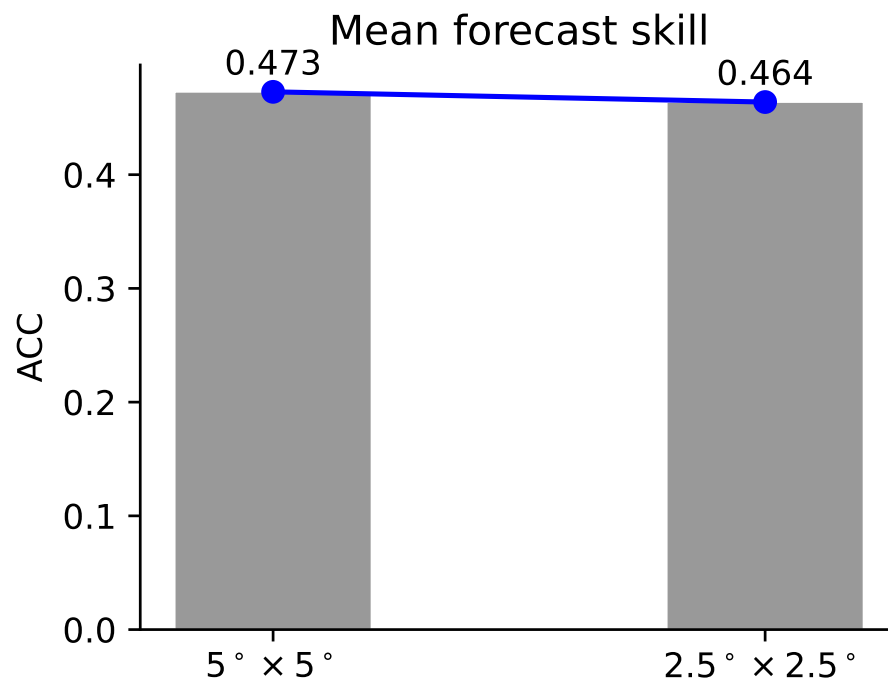
Supplementary Figure 37. Architectural ablation of the dual-branch design and prompt injection. **a.** Mean forecast skill (ACC) across climate modes comparing additive versus replacement prompt injection. **b.** Mean ACC comparing UniCM (dual-branch) against single-branch variants (GlobalFormer-only and ModeFormer-only).



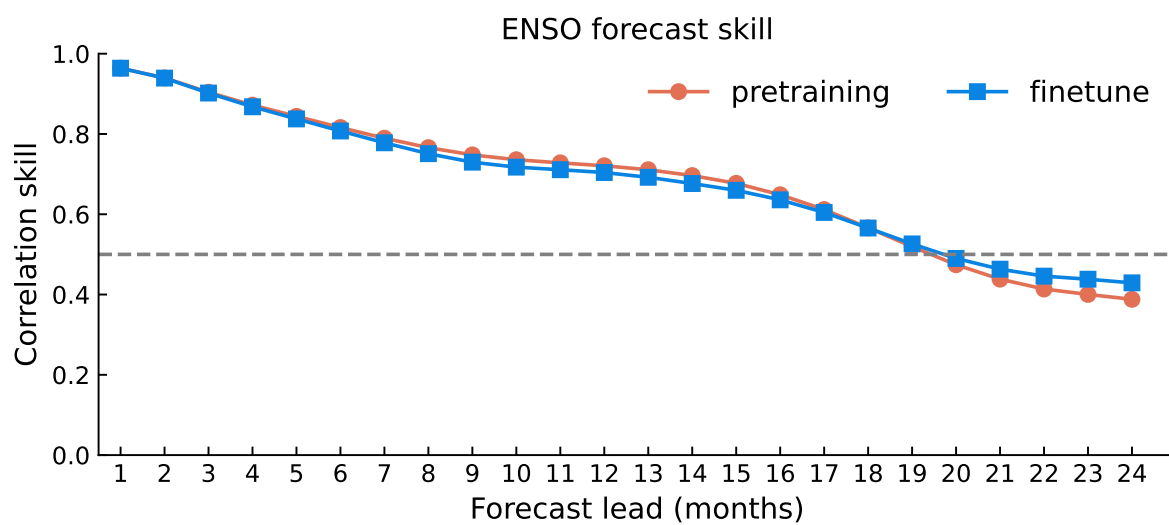
Supplementary Figure 38. Hyperparameter sensitivity of UniCM. Mean forecast skill (ACC) under **a.** different learning rates, **b.** different loss-weight settings, and **c.** different model sizes (S-1 to S-4).



Supplementary Figure 39. Training data requirements. Forecast skill (ACC) when training with reduced data. **a.** Per-mode ACC under random subsampling of the full training set. **b.** Mean ACC across all modes under random subsampling. **c.** Per-mode ACC using contiguous earlier training periods. **d.** Mean ACC across all modes using contiguous earlier training periods.



Supplementary Figure 40. Grid-resolution sensitivity. Comparison of mean ACC across climate modes for UniCM trained at 5° versus 2.5° spatial resolutions. The slight decrease in performance at 2.5° (from 0.473 to 0.464) suggests that the 5° grid is optimal for capturing the large-scale, low-frequency dynamics that dominate interannual variability.



Supplementary Figure 41. Effect of pre-training and fine-tuning on forecast skill. Analysis of the performance gap between a model trained only on CMIP6 simulations (pre-training) versus one fine-tuned on historical reanalysis data. The results highlight the model’s ability to adapt simulated physical laws to real-world observational data.

References

1. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year enso forecasts. *Nature* **573**, 568–572 (2019).
2. Lyu, P. *et al.* Resonet: Robust and explainable enso forecasts with hybrid convolution and transformer networks. *Adv. Atmospheric Sci.* **41**, 1289–1298 (2024).
3. Zhao, S. *et al.* Explainable el niño predictability from climate mode interactions. *Nature* **630**, 891–898 (2024).
4. Zhang, Z. *et al.* Enhancing the predictability limits of enso with physics-guided deep echo state networks. *npj Clim. Atmospheric Sci.* (2026).
5. Trenberth, K. E. The definition of el nino. *Bull. Am. Meteorol. Soc.* **78**, 2771–2778 (1997).
6. Wilks, D. S. *Statistical methods in the atmospheric sciences*, vol. 100 (Academic press, 2011).
7. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
8. Wilks, D. “the stippling shows statistically significant grid points”: How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Am. Meteorol. Soc.* **97**, 2263–2273 (2016).
9. Chen, L. *et al.* A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, 6425 (2024).
10. Wang, L. *et al.* Cas-canglong: A skillful 3d transformer model for sub-seasonal to seasonal global sea surface temperature prediction. *arXiv preprint arXiv:2409.05369* (2024).
11. Cai, W. *et al.* Pan-tropical climate interactions. *Science* **363**, eaav4236 (2019).
12. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. Enso as an integrating concept in earth science. *science* **314**, 1740–1745 (2006).
13. Alexander, M. A. *et al.* The atmospheric bridge: The influence of enso teleconnections on air–sea interaction over the global oceans. *J. climate* **15**, 2205–2231 (2002).
14. McKenna, S., Santoso, A., Gupta, A. S., Taschetto, A. S. & Cai, W. Indian ocean dipole in cmip5 and cmip6: characteristics, biases, and links to enso. *Sci. reports* **10**, 11500 (2020).
15. Zhang, H., Clement, A. & Di Nezio, P. The south pacific meridional mode: A mechanism for enso-like variability. *J. Clim.* **27**, 769–783 (2014).
16. Xie, S.-P. *et al.* Indian ocean capacitor effect on indo–western pacific climate during the summer following el niño. *J. climate* **22**, 730–747 (2009).
17. Saji, N., Goswami, B. N., Vinayachandran, P. & Yamagata, T. A dipole mode in the tropical indian ocean. *Nature* **401**, 360–363 (1999).
18. Jo, H.-S. *et al.* Southern indian ocean dipole as a trigger for central pacific el niño since the 2000s. *Nat. Commun.* **13**, 6965 (2022).
19. Enfield, D. B., Mestas-Núñez, A. M., Mayer, D. A. & Cid-Serrano, L. How ubiquitous is the dipole relationship in tropical atlantic sea surface temperatures? *J. Geophys. Res. Ocean.* **104**, 7841–7848 (1999).