Using Deep Learning to Identify Initial Error Sensitivity for Interpretable ENSO Forecasts

Kinya Toride,[a,b] Matthew Newman,[a] Andrew Hoell,[a] Antonietta Capotondi,[a,b] Jakob Schlör,[c] Dillon Amaya,[a]

[a] *Physical Sciences Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado*

[b] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*

[c] *Machine Learning in Climate Science, University of Tübingen, Tübingen, Germany*

Corresponding author: Kinya Toride, kinya.toride@noaa.gov

ABSTRACT

We introduce an interpretable-by-design method, optimized model-analog, that integrates deep learning with model-analog forecasting, a straightforward yet effective approach that generates forecasts from similar initial climate states in a repository of model simulations. This hybrid framework employs a convolutional neural network to estimate state-dependent weights to identify initial analog states that lead to shadowing target trajectories. The advantage of our method lies in its inherent interpretability, offering insights into initial-error-sensitive regions through estimated weights and the ability to trace the physically-based evolution of the system through analog forecasting. We evaluate our approach using the Community Earth System Model Version 2 Large Ensemble to forecast the El Niño–Southern Oscillation (ENSO) on a seasonal-to-annual time scale. Results show a 10% improvement in forecasting equatorial Pacific sea surface temperature anomalies at 9–12 months leads compared to the original (unweighted) model-analog technique. Furthermore, our model demonstrates improvements in boreal winter and spring initialization when evaluated against a reanalysis dataset. Our approach reveals state-dependent regional sensitivity linked to various seasonally varying physical processes, including the Pacific Meridional Modes, equatorial recharge oscillator, and stochastic wind forcing. Additionally, disparities emerge in the sensitivity associated with El Niño versus La Niña events. El Niño forecasts are more sensitive to initial uncertainty in tropical Pacific sea surface temperatures, while La Niña forecasts are more sensitive to initial uncertainty in tropical Pacific zonal wind stress. This approach has broad implications for forecasting diverse climate phenomena, including regional temperature and precipitation, which are challenging for the original model-analog approach.

SIGNIFICANCE STATEMENT

The purpose of this study is to demonstrate a skillful and interpretable approach for forecasting the El Niño–Southern Oscillation by combining deep learning and a simple analog forecasting method. A convolutional neural network is used to find critical areas for picking analog members. This is important because it is challenging to explain the decision-making processes of recent deep-learning approaches. The developed approach can be applied to various climate predictions.

1. Introduction

The prediction of climate variability over seasonal to interannual time scales greatly depends on the quality of El Niño–Southern Oscillation (ENSO) forecasts. The magnitude and pattern of tropical sea surface temperature (SST) anomalies associated with ENSO influence global climate through atmospheric teleconnections primarily driven by the Walker and Hadley circulations and stationary Rossby wave trains (Alexander et al. 2002; Hoell and Funk 2013; Capotondi et al. 2015; Taschetto et al. 2020). However, state-of-the-art atmosphere-ocean coupled models do not exhibit a substantial improvement over simpler linear models in predicting ENSO (Newman and Sardeshmukh 2017; Shin et al. 2021; Risbey et al. 2021).

With recent progress in deep learning, several studies have applied various neural networks to ENSO prediction (Ham et al. 2019; Petersik and Dijkstra 2020; Cachay et al. 2021; Chen et al. 2021; Ham et al. 2021; Zhou and Zhang 2023). Considering the data-intensive nature of deep learning, long-term climate simulations from multiple models are often leveraged to capture nonlinear dynamics of ENSO and mitigate model-specific biases. While these data-driven models exhibit promising performance, interpreting their decision-making processes poses a challenge due to the large number of hidden parameters. The interpretability of prediction models is crucial since models with better interpretability can enhance scientific understanding of physical processes, which can, in turn, improve prediction skill. Explainable artificial intelligence (XAI) is frequently used to elucidate neural network models in a post-hoc manner (e.g., Shin et al. 2022). However, different XAI techniques may yield different explanations for the same deep learning model (Mamalakis et al. 2022), and it remains challenging to explain complex models despite their superior accuracy in general.

Analog forecasting is a simpler method which makes predictions based on similar states that occurred in the past, assuming they follow the attractor of the dynamical system (Lorenz 1969a). While the sample size of historical records is too small to find good analogs for most climate-scale applications (Van den Dool 1989), simulated climate data allow for drawing "model-analogs" (Ding et al. 2018) from thousands of years of data. Because analog forecasting circumvents issues with initialization shock (Mulholland et al. 2015) by initializing directly in the model space, this method provides comparable skill to that of

coupled atmosphere-ocean models in forecasting seasonal tropical SST (Ding et al. 2018, 2019).

However, despite advances, finding reliable analogs within the chaotic climate system remains challenging due to both the limited sample size, even with thousands of years, and model imperfections leading to disparities between the model attractor and nature's attractor. In chaotic systems, even tiny disturbances in initial states can lead to significantly divergent trajectories (Lorenz 1963, 1969b). Fig. 1b illustrates this issue, showing that a few model-analogs, selected based only on minimal mean-square differences across the tropics, can evolve into the opposite phase of ENSO within 12 months.

Alternatively, there may exist trajectories with slightly different initial conditions that remain closer to the true trajectory over some period of time (Grebogi et al. 1990; Judd et al. 2004). Identifying these shadowing trajectories involves considering the sensitivity to initial conditions, with certain regions being more prone to initial errors while others are relatively insensitive (Errico 1997; Barsugli and Sardeshmukh 2002). For instance, the North Pacific Meridional Mode (NPMM) serves as one of key ENSO precursors (Chiang and Vimont 2004; Amaya 2019), driving the search for analogs that closely match over the NPMM region. Essentially, we aim to assign higher weights to initial-error-sensitive regions, thereby optimizing the selection of model-analogs so that their subsequent trajectories will more closely shadow the true trajectory.

In this study, we introduce a deep learning method (specifically, a convolutional neural network) that predicts state-dependent weights for selecting "optimized model-analogs". The combination of analog forecasting and machine learning has been investigated by several studies. Chattopadhyay et al. (2020) clustered surface temperature patterns into five groups and used a capsule neural network to predict the cluster indices based on states 1–5 days prior. Rader and Barnes (2023) introduced the idea of training a neural network to learn weights of a global mask to improve the selection of model-analogs for analog forecasting, and then used their mask to explore sources of predictability. However, their approach is state-independent and their forecasts struggle to predict extreme events.

Here, we find a pattern of weights identifying where the model-analogs should most closely match each initial (target) anomalous state. That is, regions with higher weights are those where initial errors may have a greater impact on subsequent anomaly evolution. Fig.

104 1c illustrates that optimized model-analogs selected using predicted weights exhibit smaller
105 error growth compared to the original model-analogs.

106 Our forecasting method is an interpretable-by-design approach, blending deep learning
107 with interpretable methods (Chen et al. 2019; Rudin 2019). We decompose the forecasting
108 processes into two components: determining the best initial state matches and tracking
109 subsequent evolution through the analog method. Specifically, this approach offers two key
110 advantages in terms of interpretability. First, the estimated weights show regions where error
111 growth is particularly sensitive to initial condition uncertainty. These weights (i.e.,
112 explanations by the network) are directly used for analog forecasting and integrated in the
113 training process (ante-hoc), unlike the post-hoc explanations provided by XAI. Second, once
114 analogs are identified using weights, we can trace the physically-based evolution of any other
115 field available in the model simulation for any lead time. This is a key advantage of the
116 model-analog technique that is unattainable with a standalone neural network unless it is
117 trained for all variables.

118 Our approach improves forecast skill of equatorial Pacific SST in both perfect-model and
119 real-world experiments. While many machine learning-driven studies typically focus on
120 predicting simple Niño indices (Ham et al. 2019; Petersik and Dijkstra 2020; Cachay et al.
121 2021; Chen et al. 2021; Ham et al. 2021; Shin et al. 2022), we aim to improve the prediction
122 of the spatial pattern of equatorial Pacific SST given the considerable diversity of individual
123 ENSO events (Capotondi et al. 2015). Additionally, we explore the connection between the
124 predicted weights and various physical processes associated with ENSO dynamics, including
125 the asymmetry in initial-error-sensitivity for El Niño and La Niña. We describe our data and
126 methods in Section 2, then evaluate forecast skill in perfect-model experiments in Section 3
127 and real-world experiments in Section 4. In Section 5, we investigate initial-error sensitivity
128 through estimated weights. The selection and effects of network size are discussed in Section
129 6. Finally, Section 7 provides a summary of our results.

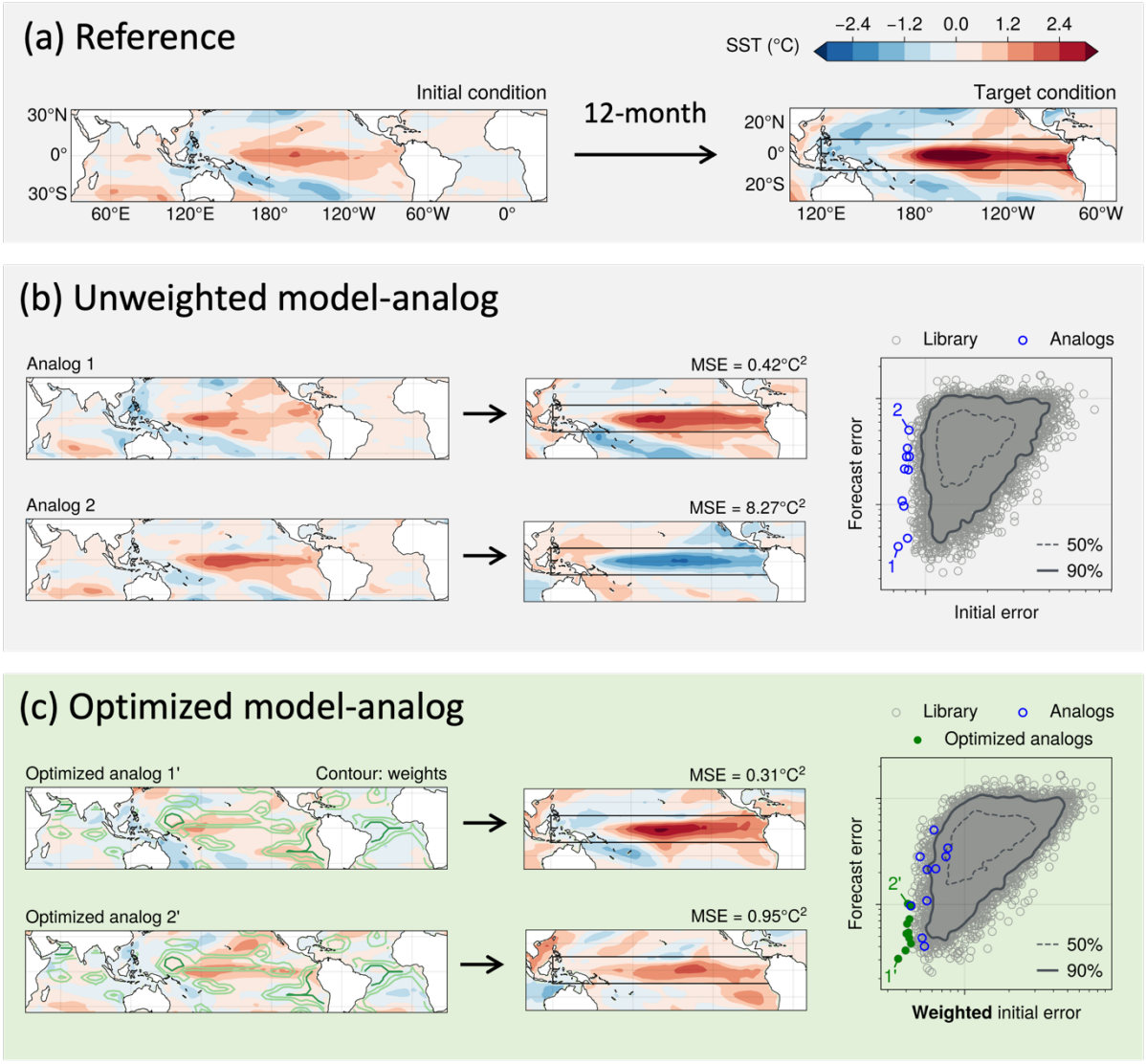


Fig. 1. Schematic method overview of the current study. (a) Reference initial condition for analog selection and target condition 12 months after. The black box in the target condition represents the equatorial Pacific, which is the focus area in this study. (b) Unweighted model-analogs and corresponding forecasts for the best and worst analogs. The mean square errors (MSEs) of the forecasts are shown in each panel. The scatter plot shows initial errors and forecast errors for all samples in the library, along with smoothed probability density curves. Blue circles show 10 analogs with the smallest initial errors. (c) As in (b), but for the optimized model-analogs which exhibit smaller error growth compared to the original analogs. This method uses deep learning to derive optimized weights for analog selection, displayed by contour lines. The scatter plot uses weighted initial errors on the x-axis. Green circles represent 10 optimized analogs, which may be compared to the original analogs represented by blue circles.

2. Methods

a. Data

We first evaluate the hybrid deep learning and model-analog approach within a perfect-model framework, with the same model generating training, validation, and test datasets. We use an ensemble of historical simulations from the Community Earth System Model Version 2 Large Ensemble (CESM2-LE; Rodgers et al. 2021). The CESM2-LE historical simulation consists of 100 ensemble members during 1850–2014, resulting in 16,500 years of data. We use monthly mean sea surface temperature (SST), sea surface height (SSH), and zonal wind stress (TAUX) data. These data are interpolated to two different resolutions, 2° × 2° and 5° × 5°. The coarser resolution data are used to train the neural network model and to select analogs, while the finer resolution data are used as forecasts after finding analogs. Detrended anomalies are determined by removing the ensemble mean temporally smoothed with a 30-year centered running mean. Throughout this study, we exclusively use anomalies. We partition the dataset into training (1865–1958; 9400 years, 70%), validation (1959–1985; 2700 years, 20%), and test (1986–1998; 1300 years, 10%) subsets. The training dataset is also used as the library to select model-analogs.

To test the trained model with observed estimates, we use the Ocean Reanalysis System 5 (ORAS5; Zuo et al. 2019) interpolated to the fine and coarse resolution grids. This evaluation uses a fair-sliding anomaly approach that refrains from using future data not available at the time of the forecast (Risbey et al. 2021). Specifically, anomalies are determined by removing the mean and linear trend during the prior 30 years up to the year of the current forecast. Note that our model is not trained on any reanalysis data.

b. Architecture of the optimized model-analog approach

We develop a deep learning method to predict weights based on a specified initial condition. To reduce computational cost, we use the coarse resolution data over 50°S–50°N (13 latitudes × 72 longitudes × 3 variables) as our input. The architecture of the optimized model-analog approach is depicted in Fig. 2. Our chosen model is the U-Net (Ronneberger et al. 2015), a fully convolutional network consisting of a symmetrically designed downsampling encoder followed by an upsampling decoder. We also experimented with variations such as U-Net with residual blocks (He et al. 2015) and with attention gates (Oktay et al. 2018), but found minimal differences.

The encoder in our architecture consists of stacked blocks, each including two convolutional layers and a max pooling operation, halving the spatial resolution while doubling the channel size (i.e., last dimension). Mirroring the encoder, the decoder includes similar stacked blocks where each incorporates a transposed convolutional layer followed by two convolutional layers. This setup reverses the encoder's blocks by doubling the spatial resolution and reducing the channel size by half. Additionally, we use skip connections, which concatenate the features from the downsampling encoder into the decoder at the corresponding level. A final 1×1 convolution aligns the output channel size with the number of input variables.

Two hyperparameters, namely depth and initial channel size, greatly influence the network size. Here, depth corresponds to the number of blocks in the encoder, set as 4 in this study. The initial channel size, set at 64 in our study, is the output channel size of the first encoder block. Either increasing the depth by one or doubling the initial channel size quadruples U-Net parameters. The sensitivity of the obtained results to the network size is discussed in Section 6.

The U-Net predicts weights that are used to determine weighted initial distances from the input initial condition for every sample within the library. The library comprises all states from the training dataset of the corresponding calendar month, which introduces seasonal cycle effects. The weighted initial distance (d_0) between the target state and each library state is defined as the sum of weighted mean square errors ($\mathrm{MSE_w}$) of standardized SST, SSH, and TAUX anomalies over 50°S–50°N,

$$d_0 = \mathrm{MSE_w}(\mathrm{SST}) + \mathrm{MSE_w}(\mathrm{SSH}) + \mathrm{MSE_w}(\mathrm{TAUX})\,, \tag{1}$$

where $\mathrm{MSE_w}$ of the standardized anomalies is defined as:

$$\mathrm{MSE_w} = \frac{\sum_i w_i \cos\phi_i \left(\frac{x_i}{\sigma_X} - \frac{y_i}{\sigma_Y}\right)^2}{\sum_i w_i \cos\phi_i} \tag{2}$$

Here, i represents a spatial degree of freedom, w represents the weight predicted by U-Net, ϕ denotes latitude, $\cos\phi$ accounts for the grid area weight, x represents the input initial state, and y represents each state in the library. Additionally, σ_X and σ_Y represent the square root of domain-averaged variance over the input domain, used for standardization purposes. Note that for $w_i = 1$, d_0 is essentially the same as the distance metric used by Ding et al. (2018) to determine unweighted model-analogs.

The most intuitive training method might be selecting analogs with the smallest weighted initial distances and defining a loss function based on analog forecast errors. However, this approach involves the complex time evolution of the climate model, with unknown analytical derivatives. Thus, we opt for a more efficient strategy to update model parameters.

Initially, the weighted initial distances are sorted, and samples with the lowest weighted initial distances are selected, specifically the top 2% of samples. We focus on these subsamples so that the network is not affected by samples that significantly deviate in initial conditions. As the network is updated and predicts different weights, a different set of subsamples is selected. Note that the sensitivity to the number of retained samples is relatively low. The loss function is defined as the mean-square-error (MSE) between the normalized weighted initial distances (d_0) and forecast errors (d_τ) of the chosen subsamples, where the forecast error is defined as the MSE of SST over the equatorial Pacific (10°S–10°N, 120°E–70°W; black box in Fig. 1) at a certain lead time (τ). The loss function L_k for the given initial condition (sample index k) can be expressed as:

$$L_k = \frac{1}{n_{sub}} \sum_{j}^{n_{sub}} \left(\frac{d_{0,j}}{\max\limits_{j \in n} d_{0,j}} - \frac{d_{\tau,j}}{\max\limits_{j \in n} d_{\tau,j}} \right)^2 \tag{3}$$

where j represents the index of samples, n_{sub} represents the number of subsamples, and n represents the number of samples in the library. The weighted initial distances and forecast errors are scaled by the respective maximums. Minimizing the loss guides the U-Net to estimate weights that prioritize samples with smaller forecast errors to have smaller weighted initial distances. Essentially, the objective is to maintain consistency in initial and forecast errors across the subsamples. This iterative process is executed for each sample in the training dataset, constituting one epoch.

Although the U-Net can be trained for various lead times (τ), it then results in identifying different analogs for different lead times. This compromises one of the advantages of analog forecasting: the ability to track the time evolution of the system. To address this, we train the U-Net using forecast errors (d_τ) defined by the mean of MSEs across 3, 6, 9, and 12-month lead times over the equatorial Pacific. This approach yields comparable skill to training for specific lead times of 6, 9, or 12 months, as detailed in Appendix B.

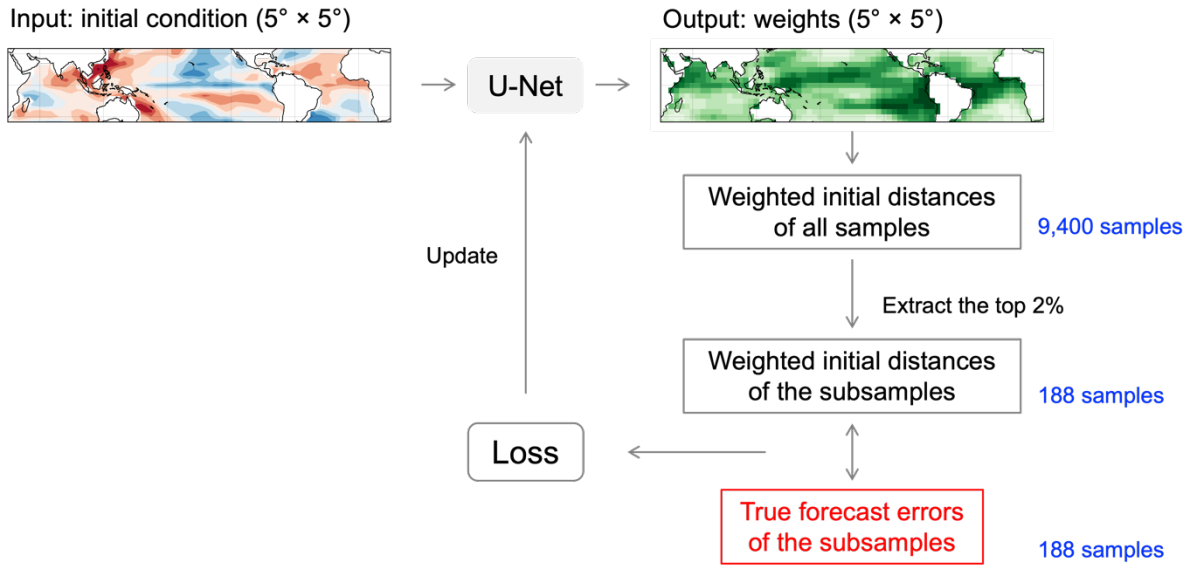


Fig. 2. Architecture of the optimized model-analog approach.

During each epoch, we monitor ensemble-mean forecast error at 12 months lead. Here, we choose 30 analog members (see Appendix A for details). The maximum number of epochs is capped at 60, and we use early stopping to prevent overfitting, i.e. training is stopped when the ensemble-mean forecast error in the validation dataset ceases to decrease. The Adam optimizer (Kingma and Ba 2017) is used to update network parameters. We train the model 10 times to account for the random initialization of U-Net parameters. Since analog selection is performed within the library of the corresponding month, we train a separate U-Net for each month. The source code is available on GitHub (https://github.com/kinyatoride/DLMA).

c. Hyperparameter tuning

Key hyperparameters considered in this study are the initial channel size, depth, learning rate, and subsample size. In the initial phase of hyperparameter tuning, we focus on January initialization with a lead time of 12 months. This choice is motivated by the largest ENSO variability observed during this month in the model. All hyperparameters are optimized based on ensemble-mean forecast error in the validation dataset with a 12-month lead time.

Upon completing the tuning process, the same set of hyperparameters is adopted for other initialization months, except for the learning rate. Due to the significant impact of the learning rate, we fine-tune this parameter independently for each month.

d. Unweighted model-analog and neural network-only approach

We compare our hybrid approach against both the original (unweighted) model-analog approach and an equivalent neural network-only approach.

The original model-analog approach draws analogs based on unweighted distance (Ding et al. 2018, 2019; Lou et al. 2023). Here, distance is defined as the sum of MSEs of standardized SST and SSH over 30°S–30°N. MSE is similar to the formulation in Eq. (2) but with a constant weight ($w_i = 1$). The number of analog members is set to 30. In contrast to the hybrid method, distances are calculated using the 2° data since no training is required. TAUX and extratropical regions are omitted in this approach, as their inclusion has been found to degrade skill of the original model-analog approach. More discussion can be found in Appendix A.

To address the question of whether combining deep learning and analog forecasting might degrade the deep learning capabilities, we compare with a neural network-only method using a similar architecture. We use the same U-Net architecture except for the final layer. The final 1×1 convolution is adjusted to generate fine-resolution SST fields over the equatorial Pacific. Consequently, this approach takes 5° SST, SSH, and TAUX fields over 50°S–50°N as input and predicts 2° SST over the equatorial Pacific. Given the discrepancy in dimension sizes between inputs and outputs, we apply additional padding and cropping of the data. The number of trainable parameters in this modified U-Net differs from the original by less than 0.01%. While the initial channel size and depth are the same as the original, we tune the learning rate separately for this model. Note that this model is only evaluated for January initialization.

e. Evaluation metrics

We use root-mean-square error (RMSE) and uncentered anomaly correlation square (AC^2) to assess the performance of ensemble-mean forecasts. AC^2 is specifically defined as $AC^2 = (\max(AC, 0))^2$, ensuring that negative correlations are treated as zero.

To test the statistical significance of the improvements achieved through the optimized analog approach over the unweighted approach, we conduct a one-sided permutation test (resampling without replacement) using the time-series of forecasts. The null hypothesis is that the true improvement is zero, which is rejected at the significance level of 5%. The null distribution is constructed through 10,000 permutations. When multiple hypotheses are simultaneously tested, as for a map of gridded data, Wilks (2016) recommends adjusting the

threshold p-value for the number of false discoveries. We use the Benjamini and Hochberg step-up procedure (Benjamini and Hochberg 1995) with a 5% false discovery rate.

To evaluate the probabilistic skill, we use the continuous ranked probability score (CRPS), which corresponds to the integral of the Brier score over all possible threshold values. CRPS can be decomposed into three components: reliability, resolution, and uncertainty (Hersbach 2000). Reliability reflects the flatness of the rank histogram and resolution is linked to the ensemble spread.

3. Forecast verification

a. January initialization

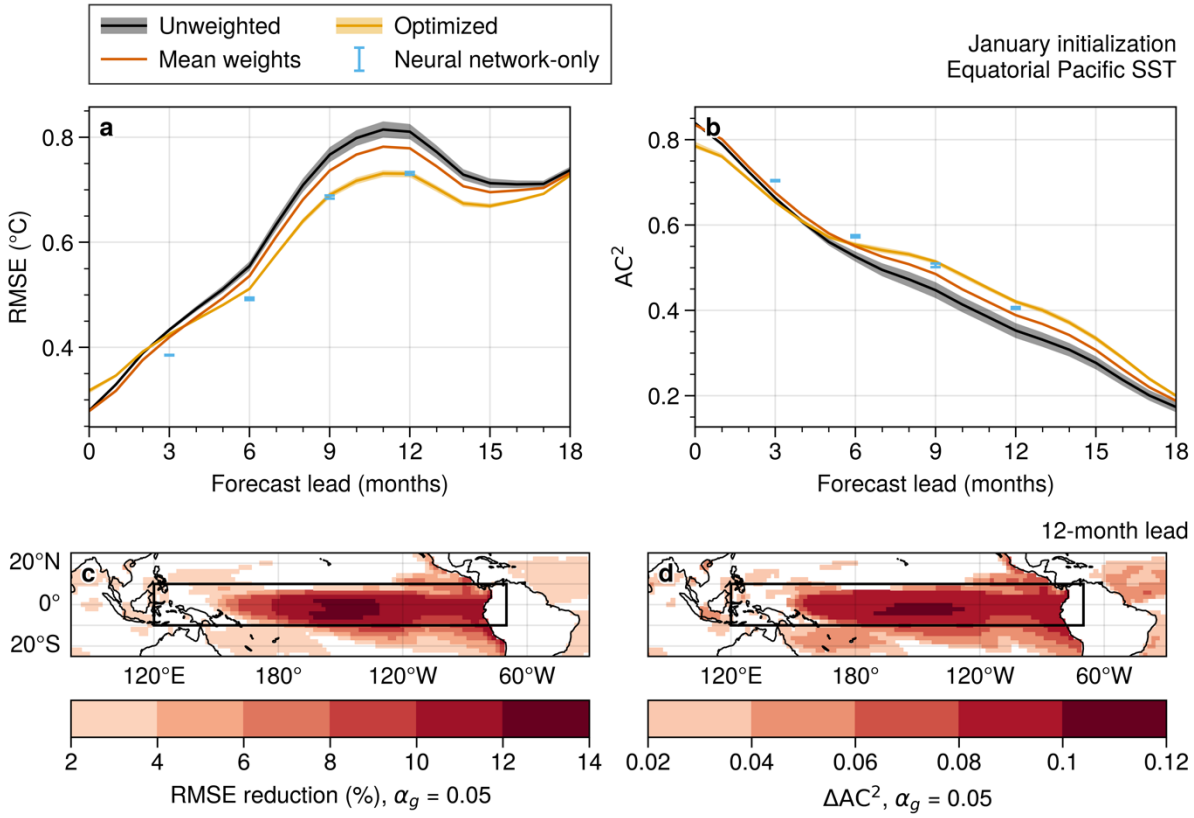
Fig. 3 shows perfect model skill using both unweighted and optimized model-analog methods for January initialization, with the test dataset spanning 1,300 years. The application of deep learning significantly enhances analog selection for forecasting SST patterns over the equatorial Pacific. RMSE is reduced by 10% for a lead time of 9–12 months (Fig. 3a), and AC^2 of 0.4 is extended by more than 2.5 months (Fig. 3b). These improvements remain robust and are minimally affected by random initialization of the training, as indicated by the orange shade. However, for shorter lead times (i.e., 1–2 months lead), the optimized approach exhibits worse forecast errors, suggesting that the neural network assigns more weights to regions beyond the target area to select analogs with better forecasts in longer leads. Consequently, the unweighted approach, which allocates relatively more weights over the equatorial Pacific, results in lower forecast errors for shorter leads.

To evaluate the contribution of the state-dependent aspect of weights to the observed skill improvements, Figs. 3a–b also present the skill of model-analogs selected using state-independent mean weights, estimated by averaging the weights from all January initializations in the test dataset (shown in Fig. 9). Although model-analogs selected with the mean weights perform better compared to the unweighted approach, the improvements are not as significant as those achieved by the optimized approach, particularly at 6–15 months leads. This finding indicates that state-dependent weights are necessary to identify shadowing trajectories.

Figs. 3c–d illustrate the spatial distribution of RMSE reduction and the increase in AC^2 achieved by the optimized approach. Skill is consistently improved east of the Maritime Continent, particularly around the Niño 3.4 region in the central equatorial Pacific. However,

315 over the Maritime Continent, neither RMSE nor AC^2 exhibits significant improvements,
316 primarily due to the small SST variability in the region and the use of MSE in the loss
317 function. The hybrid approach enhances skill in the central equatorial Pacific, where
318 unweighted model-analogs exhibit the highest skill (Ding et al. 2018).

319 Although the optimized model-analog approach significantly improves analog
320 forecasting, we might wonder whether a standalone neural network would produce better
321 forecasts. Figs. 3a–b also display the forecast skill of the equivalent neural network-only
322 method. It is important to note that this method can only generate forecasts at a single lead, so
323 it must be separately trained for 3, 6, 9, and 12 months leads. While the neural network-only
324 method exhibits better skill at 3 and 6 months leads, it demonstrates similar skill at 9 and 12
325 months leads. With respect to AC^2, the optimized model-analog approach shows better
326 accuracy at these leads, where this approach exhibits largest improvements (see Appendix B).
327 These results demonstrate that the combination of neural network and model-analog not only
328 provides an advantage for tracking full-state evolution, but also yields comparable forecast
329 skill compared to a neural network-only approach with a similar architecture and training
330 efforts.



331

Fig. 3. Forecast skill comparison among the unweighted model-analog, optimized model-analog, model-analog with the mean weights, and neural network-only approaches for January initialization using the test dataset. (a) Root-mean-square error (RMSE) of equatorial Pacific SST as a function of forecast lead. The black shading represents the 95% confidence interval estimated through the permutation test between unweighted and optimized results. The orange shading and blue error bars show the spread due to random initialization of network parameters. (b) Similar to (a), but for square anomaly correlation (AC^2) averaged over the equatorial Pacific. (c) RMSE reduction (%) of 12-month lead SST by the optimized approach compared to the unweighted approach. (d) Similar to (c), but for the increase in AC^2. In (c) and (d), color shading indicates statistically significant improvements at the 5% level with the 5% false discovery rate.

b. All-month initialization

Having tuned the hyperparameters for January initialization, we extend the application of the optimized model-analog approach to other initialization months. Fig. 4 shows the seasonal variation of perfect-model AC^2 averaged over the equatorial Pacific. In general, optimized model-analog yields consistent impacts on analog forecasting across all initialization months. While the forecast skill tends to be reduced for shorter leads typically ranging from 0 to 3 months, as the neural network places more weights outside the target region, substantial improvements are made for longer leads ranging from 6 to 18 months. These improvements are particularly notable for initialization during boreal winter and spring (Nov–Apr), with verification during boreal fall and winter (Sep–Mar).

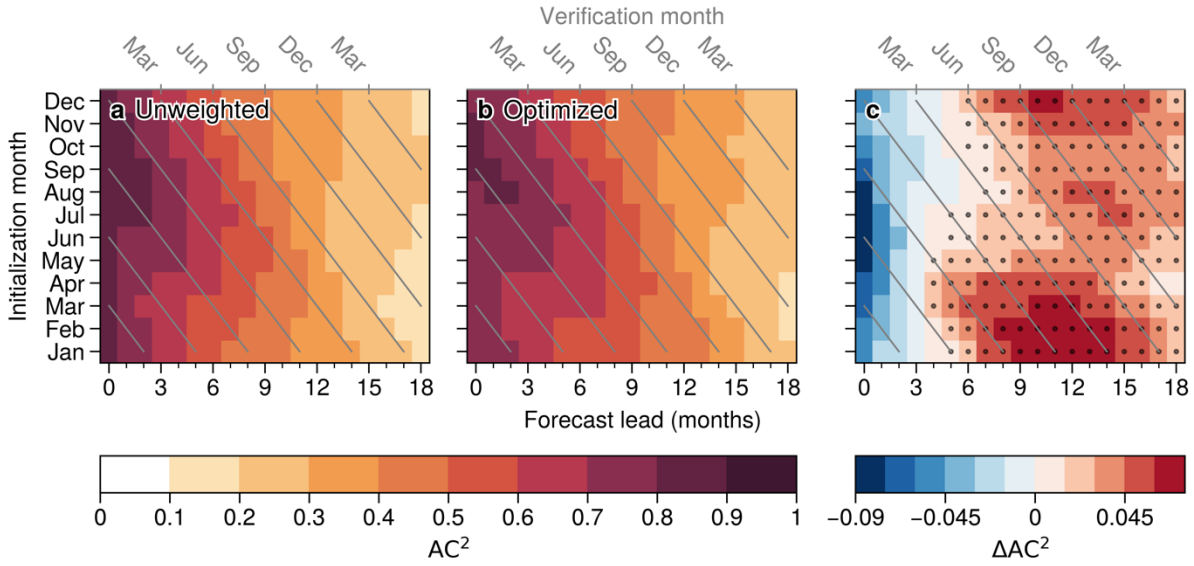
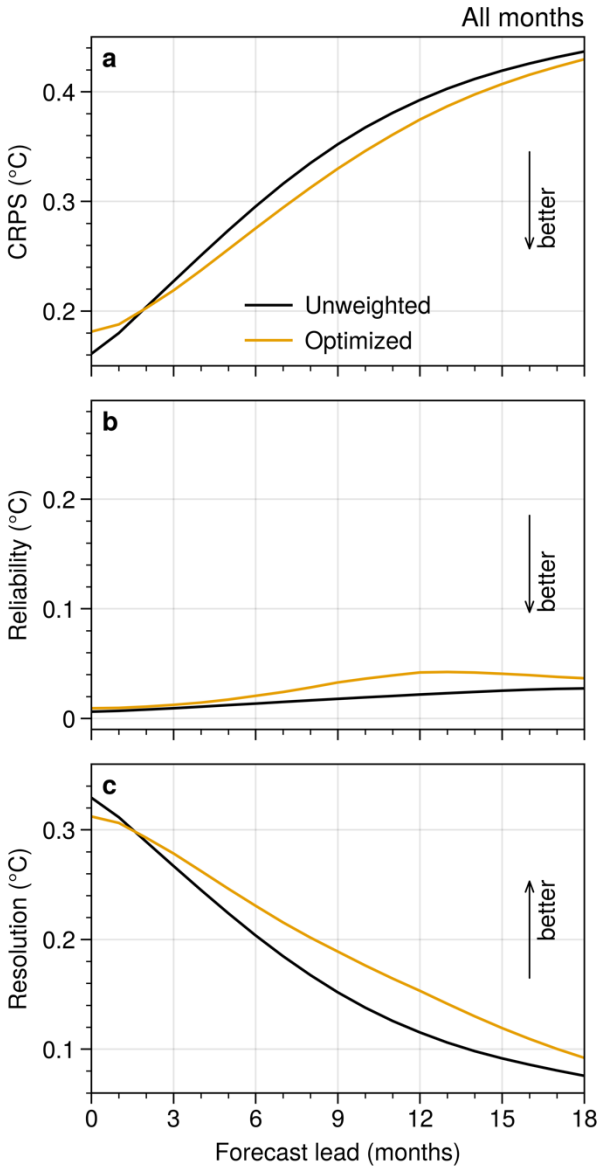


Fig. 4. The seasonality of square anomaly correlation (AC^2) of SST averaged over the equatorial Pacific as a function of forecast lead. (a) The unweighted model-analog, (b) optimized model-analog, and (c) the difference between the two approaches. Stippling in (c) indicates statistically significant improvements. The verification month is indicated by the gray diagonal lines.

Forecasting with analogs is by construction ensemble forecasting. The optimized model-analogs lead to similar probabilistic skill improvements, with reduced skill for shorter leads and enhanced skill for longer leads. This is seen in Fig. 5 which shows the all-month probabilistic forecast skill (CRPS) using 30 analog members. CRPS of 0.4°C is extended for more than 1 month in the all-month average. The improvements in CRPS are attributable to improvements in resolution (Fig. 5c), which may be anticipated given that the loss function is designed to penalize samples deviating significantly at forecast leads, resulting in narrower ensemble spreads. However, smaller ensemble spreads can deteriorate the reliability component, associated with the flatness of the rank histogram, as appears to have occurred in our results (Fig. 5b). The rank histogram is the frequency of the rank of the verification relative to sorted ensemble members. In the absence of ensemble variability, the rank histogram tends to exhibit a U-shaped distribution (Hamill 2001). Since ensemble reliability was not explicitly considered in the loss function, this stands as one of the caveats in this study.



374

375 Fig. 5. (a) Seasonally-averaged continuous ranked probability score (CRPS) of SST over
376 the equatorial Pacific as a function of forecast lead by the unweighted and optimized model-
377 analog methods. Similar to (a), but for (b) reliability and (c) resolution components of the
378 CRPS.

379

380 Once model-analogs are identified, forecasting can be extended to any field available in
381 the climate simulation. This is a distinct advantage in analog forecasting not achievable solely
382 with neural networks, where predictors and predictands must be carefully chosen based on
383 specific phenomena targeted by the model and the available computational resources. Fig. 6
384 shows the improvements in 12-month precipitation forecasting using the optimized model-

analog. Precipitation forecasting is particularly improved in DJF (Fig. 6a), with significant improvements extending beyond the target region including the central subtropical Pacific, Maritime Continent, southwest Pacific east of Australia, southeastern US, northeastern Brazil, and north of Madagascar, potentially linked to ENSO teleconnections. Similarly, forecast skill in MAM is improved both within and outside the target region, albeit with smaller magnitudes (Fig. 6b). While precipitation forecast skill in JJA and SON also displays significant improvements, the impact is primarily confined within the target region (Figs. 6c,d). It is essential to highlight that, while not always statistically significant, positive impacts on precipitation forecasting are observed in most regions across all seasons (not shown). This suggests that improving the model-analog forecasts of tropical SST contributes positively to global precipitation forecasting.

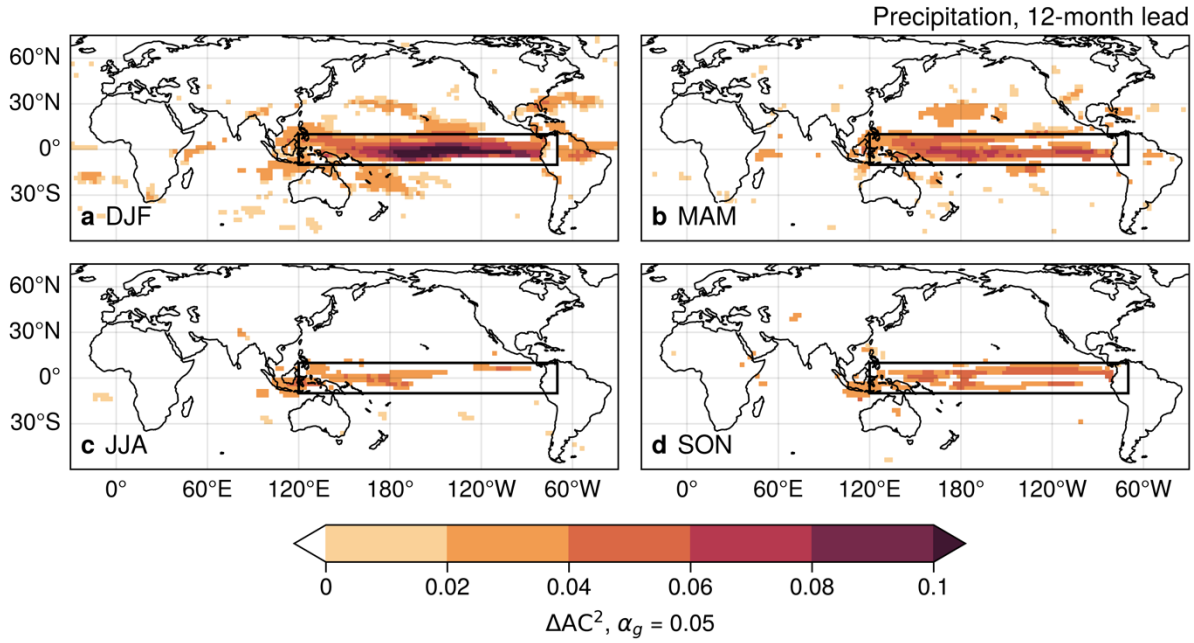


Fig. 6. Increase in square anomaly correlation (AC2) of 12-month lead precipitation by the optimized approach compared to the unweighted approach. The forecasts are initialized and verified for (a) DJF, (b) MAM, (c) JJA and (d) SON. Color shading indicates statistically significant improvements at the 5% level with the 5% false discovery rate.

4. Application to observations

We next apply the developed optimized model-analog approach to make real-world hindcasts by finding optimized model-analogs for initial anomalies drawn from the ORAS5 reanalysis dataset, using the same network but with a limited training epoch of 10 to prevent

overfitting to the CESM2 climate. Recall that we do not use any observations to train the optimized model-analog technique, nor do we employ transfer learning for these hindcasts. Fig. 7 shows the seasonal variation of hindcast skill during 1987–2020. The original (unweighted) model-analog shows lower skill than the perfect-model skill (Fig. 4) with a spring predictability barrier where skill sharply declines around March (Fig. 7a). The impact of the optimized approach varies across initialization months (Fig. 7c), in a manner that is broadly similar to its impact upon perfect model skill (Fig. 4c). However, although positive effects are observed in many initialization months, forecasts initialized in Aug–Oct display a decrease in skill. Statistically significant improvements are observed in boreal fall forecasts initialized in May and June, as well as in year 2 spring forecasts initialized in boreal winter.

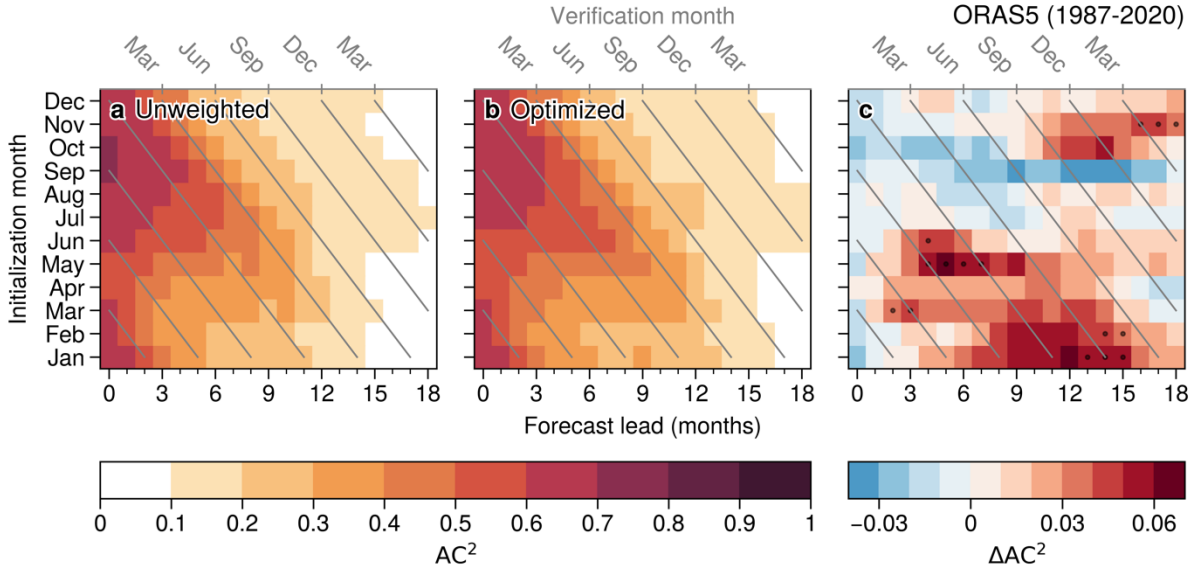


Fig. 7. Similar to Fig. 4, but for hindcast initialized during 1987–2020 using ORAS5.

Fig. 8 illustrates the ENSO conditions under which prediction skill is improved for both perfect-model and observationally-based hindcasts, initialized in January for 12 months lead. It is evident that predictions of extreme events are improved, for both El Niño and La Niña conditions (Fig. 8a), due to their large influences in the loss function. Conversely, predictions for ENSO neutral conditions (below 0.5 σ) show no discernible impacts on the median skill. Although the sample size is small, a similar relationship is observed in the observationally-based hindcasts (Fig. 8b). Apart from the La Niña event in 1996, the optimized approach reduces forecast error for all extreme events above 1 σ (darker shading). However, issues with model errors could also play a role. In Fig. 8a, the optimized approach significantly

improves extreme event forecasts, particularly those characterized by Niño 3.4 values much higher than historically observed values. This result suggests that the neural network may be learning some information with limited relevance to the real world.

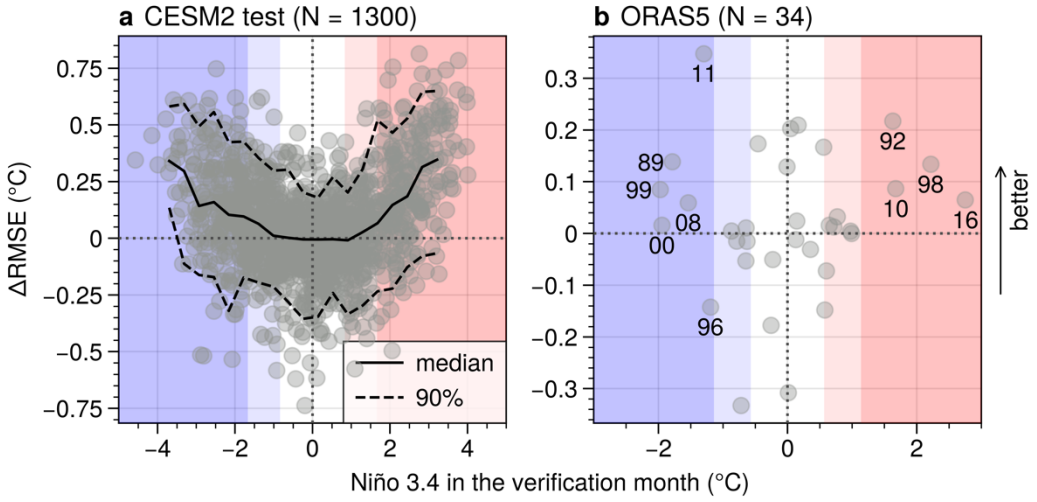


Fig. 8. Scatter plots of the RMSE reduction of SST over the equatorial Pacific and the Niño 3.4 index in the verification month for (a) the CESM2 test dataset and (b) ORAS5. The analysis focuses on 12-month forecasts initialized in January. Lighter pink/blue colors show values above 0.5 σ and darker pink/blue colors show values above 1 σ of the respective Niño 3.4 index in CESM2 and ORAS5. In (a), the median and 90% lines are estimated by binning samples according to the Niño 3.4 index. In (b), the last two digits of verification years are displayed for extreme events.

5. Interpretable weights

The neural network in the optimized model-analog approach produces interpretable weights whose state-dependence significantly impacts forecast skill (Fig. 3) and which can be regarded as indicating sensitivity to initial uncertainty. As in XAI methods, these weights do not provide causal relationships. Instead, they highlight the regions and variables where it is particularly important for the model-analogs to match the initial target anomalies, which will thereby most effectively constrain subsequent anomaly evolution through both physical processes and correlated or dependent features. Fig. 9 illustrates the mean weights for four initialization months using the CESM2 test dataset. Recall that these weights improve forecasts at 6–18 months lead (Fig. 4). Generally, the weights are allocated to similar regions year-round. However, depending on the season, the relative magnitudes of weights differ, indicating varying importance of specific processes or regions. Notably, there are nonzero

weights outside the target region (equatorial Pacific SST, indicated by the black box), although most of the weights are distributed within the tropics (30°S–30°N), suggesting that extratropical contributions are relatively small. These distributions of weights result in selecting analogs with poorer initial match (yet better subsequent trajectories) over the target region than unweighted model-analogs.

The distribution of weights among the three variables varies by calendar month, as shown in Fig. 10. From October to March, the weights are distributed relatively evenly between SST and SSH, with smaller weights for TAUX. April presents a deviation, with SST receiving the largest weights followed by SSH and TAUX. From May to September, the emphasis shifts, with TAUX receiving larger weights compared to SSH. Notably, TAUX receives the largest weights among all variables during June and July.

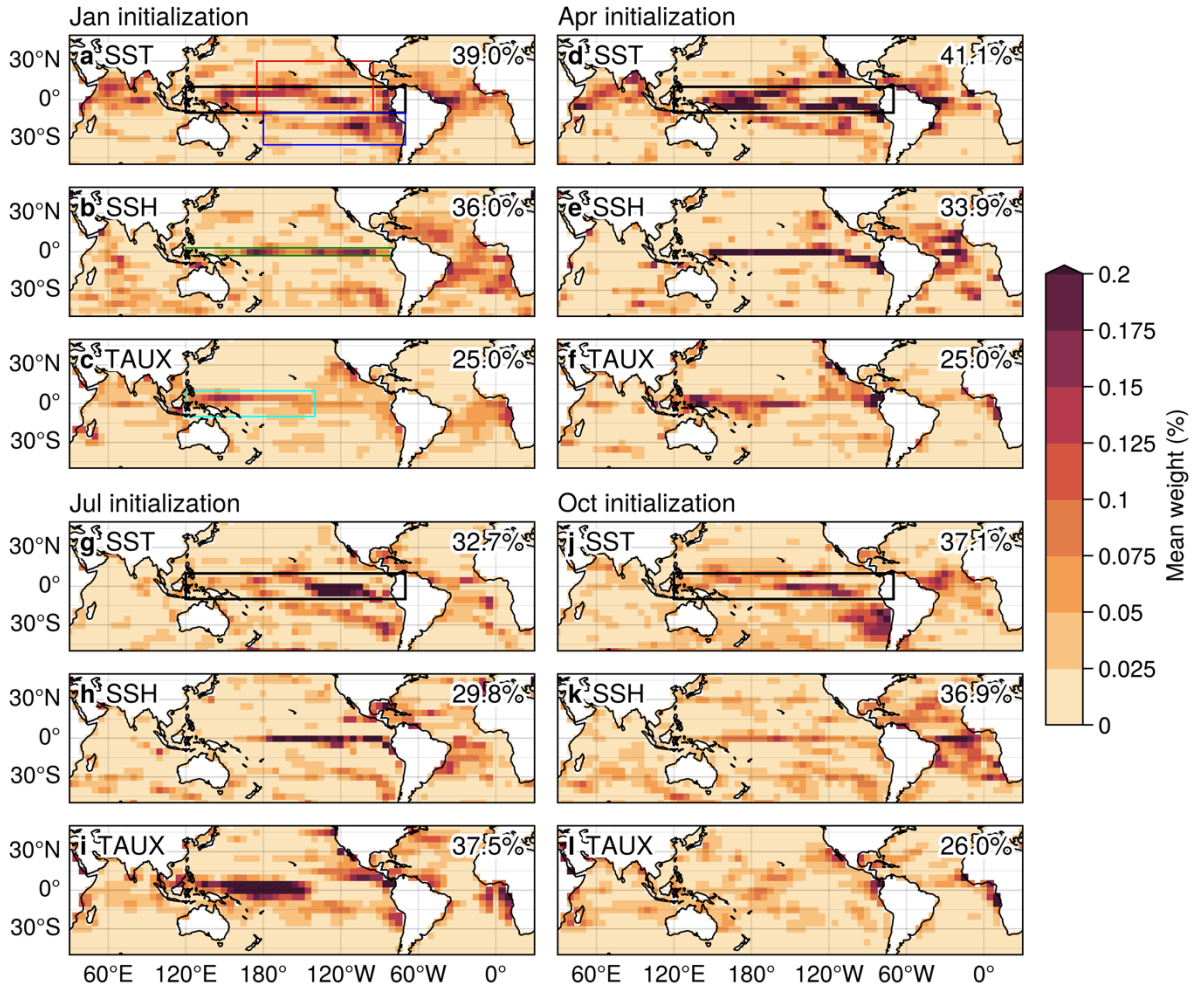
The spatial distributions of weights reveal connections to various physical processes associated with ENSO. In January (Fig. 9a) and April (Fig. 9d), SST receives weights that extend southwestward from the California coast toward the western equatorial Pacific, as well as over the eastern equatorial Pacific. This pattern closely resembles the characteristics of NPMM (Chiang and Vimont 2004; Amaya 2019), a robust predictor of ENSO conditions (Penland and Sardeshmukh 1995; Larson and Kirtman 2014; Vimont et al. 2014; Capotondi and Sardeshmukh 2015; Capotondi and Ricciardulli 2021). We find that largest weights in the NPMM region occur from April to June (Fig. 11a), which is also when the NPMM typically is strongest. Additionally, the SST weights in the subtropical southeastern Pacific resemble the pattern of the South Pacific Meridional Mode (SPMM) (Zhang et al. 2014), particularly evident in January (Fig. 9a) and October (Fig. 9j). The air-sea coupling associated with SPMM peaks in boreal winter (You and Furtado 2018), again consistent with when the SPMM weights are maximized (Fig. 11b). Regarding the July initialization (Fig. 9g), SST weights concentrate more over the eastern equatorial Pacific. This reflects the timing of ENSO events in boreal winter and their influences on subsequent seasons, which are the target leads of the July initialization.

SSH weights are consistently focused over the equatorial Pacific throughout the year, unlike SST (Figs. 9b, e, h, and k). Since SSH is dynamically linked to thermocline depth, this pattern likely relates to the recharge and discharge of upper-ocean heat content during the alternation of warm and cold ENSO phases (Jin 1997). In particular, a recharged state is conducive to the development of an El Nino, while a discharged state may likely lead to a La

Nina. The equatorial weights can constrain the zonal tilt of the equatorial thermocline concurrent with the peak of ENSO, in addition to the recharge-discharge mode which is an important precursor of ENSO (Meinen and McPhaden 2000). Notably, these weights are particularly amplified in April (Fig. 11c). Equatorial Pacific upper-ocean heat content typically precedes Niño 3.4 SST by a quarter of the ENSO cycle (McPhaden 2003), equating to about 8–10 months in CESM2 (Capotondi et al. 2020). Given that ENSO events tend to peak in boreal winter, the peak of weights in April is consistent with these established temporal dynamics.

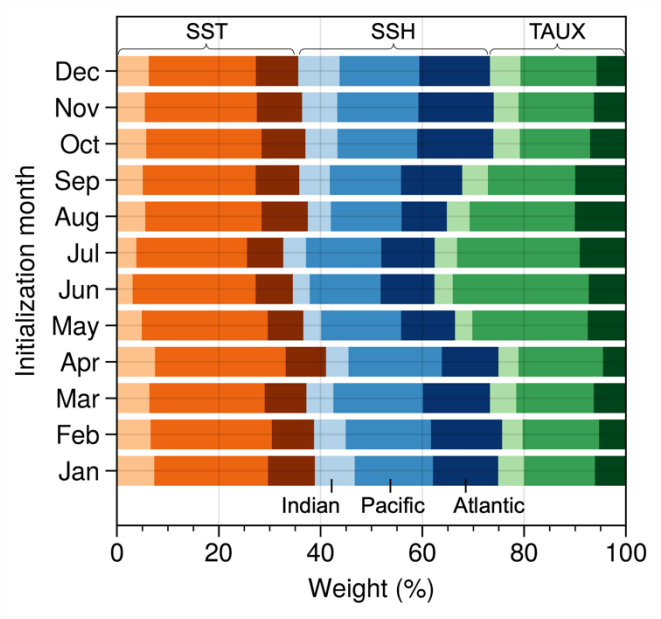
Winds play a crucial role in driving ENSO variability. TAUX weights tend to be largest in the western to central tropical Pacific throughout the year (Figs. 9c, f, i, and l), coinciding with the typical occurrence of stochastic wind forcing across the region. This stochastic forcing exhibits a broad spectrum ranging from subseasonal to interannual scales, with the lower frequency component often exerting a greater influence on ENSO evolution (Roulston and Neelin 2000; Capotondi et al. 2018). During boreal summer, the absence of the interannual component of stochastic wind can restrict ENSO growth (Menkes et al. 2014), elucidating the peak magnitude of wind weights observed in June (Fig. 11d).

Although the target region lies within the tropical Pacific, allocation of weights to the Atlantic and Indian Ocean indicates the impact of tropical interbasin interactions (Cai et al. 2019; Wang 2019). Interestingly, over the Atlantic Ocean larger weights are distributed to SSH compared to SST (Fig. 10). Our result suggests that ocean memory (i.e., upper ocean heat content) may serve as a more reliable proxy for Atlantic influences compared to SST, which measures surface heat. In contrast, large SST weights are observed over the Indian Ocean in January and April, near the Indian Ocean Dipole region.



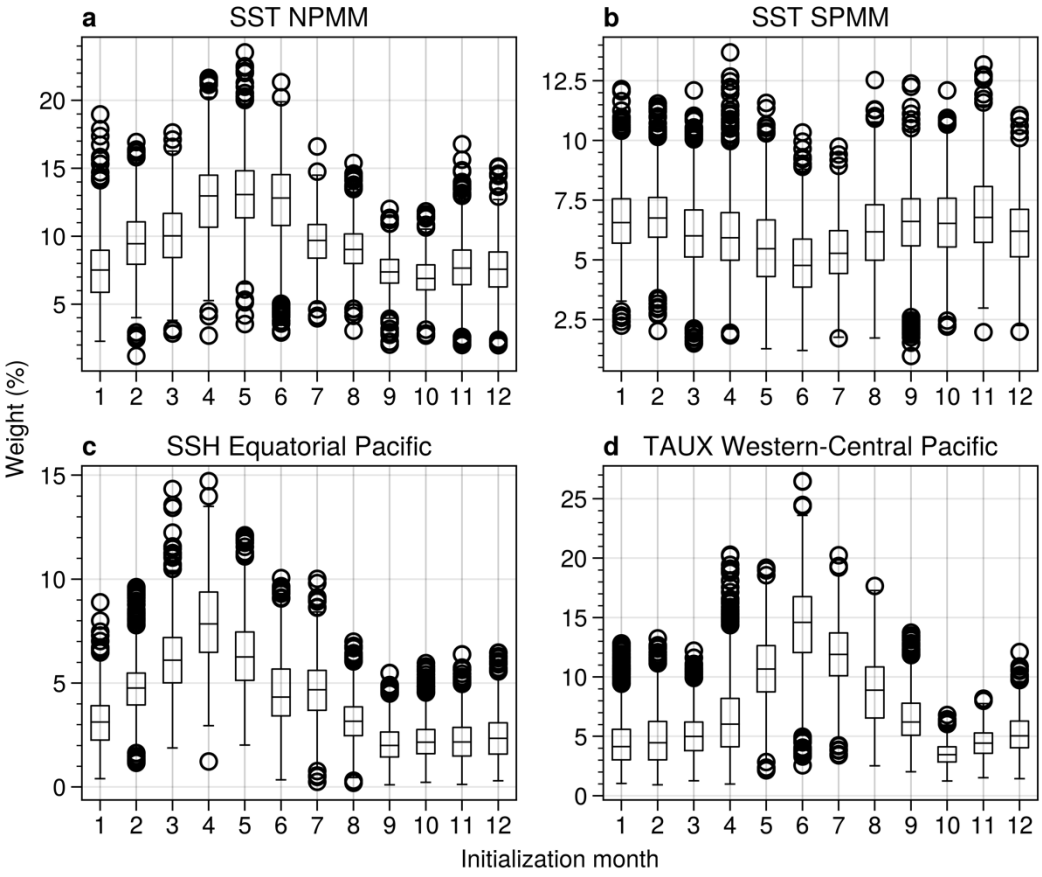
505

506 Fig. 9. Mean weights for (a–c) January, (d–f) April, (g–i) July, and (j–l) October
507 initialization in the CESM2 test dataset. These weights improve the selection of analogs for
508 forecasts with lead times of 6–18 months. Weights are unitless and scaled to ensure a sum of
509 100%. The sum of weights for each variable is displayed within each respective panel.
510 Regions of interest, denoted by red (NPMM SST), blue (SPMM SST), green (equatorial
511 Pacific SSH), and cyan (western to central tropical Pacific TAUX) boxes, are analyzed in
512 Fig. 11.



513

514 Fig. 10. Seasonal variation of mean weights in the CESM2 test dataset. Red, blue, and
515 green represent the total weights for SST, SSH, and TAUX, respectively. The intensity of
516 light, medium, and dark colors indicates the sum of weights over the Indian, Pacific, and
517 Atlantic Oceans, respectively.

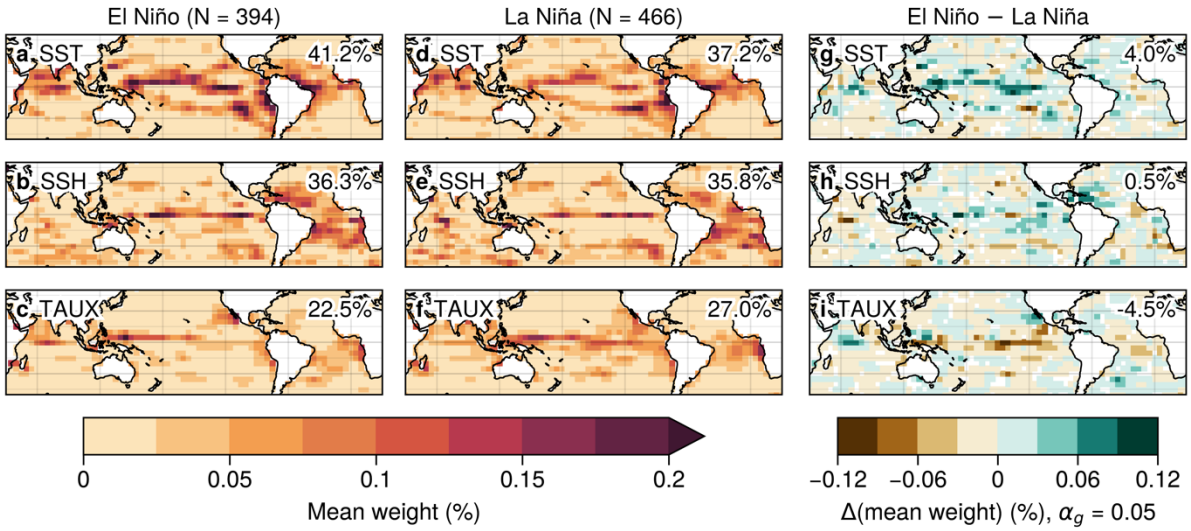


518

519 Fig. 11. Seasonal variation of (a) SST weights over the NPMM region (10°S–30°N,
520 175°E–85°W), (b) SST weights over the SPMM region (35°S–10°S, 180°–70°W), (c) SSH
521 weights over the equatorial Pacific (2.5°S–2.5°N, 120°E–80°W), and (d) TAUX weights over
522 the western to central tropical Pacific (10°S–10°N, 120°E–140°W), as observed in the
523 CESM2 test dataset. Box plots depict the minimum, maximum, median, first and third
524 quantiles, and outliers.

525

526 Since weights are state-dependent, we can analyze the asymmetry in sensitivity associated
527 with El Niño and La Niña. Fig. 12 shows the comparison of mean weights for events
528 evolving to El Niño and La Niña 12 months later, initialized in January. Here, El Niño and La
529 Niña events are defined by above and below ±0.5 σ of the Niño 3.4 index. The spatial
530 distribution of weights generally exhibits similarities to the overall mean (Fig. 9a–c), but
531 differences in magnitude can be observed. Specifically, the SST weights over the Pacific
532 exhibit larger magnitudes for El Niño and weaker magnitudes for La Niña (Fig. 12g).
533 Furthermore, Pacific TAUX weights, particularly along the NPMM region, are larger for La
534 Niña (Fig. 12i). That is, El Niño prediction (from January to the following winter) is more
535 sensitive to initial SST uncertainty, while La Niña prediction is more sensitive to initial
536 surface wind stress uncertainty in the eastern equatorial Pacific.



537

538 Fig. 12. Mean weights for events that evolve to (a–c) El Niño and (d–f) La Niña
539 conditions in 12 months using January initialization. (g–f) The difference in mean weights
540 between El Niño and La Niña. Color shading indicates statistically significant differences at
541 the 5% level with the 5% false discovery rate.

6. Network size

The complexity of a model, often indicated by the number of parameters, plays an important role in machine learning studies. Although the trend in the field leans towards more complex models with advanced skill, it is equally important to explore the potential gains achievable with simpler models, especially for those with resource constraints. As described in the Methods section, the network size is controlled by two key hyperparameters: depth and initial channel size. We employ a depth of 4 and an initial channel size of 256 in this study (referred to as 4-256), resulting in 123 million trainable parameters. This is determined through hyperparameter tuning and training cost considerations.

Either reducing the depth by 1 or halving the initial channel size decreases the number of parameters by a factor of four. We found that reducing the depth degrades model performance more than reducing the initial channel size. This may be due to the reduction in the receptive field size, which represents the region in the input space influencing an output in a single grid, associated with decreasing depth. Since forecasting ENSO requires capturing large-scale teleconnections as illustrated in the estimated weights (Fig. 9), maintaining a deep network is imperative. Although it is tempting to have a deeper network, the current input size limits the depth to 4.

Therefore, we conduct a sensitivity analysis by varying the initial channel size. Fig. 13a shows the reduction in RMSE on the validation dataset for different network sizes. As the network size increases, the skill improvement follows an asymptotic trend. Statistical tests reveal no significant difference between the 4-256 model and the 4-64 model, which has 16 times fewer parameters. Yet, a significant difference is observed between the 4-512 and 4-64 models (not shown). Hence, one needs to consider the trade-off between computational costs and model performance.

The training duration for the 4-256 model is approximately 30 minutes and 1 hour with a single NVIDIA A100 and A6000 GPU, respectively (Fig. 13b). While the training time decreases with a smaller model, the difference diminishes for models with an initial channel size smaller than 128. This is due to the sorting of samples in the library, as shown in Fig. 2. With smaller networks, sorting time dominates, while larger networks exponentially increase training time. It is essential to note that actual training time and sensitivity to network size may vary depending on the system used.

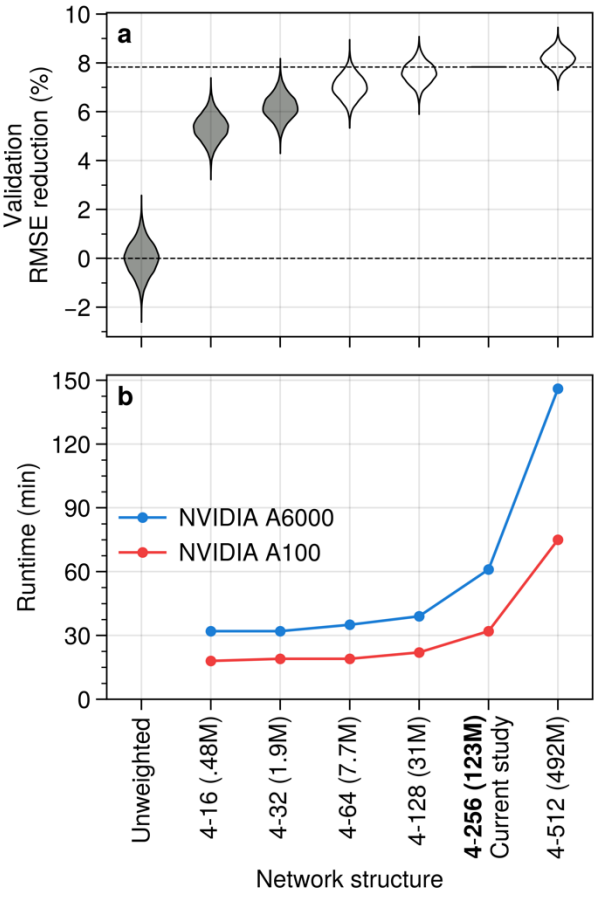


Fig. 13. (a) RMSE reduction (%) of 12-month lead SST over the equatorial Pacific in the validation dataset for different network structures. The network structure is denoted by depth-(initial channel size) with parameter counts in parentheses. Violin plots illustrate the null distribution estimated through permutation with the 4-256 model results. Gray shading indicates values are significantly different at a 5% level. (b) Approximate time taken to train U-Net models for 60 epochs using a single NVIDIA A6000 or A100 GPU in this study.

7. Conclusion

In this study, we introduce an interpretable-by-design forecasting approach called the optimized model-analog method, which integrates deep learning with model-analogs. We demonstrate how deep learning can enhance the potential of model-analog forecasting, specifically by identifying regions highly sensitive to initial uncertainty. The optimized model-analog approach yields comparable forecast skill to a standalone neural network approach, while offering additional benefits associated with analog forecasting. This approach generates interpretable, state-dependent weights that are used to select analog members. These estimated weights highlight regions that are particularly sensitive to initial

uncertainty. As a result, analogs selected with weighted distances shadow the target trajectory closer than original model-analogs. Additionally, the convolutional neural network employed in our study exhibits robust improvements across various network sizes.

The application to ENSO forecasting shows significant improvements in perfect model skill at 6–18 months leads. The most significant improvements are observed in the central equatorial Pacific region and in predicting extreme events due to the large SST variability. Once optimized model-analogs are identified based on weighted distances, their subsequent time evolution can be analyzed in any fields available in the original climate simulation dataset. We demonstrate that improving equatorial Pacific SST forecasts also results in improving precipitation forecasting beyond the target region.

We additionally show improvements in real-world applications across many initialization months and extreme events, although certain initialization months exhibit a reduction in forecast skill. Several factors contribute to the differences between real-world and perfect-model results. Climate models inherently possess systematic errors, such as the excessive westward extension of the SST anomalies associated with ENSO (Bellenger et al. 2014), which is also evident in the CESM2 model (Capotondi et al. 2020) and in all seasonal climate model forecasts (Newman and Sardeshmukh 2017; Beverley et al. 2023). If the neural network learns a model attractor that is significantly different from reality, it can deteriorate skill. A potential solution to mitigate model biases involves employing multiple climate models, as demonstrated in model-analog studies (Ding et al. 2018, 2019; Lou et al. 2023), and machine learning studies (Ham et al. 2019; Zhou and Zhang 2023). Transfer learning may also alleviate biases, although with limitations due to sample size and the effects of climate change. Additional reasons for less significant results include a limited sample size, uncertainty in the fair-sliding anomaly calculation method, and uncertainty in the reanalysis dataset used both to choose initial model-analogs and to verify the subsequent hindcasts. Future work should address these challenges by mitigating the effects of model biases, potentially through the incorporation of multiple climate models and leveraging transfer learning techniques, and by developing hindcasts based on multiple different reanalysis datasets.

The hybrid approach predicts weights linked to various known physical processes. Specifically, SST weights exhibit patterns similar to NPMM peaking in boreal spring and SPMM peaking in boreal winter. SSH weights are concentrated over the equatorial Pacific,

likely capturing states linked to the recharge-discharge of warm water volume associated with ENSO oscillatory behavior. TAUX weights are large in regions where stochastic wind forcing typically occurs, with a peak in boreal summer. Furthermore, some weights are distributed over the Atlantic and Indian Ocean, indicating the influence of the tropical interbasin interactions. These weights are generated by the neural network method used, implying that it is straightforward to integrate superior deep learning algorithms for improved weight quantification.

Our approach mirrors the principles of adjoint sensitivity, where a linearized model is used to assess the sensitivity of a specific aspect of the final forecast to initial conditions (Errico 1997). While adjoint sensitivity is effective only under the validity of the linearized approximation, our approach accommodates nonlinear evolutions of analog trajectories. Additionally, our method can be viewed as a nonlinear and flow-dependent extension of singular vectors (Diaconescu and Laprise 2012) or optimal perturbations (Penland and Sardeshmukh 1995). These methods identify perturbations with maximum growth under a specific norm over a finite time interval. Despite the conceptual similarities, our approach stands out by not requiring a predefined target once trained when forecasting from a given initial condition.

There are many possible applications of this approach. It can be used for different climate phenomena across various regions, such as regional temperature and precipitation. This has been challenging with the unweighted model-analog because the selection of input variables and input regions must be made for each target, which could be subjective. The optimized model-analog approach addresses this issue by optimizing the focus (i.e., weights) in the input space using neural networks.

Another application is evaluating the regional and variable contributions to forecasting skill, including the assessment of interactions between the tropical basins. Broadly, two approaches can be considered: 1) training neural networks with restricted regions/variables, and 2) modifying (i.e., zeroing) predicted weights of certain regions/variables. The first approach may yield results that are difficult to interpret due to correlations between used and unused features. On the other hand, the latter approach involves post-modification after model training and selects analogs without constraining a part of the input. This approach could provide interesting insights into quantifying the contribution of a specific feature by allowing error growth from that feature.

Acknowledgments.

This work was supported by the Famine Early Warning Systems Network and the NOAA Physical Sciences Laboratory. Jakob Schlör was supported by EXC number 2064/1 – Project number 390727645 and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). The authors thank Tim Smith, Jannik Thümmel, and Elizabeth Barnes for comments that improved this work.

Data Availability Statement.

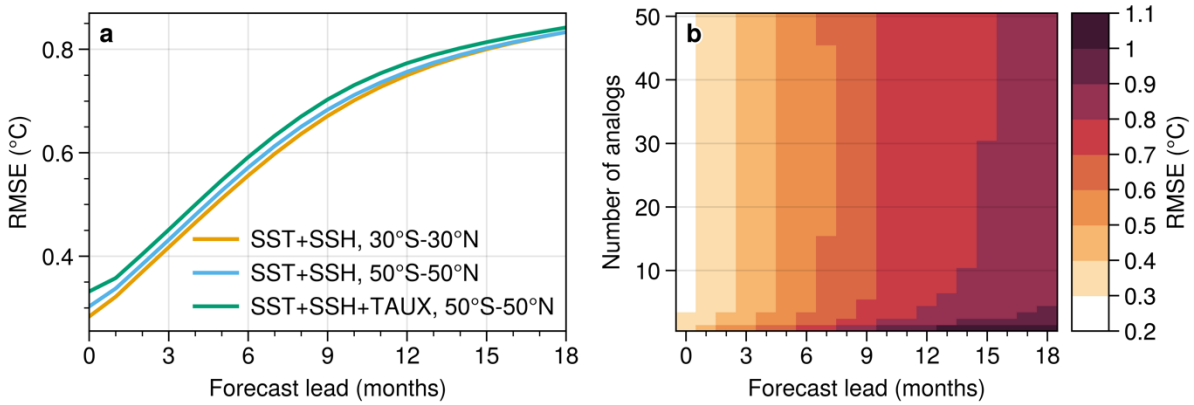
The CESM2-LE dataset is available from The National Center for Atmospheric Research (https://doi.org/10.26024/kgmp-c556). The ORAS5 dataset is available from the European Centre for Medium-Range Weather Forecasts (https://doi.org/10.24381/cds.67e8eeb7). The optimized model-analog codes are publicly available on GitHub (https://github.com/kinyatoride/DLMA).

APPENDIX

Appendix A Unweighted model-analog

This section presents the sensitivity of unweighted model-analog results to some parameters. Fig. A1a shows a skill comparison among different input regions and variables. The highest skill is achieved with SST and SSH over the tropics (30°S–30°N), as used in Lou et al. (2023). Expanding the input domain to the extratropics and including TAUX lead to a degradation in skill. Although the optimized model-analog approach assigns weights to the three variables over 50°S–50°N, we choose the one with SST and SSH over the tropics to avoid underestimating the skill of the unweighted approach.

Fig. A1b shows the sensitivity to analog member size. RMSE clearly worsens with a member size of fewer than 10. We select a member size of 30, which minimizes RMSE at lead times of 6–12 months.



680

681 Fig. A1. (a) RMSE of equatorial Pacific SST as a function of forecast lead on the test
682 dataset. Three unweighted model-analog approaches with different inputs are evaluated. (b)
683 RMSE of equatorial Pacific SST as a function of forecast lead and analog member size.

684 **Appendix B Lead time dependence**

685 Fig. B1 shows a comparison of RMSE reduction using different forecast errors in the loss
686 function. The model is trained with MSE at a specific lead time (3, 6, 9, or 12 months) in
687 addition to using averaged MSE over 3, 6, 9, and 12 months leads. Note that the learning rate
688 is fine-tuned independently. While the training results with a lead time of 3 months exhibit
689 significantly different behavior, other results display more similarity. This tendency is also
690 observed in the estimated weights, where the 3-month lead results focus more on the tropical
691 Pacific (not shown). Among longer leads, the 6-month lead results yield the highest skill,
692 especially for shorter leads. The results with the averaged MSE are slightly worse around 6-
693 month lead but generally comparable to the 6-month lead results. Considering the potential
694 dependency on the initial month for training results at specific lead times, we use the
695 averaged MSE in this study.

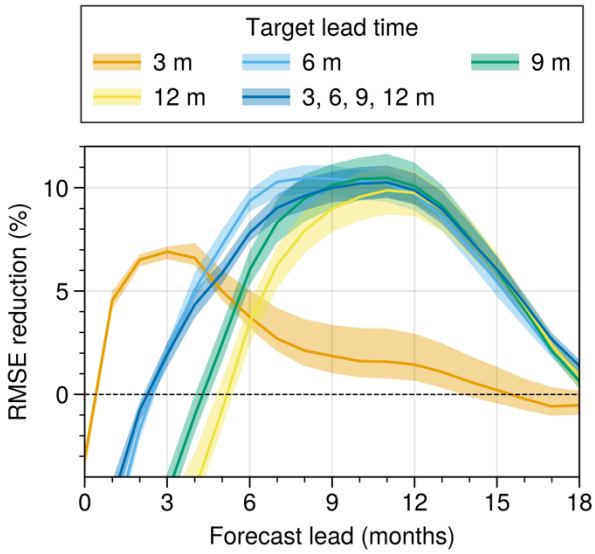


Fig. B1. RMSE reduction (%) of equatorial Pacific SST as a function of forecast lead for January initialization using the test dataset. The optimized model-analog is trained for various lead times. Shading shows the spread due to random initialization of network parameters.

REFERENCES

Alexander, M. A., I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott, 2002: The Atmospheric Bridge: The Influence of ENSO Teleconnections on Air–Sea Interaction over the Global Oceans. *Journal of Climate*, **15**, 2205–2231, https://doi.org/10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2.

Amaya, D. J., 2019: The Pacific Meridional Mode and ENSO: a Review. *Curr Clim Change Rep*, **5**, 296–307, https://doi.org/10.1007/s40641-019-00142-x.

Barsugli, J. J., and P. D. Sardeshmukh, 2002: Global Atmospheric Sensitivity to Tropical SST Anomalies throughout the Indo-Pacific Basin. *Journal of Climate*, **15**, 3427–3442, https://doi.org/10.1175/1520-0442(2002)015<3427:GASTTS>2.0.CO;2.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: ENSO representation in climate models: from CMIP3 to CMIP5. *Clim Dyn*, **42**, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Beverley, J. D., M. Newman, and A. Hoell, 2023: Rapid Development of Systematic ENSO-Related Seasonal Forecast Errors. *Geophysical Research Letters*, **50**, e2022GL102249, https://doi.org/10.1029/2022GL102249.

721 Cachay, S. R., E. Erickson, A. F. C. Bucker, E. Pokropek, W. Potosnak, S. Bire, S. Osei, and
722 B. Lütjens, 2021: The World as a Graph: Improving El Ni\~no Forecasts with Graph
723 Neural Networks. https://doi.org/10.48550/arXiv.2104.05089.

724 Cai, W., and Coauthors, 2019: Pantropical climate interactions. *Science*, **363**, eaav4236,
725 https://doi.org/10.1126/science.aav4236.

726 Capotondi, A., and P. D. Sardeshmukh, 2015: Optimal precursors of different types of ENSO
727 events. *Geophysical Research Letters*, **42**, 9952–9960,
728 https://doi.org/10.1002/2015GL066171.

729 ——, and L. Ricciardulli, 2021: The influence of pacific winds on ENSO diversity. *Sci Rep*,
730 **11**, 18672, https://doi.org/10.1038/s41598-021-97963-4.

731 ——, and Coauthors, 2015: Understanding ENSO Diversity. *Bulletin of the American*
732 *Meteorological Society*, **96**, 921–938, https://doi.org/10.1175/BAMS-D-13-00117.1.

733 ——, P. D. Sardeshmukh, and L. Ricciardulli, 2018: The Nature of the Stochastic Wind
734 Forcing of ENSO. *Journal of Climate*, **31**, 8081–8099, https://doi.org/10.1175/JCLI-
735 D-17-0842.1.

736 Capotondi, A., C. Deser, A. S. Phillips, Y. Okumura, and S. M. Larson, 2020: ENSO and
737 Pacific Decadal Variability in the Community Earth System Model Version 2.
738 *Journal of Advances in Modeling Earth Systems*, **12**, e2019MS002022,
739 https://doi.org/10.1029/2019MS002022.

740 Chattopadhyay, A., E. Nabizadeh, and P. Hassanzadeh, 2020: Analog Forecasting of
741 Extreme-Causing Weather Patterns Using Deep Learning. *Journal of Advances in*
742 *Modeling Earth Systems*, **12**, e2019MS001958,
743 https://doi.org/10.1029/2019MS001958.

744 Chen, C., O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, 2019: This Looks Like That: Deep
745 Learning for Interpretable Image Recognition.
746 https://doi.org/10.48550/arXiv.1806.10574.

747 Chen, N., F. Gilani, and J. Harlim, 2021: A Bayesian Machine Learning Algorithm for
748 Predicting ENSO Using Short Observational Time Series. *Geophysical Research*
749 *Letters*, **48**, e2021GL093704, https://doi.org/10.1029/2021GL093704.

750 Chiang, J. C. H., and D. J. Vimont, 2004: Analogous Pacific and Atlantic Meridional Modes
751 of Tropical Atmosphere–Ocean Variability. *Journal of Climate*, **17**, 4143–4158,
752 https://doi.org/10.1175/JCLI4953.1.

753 Diaconescu, E. P., and R. Laprise, 2012: Singular vectors in atmospheric sciences: A review.
754 *Earth-Science Reviews*, **113**, 161–175,
755 https://doi.org/10.1016/j.earscirev.2012.05.005.

756 Ding, H., M. Newman, M. A. Alexander, and A. T. Wittenberg, 2018: Skillful Climate
757 Forecasts of the Tropical Indo-Pacific Ocean Using Model-Analogs. *Journal of*
758 *Climate*, **31**, 5437–5459, https://doi.org/10.1175/JCLI-D-17-0661.1.

759 ——, ——, ——, and ——, 2019: Diagnosing Secular Variations in Retrospective ENSO
760 Seasonal Forecast Skill Using CMIP5 Model-Analogs. *Geophysical Research Letters*,
761 **46**, 1721–1730, https://doi.org/10.1029/2018GL080598.

762 Errico, R. M., 1997: What Is an Adjoint Model? *Bulletin of the American Meteorological*
763 *Society*, **78**, 2577–2592, https://doi.org/10.1175/1520-
764 0477(1997)078<2577:WIAAM>2.0.CO;2.

765 Grebogi, C., S. M. Hammel, J. A. Yorke, and T. Sauer, 1990: Shadowing of physical
766 trajectories in chaotic dynamics: Containment and refinement. *Phys. Rev. Lett.*, **65**,
767 1527–1530, https://doi.org/10.1103/PhysRevLett.65.1527.

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

770 ——, ——, E.-S. Kim, and K.-W. On, 2021: Unified deep learning model for El
771 Niño/Southern Oscillation forecasts by incorporating seasonality in climate data.
772 *Science Bulletin*, **66**, 1358–1366, https://doi.org/10.1016/j.scib.2021.03.009.

773 Hamill, T. M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts.
774 *Monthly Weather Review*, **129**, 550–560, https://doi.org/10.1175/1520-
775 0493(2001)129<0550:IORHFV>2.0.CO;2.

776 He, K., X. Zhang, S. Ren, and J. Sun, 2015: Deep Residual Learning for Image Recognition.
777 https://doi.org/10.48550/arXiv.1512.03385.

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

781 Hoell, A., and C. Funk, 2013: The ENSO-Related West Pacific Sea Surface Temperature
782 Gradient. *Journal of Climate*, **26**, 9545–9562, https://doi.org/10.1175/JCLI-D-12-
783 00344.1.

784 Jin, F.-F., 1997: An Equatorial Ocean Recharge Paradigm for ENSO. Part I: Conceptual
785 Model. *Journal of the Atmospheric Sciences*, **54**, 811–829,
786 https://doi.org/10.1175/1520-0469(1997)054<0811:AEORPF>2.0.CO;2.

787 Judd, K., L. Smith, and A. Weisheimer, 2004: Gradient free descent: shadowing, and state
788 estimation using limited derivative information. *Physica D: Nonlinear Phenomena*,
789 **190**, 153–166, https://doi.org/10.1016/j.physd.2003.10.011.

790 Kingma, D. P., and J. Ba, 2017: Adam: A Method for Stochastic Optimization.
791 https://doi.org/10.48550/arXiv.1412.6980.

792 Larson, S. M., and B. P. Kirtman, 2014: The Pacific Meridional Mode as an ENSO Precursor
793 and Predictor in the North American Multimodel Ensemble. *Journal of Climate*, **27**,
794 7018–7032, https://doi.org/10.1175/JCLI-D-14-00055.1.

795 Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*,
796 **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

797 ——, 1969a: Atmospheric Predictability as Revealed by Naturally Occurring Analogues.
798 *Journal of the Atmospheric Sciences*, **26**, 636–646, https://doi.org/10.1175/1520-
799 0469(1969)26<636:APARBN>2.0.CO;2.

800 ——, 1969b: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**,
801 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.

802 Lou, J., M. Newman, and A. Hoell, 2023: Multi-decadal variation of ENSO forecast skill
803 since the late 1800s. *npj Clim Atmos Sci*, **6**, 1–14, https://doi.org/10.1038/s41612-023-
804 00417-z.

805 Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the Fidelity of
806 Explainable Artificial Intelligence Methods for Applications of Convolutional Neural
807 Networks in Geoscience. *Artificial Intelligence for the Earth Systems*, **1**,
808 https://doi.org/10.1175/AIES-D-22-0012.1.

809 McPhaden, M. J., 2003: Tropical Pacific Ocean heat content variations and ENSO persistence
810 barriers. *Geophysical Research Letters*, **30**, https://doi.org/10.1029/2003GL016872.

811 Meinen, C. S., and M. J. McPhaden, 2000: Observations of Warm Water Volume Changes in
812 the Equatorial Pacific and Their Relationship to El Niño and La Niña. *Journal of*
813 *Climate*, **13**, 3551–3559, https://doi.org/10.1175/1520-
814 0442(2000)013<3551:OOWWVC>2.0.CO;2.

815 Menkes, C. E., M. Lengaigne, J. Vialard, M. Puy, P. Marchesiello, S. Cravatte, and G.
816 Cambon, 2014: About the role of Westerly Wind Events in the possible development
817 of an El Niño in 2014. *Geophysical Research Letters*, **41**, 6476–6483,
818 https://doi.org/10.1002/2014GL061186.

819 Mulholland, D. P., P. Laloyaux, K. Haines, and M. A. Balmaseda, 2015: Origin and Impact
820 of Initialization Shocks in Coupled Atmosphere–Ocean Forecasts. *Monthly Weather*
821 *Review*, **143**, 4631–4644, https://doi.org/10.1175/MWR-D-15-0076.1.

822 Newman, M., and P. D. Sardeshmukh, 2017: Are we near the predictability limit of tropical
823 Indo-Pacific sea surface temperatures? *Geophysical Research Letters*, **44**, 8520–8529,
824 https://doi.org/10.1002/2017GL074088.

825 Oktay, O., and Coauthors, 2018: Attention U-Net: Learning Where to Look for the Pancreas.
826 https://doi.org/10.48550/arXiv.1804.03999.

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

830 Petersik, P. J., and H. A. Dijkstra, 2020: Probabilistic Forecasting of El Niño Using Neural
831 Network Models. *Geophysical Research Letters*, **47**, e2019GL086423,
832 https://doi.org/10.1029/2019GL086423.

833 Rader, J. K., and E. A. Barnes, 2023: Optimizing Seasonal-To-Decadal Analog Forecasts
834 With a Learned Spatially-Weighted Mask. *Geophysical Research Letters*, **50**,
835 e2023GL104983, https://doi.org/10.1029/2023GL104983.

836 Risbey, J. S., and Coauthors, 2021: Standard assessments of climate forecast skill can be
837 misleading. *Nat Commun*, **12**, 4346, https://doi.org/10.1038/s41467-021-23771-z.

838 Rodgers, K. B., and Coauthors, 2021: Ubiquity of human-induced changes in climate
839 variability. *Earth System Dynamics*, **12**, 1393–1411, https://doi.org/10.5194/esd-12-
840 1393-2021.

841 Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for
842 Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted*
843 *Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W.M. Wells, and A.F. Frangi,
844 Eds., *Lecture Notes in Computer Science*, Cham, Springer International Publishing,
845 234–241.

846 Roulston, M. S., and J. D. Neelin, 2000: The response of an ENSO Model to climate noise,
847 weather noise and intraseasonal forcing. *Geophysical Research Letters*, **27**, 3723–
848 3726, https://doi.org/10.1029/2000GL011941.

849 Rudin, C., 2019: Stop explaining black box machine learning models for high stakes
850 decisions and use interpretable models instead. *Nat Mach Intell*, **1**, 206–215,
851 https://doi.org/10.1038/s42256-019-0048-x.

852 Shin, N.-Y., Y.-G. Ham, J.-H. Kim, M. Cho, and J.-S. Kug, 2022: Application of Deep
853 Learning to Understanding ENSO Dynamics. *Artificial Intelligence for the Earth*
854 *Systems*, **1**, https://doi.org/10.1175/AIES-D-21-0011.1.

855 Shin, S.-I., P. D. Sardeshmukh, M. Newman, C. Penland, and M. A. Alexander, 2021: Impact
856 of Annual Cycle on ENSO Variability and Predictability. *Journal of Climate*, **34**,
857 171–193, https://doi.org/10.1175/JCLI-D-20-0291.1.

858 Taschetto, A. S., C. C. Ummenhofer, M. F. Stuecker, D. Dommenget, K. Ashok, R. R.
859 Rodrigues, and S.-W. Yeh, 2020: ENSO Atmospheric Teleconnections. *El Niño*
860 *Southern Oscillation in a Changing Climate*, American Geophysical Union (AGU),
861 309–335.

862 Van den Dool, H. M., 1989: A New Look at Weather Forecasting through Analogues.
863 *Monthly Weather Review*, **117**, 2230–2247, https://doi.org/10.1175/1520-
864 0493(1989)117<2230:ANLAWF>2.0.CO;2.

865 Vimont, D. J., M. A. Alexander, and M. Newman, 2014: Optimal growth of Central and East
866 Pacific ENSO events. *Geophysical Research Letters*, **41**, 4027–4034,
867 https://doi.org/10.1002/2014GL059997.

868 Wang, C., 2019: Three-ocean interactions and climate variability: a review and perspective.
869 *Clim Dyn*, **53**, 5119–5136, https://doi.org/10.1007/s00382-019-04930-x.

870 Wilks, D. S., 2016: "The Stippling Shows Statistically Significant Grid Points": How
871 Research Results are Routinely Overstated and Overinterpreted, and What to Do
872 about It. *Bulletin of the American Meteorological Society*, **97**, 2263–2273,
873 https://doi.org/10.1175/BAMS-D-15-00267.1.

874 You, Y., and J. C. Furtado, 2018: The South Pacific Meridional Mode and Its Role in
875 Tropical Pacific Climate Variability. *Journal of Climate*, **31**, 10141–10163,
876 https://doi.org/10.1175/JCLI-D-17-0860.1.

877 Zhang, H., A. Clement, and P. D. Nezio, 2014: The South Pacific Meridional Mode: A
878 Mechanism for ENSO-like Variability. *Journal of Climate*, **27**, 769–783,
879 https://doi.org/10.1175/JCLI-D-13-00082.1.

880 Zhou, L., and R.-H. Zhang, 2023: A self-attention–based neural network for three-
881 dimensional multivariate modeling and its skillful ENSO predictions. *Science*
882 *Advances*, **9**, eadf2827, https://doi.org/10.1126/sciadv.adf2827.

883 Zuo, H., M. A. Balmaseda, S. Tietsche, K. Mogensen, and M. Mayer, 2019: The ECMWF
884 operational ensemble reanalysis–analysis system for ocean and sea ice: a description
885 of the system and assessment. *Ocean Science*, **15**, 779–808,
886 https://doi.org/10.5194/os-15-779-2019.

887