JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2024MS004422

Reconstructing the Tropical Pacific Upper Ocean Using Online Data Assimilation With a Deep Learning Model

Zilu Meng[1] and **Gregory J. Hakim**[1]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA

Key Points:

- A deep learning (DL) model exhibits superior prediction skill in the tropical Pacific compared to a linear inverse model
- Data assimilation on a sparse network of observations accurately reconstructs the monthly upper ocean spatial fields
- The superior reconstruction skill of the DL model stems from its enhanced prediction skill

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. Meng,
zilumeng@uw.edu

Citation:

Meng, Z., & Hakim, G. J. (2024). Reconstructing the tropical Pacific upper ocean using online data assimilation with a deep learning model. *Journal of Advances in Modeling Earth Systems*, *16*, e2024MS004422. https://doi.org/10.1029/2024MS004422

Received 30 APR 2024
Accepted 19 SEP 2024

Abstract A deep learning (DL) model, based on a transformer architecture, is trained on a climate-model data set and compared with a standard linear inverse model (LIM) in the tropical Pacific. We show that the DL model produces more accurate forecasts compared to the LIM when tested on a reanalysis data set. We then assess the ability of an ensemble Kalman filter to reconstruct the monthly averaged upper ocean from a noisy set of 24 sea-surface temperature observations designed to mimic existing coral proxy measurements, and compare results for the DL model and LIM. Due to signal damping in the DL model, we implement a novel inflation technique by adding noise from hindcast experiments. Results show that assimilating observations with the DL model yields better reconstructions than the LIM for observation averaging times ranging from 1 month to 1 year. The improved reconstruction is due to the enhanced predictive capabilities of the DL model, which map the memory of past observations to future assimilation times.

Plain Language Summary We use a deep learning (DL) model to better predict climate patterns in the tropical Pacific upper ocean, and to reconstruct past conditions from a sparse network of noisy observations. The DL model forecasts are more accurate than a reference Linear Inverse Model (LIM), which has approximately comparable computational demand. After we adjust DL model forecasts to better approximate errors, we show that this model can more accurately reconstruct climate fields than the LIM. This success highlights the significant potential of DL to improve our understanding and prediction of climate change through reconstructing climate variables from sparse information such as from coral proxies.

1. Introduction

Owing to the limited time span of satellite observations, our understanding of climate variability on interdecadal and longer timescales derives mainly from climate model simulations and reconstructions from paleoclimate proxies. For example, the El Niño Southern Oscillation (ENSO), one of the most significant drivers of interannual variability in the Earth's system, has a profound impact on the global climate and robust teleconnections (Cane et al., 1986; McPhaden et al., 2006; Timmermann et al., 2018). However, due to the short period of satellite observations, our understanding of ENSO variability, such as in its intensity, spatial distribution of temperature anomalies, and its interactions with other climate phenomena, is uncertain and poorly sampled (Ashok et al., 2007; D'Arrigo et al., 2005; Kug et al., 2009). To address this issue, a longer time span of data is needed to better understand these variations in ENSO (Timmermann et al., 2018). Here we assess the potential of a deep-learning model of the tropical Pacific ocean for assimilating a sparse network of noisy sea-surface temperature (SST) observations, with the goal of reconstructing past climate states from coral proxies of SST. This work provides proof of concept for this goal, and a benchmark comparison to similar approaches using linear inverse models (LIMs).

Reconstructions using paleoclimate proxy data (e.g., De Maesschalck et al., 2000; Mann et al., 1998), such as tree rings, ice core, and coral archives provide insights into past climate conditions. A major challenge with these data sources is their uneven spatial and temporal distribution, which complicates interpretation of signals in climate studies based on multiple proxies. To overcome this issue, the method of climate field reconstruction (CFR) has been used to combine information from different proxies (e.g., Mann et al., 1998). The primary objective of this approach is to use statistical methods to reconstruct regularly gridded climate fields from sparse and unevenly distributed paleoclimate data. This facilitates studies of global or regional climate variability and may significantly aid in understanding and predicting future climate change.

Recent studies have employed an objective framework for CFR based on data assimilation (DA) (e.g., Goosse et al., 2010; Hakim et al., 2016; Valler et al., 2024). One notable difference in the DA approach to climate

reconstruction as compared to weather prediction concerns the use of a model to generate the prior ("first guess") before assimilation. A key challenge is that the high operational costs of climate models render ensemble forecasting impractical, which has led to the use of "offline" DA approaches that randomly sample existing climate model simulations. This approach of offline assimilation has considerable advantages: it requires dramatically lower computational costs; random sampling from the outputs of climate model runs allows for a better estimation of the uncertainty in the assimilation results; and it can yield better outcomes, especially when the predictive capacity of climate models is limited. One key limitation of the offline method is its lack of memory in the prior. Specifically, when assimilating data at a given time step, the method relies solely on concurrent observations, thus omitting past and future observations. This is particularly disadvantageous in contexts like the ENSO, where the system's memory extends approximately beyond a year. Such historical context could significantly enhance the accuracy of DA but is unexploited in offline setups. This drawback has driven our exploration of an online approach, which we believe offers a more robust framework by incorporating a broader temporal spectrum of observational data.

Progress toward computationally feasible online paleoclimate DA has been demonstrated using a LIM (Penland & Magorian, 1993) as an emulator for climate models in the forecast step (Perkins & Hakim, 2017, 2020). This approach transfers climate information from one time step to the next, provides superior priors and more effective use of sparse proxy data. This improvement in assimilation is primarily attributed to the coupled dynamics of the ocean–atmosphere system (Perkins & Hakim, 2017), since the largest sources of proxy data, such as tree rings and ice cores (PAGES2k Consortium et al., 2017), are primarily located on continents and largely reflect atmospheric variability. Using DA with a skillful coupled atmosphere–ocean model allows for this atmospheric information to inform oceanic state estimates. Given that the predictability of the ocean is substantially higher than that of the atmosphere, using the LIM effectively transmits information through the ocean's memory onto the atmosphere, which benefits proxy assimilation at later times.

Recently, the emergence of DL provides a new approach (e.g., LeCun et al., 2015; Reichstein et al., 2019) to computationally efficient online DA. Through complex network architectures, DL can fit nonlinear relationships in data, thereby enabling more accurate predictions of future states. Moreover, many DL models have demonstrated predictive capabilities that exceed those of LIMs, and even conventional Coupled Generak Circulation Models (GCMs) (Gao et al., 2023; Ham et al., 2019; Sun et al., 2023; Zhang et al., 2024; Zhou & Zhang, 2023). For instance, simple Convolutional Neural Networks (CNNs) have shown remarkable success in predicting the time series of ENSO events with 17 months lead time (Ham et al., 2019), surpassing the best seasonal prediction models. However, neural networks in this work (Ham et al., 2019) are designed to predict a single variable, like Nino3.4 Index, not a field of variables. Therefore, to predict spatial fields for DA, we need to select network architectures that are field-to-field. Recently, the work by Zhou and Zhang (2023) introduced a neural network based on a transformer with self-attention architecture (Vaswani et al., 2017), which they call 3D-Geoformer. The 3D-Geoformer effectively forecasts monthly ocean temperature anomalies in the upper 150 m and surface wind stress fields of the tropical Pacific to produce ENSO forecasts having skill comparable to those of the aforementioned CNNs. Here we test the use of a version of this model in forecasting and DA experiments to provide proof-of-concept for use in paleoclimate DA. Although this work is motivated by paleoclimate DA, we note that the methods outlined here are also broadly applicable to assimilating instrumental observations.

The remainder of the paper is organized as follows. Section 2 details the data and methodologies employed in the construction of the LIM and DL models, as well as the DA method. Section 3 delineates the sparse observational network utilized for the DA experiments. Comparative analyses of forecasting performance between the LIM and DL models are presented in Section 4, with the DA experiments elaborated in Section 5. Finally, Section 6 offers a discussion of the findings and draws conclusions.

2. Data, Models, and Data Assimilation Methods

Here we provide a detailed description of the data and methods used in our study, including the theory and training procedures for the models, DA techniques, observations, and methods to address the loss of ensemble variance in DL model forecasts.

Table 1
Data Set and Source

Type	Source	Period
Train set	Coupled Model Inter-comparison Project Phase 6 (CMIP6)	January 1850 to December 2014
Validation set	Simple Ocean Data Assimilation (SODA) products	January 1871 to December 1979
Test set	Global Ocean Data Assimilation System (GODAS) reanalysis	January 1980 to December 2021

2.1. Data

The focus of our research is the Tropical Pacific (shown in Figure 3), with an emphasis on the dynamics of ENSO, which is the dominant source of annual to interannual variability in this region (Cane et al., 1986; Timmermann et al., 2018). Consequently, we adopt variables that are integral to understanding ENSO dynamics, namely SST, surface zonal and meridional wind stress, and the temperature of the upper ocean on seven constant layers ranging from 5 to 150 m depth (5, 20, 40, 60, 90, 120, and 150 m). The geographical scope of the data extends from 90°E to 30°W and 20°N to 20°S. The zonal grid resolution is 2° and the meridional grid resolution is 0.5° (1°) between (poleward of) 5°S and 5°N.

We train our models on data from the Coupled Model Inter-comparison Project Phase 6 (CMIP6) historical experiments (O'Neill et al., 2016), and verify and test results on data from the Simple Ocean Data Assimilation (SODA) (Carton & Giese, 2008) products and the Global Ocean Data Assimilation System (GODAS) reanalysis. We note that the reanalysis data period spanned by combining GODAS and SODA covers a span of only about 130 years, which is somewhat limited for training a neural network with a large number of parameters. Consequently, we rely on the CMIP6 model data for training our network, consisting of 23 CMIP6 models (Figure 2). We then use SODA data for validation and to fine-tune the network architecture and hyper-parameters (described in the Section 2.2.3). The GODAS data set is used as the final test set for assessing model performance. Considering the documented spin-up issues and the scarcity of subsurface observations in the early periods (Xue et al., 2012), we have excluded the first 2 years of GODAS data from our analysis metrics. It is crucial to note that although there is a temporal overlap between the training and the testing & validation sets, the ENSO variability in the CMIP6 models differs significantly from those in the reanalysis data sets. The common elements among the models are external forcings, such as solar radiation and CO2 levels. This significant difference in temporal evolution ensures that the overlap does not result in overfitting. In this context, the training, validation, and testing data sets are derived from different sources, which helps to prevent overfitting and independently assess results. Summary information about these data sets is provided in Table 1.

2.2. Models

2.2.1. Linear Inverse Model (LIM)

Linear Inverse Modeling (LIM) is an efficient, widely applied, and powerful model for SST prediction and assimilation, especially for ENSO (e.g., Newman, 2013; Penland & Magorian, 1993; Penland & Sardeshmukh, 1995). The LIM is an empirically determined estimate of a dynamical system linearized about its mean state:

$$\frac{d\boldsymbol{x}}{dt} = \mathbf{L}\boldsymbol{x} + \boldsymbol{\xi}, \tag{1}$$

in which $\boldsymbol{x}$ is the state vector, typically cast in terms of a truncated set of Empirical Orthogonal Functions (EOFs), $\mathbf{L}$ is the linear system operator, and $\boldsymbol{\xi}$ represents noise that is white in time but correlated in the state variables. $\mathbf{L}\boldsymbol{x}$ represent the deterministic tendency of the system, and ξ a stochastic forcing that represents the net effect of nonlinearity and unresolved processes. Integrating (Equation 1) over $t = 0 : \tau$, for any τ, and taking the expected value, gives

$$\boldsymbol{x}(\tau) = \mathbf{G}_{\tau}\boldsymbol{x}(0), \tag{2}$$

where $\mathbf{G}_\tau = \exp(\mathbf{L}\tau)$. Given sample training data, $\mathbf{G}_\tau$ may be determined for a single τ from regular least-squares regression:

$$\mathbf{G}_\tau = \mathbf{C}(\tau)\mathbf{C}(0)^{-1}. \tag{3}$$

here $\mathbf{C}(\tau)$ is the lag covariance matrix of the state vector at lag time τ,

$$\mathbf{C}(\tau) = <\boldsymbol{x}(\tau)\boldsymbol{x}^T(0)>$$

and "<>" represents a sample average. Matrix $\mathbf{L}$ is then determined from $\mathbf{G}_\tau$. The stochastic noise ξ is assumed to be a white noise process with a covariance matrix $\mathbf{Q}$, meaning $<\boldsymbol{\xi}\boldsymbol{\xi}^T> = \mathbf{Q}$. Assuming stationary statistics, matrix $\mathbf{Q}$ is constant and defined by

$$\frac{d\mathbf{C}(0)}{dt} = \mathbf{L}\mathbf{C}(0) + \mathbf{C}(0)\mathbf{L}^T + \mathbf{Q} = 0. \tag{4}$$

Forecasts are computed using

$$\boldsymbol{x}_{t+\delta t} = (\mathbf{L}\delta t + \mathbf{I})\boldsymbol{x}_t + \hat{\boldsymbol{Q}}\sqrt{\boldsymbol{\Lambda}\delta t}\boldsymbol{\alpha} \tag{5}$$

$$\boldsymbol{x}_{t+\delta t/2} = \frac{1}{2}(\boldsymbol{x}_{t+\delta t} + \boldsymbol{x}_t), \tag{6}$$

in which $\boldsymbol{\alpha}$ is a vector of independent standard normal random variables, $\boldsymbol{\Lambda}$ and $\hat{\mathbf{Q}}$ are the eigenvalues and eigenvectors of $\mathbf{Q}$, respectively. In this study, τ is set to 1 month for the monthly DA and δt is set to 6 hr. In this context, the LIM can be viewed as a linear system driven by spatially correlated, temporally white, noise, representing nonlinear and unresolved fast processes. More details of the LIM can be found in Penland and Magorian (1993) and Newman et al. (2011).

2.2.2. Deep Learning Model (DL)

Zhou and Zhang (2023) introduce a novel self-attention-based neural network specifically designed for predicting the tropical Pacific upper ocean. The model is structured to take a 12-month state vector as input, with fields including SST, surface wind stresses, and upper ocean temperature on 7 vertical levels. The output of this model is a state vector for the subsequent 12 months, which can be autoregressively extended indefinitely into the future:

$$\boldsymbol{X}^{\text{out}}_{t+1:t+12} = \boldsymbol{DL}\left(\boldsymbol{X}^{\text{in}}_{t-12:t}\right), \tag{7}$$

in which $\boldsymbol{X}^{\text{in}}_{t-12:t}$ is the input state vector from time $t-12$ to t, $\boldsymbol{X}^{\text{out}}_{t+1:t+12}$ denotes the forecasted state vector covering the period from time $t+1$ to $t+12$, and $\boldsymbol{DL}$ is the DL model operator.

This model initially employs an embedding layer to encode the data into smaller blocks based on latitude and longitude, into a vector of length 256. Subsequently, it utilizes temporal self-attention and spatial self-attention modules to extract features. Finally, the model outputs the predicted results through a fully connected layer. More details of the model can be found in Zhou and Zhang (2023).

Compared with the LIM, there are two main advantages of DL model. First, the DL model can capture the deterministic non-linear relationships between the current and future states. Second, the DL model employs nonlinear dimensional reduction, which retains more predictive capability. As we will show, the DL model has better prediction skill than the LIM (see Section 4 for details).

2.2.3. Model Training and Configuration

LIM. The first step in LIM training employs multivariate EOFs (e.g., Hannachi et al., 2007; Lorenz, 1956). This is achieved by first consolidating the normalized variables—Sea Surface Temperature (SST), wind stress, and ocean temperature—into a unified matrix $\mathbf{X}$:

$$\mathbf{X} = \left[\frac{\mathbf{X}_{SST}^{T}}{\sigma_{SST}}, \frac{\mathbf{X}_{\text{wind stress}}^{T}}{\sigma_{\text{wind stress}}}, \frac{\mathbf{X}_{\text{ocean temperature}}^{T}}{\sigma_{\text{ocean temperature}}}\right]^{T} \tag{8}$$

here, σ represents the standard deviation of each variable. Subsequently, Singular Value Decomposition (SVD) is employed to extract the leading modes $\mathbf{U}$, such that $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{T}$. The fields are then projected into this dimensionally reduced Principal Component (PC) space using: $\hat{\mathbf{X}} = \hat{\mathbf{U}}^{\mathbf{T}}\mathbf{X}$.

We then use this PC time sequences $(\hat{\mathbf{X}})$ to construct the LIM from Equation 3. To ensure that the LIM achieves its best forecasting ability, we perform dimensional reduction on the data from each of the 23 CMIP6 models and conduct an exhaustive search to identify the optimal number of PCs for training a LIM from each model. The search objective is to identify the highest 12-month mean Nino3.4 Index forecast correlation skill for each model. We find that the first 12 PC sequences from the GFDL-CM4 model yield the best forecasting results of the 23 LIMs evaluated. Therefore, we use the first 12 PC sequences from the GFDL-CM4 model (Adcroft et al., 2019) to build our LIM. It is noteworthy that previous studies have constructed LIMs using only SST or a combination of SST and Sea Surface Height (SSH) (Lou et al., 2020; Shin et al., 2020), rather than also predicting the mixed layer as we do here. Thus, to provide a basis for comparison with previous work, we also construct a LIM using only SST, which we refer to as LIMOsst. A comprehensive search again reveals that the LIM built with the first 12 PC sequences of GFDL-CM4 SST yields the best forecasting performance.

DL Model. As our objective is to predict the state of the entire field for the next time step, we have modified the loss function of the original model described in (Zhou & Zhang, 2023). Instead of incorporating the Root Mean Squared Error (RMSE) of the Nino3.4 index into the loss function as in the original study (Zhou & Zhang, 2023), we have adopted the RMSE of the entire state vector as our loss function:

$$\text{Loss} = \frac{1}{T_{\text{out}}}\sum_{t=1}^{T_{\text{out}}}\sqrt{\frac{1}{N_{\text{lat}} \times N_{\text{lon}} \times C}\sum_{i=1}^{N_{\text{lon}}}\sum_{j=1}^{N_{\text{lat}}}\sum_{k=1}^{C}\left(x_{t,k,j,i}^{\text{out}} - x_{t,k,j,i}^{\text{true}}\right)^{2}}, \tag{9}$$

where x^{out} represents the forecast states, which equals 12 in this research, while T_{out}, N_{lat}, and N_{lon} denote the number of output time steps, the number of latitude grids, and the number of longitude grids, respectively. Although this modification led to a slight decrease in Nino3.4 index prediction skill, it improves the Nino3.4 index DA skill by 5%. This adjustment aligns with our broader goal of enhancing DA through more robust covariance relationships, reflecting teleconnections across the entire field.

For training we use the PyTorch framework (Paszke et al., 2019) and source code from Zhou and Zhang (2023). To achieve optimal tuning of the network parameters, we use the Adam optimization algorithm (Kingma & Ba, 2014), implementing a decaying learning rate strategy that starts at 0.0005 and decreases with ongoing training. Moreover, we incorporate an early stopping mechanism (Prechelt, 2002) that halts training if there is no reduction in the validation RMSE over four consecutive epochs. All training strategies are detailed with the code accompanying this study (see Section 6).

2.3. Data Assimilation Methods

This section describes the assimilation methods used in this study, including augmenting the error of the DL model ensemble forecasts.

2.3.1. Ensemble, Online and Offline Assimilation

We perform DA using an Ensemble Kalman Filter (EnKF) (Evensen, 2009), which has been shown to perform well in paleo-data assimilation tasks (Hakim et al., 2016; Perkins & Hakim, 2017; Tardif et al., 2019; Zhu et al., 2023). The first part of the Kalman filter is the update step:

$$\boldsymbol{x}_a = \boldsymbol{x}_p + \mathbf{K}[\,\boldsymbol{y} - \mathcal{H}(\boldsymbol{x}_p)], \tag{10}$$

in which $\boldsymbol{x}_a$ is the analysis state vector, $\boldsymbol{x}_p$ is the prior state vector, $\boldsymbol{y}$ is the observation vector, $\mathcal{H}$ is the observation operator, and $\mathbf{K}$ is the Kalman gain matrix defined by:

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathbf{T}}\left[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathbf{T}} + \mathbf{R}\right]^{-1}, \tag{11}$$

here, $\mathbf{B}$ is the prior covariance matrix, $\mathbf{H}$ is a linearized version of $\mathcal{H}$, and $\mathbf{R}$ is the observation error covariance matrix. In the Last Millennium Reanalysis (Hakim et al., 2016, LMR) framework, $\mathbf{R}$ is a diagonal matrix with diagonal elements equal to the observational error variance. In this study, we use the Ensemble Square Root Filter (EnSRF) method (Tippett et al., 2003) to solve Equations 10 and 11, including serial observation processing. Since the covariance spatial length scales on annual to monthly time scales in this region are relatively long, we do not impose covariance localization. The EnSRF method for the kth proxy, y_k, is described in the following equations. We separate the ensemble into two parts, the ensemble mean and ensemble perturbation. First, we calculate the analysis ensemble mean by:

$$\overline{x_a} = \overline{x_p} + \frac{\mathrm{cov}(\boldsymbol{x}_p, y_{e,k})}{\left(\mathrm{var}(y_{e,k}) + R_k\right)}\left(y_k - \overline{y_{e,k}}\right). \tag{12}$$

here the overbar $(\overline{x})$ denotes an ensemble mean, and primes (x') denotes an ensemble perturbation. Subscript "p" denotes the prior state vector (forecast), subscript "a" the analysis state vector (posterior), $\boldsymbol{y}_{e,k}$ is the kth proxy estimate from the ensemble, and R_k is the kth proxy error variance. The "var" and "cov" are the variance and covariance operators on the ensemble number dimension. Second, we calculate the analysis ensemble perturbations from:

$$\boldsymbol{x}'_a = \boldsymbol{x}'_b - \left[1 + \sqrt{\frac{R_k}{\mathrm{var}(y_{e,k}) + R_k}}\right]^{-1} \frac{\mathrm{cov}(\boldsymbol{x}_p, y_{e,k})}{\left(\mathrm{var}(y_{e,k}) + R_k\right)}\left(y'_{e,k}\right) \tag{13}$$

Finally, each ensemble member analysis state vector is calculated by adding the ensemble mean and ensemble perturbation:

$$\boldsymbol{x}_a = \overline{\boldsymbol{x}_a} + \boldsymbol{x}'_a \tag{14}$$

One objective of this paper is to demonstrate that low-frequency observations (averaged more than 1 month) can be effectively assimilated to reconstruct monthly averaged climate fields. By employing the method mentioned previously, we can update the monthly variables using 3, 6, or 12-month averaged observations through the covariance matrix $\mathrm{cov}(\boldsymbol{x}_p, y_{e,k})$.

To assess the impact of the models on the DA results, we perform experiments using both online and offline DA, as subsequently described. All experiments in this paper use 100 ensemble members, a number chosen to balance good results with the constraints of limited computational resources. Results for the DL model improve modestly for larger ensembles (tests for 50–400 members shown in Figure S3 in Supporting Information S1).

Online Assimilation. After the update step, in the online assimilation method, we perform the forecast step, meaning a forecast initialized with the result of the update step:

$$\boldsymbol{x}_{p,t+1} = \mathcal{M}(\boldsymbol{x}_{a,t}), \tag{15}$$

where $\mathcal{M}$ represents the model operator. We employ both the LIM and the DL model as the model operators. Additionally, we evaluate the effect of different observation-averaging periods on the analyses, since these vary by climate proxy. Specifically, we consider time averages including 1, 3, 6, and 12 months. Consequently, when assimilating observations, we adapt the model's forward prediction to align with the temporal length of the averaging time of the observations. This allows us to compute the prior values for the proxy data, which are subsequently assimilated using method described above (Huntley & Hakim, 2010; Steiger et al., 2014). To initialize the ensemble at the start of each experiment, we randomly select 100 model output data fields from the

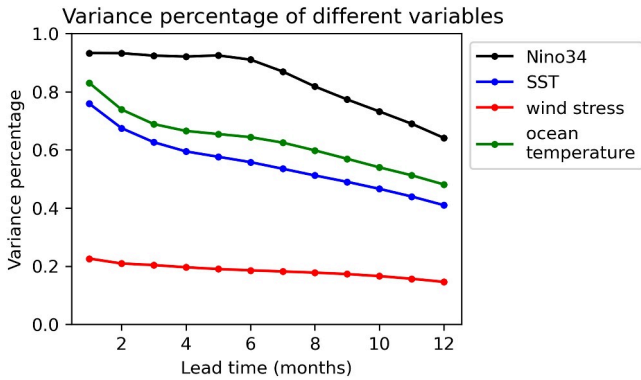

Figure 1. Variance proportion of deep learning (DL) forecasts compared to Global Ocean Data Assimilation System (GODAS) observations across different variables over time. Illustrated here are the variance ratios for predictions made by the DL model relative to actual observations from the GODAS data set, covering variables such as the Nino3.4 Index, sea-surface temperature, wind stress, and ocean temperature, as a function of varying lead times in months.

CMIP6 data that correspond to the current month. After completing assimilation for this randomly drawn ensemble, we make a forecast using this assimilated field as initial condition to the next time for assimilation. This forecast–assimilation cycle continues until the final time with observations.

Offline Assimilation. In the original LMR framework (Hakim et al., 2016; Tardif et al., 2019), assimilation is executed offline. In the offline case, the prior state vector comprises the same random ensemble drawn from a single CMIP6 model, with no intervening forecast step. To optimize the performance of this offline assimilation, we conduct the assimilation process using data from all 23 CMIP6 models as the prior ensemble and select the most accurate outcome as our final result, as defined by the highest reconstructed Nino3.4 Index correlation. We find that the MRI-ESM2-0 simulation is the optimal source for the offline prior ensemble. This approach ensures that all experiments can be used to validate the relative outperformance of the Deep-Learning–assimilation method for reconstructing tropical Pacific climate fields.

2.3.2. DL Model Ensemble Inflation

In the case of DL networks, particularly for tasks involving monthly or annual forecasting, DL models tend to lose error associated with the unpredictable signal. For example, when forecasts are initialized and verified using the GODAS data set, the ratio of the SST variance between the predictions of the DL model and the target data in the GODAS data set is less than 1. This ratio decreases with increasing prediction lead time, as illustrated by the blue line in Figure 1. The variance of the forecast wind stress is also significantly lower than that of GODAS, and the variance ratio of other variables are similarly smaller. The primary reason for this discrepancy is that the DL model is not trained to capture unpredictable signals, which is especially evident in the atmosphere as seen in Figure S1 in Supporting Information S1. This unpredictable noise may be caused by unresolved processes, or by signals that originate outside of the forecast domain, such as the Pacific Meridional Mode (PMM) (Meng & Li, 2024; Vimont et al., 2003) from mid-latitudes, the Indian Ocean Dipole (IOD) (Saji et al., 1999), and from other ocean basin and deep-ocean dynamics.

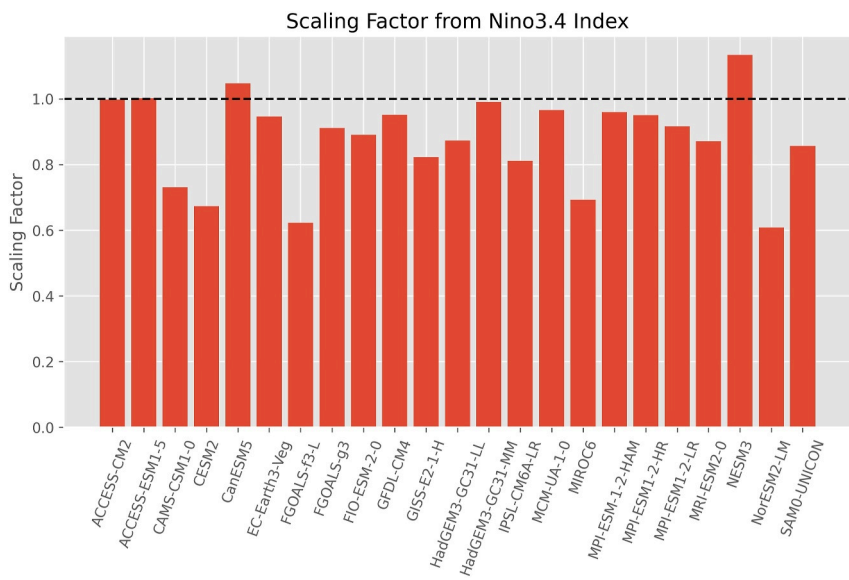

Figure 2. Comparison of standard deviation ratios: Simple Ocean Data Assimilation (SODA) versus CMIP6 Models for the Nino3.4 Index. This graph displays the ratio of the Nino3.4 index standard deviation from the SODA data set to that of various CMIP6 models, which is utilized as a scaling factor for noise based on the CMIP6 model data.

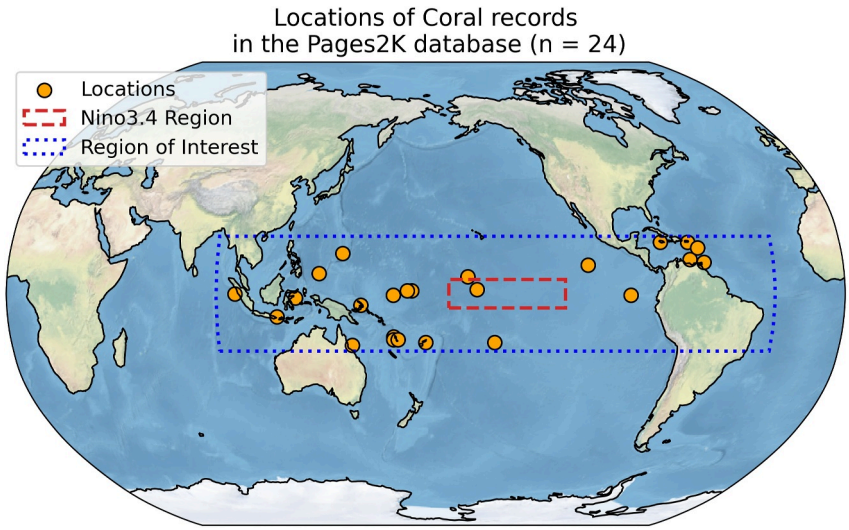

Figure 3. Geographical distribution of coral $\delta^{18}O$ proxy records in the tropical Pacific region from the PAGES2K database (Outlined by the dashed blue box). The dashed blue box delineates the area of focused modeling and interest, the dashed red box indicates the Nino3.4 region, and the brownish-yellow dots represent the locations of coral $\delta^{18}O$ proxy records.

While this is not an issue for forecasting tasks aimed at providing only the predictable future signal, it presents a challenge for DA since it involves estimating errors. Underestimating errors in the forecast can lead to under weighting the information from observations. To address this issue, we employ a variance inflation technique (Evensen, 2009) by adding random errors to the DL forecast. Specifically, we assume that the unpredicted error is a random process. We sample from an ensemble of these random vectors obtained by calculating the difference between the DL forecast and the verifying fields from hindcasting experiments. The sample size of GODAS and SODA data is not sufficient to obtain the statistical characteristics of this noise, so we use DL forecast errors from initializing and predicting on the CMIP6 models. Since there is a significant difference between the intensity of ENSO in CMIP6 and observations (SODA) (Beobide-Arsuaga et al., 2021), we use the standard deviation of the Nino3.4 Index (ENSO intensity) as a scaling factor on the errors from CMIP6 models hindcast. Specifically, the random errors added to the DL forecasts are calculated by the following steps.

First, we calculate the ratio between the standard deviation of SODA Nino3.4 index and each CMIP6 model's Nino3.4 index as the scaling factor α_i,

$$\alpha_i = \frac{\sigma_{SODA}}{\sigma_i}. \tag{16}$$

Here, σ_{SODA} is the standard deviation of the SODA Nino3.4 index and σ_i is the standard deviation of the i-th CMIP6 model's Nino3.4 index. The major reason for using Nino3.4 index as scale factor is that the Nino3.4 index standard deviation is representative of the intensity of ENSO. Ratios for most models are less than 1 (Figure 2), which means the variance of the CMIP6 models ENSO intensity is larger than in SODA. We apply the scaling factor to the forecast errors from CMIP6 models, $\boldsymbol{\eta}_{m,l,i}$, for the ith CMIP6 model, lth lead time and mth ensemble member:

$$\boldsymbol{\eta}_{m,l,i} = \alpha_i\left(\boldsymbol{x}^{\text{true}}_{m,l,i} - \boldsymbol{x}^{\text{out}}_{m,l,i}\right). \tag{17}$$

The corrected forecast from the DL model is then defined by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}^{\text{out}}_{m,l} + \boldsymbol{\eta}_{m,l}. \tag{18}$$

We note that, in comparison to adding noise directly, scaling the noise leads to an approximately 5% improvement in the reconstruction results in terms of correlation. However, the scaling method cannot correct the large

discrepancies in the power spectral density of the Nino3.4 index (Brown et al., 2020); which is a subject for future research.

On the contrary, in the LIM, the variance of the deterministic part of the state vector also decays with time, because the real part of the eigenvalues of $\mathbf{L}$ are less than 0. However, the random noise forcing component $\boldsymbol{\xi}$ yields an unbiased forecast covariance (as seen in Equation 4). Therefore, the variance inflation technique is not used for the LIM forecasts.

2.3.3. Evaluation Criteria

The major evaluation criteria used in this study for the prediction skill and reconstruction skill are sample time-series correlation and root-mean-squared error (RMSE). The correlation is calculated by the following equation:

$$\text{corr} = \frac{1}{T}\sum_{t=1}^{T}\frac{\left(f_t - \bar{f}\right)\left(v_t - \bar{v}\right)}{\sigma_f \sigma_v}, \tag{19}$$

in which f_t is the prediction result or reconstruction result, v_t is the ground truth at time t, $\bar{f}$ and $\bar{v}$ are the mean of prediction result and ground truth, σ_f and σ_v are the standard deviation of prediction result and ground truth. RMSE is calculated by:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(f_t - v_t\right)^2}. \tag{20}$$

Correlation and RMSE results are averaged over the entire domain to provide summary measures of skill. To quantify the improvement of DL over the LIM and offline DA approaches, we use the improvement ratio (IR):

$$\text{IR} = \frac{S_{DL} - S_{traditional}}{S_{traditional}}, \tag{21}$$

in which S_{DL} and $S_{traditional}$ are the skill scores of the DL model and the traditional model (the LIM, or the offline method), respectively. The skill score is defined by first computing the domain-averaged correlation or RMSE, and then the IR. For consistency with RMSE, the IR is multiplied by -1, so that smaller values mean better skill.

3. Observing Network

Here we describe the design of the DA experiments, including the locations of the pseudoproxies, the types of pseudoproxies, the error characteristics for the pseudoproxies, and the experimental setup.

To simulate the real assimilation process as closely as possible, we use the locations of stable oxygen isotope composition $\left(\delta^{18}O\right)$ coral proxy locations from the widely utilized PAGES2K database. We take the average number of coral sites available in the tropical Pacific domain from 1600 to 2000, which amounts to 24, as the locations for the pseudoproxy data (Figure 3). In the assimilation of paleoclimate data, the proxy data averaging time varies from monthly to annual. Therefore, assimilating these proxies, which represent climatic data over periods longer than a month, to obtain monthly average climate data poses a significant challenge to the forecast model's ability to predict the duration of climate information. To test this capability, we set the duration of the pseudoproxy data averaging time to 1, 3, 6, and 12 months, respectively, and conduct separate assimilation experiments for each duration.

We take observations from the GODAS data set as ground truth, interpolating to the proxy location and time averaging:

$$y_{\text{avg},N} = \frac{1}{N}\sum_{i=k+1}^{k+N} y_i, \tag{22}$$

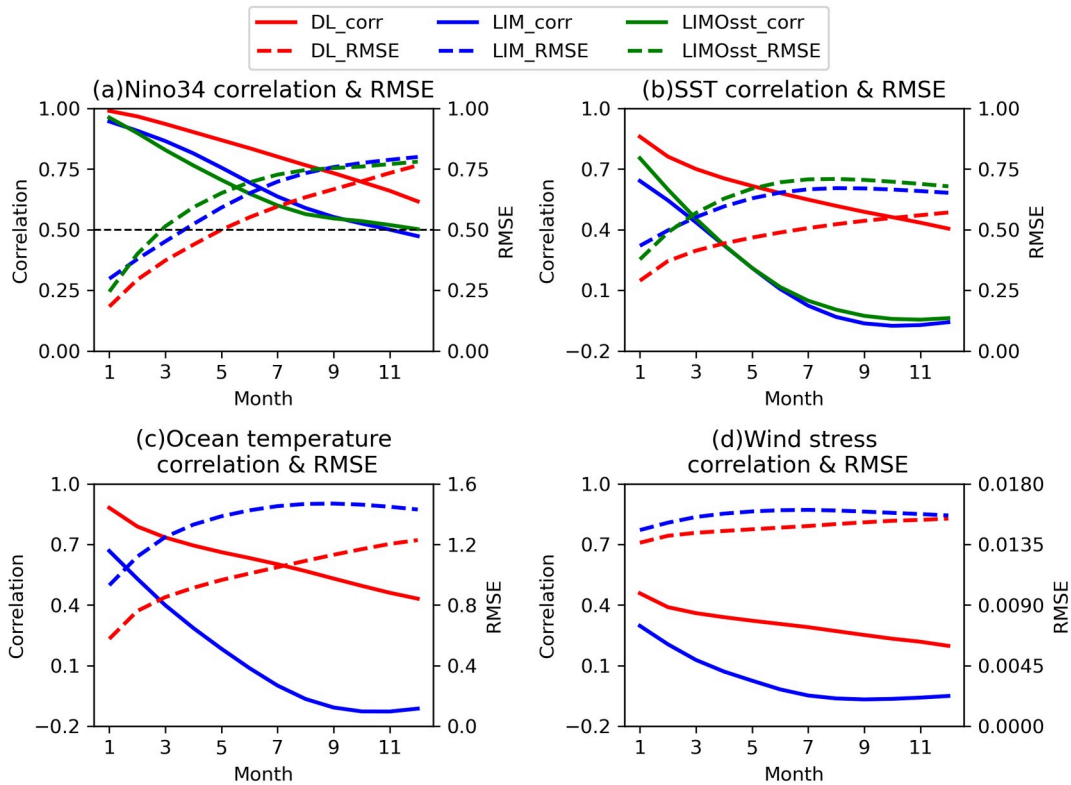

Figure 4. Forecast skill of Deep Learning model (red), Linear Inverse Model (blue) & LIMOsst (green) in terms of correlation (solid lines) and Root Mean Squared Error (RMSE) (dashed lines) across lead time for the Nino3.4 Index (a), sea-surface temperature field (b), ocean temperature field (c), and the wind stress field (d). The correlation scale is provided on the left *y*-axis and RMSE on the right *y*-axis, as the function of forecast lead time.

with N taking values of 1, 3, 6, and 12 in this study. To synchronize with ENSO seasonality, we compute 3-month averages corresponding to MAM (March–April–May), JJA (June–July–August), SON (September–October–November), and DJF (December–January–February). The 6-month averages are MAMJJA (March to August) and SONDJF (September to February), while the 12-month average spans MAMJJASONDJF (March to February).

After completing interpolation and averaging, we simulate random errors in the data drawn from a Gaussian distribution with a mean of 0. The Signal-to-Noise Ratio (SNR) (Zhu et al., 2023), defined in terms of standard deviation, is set to 1. Additionally, we tested the sensitivity to SNR by conducting additional experiments for SNR = 0.2, 0.5, 2, and 5, and find that the amplitude of the SNR does not significantly impact the results (Supporting Information S1). We simulate observation error by

$$y'_{\text{avg},N} = y_{\text{avg},N} + \zeta, \tag{23}$$

in which $\zeta \sim N(0,\sigma^2)$, and σ is the standard deviation of the real data divided by the SNR. The error simulation is performed for each proxy location and each time-averaging duration.

4. Forecasting Results

We now compare the forecast skill of DL, LIM and LIMOsst forecasts by initializing and verifying with the GODAS data set. In terms of the domain-averaged correlation and RMSE metric of all variables, and the Nino3.4 Index, the DL forecasts consistently outperform both the LIM and LIMOsst forecasts across all variables and at all lead times (Figure 4). Spatial maps of DL-forecast skill improvement reveal SST and ocean temperature spatial patterns similar to El-Niño (La-Niña) (Figures 5 and 6). Specifically, from 1-month to 12-month lead time, the region with improved predictions evolves from off the equator to the equatorial region, accompanied by ocean

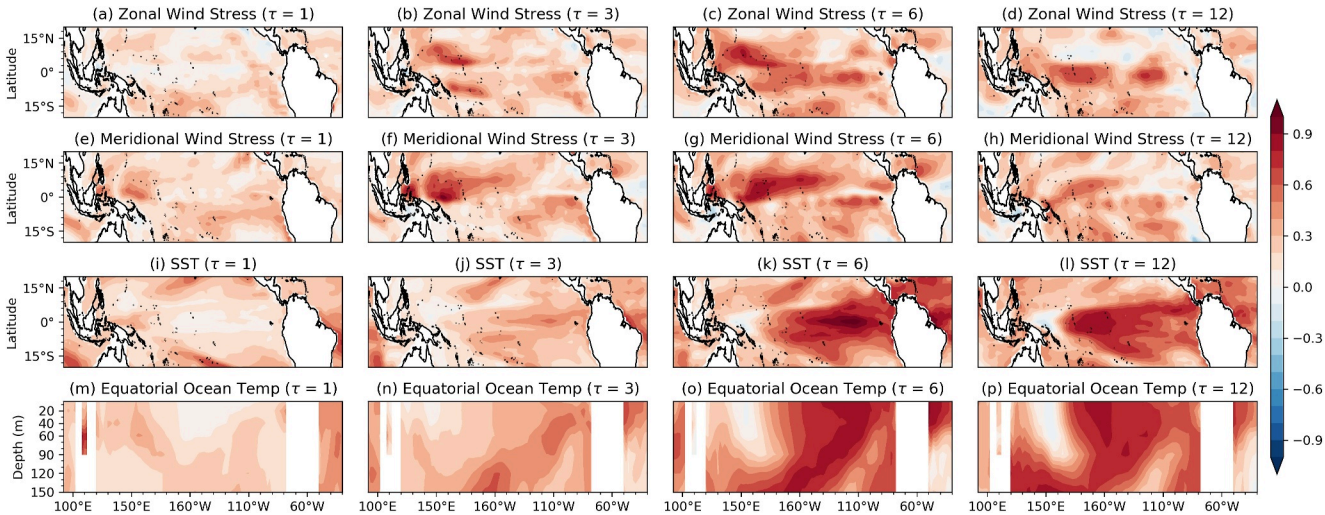

Figure 5. Difference in forecast skill as measured by correlation between Deep Learning and Linear Inverse Models as a function of lead time (τ). The first row (a–d) shows zonal-wind stress, the second row (e–h) meridional wind stress, the third row (i–l) sea-surface temperature, and fourth row (m–p) equatorial (5°N–5°S) ocean temperature.

temperature anomalies that tilt from the lower western to the upper eastern surface along the sloping thermocline (also as shown in the Figures S5–S8 in Supporting Information S1). For surface wind stress, the DL model forecast-skill improvement is located in the central and western equatorial Pacific region, aligning with the region of improved skill for ocean temperature. This indicates that the DL model is able to better simulate the dynamics of ENSO compared to the LIMs.

Compared to the correlation metric, the improvement in RMSE by the DL model is not as significant, as illustrated in Figure 4. An inherent advantage of the LIM at long lead times derives from the negative eigenvalues of **L**, trending the forecasts toward zero anomaly as lead time increases. This means that the RMSE converges on the climatological standard deviation. In contrast, the DL model lacks this constraint and exhibits systematic errors that increase the RMSE as shown in Figure S9 in Supporting Information S1, potentially to values larger than climatology.

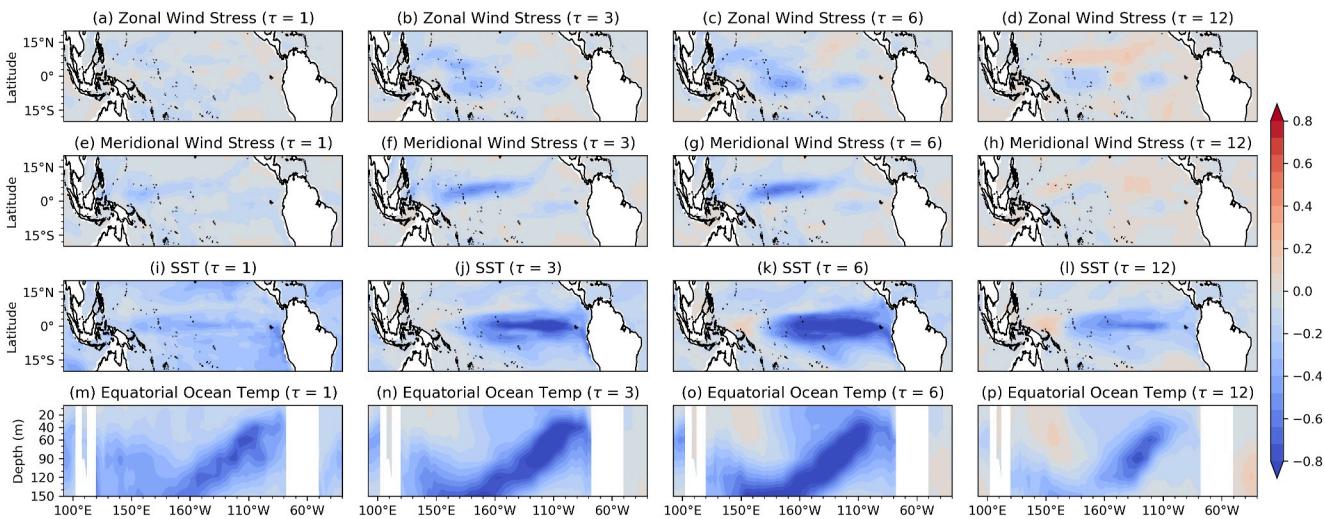

Figure 6. Comparison of spatial patterns in normalized Root Mean Squared Error (normalization by the domain-averaged standard deviation of the corresponding variables) skill between Deep Learning and Linear Inverse Models as a function of lead time (τ). The first row (a–d) shows zonal-wind stress, the second row (e–h) meridional wind stress, the third row (i–l) sea-surface temperature, and fourth row (m–p) equatorial (5°N–5°S) ocean temperature.

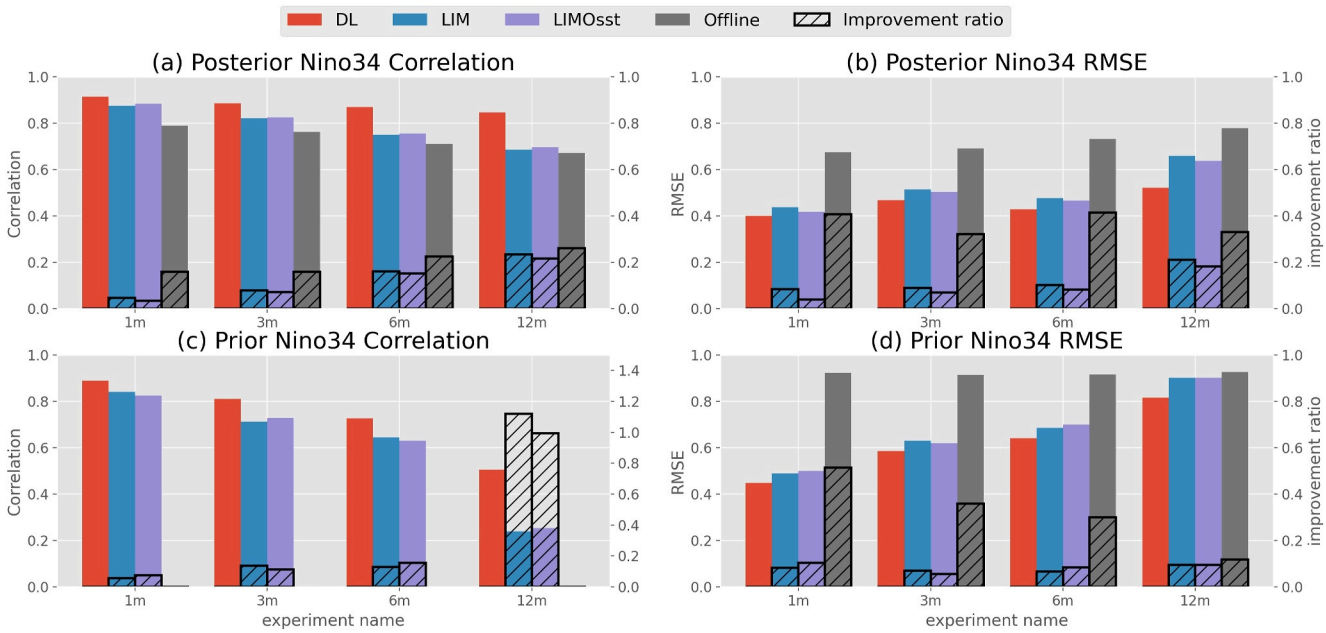

Figure 7. Correlation (left panel, left *y*-axis) and Root Mean Squared Error (right panel, left *y*-axis) for the prior (lower panel)& posterior (lower pannel) Nino3.4 index and Global Ocean Data Assimilation System data set Nino3.4 index in 1-month, 3-month, 6-month and 12-month experiment. Red bars show deep learning (DL) model result, blue bars are linear inverse model (LIM) result, the purple bars are the LIMOsst model result, and the gray bars are the offline method result. The bars with diagonal hatching (right *y*-axis) show the improvement ratio of the DL model compared to the LIM, LIMOsst model and offline method.

It is noteworthy that in the equatorial western Pacific (near 135°E), DL forecast skill for SST and ocean temperature fields decreases with lead time. This appears to be a consequence of a systematic bias due to the exaggerated westward extension of the equatorial cold tongue in the CMIP6 models (Beobide-Arsuaga et al., 2021; Jiang et al., 2021; Zhou & Zhang, 2023). The LIM, trained exclusively on GFDL-CM4 data, demonstrates a mitigated version of this bias, owing to the relatively minor extent of the issue in GFDL-CM4 simulations. Conversely, the DL model, informed by a wider array of CMIP6 historical outputs, tends to accentuate the cold-tongue bias.

The strength of our DL forecasts partly stems from the use of multi-time input and auto-regression outputs. Unlike recurrent neural network (RNN) models (Medsker & Jain, 2001) and LIM, which incorporate outputs back to inputs, our approach ensures that each output is derived from a complete 12-month analysis of inputs, capturing the dynamic processes of the Pacific more effectively than single-instance inputs.

5. Data Assimilation Results

Assimilation experiments for a sparse network of noisy GODAS-sampled SST observations show that cycling with the DL model outperforms the others (LIM, LIMOsst, and offline) by around from 10% to 30% in reconstructing the Nino3.4 index (Figure 7, top panels). Similar results are found for skill in the prior forecast before DA (Figure 7, bottom panels). Improvement using the DL model increases with observation averaging time, which we attribute to the increase of forecast skill with lead time shown in Section 4. We find that these results are not sensitive to the SNR (see Figure S2 in Supporting Information S1). In terms of skill across the entire domain, the DL model outperforms the LIMs for all observation averaging times and variables, most notably for correlation, and less so for RMSE Figure 8. Two key factors contribute to the modest improvement in the RMSE metrics. First, the enhancement in forecast skill in terms of RMSE is limited, as demonstrated in Figure 4. The RMSE improvement for SST, upper ocean temperature, and wind stress is relatively smaller compared to the improvement in correlation, as discussed in Section 4. Secondly, as discussed previously, the LIM system is constrained by the fluctuation–dissipation relationship (Equation 4), which limits RMSE growth; the DL model does not have such constraint. Another possible contribution is the noise we have introduced to manage the loss of forecast variance and the limited number of ensemble members, since the DL model results improve modestly for larger ensembles (see Figure S3 in Supporting Information S1).

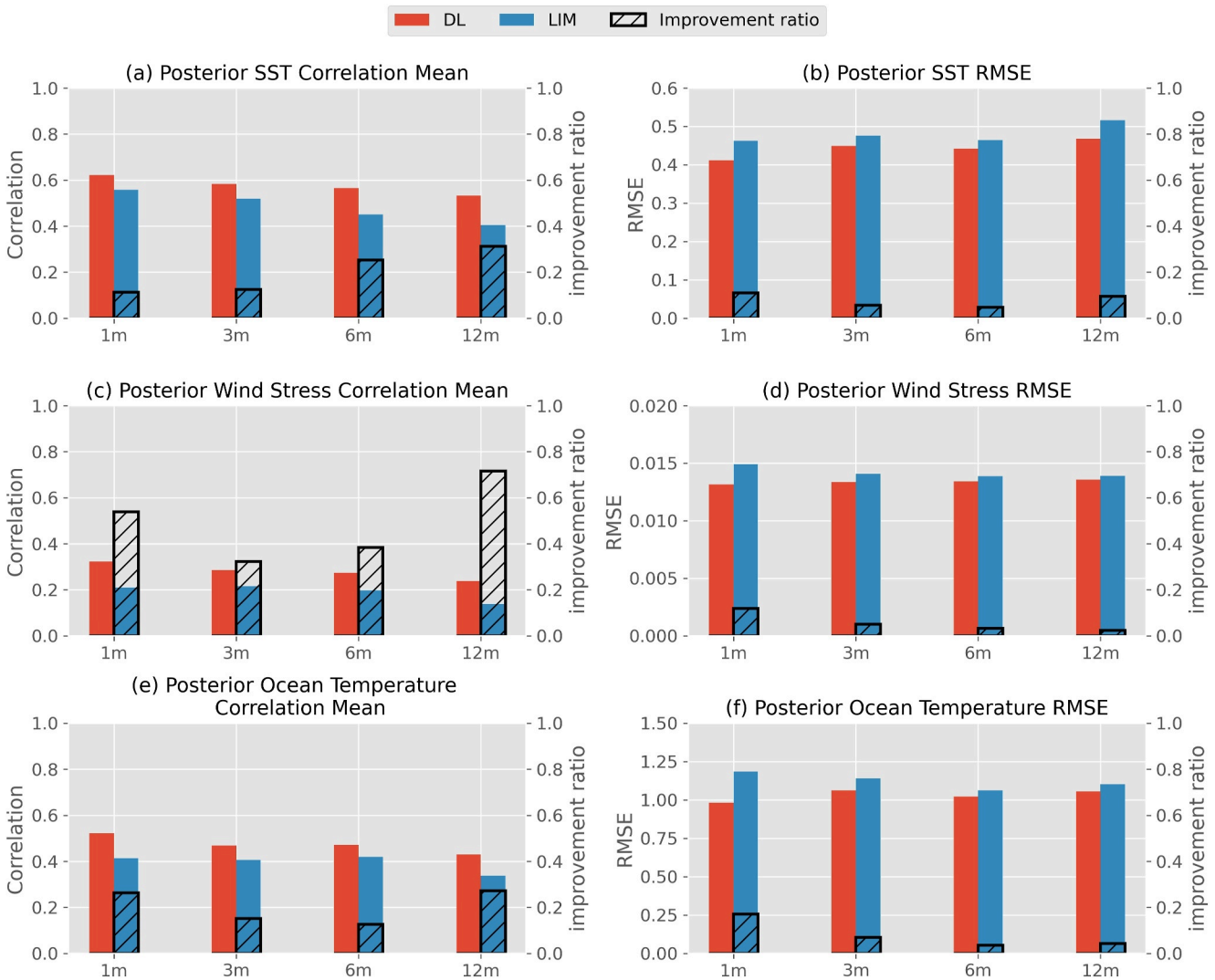

Figure 8. Domain-averaged reconstructed (posterior) correlation and Root Mean Squared Error results for the Deep-learning model & linear inverse model (LIM) verified on the Global Ocean Data Assimilation System data set. Red bars represent the Deep Learning (DL) model, blue bars the LIM (LIM); bars with diagonal hatching indicate the improvement ratio of the results for the DL model relative to those for the LIM. The upper panel shows the sea-surface temperature field, the middle panel shows the wind stress field, and the lower panel shows the ocean temperature field.

The reconstruction correlation and RMSE spatial differences between DL and LIM is shown in the Figures 9 and 10, respectively. Improvement of the DL results over the LIM in zonal wind stress and SST are located primarily off the equator, which is not the same as for the forecasting experiments (cf. Figures 5 and 6). This suggests that, relative to the LIM, the DL covariance estimates allow for more information extraction from the observations, which are located primarily closer to the equator. Improvements in the meridional wind stress are more closely confined to the equator relative to the forecasting experiments, suggesting a local influence of the observations. For the ocean temperature field, the reconstruction skill improvement of the DL model predominantly manifests in the mid-Pacific region at a depth of 100–150 m, with extensions upward and eastward along the sloping thermocline. This spatial concentration roughly corresponds with the forecast skill improvements, especially for the 12-month observation averaging time. Areas where the DL model results are worse than the LIM are concentrated near South America and the land areas around the western Pacific warm pool, which corresponds with the forecast skill differences in RMSE and correlation (Figures 5 and 6), but these areas are smaller in magnitude compared with skill enhancements elsewhere.

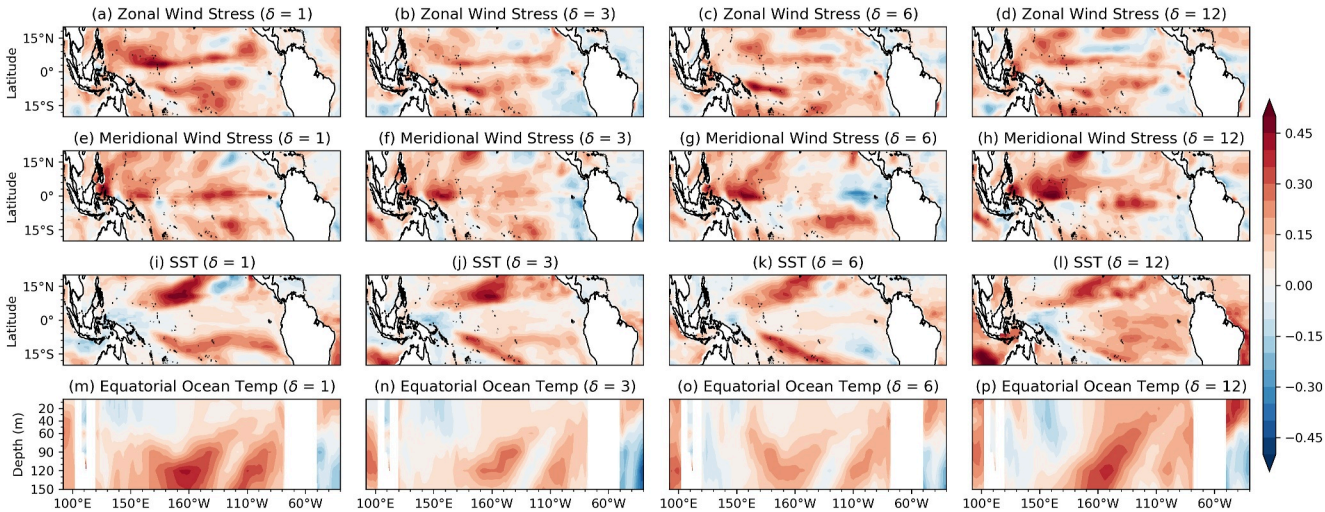

Figure 9. The reconstructed result correlation skill difference between Deep-learning model and Linear Inverse Model of δ–month experiment in zonal wind stress (a–d), meridional wind stress (e–h), sea-surface temperature (i–l) and equatorial ocean temperature (from 5°N to 5°S) (m–p) field in 1,3,6 and 12 months-averaged experiments.

Figure 11 presents a comparison of the evolution of SST and zonal wind stress in the GODAS reanalysis and the DL DA results. The reconstruction achieves remarkable accuracy, faithfully capturing the peaks and troughs of central-east Pacific SST as observed in the reanalysis, along with the corresponding zonal-wind stress. This is particularly notable in the 12-month averaged experiment, which utilizes only 24 observations per year, yet still largely captures the observed patterns. Notable differences include excessive easterly wind stress, and cooler SSTs, around 1996 and 2000, in the DL results when compared to GODAS.

6. Conclusion and Discussion

We have evaluated the potential of using a deep-learning model for cycling DA on sparse observations to reconstruct the upper ocean and surface wind stress of the tropical Pacific ocean. The DL model is trained on CMIP6 model data following Zhou and Zhang (2023), and validated by forecasting on SODA reanalysis data. A significant drawback of the DL model for DA is a bias for small forecast error variance. Therefore, we employed a

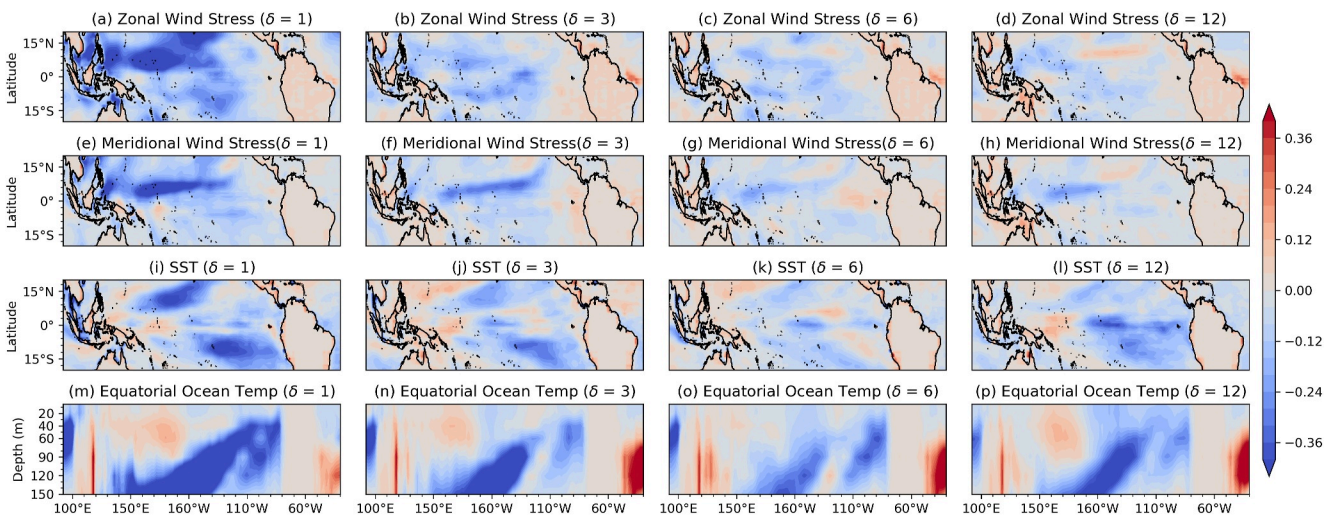

Figure 10. The reconstructed, normalized Root Mean Squared Error (normalization by the domain-averaged standard deviation of the corresponding variables) skill difference between Deep-learning model and Linear Inverse Model of δ–month experiment in zonal wind stress (a–d), meridional wind stress (e–h), sea-surface temperature (i–l) and equatorial ocean temperature (from 5°N to 5°S) (m–p) field in 1,3,6 and 12-month experiment.

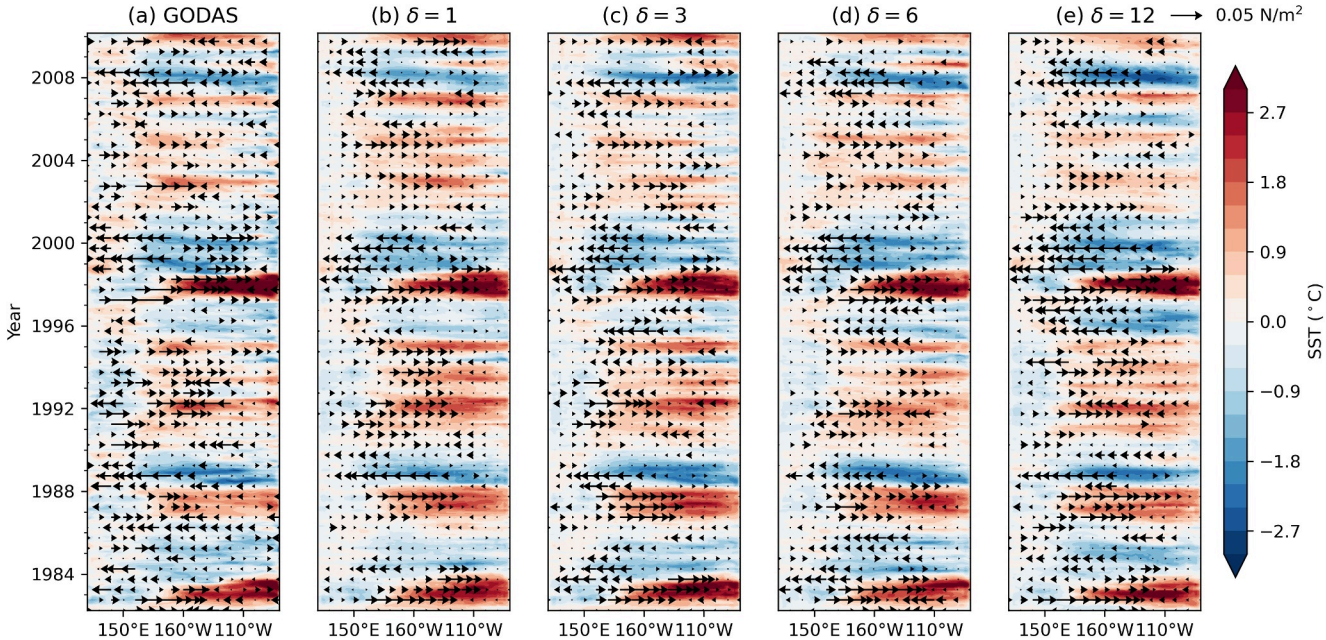

Figure 11. The equatorial (5°N–5°S) evolution of sea-surface temperature (shading) and zonal wind stress anomalies (vectors) during 1980–2010 in δ–month experiment. (a) Global Ocean Data Assimilation System data set. (b–e) DL-model reconstruction for experiments assimilating observations averaged over 1, 3, 6, and 12 months.

approach to restore ensemble forecast variance by adding scaled samples from a library of DL model forecast errors to the DL model forecasts. We compare the performance of the DL model in forecasting and DA experiments to control experiments using LIMs trained on the same CMIP6 data set, and an offline DA experiment that samples only from CMIP6 without a forecast model.

Overall, the results show that the DL model provides better forecasts compared to the LIM, especially in the central and eastern Pacific where ENSO dominates variability. The DL model outperformance is most notable in correlation (signal timing), and smaller in RMSE. For the DA experiments, the results also show improvement using the DL model relative to the LIM, but the spatial distribution of these improvements differs from the forecasting results. In particular, relative to the LIM DA results, we find larger improvements off-equator in zonal-wind stress and SST, near the equator in ocean temperatures, and near the thermocline in the mid Pacific. These improvements reflect a combination of the forecast-skill improvements, which better retain the memory of past observations, and improved spatial covariance, which spread information from the sparse network of observations, which are more abundant over the equatorial western Pacific.

Based on this proof-of-concept study, we conclude that a deep-learning model can provide computationally efficient forecast priors for online paleoclimate DA, leading to improved reconstruction outcomes. Future research will consider the application of these models to assimilating real proxy data, and extending the approach outside the tropics.

Data Availability Statement

GODAS data set can be found: https://psl.noaa.gov/data/gridded/data.godas.html. SODA data set can be found: http://www.soda.umd.edu. CMIP6 Data set can be found: https://pcmdi.llnl.gov/CMIP6/. Plotting tools can be found in SACPY: https://zenodo.org/records/13227070 (Meng et al., 2023). The main codes can be found: https://zenodo.org/records/13896365.

Acknowledgments
We thank Eric J. Steig and Zhanxiang Hua from the University of Washington, Ronghua Zhang from Nanjing University of Information Science and Technology (NUIST), and Chris Snyder from the National Center for Atmospheric Research (NCAR) for valuable suggestions and conversations related to this work. We gratefully acknowledge constructive and insightful suggestions from two anonymous reviewers, which improved the clarity of the original manuscript. The Zhou & Zhang, 2023 data set (https://zenodo.org/records/7445611), provided the source code and training data for the DL model trained in this research. We would like to acknowledge high-performance computing support (https://doi.org/10.5065/qx9a-pg09) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. This research was supported by NSF awards 2202526 and 2105805, and Heising-Simons Foundation award 2023-4715.

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167–3211. https://doi.org/10.1029/2019MS001726

Ashok, K., Behera, S. K., Rao, S. A., Weng, H., & Yamagata, T. (2007). El Niño Modoki and its possible teleconnection. *Journal of Geophysical Research*, *112*(C11), C11007. https://doi.org/10.1029/2006JC003798

Beobide-Arsuaga, G., Bayr, T., Reintges, A., & Latif, M. (2021). Uncertainty of ENSO-amplitude projections in CMIP5 and CMIP6 models. *Climate Dynamics*, *56*(11), 3875–3888. https://doi.org/10.1007/s00382-021-05673-4

Brown, J. R., Brierley, C. M., An, S.-I., Guarino, M.-V., Stevenson, S., Williams, C. J., et al. (2020). Comparison of past and future simulations of ENSO in CMIP5/PMIP3 and CMIP6/PMIP4 models. *Climate of the Past*, *16*(5), 1777–1805. https://doi.org/10.5194/cp-16-1777-2020

Cane, M. A., Zebiak, S. E., & Dolan, S. C. (1986). Experimental forecasts of el niño. *Nature*, *321*(6073), 827–832. https://doi.org/10.1038/321827a0

Carton, J. A., & Giese, B. S. (2008). A reanalysis of ocean climate using simple ocean data assimilation (soda). *Monthly Weather Review*, *136*(8), 2999–3017. https://doi.org/10.1175/2007MWR1978.1

D'Arrigo, R., Cook, E. R., Wilson, R. J., Allan, R., & Mann, M. E. (2005). On the variability of ENSO over the past six centuries. *Geophysical Research Letters*, *32*(3), 2004GL022055. https://doi.org/10.1029/2004GL022055

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

Evensen, G. (2009). *Data assimilation: The ensemble Kalman filter*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03711-5

Gao, C., Zhou, L., & Zhang, R.-H. (2023). A transformer-based deep learning model for successful predictions of the 2021 second-year La Niña condition. *Geophysical Research Letters*, *50*(12), e2023GL104034. https://doi.org/10.1029/2023gl104034

Goosse, H., Crespin, E., De Montety, A., Mann, M. E., Renssen, H., & Timmermann, A. (2010). Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation. *Journal of Geophysical Research*, *115*(D9), 2009JD012737. https://doi.org/10.1029/2009JD012737

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., et al. (2016). The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres*, *121*(12), 6745–6764. https://doi.org/10.1002/2016JD024751

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(9), 1119–1152. https://doi.org/10.1002/joc.1499

Huntley, H. S., & Hakim, G. J. (2010). Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model. *Climate Dynamics*, *35*(6), 995–1009. https://doi.org/10.1007/s00382-009-0714-5

Jiang, W., Huang, P., Huang, G., & Ying, J. (2021). Origins of the excessive westward extension of ENSO SST simulated in CMIP5 and CMIP6 models. *Journal of Climate*, *34*(8), 2839–2851. https://doi.org/10.1175/JCLI-D-20-0551.1

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kug, J.-S., Jin, F.-F., & An, S.-I. (2009). Two types of El Niño events: Cold tongue El Niño and warm pool El Niño. *Journal of Climate*, *22*(6), 1499–1515. https://doi.org/10.1175/2008JCLI2624.1

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lorenz, E. N. (1956). *Empirical orthogonal functions and statistical weather prediction* (Vol. 1). Massachusetts Institute of Technology, Department of Meteorology Cambridge.

Lou, J., O'Kane, T. J., & Holbrook, N. J. (2020). A linear inverse model of tropical and South Pacific seasonal predictability. *Journal of Climate*, *33*(11), 4537–4554. https://doi.org/10.1175/jcli-d-19-0548.1

Mann, M. E., Bradley, R. S., & Hughes, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, *392*(6678), 779–787. https://doi.org/10.1038/33859

McPhaden, M. J., Zebiak, S. E., & Glantz, M. H. (2006). Enso as an integrating concept in earth science. *Science*, *314*(5806), 1740–1745. https://doi.org/10.1126/science.1132588

Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*(64–67), 2.

Meng, Z., & Li, T. (2024). Why is the pacific meridional mode most pronounced in boreal spring? *Climate Dynamics*, *62*(1), 459–471. https://doi.org/10.1007/s00382-023-06914-4

Meng, Z., Zhu, F., & Hakim, G. J. (2023). Sacpy-a python package for statistical analysis of climate. *Agu fall meeting abstracts*, *2023*, PP13D–1252.

Newman, M. (2013). An empirical benchmark for decadal forecasts of global surface temperature anomalies. *Journal of Climate*, *26*(14), 5260–5269. https://doi.org/10.1175/JCLI-D-12-00590.1

Newman, M., Alexander, M. A., & Scott, J. D. (2011). An empirical model of tropical ocean dynamics. *Climate Dynamics*, *37*(9–10), 1823–1841. https://doi.org/10.1007/s00382-011-1034-0

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

PAGES2k Consortium, Emile-Geay, J., McKay, N. P., Kaufman, D. S., Von Gunten, L., Wang, J., et al. (2017). A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data*, *4*(1), 170088. https://doi.org/10.1038/sdata.2017.88

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Penland, C., & Magorian, T. (1993). Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *Journal of Climate*, *6*(6), 1067–1076. https://doi.org/10.1175/1520-0442(1993)006⟨1067:PONSST⟩2.0.CO;2

Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, *8*(8), 1999–2024. https://doi.org/10.1175/1520-0442(1995)008<1999:togots>2.0.co;2

Perkins, W. A., & Hakim, G. (2020). Linear inverse modeling for coupled atmosphere-ocean ensemble climate prediction. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001778. https://doi.org/10.1029/2019MS001778

Perkins, W. A., & Hakim, G. J. (2017). Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. *Climate of the Past*, *13*(5), 421–436. https://doi.org/10.5194/cp-13-421-2017

Prechelt, L. (2002). Early stopping-but when? In *Neural networks: Tricks of the trade* (pp. 55–69). Springer.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., & Yamagata, T. (1999). A dipole mode in the tropical Indian ocean. *Nature*, *401*(6751), 360–363. https://doi.org/10.1038/43854

Shin, J., Park, S., Shin, S.-I., Newman, M., & Alexander, M. A. (2020). Enhancing ENSO prediction skill by combining model-analog and linear inverse models (MA-LIM). *Geophysical Research Letters*, *47*(1), e2019GL085914. https://doi.org/10.1029/2019gl085914

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., & Roe, G. H. (2014). Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, *27*(1), 426–441. https://doi.org/10.1175/JCLI-D-12-00693.1

Sun, M., Chen, L., Li, T., & Luo, J.-J. (2023). Cnn-based Enso forecasts with a focus on SSTA zonal pattern and physical interpretation. *Geophysical Research Letters*, *50*(20), e2023GL105175. https://doi.org/10.1029/2023gl105175

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., et al. (2019). Last millennium reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, *15*(4), 1251–1273. https://doi.org/10.5194/cp-15-1251-2019

Timmermann, A., An, S.-I., Kug, J.-S., Jin, F.-F., Cai, W., Capotondi, A., et al. (2018). El Niño–southern oscillation complexity. *Nature*, *559*(7715), 535–545. https://doi.org/10.1038/s41586-018-0252-6

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., & Whitaker, J. S. (2003). Ensemble square root filters. *Monthly Weather Review*, *131*(7), 1485–1490. https://doi.org/10.1175/1520-0493(2003)131⟨1485:ESRF⟩2.0.CO;2

Valler, V., Franke, J., Brugnara, Y., Samakinwa, E., Hand, R., Lundstad, E., et al. (2024). Mode-ra: A global monthly paleo-reanalysis of the modern era 1421 to 2008. *Scientific Data*, *11*(1), 36. https://doi.org/10.1038/s41597-023-02733-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Vimont, D. J., Wallace, J. M., & Battisti, D. S. (2003). The seasonal footprinting mechanism in the pacific: Implications for Enso. *Journal of Climate*, *16*(16), 2668–2675. https://doi.org/10.1175/1520-0442(2003)016⟨2668:TSFMIT⟩2.0.CO;2

Xue, Y., Balmaseda, M. A., Boyer, T., Ferry, N., Good, S., Ishikawa, I., et al. (2012). A comparative analysis of upper-ocean heat content variability from an ensemble of operational ocean reanalyses. *Journal of Climate*, *25*(20), 6905–6929. https://doi.org/10.1175/jcli-d-11-00542.1

Zhang, R.-H., Zhou, L., Gao, C., & Tao, L. (2024). A transformer-based coupled ocean-atmosphere model for ENSO studies. *Science Bulletin*, S2095–S9273.

Zhou, L., & Zhang, R.-H. (2023). A self-attention–based neural network for three-dimensional multivariate modeling and its skillful Enso predictions. *Science Advances*, *9*(10), eadf2827. https://doi.org/10.1126/sciadv.adf2827

Zhu, F., Emile-Geay, J., Anchukaitis, K. J., McKay, N. P., Stevenson, S., & Meng, Z. (2023). A pseudoproxy emulation of the pages 2K database using a hierarchy of proxy system models. *Scientific Data*, *10*(1), 624. https://doi.org/10.1038/s41597-023-02489-1